Systematic Evaluation of Neural Retrieval Models on the Touché 2020 Argument Retrieval Subset of BEIR

Nandan Thakur*
University of Waterloo
Waterloo, Canada

Luiz Bonifacio
UNICAMP and University of Waterloo
Campinas, Brazil

Maik Fröbe
Friedrich-Schiller-Universität Jena
Jena, Germany

Alexander Bondarenko
Leipzig University and Friedrich-Schiller-Universität Jena
Leipzig, Germany

Ehsan Kamalloo
University of Waterloo
Waterloo, Canada

Martin Potthast
University of Kassel, hessian.AI, and ScaDS.AI
Kassel, Germany

Matthias Hagen
Friedrich-Schiller-Universität Jena
Jena, Germany

Jimmy Lin
University of Waterloo
Waterloo, Canada

ABSTRACT

The zero-shot effectiveness of neural retrieval models is often evaluated on the BEIR benchmark—a combination of different IR evaluation datasets. Interestingly, previous studies found that particularly on the BEIR subset Touché 2020, an argument retrieval task, neural retrieval models are considerably less effective than BM25. Still, so far, no further investigation has been conducted on what makes argument retrieval so "special". To more deeply analyze the respective potential limits of neural retrieval models, we run a reproducibility study on the Touché 2020 data. In our study, we focus on two experiments: (i) a black-box evaluation (i.e., no model retraining), incorporating a theoretical exploration using retrieval axioms, and (ii) a data denoising evaluation involving post-hoc relevance judgments. Our black-box evaluation reveals an inherent bias of neural models towards retrieving short passages from the Touché 2020 data, and we also find that quite a few of the neural models' results are unjudged in the Touché 2020 data. As many of the short Touché passages are not argumentative and thus non-relevant per se, and as the missing judgments complicate fair comparison, we denoise the Touché 2020 data by excluding very short passages (less than 20 words) and by augmenting the unjudged data with post-hoc judgments following the Touché guidelines. On the denoised data, the effectiveness of the neural models improves by up to 0.52 in nDCG@10, but BM25 is still more effective. Our code and the augmented Touché 2020 dataset are available at https://github.com/castorini/touche-error-analysis.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; **Evaluation of retrieval results**.

*Corresponding author: <nandan.thakur@uwaterloo.ca>

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

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657861

KEYWORDS

Argument retrieval; Neural retrieval models; Model evaluation

ACM Reference Format:
Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic Evaluation of Neural Retrieval Models on the Touché 2020 Argument Retrieval Subset of BEIR. 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, 11 pages. https://doi.org/10.1145/3626772.3657861

1 INTRODUCTION

Substantial progress has been made in developing different types of neural retrieval models, including dense (e.g., [27, 31, 37, 71, 73]), sparse (e.g., [14, 21, 40, 75]), and multi-vector models (e.g., [23, 32, 36, 55]). However, evaluations on the BEIR retrieval benchmark [61] show that the effectiveness of neural models substantially varies across different tasks and especially drops for some that lack dedicated training data (e.g., argument retrieval), while simple lexical BM25 retrieval tends to be more robust [61]. To address this problem, numerous efforts have spurred to improve the neural models' effectiveness by optimizing the training stage via knowledge transfer from high-resource datasets (e.g., MS MARCO [43]), and with better mined hard negatives [4, 21, 44, 51, 55] by including an additional pretraining objective [23, 29, 69] or by using data augmentation via synthetic query generation [15, 61, 62]. Surprisingly, *all* neural models continue to be less effective than BM25 on the Touché 2020 [6] subset of BEIR, an argument retrieval task; cf. Table 1 with results for BM25 and state-of-the-art neural retrieval models like $E5_{large}$ [67], CITADEL+ [39], SPLADEv2 [21], etc.

Motivated by this observation, we conduct a two-stage reproducibility study on the Touché 2020 data to understand the potential respective limits of current neural retrieval models. Our first stage are black-box evaluations (i.e., without requiring model retraining) to examine and possibly somewhat correcting errors incurred by the neural models. In first analyses, we find that the neural models on average retrieve much shorter arguments than BM25 (Sections 4.1). For instance, about half of the top-10 results of dense retrievers (e.g., TAS-B [27]) contain at most two sentences that often are not even argumentative (e.g., "Pass" or "I agree with lannan13") yielding low effectiveness scores. To possibly improve the neural

Table 1: The motivation of our work: dense (left), multi-vector (top right), and sparse retrieval models (bottom right) are less effective than BM25 on the BEIR subset Touché 2020; nDCG@10 scores taken from the referenced publications.

Model	Reference	Type	nDCG@10
BM25 (BEIR)	Thakur et al. [61]	lexical	**0.367**
$E5_{large}$	Wang et al. [67]	dense	0.272
BGE-large	Xiao et al. [70]	dense	0.266
Promptagator	Dai et al. [15]	dense	0.266
DRAGON+	Lin et al. [41]	dense	0.263
GTR-XXL	Ni et al. [44]	dense	0.256
GPL	Wang et al. [66]	dense	0.255
RocketQAv2	Ren et al. [51]	dense	0.247
ANCE	Xiong et al. [71]	dense	0.240
RetroMAE	Xiao et al. [69]	dense	0.237
Contriever	Izacard et al. [29]	dense	0.204
TART-dual	Asai et al. [4]	dense	0.201
TAS-B	Hoffstätter et al. [27]	dense	0.162

Model	Reference	Type	nDCG@10
BM25 (BEIR)	Thakur et al. [61]	lexical	**0.367**
CITADEL+	Li et al. [39]	mult.-vec.	0.342
XTR (XXL)	Lee et al. [36]	mult.-vec.	0.309
CITADEL	Li et al. [39]	mult.-vec.	0.294
COIL-full	Gao et al. [24]	mult.-vec.	0.281
ColBERTv2	Santharam et al. [55]	mult.-vec.	0.263
ColBERT	Khattab et al. [32]	mult.-vec.	0.202
uniCOIL	Lin et al. [40]	sparse	0.298
SPLADEv2	Formal et al. [21]	sparse	0.272
SPLADE++	Lassance et al. [34]	sparse	0.244
DeepCT	Dai et al. [14]	sparse	0.175
SPARTA	Zhao et al. [75]	sparse	0.156

model's effectiveness, we then repeat the evaluation on augmented versions of the Touché 2020 corpus: (i) via document expansion[1] using DocT5query [45] (lengthening short arguments) and (ii) via document summarization with GPT-3.5 [46] (shortening longer arguments). The corpus augmentation does not require a retraining of the neural models and our results show that the effectiveness indeed increases for a majority of the models (Section 4.2).

