QUT ielab at CLEF eHealth 2017 Technology Assisted Reviews Track: Initial Experiments with Learning To Rank

Harrisen Scells[1], Guido Zuccon[1], Anthony Deacon[1], Bevan Koopman[2]

[1] Queensland University of Technology, Brisbane, Australia
[2] Australian E-Health Research Centre, CSIRO, Brisbane, Australia
harrisen.scells@hdr.qut.edu.au, g.zuccon@qut.edu.au
aj.deacon@qut.edu.au, bevan.koopman@csiro.au

Abstract. In this paper we describe our participation to the CLEF eHealth 2017 Technology Assisted Reviews track (TAR). This track aims to evaluate and advance search technologies aimed at supporting the creation of biomedical systematic reviews. In this context, the track explores the task of screening prioritisation: the ranking of studies to be screened for inclusion in a systematic review. Our solution addresses this challenge by developing ranking strategies based on learning to rank techniques and exploiting features derived by the use of the PICO framework. PICO (Population, Intervention, Control or comparison and Outcome) is a technique used in evidence based practice to frame and answer clinical questions and is used extensively in the compilation of systematic reviews. Our experiments show that the use of the PICO-based feature within learning to rank provides improvements over the use of baseline features alone.

1 Introduction

A systematic review is a type of literature review that appraises and synthesises the work of primary research studies to answer one or more research questions. Most authors follow the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) method for conducting and reporting these reviews. This includes the definition of a formal search strategy to retrieve studies which are to be considered for inclusion in the review.

Given a research question and a set of inclusion/exclusion criteria, researchers undertaking a systematic review define a search strategy (the query) to be issued to one or more search engines that index published literature (e.g. PubMed). In medical and biomedical research, search strategies are commonly expressed as (large) boolean queries. After the search strategy has been executed, the title, and then abstract, of studies retrieved by it are reviewed in a process known as *screening*. Where the study appears relevant the full-text is then retrieved for more detailed examination.

The compilation of systematic reviews can take significant time and resources, hampering their effectiveness. Tsafnat et al. report that it can take several years to complete and publish a systematic review [4]. When systematic reviews take such significant time to complete, they can become out-of-date even at time of publishing. While the compilation of a systematic review involves several steps, one of the most time-consuming is screening. Thus, the development of IR methods that decrease the number of documents to be screened, would have a major impact on the time and resources required to undertake systematic reviews. Similarly, the ordering of studies to be screened according to the likelihood of satisfying the inclusion criteria of the systematic reviews (*screening prioritisation*) would allow relevant studies to be identified early on in the screening process, thus providing a feedback loop to improve the development of search strategies. Screening prioritisation is typically done as a two-stage process. An initial set of studies are retrieved using a boolean retrieval process; these are then ranked according to some relevance measure.

The challenge of compiling systematic reviews can be fertile ground for information retrieval (IR) research, as this can provide techniques to improve current screening and screening prioritisation processes. The CLEF eHealth 2017 Technology Assisted Reviews track (TAR) [1,2] joins our recent work [3] in devising evaluation resources for evaluation of information retrieval techniques that attempt to automate and improve processes involved in the creation of systematic reviews. The TAR track considers two tasks: (1) to produce an the efficient ordering of studies retrieved by a boolean search strategy, such that all of the relevant abstracts are retrieved as early as possible, and (2) to identify a subset of the ranked studies which contains all or as many of the relevant abstracts for the least effort (i.e. total number of abstracts to be assessed). In our submissions, we tackle the first task, and use learning to rank to produce a re-ranking of the initial set of studies retrieved for screening by the systematic review's boolean search strategy.

2 Our Approach for TAR

We trained a learning to rank model using domain specific features to provide an efficient ordering of studies retrieved by a systematic review. Specifically, we aim to observe what effect PICO features have with respect to learning to rank algorithms. PICO (Population, Intervention, Control or comparison and Outcome) is a technique used in evidence based practice to frame and answer clinical questions and is used extensively in the compilation of systematic reviews. We investigated several learning to rank algorithms and observed the effect queries annotated with PICO elements had on the reordering of results compared to the original Boolean queries.

