Neural Rankers for Effective Screening Prioritisation in Medical Systematic Review Literature Search

Shuai Wang
University of Queensland
Brisbane, Australia
shuai.wang2@uq.edu.au

Bevan Koopman
CSIRO
Brisbane, Australia
b.koopman@csiro.com

Harrisen Scells
Leipzig University
Leipzig, Germany
harry.scells@uni-leipzig.de

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

ABSTRACT

Medical systematic reviews typically require assessing all the documents retrieved by a search. The reason is two-fold: the task aims for "total recall"; and documents retrieved using Boolean search are an unordered set, and thus it is unclear how an assessor could examine only a subset. *Screening prioritisation* is the process of ranking the (unordered) set of retrieved documents, allowing assessors to begin the downstream processes of the systematic review creation earlier, leading to earlier completion of the review, or even avoiding screening documents ranked least relevant.

Screening prioritisation requires highly effective ranking methods. Pre-trained language models are state-of-the-art on many IR tasks but have yet to be applied to systematic review screening prioritisation. In this paper, we apply several pre-trained language models to the systematic review document ranking task, both directly and fine-tuned. An empirical analysis compares how effective neural methods compare to traditional methods for this task. We also investigate different types of document representations for neural methods and their impact on ranking performance.

Our results show that BERT-based rankers outperform the current state-of-the-art screening prioritisation methods. However, BERT rankers and existing methods can actually be complementary, and thus, further improvements may be achieved if used in conjunction.

KEYWORDS

Systematic Reviews, Neural Ranker, Screening Prioritisation

1 INTRODUCTION

In medicine, systematic reviews are considered the most comprehensive and reliable instrument to synthesise evidence for a specific research question. When searching for documents for a systematic review, all retrieved documents are assessed to ensure the systematic review is comprehensive and correct. Arguably, however, the only reason that all documents must be assessed is that the documents retrieved from typical databases such as PubMed are returned as an unordered set, thus all equally relevant to the query. A typical systematic review requires upwards and beyond 10,000 documents to be assessed [45]. Naturally, this assessment process, called *screening*, is costly and is further exasperated given the requirement that the documents should be screened multiple times by different assessors to account for biases and disagreements [17, 45]. Screening is executed by assessing the title and abstract of a retrieved publication. Once a document is considered relevant at a title-abstract level, it must also be assessed at the full-text level. Overall, it is a laborious process.

As a way to reduce systematic review creation effort, time delays and costs, the systematic review community has looked at adopting automation tools [32]. One of the tasks for which automation tools can be helpful is *screening prioritisation*. Here, retrieved documents are ordered by their relevance to the review. Screening prioritisation can reduce the time and cost factors associated with systematic review creation in two ways: (1) researchers can begin and complete downstream tasks, such as full-text screening in parallel, earlier than if publications were assessed in random order; and (2) researchers can stop screening early by only considering the top-k ranked documents, possibly with a certain level of confidence that "total recall" (or some approximation of) has been achieved. Numerous methods have been proposed by the Information Retrieval community to address this problem [1, 2, 6, 8, 14, 15, 21, 39, 49, 51, 55, 56]. These methods can be separated into two classes: (1) methods that directly use queries to rank, where the queries can be the title [1, 2], the review's Boolean query [1–3, 38, 55], the objectives of the review [39, 43], or a set of studies known a priori [21, 49, 51]; and (2) use different ranking techniques, including relevance feedback [3], and various forms of active learning [6, 8, 14, 15], without the need for a query. While initial attempts to use new state-of-the-art pre-trained language models have been made in the second-ranking

category of methods [56], no research has considered using pre-trained language models for the first, query-focused category; this is the focus of this paper.

In this paper, we investigate the use of rankers based on pre-trained-language models for the systematic review screening prioritisation task. Specifically, we apply different types of zero-shot pre-trained models to investigate the effectiveness of different pre-training types. We also fine-tune the ranker for the screening prioritisation task and use different fine-tuning approaches to investigate the effectiveness of fine-tuning. Lastly, we perform extensive analysis to show how the effectiveness of neural methods differs from current state-of-the-art non-neural screening prioritisation methods. We make the following contributions:

(1) We conduct experiments using multiple neural rankers both under zero-shot and fine-tuned settings. We compare these rankers with the state-of-the-art methods for the task to show that they significantly outperform the current best ranking methods. We obtain similar results to methods that use interactive ranking approaches with relevance feedback (i.e., active learning).

(2) We use two representations for candidate documents to fine-tune our pre-trained language models: Title (title only) and TiAb (title+abstract), and we show the effectiveness difference between these two representations. We find that encoding the abstract with the title (TiAb) is far better than title only to represent candidate documents.

(3) We perform a query-by-query comparative analysis on the best neural and state-of-the-art methods. We show that although the average performance of fine-tuned BioBERT and the current state-of-the-art iterative methods obtain similar evaluation results, the effectiveness differs significantly at the topic level. The finding indicates that neural methods may achieve even higher performance if the current state-of-the-art active learning methods are combined with neural rankers.

2 RELATED WORK

2.1 IR for Medical Systematic Review Automation

Medical systematic reviews follow a standardised process to ensure consistency and quality. The most laborious part of the process is the screening of documents [45]. These documents are retrieved using a complex Boolean query that attempts to precisely encode the information need required to answer the research question of the systematic review and to ensure the reproducibility of the review (i.e., re-running the search in the future should produce the same set of documents) [13, 28, 44]. The Boolean query retrieves an unordered set of documents. However, ranking the set of retrieved documents has two main advantages:

(1) Systematic review creation often comes with a budget, which means screening of all retrieved documents may be impossible

with a small budget or limited time (i.e. rapid reviews) [29, 30]. To help systematic reviewers obtain the most relevant documents within a budget, an effective ranking of documents can ensure that the most relevant documents are found without having to screen all documents.

(2) Systematic reviews require a two-stage screening process: first, documents are assessed by only title and abstract; only if relevant are full-text documents reviewed. These two steps are run sequentially, and for each document, full-text screening can start as soon as that document has undergone the title-abstract level screening. Therefore, exhausting the screening of all relevant documents before starting the screening of non-relevant ones (a task achieved through high-quality screening prioritisation) leads to downstream steps taking place in parallel with the remaining screening. This ultimately allows the whole systematic review to conclude earlier than if screening prioritisation was not implemented.