In our second reproducibility stage, we analyze intrinsic characteristics of the Touché 2020 corpus and find that, unlike for other BEIR subsets, about 20% of the Touché 2020 corpus are very short documents (and thus mostly non-argumentative) and that at least 50% of the documents retrieved by neural models (and even BM25) are actually unjudged and thus considered non-relevant in standard evaluation setups. To counter these effects, we carefully remove short documents (less than 20 words) from the Touché 2020 corpus (Section 4.3) and we add missing judgments following the Touché guidelines (Section 4.4). Our experimental results show that without very short documents in the corpus and with added post-hoc judgments, the effectiveness of all neural models substantially improves by up to 0.52 in terms of nDCG@10. Yet, even after denoising and post-hoc judgments, BM25 remains the most effective.

We finally supplement our findings with a theoretical analysis using information retrieval axioms [9] on the Touché 2020 data (Section 4.5) and find that all neural models violate the document length normalization axiom LNC2 [18], which is well-supported by BM25. Overall, our contributions are the following:

- We reproduce dense, sparse, and multi-vector neural retrieval models on the BEIR subset Touché 2020 (argument retrieval) and find that short and low-quality arguments substantially harm the effectiveness of many neural models.
- After carefully denoising the Touché 2020 corpus and adding post-hoc judgments, the effectiveness of all neural models substantially improves. However, BM25 remains more effective.
- Our code and the denoised, post-hoc judged dataset are available at https://github.com/castorini/touche-error-analysis.

[1]In argument retrieval, the terms 'argument' and 'document' are used interchangeably.

2 BACKGROUND AND RELATED WORK

Argument retrieval is the task of ranking documents based on the topical relevance to argumentative queries (i.e., queries about debated topics like "Should bottled water be banned?"), i.e., the documents should contain appropriate arguments pertinent to the query. An argument is often modeled as a conclusion (i.e., a claim that can be accepted or rejected) and a set of supporting or attacking premises (i.e., reasons to accept or reject the conclusion like statistical evidence, an anecdotal example, etc.) [58, 65].

Previous works on argument retrieval [47, 56] majorly made use of lexical retrieval models such as BM25 [53], DirichletLM [74], DPH [2], and TF-IDF [30]. These models were also commonly used to retrieve argumentative documents in argument search engines. For instance, popular argument search engines such as args.me [65], ArgumenText [58], and TARGER [12], all utilize BM25 for retrieving argumentative documents. Further, a large body of work to study argument retrieval approaches was carried out as part of the Touché's shared task on argument retrieval for controversial questions [6]. Most of the submitted approaches by the task participants also used lexical retrieval models (e.g., BM25 and DirichletLM) for document retrieval combined with various query processing, query reformulation, and expansion techniques. In our work, we focus on evaluating neural retrieval models as lexical retrieval models have already been well examined and utilized in argument retrieval.

The Touché 2020 dataset (queries, document collection, and relevance judgments) was later included as an argument retrieval subset in the BEIR benchmark for zero-shot evaluation of neural retrieval models in Thakur et al. [61]. Interestingly, none of the tested neural retrieval models, trained on MS MARCO [43], outperform BM25 on the Touché 2020 argument retrieval task, as shown in Table 1. But neural models outperform BM25 on a majority of the other datasets included in the BEIR benchmark (e.g., MS MARCO [43] or Natural Questions [33]). Subsequent works improving model generalization on BEIR such as $E5_{large}$ [67], CITADEL+ [39] or DRAGON+ [41] continue to underperform on Touché 2020.

The study in Thakur et al. [61] was one of the earliest works to observe the tendency of dense retrievers to retrieve short documents in Touché 2020 and provided a theoretical explanation using different similarity measures in the training loss function. In our work, we extend the idea from Thakur et al. [61] and conduct a more thorough systematic evaluation by including diverse neural model architectures and examining the Touché 2020 corpus.

Prior works have suggested several ways to understand the relationship between retrieval effectiveness and quality of test collections via empirical analyses. For instance, train–test leakage [38], retrievability bias due to query length [68], sampling bias due to near-duplicates [22], or saturated leaderboards unable to distinguish any meaningful improvements [3] were examined. However, prior work has missed out on evaluating the impact of document corpora on retrieval effectiveness, i.e., the potential impact of non-relevant documents present within a corpus on neural models. In our work, we conduct a comprehensive evaluation by independently evaluating both the Touché 2020 dataset and retrieval models to help devise targeted strategies for model improvement or data cleaning.

3 EXPERIMENTAL SETUP

In this section, we review the Touché 2020 dataset used for argument retrieval and provide details on the baseline retrieval models. Next, we provide details on model evaluation and implementation.

Touché 2020. The Touché 2020 task on controversial argument retrieval [6] uses a focused crawl of arguments for 49 test queries addressing socially important (and often controversial) issues like "Should bottled water be banned?". The document collection is the args.me corpus [1] containing 382,545 arguments. Each argument has a title in the form of a conclusion (i.e., a claim that an arguer could make) and a context containing several premises (reasons, opinions, or evidence that support or attack the claim). The task organizers also published relevance judgments (non-relevant, relevant, and highly relevant) for 2,214 documents (cf. Table 5 for dataset characteristics). The documents were pooled using the top-5 pooling strategy from 12 ranked results submitted by participants. While the argument retrieval track in Touché has more recent versions [7, 8, 10], in our work, we use the Touché 2020 dataset, due to its availability in the BEIR benchmark [61].[2]

Retrieval Models. For our experiments, we select different open-source retrieval models to better understand errors across different neural architectures. The selected models are the lexical model BM25 [53], the dense retrievers DRAGON+ [41], Contriever [29], and TAS-B [27], the sparse retriever SPLADEv2 [21], and the multi-vector retriever CITADEL+ [39].

Evaluation. To evaluate retrieval effectiveness on Touché 2020, we use nDCG@10 metric as it has been widely adopted in the BEIR benchmark [61]. In addition, we use the hole@k rate (i.e., the ratio of results retrieved by a model at cutoff k that do not have relevance judgments) to estimate the proportion of unjudged documents.

