DeepCT-enhanced Lexical Argument Retrieval

Alexander Bondarenko
Leipzig University and
Friedrich-Schiller-Universität Jena

Maik Fröbe
Friedrich-Schiller-Universität Jena

Danik Hollatz
Martin-Luther-Universität Halle-Wittenberg

Jan Heinrich Merker
Friedrich-Schiller-Universität Jena

Matthias Hagen
Friedrich-Schiller-Universität Jena

Abstract

The recent Touché lab's argument retrieval task focuses on controversial topics like 'Should bottled water be banned?' and asks to retrieve relevant pro/con arguments. Interestingly, the most effective systems submitted to that task still are based on lexical retrieval models like BM25. In other domains, neural retrievers that capture semantics are more effective than lexical baselines. To add more "semantics" to argument retrieval, we propose to combine lexical models with DeepCT-based document term weights. Our evaluation shows that our approach is more effective than all the systems submitted to the Touché lab while being on par with modern neural re-rankers that themselves are computationally more expensive.

1 Introduction

Lexical retrieval models like BM25 (Robertson et al., 1994) or DirichletLM (Zhai and Lafferty, 2001) are the basis of many of the early argument retrieval approaches (Chernodub et al., 2019; Potthast et al., 2019; Stab et al., 2018; Wachsmuth et al., 2017) and also were the most common choice of many participants of the Touché lab's shared task on argument retrieval for controversial questions (Bondarenko et al., 2020, 2021). A few neural rankers like K-NRM (Xiong et al., 2017) and CEDR (MacAvaney et al., 2019) were used by the task participants but showed to be less effective than the task's official DirichletLM-based baseline. Interestingly, also newer neural retrieval models like ColBERTv2 (Santhanam et al., 2022) and LaPraDoR (Xu et al., 2022) are less effective than BM25 on the Touché subset of the BEIR benchmark for zero-shot retrieval (Thakur et al., 2021).

In this paper, we propose to improve the effectiveness of lexical argument retrieval models by adding a semantic document expansion step that uses term weights calculated by DeepCT (Dai and Callan, 2020b). For term weighting, DeepCT utilizes contextualized word representations generated by BERT (Devlin et al., 2019) and is then fine-tuned to predict whether a document term is likely to appear in "relevant" queries. At the inference step, a fine-tuned model is applied to every document in the retrieval collection independently from the test queries. Hence, an advantage of DeepCT is that the inference can be done offline before indexing the corpus. Another advantage is that DeepCT does not necessarily require relevance judgments either for training or for inference making DeepCT beneficial for retrieval tasks in specialized domains that have no or little training data.

For our experiments, we use the lexical retrieval models BM25 and DirichletLM and their combination with the query expansion model RM3 (Abdul-Jaleel et al., 2004). We test these models on the Touché 2020 and 2021 test collections consisting of 49 and 50 test queries on controversial topics respectively, the args.me corpus (Ajjour et al., 2019) as a document collection (about 400,000 documents, i.e., English arguments crawled from online debate portals), and 6,000 graded relevance judgments (not relevant, relevant, and highly relevant) from Touché. Additionally, we expand the documents in the retrieval corpus based on the term weights predicted by fine-tuned DeepCT models. To fine-tune the DeepCT models (originally trained on the MS MARCO dataset (Nguyen et al., 2016)) specifically for the argument retrieval task, we make use of the args.me structured documents, consisting of the argument premises, the conclusion, and the main debate topic. We use either the conclusions or the debate topics combined with the conclusions as ground truth terms in the reference field of DeepCT. Afterwards, we apply the fine-tuned DeepCT model to the whole args.me corpus and expand the document's premises by repeating terms based on the learned DeepCT term weights.

We compare our approaches with the following

Proceedings of the 11th Workshop on Argument Mining, pages 29–35
August 15, 2024 ©2024 Association for Computational Linguistics

baselines: (1) the most effective Touché systems (that use BM25 and DirchletLM combined with query and document processing) and (2) BM25 combined with neural re-rankers: (a) a contextualized late interaction model ColBERT (Khattab and Zaharia, 2020), (b) pointwise cross-encoders monoBERT (Nogueira and Cho, 2019) and monoT5 (Nogueira et al., 2020), and (c) a zero-shot listwise re-ranker LiT5 (Tamber et al., 2023).

