Learning-to-Rank and Relevance Feedback for Literature Appraisal in Empirical Medicine

Athanasios Lagopoulos[0000−0002−1979−3915], Antonios Anagnostou[0000−0002−2451−3981], Adamantios Minas[0000−0001−6752−2338], and Grigorios Tsoumakas[0000−0002−7879−669X]

Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
{lathanag,anagnoad,adamantcm,greg}@csd.auth.gr

Abstract. The constantly expanding medical libraries contain immense amounts of information, including evidence from healthcare research. Gathering and interpreting this evidence can be both challenging and time-consuming for researchers conducting systematic reviews. Technologically assisted review (TAR) aims to assist this process by finding as much relevant information as possible with the least effort. Toward this, we present an incremental learning method that ranks documents, previously retrieved, by automating the process of title and abstract screening. Our approach combines a learning-to-rank model trained across multiple reviews with a model focused on the given review, incrementally trained based on relevance feedback. The classifiers use as features several similarity metrics between the documents and the research topic, such as Levensthein distance, cosine similarity and BM25, and vectors derived from word embedding methods such as Word2Vec and Doc2Vec. We test our approach using the dataset provided by the Task II of CLEF eHealth 2017 and we empirically compare it with other approaches participated in the task.

Keywords: Learning to Rank · Relevance Feedback · Technology-Assisted Reviews · Empirical Medicine

1 Introduction

Evidence-Based Medicine (EBM) is an approach to medical practice that makes thorough and explicit use of the current best evidence in making decisions about the care and treatment of patients. Clinicians practice EBM by integrating their expertise with the best available external clinical evidence from systematic reviews [25]. A systematic review attempts to collect all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question by minimizing the bias and thus providing more reliable findings [9]. The creation of a systematic review usually includes the following three stages [13]:

1. **Document retrieval**: Information specialists build a Boolean query and submit it to a medical database, which returns a set of possibly relevant studies. Boolean queries typically have very complicated syntax and consist

Listing 1.1. Part of a boolean query constructed by Cochrane experts. Retrieved from Task II of CLEF eHealth 2017 (Topic ID: CD007394).

```
exp Ovarian Neoplasms/
Fallopian Tube Neoplasms/
((ovar* or fallopian tube*) adj5 (cancer* or tumor*
or tumour* or adenocarcinoma* or carcino* or
cystadenocarcinoma* or choriocarcinoma* or malignan*
or neoplas* or metasta* or mass or masses)).tw,ot.
```

of multiple lines. An example of such a query can be found for reference in Listing 1.1.

2. **Title and abstract screening**: Domain experts go through the title and abstract of the set of documents retrieved by the previous stage, perform a first level of screening and remove irrelevant studies.
3. **Document screening**: Experts go through the full text of each document that passes the screening of the previous stage to decide whether it will be included in their systematic review.

Considering the rapid pace with which medical databases are expanding and the amount of information they contain, collecting and interpreting evidence into reviews requires time, skills and resources making it very challenging for health care providers and researchers. Organizations such as Cochrane[1], the Centre for Reviews and Dissemination[2] and the Joanna Briggs Institute[3] respond to this challenge by producing high-quality systematic reviews in health care. However, the specificity of boolean searches is usually low, hence the reviewers often need to look manually through thousands of articles, in tight timescales, in order to identify only the relevant ones [17]. Therefore, identifying *all* relevant studies and minimizing the bias in the selection are still very complex tasks [20,3].

This paper presents an approach for assisting experts in the second stage of creating systematic reviews, by ranking the set of documents retrieved by a Boolean query search. Our approach is based on text mining techniques and combines an inter-review learning-to-rank method with an intra-review incremental training method. Both similarity measures and vectors extracted by word embedding methods are used as features to the classifiers. We test our approach using the dataset provided by Task II [13] of the CLEF eHealth 2017 lab [7] and compare it with other approaches submitted to the task. Finally, we evaluate the performance of the different features extracted. A preliminary version of this work [2] was presented at the CLEF eHealth 2017 lab.