To achieve effective ranking in systematic review document screening, research has investigated using (1) *One-off Ranking:* methods that use queries to obtain one-off rankings of candidate documents [1, 2, 21, 39, 43, 49, 51], and (2) *Iterative Ranking:* methods that iteratively acquire user feedback during document screening and re-rank the unjudged documents [6, 8, 14, 15]. These methods could be used in conjunction; e.g., start off with a one-off ranking and then update the ranking as feedback is received.

Aside from ranking, IR research has considered techniques for query improvement and reformulation that actually reduce the number of documents needing screening [40–42, 48, 50]. Although these methods affect screening, and could be used in conjunction with screening prioritisation, they take place before the screening task, and thus we do not consider them further in this paper.

2.2 Pre-trained Language Models

Advances in pre-trained language models like BERT [9], Roberta [27] and T5 [36] show effective on multiple downstream tasks: document ranking [11, 25, 52], question-answering [35], and conversational search [10, 35]. These language models are pre-trained on large text corpora (Wikipedia [9] or PubMed [16, 22, 34]) to learn textual features such as sentence structure, word semantics and sentence semantics, before being applied to specific tasks, such as document ranking. There are typically two ways to apply pre-trained language models to downstream tasks: (1) apply the model directly to the task, i.e., *zero-shot*; or (2) using training samples from the downstream task to *fine-tune* the pre-trained language model.

Currently, these neural rankers have yet to be investigated for the systematic review screening prioritisation task. The only related work is Yang et al. [56], which trains classifiers based on pre-trained language models to perform automatic Technology Assisted Reviews in domains other than systematic review literature search. In this paper, we propose methods and deliver initial experiments to examine both zero-shot and fine-tuned neural methods for the screening prioritisation task in the systematic review and their corresponding effectiveness, thus filling this gap.

3 NEURAL RANKERS FOR SCREENING PRIORITISATION

3.1 Model Architecture

In this paper, we examine two different avenues for using pre-trained language model based rankers for the screening prioritisation task: (1) zero-shot and (2) fine-tuned. Both approach rely on the typical monoBERT cross-encoder architecture to compute scores [31]. For each query-document pair, we concatenate the text of a document d with that of the query q (separated by a SEP token) and encode this input to obtain the relevance score of d given q.

In the *zero-shot* setting, we use the pre-trained language models directly on the screening prioritisation task. In our experiments, we consider the BERT base model [9], the BERT base model fine-tuned on the MS MARCO dataset [12], and an array of BERT models pre-trained on different medical-specific text corpora: BioBERT [22], PubMedBERT [16] and BlueBERT [34].

In the *fine-tuned* setting, we further fine-tune the BERT model using the training portion in our dataset and then apply the resulting ranker on the screening prioritisation task on the test portion of the dataset. For fine-tuning, we experiment with BERT base, BERT fine-tuned on MS MARCO and BioBERT. We use localised contrastive loss with triples of $< title, D^+, set(D^-) >$ where D is the representation of a candidate document, D^+ is a document judged relevant at the abstract level, D^- is conversely a document judged not-relevant at the abstract level. We use the framework proposed by Gao et al. [12] and default learning parameters to develop our model.

3.2 Documents Representation

We also investigate two ways to represent a document: (1) *Title*, and (2) Title and Abstract (*TiAb*) concatenated and separated by BERT's [*SEP*] token. Using the *Title* representation instead of the *TiAb* may be reasonable because BERT has an input limit of 512 tokens; concatenating the title and abstract with the query may exceed the length of the BERT input, and the text input will then be truncated (i.e., the exceeded tokens discarded). This truncation may remove information in some of the long training samples, and possibly cause training mismatch as some samples may be truncated and others not. A shorter input size also results in slightly faster inference (i.e., reduced latency). We perform the comparison between the two representations for the fine-tuned setting only; for the zero-shot setting, we only investigate the *TiAb* representation – this is because our early experiments showed this to be highly ineffective for zero-shot; thus, we do not retain this of interest for the paper.

4 EXPERIMENTAL SETUP

4.1 Dataset & Evaluation

We use three CLEF Technological Assisted Review (TAR) datasets to evaluate the effectiveness of the screening prioritisation task [18–20]. The CLEF TAR datasets contain 50 systematic review topics in 2017 (20 training, 30 testing) [18] and 80 topics in 2018 (50 training, 30 testing) [20]. All topics in CLEF TAR 2017 and 2018 datasets are Diagnostic Test Accuracy (DTA) systematic reviews. The CLEF TAR 2019 dataset includes 40 intervention systematic review topics (20 training, 20 testing), 88 DTA topics (80 training, 8 testing), one

prognosis review and one qualitative review [19]. For each topic, the collection contains the title of the review, the Boolean query used for document retrieval, the documents (retrieved by issuing the Boolean query), and the relevance labels for the documents (abstract-level and full-text level). Our experiments use the training and testing portions pre-defined in the above datasets. We use all systematic review topics from CLEF TAR 2017 and 2018. For CLEF TAR 2019, we only use intervention (2019-intervention) and DTA (2019-dta) topics as there is no training data for the other types of reviews.

We use the title of the review for each topic as the query to rank documents. In the CLEF TAR dataset, documents are represented by their PubMed PMID (i.e., a document identifier). In our experiment, we obtain the title and abstract of these documents directly from the PubMed index.

We consider only abstract-level relevance for our evaluation, as PubMed full-text is not publicly available from its index. We use the same evaluation measures as CLEF TAR: the rank position of the last relevant document in the ranking (Last_Rel), AP, Recall@p% ($p = 1, 5, 10, 20$), and Work Saved Over Sampling (WSS) at k% ($k = 95\%, 100\%$). Intuitively, WSS@k measures the fraction of the screening workload saved if one stops examining the ranking once k% of the relevant documents have been found compared to screening the entire set of documents [5]. All evaluation measures are computed using the script provided in CLEF TAR 2018.