Implementation Details. In our work, we conduct a reproducibility study with previously available models' checkpoints. We did not retrain any neural model and use up to a maximum of A6000 × 4 GPUs for inference. For BM25, we follow Thakur et al. [61] and use multi-field (title and body indexed separately with equal weights) version[3] available in Anserini [72] with default parameters ($k_1 = 0.9$ and $b = 0.4$). For our dense models, Contriever (mean pooling with dot product), TAS-B, and DRAGON+ (both `[CLS]` token pooling with dot product), we reproduce the results by converting model checkpoints using `sentence-transformers`[4] and evaluate them on Touché 2020 using BEIR evaluation.[5] For SPLADEv2 (max aggregation), we reproduce the model using the SPRINT toolkit [62].[6] Finally, for CITADEL+ (with distillation and hard negative mining), we use the original `dpr-scale` repository for reproduction.[7] Apart from DRAGON+, in our work, we successfully reproduce the nDCG@10 on Touché 2020.[8]

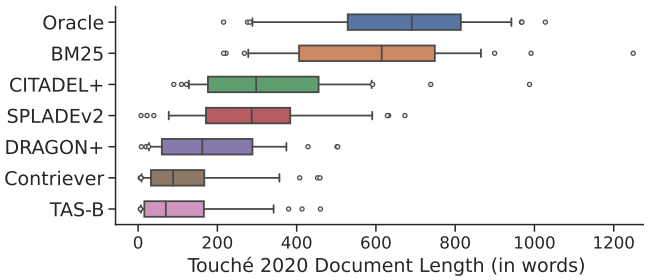


Figure 1: Boxplots showing the average length in words (x-axis) of the top-10 Touché 2020 results retrieved by the models on the y-axis (sorted by decreasing nDCG@10; oracle: avg. length of all documents judged as relevant). The results of the neural models are much shorter in comparison to BM25.

4 EVALUATION EXPERIMENTS

In this section, we describe our evaluation experiments consisting of two independent parts. First, we conduct a black-box evaluation to understand the limitations of neural models on Touché 2020 (Section 4.1) and propose two methods to improve the neural model effectiveness at inference time (Section 4.2). Next, we denoise the data by filtering out short documents (Section 4.3) and conduct post-hoc relevance judgments (Section 4.4) to measure the unbiased nDCG@10 of neural models versus BM25 on Touché 2020. Finally, we attempt to theoretically understand our findings using axioms for information retrieval (Section 4.5).

4.1 Black-Box Model Evaluation on Touché 2020

The neural retrieval model's training often involves one or several of the following steps, a particular training dataset selection [33, 43], choosing a training optimization objective [26, 31] and deciding whether to train with specialized hard negatives [27, 48]. These configurations are crucial for neural model effectiveness but lack explainability. Hence, our objective is to uncover the reasons for errors of retrieval models (BM25 vs. neural models) on Touché 2020,

[2] ukp.informatik.tu-darmstadt.de/thakur/BEIR/datasets/webis-touche2020.zip

[3] https://github.com/castorini/anserini
[4] https://github.com/UKPLab/sentence-transformers
[5] https://github.com/beir-cellar/beir
[6] https://github.com/thakur-nandan/sprint
[7] https://github.com/facebookresearch/dpr-scale/tree/citadel
[8] For DRAGON+, we suspect the difference being caused by using A100 vs. A6000 GPUs.

Table 2: Example of the top-ranked document for a randomly selected query showing that neural models may retrieve documents with a relevant conclusion / title (within the < >) but a non-relevant premise / body. Green: relevant document; red: non-relevant document.

Query (qid=5): *Should social security be privatized?*
BM25: <Social security should be privatized> Social Security has serious issues [...] First, privatization has a shaky track record. A 2004 report from the World Bank (http://wbln1018.worldbank.org) [...]
CITADEL+: <Social security should be privatized> - Social security is a complete joke. Although it was originally designed [...] the young are forced to subsidize the old, a facet of socialism [...]
SPLADEv2: <Social Security R.I.B Should be Privatized> Thank you lannan13 for an invigorating debate.
DRAGON+: <Social Security R.I.B Should be Privatized> Pass
Contriever: <Social Security R.I.B Should be Privatized> Thank you lannan13 for an invigorating debate.
TAS-B: <Privatizing social security> Social security is in crisis

Table 3: Error rates as the percentage of a model's top-k results that are non-relevant (judgment of 0 or unjudged) and shorter than 20 words. Lower error rates are better.

Model	BM25	CITADEL+	SPLADEv2	DRAGON+	Contriever	TAS-B
Top-1	0.0%	6.1%	22.4%	40.8%	55.1%	59.2%
Top-5	0.4%	3.3%	15.9%	32.7%	40.4%	59.2%
Top-10	0.8%	4.5%	14.6%	26.5%	35.9%	51.6%

by treating models as black-boxes (without modifying parameters). Specifically, we ask the following research question:

RQ1 *Does the non-uniformity in document lengths affect neural model effectiveness on the Touché 2020 dataset?*

Quantitative Results. Figure 1 shows boxplots depicting the average document lengths of the top-10 retrieved documents by the models under investigation, where the whiskers plot the 95% confidence interval. The lengths are computed as word counts after applying `nltk` word tokenizer [5]. All neural models, on average, retrieve shorter documents containing less than 350 words (visible from medians and whiskers in Figure 1) in contrast to BM25, which retrieves longer documents on average containing more than 600 words which best mimics the Oracle distribution. Dense models (TAS-B, Contriever, and DRAGON+) appear to retrieve the shortest arguments, followed by sparse (SPLADEv2) and muti-vector (CITADEL+). The decrease in nDCG@10 on Touché 2020 is found to be *perfectly correlated* with the increase in shorter top-10 retrieved documents (Spearman correlation $\rho = 1.0$). Overall, this provides positive evidence for our hypothesis that the shorter documents present in Touché 2020 (cf. Figure 4) negatively affect neural models in terms of retrieval effectiveness.

Empirical Evidence. Upon a careful analysis of the retrieved documents by the models under investigation, we observe an interesting

```
system:
You are an Argument Summarizer, an intelligent assistant
that can summarize an argument. The output summary must
be written using an argument nature.

assistant:
Okay, please provide the argument.

user:
{argument_text}
```

Figure 2: Vanilla zero-shot prompt template used in our work with GPT-3.5 [46] to summarize Touché 2020 documents.

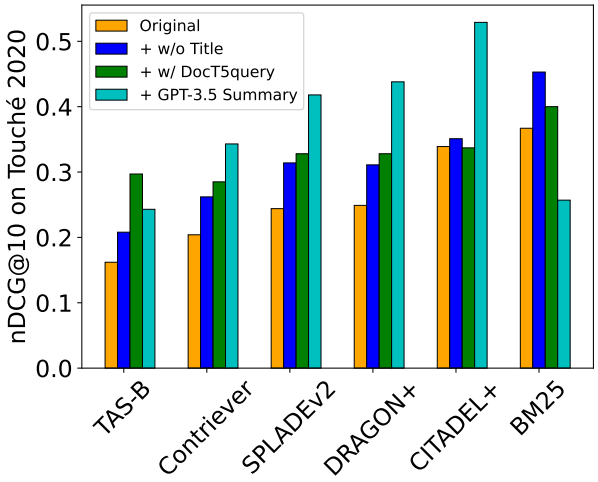


Figure 3: Change in effectiveness with DocT5query [45] query expansions and GPT-3.5 [46] summary replacement on Touché 2020. Both techniques improve the nDCG@10 for a majority of the neural models.

pattern across the retrieved documents. We find that documents retrieved in Touché 2020 by neural models show a high overlap of the query terms with the argument conclusion (document title) which is often relevant, but includes a rather "noisy", i.e., short argument premise (document body) which is non-relevant, e.g., a single word "Pass" or "social security is in crisis" (an example for a test query is shown in Table 2). To quantify the empirical evidence, we compute an error rate (in %) by counting a mistake the model makes if the document retrieved (i) is non-relevant (relevance either 0 or unjudged), and (ii) is shorter than 1–2 sentences (a maximum of 20 words). From Table 2, we observe that dense retrievers suffer the most with TAS-B with the highest 51.6% error rate in top-10 retrieved documents. CITADEL+ contains a lower percentage of shorter non-relevant documents with a low 4.5% error rate. BM25 has the lowest error rate of 0.8%, which suggests that BM25 is empirically found to be robust against non-uniformity in document lengths present within the Touché 2020 corpus.