The rest of the paper is organized as follows: After providing related work in Section 2, we introduce our approach in ranking documents retrieved by a boolean query in Section 3. In Section 4, we describe our empirical study by

[1] http://www.cochrane.org/
[2] https://www.york.ac.uk/crd/
[3] http://joannabriggs.org/

presenting the data and the evaluation process we followed for our classification methods, while final conclusions and future work are outlined in Section 5.

2 Related Work

Several approaches have been proposed in the past to automate the different processes of creating a systematic review. Most of them are particularly focused on reducing the burden of screening for reviewers. These approaches are based on text mining [31,20,11] along with active learning [8,30] or learning-to-rank [22]. Furthermore, different systems and platforms have been developed. Abstrackr [23] and Rayyan [21] use a semi-automatic active learning way to perform citation screening, while Cochrane Crowd[4] is an online collaborative platform that categorizes health care evidence.

The recently organized task on Technologically Assisted Reviews in Empirical Medicine [13] of CLEF eHealth 2017 [7], with a focus on Diagnostic Test Accuracy (DTA), aimed to bring together academic, commercial, and government researchers that conduct experiments and share results on automatic methods to retrieve relevant studies. Specifically, a set of research topics were provided to the participants. The topics were constructed by Cochrane experts and each topic contained the title of a systematic review and the corresponding boolean query. The set of documents returned from the query were also provided. The participants were asked to rank the documents so as: (i) to produce an efficient ordering of the documents such that all of the relevant abstracts are retrieved as early as possible, and (ii) to identify a subset of documents which contains all or as many of the relevant abstracts for the least effort (i.e. total number of abstracts to be assessed). Fourteen teams participated in the task and presented their work. Several teams developed Learning-to-Rank approaches [10,26,4], while others adopted active learning techniques [6,32]. Two teams worked with neural networks and deep learning [27,16]. Furthermore, participants represented the textual data in a variety of ways, including topic models [29,12], TF-IDF [1] and n-grams [19].

3 Our Approach

Our approach comprises two consecutive supervised learning models. The first model is a learning-to-rank binary classifier that considers a topic-document pair as input and whether the document is relevant to the systematic review or not as output (Figure 1). This inter-review model is used at the first stage of our approach in order to obtain an initial ranking of all documents returned by the Boolean query of an unseen test topic. The second model is a standard binary classifier that considers a document of the given test topic as input and whether this document is relevant to the test topic as output. This intra-review model is incrementally trained based on relevance feedback that it requests after

[4] `http://crowd.cochrane.org/`

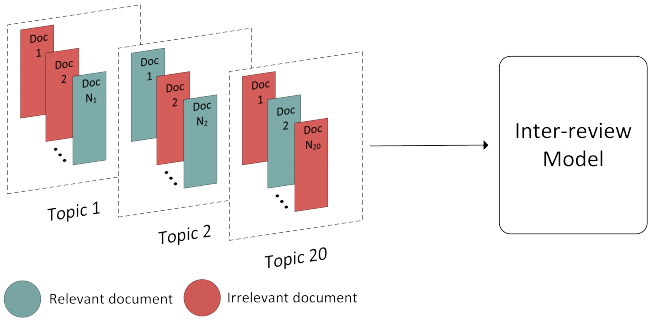

Fig. 1. Training of the inter-review model.

returning one or more documents to the user. The first version of this model is trained based on feedback obtained from the top k ranked documents by the inter-review model (Figure 2). The re-ranking of subsequent documents is from then on based solely on the intra-review model (Figure 3).

3.1 Inter-review model

The inter-review model is a learning-to-rank model that ranks the set of documents according to their relevance and importance to the topic. Each topic-document pair is represented by a multi-dimensional feature vector, and each dimension of the vector is a feature indicating how relevant or important the document is with respect to the topic [22]. In total, 31 features were extracted. Most of the features (1-26) are simple similarity features and they are computed by considering the similarity of different fields of the document (title, abstract), with different fields of the topic (title, boolean query), using a variety of similarity measures, such as the number of common terms between the topic and the document parts, Levenshtein distance, cosine similarity and BM25 [28]. The text

