Dense Retrieval with Continuous Explicit Feedback for Systematic Review Screening Prioritisation

Xinyu Mao
The University of Queensland
Brisbane, Australia
xinyu.mao@uq.edu.au

Shengyao Zhuang
CSIRO
Brisbane, Australia
shengyao.zhuang@csiro.au

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

ABSTRACT

The goal of screening prioritisation in systematic reviews is to identify relevant documents with high recall and rank them in early positions for review. This saves reviewing effort if paired with a stopping criterion, and speeds up review completion if performed alongside downstream tasks. Recent studies have shown that neural models have good potential on this task, but their time-consuming fine-tuning and inference discourage their widespread use for screening prioritisation. In this paper, we propose an alternative approach that still relies on neural models, but leverages dense representations and relevance feedback to enhance screening prioritisation, without the need for costly model fine-tuning and inference. This method exploits continuous relevance feedback from reviewers during document screening to efficiently update the dense query representation, which is then applied to rank the remaining documents to be screened. We evaluate this approach across the CLEF TAR datasets for this task. Results suggest that the investigated dense query-driven approach is more efficient than directly using neural models and shows promising effectiveness compared to previous methods developed on the considered datasets. Our code is available at https://github.com/ielab/dense-screening-feedback.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Relevance Feedback, Dense Retrieval, Systematic Review Automation.

ACM Reference Format:
Xinyu Mao, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2024. Dense Retrieval with Continuous Explicit Feedback for Systematic Review Screening Prioritisation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3626772.3657921

1 INTRODUCTION

A medical systematic review aims to answer a specific medical question by collecting and appraising relevant studies as evidence. These reviews often require screening. Screening involves assessing documents for inclusion or exclusion in the review. A large number of retrieved documents, a task that is labour-intensive and time-consuming [6, 33, 34]. Screening is done in two phases: 1) document titles and abstracts are assessed first for relevance; and 2) full-text articles from phase 1 are assessed for relevance [20]. These two phases are conceptually executed sequentially; i.e., after filtering out a large number of irrelevant documents based on title and abstract, full text screening is started on the remaining documents.

Screening prioritisation can reduce the time needed to complete the review and becomes a critical task for systematic review automation methods, such as technology-assisted review (TAR) [26]. In screening prioritisation, documents are ranked, ideally in a way that places relevant documents above non-relevant ones, allowing them to be screened first. This enables downstream tasks such as full-text screening to begin as soon as a relevant document is found. If all relevant documents are encountered immediately, then downstream tasks might be completed concurrently as other researchers continue screening the remaining irrelevant documents, thus shortening the time to write the review [38]. Screening prioritisation can also save overall effort (and thus cost) when early stopping is applied to halt screening after top-k documents are reviewed, thereby avoiding the exhaustive screening of all documents [9, 42, 48].

There are two main approaches to screening prioritisation [45]: (1) **Query-based:** The query associated to the systematic review is used to rank documents, often using sparse representations such as BM25 [3, 13] and TF-IDF [2, 40]. Queries are often formed from the working title of the review [1–3], its research questions [40], or the Boolean query used to retrieve the documents to screen [1–3, 7]. (2) **Model-based:** A classification model is trained to discriminate documents into inclusion or exclusion classes. Most work on this task has focused on the use of traditional machine learning models such as SVM [4, 52] and logistic regression [28, 36, 46]. Recent studies have shown the potential of BERT-based models for this task [35, 45, 49]. Many techniques are also leveraged together with these models, such as active learning [10, 11, 18, 28, 37, 42, 52], and relevance feedback [4, 10, 11, 21, 52], which is also seen in combination with query-based methods [1, 3, 12–14].

Model-based approaches in TAR often incorporate *active learning*, such as CAL [8], where a SVM classifier is trained with continuous feedback from top-ranked documents. Specifically, linear classifiers require relevant seed documents to initiate training. However, in systematic reviews, seed studies are served for query formulation and are not always relevant to the topic [44], which may harm retrieval effectiveness when initiating the active learning workflow in a real systematic review setting. Additionally, the size of the training set also increases due to the active learning mechanism, though the

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657921

classifiers can still benefit by combining knowledge from retrieval (such as search keywords) regardless of the varying size of the training set [47]. Recently, a BERT-based TAR workflow [49] has shown a promising trend, delivering higher effectiveness when prioritising documents for screening in systematic reviews, especially if applied with a domain-specific backbone without any further pre-training [32]. A drawback is the method incurs higher costs in time and computation compared to linear models [32]. Query-based approaches have mainly been investigated in static, once-off rankings. These methods could be promising in a continuous feedback setting, especially if they can deliver similar effectiveness to model-based approaches but at a lower computational cost.

In this paper, we follow the query-based approach but adopt neural encoder models, such as BERT-based dense retrievers, to perform screening prioritisation that exploits the human reviewer's iterative feedback. While BERT-based rankers have been shown highly effective for screening prioritisation, no relevance feedback mechanism has been investigated [45]. Methods for using relevance feedback in combination with BERT-based rankers have been devised for ad-hoc retrieval [43, 51, 53]; however, their limitations include (1) only considering pseudo-relevance feedback, (2) only implementing once-off setting (i.e., the feedback mechanism is used only once to produce a ranking), (3) most methods re-train the ranker after feedback (through additional fine-tuning) [5, 30, 51]. An exception is Li et al. [29], which we adapt here to screening prioritisation. The task of screening prioritisation differs in that the feedback is explicit, and is provided in a continuous manner as reviewers perform screening. Typically, adapting current BERT-based rankers for screening prioritisation with relevance feedback is computationally expensive due to iterative re-fine-tuning. We therefore explore an unexamined alternative: using dense retrievers with an efficient strategy for leveraging feedback [29]. Our results show that this approach is not only efficient but also matches or exceeds the performance of specialised methods for screening prioritisation.

2 DENSE RETRIEVAL WITH CONTINUOUS EXPLICIT FEEDBACK

Figure 1 is an overview of our proposed dense retrieval framework for screening prioritisation with continuous feedback. In Stage 1, users formulate Boolean queries to retrieve documents from databases such as PubMed. These documents form the pool for title and abstract screening in Stage 2. A protocol, defining key questions of the review and the inclusion/exclusion criteria, is also available at Stage 1. We then propose to utilise topic-related information from the protocol as the query for dense retrieval against the pool, and then select top-k documents for user examination. Users assess these documents as relevant or non-relevant using titles and abstracts. These binary judgments then serve as explicit relevance feedback, updating the query with Rocchio's algorithm on the dense representations [29]: $\overrightarrow{q}_{\text{update}} = \alpha * \overrightarrow{q} + \beta * \text{avg}(\overrightarrow{d}^{+}_{1}, ..., \overrightarrow{d}^{+}_{m}) + \gamma * \text{avg}(\overrightarrow{d}^{-}_{1}, ..., \overrightarrow{d}^{-}_{n})$ where $\overrightarrow{d}^{+}$ and $\overrightarrow{d}^{-}$ are relevant and non-relevant dense representations (as judged by reviewers) and $\overrightarrow{q}$ is the dense query. α, β, and γ are the Rocchio weights of the previous query, relevant documents, and non-relevant documents' dense representation, respectively. The refined dense query representation is

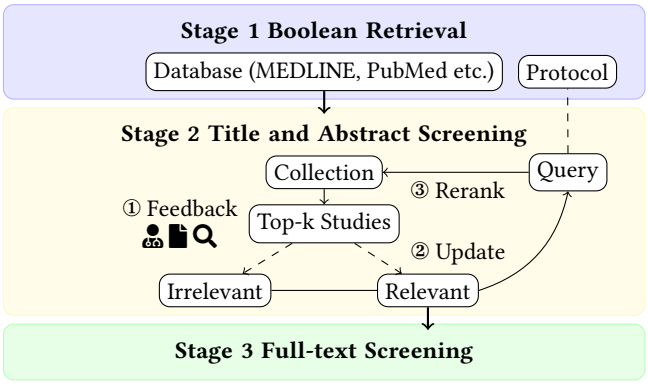

Figure 1: Screening prioritisation with feedback.

used to obtain a new ranking over the remaining documents to be screened. This process is repeated until all documents are reviewed or a stopping point is reached. Documents identified as relevant at each iteration can be directly passed to the downstream tasks (e.g., full-text screening), accelerating the systematic review process. The effectiveness and efficiency of Rocchio's algorithm on dense representation of queries and documents have been studied for once-off, static creation of a ranking based on pseudo-relevance feedback and for ad-hoc retrieval; however, its use in continuous feedback and for systematic review screening has not been explored.

3 EXPERIMENT SETUP

Datasets. We rely on the CLEF-TAR 2017, 2018 and 2019 Subtask 2 datasets [22–24] (abbreviated as CLEF 17-19). These datasets contain queries (topics), documents (title and abstract only) and relevance assessments associated with real systematic reviews. We use the working title of the review as the query, which is available in the protocol file under each topic in the datasets. Each dataset has one training set and one test set. We use the training set to sample positive and negative documents for training a dense retriever; while we use the test set for the iterative ranking and the evaluation.

Dense Retriever Training and Retrieval. For dense retrievers, we use domain-specific (BioBERT [27], PubMedBERT [19], BiolinkBERT [50]) and task-specific (coCondenser [16]) as backbones. Models were trained using Tevatron [17] with 10 training passages per topic (comprising 1 positive and 9 negatives) in a triplet loss $< \text{topic}, d^{+}, d^{-} >$ [25], on a single NVIDIA V100 with 32GB memory for 60 epochs. For dense retrieval with relevance feedback, we used Pyserini [31] for retrieval and FAISS [15] for encoding, and indexing the queries and corpus. Average runtime per collection was 2.6 minutes for training, 1.2 minutes for encoding/indexing, and 5 minutes per retrieval setting (model, Rocchio setting, feedback size), with time increasing for smaller feedback sizes.

Relevance Feedback Settings. Rocchio's algorithm has three parameters α, β, γ, for which we have four settings: (1, 1, 1), (1, 0.8, 0.2), (1, 0.5, 0.5), (1, 1, 0). This allows us to explore the impact of including only positive feedback and varying degrees of negative feedback. Another parameter is the number of documents included in each feedback iteration, denoted as k, which is typically fixed throughout an experiment. We set $k = 25$ in line with previous work

Table 1: Initial ranking refers to the results obtained before screening prioritisation with continuous feedback. (α, β, γ) presents different feedback settings. † indicates statistical significant difference between dense retrievers and BM25+RM3 for initial rankings, while * between no feedback and continuous feedback. Statistical significance has been computed using paired t-test with Bonferroni correction, $p < 0.05$.

Collections	Methods	initial ranking		(1,1,1)		(1,0.5,0.5)		(1,0.8,0.2)		(1,1,0)	
		AP	Last Rel	AP	Last Rel	AP	Last Rel	AP	Last Rel	AP	Last Rel
CLEF 17	**BM25+RM3**	.1439	2228	-	-	-	-	-	-	-	-
	coCondenser	.2064	2477	.2335	*3770	*.2357	*3734	*.2404	2110	***.2370**	**1979**
	BioLinkBERT	.2402	**1917**	***.2738**	*3620	*.2707	*3538	.2543	2087	.2431	2314
	PubMedBERT	.1597	**2669**	***.1914**	3560	*.1830	3598	*.1227	2850	*.1144	*3423
	BioBERT	*.1083	†3278	.1161	*3767	**.1175**	*3777	.1095	3315	.1076	**3113**
	auth.simple.run1 [4] AP: **.2970** Last Rel: **2143**										
CLEF 18	**BM25+RM3**	.1958	3276	-	-	-	-	-	-	-	-
	coCondenser	†.2798	5291	*.3420	*4476	***.3472**	*4315	*.3520	**2590**	*.3456	2609
	BioLinkBERT	†.3601	5047	***.4274**	*4027	*.4213	*4015	.3816	**2639**	.3656	2954
	PubMedBERT	†.3284	5673	***.3916**	3764	*.3818	3773	*.2047	**3304**	*.1826	*4236
	BioBERT	.1748	†6335	***.2331**	4536	*.2287	4552	.1783	3839	.1692	**3701**
	cnrs_comb [37] AP: **.3470** Last Rel: **2406**										
CLEF 19 dta	**BM25+RM3**	.1677	2633	-	-	-	-	-	-	-	-
	coCondenser	.1904	**1232**	.2086	2932	.2158	2893	.2337	2178	***.2400**	1288
	BioLinkBERT	.2239	1255	.2567	2822	.2529	2759	***.2568**	**876**	.2534	1143
	PubMedBERT	.2288	**891**	***.2536**	2670	.2485	2636	.1914	1256	.1763	2000
	BioBERT	.1723	**1577**	.1925	2912	.1927	2857	.1865	1638	***.1824**	1722
CLEF 19 int.	**BM25+RM3**	.1620	1178	-	-	-	-	-	-	-	-
	coCondenser	†.3459	1172	.3760	1886	.3725	1872	***.3794**	**658**	*.3747	830
	BioLinkBERT	†.3876	964	***.4151**	1628	.4087	1734	*.4088	**707**	*.4041	961
	PubMedBERT	†.2960	**946**	***.3261**	1891	*.3178	1619	*.2702	1113	*.2666	1416
	BioBERT	.1251	1713	.1475	1959	**.1495**	1957	*.1367	**1576**	.1332	1579

Table 2: Comparison between our dense retrieval method and TAR active learning with relevance feedback for screening prioritisation. B stands for BioLinkBERT, L for logistic regression; title and pos for different seed settings; * for statistical significance computed as in Table 1.

20 iteration (cut-off @ 500)			
Collections	**Runs**	**AP**	**Last Rel**
CLEF 17	dense - B - best	**.2660**	**241**
	tar - B - title	*.0674	*358
	tar - B - pos	***.1431**	***331**
	tar - L - title	.2240	*322
	tar - L - pos	**.2407**	***318**
CLEF 18	dense - B - best	**.4071**	**269**
	tar - B - title	*.1292	*378
	tar - B - pos	***.1971**	***342**
	tar - L - title	*.2709	*361
	tar - L - pos	**.3005**	**334**
CLEF 19 dta	dense - B - best	**.2547**	**268**
	tar - B - title	*.1177	**331**
	tar - B - pos	**.1681**	356
	tar - L - title	.1898	343
	tar - L - pos	**.2815**	**314**
CLEF 19 int.	dense - B - best	**.4100**	**175**
	tar - B - title	*.0937	*281
	tar - B - pos	***.1410**	***268**
	tar - L - title	*.2352	*257
	tar - L - pos	***.2578**	*257

in the systematic reviews field [41], but also explore the effects of varying k among $\{5, 10, 15, 25, 50\}$ in one of our research questions.

Baselines. We first compare our method to BM25+RM3 (pseudo relevance feedback) for the effectiveness in initial rankings, where the corresponding dense retrievers do not use any feedback. We then select the best run (AP) submitted to CLEF for the feedback settings, with feedback and without stopping strategies. Specifically: From CLEF 17, *auth.simple.run1* [4], which was a hybrid classifier with LTR features iteratively trained with explicit feedback. From CLEF 18, *cnrs_comb* [37], which was a neural network trained from a logistic regression and a CAL model, using task description as seed. CLEF 19 had no suitable runs to compare with. We also compare against a recent BERT-based active learning workflow proposed for TAR [49]. We consider both neural and traditional linear classifiers for the TAR method. Specifically, we select BioLinkBERT [50] without further pre-training as suggested by [32] and logistic regression as in [49]. For a fair comparison, we apply the same recording approach to the TAR methods by concatenating the rank of feedback documents in each iteration as introduced below, which is different from the previous recording for TAR detailed in [32]. We test different seed document settings: the review title (title) as above and one relevant document (pos). We limit both the active learning and our feedback method to 20 iterations as running BERT-based active learning is computationally expensive, especially on large topics. For this setting, we run experiments on an NVIDIA H100 (80GB).

Evaluation. We examine a continuous, iterative relevance feedback task where subsets of documents are progressively re-ranked. Specifically, if n documents are ranked at iteration i, only $n - k$ are re-ranked at $i + 1$, with k being the feedback batch size. This results in $\lceil n/k \rceil$ rankings by the end. We unify these into a single ranking, maintaining the order from each iterative batch without revisions. This is distinct from some CLEF-TAR methods that may re-order already examined rank positions based on new relevance labels, potentially overestimating effectiveness. Our task focuses on whether relevance feedback can benefit screening prioritisation. Therefore, we exhaust all candidate documents under each topic and record the reviewed documents in each iteration. We then measure on the concatenated list of reviewed documents (feedback) generated within the dense retrieval framework, instead of examining a new ranked list as with a reranker. We use Average Precision (AP) and Last Relevant Found (Last Rel – the position of the last retrieved relevant document) to measure the effectiveness of the ranking methods. We also report the run time as the measure of efficiency.

4 RESULTS

In Table 1 we report the results obtained by BM25+RM3, the dense retrieval methods with/without feedback, and the best runs from CLEF when available. The results show that all dense retrievers except BioBERT outperform BM25+ RM3 in terms of AP when no feedback is considered, strengthening the use case of dense rankers in systematic review screening. Interestingly, the task-specific dense retriever (coCondenser) sometimes outperforms domain-specific retrievers such as BioBERT, which use a biomedical BERT. However, effectiveness differences are observed in Last Rel: methods that obtain high AP do not always obtain low Last Rel. For example, in CLEF 2018, despite significant AP gains by BioLinkBERT and

PubMedBERT, these are not reflected in Last Rel where, compared to BM25+RM3, reviewers have to screen about 2000 more documents.

When examining results obtained with explicit relevance feedback, we observe that this always improves AP across all dense retrievers, given a suitable weight setting. We then consider the impact of different weight combinations have on the dense relevance feedback; for this, we again analyse Table 1. We identify that most of the top-performing results for the dense retrievers that use domain-specific backbones are obtained when the same weight is assigned to the query, relevant documents and non-relevant document representations (i.e. setting (1,1,1)). For coCondenser, instead, improvements are generally observed when non-relevant documents are given lower weight (i.e. (1,1,0) and (1,0.8,0.2)).

In Table 1 we also contextualise the effectiveness of the examined relevance feedback technique with that of compatible runs submitted to the CLEF shared tasks. Direct comparison is difficult as we are not aware of the exact settings of feedback used by the CLEF runs (e.g., whether we use the same feedback size, or if examined documents are reordered once their relevance is observed). However, the results suggest that the feedback method studied here provides similar effectiveness, and warrants further comparative analysis with previous methods (for which code is often not released). We then would like to see how the dense retrieval method, which does not involve iterative fine-tuning, compares to the popular TAR method, where a classifier is continuously trained. The TAR workflow consists of an active learning strategy and a classifier. Previous studies report higher effectiveness with a relevance feedback strategy [32], that is, screening the top-k documents suggested by the model. We also follow this practice, as another common strategy, uncertainty sampling [39] does not promote documents for higher relevance and is therefore not suitable for screening prioritisation. Another typical feature of the TAR approach is that it requires at least one relevant document as the seed to initialise. Typically, the seed documents are randomly sampled from the relevant document pool for experiments [49]. This practice may fit scenarios where certain related studies are identified prior to screening for a systematic review and can lead to higher effectiveness of the classifier. However, it is not fair to directly compare to the proposed dense retrieval method, where only topic-related information is used as the query. Additionally, to measure how these methods can prioritise relevant documents during the screening phase, we keep tracking those feedback sets and concatenate them into an overall ranking.

We report our results in Table 2, where a cutoff is set at 20 feedback iterations (totalling 25 x 20 = 500 documents). For dense retrieval, we select BioLinkBERT as the backbone and report the best result from the Rocchio settings we use. For the TAR workflow, we examine BioLinkBERT and logistic regression as the classifiers. Generally, we find the linear model to be more effective compared to the BERT-based model for TAR, with a gap of at least 10% in AP across all the CLEF collections, showing less difference in terms of Last Rel. When initiated with a relevant document, both methods show large improvements compared with using the title, which suggests TAR methods rely on a good seed to start. When turning to the dense retrieval method, however, it shows significantly higher effectiveness on both AP and Last Rel, except for in CLEF 19 dta, where TAR with logistic regression has 0.2815 in AP and 314 in Last Rel, whereas the dense retrieval method has 0.2547 and 268, respectively. This shows that the query-driven dense retrieval methods can effectively prioritise relevant documents with fewer iterations, with no prior reviewed relevant document involved. In terms of efficiency, in the 20 iteration setting, the dense retrieval method takes ≈9 seconds per topic, which is competitive with that of TAR using logistic regression, which takes on average 5 seconds per topic. However, TAR with BioLinkBERT requires ≈10 minutes per topic (including fine-tuning and inference in each iteration).

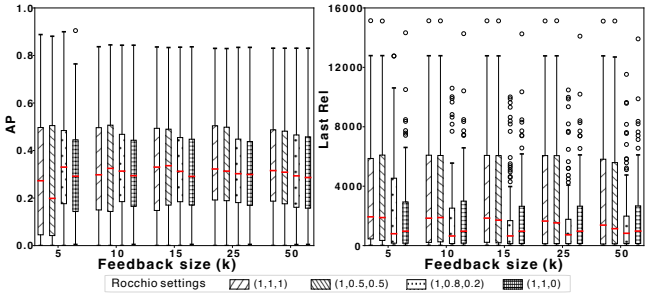

Figure 2: Distributions of Average Precision (left) and Last Relevant (right) across all CLEF collections, faceted by Rocchio setting and relevance feedback size (i.e. value of k).

