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>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR ’24, July 14–18, 2024, Washington, DC, USA © 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0431-4/24/07 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 E5large [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 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 E5large [67], CITADEL+ [39], SPLADEv2 [21], etc.

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Reference</th>
<th>Type</th>
<th>nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25 (BEIR)</td>
<td>Thakur et al. [61]</td>
<td>lexical</td>
<td>0.367</td>
</tr>
<tr>
<td>E5\textsubscript{large}</td>
<td>Wang et al. [67]</td>
<td>dense</td>
<td>0.272</td>
</tr>
<tr>
<td>BGE-large</td>
<td>Xiao et al. [70]</td>
<td>dense</td>
<td>0.266</td>
</tr>
<tr>
<td>Promptagator</td>
<td>Dai et al. [15]</td>
<td>dense</td>
<td>0.266</td>
</tr>
<tr>
<td>DRAGON+</td>
<td>Lin et al. [41]</td>
<td>dense</td>
<td>0.263</td>
</tr>
<tr>
<td>GTR-XXL</td>
<td>Ni et al. [44]</td>
<td>dense</td>
<td>0.256</td>
</tr>
<tr>
<td>GPL</td>
<td>Wang et al. [66]</td>
<td>dense</td>
<td>0.255</td>
</tr>
<tr>
<td>RocketQA\textsubscript{v2}</td>
<td>Ren et al. [51]</td>
<td>dense</td>
<td>0.247</td>
</tr>
<tr>
<td>ANCE</td>
<td>Xiong et al. [71]</td>
<td>dense</td>
<td>0.240</td>
</tr>
<tr>
<td>RetroMAE</td>
<td>Xiao et al. [69]</td>
<td>dense</td>
<td>0.237</td>
</tr>
<tr>
<td>Contriever</td>
<td>Izacard et al. [29]</td>
<td>dense</td>
<td>0.204</td>
</tr>
<tr>
<td>TART-dual</td>
<td>Asai et al. [4]</td>
<td>dense</td>
<td>0.201</td>
</tr>
<tr>
<td>TAS-B</td>
<td>Hoffstätter et al. [27]</td>
<td>dense</td>
<td>0.162</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Reference</th>
<th>Type</th>
<th>nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25 (BEIR)</td>
<td>Thakur et al. [61]</td>
<td>lexical</td>
<td>0.367</td>
</tr>
<tr>
<td>CITADEL+</td>
<td>Li et al. [39]</td>
<td>mult.-vec.</td>
<td>0.342</td>
</tr>
<tr>
<td>XTR (XXL)</td>
<td>Lee et al. [36]</td>
<td>mult.-vec.</td>
<td>0.309</td>
</tr>
<tr>
<td>CITADEL</td>
<td>Li et al. [39]</td>
<td>mult.-vec.</td>
<td>0.294</td>
</tr>
<tr>
<td>COIL-full</td>
<td>Gao et al. [24]</td>
<td>mult.-vec.</td>
<td>0.281</td>
</tr>
<tr>
<td>ColBERT\textsubscript{v2}</td>
<td>Santharam et al. [55]</td>
<td>mult.-vec.</td>
<td>0.263</td>
</tr>
<tr>
<td>ColBERT</td>
<td>Khattab et al. [32]</td>
<td>mult.-vec.</td>
<td>0.202</td>
</tr>
<tr>
<td>uniCOIL</td>
<td>Lin et al. [40]</td>
<td>sparse</td>
<td>0.298</td>
</tr>
<tr>
<td>SPLADE\textsubscript{v2}</td>
<td>Formal et al. [21]</td>
<td>sparse</td>
<td>0.272</td>
</tr>
<tr>
<td>SPLADE\textsuperscript{++}</td>
<td>Lassance et al. [34]</td>
<td>sparse</td>
<td>0.244</td>
</tr>
<tr>
<td>DeepCT</td>
<td>Dai et al. [14]</td>
<td>sparse</td>
<td>0.175</td>
</tr>
<tr>
<td>SPARTA</td>
<td>Zhao et al. [75]</td>
<td>sparse</td>
<td>0.156</td>
</tr>
</tbody>
</table>

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\textsubscript{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 documents were pooled using the top-5 pooling strategy from 12 ranked results submitted by participants. The documents were indexed using two multi-field (title and body indexed separately with equal weights) configurations: one 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 $\text{CLS}$ 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.$^6$ 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.$^8$