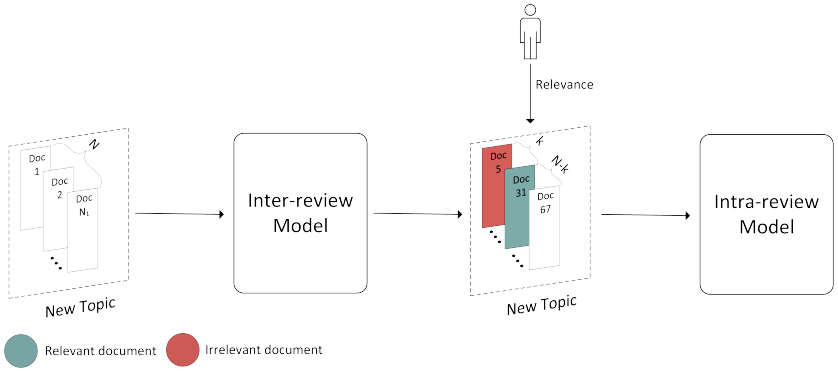

Fig. 2. Ranking with the inter-review model. Initial training of the intra-review model.

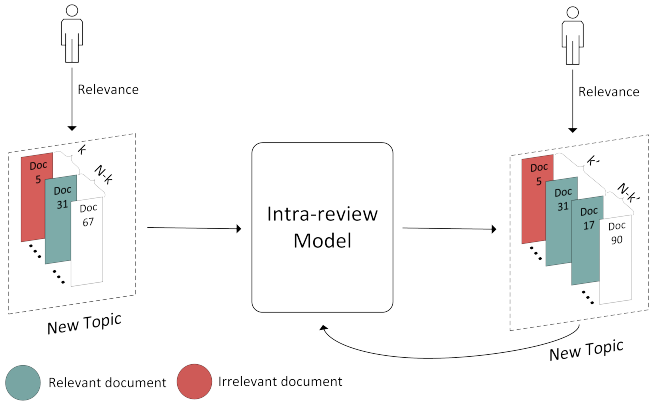

Fig. 3. Continuous re-ranking of subsequent documents and incremental re-training of the intra-review model.

in these cases is represented either as simple word tokens or as TF-IDF vectors. The remaining 5 features (27-31) are also similarity measures between the topic and the document but the text representations are word embeddings extracted from methods such as Word2Vec [18] and Doc2Vec [15].

Table 1 presents the features which we employed in our model. Two of these features depend only on the topic, denoted with T in the *Category* column of Table 1, as opposed to the rest of the features that are dependent on both the topic and the document, denoted with $T - D$. Details about the features are listed below.

1. We consider two fields of a document d: the title and the abstract. The column *Document field* indicates which field is used by the feature.
2. We consider two fields of a topic: the title t, consisting of tokens t_i, and the Medical Subject Headings (MeSH) m extracted from the boolean query.
3. $|C|$ is the total number of documents in the document collection. $|d|$ denotes the length, the number of tokens d_j, of a document d considering a specific field. Document frequency $df(t_i)$ is the number of documents containing t_i.
4. The number of occurrences of title tokens or MeSH of the topic in a document d is denoted as $c(t, d)$ and $c(m, d)$, respectively.
5. In features 1-20 a simple string tokenization of the text is considered.
6. The $levenshtein(x, y)$ stands for the Levenshtein distance string metric. The v value is user defined.
7. The BM25 score is computed as in [24].
8. The vocabulary and inverse-term frequency (idf) of tf-idf is fitted on the topic's title.
9. In feature 25-26, we follow a standard Latent Semantic Analysis (LSA). A Singular Value Decomposition (SVD) is performed upon the tf-idf, which is fitted on the documents' title and abstract. The cosine similarity is estimated from the reduced vectors of the two fields.

10. In feature 27-28, the vector of each field is the averaging vector of the word vectors produced by a Word2Vec model.
11. In feature 29-30, the Word Mover's Distance (WMD) of the word vectors is computed as in [14].
12. In feature 31, the vector of each field is produced by a Doc2Vec model [15].

Table 1. Set of features employed by the inter-review model.