4.2 Baseline Methods

We use traditional exact word-matching methods, including the Query Likelihood Model (QLM) and BM25, as baseline methods. To compute the QLM score of a query-document pair, we use Jelinek-Mercer (JM) smoothing [58]. To compute BM25 scores, we used the Gensim toolkit [37].

Additionally, for each CLEF TAR dataset, we obtained the best-performing runs from the campaign results and available data — we deem these as being the current state-of-the-art (noting there does not seem to be follow-up work that significantly outperformed these runs). Note that we can only use runs that do not cut off the document list for a fair comparison. The runs selected are then:

- CLEF TAR 2017: `waterloo.A-rank-normal` [6];
- CLEF TAR 2018: `UWB` [7];
- CLEF TAR 2019 dta: `Sheffield/DTA/DTA_sheffield-Odds_Ratio` [3];
- CLEF TAR 2019 Intervention: `Sheffield/DTA/DTA_sheffield-Log_Likellihood` [2].

CLEF TASK participants were allowed to use *Iterative Ranking*, that is, use the relevance assessments of the topics for explicit relevance feedback. Thus simulating the user provides feedback on results in an interactive manner. However, we only consider the setting of the screening prioritisation task with no feedback. For the CLEF TAR 2017 and 2018 datasets, we could identify the runs that did not use feedback via the task overview papers. The selected runs for CLEF-2017 and CLEF-2018 are:

- CLEF TAR 2017: `sheffield.run4` [2];
- CLEF TAR 2018: `shef-general` [1];

However, for CLEF TAR 2019, we could not distinguish between feedback and non-feedback runs; therefore, we consider the overall

best run as the state-of-the-art, noting this run may or may not be using feedback.

4.3 Fine-tuning Details

In our experiments that involve fine-tuning, we fine-tune the BERT models using a group size of 10: for each training step, one positive sample and nine negative samples are used to compute the loss. We use a batch size of 3. Once the models are fine-tuned, we concatenate the systematic review title with the representation of the document to compute their relevance scores. In the experiments, we report the models' results using the last model checkpoint that has been fine-tuned for 100 epochs. We also discuss model convergence by investigating the test effectiveness on every saved checkpoint (saved every 100 training steps).

5 RESULTS

5.1 Zero-shot Prioritisation

Table 1 reports the effectiveness of the neural methods under the zero-shot setting, along with that of the baselines considered. In the table, BERT-M refers to BERT fine-tuned on the MS MARCO dataset[1]. For these experiments, we consider the title and abstract (TiAb) representation for documents.

QLM and BM25 outperform the zero-shot neural methods across all evaluation metrics, with a few exceptions: (1) the zero-shot BERT outperforms the other methods in all datasets but 2019-dta for WSS100, in 2019-intervention for WSS95, and in 2017 and 2019-intervention for Last_Rel; (2) the zero-shot BERT-M outperforms the other methods in 2017 and 2018 for WSS95, and in 2018 for Last_Rel.

Comparing the different zero-shot neural rankers, we find that generally, BERT and BERT-M perform similarly and better than the remaining domain-specific models; among these remaining models, BioBERT is the one that performs best. We further note that the fine-tuning on the MS MARCO dataset (BERT-M), and thus to a different but related task to screening prioritisation, does not lead to significant effectiveness boosts compared to BERT. Overall, the use of zero-shot neural rankers for the task of screening prioritisation does not appear to be a competitive and viable approach to the task.

5.2 Fine-tuned Prioritisation

Table 2 reports the effectiveness of the neural rankers under the fine-tuned settings. For these experiments, we only considered fine-tuning BERT, BERT-M and BioBERT, the three best-performing rankers in the zero-shot setting. The results in the table refer to using the title and abstract (TiAb) representation.

Firstly, we observe that the fine-tuning regime greatly improves the effectiveness of the considered neural rankers over the zero-shot setting of Table 1. While this may be somewhat expected, we highlight that the considered datasets contain only a small portion of training samples available for fine-tuning — thus, even this small signal is enough to train far more effective neural rankers.

Next, we compare the three pre-trained language rankers. Even though BERT and BERT-M obtained higher effectiveness in the

zero-shot setting, when fine-tuning is performed, BioBERT often obtains higher effectiveness among the three rankers.

We now turn our attention to comparing the neural rankers with the other considered baselines. The fine-tuned neural rankers are now able to outperform the baseline methods across all measures and datasets significantly. This finding strengthens our previous insights that pre-trained neural models should be fine-tuned for the systematic review screening prioritisation task, even if just on a few training samples.

Table 2 also reports the effectiveness of the best no-feedback methods submitted to the respective CLEF TAR tasks[2]. We find that all the fine-tuned neural rankers considered here achieve higher effectiveness than the best no-feedback runs for the corresponding CLEF year, with minor exceptions. The table also reports the effectiveness of the best iterative run (i.e., feedback via relevance assessment) at CLEF 2017 and 2018. We find that the fine-tuned BioBERT ranker consistently achieves higher effectiveness than the best iterative runs for Recall@1%, Recall@5%, AP and WSS100. We also find that the difference in effectiveness is significantly higher for shallow evaluation metrics while it is marginally higher for deep evaluation metrics. This finding can be explained by the fact that the iterative runs use feedback and some form of active learning, and thus, while effectiveness may be low in the first rank positions[3], as shown for example by Recall@1% and 5%, they improve as more feedback is accumulated. This point is interesting because it suggests that the neural rankers investigated here, which are highly effective, especially at the top of the ranking, could be used within an iterative, active learning loop in a bid to further improve effectiveness throughout the whole of the ranking.

5.3 Model Convergence

We investigate the neural rankers' convergence during the fine-tuning process. Figure 1 plots the AP values achieved on the test sets across subsequent steps of fine-tuning. From the figure, we deduce that ranker fine-tuning effectiveness saturates after about 100 epochs; this is confirmed by a statistical analysis that shows no significant differences are found in terms of test effectiveness across rankers' checkpoints beyond this training step (paired two-tailed t-test with Bonferroni correction, $p < 0.05$).