We trained two learning to rank models using both the original queries provided by the task organiser, and another modified set of queries which contains annotations from the PICO framework. In total, we used seven features to train our learning to rank model. Table 1 summarises the features used. The first four features (IDFSum, IDFStd, IDFMax, and IDFAvg) calculate the inverse document frequency (idf) for each of the terms in the document that also appear

Id	Feature
1	IDFSum
2	IDFStd
3	IDFMax
4	IDFAvg
5	PopulationCount
6	InterventionCount
7	OutcomeCount

Table 1: Features used as training for our learning to rank model.

in the query. IDFSum sum of all idf scores, IDFStd is the standard deviation of the idf scores, IDFMax is the maximum idf score and IDFAvg is the mean idf score. The other three features (PopulationCount, InterventionCount, and OutcomeCount) are the number of terms in the document and in the query that also appear in the respective PICO annotation. PICO annotations for documents were automatically extracted using RobotReviewer [5]. This automatic process only annotates the Population, Intervention, and Outcome for studies (the Control element is not annotated). PICO annotations for queries were manually collected by one of the team members, who is a clinician (AD). Afterwards, search strategies (both the original boolean query, and the new boolean query with PICO annotations) were manually transformed into Elasticsearch queries. The result is two Elasticsearch queries per topic — one which is representative of the original query made by the systematic review authors, and another annotated with PICO elements.

Initial testing on a recent collection we developed [3] allowed us to select a number of candidate learning to rank algorithms that may be effective in the screening prioritisation of systematic reviews.

We then empirically evaluated the selected five learning to rank algorithms listed in Table 2 and found that Coordinate Ascent provided us with the best MAP score on validation data for CLEF eHealth 2017 over the other models. Each time we trained a model, we used the default values for that model[1] and we set aside the same 30% of queries for validation. Table 2 summarises the NDCG@10 and average precision (AP) scores for both the original Boolean queries and the annotated PICO queries. We found that Coordinate Ascent was the best algorithm for learning to rank these types of studies. Additionally, we found that Random Forests and MART methods both had similar levels of NCG@10 and AP.

Additionally, we also used Elasticsearch (version 5.3) to produce a re-ranking. We did this by issuing the Boolean and PICO query to Elasticsearch and limited the results to only the PubMed identifiers contained in the topic file for each query. We then let Elasticsearch rank these documents using BM25 with the default settings[2]. We considered the Elasticsearch runs as our baseline.

[1] The default values for each model can be found at the following URL `https://sourceforge.net/p/lemur/wiki/RankLib%20How%20to%20use/`

[2] $k_1 = 1.2$, $b = 0.75$.

	NCG@10		AP	
	Boolean	PICO	Boolean	PICO
Elasticsearch	**0.397**	**0.409**	**0.104**	0.102
MART	0.237	0.327	0.066	0.086
AdaRank	0.0875	0.2197	0.0255	0.0619
Coordinate Ascent*	0.305	0.378	0.076	**0.114**
LambdaMART	0.259	0.377	0.068	0.097
Random Forests*	0.247	0.275	0.061	0.088

Table 2: Evaluation of each learning to rank model using the features listed in Table 1 on the test data compared to the Elasticsearch baseline. Algorithms marked with * indicate our submitted runs.

3 Results and Analysis

We found that a learning to rank approach to re-ranking studies for systematic reviews shows promising results. Table 2 illustrates our submitted runs compared to the baseline Elasticsearch ranking and additional runs performed post-submission. The models trained using the search strategies annotated using PICO achieved slightly better results than the provided Boolean search strategies. None of our models were able to score higher than the baseline in NCG@10, however the Coordinate Ascent model trained using PICO annotations outperformed the baseline in AP.

Additionally, we report AP, NCG@10, WSS@100 and the position of the last relevant document (last_rel) in Figure 1. These visualisations show that the Coordinate Ascent model provides the most effective ranking of documents (in terms of AP and WSS@100) and scores the highest amongst the learning to rank models for recall based measurements (NCG@10). Figure 1c shows that learning to rank models trained on the Boolean search strategies positioned the last relevant document in the re-ranked list the highest; and that the baseline Elasticsearch runs do not do this as well.

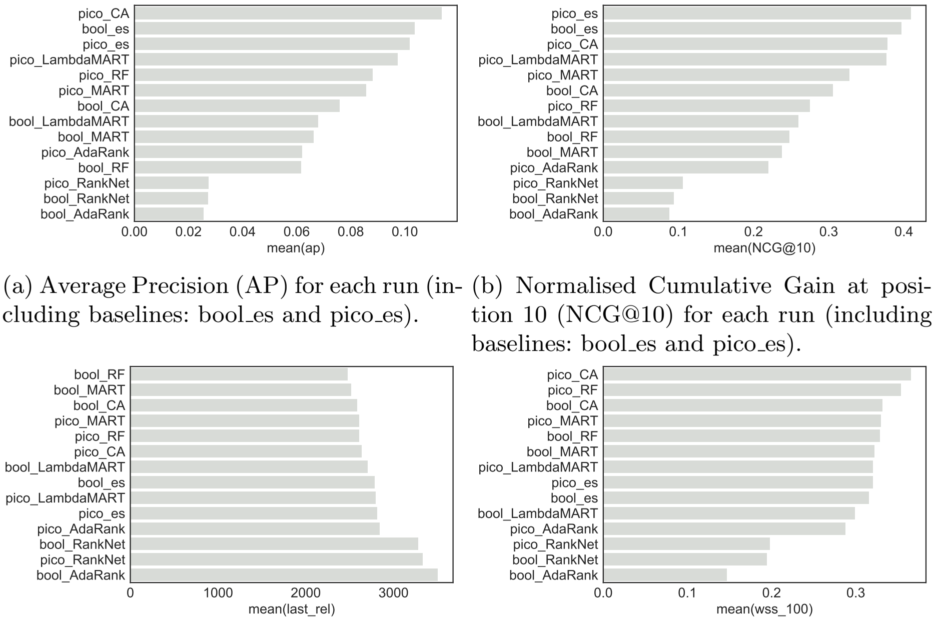
Figure 2 examines the effect PICO had on re-ranking. The effect appears negligible on the baseline, however, we notice an increase in precision when PICO annotations are used as training data for learning to rank models. This suggests that the use of PICO provides a trade off between precision and recall. Our results illustrate this clearly when precision-based measures are compared against recall-based measures.