ID	Description	Category	Topic field	Document field
1	$\sum_{t_i \in t \cap d} c(t_i, d)$	$T-D$	Title	Title
2	$\sum_{t_i \in t \cap d} \log(c(t_i, d))$	$T-D$	Title	Title
3	$\sum_{t_i \in t \cap d} c(t_i, d)$	$T-D$	Title	Abstract
4	$\sum_{t_i \in t \cap d} \log(c(t_i, d))$	$T-D$	Title	Abstract
5	$\sum_{m_i \in t \cap d} c(m_i, d)$	$T-D$	Query	Title
6	$\sum_{m_i \in t} \sum_{d_j \in d} levenshtein(m_i, d_j)$	$T-D$	Query	Title
7	$\sum_{m_i \in t} \sum_{d_j \in d} levenshtein(m_i, d_j)$ if $levenshtein(m_i, d_j) < k$	$T-D$	Query	Title
8	$\sum_{m_i \in t \cap d} \log(c(m_i, d))$	$T-D$	Query	Title
9	$\sum_{m_i \in t \cap d} c(m_i, d)$	$T-D$	Query	Abstract
10	$\sum_{m_i \in t \cap d} \log(c(m_i, d))$	$T-D$	Query	Abstract
11	$\sum_{m_i \in t} \log(\frac{\lvert C \rvert}{df(t_i)})$	T	Title	-
12	$\sum_{m_i \in t} \log(\log(\frac{\lvert C \rvert}{df(t_i)}))$	T	Title	-
13	BM25	$T-D$	Title	Title
14	BM25	$T-D$	Title	Abstract
15	BM25	$T-D$	Query	Title
16	BM25	$T-D$	Query	Abstract
17	log(BM25)	$T-D$	Title	Title
18	log(BM25)	$T-D$	Title	Abstract
29	log(BM25)	$T-D$	Query	Title
20	log(BM25)	$T-D$	Query	Abstract
21	*cos*(tf-idf)	$T-D$	Title	Title
22	*cos*(tf-idf)	$T-D$	Title	Abstract
23	*cos*(tf-idf)	$T-D$	Query	Title
24	*cos*(tf-idf)	$T-D$	Query	Abstract
25	*cos*(SVD(tf-idf))	$T-D$	Title	Title
26	*cos*(SVD(tf-idf))	$T-D$	Title	Abstract
27	*cos*(Word2Vec)	$T-D$	Title	Title
28	*cos*(Word2Vec)	$T-D$	Title	Abstract
29	WMD(Word2Vec)	$T-D$	Title	Title
30	WMD(Word2Vec)	$T-D$	Title	Abstract
31	*cos*(Doc2Vec)	$T-D$	Title	Abstract

3.2 Intra-review model

The intra-review model is a standard binary model which classifies a document as relevant or not to a certain topic. Initially, the intra-review model is trained based on the top k documents as ranked by the inter-review model. It then iteratively re-ranks the rest of the documents, expanding the training set of the intra-review model with the top-ranked document, until the whole list has been added to the training set or a certain threshold is reached. The expansion of the training set can be configured with user-defined steps. After the initial training with k documents, an initial expansion step is defined ($step_{init}$) until a certain threshold (t_{step}) is reached. Then, the step is increased to a secondary step ($step_{secondary}$). The secondary step is used until the final threshold (t_{final}). This iterative feedback and re-ranking mechanism is described in detail in Algorithm 1. The use of different steps and thresholds reduces the cost of feedback and the time needed to produce predictions since the classifier is considered sufficiently trained when a certain amount of documents is used in the training set. For this classifier, a standard *tf-idf* vectorization was used, enhanced with English stop word removal.

4 Empirical Study

This section initially describes the data we used for our empirical study and gives details about the implementation and the technologies underneath our approach. It then presents the evaluation process we followed for our models and, finally, it discusses the evaluation results and compares them with the results presented in Task II of CLEF eHealth 2017 lab.

4.1 Data & Preprocessing

We experimented with the development set distributed by the Task II of CLEF eHealth 2017 lab. In total, the set contains 50 topics, 20 topics in the training set and 30 topics in the test set. However, 8 topics were later marked as unreliable from the organizers, reducing the number of total topics to 42. Each topic contains an ID, a systematic review title, a boolean query in Ovid MEDLINE format and set of MEDLINE document's PIDs returned from the boolean query. The title and the boolean query are constructed by Cochrane experts. Each MEDLINE document contains the title, the abstract text and the MeSH headings. Along with the topics, the corresponding relevance sheet were also provided, denoting the positive or negative relevance of a document to a topic as derived from an abstract-level screening. The percentage of relevant documents at abstract level for the 42 topics is 4.07%. The full dataset is publicly available at the official GitHub repository of the task [5].