Finally, we study how relevance feedback effectiveness changes with the batch size (previously $k = 25$). We use the BiolinkBERT retriever and vary k from 5 to 50; results are shown in Figure 2. For AP, in general, the larger the feedback size the higher the effectiveness. Interestingly, we find that a small feedback size (e.g., $k = 5$), which considers less feedback between query representation updates, is not more effective than larger ones. This might be because relevant documents are rare in the assessment pool and thus smaller batches are more likely to contain only non-relevant documents. As in previous results, this reduces ranking effectiveness. Further, using larger feedback sizes is computationally beneficial as it requires fewer updates and re-rankings of the query representation. We also observe that for AP, the best weight setting for the representations varies across feedback sizes, though most differences are not significant. For Last Rel it appears that settings that put less importance on the feedback (and especially on the negative one) consistently yield higher effectiveness than other settings.

5 CONCLUSION

We considered the context of screening prioritisation for systematic review automation and adapted a generic relevance feedback mechanism that exploits dense retrieval. Unique to our settings is the fact that feedback is explicit and continuous, i.e. is provided iteratively as users screen documents. Through extensive empirical experimentation, we reported that this method can achieve similar or better effectiveness in terms of AP and Last Rel compared to methods specifically designed for this task. In addition, it is computationally efficient as there is no need to re-train the ranker at each relevance feedback iteration, making it suitable for use in practice.

ACKNOWLEDGMENT

Xinyu Mao is supported by the Australian Research Council Discovery Project DP210104043.

REFERENCES

[1] Amal Alharbi, William Briggs, and Mark Stevenson. 2018. Retrieving and Rranking Studies for Systematic Reviews: University of Sheffield's Approach to CLEF eHealth 2018 Task 2. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Vol. 2125.

[2] Amal Alharbi and Mark Stevenson. 2017. 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*, Vol. 1866.

[3] Amal Alharbi and Mark Stevenson. 2019. Ranking Studies for Systematic Reviews using Query Adaptation: University of Sheffield's Approach to CLEF eHealth 2019 Task 2. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, Vol. 2380.

[4] Antonios Anagnostou, Athanasios Lagopoulos, Grigorios Tsoumakas, and Ioannis Vlachavas. 2017. Combining Inter-Review Learning-to-Rank and Intra-Review Incremental Training for Title and Abstract Screening in Systematic Reviews. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*, Vol. 1866.

[5] Tim Baumgärtner, Leonardo FR Ribeiro, Nils Reimers, and Iryna Gurevych. 2022. Incorporating Relevance Feedback for Information-Seeking Retrieval using Few-Shot Document Re-Ranking. *arXiv preprint arXiv:2210.10695* (2022).

[6] Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open* 7, 2 (2017).

[7] Jiayi Chen, Su Chen, Yang Song, Hongyu Liu, Yueyao Wang, Qinmin Hu, Liang He, and Yan Yang. 2017. ECNU at 2017 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Vol. 1866.

[8] Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*. 153–162.

[9] Gordon V Cormack and Maura R Grossman. 2016. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 75–84.

[10] Gordon V Cormack and Maura R Grossman. 2017. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2017. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*, Vol. 1866.

[11] Gordon V Cormack and Maura R Grossman. 2018. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Vol. 2125.

[12] Giorgio Maria Di Nunzio. 2019. A Distributed Effort Approach for Systematic Reviews. IMS Unipd at CLEF 2019 eHealth Task 2. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, Vol. 2380.

[13] Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. 2017. An Interactive Two-Dimensional Approach to Query Aspects Rewriting in Systematic Reviews. IMS Unipd At CLEF eHealth Task 2. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*, Vol. 1866.

[14] Giorgio Maria Di Nunzio, Giacomo Ciuffreda, and Federica Vezzani. 2018. Interactive Sampling for Systematic Reviews. IMS Unipd At CLEF 2018 eHealth Task 2. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Vol. 2125.

[15] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281* (2024).

[16] Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540* (2021).

[17] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval. *arXiv preprint arXiv:2203.05765* (2022).

[18] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2017. Automatic and semi-automatic document selection for technology-assisted review. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 905–908.

[19] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[20] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

[21] Vassil Kalphov, Georgios Georgiadis, and Leif Azzopardi. 2017. Sis at clef 2017 ehealth tar task. In *CEUR Workshop Proceedings*, Vol. 1866. 1–5.

[22] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*, Vol. 1866. 1–29.

[23] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 Technologically Assisted Reviews in Empirical Medicine Overview. In *CEUR Workshop Proceedings*, Vol. 2125.

[24] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*, Vol. 2380.

[25] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[26] Grace E Lee and Aixin Sun. 2018. Seed-driven document ranking for systematic reviews in evidence-based medicine. In *The 41st international ACM SIGIR Conference on Research and Development in information retrieval*. 455–464.

[27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[28] Dan Li and Evangelos Kanoulas. 2019. Automatic Thresholding by Sampling Documents and Estimating Recall. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, Vol. 2380.

[29] Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2023. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Transactions on Information Systems* 41, 3 (2023), 1–40.

[30] Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In *European Conference on Information Retrieval*. Springer, 599–612.

[31] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2356–2362.

[32] Xinyu Mao, Bevan Koopman, and Guido Zuccon. 2024. A Reproducibility Study of Goldilocks: Just-Right Tuning of BERT for TAR. In *European Conference on Information Retrieval*. Springer, 132–146.

[33] Jessie McGowan and Margaret Sampson. 2005. Systematic reviews need systematic searchers (IRP). *Journal of the Medical Library Association* 93, 1 (2005), 74.

[34] Matthew Michelson and Katja Reuter. 2019. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary clinical trials communications* 16 (2019), 100443.

[35] Alessio Molinari and Evangelos Kanoulas. 2022. Transferring knowledge between topics in systematic reviews. *Intelligent Systems with Applications* 16 (2022), 200150.

[36] Christopher Norman, Mariska Leeflang, and Aurélie Névéol. 2017. LIMSI@ CLEF eHealth 2017 Task 2: Logistic Regression for Automatic Article Ranking. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*, Vol. 1866.

[37] Christopher Norman, Mariska Leeflang, and Aurélie Névéol. 2018. LIMSI@ CLEF eHealth 2018 Task 2: Technology Assisted Reviews by Stacking Active and Static Learning. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Vol. 2125.

[38] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 4, 1 (2015), 5.

[39] Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American society for information science* 41, 4 (1990), 288–297.

[40] Harrisen Scells, Guido Zuccon, Anthony Deacon, and Bevan Koopman. 2017. QUT ielab at CLEF eHealth 2017 technology assisted reviews track: initial experiments with learning to rank. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*, Vol. 1866. 1–6.

[41] Gaurav Singh, James Thomas, and John Shawe-Taylor. 2018. Improving active learning in systematic reviews. *arXiv preprint arXiv:1801.09496* (2018).

[42] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 11, 1 (2010), 1–11.

[43] Junmei Wang, Min Pan, Tingting He, Xiang Huang, Xueyan Wang, and Xinhui Tu. 2020. A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Information Processing & Management* 57, 6 (2020), 102342.

[44] Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. From little things big things grow: A collection with seed studies for medical systematic review literature search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3176–3186.

[45] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2022. Neural Rankers for Effective Screening Prioritisation in Medical Systematic Review Literature Search. In *Proceedings of the 26th Australasian Document Computing Symposium*. 1–10.

[46] Huaying Wu, Tingting Wang, Jiayi Chen, Su Chen, Qinmin Hu, and Liang He. 2018. ECNU at 2018 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Vol. 2125.

[47] Eugene Yang, David D Lewis, and Ophir Frieder. 2019. Text retrieval priors for Bayesian logistic regression. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1045–1048.

[48] Eugene Yang, David D Lewis, and Ophir Frieder. 2021. Heuristic stopping rules for technology-assisted review. In *Proceedings of the 21st ACM Symposium on Document Engineering*. 1–10.

[49] Eugene Yang, Sean MacAvaney, David D Lewis, and Ophir Frieder. 2022. Goldilocks: Just-right tuning of bert for technology-assisted review. In *European Conference on Information Retrieval*. Springer, 502–517.

[50] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827* (2022).

[51] Hongchien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3592–3596.

[52] Zhe Yu and Tim Menzies. 2017. Data Balancing for Technologically Assisted Reviews: Undersampling or Reweighting. In *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*, Vol. 1866.

[53] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2021. Contextualized query expansion via unsupervised chunk selection for text retrieval. *Information Processing & Management* 58, 5 (2021), 102672.