We further note that the results reported in Table 2 are not necessarily the best results these neural rankers could achieve. In fact, in Table 2, for each ranker, we reported the effectiveness of the last checkpoint – obtained after 100 fine-tuning epochs. However, the best test effectiveness is actually achieved by earlier checkpoints. If an effective way to detect the optimal point at which fine-tuning should be stopped, higher test effectiveness than that reported in Table 2 would be achieved. We note that one approach to this is using a validation dataset (though not guaranteeing optimal convergence): checkpoints would be measured against this dataset, and fine-tuning stopped after improvements below a threshold have been observed. No validation set was provided in CLEF, and we decided that splitting the training set to have a small validation set would have resulted in too little data for training (and too little data for validation, making validation unreliable).

---

[1]We consider this a zero-shot method because fine-tuning is done for a different dataset than the target one, and a task of Adhoc web search is different to screening prioritisation.

[2]Recall that it is not possible to determine a CLEF 2019 run uses feedback data or not.
[3]Which correspond to the first few iterations.

Dataset	Method	Last_Rel	AP	Recall@1%	Recall@5%	Recall@10%	Recall@20%	WSS95	WSS100
2017	BM25	2999.7000	$0.1497^\dagger$	$0.0931^\dagger$	$0.2717^\dagger$	$0.3851^\dagger$	$0.5737^\dagger$	0.3518	0.2520
	QLM	2999.5333	$\mathbf{0.1721}^\dagger$	$\mathbf{0.1071}^\dagger$	$\mathbf{0.2849}^\dagger$	$\mathbf{0.4067}^\dagger$	$\mathbf{0.6340}^\dagger$	0.3588	0.2588
	BERT	**2898.0333**	0.0917	0.0119	0.1052	0.2594	0.4386	0.3565	**0.3082**
	BERT-M	3161.5667	0.1247	0.0220	0.1324	0.2682	0.5265	**0.3627**	0.2606
	BioBERT	3358.2667	0.0998	0.0301	0.1801	0.2922	0.4372	$0.2373^\dagger$	$0.1836^\dagger$
	BlueBERT	$3824.4667^\dagger$	$0.0365^\dagger$	$0.0001^\dagger$	$0.0113^\dagger$	$0.0349^\dagger$	$0.0935^\dagger$	$0.0233^\dagger$	$0.0407^\dagger$
	PubMedBERT	$3628.5000^\dagger$	0.0670	0.0156	0.0667	$0.1433^\dagger$	$0.2717^\dagger$	$0.1205^\dagger$	$0.1018^\dagger$
2018	BM25	6095.2000	$\mathbf{0.1683}^\dagger$	$\mathbf{0.0972}^\dagger$	$\mathbf{0.2947}^\dagger$	$0.4444^\dagger$	$0.6441^\dagger$	0.4202	0.2336
	QLM	5956.9667	$0.1660^\dagger$	$0.0896^\dagger$	$0.2904^\dagger$	$\mathbf{0.4544}^\dagger$	$\mathbf{0.6456}^\dagger$	0.4322	0.2540
	BERT	6158.2333	0.1249	0.0222	0.1348	0.2960	0.5422	0.4037	**0.2622**
	BERT-M	**5805.4667**	0.1398	0.0188	0.1465	0.2997	0.5894	**0.4476**	0.2567
	BioBERT	6696.9333	0.0881	0.0236	0.1526	0.2580	$0.4059^\dagger$	$0.1499^\dagger$	$0.0836^\dagger$
	BlueBERT	$7204.3667^\dagger$	$0.0329^\dagger$	$0.0014^\dagger$	$0.0093^\dagger$	$0.0198^\dagger$	$0.0493^\dagger$	$0.0017^\dagger$	$0.0193^\dagger$
	PubMedBERT	$6955.8333^\dagger$	$0.0700^\dagger$	0.0112	$0.0636^\dagger$	$0.1593^\dagger$	$0.3138^\dagger$	$0.1424^\dagger$	$0.0883^\dagger$
2019-dta	BM25	2722.7500	0.1185	0.0479	0.2129	**0.3290**	0.5276	0.3138	0.2080
	QLM	**2318.2500**	**0.1223**	**0.0644**	**0.2164**	0.3270	**0.5335**	**0.3470**	**0.2477**
	BERT	2513.8750	0.0922	0.0244	0.1318	0.2381	0.3906	0.2577	0.2095
	BERT-M	3233.7500	0.0955	0.0105	0.0792	0.1979	0.3793	0.2629	0.1232
	BioBERT	3264.0000	0.0810	0.0160	0.1290	0.2294	0.3365	0.1370	0.0950
	BlueBERT	3771.0000	0.0688	0.0010	0.0256	0.0526	0.1050	0.0227	0.0160
	PubMedBERT	3330.2500	0.1044	0.0335	0.1226	0.2144	0.3119	0.2016	0.0979
2019-int.	BM25	1715.6000	0.2112	0.0968	**0.3053**	**0.3989**	**0.5542**	0.3510	0.2955
	QLM	1724.0500	**0.2123**	**0.0981**	0.2793	0.3851	0.5110	0.3397	0.2939
	BERT	**1398.5500**	0.1603	0.0536	0.2104	0.3282	0.5041	**0.3624**	**0.3330**
	BERT-M	1836.2000	0.1769	0.0384	0.1951	0.3545	0.5268	0.3228	0.2663
	BioBERT	1832.8500	0.1463	0.0530	0.1346	0.1982	$0.3074^\dagger$	$0.1585^\dagger$	$0.1631^\dagger$
	BlueBERT	2057.0000	0.0462	0.0063	$0.0275^\dagger$	$0.0513^\dagger$	$0.1066^\dagger$	$0.0083^\dagger$	$0.0361^\dagger$
	PubMedBERT	1974.2500	0.0780	0.0124	0.0502	$0.0905^\dagger$	$0.2748^\dagger$	$0.1207^\dagger$	$0.0944^\dagger$