Reasoning. We hypothesize reasons for the observed error pattern. We start by assuming query q_i and the short non-relevant

Table 4: Example queries and Touché 2020 documents: original and modified by replacing with a GPT-3.5 summary [46] or expanded by DocT5query queries [45]. Green: a relevant document; red: a non-relevant document.

Query (qid=13): *Can alternative energy effectively replace fossil fuels?*
Original: ⟨fossil fuel⟩ [...] there are many alternatives to fossil fuel [...] some of these alternatives are Nuclear fusion geothermal energy wind and solar power [...] Nuclear fusion is a very effective way for one to create a mass amount of energy [...]
GPT-3.5 Summary: The argument presented is that there are many alternatives to fossil fuel and that these alternatives, such as nuclear fusion, geothermal energy, and solar and wind power, are both efficient and cost-effective. The argument emphasizes the need for a new and better source of energy [...]
Query (qid=2): *Is vaping with e-cigarettes safe?*
Original: ⟨Cigarettes should be banned⟩ They are bad
DocT5query: They are bad why are oohs swollen and puffy why do bad people make up names are narcotics bad why are morgans bad what is the reason they are bad are the oxen bad why are fish really bad for kids why do humans keep bad odours? are spiders bad [...]

document $\hat{d}_i$ have a high word overlap due to similarity with the conclusion (title), i.e., the document is a good paraphrase of the query but does not contain information to answer the question. As shown previously in Ram et al. [49], lexical overlap remains a highly dominant signal for relevance in dense retrievers, which we suspect causes the non-relevant short document, with a similar length to the query, closer within the dense embedding space representation. For sparse and multi-vector retrievers, the token overlap of query q_i and the shorter non-relevant document is high, which results in a higher similarity score. However, in contrast, the BM25 algorithm accounts for document length normalization within its parameter b [57]. As longer documents tend to have more term occurrences, leading to potential bias, document length normalization in BM25 acts as a normalization parameter, improving robustness against sensitivity toward short document errors. Neural models, conversely, suffer from noise present in the form of short and non-relevant documents in Touché 2020.

4.2 Improving Effectiveness at Inference Time

Information retrieval datasets such as Touché 2020 (unlike MS MARCO or Natural Questions), may not be uniform in document length. Ideally, models should be explicitly trained to be robust against noisy short documents, but practitioners lack access to these setups, and retraining is often computationally expensive. Based on these observations, we ask the following research question:

RQ2 *Can we improve neural model effectiveness at inference time without expensive retraining of models?*

We experiment with two techniques to improve neural model effectiveness at inference time: (i) expanding documents with synthetic DocT5query queries, and (ii) shortening documents by replacing them with GPT-3.5-generated summaries.

DocT5query Expansion. We reuse the DocT5query [45] model[9] from BEIR [61] to expand documents in Touché 2020 with generated queries. We focus solely on generating queries using the premise (body) and not the conclusion (title) for all 382,545 documents in Touché 2020. We hypothesize that noisy, shorter documents that negatively affect retrievers will increase the document length and decrease relevance as they now contain additional non-relevant terms (generated queries would repeat these terms). For our experiments, we generate 10 synthetic queries for each document within Touché 2020, append these synthetic queries to their respective argument (cf. Table 4), and re-evaluate all tested models.

GPT-3.5 Summarization. Furthermore, we explore an additional technique by using shorter and relevant summaries to replace lengthy documents in Touché 2020. Using GPT-3.5-turbo [46], we generate concise summaries of all the 2,214 originally judged documents in Touché 2020 available as a proxy[10] using a zero-shot vanilla prompt template shown in Figure 2. We replace the original judged document with the summarized version. The synthetic summaries typically follow a uniform structure, starting with an introductory overview of the topic, followed by supporting or opposing premises, with examples and evidence originally discussed in the source document (cf. Table 4).

Experimental Results. As shown in Figure 3, removing conclusions (document titles) from arguments improves the nDCG@10 on Touché 2020 across all models, with a particularly pronounced effect on BM25. We discuss more about this later in Section 4.3. The DocT5query-based expansion improves TAS-B on Touché 2020 with minor improvements for other neural models, except CITADEL+. As hypothesized, document expansion with generated queries helps neural models to smartly avoid retrieving short and non-relevant documents by extending them with additional non-relevant terms (see Table 4 for reference). With GPT-3.5 replaced summaries, BM25 shows a decline in nDCG@10, whereas other neural models like DRAGON+ and CITADEL+ show significant improvements in nDCG@10 on Touché 2020. The absence of query terms in the GPT-3.5 summary may impact BM25's ability to effectively match query terms, unlike neural models' semantic representation, which can fit more relevant information within their (maximum) sequence length constraint of 512 tokens, thereby helping neural models to retrieve better documents as summaries.

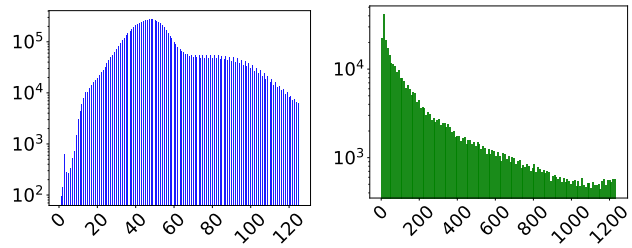
4.3 Denoising the Touché 2020 Corpus

As discussed in Section 3, the args.me corpus in Touché 2020 contains web-crawled arguments from various debating portals and thereby may contain noise as non-valid arguments. However, a valid document premise (or body) should provide evidence or reasoning that can be used to back up the conclusion (or title) as an argument [59, 60, 63]. But very short premises that are less than 1–2 sentences (e.g., "Pass" or "I agree") do not contain enough evidence to be classified as a valid argument.