To evaluate the retrieval effectiveness, we use nDCG@5 (Järvelin and Kekäläinen, 2002), the official evaluation measure of the Touché task. To account for missing relevance judgments (up to 60%; see Table 2), we measure nDCG@5 after removing unjudged documents from ranked results as proposed by Sakai (2007) and use the bpref measure (Buckley and Voorhees, 2004) which is robust to missing relevance judgments.

The evaluation results show that our DeepCT-enhanced lexical argument retrieval approach is more effective than all the systems submitted to the Touché task while being on par with modern neural re-rankers that are more computationally expensive. Our findings thus may indicate the potential of combining lexical models with semantic document expansion for specialized retrieval tasks like argument retrieval, where little or no training data (in terms of relevance judgments) is available.[1]

2 Related Work

Retrieving relevant arguments from the Web is essential to support discussions on controversial topics like 'Should bottled water be banned?' (Ajjour et al., 2019). Until now, lexical retrieval models like BM25 (Robertson et al., 1994) and DirichletLM (Zhai and Lafferty, 2001) have been the most effective retrievers for this task (Potthast et al., 2019). For instance, argument search engines args.me (Wachsmuth et al., 2017), ArgumenText (Stab et al., 2018), and TARGER (Chernodub et al., 2019), all use BM25 for initial retrieval. However, even though neural retrievers like ColBERTv2 (Santhanam et al., 2022), LaPraDoR (Xu et al., 2022), or COCO-DR (Yu et al., 2022) have led to effectiveness improvements in many domain-specific retrieval tasks as evaluated in the BEIR benchmark (Thakur et al., 2021), for argument retrieval (e.g., the Touché subset of BEIR), lexical retrievers still outperform neural models.

Studying argument retrieval approaches was also carried out as part of the Touché lab's shared tasks on argument retrieval for controversial questions (Bondarenko et al., 2020, 2021). Most of the participant's approaches used lexical retrieval models (i.e., BM25 and DirichletLM) for initial document retrieval combined with various query processing and reformulation techniques. The initial document candidates were further re-ranked based on the estimated document argumentativeness (i.e., the presence of conclusions and premises) and argument quality. Several tested neural rankers, like K-NRM (Xiong et al., 2017) and CEDR (MacAvaney et al., 2019) were less effective (measured with nDCG@5) than the lexical models.

Lexical retrieval models (that rely on an exact match between the query and document terms), conversely, may suffer from "ignoring" the semantic similarity between the query and document terms. Hence, we propose to combine lexical retrievers (that are effective for argument retrieval) with document expansion based on estimated semantic term importance (term weights) predicted by DeepCT (Dai and Callan, 2020b,a). The DeepCT model exploits the BERT (Devlin et al., 2019) fine-tuning paradigm by fine-tuning a pre-trained BERT model to predict the importance of words in documents w.r.t. reference terms (e.g., query terms). Fine-tuning aims to minimize the mean square error between the predicted term weights and the ground truth term weights. The ground truth labels can be generated using documents only, relevance labels, or pseudo-relevance feedback. In our work, we use the documents-only approach which does not require manual relevance labels for fine-tuning DeepCT (cf. Section 3 for more details). The fine-tuned DeepCT is then applied to the documents and predicts the document term importance scores. Finally, the documents are modified by repeating terms proportionally to predicted weights ($w * 100$, where term weights $w \in [0, 1]$), thus boosting the term frequency of the repeated terms in the inverted index. Thus, lexical retrieval models that rely on the term frequency as a relevance signal can benefit from repeated "relevant" document terms. It has been shown that combining DeepCT with lexical models improves the effectiveness of ad hoc retrieval (Dai and Callan, 2020b,a) on general-domain document collections, e.g., MS MARCO (Nguyen et al., 2016) or Wikipedia articles (Dietz et al., 2017). Thus, we aim to test a combination of lexical re-

[1] Code and data are at https://github.com/webis-de/argmining24-deepct-lexical-argument-retrieval/

trieval models with semantic document expansion for the argument retrieval task.

