Ranking Studies for Systematic Reviews Using Query Adaptation: University of Sheffield's Approach to CLEF eHealth 2019 Task 2

Working Notes for CLEF 2019

Amal Alharbi[1,2] and Mark Stevenson[1]

[1] University of Sheffield, UK
[2] King Abdulaziz University, Saudi Arabia
{ahalharbi1,mark.stevenson}@sheffield.ac.uk

Abstract. This paper describes the University of Sheffield's approach to the CLEF 2019 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 subtask 2 (*Abstract and Title Screening*). Our approach used lexical statistics (Log-Likelihood, Chi-Squared and Odds-Ratio) to identify terms that retrieve specific types of evidence. A total of 12 official runs were submitted.

1 Introduction

Systematic reviews aim to collect, synthesise and summarise all available evidence that answers a specific research question. Medical practitioners and decision makers rely on the information they contain to guide treatment decisions.

Cochrane is one the key producers of medical systematic reviews. Its library contains 7,987 reviews[1] which fall into five categories [1]:

1. **Intervention reviews** assess the benefits and harms of interventions used in healthcare and health policy.
2. **Diagnostic test accuracy reviews (DTA)** assess the accuracy of a diagnostic test when used to detect a particular disease.
3. **Methodology reviews** explore issues about the processes associated with conducting systematic reviews and clinical trials.
4. **Qualitative reviews** address questions related to healthcare interventions other than effectiveness by synthesizing qualitative evidence.
5. **Prognosis reviews** address the probable course or future outcome(s) of people with a health problem.

[1] At the date of writing this paper May 2019

Systematic reviews are time-consuming to create, it may take up to a year to conduct a single review [7]. One of the most time consuming steps is evidence collection. The main stages in this process are: *(1) Boolean Search:* A Boolean query is created and applied to a medical database, such as MEDLINE, to retrieve a set of candidate citations. *(2) Title and Abstract Screening:* The title and abstract of all candidate citations returned by the Boolean query are screened to decide which ones should be considered for inclusion in the review. *(3) Content Screening:* The full text of the remaining citations are examined to determine the final set that will be included in the review [2].

CLEF eHealth 2019 Task 2 Subtask 2 [9] focuses on the second stage of evidence collection ('*Title and Abstract Screening*'). The dataset contains four of the five types of review produced by Cochrane: DTA, Intervention, Prognosis and Qualitative. Participants are asked to rank the list of PubMed Document Identifiers (PMIDs) returned from the Boolean query so that relevant citations appear as early as possible.

This paper is structured as follows: Section 2 describes the datasets and approaches used, Section 3 the experiments conducted and Section 4 states and discusses the results obtained.

2 Method

2.1 Datasets

CLEF2019 dataset is partitioned into training and testing datasets. The training dataset contains two types of review (72 DTA and 20 Intervention) while the test dataset contains four types (eight DTA, 20 Intervention, two Qualitative and one Prognosis). For each review participants are provided with the review title, Boolean query, set of PMIDs and relevance judgements for both title/abstract and content level screening.

2.2 University of Sheffield's Approach to Subtask 2

Sheffield's submission extended the approach that had been developed for our previous entries to the task. The core of our approach extracts terms from the Boolean query and uses them to rank the studies [3]. In addition, these terms are augmented with additional ones designed to identify the DTA reviews that formed the majority of the studies in previous editions of the task (e.g. 'sensitivity', 'specificity' and 'diagnosis') [2].

Our submissions to the 2019 task extended this approach to multiple review types by developing lists of terms that indicate the relevant evidence for a specific review type. Lexical statistics were used to automatically derive these lists of key terms. Three lexical statistics were applied: Log-Likelihood, Chi-Squared and Odds-Ratio [4,6,12,8,10,11]. These lexical statistics are computed using a contingency table created for each term (see Table 1). This table assumes that the collection is partitioned into relevant and irrelevant documents and encodes

information about the frequency with which the term appears in each. For example, O_{rel} represents the number of times the term occurs within the entire set of relevant documents and N_{rel} the sum of the occurrences of all terms.

Table 1: Contingency table for computing lexical statistics.

	Relevant	Irrelevant
Frequency of term	O_{rel}	O_{irrel}
Total tokens	N_{rel}	N_{irrel}

Log-Likelihood is computed as

$$\textit{Log-Likelihood} = 2 \times \left(O_{rel} \times \log \frac{O_{rel}}{E_{rel}} + O_{irrel} \times \log \frac{O_{irrel}}{E_{irrel}} \right) \tag{1}$$

where O_{rel} and O_{irrel} are the observed frequency of the term in different subsets of the collection (e.g. relevant and irrelevant documents). E_{rel} and E_{irrel} are the term's expected frequencies, calculated as

$$E_{rel} = N_{rel} \times \frac{O_{rel} + O_{irrel}}{N_{rel} + N_{irrel}} \quad , \quad E_{irrel} = N_{irrel} \times \frac{O_{rel} + O_{irrel}}{N_{rel} + N_{irrel}} \tag{2}$$

where N_{rel} and N_{irrel} represent sub-corpus size (e.g. relevant and irrelevant documents). Terms are assigned high Log-Likelihood scores for a particular corpus when their observed frequency is (much) higher than the expected frequency.

Chi-Squared is computed as

$$\textit{Chi-Squared} = \frac{(O_{rel} - E_{rel})^2}{E_{rel}} + \frac{(O_{irrel} - E_{irrel})^2}{E_{irrel}} \tag{3}$$

where O_{rel} and O_{irrel} are the observed values and E_{rel} and E_{irrel} are expected values calculated using equation 2.

Odds-Ratio is computed as

$$\textit{Odds-Ratio} = \frac{O_{rel} \times (N_{irrel} - O_{irrel})}{O_{irrel} \times (N_{rel} - O_{rel})} \tag{4}$$

where O_{rel} and O_{irrel} are the frequency counts of the term in the relevant and irrelevant sub-corpus and N_{rel} and N_{irrel} are the total number of terms in each of those sub-corpus.

3 Experiments

Four official runs were submitted for DTA and Intervention reviews: *sheffield-baseline, sheffield-Log_Likelihood, sheffield-Chi_Squared* and *sheffield-Odds_Ratio*. Two official runs were submitted for Prognosis and Qualitative reviews: *sheffield-baseline* and *sheffield-relevance_feedback*.

3.1 sheffield-baseline

A baseline query was formed using the review title and terms extracted from the Boolean query. Studies were ranked using this query and BM25 [5]. This approach was applied to all reviews types and was also used in our submissions to previous editions of the task [2,3].

3.2 sheffield-Log_Likelihood, sheffield-Chi_Squared, sheffield-Odds_Ratio

Training data was available for two review types (DTA and Intervention). Where this is available lexical statistics were applied to derive a list of terms that indicate evidence relevant to a specific type of review.

Studies in the training dataset were partitioned into relevant and irrelevant sets depending upon whether they were included in the systematic review. The three lexical statistics described in Section 2 were calculated and the terms with the highest scores added to the baseline query. The number of terms added was determined from experiments conducted using the training data[2]. The studies in the test dataset are ranked by matching terms from the expanded queries against those in the abstracts using BM25. Note that sets of additional terms were generated for each review type separately, i.e. once for DTA reviews and again for Intervention reviews.

3.3 sheffield-relevance_feedback

No training data was provided for Prognosis and Qualitative reviews. Consequently it was not possible to apply the lexical statistics and a relevance feedback approach was used instead. Studies in the test dataset are ranked using BM25 and the top 5% extracted. The Chi-Squared statistic was then applied using relevance judgements to divide the studies into relevant and irrelevant sets. The top 20 terms were added to the query which is then used to re-rank the remaining 95% of the studies.

[2] For DTA reviews, 10 terms were added when the Log-Likelihood and Chi-Squared lexical statistics were used and 50 when Odds-Ratio was used. For Intervention reviews, 20 terms were added for the Log-Likelihood statistic, 5 for Chi-Squared and 50 for Odds-Ratio.

4 Results and Discussion

Table 2 shows the results for DTA reviews (computed using the script provided by the task organisers[3]). All three lexical statistics outperform the baseline, as expected. This improvement is consistent across all metrics for both abstract and content level screening. The best result was achieved by applying Odds-Ratio. Results demonstrate that expanding query with terms generated to identify DTA studies helps improve performance.

Table 2: Performance ranking abstracts for **DTA reviews** at (a) abstract and (b) content levels.

Approach	**MAP**	**WSS@100**	**WSS@95**
(a) abstract level			
sheffield-baseline	0.175	33.80%	45.10%
sheffield-Log_Likelihood	0.234	**38.10%**	48.70%
sheffield-Chi_Squared	0.222	37.50%	47.50%
sheffield-Odds_Ratio	**0.248**	34.70%	**49.00%**
(b) content level			
sheffield-baseline	0.066	51.90%	55.30%
sheffield-Log_Likelihood	0.120	**56.70%**	58.80%
sheffield-Chi_Squared	0.113	54.10%	58.30%
sheffield-Odds_Ratio	**0.129**	54.90%	**63.50%**

Results for Intervention reviews are shown in Table 3. Log-Likelihood performs strongly, with the best results for all metrics using content level judgements and using MAP for abstract level judgements. However, it is noteworthy that the baseline approach achieves the best performance using the WSS@95 metric and abstract level judgements.

Tables 4 and 5 show the results produced by applying the baseline and relevance-feedback approaches to the Qualitative and Prognosis reviews, respectively. The use of relevance feedback produced a slight improvement in the results, particularly MAP. The modest level of improvement may be down to the small number of relevant studies found in the top 5% of the ranked documents. For example, for the Qualitative review (CD011558) only two of the 2,168 (0.09%) studies are relevant.

[3] `https://github.com/leifos/tar`

Table 3: Performance ranking abstracts for **Intervention reviews** at (a) abstract and (b) content levels.

Approach	MAP	WSS@100	WSS@95
(a) abstract level			
sheffield-baseline	0.245	38.60%	**47.00%**
sheffield-Log_Likelihood	**0.293**	38.10%	45.80%
sheffield-Chi_Squared	0.262	**41.50%**	46.90%
sheffield-Odds_Ratio	0.261	38.40%	46.20%
(b) content level			
sheffield-baseline	0.185	49.80%	50.00%
sheffield-Log_Likelihood	**0.272**	**57.90%**	**56.80%**
sheffield-Chi_Squared	0.223	51.70%	52.80%
sheffield-Odds_Ratio	0.254	53.60%	54.20%

Table 4: Performance ranking abstracts for **Qualitative reviews** using baseline and relevance feedback at (a) abstract and (b) content levels.

Approach	MAP	WSS@100	WSS@95
(a) abstract level			
sheffield-baseline	0.051	8.20%	13.50%
sheffield-relevance_feedback	**0.060**	**10.30%**	**18.50%**
(b) content level			
sheffield-baseline	0.035	5.40%	30.10%
sheffield-relevance_feedback	**0.041**	**36.00%**	**35.70%**

Table 5: Performance ranking abstracts for **Prognosis review** using baseline and relevance feedback at (a) abstract and (b) content levels.

Approach	MAP	WSS@100	WSS@95
(a) abstract level			
sheffield-baseline	0.126	11.20%	24.70%
sheffield-relevance_feedback	**0.141**	**17.60%**	**30.50%**
(b) content level			
sheffield-baseline	0.077	11.20%	27.90%
sheffield-relevance_feedback	**0.086**	**18.70%**	**36.70%**

5 Conclusions

This paper presented the University of Sheffield's approach to CLEF2019 task 2 subtask 2. Studies were ranked by supplementing terms extracted from the Boolean query with ones specific to the review type. Three lexical statistics were used to generate these list of supplementary terms. Results demonstrated that adding these additional terms improved performance although there was no clear picture of which lexical statistics was most effective.

Bibliography

1. About Cochrane Reviews — Cochrane Library. https://www.cochranelibrary.com/about/about-cochrane-reviews, accessed: 2019-05-1
2. Alharbi, A., Briggs, W., Stevenson, M.: Retrieving and ranking studies for systematic reviews: University of sheffield's approach to clef ehealth 2018 task 2. In: CLEF 2018 Labs Working Notes. Avignon, France (2018)
3. Alharbi, A., Stevenson, M.: 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, CEUR-WS.org, Dublin, Ireland (September 11-14 2017)
4. Alharbi, A., Stevenson, M.: Improving ranking for systematic reviews using query adaptation. In: Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019). Springer, Lugano, Switzerland (2019)
5. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology Behind Search. Addison-Wesley Publishing Company, USA, 2nd edn. (2011)
6. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational linguistics **19**(1), 61–74 (1993)
7. Ganann, R., Ciliska, D., Thomas, H.: Expediting systematic reviews: methods and implications of rapid reviews. Implementation science : IS **5**, 56 (jul 2010). https://doi.org/10.1186/1748-5908-5-56
8. Gelbukh, A., Sidorov, G., Lavin-Villa, E., Chanona-Hernandez, L.: Automatic term extraction using log-likelihood based comparison with general reference corpus. In: Hopfe, C.J., Rezgui, Y., Métais, E., Preece, A., Li, H. (eds.) Natural Language Processing and Information Systems. pp. 248–255. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
9. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In: CLEF 2019 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2019)
10. Oakes, M., Farrow, M.: Use of the chi-squared test to examine vocabulary differences in english language corpora representing seven different countries. Literary and Linguistic Computing **22**(1), 85–99 (2007)
11. Pojanapunya, P., Todd, R.W.: Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Ling. Theory **14**(1), 133–167 (2018)
12. Rayson, P.: From key words to key semantic domains. International Journal of Corpus Linguistics **13**(4), 519–549 (2008)