To better understand the Touché 2020 corpus, we compare its document length distribution to the standard retrieval dataset

[9] https://huggingface.co/BeIR/query-gen-msmarco-t5-base-v1

[10] Generating summaries using GPT-3.5 for all 382,545 documents in Touché 2020 is expensive and not feasible within our computational budget.



(a) MS MARCO [43] (b) Touché 2020 [6]

Figure 4: Document length distribution in Touché 2020 vs. MS MARCO (x-axis: document length in words; log-scaled y-axis: frequency of document lengths). Touché 2020 has a monotonically decreasing broad distribution, while the MS MARCO distribution is much narrower.

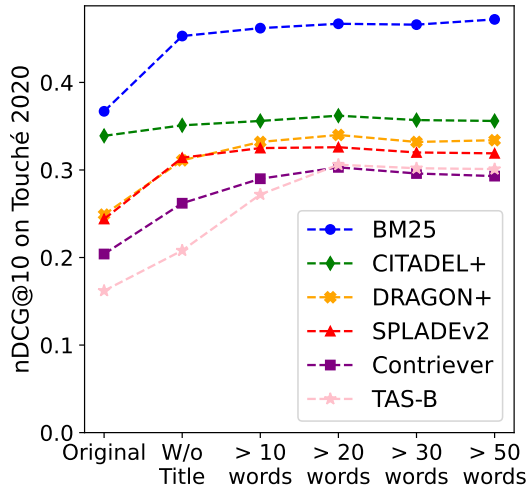


Figure 5: Denoising experiment to determine the best threshold n for filtering out short documents in Touché 2020. All models improve (until a maximum of 20 words) in effectiveness with data denoising in Touché 2020.

MS MARCO [43]. The plots in Figure 4 show that the length distribution of Touché 2020 monotonically decreases with a high frequency of extremely short arguments (spike in the graph at 20–30 words) and a long tail of long arguments (even exceeding 1200 words), while the MS MARCO length distribution is much narrower with relatively few extremely short outliers.

RQ3 *Does neural retrieval model effectiveness improve by denoising the Touché 2020 document corpus?*

The hypothesis for investigating this research question lies in whether the effectiveness of neural models can be improved by cleaning, i.e., reducing noise in the Touché 2020 document corpus. To validate this, we experiment by reducing noise in Touché 2020, i.e., filtering out non-argumentative documents from the corpus.

Table 5: The Touché 2020 dataset characteristics before and after denoising and post-hoc judgments. Reported are the total number of documents, the average document length, the number of queries, the number of relevance-judged documents, and the number of documents per relevance grade: non-relevant (0), relevant (1), and highly relevant (2).

	Original	Denoised	Post-hoc
# Documents	382,545	303,732	303,732
Avg. length	293.5	358.7	358.7
# Queries	49	49	49
# Judgments	2,214	1,785	2,849
# Relevance = 2	636	620 (16 ↓)	1,136 (516 ↑)
# Relevance = 1	296	265 (31 ↓)	576 (311 ↑)
# Relevance = 0	1,282	900 (382 ↓)	1,137 (237 ↑)

One way is to use argument classification to classify each document [16, 50] as either a valid or non-valid argument, however, it is computationally expensive to classify all arguments in Touché 2020 [25, 50]. Instead, we follow a simple heuristic and filter out potentially non-valid arguments based on the document length. Our denoising technique removes the conclusion (across all documents in Touché 2020) and only carefully selects documents with premises greater than a threshold of at least n words in length.

Results after Denoising. Figure 5 shows that our heuristic denoising improves the nDCG@10 for all models. That removing the argument conclusion (i.e., title) alone improves the nDCG@10 for all models is probably caused by the inherent nature of argument retrieval, where premises are more important for a document to be classified as a valid argument than the conclusion. Without the conclusion, often also the lexical overlap with the query that confuses neural models (cf. Section 4.1) is decreased. As for a length threshold for removing documents, $n = 20$ words empirically provides the best nDCG@10 across all tested models, as the effectiveness saturates when removing premises with more than 20 words.

A limitation of denoising Touché 2020 is that we miss out on a few human-judged query-document pairs with document lengths shorter than 20 words. However, as Table 5 shows, overall 89% (382 out of 429) of the missed judgments were originally non-relevant (score 0), and only 3.7% (16 out of 429) are highly relevant (score 2). This suggests that shorter documents in the Touché 2020 corpus are likely to be non-relevant, hence denoising based on document length is a good and simple heuristic for checking valid arguments in the argument retrieval task.

4.4 Adding Post-hoc Relevance Judgments

Retrieval datasets can contain multiple biases induced by either the annotation guidelines, annotation setup, or human annotators. For instance, to avoid selection bias [42] in later studies using some retrieval dataset, popular information retrieval challenges, for instance at TREC [13, 35], aim to encouragethe submission of diverse retrieval approaches to yield diverse judgment pools.

Table 6: Retrieval effectiveness as nDCG@10 and missing judgments as hole@10 on the original, denoised (cf. Section 4.3), and post-hoc judged (cf. Section 4.4) Touché 2020 data showing that BM25 still outperforms the neural retrievers even after denoising and after post-hoc judgments.

	Original		+ Denoised		++ Post-hoc	
Model	nDCG@10	hole@10	nDCG@10	hole@10	nDCG@10	δ inc.
BM25	**0.367**	61.6%	**0.467**	51.8%	**0.785**	△ 0.418
CITADEL+	0.339	60.2%	0.362	62.5%	0.703	△ 0.364
SPLADEv2	0.272	66.3%	0.326	63.3%	0.679	△ 0.407
DRAGON+	0.249	69.2%	0.340	63.9%	0.718	△ 0.469
Contriever	0.205	71.4%	0.303	65.9%	0.650	△ 0.445
TAS-B	0.162	77.8%	0.306	67.5%	0.682	△ 0.520

To quantify the selection bias in Touché 2020, we compute how many of the top-10 results of our tested models are unjudged in the original and denoised corpus versions. Table 6 shows that the respective hole@10 values all are greater than 50% (i.e., more than half of the top results of every model are unjudged in the Touché 2020 data). Therefore, we ask the following research question:

RQ4 *Are neural retrieval models unfairly penalized on Touché 2020 due to a selection bias?*

Annotation Details. We conduct a post-hoc relevance judgment study to fill up the hole@10 across all tested models, i.e., annotating originally unjudged arguments, as filling up holes would account for denser judgments and a better estimate of nDCG@10. We hired 5+ annotators with prior debating experience and follow annotation guidelines available in Bondarenko et al. [6]. We conduct the post-hoc judgments and fill up hole@10 for all tested models by evaluating each unjudged document with three relevance labels: 0 (non-relevant), 1 (relevant), and 2 (highly relevant). We cumulatively took around 10–15 hours to judge 1,064 judgment pairs and paid each annotator a competitive hourly rate of 14.86 USD per hour. Table 5 contains Touché 2020 statistics before and after the denoising and post-hoc judgment rounds. In our post-hoc judgment round, over 78% of the judgment pairs were judged relevant (with 48% highly relevant and 30% relevant), indicating that many "relevant" documents are retrieved by models but unjudged originally in Touché 2020. We measure the inter-annotator agreement score with Fleiss' κ [20]. Since argument retrieval is highly subjective and biased towards annotator preferences and beliefs as discussed in [6, 28], we achieve a comparable score of $\kappa = 0.31$.[11]