For the evaluation of argument retrieval approaches, several datasets (Abbott et al., 2016; Hidey et al., 2017; Ajjour et al., 2019) and the Touché test collections (Bondarenko et al., 2020, 2021) emerged. By far the largest and hence one of the most frequently used document collections, the args.me corpus, contains about 400,000 arguments crawled from online debates on controversial topics (Ajjour et al., 2019). The Touché shared tasks on argument retrieval for controversial questions also used the args.me corpus. Additionally, the task organizers created and published manual relevance judgments and runs (ranked results) submitted by the task participants. Our experiments use the Touché data from the years 2020 and 2021.

3 Data and Approach

Data. In this work, we use the datasets from the Touché 2020 and 2021 tasks on controversial argument retrieval (Bondarenko et al., 2020, 2021). The task was to retrieve and rank relevant argumentative documents for queries addressing socially important (and often controversial) topics like ‘Should bottled water be banned?’. The document collection was the args.me corpus (Ajjour et al., 2019) containing about 400,000 arguments crawled from different online debate portals. Each document is structured and contains a debate topic field (e.g., ‘Pollution’), an argument conclusion (e.g., ‘Plastic bottles should be banned’), and a main content containing several premises (i.e., reasons, opinions, or evidence that support or attack the conclusion). We access all the data inside the PyTerrier framework (Macdonald and Tonellotto, 2020) via `ir_datasets` (MacAvaney et al., 2021), including queries, document collection, and available document-level manual relevance judgments (the participants' systems are available at the Touché task website `https://touche.webis.de`).

DeepCT-based term weighting. Our pilot experiments using the original DeepCT model pretrained on the MS MARCO dataset showed that the retrieval effectiveness of lexical models degrades (cf. DirichletLM + DeepCT achieves nDCG@5 of 0.59 vs. 0.83 that DirichletLM achieves on the unmodified args.me corpus). This is likely due to the document domain change (general domain vs. argument retrieval). We thus opt for fine-tuning DeepCT for the argument retrieval task on the args.me corpus. To create training samples for fine-tuning, we use a content-based weak-supervision strategy proposed by Dai and Callan (2020a) that determines the target important terms by utilizing the document's structure (i.e., different fields like debate topic, conclusion, and premises in our case). Since the Touché queries are used for testing, as a reference field for fine-tuning, we use either an argument's conclusion field of the args.me document or a concatenation of a debate topic and conclusion.

Table 1: Example of a training sample to fine-tune DeepCT created using the conclusion ‘Banning bottled water would reduce waste and protect the environment’ as a reference field. The important terms identified by DeepCT at the inference step are in bold; superscripts indicate the number of times each term is repeated in the document (term weights predicted by DeepCT (from 0 to 1) multiplied by 100).

Passage:	Plastic **water**$^{(48)}$ **bottles**$^{(23)}$ were the third most commonly collected **waste**$^{(34)}$ during the Ocean Conservancy's International Coastal Cleanup behind cigarette butts and plastic food wrappers. By 2050, estimates suggest there will be more plastic **waste**$^{(14)}$ by weight in the oceans than fish. [...] A nation-wide **ban**$^{(27)}$ on **bottled**$^{(21)}$ **water**$^{(17)}$ would lead to an estimated 68 billion fewer plastic **water**$^{(14)}$ **bottles**$^{(19)}$ being manufactured, purchased, used, and discarded.
Reference:	water: 1.0, bottles: 1.0, waste: 1.0, ban: 1.0, bottled: 1.0.

Following the original DeepCT fine-tuning strategy (Dai and Callan, 2020b), we split the premises of the args.me documents into passages of 500 tokens to comply with the DeepCT input limit of 512 tokens. To identify the reference field's ground truth terms, we remove stop words using NLTK (Bird, 2006) from passages, conclusions, and debate topics. Afterwards, we apply stemming using the NLTK's Porter stemmer (Porter, 1980). The reference field's ground truth terms are selected as follows: If there is a stem of a word from a passage and this stem also appears in the stemmed conclusion (or debate topic + conclusion), the original form of the word is added to the reference field. The target term weights are assigned 1.0 (see Table 1 for an example).

