Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and **Jimmy Lin**
David R. Cheriton School of Computer Science
University of Waterloo

Abstract

This paper applies BERT to *ad hoc* document retrieval on news articles, which requires addressing two challenges: relevance judgments in existing test collections are typically provided only at the document level, and documents often exceed the length that BERT was designed to handle. Our solution is to aggregate sentence-level evidence to rank documents. Furthermore, we are able to leverage passage-level relevance judgments *fortuitously* available in other domains to fine-tune BERT models that are able to capture cross-domain notions of relevance, and can be directly used for ranking news articles. Our simple neural ranking models achieve state-of-the-art effectiveness on three standard test collections.

1 Introduction

The dominant approach to *ad hoc* document retrieval using neural networks today is to deploy the neural model as a reranker over an initial list of candidate documents retrieved using a standard bag-of-words term-matching technique. Despite the plethora of neural models that have been proposed for document ranking (Mitra and Craswell, 2019), there has recently been some skepticism about whether they have truly advanced the state of the art (Lin, 2018), at least in the absence of large amounts of behavioral log data only available to search engine companies.

In a meta-analysis of over 100 papers that report results on the dataset from the Robust Track at TREC 2004 (Robust04), Yang et al. (2019a) found that most neural approaches do not compare against competitive baselines. To provide two recent examples, McDonald et al. (2018) report a best AP score of 0.272 and Li et al. (2018) 0.290, compared to a simple bag-of-words query expansion baseline that achieves 0.299 (Lin, 2018). Further experiments by Yang et al. (2019a) achieve 0.315 under more rigorous experimental conditions with a neural ranking model, but this is still pretty far from the best-known score of 0.3686 on this dataset (Cormack et al., 2009).

Although Sculley et al. (2018) remind us that the goal of science is not *wins*, but knowledge, the latter requires first establishing strong baselines that accurately quantify proposed contributions. Comparisons to weak baselines that inflate the merits of an approach are not new problems in information retrieval (Armstrong et al., 2009), and researchers have in fact observed similar issues in the recommender systems literature as well (Rendle et al., 2019; Dacrema et al., 2019).

Having placed evaluation on more solid footing with respect to well-tuned baselines by building on previous work, this paper examines how we might make neural approaches "work" for document retrieval. One promising recent innovation is models that exploit massive pre-training (Peters et al., 2018; Radford et al., 2018), leading to BERT (Devlin et al., 2019) as the most popular example today. Researchers have applied BERT to a broad range of NLP tasks with impressive gains: most relevant to our document ranking task, these include BERTserini (Yang et al., 2019b) for question answering and Nogueira and Cho (2019) for passage reranking.

Extending our own previous work (Yang et al., 2019c), the main contribution of this paper is a successful application of BERT to yield large improvements in *ad hoc* document retrieval. We introduce two simple yet effective innovations: First, we focus on integrating *sentence-level evidence* for document ranking to address the fact that BERT was not designed for processing long spans of text. Second, we show, quite surprisingly, that it is possible to transfer models of relevance across different domains, which nicely solves the problem of the lack of passage-level relevance anno-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3490–3496,
Hong Kong, China, November 3–7, 2019.

tations. Combining these two innovations allows us to achieve 0.3697 AP on Robust04, which is the highest reported score that we are aware of (neural or otherwise). We establish state-of-the-art effectiveness on two more recent test collections, Core17 and Core18, as well.

2 Background and Approach

To be clear, our focus is on neural ranking models for *ad hoc* document retrieval, over corpora comprising news articles. Formally, in response to a user query Q, the system's task is to produce a ranking of documents from a corpus that maximizes some ranking metric—in our case, average precision (AP). We emphasize that this problem is quite different from web search, where there is no doubt that large amounts of behavioral log data, along with other signals such as the web-graph, have led to large improvements in search quality (Mitra and Craswell, 2019). Instead, we are interested in limited data conditions—what can be achieved with modest resources outside web search engine companies such as Google and Microsoft who have large annotation teams—and hence we only consider "classic" TREC newswire test collections.