In order to use the rich information available in the boolean query field of the topics and be able to construct the features described in Table 1 we used

[5] `https://github.com/CLEF-TAR/tar`

Algorithm 1: Iterative relevance feedback algorithm of the intra-review model

Input : The ranked documents R, of length n, as produced by the inter-review model, initial training step k, initial local training step $step_{init}$, secondary local training step $step_{secondary}$, step change threshold t_{step}, final threshold t_{final} (optional)

Output: Final ranking of documents R - $finalRanking$

$finalRanking \leftarrow ()$; `// empty list`

for $i = 1$ ***to*** k **do**

  $finalRanking_i \leftarrow R_i$

$k' \leftarrow k$;

while *not* $finalRanking$ *contains both* ***relevant*** *and* ***irrelevant*** *documents* **do**

  $k' \leftarrow k' + 1$;

  $finalRanking_{k'} = R_{k'}$;

while *not* $length(finalRanking) == n$ ***OR*** $length(finalRanking) == t_{final}$ **do**

  $train(finalRanking)$; `// Train a local classifier by asking for abstract or document relevance for these documents`

  $localRanking = rerank(R - finalRanking)$; `// Rerank the rest of the initial list` R `from the predictions of the local classifier`

  if $length(finalRanking) < t_{step}$ **then**

    $step = step_{init}$;

  else

    $step = step_{secondary}$;

  for $i = k'$ ***to*** $k' + step$ **do**

    $finalRanking_i \leftarrow localRanking_{i-k'}$;

return $finalRanking$;

Polyglot[6], a JavaScript tool that can parse and produce a full syntactical tree of Ovid MEDLINE boolean queries. In particular, we extracted those MeSH that *should* characterize the retrieved documents, avoiding the ones that are negated in the query syntax.

4.2 Evaluation Process & Results

We split our evaluation process into two stages. The first stage is focused solely on the evaluation of the inter-review model and how different sets of features affect its performance. In the second stage we try to utilize the parameters of the intra-review model to make better use of the output, the initial ranking, of the inter-review model.

For all our experiments, we employ the XGBoost algorithm [5] to learn the inter-review model and linear support vector machines (SVMs), from the scikit-learn library[7], to learn the intra-review models. We use the default parameter

[6] `https://github.com/CREBP/sra-polyglot`

[7] `http://scikit-learn.org/`

settings for the XGBoost classifier and we set the C parameter of linear SVM to 0.1. Furthermore, for feature 7 we set v to 5 and for features 25-26 we set the number of output dimensions of SVD to 200. The Word2Vec model used for features 27-30 was obtained from the BioASQ challenge[8]. This model has been trained on 10,876,004 English abstracts of biomedical articles from PubMed resulting in 1,701,632 distinct word vectors. The Doc2Vec model used in feature 31 is trained with all the documents associated with a topic retrieved from PubMed Central (PMC). Finally, we use all 42 topics for our evaluation and we perform cross validation using the Leave-One-(Topic)-Out method. Evaluation measures are computed using the script provided by the task[9] based on relevant judgment at the abstract level.

The first stage of our evaluation process focuses on the inter-review model. We first evaluate the features used in this model by performing an Anova F-test between each feature and the class. Table 2 shows the top-10 features along with their scores. We notice that all the features in which tf-idf is computed (21-26) are included in the top-10 with the ones using SVD to be the highest-ranked, which highlights the importance of semantic analysis. The list is completed with two Word2Vec features (27,30) and two features that depend only on the topic (11-12). These features are related to the frequency of terms t_i in the title of the topic and are most probably regulating the importance of other features based on t_i, such as features 1-4.