Results after Post-hoc Judgments. Re-evaluation scores of the retrieval models after post-hoc judgment rounds are shown in Table 6 (column '++ Post-hoc'). The maximum increase in nDCG@10 is observed in dense retrievers (TAS-B, Contriever, and DRAGON+) and the least in multi-vector retrieval with CITADEL+. This provides evidence that post-hoc relevance judgments to fill up holes are necessary for a fair evaluation of models. In our hypothesis, we suspected a bias towards lexical retrievers due to their dominance in the original candidates during original Touché 2020 judgment rounds. However, even after post-hoc relevance judgments with more and better "semantic", i.e., neural retrieval models, and denoising Touché 2020, BM25 continues to outperform all neural models by a margin of at least 6.7 points on nDCG@10, thereby making it still a robust baseline for argument retrieval.

[11]We earlier observed a lower κ due to mistakes from a single annotator, which we discussed internally and rectified.

Table 7: Agreement (in %) with the length normalization axiom LNC2 when retrieving with (w/) or without (w/o) the title on Touché 2020. BM25 agrees perfectly with LNC2.

	BM25	CITADEL+	SPLADEv2	DRAGON+	Contriever	TAS-B
w/ title	99.6	75.3	60.6	39.2	41.8	35.2
w/o title	99.5	79.1	68.2	39.5	40.8	38.9

4.5 Axiomatic Error Analysis on Touché 2020

To contrast our previous empirical evaluation of neural retrieval models on Touché 2020 with well-grounded theoretical foundations of information retrieval, we investigate if we can observe similar trends using axiomatic analysis. Therefore, we measure the agreement of the neural models under investigation with information retrieval axioms. A higher agreement indicates that a retrieval model fulfills the theoretical constraint introduced in the axiom. These axioms can highlight the problems in neural models, and fixing these problems can improve the model's effectiveness [9], even when there is no strong correlation between axioms and relevance judgments [11]. While retrieval axioms can increase the effectiveness of neural retrieval models (e.g., when used for regularization [54]), dedicated axioms for neural retrieval models are still missing [64]. Consequently, our axiomatic error analysis aims to answer the following research question:

RQ5 *Can retrieval axioms explain why BM25 is better at effectiveness on Touché 2020 than neural retrieval models?*

Setup and Background. We conduct our axiomatic analysis using the `ir_axioms` framework [9].[12] Because most axioms require theoretical preconditions that are rarely met in real-world datasets (e.g., requiring document pairs retrieved for the same query of identical length) [9], we first use synthetic document pairs derived from real documents and subsequently use real document pairs with the default length relaxation from `ir_axioms`. Given more than 20 previously proposed retrieval axioms [9], we include a subset of all axioms related to document length, term frequency, and semantic similarity in our analysis. We focus on document length axioms following our observation that document length plays an important role in Touché 2020, while we include term frequency and semantic similarity because they are the specialty of lexical and neural retrieval models. In all cases, we report the agreement in the percentage of the model under investigation with the preferences of an axiom as implemented in `ir_axioms`.

Axiomatic Analysis on Synthetic Document Pairs. Table 7 shows the agreement of the tested models with the document length normalization axiom LNC2 that (somewhat artificially) states that

[12]https://github.com/webis-de/ir_axioms

Table 8: Agreement (in %) with the term frequency, document length, and semantic similarity axioms for all tested models on the original (O) and the denoised (+D) Touché 2020 data.

Model	Term Frequency			Doc. Length		Semantic Sim.	
	TFC1	TFC3	M-TDC	LNC1	TF-LNC	STMC1	STMC2
BM25 (O)	61.6	100.0	51.8	37.8	58.5	48.3	54.9
BM25 (+D)	68.5	100.0	55.6	32.8	57.4	48.4	50.9
CITADEL+ (O)	59.2	88.9	56.6	54.3	60.7	50.7	57.9
CITADEL+ (+D)	62.6	72.7	47.6	56.1	57.1	51.0	57.8
Contriever (O)	59.7	100.0	46.5	52.7	55.9	52.5	59.1
Contriever (+D)	59.4	80.0	51.4	52.5	57.7	52.6	54.3
DRAGON+ (O)	61.1	100.0	50.6	55.3	59.0	52.1	58.2
DRAGON+ (+D)	63.2	92.3	54.7	53.1	55.4	52.2	54.5
SPLADEv2 (O)	59.8	50.0	57.1	47.8	56.2	50.8	56.6
SPLADEv2 (+D)	62.9	91.7	53.0	51.5	57.8	51.3	55.2
TAS-B (O)	60.1	33.3	55.4	50.3	55.8	52.3	60.5
TAS-B (+D)	62.2	33.3	50.6	54.0	53.1	52.4	54.2

the relevance score of an m-times self-concatenation of a document should not be lower than the original document's relevance score [18]. We synthetically create document pairs that fulfill this precondition by randomly sampling 250 query–document pairs from the top-10 ranked results by all models under investigation. For each query–document pair, we create pairs for $m = 1, 2, 3$, and 4. We observe that BM25 almost perfectly agrees with the LNC2 axiom (agreement above 99%), whereas neural models substantially violate LNC2, with TAS-B having the highest disagreement, which is an expected shortcoming of TAS-B as all documents are, independent of their length, represented by vectors of the same length.

Axiomatic Analysis on Real Document Pairs. Table 8 shows the results of our axiomatic analysis on all document pairs from the top-50 ranked results for each test query on both the original (O) and the denoised (+D) Touché 2020 corpus. We report the term frequency axioms TFC1 [17], TFC3 [18] (we leave out TFC2 [18] because this axiom can only be applied on synthetic documents), and TDC [18], the document length axioms LNC1 [18] and TF-LNC [18], and the semantic similarity axioms STMC1 [19] and STMC2 [19]. We observe that BM25 has the highest agreement with the term frequency axioms TFC1 and TFC3 which are more frequently violated by the other neural models. For the M-TDC, LNC1, and TF-LNC axioms, BM25 achieves only mediocre agreement. Similarly, BM25 does not agree well with the semantic similarity axioms STMC1 and STMC2, where neural models outperform BM25, for which this could be expected (BM25 alone suffers from vocabulary mismatch in contrast to neural models), which indicates that those axioms play a subordinate role on Touché 2020.