To fine-tune DeepCT, we use three variants of the args.me corpus: (1) all documents in the corpus; and to analyze the effect of possible train-test leakage: (2) judged documents from the Touché 2020 and 2021 tasks are removed, and (3) top-50 documents from all systems submitted to Touché 2020

Table 2: Retrieval effectiveness of the best data transformation technique to fine-tune DeepCT (as per nDCG@5) per retrieval model (BM25, DirichletLM (DLM), and their combinations with RM3; model parameters tuned, see Column 'PT') on the Touché 2020 and 2021 datasets: (1) all args.me documents, (2) judged documents are removed, and (3) top-50 documents are removed. Document fields used as the reference field for DeepCT: debate topic and conclusion (TC), or conclusion only (C). Both retrieval models without DeepCT doc. term weighting, best Touché systems, and neural baselines from TIREx are reported for comparison. The nDCG@5 scores are evaluated after removing unjudged documents (cf. the ratio of retrieved documents with relevance judgments, 'judged@5'). The bpref score is robust to unjudged documents. Underlines denote the best system per metric; bold indicates significant equivalence to the best system within ±0.1 (two one-sided t-tests, $p < 0.05$, Bonferroni correction).

	Retrieval model	PT	Data transf.	nDCG@5	bpref	judged@5
Touché 2020	DeepCT + DLM + RM3	✓	(1), C	**0.88**	0.71	0.45
	BM25 + monoT5	×	n/a	**0.87**	**0.81**	0.41
	DeepCT + BM25 + RM3	✓	(2), TC	**0.87**	**0.77**	0.46
	BM25 + RM3	✓	n/a	**0.87**	0.71	0.43
	BM25 + LiT5	×	n/a	0.86	0.51	0.39
	BM25 + monoBERT	×	n/a	0.85	**0.79**	0.41
	DeepCT + BM25	✓	(2), TC	0.84	0.71	0.47
	BM25 + ColBERT	×	n/a	0.83	**0.77**	0.42
	Best Touché	×	n/a	0.83	0.70	1.00
	DeepCT + DLM	✓	(2), TC	0.82	0.68	0.47
	DLM + RM3	✓	n/a	0.82	0.58	0.51
	BM25	✓	n/a	0.80	0.64	0.44
	DLM	✓	n/a	0.78	0.57	0.56
Touché 2021	BM25 + monoT5	×	n/a	**0.77**	**0.80**	0.70
	DeepCT + BM25	✓	(3), TC	0.74	0.74	0.78
	DeepCT + BM25 + RM3	✓	(2), TC	0.74	0.74	0.70
	Best Touché	×	n/a	0.74	0.73	1.00
	DeepCT + DLM	✓	(1), TC	0.74	0.72	0.79
	BM25 + monoBERT	×	n/a	0.73	**0.77**	0.69
	BM25 + LiT5	×	n/a	0.73	0.59	0.79
	DeepCT + DLM + RM3	✓	(1), TC	0.70	0.73	0.72
	BM25 + RM3	✓	n/a	0.70	0.65	0.82
	BM25 + ColBERT	×	n/a	0.69	**0.75**	0.63
	BM25	✓	n/a	0.67	0.62	0.95
	DLM	✓	n/a	0.67	0.62	0.94
	DLM + RM3	✓	n/a	0.64	0.56	0.75

and 2021 are removed from args.me. After expanding the passages using the fine-tuned DeepCT models, the passages are concatenated back into complete documents. The original args.me corpus is then modified with the three differently fine-tuned DeepCT models, resulting in three corpus variants.

Retrieval models. For every variant of the modified corpus, we test the effectiveness of BM25 and DirichletLM and their combination with the query expansion model RM3. We select the model's parameters using grid search and two-fold cross-validation (each fold is either the Touché 2020 or 2021 relevance judgments) implemented in PyTerrier (Macdonald and Tonellotto, 2020).

4 Evaluation

We compare our approaches (lexical retrieval models with DeepCT-based corpus transformations) with the most effective systems at Touché 2020 (49 queries, and 2,298 relevance judgments) and 2021 (50 queries, and 3,711 judgments) as well as with four strong neural retrieval baselines implemented in TIREx (Fröbe et al., 2023).