Beyond our own previous work (Yang et al., 2019c), which to our knowledge is the first reported application of BERT to document retrieval, there have been several other proposed approaches, including MacAvaney et al. (2019) and unrefereed manuscripts (Qiao et al., 2019; Padigela et al., 2019). There are two challenges to applying BERT to document retrieval: First, BERT has mostly been applied to sentence-level tasks, and was not designed to handle long spans of input, having a maximum of 512 tokens. For reference, the retrieved results from a typical bag-of-words query on Robust04 has a median length of 679 tokens, and 66% of documents are longer than 512 tokens. Second, relevance judgments in nearly all newswire test collections are annotations on documents, not on individual sentences or passages. That is, given a query, we only know what documents are relevant, not spans within those documents. Typically, a document is considered relevant as long as some part of it is relevant, and in fact most of the document may not address the user's needs. Given these two challenges, it is not immediately obvious how to apply BERT to document ranking.

2.1 Key Insights

We propose two innovations that solve the above-mentioned challenges. First, we observe the existence of test collections that *fortuitously* contain passage-level relevance evidence: the MAchine Reading COmprehension (MS MARCO) dataset (Bajaj et al., 2018), the TREC Complex Answer Retrieval (CAR) dataset (Dietz et al., 2017), and the TREC Microblog (MB) datasets (Lin et al., 2014), described below.

MS MARCO features user queries sampled from Bing's search logs and passages extracted from web documents. Each query is associated with sparse relevance judgments by human editors. TREC CAR uses queries and paragraphs extracted from English Wikipedia: each query is formed by concatenating an article title and a section heading, and passages in that section are considered relevant. This makes CAR, essentially, a synthetic dataset. TREC Microblog datasets draw from the Microblog Tracks at TREC from 2011 to 2014, with topics (i.e., queries) and relevance judgments over tweets. We use the dataset prepared by Rao et al. (2019).

Note, however, that all three datasets are out of domain with respect to our task: MS MARCO passages are extracted from web pages, CAR paragraphs are taken from Wikipedia, and MB uses tweets. These corpora clearly differ from news articles to varying degrees. Furthermore, while MS MARCO and MB capture search tasks (although queries over tweets are qualitatively different), CAR "queries" (Wikipedia headings) do not reflect search queries from real users. Nevertheless, our experiments arrive at the surprising conclusion that these datasets are useful to train neural ranking models for news articles.

Our second innovation involves the aggregation of sentence-level evidence for document ranking. That is, given an initial ranked list of documents, we segment each into sentences, and then apply inference over *each sentence* separately, after which sentence-level scores are aggregated to yield a final score for ranking documents. This approach, in fact, is well motivated: There is a long thread of work in the information retrieval literature, dating back decades, that leverages passage retrieval techniques for document ranking (Hearst and Plaunt, 1993; Callan, 1994; Kaszkiel and Zobel, 1997; Clarke et al., 2000). In addition, recent studies of human searchers (Zhang et al., 2018b,a)

revealed that the "best" sentence or paragraph in a document provides a good proxy for document-level relevance.

2.2 Model Details

The core of our model is a BERT *sentence-level* relevance classifier. Following Nogueira and Cho (2019), this is framed as a binary classification task. We form the input to BERT by concatenating the query Q and a sentence S into the sequence `[[CLS]`, Q, `[SEP]`, S, `[SEP]]` and padding each sequence in a mini-batch to the maximum length in the batch. We feed the final hidden state corresponding to the `[CLS]` token in the model to a single layer neural network whose output represents the probability that sentence S is relevant to the query Q.

To determine *document* relevance, we apply inference over each individual sentence in a candidate document, and then combine the top n scores with the original document score as follows:

$$S_f = a \cdot S_{doc} + (1 - \alpha) \cdot \sum_{i=1}^{n} w_i \cdot S_i \quad (1)$$

where S_{doc} is the original document score and S_i is the i-th top scoring sentence according to BERT. In other words, the relevance score of a document comes from the combination of a document-level term-matching score and evidence contributions from the top sentences in the document as determined by the BERT model. The parameters α and the w_i's can be tuned via cross-validation.

3 Experimental Setup

We begin with $\text{BERT}_{\text{Large}}$ (uncased, 340m parameters) from Devlin et al. (2019), and then fine-tune on the collections described in Section 2.1, individually and in combination. Despite the fact that tweets aren't always sentences and that MS MARCO and CAR passages, while short, may in fact contain more than one sentence, we treat all texts as if they were sentences for the purposes of fine-tuning BERT. For MS MARCO and CAR, we adopt exactly the procedure of Nogueira and Cho (2019). For MB, we tune on 2011–2014 data, with 75% of the total data reserved for training and the rest for validation.