5 DISCUSSION AND FUTURE WORK

Our systematic evaluation reveals the limitations of existing neural retrieval models for argument retrieval. These limitations largely stem from (i) the noise (short arguments) present within Touché 2020 and (ii) the nature of the task that ties relevance with argument quality. Ensuring that neural models do not merely focus on the high-lexical overlap between the query and retrieved document remains a challenge. To tackle this problem, it is critical to teach retrieval models potentially via further training, to identify documents that are not just lexically similar but semantically relevant. We leave it as future work to investigate strategies for updating the training loss function with regularization terms that penalize short documents in Touché 2020, a concept borrowed from document length normalization [57], to improve robustness in retrieval systems against noise present within document corpus.

Our evaluation also reveals that Touché 2020 corpus is rather noisy (similar to real-world test collections) containing many low-quality arguments and a lot of unjudged documents. Noisy data can create several problems that lead to the drawing of false conclusions. As shown in this work, enhancing data quality through careful denoising and post-hoc judgments leads to substantial improvements in the effectiveness of all retrieval models. We hope the community adopts similar insights from our work and potentially evaluate future model effectiveness on our denoised and post-hoc relevance judged Touché 2020 dataset is publicly available at https://github.com/castorini/touche-error-analysis.

Limitations. We acknowledge that our work is not perfect and contains limitations. In our work, we conduct an in-depth study of argument retrieval. TREC-COVID [52], a bio-medical dataset in the BEIR benchmark observes a similar spike in short document distribution, as a large number of documents in the corpus do not contain an abstract (i.e., body) [61]. We leave it as future work, to similarly investigate denoising and black-box model evaluation on TREC-COVID. Similarly, in our work, we investigate only the retrieval model's effectiveness in the first-stage argument retrieval. We did not evaluate cross-encoders or neural models at the second, i.e., reranking stage, in argument retrieval. Lastly, in our work, we did not retrain any model due to the additional computation costs. In the future, we would like to explore training robust neural models and implementing document length normalization as a regularization objective to make neural models less sensitive against noisy short documents in Touché 2020.

6 CONCLUSION

In this paper, we addressed the question of why neural models are subpar, compared to BM25, on the BEIR subset Touché 2020, an argument retrieval task. To this end, we conducted a systematic error analysis and found that neural models often retrieve short and non-relevant arguments. To alleviate this issue, we enhanced data quality by filtering out noisy and short arguments in Touché 2020 and included post-hoc judgments to fill up holes for a fair evaluation of all tested models. Although our amendments improve the effectiveness of neural models by up to a margin of 0.52 in terms of nDCG@10 scores, they still lag behind BM25. Coupled with our theoretical analysis, we highlight that all neural models violate the document length normalization LNC2 axiom, intuitively explainable as documents are mapped to equal-size vectors. Addressing these shortcomings demands improved training strategies to adapt neural models for argument retrieval. Drawing insights from our findings, future work may focus on instructing models to favor longer and high-quality argumentative documents or to better support traditional retrieval axioms.

ACKNOWLEDGMENTS

We thank our annotators for the post-hoc relevance judgments, Jack Lin for helping out reproducing DRAGON+, and Minghan Li for helping out reproducing CITADEL+. Our research was partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada; computational resources were provided by Compute Canada; by the DFG (German Research Foundation) through the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999); by the European Union's Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014); and by the Stiftung für Innovation in der Hochschullehre under the "freiraum 2022" call (FRFMM-58/2022).

REFERENCES

[1] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me Corpus. In *Proceedings of the 42nd German Conference on AI, KI 2019*. Springer, 48–59. https://doi.org/10.1007/978-3-030-30179-8_4

[2] Giambattista Amati. 2006. Frequentist and Bayesian Approach to Information Retrieval. In *Proceedings of the 28th European Conference on IR Research, ECIR 2006 (Lecture Notes in Computer Science, Vol. 3936)*. Springer, 13–24. https://doi.org/10.1007/11735106_3

[3] Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. 2022. Shallow Pooling for Sparse Labels. *Inf. Retr. J.* 25, 4 (2022), 365–385. https://doi.org/10.1007/s10791-022-09411-0

[4] Akari Asai, Timo Schick, Patrick S. H. Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware Retrieval with Instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*. ACL, 3650–3675. https://doi.org/10.18653/v1/2023.findings-acl.225

[5] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. http://www.oreilly.de/catalog/9780596516499/index.html

[6] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval. In *Proceedings of the 11th International Conference of the CLEF Association, CLEF 2020*. Springer, 384–395. https://doi.org/10.1007/978-3-030-58219-7_26

[7] Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barriere, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of Touché 2023: Argument and Causal Retrieval. In *Proceedings of the 14th International Conference of the CLEF Association, CLEF 2023*. Springer, 507–530. https://doi.org/10.1007/978-3-031-42448-9_31

[8] Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of Touché 2022: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 311–336.

[9] Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. 2022. Axiomatic Retrieval Experimentation with ir_axioms. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 3131–3140. https://doi.org/10.1145/3477495.3531743

[10] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of Touché 2021: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 450–467.

[11] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12035)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 605–618.

[12] Artem N. Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural Argument Mining at Your Fingertips. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*. ACL, 195–200. https://doi.org/10.18653/v1/p19-3031

[13] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2024. Overview of the TREC 2023 Deep Learning Track. In *Text REtrieval Conference (TREC)*. NIST, TREC. https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2023-deep-learning-track/

[14] Zhuyun Dai and Jamie Callan. 2019. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval. arXiv: 1910.10687. http://arxiv.org/abs/1910.10687

[15] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *Proceedings of the Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net. https://openreview.net/pdf?id=gmL46YMpu2J

[16] Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. 2020. ArgumenText: Argument Classification and Clustering in a Generalized Search Scenario. *Datenbank-Spektrum* 20, 2 (2020), 115–121. https://doi.org/10.1007/s13222-020-00347-7

[17] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza (Eds.). ACM, 49–56.

[18] Hui Fang, Tao Tao, and ChengXiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2 (2011), 7:1–7:42.

[19] Hui Fang and ChengXiang Zhai. 2006. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin (Eds.). ACM, 115–122.