Due to the high portion of missing judgments for systems not in the Touché's original pool (cf. column 'judged@5' in Table 2), we measure nDCG@5 (Järvelin and Kekäläinen, 2002), the official evaluation measure of the Touché task, after removing unjudged documents as proposed by Sakai (2007). In our evaluation, we also include the bpref measure (Buckley and Voorhees, 2004) that is invariant to missing judgments. While removing unjudged documents and using bpref have been accepted in IR evaluation, filling in missing judgments by manual annotation can provide more robust evaluation results in future work. We use the effectiveness measures implemented in ir_measures (MacAvaney et al., 2022).

At Touché 2020, the most effective system (highest nDCG@5 and highest bpref; cf. Table 2) was the official Touché task baseline that used Lucene's (Bialecki et al., 2012) DirichletLM implementation without any query or document processing (all the participants' systems were less effective). In 2021, the most effective participants' systems were the following: (1) Lucene's BM25, stop word removal, and boolean OR query (highest nDCG@5), and (2) Lucene's DirichletLM, stop word removal, and stemming using the Krovetz stemmer (Krovetz, 1993) (highest bpref).

We complement the best systems at Touché (which are all based on lexical retrieval) with four neural re-rankers: (1) ColBERT (Khattab and Zaharia, 2020), a contextualized late interaction model that uses BERT (Devlin et al., 2019), (2–3) monoBERT (Nogueira and Cho, 2019) and monoT5 (Nogueira et al., 2020), two pointwise cross-encoder models based on BERT and T5 (Raffel et al., 2020), and (4) LiT5 (Tamber et al., 2023), a zero-shot listwise re-ranker using T5. All four models were used in a re-ranking setting using TIREx (Fröbe et al., 2023), to re-rank the top-1000 documents retrieved by BM25.

Results. With respect to both nDCG@5 and bpref, our approach of using DeepCT for semantic document term weighting improves over the Touché best systems when using the BM25 retrieval model. When using DirichletLM, the DeepCT term weighting does not outperform the participants' systems on the Touché 2021 data. We also find that the best neural baseline, monoT5 as a re-ranker, is also more effective than the best Touché systems of 2021, while the other neural re-rankers fall back behind. Our most effective DeepCT-based approach does not outperform monoT5; yet, it is on par with monoT5 for Touché 2020 data (significantly equivalent to the best system within a ±0.1 band, see Table 2) and not far off on the 2021 data. The promising effectiveness indicates the potential of combining lexical models with semantic document term weighting for argument retrieval tasks. In contrast to neural models, however, DeepCT is applied at index time and does not require model inference at query time. As we also showed, fine-tuning DeepCT does not require manual relevance judgments. Thus, our approach can have beneficial properties for deployment in low-resource environments which is common for specialized tasks like argument retrieval.

Furthermore, in at least half of the retrieval scenarios, fine-tuning DeepCT on the args.me documents after removing the judged ones, results in the highest evaluation scores. Thus, we do not observe strong evidence of the train-test leakage influence on the retrieval results. Moreover, combining an argument conclusion with a debate topic for fine-tuning DeepCT often benefits the retrieval effectiveness of lexical models.

5 Conclusion

In this paper, we proposed to combine lexical retrieval models with semantic document expansion for argument retrieval. Specifically, to calculate the term weights, we fine-tuned DeepCT on the args.me corpus. The main advantages of DeepCT are that the calculation of term weights can be done in an offline fashion before document indexing and that its training does not require manual relevance judgments. This is especially important in the specialized domains (e.g., argument retrieval), where no or little training data is available. Furthermore, at query time only lexical retrieval models are used on the expanded documents that require less computational resources than neural models.

Our evaluation results showed that adding some "semantics" to strong lexical argument retrieval approaches improves the overall effectiveness over the lexical retrieval alone. Additionally, we showed that our approach is on par with modern neural rerankers, which themselves can be more computationally expensive. However, we also indicated that for a more robust conclusion, further experiments should be conducted, where the missing relevance judgments are filled.