To evaluate our model we perform three experiments using different sets of features. The first experiment makes use of features 1-24 which are standard LtR features. Our submission in Task II of CLEF eHealth 2017 lab also included the same features [2]. The second experiment consists of the full list of features 1-31 which includes advanced text representations derived from word embedding methods. The final experiment uses the top-10 features determined by the Anova F-test. Table 3 shows the Average Precision (AP), the normalized cumulative gain (NCG) at 10% and 20% and the minimum number of documents returned to retrieve all relevant documents (Last Relative - LR) of the three models described above. We first notice that using the full list of features achieves better scores than using just the top-10 features or the simple LtR features, beating our previous approach. Besides the increase in average precision, we also see an increase of NCG@10 and NCG@20 which indicates that more relevant documents appear first when using the additional features. This also hints at the need of a highly complex model that can overcome the high bias due to our very unbalanced dataset. Furthermore, the fact that the model using just the top-10 features achieves better results than the model using the simple LtR features highlights the strong influence of these specific features.

In the second stage of our evaluation process we explore the parameter space of the intra-review model as described in Section 3.2. Table 4 presents the final results of our approach using different parameter sets. The inter-review model using the full list of features is employed. We notice that integrating the intra-

[8] `http://bioasq.org/`
[9] `https://github.com/CLEF-TAR/tar`

Table 2. The scores of the top-10 features as measured by the F-test in ANOVA.

Rank	Feature ID	F-Score	Rank	Feature ID	F-score
1	25	7013.55	6	12	2682.01
2	26	6363.41	7	11	2613.24
3	21	5252.00	8	30	1541.90
4	22	3289.95	9	24	501.01
5	27	2700.76	10	23	373.70

Table 3. Results concerning the inter-review model using different sets of features.

Features	AP	NCG@10	NCG@20	LR
Simple LtR (1-24)	0.171	0.363	0.594	4085.643
Full list (1-31)	0.187	0.382	0.613	3776.262
Top-10	0.177	0.372	0.601	3993.167

review model greatly increases the scores in all four metrics compared with the sole use of the inter-review model. The intra-review model not only ranks the relevant documents higher, as indicated by the NCG@ metrics, but also decreases, almost in half, the total number of documents returned to retrieve all relevant documents (LR metric).

Table 4. Results of our approach using different parameters of the intra-review model.

Run	k	$step_{init}$	t_{step}	$step_{sec}$	t_{final}	AP	NCG@10	NCG@20	LR
1	5	1	200	100	2000	0.309	0.533	0.819	2109.83
2	10	1	200	100	2000	0.309	0.536	0.820	2106.43
3	15	1	200	100	2000	0.304	0.533	0.820	2109.95
4	10	1	300	100	2000	0.310	0.534	0.824	2104.97
5	10	1	500	100	2000	0.311	0.538	0.822	2108.93

5 Conclusion and future work

We introduced a classification approach for automatic title and abstract screening for systematic reviews. Our approach constructs a global inter-review classification model based on LtR features of the topics and documents, produces an initial ranking for the test documents and then a second model iteratively asks for feedback and re-ranks them based on the acquired relevance feedback.

In the future, we plan to work more on the tuning and extraction of better features for the inter-review model and produce a better representation for the intra-review model using word embedding methods. Moreover, it would be worthy to experiment with other classification approaches as well, such as convolutional and recurrent neural networks.