[20] Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 76 (1971), 378–382. https://psycnet.apa.org/fulltext/1972-05083-001.pdf

[21] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. *CoRR* abs/2109.10086 (2021). arXiv:2109.10086 https://arxiv.org/abs/2109.10086

[22] Maik Fröbe, Janek Bevendorff, Jan Heinrich Reimer, Martin Potthast, and Matthias Hagen. 2020. Sampling Bias Due to Near-Duplicates in Learning to Rank. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1997–2000. https://doi.org/10.1145/3397271.3401212

[23] Luyu Gao and Jamie Callan. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*. ACL, 2843–2853. https://doi.org/10.18653/v1/2022.acl-long.203

[24] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,NAACL-HLT 2021*. ACL, 3030–3042. https://doi.org/10.18653/v1/2021.naacl-main.241

[25] Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient Pairwise Annotation of Argument Quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5772–5781. https://doi.org/10.18653/v1/2020.acl-main.511

[26] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation. *CoRR* abs/2010.02666 (2020). arXiv:2010.02666 https://arxiv.org/abs/2010.02666

[27] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of the 44th International Conference on Research and Development in Information Retrieval, SIGIR 2021*. ACM, 113–122. https://doi.org/10.1145/3404835.3462891

[28] Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument Generation with Retrieval, Planning, and Realization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2661–2672. https://doi.org/10.18653/v1/p19-1255

[29] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=jKN1pXi7b0

[30] Karen Spärck Jones. 2004. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *J. Documentation* 60, 5 (2004), 493–502. https://doi.org/10.1108/00220410410560573

[31] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[32] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International Conference on Research and Development in Information Retrieval, SIGIR 2020*. ACM, 39–48. https://doi.org/10.1145/3397271.3401075

[33] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[34] Carlos Lassance and Stéphane Clinchant. 2022. An Efficiency Study for SPLADE Models. In *Proceedings of the 45th International Conference on Research and Development in Information Retrieval, SIGIR 2022*. ACM, 2220–2226. https://doi.org/10.1145/3477495.3531833

[35] Dawn J. Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2022. Overview of the TREC 2022 NeuCLIR Track. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/Overview_neuclir.pdf

[36] Jinhyuk Lee, Zhuyun Dai, Sai Meher Karthik Duddu, Tao Lei, Iftekhar Naim, Ming-Wei Chang, and Vincent Zhao. 2023. Rethinking the Role of Token Retrieval in Multi-Vector Retrieval. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/31d997278ee9069d6721bc194174bb4c-Abstract-Conference.html

[37] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6086–6096. https://doi.org/10.18653/v1/p19-1612

[38] Patrick S. H. Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 1000–1008. https://doi.org/10.18653/v1/2021.eacl-main.86

[39] Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. CITADEL: Conditional Token Interaction via Dynamic Lexical Routing for Efficient and Effective Multi-Vector Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. ACL, 11891–11907. https://doi.org/10.18653/v1/2023.acl-long.663

[40] Jimmy Lin and Xueguang Ma. 2021. A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques. arXiv: 2106.14807. https://arxiv.org/abs/2106.14807

[41] Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 6385–6400. https://doi.org/10.18653/v1/2023.findings-emnlp.423

[42] Aldo Lipani. 2018. On Biases in Information Retrieval Models and Evaluation. *SIGIR Forum* 52, 2 (2018), 172–173. https://doi.org/10.1145/3308774.3308804

[43] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016) (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[44] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. ACL, 9844–9855. https://doi.org/10.18653/v1/2022.emnlp-main.669

[45] Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery. *Online preprint* 6 (2019), 2.

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[47] Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument Search: Assessing Argument Relevance. In *Proceedings of the 42nd International Conference on Research and Development in Information Retrieval, SIGIR 2019*. ACM, 1117–1120. https://doi.org/10.1145/3331184.3331327

[48] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 5835–5847. https://doi.org/10.18653/v1/2021.naacl-main.466

[49] Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 2481–2498. https://doi.org/10.18653/v1/2023.acl-long.140

[50] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 567–578. https://doi.org/10.18653/v1/p19-1054

[51] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. ACL, 2825–2835. https://doi.org/10.18653/v1/2021.emnlp-main.224

[52] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. 2021. Searching for Scientific Evidence in a Pandemic: An Overview of TREC-COVID. *J. Biomed. Informatics* 121 (2021), 103865. https://doi.org/10.1016/J.JBI.2021.103865

[53] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994*. NIST, 109–126. http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

[54] Corby Rosset, Bhaskar Mitra, Chenyan Xiong, Nick Craswell, Xia Song, and Saurabh Tiwary. 2019. An Axiomatic Approach to Regularizing Neural Ranking Models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 981–984.

[55] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*. ACL, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

[56] Mahsa S. Shahshahani and Jaap Kamps. 2020. Argument Retrieval from Web. In *Proceedings of the11th International Conference of the CLEF Association, CLEF 2020*. Springer, 75–81. https://doi.org/10.1007/978-3-030-58219-7_7

[57] Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted Document Length Normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson (Eds.). ACM, 21–29. https://doi.org/10.1145/243199.243206

[58] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018*. ACL, 21–25. https://doi.org/10.18653/v1/n18-5005

[59] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3664–3674. https://doi.org/10.18653/v1/d18-1402

[60] Xichen Sun, Wenhan Chao, and Zhunchen Luo. 2021. Syntax and Coherence - The Effect on Automatic Argument Quality Assessment. In *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13029)*, Lu Wang, Yansong Feng, Yu Hong, and Ruifang He (Eds.). Springer, 3–12. https://doi.org/10.1007/978-3-030-88483-3_1

[61] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html

[62] Nandan Thakur, Kexin Wang, Iryna Gurevych, and Jimmy Lin. 2023. SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval. In *Proceedings of the 46th International Conference on Research and Development in Information Retrieval, SIGIR 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 2964–2974. https://doi.org/10.1145/3539618.3591902

[63] Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 5624–5634. https://doi.org/10.18653/v1/D19-1564

[64] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, Faegheh Hasibi, Yi Fang, and Akiko Aizawa (Eds.). ACM, 13–22.

[65] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017*. ACL, 49–59. https://doi.org/10.18653/v1/w17-5106

[66] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*. ACL, 2345–2360. https://doi.org/10.18653/v1/2022.naacl-main.168

[67] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *CoRR* abs/2212.03533 (2022). https://doi.org/10.48550/ARXIV.2212.03533 arXiv:2212.03533

[68] Colin Wilkie and Leif Azzopardi. 2015. Query Length, Retrievability Bias and Performance. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 1787–1790. https://doi.org/10.1145/2806416.2806604

[69] Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. ACL, 538–548. https://doi.org/10.18653/v1/2022.emnlp-main.35

[70] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. *CoRR* abs/2309.07597 (2023). https://doi.org/10.48550/arXiv.2309.07597 arXiv:2309.07597

[71] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[72] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 1253–1256. https://doi.org/10.1145/3077136.3080721

[73] Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning Discriminative Projections for Text Similarity Measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011*, Sharon Goldwater and Christopher D. Manning (Eds.). ACL, 247–256. https://aclanthology.org/W11-0329/

[74] Chengxiang Zhai and John D. Lafferty. 2017. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. *SIGIR Forum* 51, 2 (2017), 268–276. https://doi.org/10.1145/3130348.3130377

[75] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*. ACL, 565–575. https://doi.org/10.18653/v1/2021.naacl-main.47