SiS at CLEF 2017 eHealth TAR Task

Vassil Kalphov[1], Georgios Georgiadis[1], Leif Azzopardi[1]
vassil.kalphov.2013@uni.strath.ac.uk,
georgios.georgiadis.2013@uni.strath.ac.uk, and
Leif.Azzopardi@strath.ac.uk

University of Strathclyde, Glasgow, UK

Abstract. This paper presents Strathclyde iSchool's (SiS) participation in the Technological Assisted Reviews in Empirical Medicine Task. For the ranking task, we explored two ways in which assistance to reviewers could be provided during the assessment process: (i) topic models, where we use Latent Dirichlet Allocation to identify topics within the set of retrieved documents, ranking documents by the topic most likely to be relevant and (ii) relevance feedback, where we use Rocchio's algorithm to update the query model for subsequent rounds of interaction. A third approach combines the topic and relevance feedback to quickly identify the relevant abstracts. For the thresholding task, we apply a score threshold, and exclude documents which did not exceed the threshold given BM25.

1 Introduction

CLEF 2017 introduced a new eHealth retrieval problem - that of providing technological assistance to reviewers of systematic reviews - where the goals of the task were to explore how Information Retrieval techniques could be used to: (i) identify relevant material more quickly in the ranking challenge and (ii) identify when reviewers could stop processing documents in the thresholding challenge [3, 2]. During the review process, reviewers will routinely examine hundreds to thousands of abstracts to decided if the document (and evidence it contains) could be included in the systematic review that they are conducting [5]. Once they have identified a subset of abstracts, which are potentially relevant, they examine the document's contents to decide whether the document should be included or excluded. The track focused on the first part, identifying potential relevant documents during the, so called, screening phase.

In this work we considered two different approaches - one which uses topic modelling and the other which uses relevance feedback. In selecting these approaches we thought that such techniques could be used in the following way. For topic modelling, we envisaged that the download abstracts could be semantically clustered - and the different clusters could be presented to the reviewer - the reviewer could then start the review process by selecting a cluster that they felt was most likely to contain the relevant documents. Since we did not have recourse to reviewers, we explored a number of different ways to automatically select the best cluster. For relevance feedback, we envisaged that as the reviewer

starts to examine documents, the query could be updated to bring back the next most relevant documents, so that they would quickly find all the relevant material as soon as possible. Obviously, if the aim is to reduce the workload of the reviewers, then we need to be able to select a point where the reviewer can stop assessing documents - however, this runs the risk of losing relevant documents. To this end, we explore various heuristics to select the threshold such that we minimize effort and maximise recall (but ideally obtain total recall).

2 Experimental Set-up

Data: Given the list of topic descriptions the PubMed IDs were extracted, and a scripted fetched the Abstract and associated Metadata from the PubMed API. From the topics, we extracted the title for each topic and use that as the query.
Indexing and Retrieval System: We used Lucene 6.2 to create a separate index for each of the topic (where stop words were removed, no stemming was applied). A Lucene Document was created where the following fields were index: Title, Abstract, Author, and Publication Name. The baseline retrieval algorithm we employed was fielded BM25 with standard parameters settings i.e. $b = 0.75$, and equal weights between fields (denoted as **BM25**).
Relevance Feedback: We implemented Rocchio's Algorithm [4] in Lucene - where feedback was used to provide relevance information. In each round of feedback, 30 documents were examined, and the query model updated, to provide a re-ranking of the subsequent documents. This was performed on the first 10%, 20%, etc of documents associated with the topic. Here, we only report the 30% runs (AL30) as these generally performed the best and little change in performance was observed on the training set using more feedback.
Topic Modelling: We used MALLET toolkit, and thus Latent Dirichlet Allocation [1] to semantically cluster the documents within each topic. We set the number of latent topics to 5, and $\alpha =$. To rank the documents we selected one of the latent topics z_i, and ordered the documents by the probability $p(z_i|d|)$. In an attempt to select the cluster that provides the best ranking, we took the **BM25** ranking from above, and use the top 100 ranked documents as pseudo-relevance feedback. Then we ranked each latent topic (given the ordering by $p(z_i|d|)$) and select the one with the highest overlap with BM25 (**TMBM**).
Combined: Since our topic modelling approaches do not use feedback, we decided to see whether we could start the process of relevance feedback using the topic modelling run, and refine the query model accordingly. Thus, we selected the best performing method **TMAL**, given the training data, and used this as starting point for the active learning. Again we considered obtained feedback for the first 10%,20% and 30% of the documents per topic.
Thresholded Runs To create thresholded runs, we took the BM25 run and applied a simple score based threshold. Using Lucene the scores for a query given BM25 are from 1.0 or greater, so we used thresholds 1.0, 1.5, 2.0 and 2.5. This led to a reasonable reduction in the number of documents without sacrificing much recall.