References

1. 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: Working notes for CLEF 2017. In: CEUR Workshop Proceedings. vol. 1866 (2017)
2. Anagnostou, A., Lagopoulos, A., Tsoumakas, G., Vlahavas, I.: Combining inter-review learning-to-rank and intra-review incremental training for title and abstract screening in systematic reviews. In: CEUR Workshop Proceedings. vol. 1866 (2017)
3. Bastian, H., Glasziou, P., Chalmers, I.: Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? PLoS Medicine (2010). https://doi.org/10.1371/journal.pmed.1000326
4. Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., He, L., Yang, Y.: ECNU at 2017 eHealth task 2: Technologically assisted reviews in empirical medicine. In: CEUR Workshop Proceedings. vol. 1866 (2017)
5. Chen, T., Guestrin, C.: XGBoost : Reliable Large-scale Tree Boosting System. arXiv pp. 1–6 (2016). https://doi.org/10.1145/2939672.2939785
6. Cormack, G.V., Grossman, M.R.: Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017. In: CEUR Workshop Proceedings. vol. 1866 (2017)
7. 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: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 10456 LNCS, pp. 291–303 (2017). https://doi.org/10.1007/978-3-319-65813-1_26
8. Hashimoto, K., Kontonatsios, G., Miwa, M., Ananiadou, S.: Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of Biomedical Informatics **62**, 59–65 (2016). https://doi.org/10.1016/j.jbi.2016.06.001
9. Higgins JP, G.S.: Cochrane Handbook for Systematic Reviews of Interventions (2011), `www.handbook.cochrane.org`
10. Hollmann, N., Eickhoff, C.: Ranking and feedback-based stopping for recall-centric document retrieval. In: CEUR Workshop Proceedings. vol. 1866 (2017)
11. Howard, B.E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M.R., Holmgren, S., Pelch, K.E., Walker, V., Rooney, A.A., Macleod, M., Shah, R.R., Thayer, K.: SWIFT-Review: A text-mining workbench for systematic review. Systematic Reviews **5**(1) (2016). https://doi.org/10.1186/s13643-016-0263-z
12. Kalphov, V., Georgiadis, G., Azzopardi, L.: SiS at CLEF 2017 eHealth tar task. In: CEUR Workshop Proceedings. vol. 1866 (2017)
13. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings. vol. 1866 (2017)
14. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From Word Embeddings To Document Distances. Proceedings of The 32nd International Conference on Machine Learning **37**, 957966 (2015)
15. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. International Conference on Machine Learning - ICML 2014 **32**, 11881196 (2014). https://doi.org/10.1145/2740908.2742760
16. Lee, G.E.: A study of convolutional neural networks for clinical document classification in systematic reviews: Sysreview at CLEF eHealth 2017. In: CEUR Workshop Proceedings. vol. 1866 (2017)

17. Lefebvre, C., Manheimer, E., Glanville, J.: Searching for Studies. In: Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series, pp. 95–150 (2008). https://doi.org/10.1002/9780470712184.ch6
18. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013) pp. 1–12 (2013). https://doi.org/10.1162/153244303322533223
19. Norman, C., Leeflang, M., Névéol, A.: LIMSI@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking. In: CEUR Workshop Proceedings. vol. 1866 (2017)
20. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: A systematic review of current approaches. Systematic Reviews **4**(1), 1–22 (2015). https://doi.org/10.1186/2046-4053-4-5
21. Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A.: Rayyan-a web and mobile app for systematic reviews. Systematic Reviews **5**(1) (2016). https://doi.org/10.1186/s13643-016-0384-4
22. Qin, T., Liu, T.Y., Xu, J., Li, H.: LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval **13**(4), 346–374 (2010). https://doi.org/10.1007/s10791-009-9123-y
23. Rathbone, J., Hoffmann, T., Glasziou, P.: Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Systematic Reviews **4**(1) (2015). https://doi.org/10.1186/s13643-015-0067-6
24. Robertson, S.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2010). https://doi.org/10.1561/1500000019
25. Sackett, D.L.: Evidence-based medicine. Seminars in Perinatology **21**(1), 3–5 (1997). https://doi.org/10.1016/S0146-0005(97)80013-4
26. Scells, H., Zuccon, G., Deacon, A., Koopman, B.: QUT ielab at CLEF eHealth 2017 technology assisted reviews track: Initial experiments with learning to rank. In: CEUR Workshop Proceedings. vol. 1866 (2017)
27. Singh, G., Marshall, I., Thomas, J., Wallace, B.: Identifying diagnostic test accuracy publications using a deep model. In: CEUR Workshop Proceedings. vol. 1866 (2017)
28. Sparck Jones, K., Sparck Jones, K., Walker, S., Walker, S., Robertson, S.E., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments Part 2. Information Processing and Management **36**, 809–840 (2000). https://doi.org/10.1016/S0306-4573(00)00016-9
29. Van Altena, A.J., Olabarriaga, S.D.: Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling. In: CEUR Workshop Proceedings. vol. 1866 (2017)
30. Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A.: Active learning for biomedical citation screening. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10 p. 173 (2010). https://doi.org/10.1145/1835804.1835829
31. Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C.E., Schmid, C.H.: Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics **11**(1), 55 (2010). https://doi.org/10.1186/1471-2105-11-55
32. Yu, Z., Menzies, T.: Data balancing for technologically assisted reviews: Undersampling or reweighting. In: CEUR Workshop Proceedings. vol. 1866 (2017)