Table 1: Results obtained using pre-trained language models in a zero-shot setting. Statistical significant differences (Student's two-tailed paired t-test with Bonferonni correction, $p < 0.05$) between BERT and all other methods are indicated by $\dagger$.
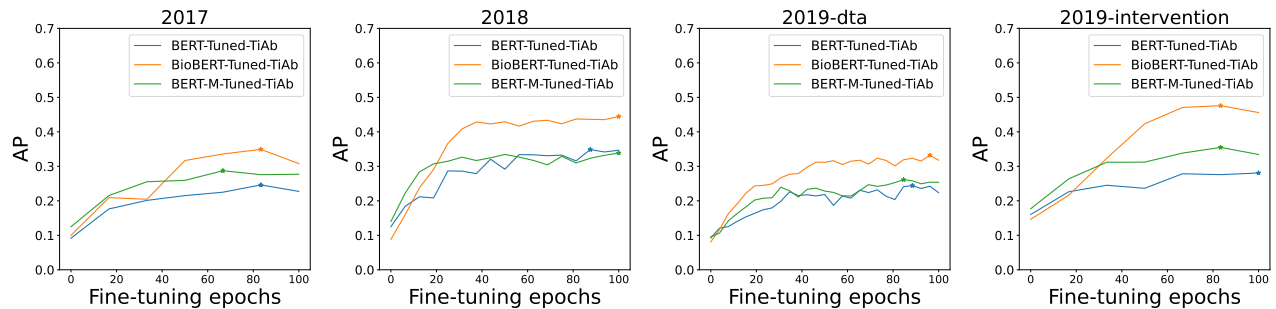

Figure 1: Convergence of neural rankers during fine tuning. The y-axis reports AP measured on the test set, while the x-axis corresponds to subsequent fine-tuning steps. AP measurements are taken every 100 training steps. For each neural ranker, the checkpoint with the highest test AP is marked with ∗.

Dataset	Method	Last_Rel	AP	Recall@1%	Recall@5%	Recall@10%	Recall@20%	WSS95	WSS100
2017	BEST-No-Feedback	2382.4667	$0.2179^\dagger$	0.1308	$0.3325^\dagger$	$0.4993^\dagger$	$0.6877^\dagger$	$0.4880^\dagger$	$0.3946^\dagger$
	BEST-Iterative	1469.4000	**0.3183**	0.1707	**0.5434**	**0.7322**	**0.8863**	**0.7009**	**0.6106**
	BM25	$2999.7000^\dagger$	$0.1497^\dagger$	$0.0931^\dagger$	$0.2717^\dagger$	$0.3851^\dagger$	$0.5737^\dagger$	$0.3518^\dagger$	$0.2520^\dagger$
	QLM	$2999.5333^\dagger$	$0.1721^\dagger$	$0.1071^\dagger$	$0.2849^\dagger$	$0.4067^\dagger$	$0.6340^\dagger$	$0.3588^\dagger$	$0.2588^\dagger$
	BERT-Tuned	$2418.8333^\dagger$	$0.2273^\dagger$	$0.1066^\dagger$	$0.4088^\dagger$	$0.5888^\dagger$	0.7789	0.5435	$0.4340^\dagger$
	BERT-M-Tuned	$2475.2000^\dagger$	0.2770	0.1419	$0.3727^\dagger$	$0.5843^\dagger$	0.7707	$0.5234^\dagger$	$0.4187^\dagger$
	BioBERT-Tuned	**1461.6667**	0.3078	**0.1845**	0.4903	0.6816	0.8355	0.6530	0.5913
2018	BEST-No-Feedback	5519.2000	$0.2584^\dagger$	$0.1287^\dagger$	$0.3827^\dagger$	$0.5449^\dagger$	$0.7295^\dagger$	$0.5520^\dagger$	$0.4314^\dagger$
	BEST-Iterative	**2655.0000**	0.3776	0.1854	0.5940	**0.7696**	**0.9149**	**0.7558**	**0.6104**
	BM25	$6095.2000^\dagger$	$0.1683^\dagger$	$0.0972^\dagger$	$0.2947^\dagger$	$0.4444^\dagger$	$0.6441^\dagger$	$0.4202^\dagger$	$0.2336^\dagger$
	QLM	$5956.9667^\dagger$	$0.1660^\dagger$	$0.0896^\dagger$	$0.2904^\dagger$	$0.4544^\dagger$	$0.6456^\dagger$	$0.4322^\dagger$	$0.2540^\dagger$
	BERT-Tuned	5581.7667	$0.3467^\dagger$	$0.1981^\dagger$	$0.5028^\dagger$	$0.6772^\dagger$	$0.8196^\dagger$	$0.6188^\dagger$	$0.4121^\dagger$
	BERT-M-Tuned	5185.5000	$0.3387^\dagger$	$0.1934^\dagger$	$0.4833^\dagger$	$0.6515^\dagger$	$0.8265^\dagger$	0.6559	$0.4815^\dagger$
	BioBERT-Tuned	4108.4000	**0.4444**	**0.2768**	**0.5975**	0.7574	0.8946	0.7194	0.6103
2019-dta	BEST	2183.5000	0.2477	0.1685	0.4391	0.5940	0.7421	0.4899	0.3470
	BM25	2722.7500	$0.1185^\dagger$	0.0479	0.2129	$0.3290^\dagger$	$0.5276^\dagger$	$0.3138^\dagger$	$0.2080^\dagger$
	QLM	2318.2500	$0.1223^\dagger$	0.0644	0.2164	$0.3270^\dagger$	$0.5335^\dagger$	$0.3470^\dagger$	$0.2477^\dagger$
	BERT-Tuned	1399.3750	0.2234	0.1580	0.4390	0.6013	0.7620	0.5870	0.4600
	BERT-M-Tuned	1178.0000	0.2535	0.2049	0.4474	0.5904	0.7536	0.6151	0.4997
	BioBERT-Tuned	**852.7500**	**0.3177**	**0.2604**	**0.4998**	**0.6710**	**0.8171**	**0.6857**	**0.5845**
2019-int.	BEST	1132.0000	$0.2929^\dagger$	0.1655	$0.4192^\dagger$	$0.5424^\dagger$	$0.7225^\dagger$	$0.4582^\dagger$	$0.3808^\dagger$
	BM25	1715.6000	$0.2112^\dagger$	$0.0968^\dagger$	$0.3053^\dagger$	$0.3989^\dagger$	$0.5542^\dagger$	$0.3510^\dagger$	$0.2955^\dagger$
	QLM	1724.0500	$0.2123^\dagger$	$0.0981^\dagger$	$0.2793^\dagger$	$0.3851^\dagger$	$0.5110^\dagger$	$0.3397^\dagger$	$0.2939^\dagger$
	BERT-Tuned	1374.3000	$0.2808^\dagger$	$0.1646^\dagger$	$0.3736^\dagger$	$0.5274^\dagger$	$0.6586^\dagger$	$0.3629^\dagger$	$0.3011^\dagger$
	BERT-M-Tuned	1571.2500	$0.3343^\dagger$	0.1614	$0.4021^\dagger$	$0.5649^\dagger$	$0.7061^\dagger$	$0.4461^\dagger$	$0.3623^\dagger$
	BioBERT-Tuned	**706.8500**	**0.4559**	**0.2155**	**0.5805**	**0.7374**	**0.8417**	**0.6462**	**0.5794**