**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 x 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 $\text{CLS}$ 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.$^6$ 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.$^8$

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,

---

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

<table>
<thead>
<tr>
<th>Model</th>
<th>BM25</th>
<th>CITADEL+</th>
<th>SPLADEv2</th>
<th>DRAGON+</th>
<th>Contriever</th>
<th>TAS-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1</td>
<td>0.0%</td>
<td>6.1%</td>
<td>22.4%</td>
<td>40.8%</td>
<td>55.1%</td>
<td>59.2%</td>
</tr>
<tr>
<td>Top-5</td>
<td>0.4%</td>
<td>3.3%</td>
<td>15.9%</td>
<td>32.7%</td>
<td>40.4%</td>
<td>59.2%</td>
</tr>
<tr>
<td>Top-10</td>
<td>0.8%</td>
<td>4.5%</td>
<td>14.6%</td>
<td>26.5%</td>
<td>35.9%</td>
<td>51.6%</td>
</tr>
</tbody>
</table>

系统的自然语言描述如下：

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 multi-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 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 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
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

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.

4.3 Denoising the Touché 2020 Corpus

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.
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.

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.

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).

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

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 encourage the 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.

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

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 51% highly relevant and 30% relevant), indicating that many "relevance judgments" are retrieved by models but unjudged originally in Touché 2020.

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]. 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.

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>w/ title</th>
<th>w/o title</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>99.6</td>
<td>99.5</td>
</tr>
<tr>
<td>CITADEL+</td>
<td>75.3</td>
<td>79.1</td>
</tr>
<tr>
<td>SPLADEv2</td>
<td>60.6</td>
<td>68.2</td>
</tr>
<tr>
<td>DRAGON+</td>
<td>39.2</td>
<td>39.5</td>
</tr>
<tr>
<td>Contriever</td>
<td>41.8</td>
<td>40.8</td>
</tr>
<tr>
<td>TAS-B</td>
<td>35.2</td>
<td>38.9</td>
</tr>
</tbody>
</table>


\[\text{\footnotesize \texttt{https://github.com/webis-de/ir_axioms}}\]
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Term Frequency</th>
<th>Doc. Length</th>
<th>Semantic Sim.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TFC1</td>
<td>TFC3</td>
<td>M-TDC</td>
</tr>
<tr>
<td>BM25 (O)</td>
<td>61.6</td>
<td>100.0</td>
<td>51.8</td>
</tr>
<tr>
<td>BM25 (+D)</td>
<td>68.5</td>
<td>100.0</td>
<td>55.6</td>
</tr>
<tr>
<td>CITEADEL+ (O)</td>
<td>59.2</td>
<td>88.9</td>
<td>56.6</td>
</tr>
<tr>
<td>CITEADEL+ (+D)</td>
<td>62.6</td>
<td>72.7</td>
<td>47.6</td>
</tr>
<tr>
<td>Contriever (O)</td>
<td>59.7</td>
<td>100.0</td>
<td>46.5</td>
</tr>
<tr>
<td>Contriever (+D)</td>
<td>59.4</td>
<td>80.0</td>
<td>51.4</td>
</tr>
<tr>
<td>DRAGON+ (O)</td>
<td>61.1</td>
<td>100.0</td>
<td>50.6</td>
</tr>
<tr>
<td>DRAGON+ (+D)</td>
<td>63.2</td>
<td>92.3</td>
<td>54.7</td>
</tr>
<tr>
<td>SPLADEv2 (O)</td>
<td>59.8</td>
<td>50.0</td>
<td>57.1</td>
</tr>
<tr>
<td>SPLADEv2 (+D)</td>
<td>62.9</td>
<td>91.7</td>
<td>53.0</td>
</tr>
<tr>
<td>TAS-B (O)</td>
<td>60.1</td>
<td>33.3</td>
<td>55.4</td>
</tr>
<tr>
<td>TAS-B (+D)</td>
<td>62.2</td>
<td>33.3</td>
<td>50.6</td>
</tr>
</tbody>
</table>

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 [32], 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 strategies.
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 10170014 (OpenWebSearchEU), https://doi.org/10.1030/10170014; and by the Stiftung für Innovation in der Hochschullehre under the “freiraum 2022” call (FRFMM-58-2022).

REFERENCES