Another potentially interesting future direction can be to include the argument mining step in the document expansion process, for instance, using only argumentative parts (conclusions and premises) of documents for fine-tuning DeepCT.

Acknowledgments

This work was partially supported 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) and by the Stiftung für Innovation in der Hochschullehre under the "freiraum 2022" call (FRFMM-58/2022).

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of LREC 2016*, pages 4445–4452, Paris. ELRA.

Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC 2004*, volume 500-261 of *NIST Special Publication*, Gaithersburg. NIST.

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 KI 2019*, volume 11793 of *LNCS*, pages 48–59, Berlin. Springer.

Andrzej Bialecki, Robert Muir, and Grant Ingersoll. 2012. Apache Lucene 4. In *Proceedings of OSIR@SIGIR 2012*, pages 17–24, Otago. University of Otago.

Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of ACL 2006*, pages 69–72, Kerrville. ACL.

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 CLEF 2020*, volume 12260 of *LNCS*, pages 384–395, Berlin. Springer.

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 *Proceedings of CLEF 2021*, volume 12880 of *LNCS*, pages 450–467, Berlin. Springer.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR 2004*, pages 25–32, New York. ACM.

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 ACL 2019*, pages 195–200, Kerrville. ACL.

Zhuyun Dai and Jamie Callan. 2020a. Context-aware document term weighting for ad-hoc search. In *Proceedings of WWW 2020*, pages 1897–1907, Geneva. IW3C2.

Zhuyun Dai and Jamie Callan. 2020b. Context-aware term weighting for first stage passage retrieval. In *Proceedings of SIGIR 2020*, pages 1533–1536, New York. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Kerrville. ACL.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC complex answer retrieval overview. In *Proceedings of TREC 2017*, volume 500-324 of *NIST Special Publication*, Gaithersburg. NIST.

Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The information retrieval experiment platform. In *Proceedings of SIGIR 2023*, pages 2826–2836, New York. ACM.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of ArgMining@EMNLP 2017*, pages 11–21, Kerrville. ACL.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR 2020*, pages 39–48, New York. ACM.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of SIGIR 1993*, pages 191–202, New York. ACM.

Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2022. Streamlining evaluation with ir-measures. In *Proceedings of ECIR 2022*, volume 13186, pages 305–310, Berlin. Springer.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of SIGIR 2019*, pages 1101–1104, New York. ACM.

Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified data wrangling with ir_datasets. In *Proceedings of SIGIR 2021*, pages 2429–2436, New York. ACM.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using PyTerrier. In *Proceedings of ICTIR 2020*, pages 161–168, New York. ACM.

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 CoCo@NIPS 2016*, volume 1773 of *CEUR Workshop Proceedings*, online. CEUR-WS.org.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. arXiv 1901.04085.

Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of EMNLP 2020*, pages 708–718, Kerrville. ACL.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

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 SIGIR 2019*, pages 1117–1120, New York. ACM.

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

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of TREC 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126, Gaithersburg. NIST.

Tetsuya Sakai. 2007. Alternatives to bpref. In *Proceedings of SIGIR 2007*, pages 71–78, New York. ACM.

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 NAACL 2022*, pages 3715–3734, Kerrville. ACL.

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 NAACL-HLT 2018*, pages 21–25, Kerrville. ACL.

Manveer Singh Tamber, Ronak Pradeep, and Jimmy Lin. 2023. Scaling down, LiTting up: Efficient zero-shot listwise reranking with seq2seq encoder-decoder models. arXiv 2312.16098.

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 NeurIPS Datasets and Benchmarks 2021*, online. Proceedings.org.

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 ArgMining@EMNLP 2017*, pages 49–59, Kerrville. ACL.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of SIGIR 2017*, pages 55–64, New York. ACM.

Canwen Xu, Daya Guo, Nan Duan, and Julian J. McAuley. 2022. LaPraDoR: Unsupervised pretrained dense retriever for zero-shot text retrieval. In *Findings of ACL 2022*, pages 3557–3569, Kerrville. ACL.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. arXiv 2210.15212.

ChengXiang Zhai and John D. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR 2001*, pages 334–342, New York. ACM.