Table 2: Results obtained when using pre-trained language models in the fine-tuned setting. Statistical significant differences (Student's two-tailed paired t-test with Bonferonni correction, $p < 0.05$) between BioBERT-Tuned and all other methods are indicated by †.

5.4 Document Representation

Document representation was title only (Title) and title and abstract (TiAb)[4] representations. Table 3 shows how document representation impacts neural rankers' effectiveness. We only consider fine-tuned rankers, as zero-shot rankers did not correspond to viable effectiveness for the screening prioritisation task.

From these results, we find that using title and abstract within the neural rankers significantly outperforms using the title only representation, regardless of the underlying pre-trained language model employed. This finding generalises across all CLEF TAR datasets and all evaluation metrics, except for Recall@1% for the BERT ranker on CLEF TAR 2017. This finding indicates that the abstracts contain essential information for the task.

5.5 Topic by topic analysis

The results in Section 5.2 and Table 2 show that the fine-tuned BioBERT ranker achieves comparable effectiveness to the best iterative methods submitted to the CLEF tasks. We note this result is achieved by averaging effectiveness across topics in each dataset, and the datasets have very few topics (60 topics overall in 2017 and 2018). Thus, the average may be highly influenced by outliers, and because of this, we perform a deeper, topic-by-topic analysis of the results. In particular, we compare the BioBERT ranker to the best iterative run for the respective CLEF datasets. This analysis is shown in Figure 2 as a gain-loss plot.

Topics with high effectiveness differ significantly between the two methods. Nearly half of the topics obtain higher effectiveness using the neural ranker that does not exploit feedback. Methods that use feedback will rely heavily on high effectiveness at early ranks. This is somewhat captured by comparing the WSS100 metric (a deep metric) with the Recall@1% metric (a relatively shallow

---

[4]Note the results of the TiAb representation correspond to those also reported in Table 2.

Dataset	Method	Last_Rel	AP	Recall@1%	Recall@5%	Recall@10%	Recall@20%	WSS95	WSS100
2017	BERT-Tuned-TiAb	**2418.8333**	**0.2273**	0.1066	**0.4088**	**0.5888**	**0.7789**	**0.5435**	**0.4340**
	BERT-Tuned-Title	3540.6667†	0.2037	**0.1119**	0.3429	0.4819†	0.6468†	0.2626†	0.1700†
	BERT-M-Tuned-TiAb	**2475.2000**	**0.2770**	**0.1419**	**0.3727**	**0.5843**	**0.7707**	**0.5234**	**0.4187**
	BERT-M-Tuned-Title	2953.4667	0.2299†	0.1082	0.3388	0.5084†	0.7027	0.4361†	0.3043†
	BioBERT-Tuned-TiAb	**1461.6667**	**0.3078**	**0.1845**	**0.4903**	**0.6816**	**0.8355**	**0.6530**	**0.5913**
	BioBERT-Tuned-Title	2072.1000†	0.2789†	0.1565	0.4176†	0.5937†	0.7919	0.5876†	0.4655†
2018	BERT-Tuned-TiAb	**5581.7667**	**0.3467**	**0.1981**	**0.5028**	**0.6772**	**0.8196**	**0.6188**	**0.4121**
	BERT-Tuned-Title	6087.9667	0.2652†	0.1303†	0.4054†	0.5667†	0.7328†	0.4898†	0.3004†
	BERT-M-Tuned-TiAb	**5185.5000**	**0.3387**	**0.1934**	**0.4833**	**0.6515**	**0.8265**	**0.6559**	**0.4815**
	BERT-M-Tuned-Title	5757.7667†	0.2661†	0.1466†	0.3967†	0.5738†	0.7523†	0.5333†	0.3356†
	BioBERT-Tuned-TiAb	**4108.4000**	**0.4444**	**0.2768**	**0.5975**	**0.7574**	**0.8946**	**0.7194**	**0.6103**
	BioBERT-Tuned-Title	4928.0000†	0.3557†	0.1862†	0.5202†	0.6952†	0.8524†	0.6527†	0.4813†
2019-dta	BERT-Tuned-TiAb	**1399.3750**	**0.2234**	**0.1580**	**0.4390**	**0.6013**	**0.7620**	**0.5870**	**0.4600**
	BERT-Tuned-Title	1597.2500	0.1851	0.1436	0.3802	0.4993†	0.6708†	0.5247	0.3892
	BERT-M-Tuned-TiAb	**1178.0000**	**0.2535**	**0.2049**	**0.4474**	**0.5904**	**0.7536**	**0.6151**	**0.4997**
	BERT-M-Tuned-Title	1798.6250	0.1858†	0.1195	0.3250†	0.5229	0.6831	0.5136	0.3729
	BioBERT-Tuned-TiAb	**852.7500**	**0.3177**	**0.2604**	**0.4998**	**0.6710**	**0.8171**	**0.6857**	**0.5845**
	BioBERT-Tuned-Title	1091.0000	0.2566†	0.1834	0.4761	0.6308	0.7876	0.6244	0.5095
2019-int.	BERT-Tuned-TiAb	**1374.3000**	**0.2808**	**0.1646**	**0.3736**	**0.5274**	**0.6586**	**0.3629**	**0.3011**
	BERT-Tuned-Title	1883.8500	0.2476	0.0953†	0.3231	0.4609	0.6134	0.3103	0.2763
	BERT-M-Tuned-TiAb	**1571.2500**	**0.3343**	**0.1614**	**0.4021**	**0.5649**	**0.7061**	**0.4461**	**0.3623**
	BERT-M-Tuned-Title	1653.6500	0.2702†	0.1045†	0.3488	0.5176	0.6891	0.3924	0.3288
	BioBERT-Tuned-TiAb	**706.8500**	**0.4559**	**0.2155**	**0.5805**	**0.7374**	**0.8417**	**0.6462**	**0.5794**
	BioBERT-Tuned-Title	1145.5000†	0.3677†	0.1694	0.4868†	0.6447†	0.7572†	0.5222†	0.4622