3 Results and Discussion

Tables 1 and 3 report the performance on the *Ranking Task*, while Tables 2 and 4 report the performance on the *Threshold Task*. Our best performing run on the Ranking Task, in terms of normalized area under the gain curve, was *AL30*, which used 30% of the documents as feedback. From our results, it is clear that the topic modelling approach, as we have employed it, has not lead to significantly better improvements over the BM25 baseline. However, when inspecting the individual topic based ranking, the most probably topic, was not always the best performing topic. So we will direct more research into topic selection. This is because, when formulating queries, reviewers could use topic modelling to understand the space of documents retrieved, and then refine their query further - and thus make savings *a priori* rather than have to trawl through hundreds and thousands of documents. Another factor that could significantly improve our results is that for these initial runs we used the title as the query, as opposed to the boolean query provided within the topics. It is quite possible that the more complex and verbose boolean queries could lead to a better initial ranking - and so when used in conjunction with relevance feedback the relevant items could be found sooner. We leave these directions for further work.

Table 1: Results on Training Data for Ranking Task

Run	**Random**	**BM25**	**AL30**	**TMBM**	**TMAL30**
NumRels	2494	2494	2494	2494	2494
NumFeed	0	0	44939	0	44940
RelsFound	2494	2494	2494	2494	2494
AP	0.044	0.171	0.273	0.15	0.24
MinRel	6947	5174	3210	4221	2835
WSS100	0.05	0.25	0.33	0.30	0.37
Area	0.51	0.83	0.91	0.80	0.90
NCG10	0.07	0.30	0.41	0.30	0.41
NCG20	0.18	0.51	0.73	0.57	0.73
NCG30	0.29	0.67	0.86	0.74	0.86
TotalCost	7453	7453	11947	7453	11947
TotalCostUniform	7453	7453	11947	7453	11947
TotalCostWeighted	7453	7453	11947	7453	11947
loss_er	0.38	0.38	0.38	0.38	0.38
loss_r	0.0	0.0	0.0	0.0	0.0
loss_e	0.38	0.38	0.38	0.38	0.38

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)

Table 2: Results on Training Data for Threshold Task

Run	**BM25**	**T1**	**T1.5**	**T2**	**T2.5**
NumRels	2494	2949	2494	2494	2494
NumFeed	0	0	0		0
RelsFound	2494	2439	2419	2318	2187
AP	0.17	0.171	0.17	0.17	0.17
MinRel	5174	4531	3930	3453	3236
WSS100	0.25	0.23	0.22	0.19	0.19
Area	0.83	0.83	0.83	0.82	0.80
NCG10	0.30	0.30	0.30	0.30	0.30
NCG20	0.51	0.51	0.51	0.51	0.51
NCG30	0.67	0.67	0.67	0.67	0.67
TotalCost	7453	6010	5391	4700	4144
TotalCostWeighted	7453	6811	6978	6911	6985
TotalCostUniform	7453	6049	5482	5008	4714
loss_er	0.38	0.30	0.24	0.20	0.16
loss_r	0.0	0.0	0.001	0.006	0.016
loss_e	0.38	0.29	0.24	0.19	0.15

Table 3: Results on Test Data for Ranking Task

Run	**Random**	**BM25**	**AL30**	**TMBM**	**TMAL30**
NumRels	1857	1857	1857	1857	1857
NumFeed	0	0	35730	0	35432
RelsFound	1857	1857	1857	1857	1857
AP	0.05	0.17	0.22	0.12	0.16
MinRel	3722	2851	2290	3124	2305
WSS100	0.04	0.29	0.41	0.27	0.40
Area	0.48	0.81	0.86	0.73	0.84
NCG10	0.09	0.45	0.62	0.31	0.54
NCG20	0.19	0.65	0.79	0.55	0.77
NCG30	0.28	0.75	0.88	0.68	0.86
TotalCost	3918	3918	6300	3918	6280
TotalCostUniform	3918	3918	6300	3918	6280
TotalCostWeighted	3918	3918	6300	3918	6280
loss_er	0.54	0.54	0.54	0.54	0.54
loss_r	0.00	0.00	0.00	0.00	0.00
loss_e	0.54	0.54	0.54	0.54	0.54

2. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth evaluation lab overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer (September 2017)
3. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: echnologically assisted reviews in empirical medicine. In: CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2017)
4. Rocchio, J.J.: Relevance feedback in information retrieval (1971)

Table 4: Results on Test Data for Threshold Task

Run	**BM25**	**T1**	**T1.5**	**T2**	**T2.5**
NumRels	1857	1857	1857	1857	1857
NumFeed	0	0	0	0	0
RelsFound	1857	1828	1809	1784	1758
AP	0.17	0.17	0.17	0.17	0.17
MinRel	2851	2503	2333	2068	1877
WSS100	0.29	28	0.27	0.23	0.22
Area	0.81	0.81	0.80	0.80	0.79
NCG10	0.45	0.45	0.45	0.45	0.45
NCG20	0.65	0.65	0.65	0.65	0.65
NCG30	0.75	0.75	0.75	0.75	0.75
TotalCost	3918	3435	3165	2824	2536
TotalCostUniform	3918	3786	3865	3748	3902
TotalCostWeighted	3918	3454	3280	3117	2905
loss_er	0.54	0.54	0.38	0.33	0.27
loss_r	0.00	0.001	0.005	0.01	0.014
loss_e	0.54	0.54	0.38	0.32	0.26

5. Shemilt, Khan, Park, Thomas: Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Systematic reviews (2016)