4 Future Work

We plan to further increase the precision of our experiments by tuning the hyper parameters of the best performing learning to rank models. Our learning to rank models were trained using only a small number of features. We will investigate the effects of other features that are commonly used for learning to rank, and explore more domain specific features in addition to PICO.

References

1. 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 (2017)
2. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
3. Scells, H., Zuccon, G., Koopman, B., Deacon, A., Azzopardi, L., Geva, S.: A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In: Proceedings of SIGIR '17 (2017)
4. Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F., Coiera, E.: Systematic review automation technologies. Systematic reviews 3(1), 74 (2014)
5. Wallace, B.C., Kuiper, J., Sharma, A., Zhu, M.B., Marshall, I.J.: Extracting PICO sentences from clinical trial reports using supervised distant supervision. Journal of Machine Learning Research 17(132), 1–25 (2016)



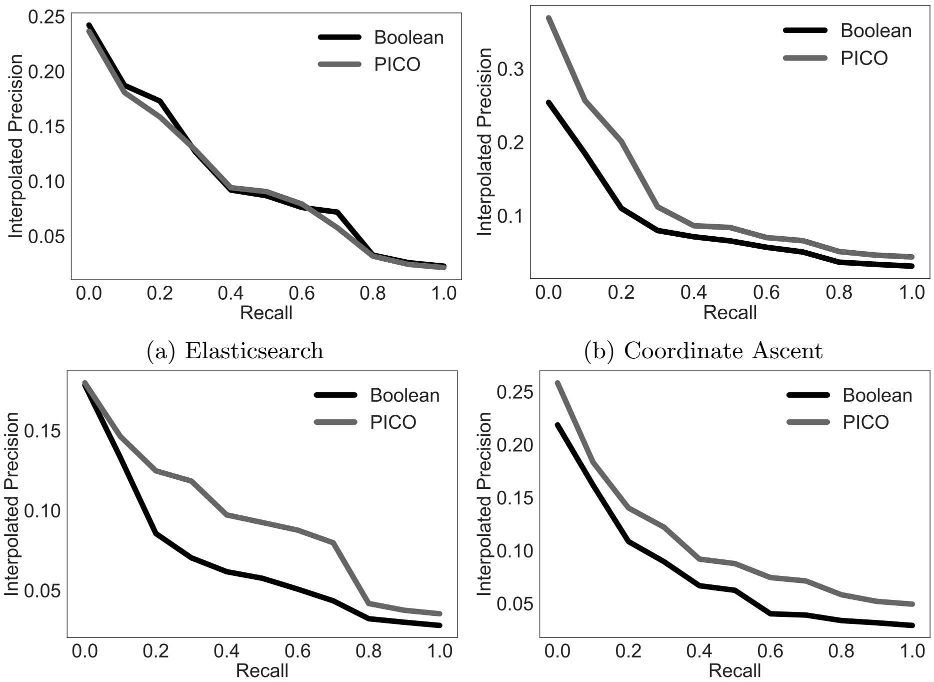
(a) Average Precision (AP) for each run (including baselines: bool_es and pico_es).

(b) Normalised Cumulative Gain at position 10 (NCG@10) for each run (including baselines: bool_es and pico_es).

(c) Position in the re-ranked list of the last study retrieved (including baselines: bool_es and pico_es).

(d) Position in the re-ranked list of the last study retrieved (including baselines: bool_es and pico_es).

Fig. 1: Comparison of the effects each algorithm had on different measures.



(a) Elasticsearch

(b) Coordinate Ascent

(c) Random Forests

(d) LambdaMART

Fig. 2: Precision-recall curves for the Elasticsearch baselines and the Coordinate Ascent models.