Table 3: Comparison of using *Title* vs *TiAb* as document representation. Statistical significance (Student's two-tailed paired t-test $p < 0.05$) between representation of two models is indicated by †.

metric, as it stops considering the ranking once 1% of the relevant documents have been retrieved, which in most cases is very early into the ranking and quite far from the end of the ranking). Similarly, observations can be made if precision at rank 5 (P@5, another shallow metric) was considered (not shown in the figure).

This illustrates that relevant feedback may not be needed for some topics as using neural rankers can already achieve significantly higher performance. Furthermore, this finding suggests that some topics may need feedback to ensure a higher screening prioritisation effectiveness. However, most relevance feedback ranking pipelines rely on the effectiveness of early ranked documents, especially in systematic review screening prioritisation, as relevant documents signals may be much harder to get than irrelevant ones. In Figure 2, we use Recall@1% to show the effectiveness of early ranks; we find that on the topic level, neural rankers achieved higher performance in the majority of the topics, suggesting that neural rankers may give much better relevance signal in early ranks, thus lead to further improvements when the iterative method is applied.

6 CONCLUSION

In this paper, we investigate the effectiveness of rankers based on pre-trained language models for the task of screening prioritisation for systematic review creation. We focus on non-iterative rankers:

those that produce a one-off ranking. In this context, we investigated neural rankers across two settings: zero-shot (apply the neural models without further training) and fine-tuned (apply the neural models after further training).

Our experiments show that while zero-shot neural rankers perform poorly, rankers fine-tuned on even a small amount of training data achieve significantly higher effectiveness than the current state-of-the-art non-iterative methods and comparable effectiveness to this task's current state-of-the-art iterative ranking methods. We also experimented with two different document representations (title only and title and abstract) and found that the abstract was essential for effective ranking.

An interesting finding is that state-of-the-art iterative methods are better than the neural rankers for topics with good-quality initial rankings. On the other hand, the neural methods provide far better early rankings for most of the topics and, for many of these topics, these early wins result in overall better deep metrics (WSS, AP). It seems reasonable then to hypothesise that neural rankers could be further improved if cast into the iterative setting. For this to be possible, effective ways to exploit relevance assessments in the context of these neural rankers are required. We note that Yang et al [56] have proposed a method for iteratively exploiting relevance assessments in the context of a classifier based on pre-trained
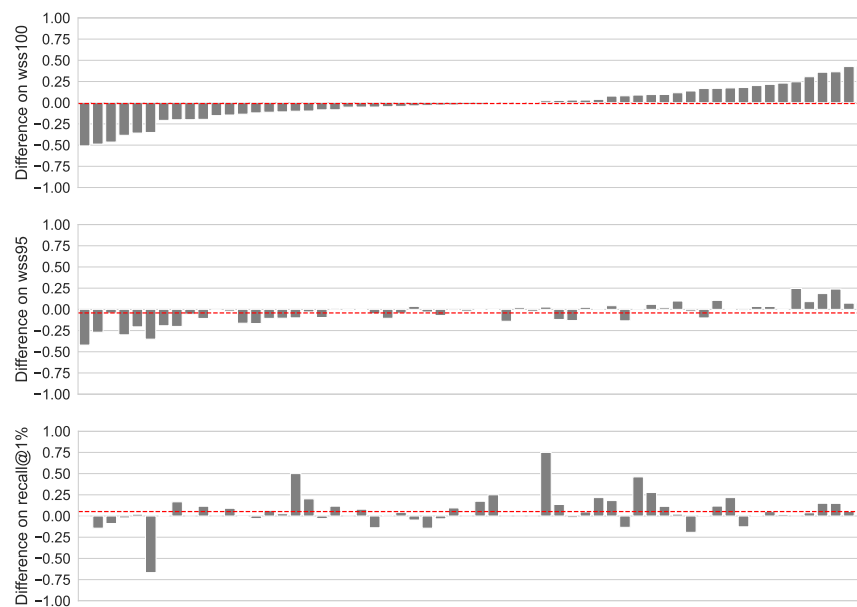
Figure 2: Per-topic effectiveness difference between BioBERT fine-tuned model and the best iterative runs for in CLEF TAR 2017 and 2018 (i.e., best active learning result of the year).

language models for Technology Assisted Review. This classifier bears similarity with our neural rankers. However, a clear drawback of their method is the high computational costs involved with their continuous learning approach and the consequent high latency imposed on the user. We highlight that expert screeners can return an assessment every 20–30 seconds [4, 46], a timeframe sensibly lower than the latency of Yang et al.'s method (1-2 hours). An alternative direction is considering methods for relevance feedback in the context of rankers based on pre-trained language models. Two main classes of methods have been proposed in this respect. The first class of methods combines the text input of the feedback with that of the query and the document to be scored [33, 47, 53]: these approaches are severely impacted by the language model's input size limit and would not be viable for the screening prioritisation task. The second class of methods instead combines the representations of feedback documents, not their text [23, 24, 26, 54, 57]: these methods trade-off some loss in effectiveness compared to the first class of methods for the ability to model indefinitely long feedback and for lower latency and computational costs. Immediate future work on the use of rankers based on pre-trained language models for screening prioritisation should be directed towards investigating and adapting these two classes of approaches.

Neural rankers show promise in terms of effectiveness for screening prioritisation in systematic review creation. These rankers have the potential to greatly reduce the effort of compiling systemic reviews. This impact is better quality and a greater number of systematic reviews being produced, which improves the important medical and policy decisions based on these reviews.