We use the maximum sequence length of 512 tokens in all experiments. We train all models using cross-entropy loss for 5 epochs with a batch size of 16. We use Adam (Kingma and Ba, 2014) with an initial learning rate of 1×10^{-5}, linear learning rate warmup at a rate of 0.1 and decay of 0.1. All experiments are conducted on NVIDIA Tesla P40 GPUs with PyTorch v1.2.0.

During inference, we first retrieve an initial ranked list of documents to depth 1000 from the collection using the Anserini toolkit[1] (post-v0.5.1 commit from mid-August 2019, based on Lucene 8.0). Following Lin (2018) and Yang et al. (2019a), we use BM25 with RM3 query expansion (default parameters), which is a strong baseline, and has already been shown to beat most existing neural models. We clean the retrieved documents by stripping any HTML/XML tags and splitting each document into its constituent sentences with NLTK. If the length of a sentence with the meta-tokens exceeds BERT's maximum limit of 512, we further segment the spans into fixed size chunks. All sentences are then fed to the BERT model.

We conduct end-to-end document ranking experiments on three TREC newswire collections: the Robust Track from 2004 (Robust04) and the Common Core Tracks from 2017 and 2018 (Core17 and Core18). Robust04 comprises 250 topics, with relevance judgments on a collection of 500K documents (TREC Disks 4 and 5). Core17 and Core18 have only 50 topics each; the former uses 1.8M articles from the New York Times Annotated Corpus while the latter uses around 600K articles from the TREC Washington Post Corpus. Note that none of these collections were used to fine-tune the BERT relevance models; the only learned parameters are the weights in Eq (1).

Based on preliminary exploration, we consider up to the top three sentences; any more does not appear to yield better results. For Robust04, we follow the five-fold cross-validation settings in Lin (2018) and Yang et al. (2019a); for Core17 and Core18 we similarly apply five-fold cross validation. The parameters α and the w_i's are learned via exhaustive grid search as follows: we fix $w_1 = 1$ and then vary $a, w_2, w_3 \in [0, 1]$ with a step size 0.1, selecting the parameters that yield the highest average precision (AP). Retrieval results are reported in terms of AP, precision at rank 20 (P@20), and NDCG@20.

Code for replicating all the experiments described in this paper is available as part of our recently-developed Birch IR engine.[2] Additional

[1] http://anserini.io/
[2] http://birchir.io/

	Robust04			Core17			Core18		
Model	**AP**	**P@20**	**NDCG@20**	**AP**	**P@20**	**NDCG@20**	**AP**	**P@20**	**NDCG@20**
BM25+RM3	0.2903	0.3821	0.4407	0.2823	0.5500	0.4467	0.3135	0.4700	0.4604
1S: BERT(MB)	0.3408†	0.4335†	0.4900†	0.3091†	0.5620	0.4628	0.3393†	0.4930	0.4848†
2S: BERT(MB)	0.3435†	0.4386†	0.4964†	0.3137†	0.5770	0.4781	0.3421†	0.4910	0.4857†
3S: BERT(MB)	0.3434†	0.4422†	0.4998†	0.3154†	0.5880	0.4852†	0.3419†	0.4950†	0.4878†
1S: BERT(CAR)	0.3025†	0.3970†	0.4509	0.2814†	0.5500	0.4470	0.3120	0.4680	0.4586
2S: BERT(CAR)	0.3025†	0.3970†	0.4509	0.2814†	0.5500	0.4470	0.3116	0.4670	0.4585
3S: BERT(CAR)	0.3025†	0.3970†	0.4509	0.2814†	0.5500	0.4470	0.3113	0.4670	0.4584
1S: BERT(MS MARCO)	0.3028†	0.3964†	0.4512	0.2817†	0.5500	0.4468	0.3121	0.4670	0.4594
2S: BERT(MS MARCO)	0.3028†	0.3964†	0.4512	0.2817†	0.5500	0.4468	0.3121	0.4670	0.4594
3S: BERT(MS MARCO)	0.3028†	0.3964†	0.4512	0.2817†	0.5500	0.4468	0.3121	0.4670	0.4594
1S: BERT(CAR $\rightarrow$ MB)	0.3476†	0.4380†	0.4988†	0.3103†	0.5830	0.4758	0.3385†	0.4860	0.4785
2S: BERT(CAR $\rightarrow$ MB)	0.3470†	0.4400†	0.5015†	0.3140†	0.5830	0.4817†	0.3386†	0.4810	0.4755
3S: BERT(CAR $\rightarrow$ MB)	0.3466†	0.4398†	0.5014†	0.3143†	0.5830	0.4807	0.3382†	0.4830	0.4731
1S: BERT(MS MARCO $\rightarrow$ MB)	0.3676†	0.4610†	0.5239†	0.3292†	0.6080†	0.5061†	0.3486†	**0.4920**	**0.4953**†
2S: BERT(MS MARCO $\rightarrow$ MB)	**0.3697**†	0.4657†	0.5324†	**0.3323**†	0.6170†	**0.5092**†	0.3496†	0.4830	0.4899†
3S: BERT(MS MARCO $\rightarrow$ MB)	0.3691†	**0.4669**†	**0.5325**†	0.3314†	**0.6200**†	0.5070†	**0.3522**†	0.4850	0.4899†

Table 1: Ranking effectiveness on Robust04, Core17, and Core18 in terms of AP, P@20, and NDCG@20.

details about the technical design of our system are presented in Yilmaz et al. (2019), a companion demonstration paper.

4 Results and Discussion

Our main results are shown in Table 1. The top row shows the BM25+RM3 query expansion baseline using default Anserini parameters.[3] The remaining blocks display the ranking effectiveness of our models on Robust04, Core17, and Core18. In parentheses we describe the fine-tuning procedure: for instance, MSMARCO $\rightarrow$ MB refers to a model that was first fine-tuned on MS MARCO and then on MB. The nS preceding the model name indicates that inference was performed using the top n scoring sentences from each document.

Table 1 also includes results of significance testing using paired t-tests, comparing each condition with the BM25+RM3 baseline. We report significance at the $p < 0.01$ level, with appropriate Bonferroni corrections for multiple hypothesis testing. Statistically significant differences with respect to the baseline are denoted by †.

We find that BERT fine-tuned on MB alone significantly outperforms the BM25+RM3 baseline for all three metrics on Robust04. On Core17 and Core18, we observe significant increases in AP as well (and other metrics in some cases). In other words, relevance models learned from tweets successfully transfer over to news articles despite large differences in domain. This surprising result highlights the relevance matching power introduced by the deep semantic information learned by BERT.

Fine-tuning on MS MARCO or CAR alone yields at most minor gains over the baselines across all collections, and in some cases actually hurts effectiveness. Furthermore, the number of sentences considered for final score aggregation does not seem to affect effectiveness. It also does not appear that the synthetic nature of CAR data helps much for relevance modeling on newswire collections. Interestingly, though, if we fine-tune on CAR and then MB (CAR $\rightarrow$ MB), we obtain better results than fine-tuning on either MS MARCO or CAR alone. In some cases, we slightly improve over fine-tuning on MB alone. One possible explanation could be that CAR has an effect similar to language model pre-training; it alone cannot directly help the downstream document retrieval task, but it provides a better representation that can benefit from MB fine-tuning.

However, we were surprised by the MS MARCO results: since the dataset captures a search task and the web passages are "closer" to our newswire collections than MB in terms of domain, we would have expected relevance transfer to be more effective. Results show, however, that fine-tuning on MS MARCO alone is far less effective than fine-tuning on MB alone.

Looking across all fine-tuning configurations, we see that the top-scoring sentence of each candi-

[3]Even though Lin (2018) and Yang et al. (2019a) report slightly higher AP scores for *tuned* BM25+RM3 on Robust04, for consistency we use default parameters because no careful tuning has been performed for Core17 and Core18.

date document alone seems to be a good indicator of document relevance, corroborating the findings of Zhang et al. (2018a). Additionally considering the second ranking sentence yields at most a minor gain, and in some cases, adding a third actually causes effectiveness to drop. This is quite a surprising finding, since it suggests that the document ranking problem, at least as traditionally formulated by information retrieval researchers, can be distilled into relevance prediction primarily at the sentence level.

In the final block of the table, we present our best model, with fine-tuning on MS MARCO and then on MB. We confirm that this approach is able to exploit *both* datasets, with a score that is higher than fine-tuning on each dataset alone. Let us provide some broader context for these scores: For Robust04, we report the highest AP score that we are aware of (0.3697). Prior to our work, the meta-analysis by Yang et al. (2019a), which analyzed over 100 papers up until early 2019,[4] put the best neural model at 0.3124 (Dehghani et al., 2018).[5] Furthermore, our results exceed the previous highest known score of 0.3686, which is a non-neural method based on ensembles (Cormack et al., 2009). This high water mark has stood unchallenged for ten years.

Recently, MacAvaney et al. (2019) reported 0.5381 NDCG@20 on Robust04 by integrating contextualized word representations into existing neural ranking models; unfortunately, they did not report AP results. Our best NDCG@20 on Robust04 (0.5325) approaches their results even though we optimize for AP. Finally, note that since we are only using Robust04 data for learning the document and sentence weights in Eq (1), and not for fine-tuning BERT itself, it is less likely that we are overfitting.

Our best model also achieves a higher AP on Core17 than the best TREC submission that does not make use of past labels or human intervention (`umass_baselnrm`, 0.275 AP) (Allan et al., 2017). Under similar conditions, we beat every TREC submission in Core18 as well (with the best run being `uwmrg`, 0.276 AP) (Allan et al., 2018). Core17 and Core18 are relatively new and thus have yet to receive much attention from researchers, but to our knowledge, these figures represent the state of the art.

[4] https://github.com/lintool/robust04-analysis

[5] Setting aside our own previous work (Yang et al., 2019c).

5 Conclusion

This paper shows how BERT can be adapted in a simple manner to yield large improvements in *ad hoc* document retrieval on "classic" TREC newswire test collections. Our results demonstrate two surprising findings: first, that relevance models can be transferred quite straightforwardly across domains by BERT, and second, that effective document retrieval requires only "paying attention" to a small number of "top sentences" in each document. Important future work includes more detailed analyses of these transfer effects, as well as a closer look at the contributions of document-level and sentence-level scores. Nevertheless, we believe that both findings pave the way for new directions in document ranking.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and enabled by computational resources provided by Compute Ontario and Compute Canada.

References

James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen Voorhees. 2017. TREC 2017 Common Core Track overview. In *Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017)*, Gaithersburg, Maryland.

James Allan, Donna Harman, Evangelos Kanoulas, and Ellen Voorhees. 2018. TREC 2018 Common Core Track overview. In *Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018)*, Gaithersburg, Maryland.

Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM 2009)*, pages 601–610, Hong Kong, China.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv:1611.09268v3*.

James P. Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 302–310, Dublin, Ireland.

Charles L. A. Clarke, Gordon V. Cormack, and Elizabeth Tudhope. 2000. Relevance ranking for one to three term queries. *Information Processing and Management*, 36:291–311.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 758–759, Boston, Massachusetts.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. *arXiv:1907.06902*.

Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2018. Fidelity-weighted learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval overview. In *Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017)*, Gaithersburg, Maryland.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pages 56–68, Pittsburgh, Pennsylvania.

Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997)*, pages 178–185, Philadelphia, Pennsylvania.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Canjia Li, Yingfei Sun, Ben He, Le Wang, Kai Hui, Andrew Yates, Le Sun, and Jungang Xu. 2018. NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4482–4491, Brussels, Belgium.

Jimmy Lin. 2018. The neural hype and comparisons against weak baselines. *SIGIR Forum*, 52(2):40–51.

Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. Overview of the TREC-2014 Microblog Track. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*, Gaithersburg, Maryland.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1101–1104, Paris, France.

Ryan McDonald, George Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1860, Brussels, Belgium.

Bhaskar Mitra and Nick Craswell. 2019. An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 13(1):1–126.

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

Harshith Padigela, Hamed Zamani, and W. Bruce Croft. 2019. Investigating the successes and failures of BERT for passage re-ranking. *arXiv:1905.01758*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana.

Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of BERT in ranking. *arXiv:1904.07531*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report.

Jinfeng Rao, Wei Yang, Yuhao Zhang, Ferhan Ture, and Jimmy Lin. 2019. Multi-perspective relevance matching with hierarchical ConvNets for social media search. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 232–240, Honolulu, Hawaii.

Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the difficulty of evaluating baselines: a study on recommender systems. *arXiv:1905.01395*.

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. In *Proceedings of the 6th International Conference on Learning Representations, Workshop Track (ICLR 2018)*.

Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019a. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1129–1132, Paris, France.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019b. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019c. Simple applications of BERT for ad hoc document retrieval. *arXiv:1903.10972*.

Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Applying BERT to document retrieval with Birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2018a. Effective user interaction for high-recall retrieval: less is more. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, pages 187–196, Torino, Italy.

Haotian Zhang, Gordon V. Cormack, Maura R. Grossman, and Mark D. Smucker. 2018b. Evaluating sentence-level relevance feedback for high-recall information retrieval. *arXiv:1803.08988*.