Acknowledgements.

REFERENCES

[1] Amal Alharbi, William Briggs, and Mark Stevenson. 2018. Retrieving and ranking studies for systematic reviews: University of Sheffield's approach to CLEF eHealth 2018 Task 2. In *CEUR Workshop Proceedings*, Vol. 2125. CEUR Workshop Proceedings.

[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 *CLEF (Working Notes)*.

[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 working notes for CLEF 2019. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, Vol. 2380. CEUR Workshop Proceedings.

[4] Justin Clark, Paul Glasziou, Chris Del Mar, Alexandra Bannach-Brown, Paulina Stehlik, and Anna Mae Scott. 2020. A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of clinical epidemiology* 121 (2020), 81–90.

[5] Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13, 2 (2006), 206–219.

[6] Gordon V Cormack and Maura R Grossman. 2017. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2017.. In *CLEF (Working Notes)*.

[7] Gordon V Cormack and Maura R Grossman. 2018. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2018. In *CLEF (Working Notes)*.

[8] Gordon V Cormack and Maura R Grossman. 2019. Systems and methods for conducting a highly autonomous technology-assisted review classification. US Patent 10,229,117.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Rafael Ferreira, Mariana Leite, David Semedo, and Joao Magalhaes. 2022. Open-domain conversational search assistants: the Transformer is all you need. *Information Retrieval Journal* 25, 2 (2022), 123–148.

[11] Luyu Gao and Jamie Callan. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 981–993.

[12] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In *European Conference on Information Retrieval*. Springer, 280–286.

[13] Chantelle Garritty, Gerald Gartlehner, Barbara Nussbaumer-Streit, Valerie J King, Candyce Hamel, Chris Kamel, Lisa Affengruber, and Adrienne Stevens. 2021. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. *Journal of clinical epidemiology* 130 (2021), 13–22.

[14] Maura R Grossman and Gordon V Cormack. 2011. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law & Technology* 17, 3 (2011), 11.

[15] 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.

[16] 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.

[17] 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.

[18] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. 2017. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In *CLEF'17*.

[19] 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.

[20] Evangelos Kanoulas, Rene Spijker, Dan Li, and Leif Azzopardi. 2018. CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*.

[21] 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 & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. ACM, New York, NY, USA, 455–464. https://doi.org/10.1145/3209978.3209994

[22] 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.

[23] Hang Li, Ahmed Mourad, Bevan Koopman, and Guido Zuccon. 2022. How Does Feedback Signal Quality Impact Effectiveness of Pseudo Relevance Feedback for Passage Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2154–2158. https://doi.org/10.1145/3477495.3531822

[24] Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2021. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *arXiv preprint arXiv:2108.11044* (2021).

[25] Hang Li, Shuai Wang, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. To Interpolate or Not to Interpolate: PRF, Dense and Sparse Retrievers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2495–2500. https://doi.org/10.1145/3477495.3531884

[26] 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 *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 599–612.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[28] Andrew MacFarlane, Tony Russell-Rose, and Farhad Shokraneh. 2022. Search Strategy Formulation for Systematic Reviews: issues, challenges and opportunities. *Intelligent Systems with Applications* (2022), 200091.

[29] Iain J Marshall, Rachel Marshall, Byron C Wallace, Jon Brassey, and James Thomas. 2019. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. *Journal of clinical epidemiology* 109 (2019), 30–41.

[30] 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. https://doi.org/10.1016/j.conctc.2019.100443

[31] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

[32] 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), 1–22.

[33] Ramith Padaki, Zhuyun Dai, and Jamie Callan. 2020. Rethinking query expansion for BERT reranking. In *European conference on information retrieval*. Springer, 297–304.

[34] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474* (2019).

[35] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 1133–1136.

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[37] Radim Rehurek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org* (2011).

[38] Harrisen Scells and Guido Zuccon. 2020. You Can Teach an Old Dog New Tricks: Rank Fusion Applied to Coordination Level Matching for Ranking in Systematic Reviews. In *42nd European Conference on IR Research, ECIR 2020*.

[39] 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 [CEUR Workshop Proceedings, Volume 1866]*. Sun SITE Central Europe, 1–6.

[40] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2019. Automatic Boolean query refinement for systematic review literature search. In *The world wide web conference*. 1646–1656.

[41] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2021. A comparison of automatic Boolean query formulation for systematic reviews. *Information Retrieval Journal* 24, 1 (2021), 3–28.

[42] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. Automatic boolean query formulation for systematic review literature search. In *Proceedings of The Web Conference 2020*. 1071–1081.

[43] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. Integrating the framing of clinical questions via PICO into the retrieval of medical literature for systematic reviews. In *CIKM'17*.

[44] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Shlomo Geva, and Leif Azzopardi. 2017. A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews. In *SIGIR'2017*.

[45] Ian Shemilt, Nada Khan, Sophie Park, and James Thomas. 2016. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic reviews* 5, 1 (2016), 140.

[46] 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.

[47] 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.

[48] Shuai Wang, Hang Li, Harrisen Scells, Daniel Locke, and Guido Zuccon. 2021. MeSH Term Suggestion for Systematic Review Literature Search. In *Proceedings of the 25th Australasian Document Computing Symposium*. 1–8.

[49] 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* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3176–3186. https://doi.org/10.1145/3477495.3531748

[50] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2022. Automated MeSH Term Suggestion for Effective Query Formulation in Systematic Reviews Literature Search. *arXiv preprint arXiv:2209.08687* (2022).

[51] Shuai Wang, Harrisen Scells, Ahmed Mourad, and Guido Zuccon. 2022. Seed-Driven Document Ranking for Systematic Reviews: A Reproducibility Study. In *European Conference on Information Retrieval*. Springer, 686–700.

[52] Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 317–324.

[53] Xiao Wang. 2022. Neural Pseudo-Relevance Feedback Models for Sparse and Dense Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3497–3497.

[54] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2021. Pseudo-relevance feedback for multiple representation dense retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 297–306.

[55] 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. *Methods* 4, 5 (2018), 7.

[56] 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.

[57] 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.

[58] ChengXiang Zhai and Sean Massung. 2016. *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool.