

# Simple Applications of BERT for Ad Hoc Document Retrieval

Wei Yang,<sup>\*</sup> Haotian Zhang,<sup>\*</sup> and Jimmy Lin

David R. Cheriton School of Computer Science  
University of Waterloo

## Abstract

Following recent successes in applying BERT to question answering, we explore simple applications to *ad hoc* document retrieval. This required confronting the challenge posed by documents that are typically longer than the length of input BERT was designed to handle. We address this issue by applying inference on sentences individually, and then aggregating sentence scores to produce document scores. Experiments on TREC microblog and newswire test collections show that our approach is simple yet effective, as we report the highest average precision on these datasets by neural approaches that we are aware of.

## 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. Researchers have proposed many neural ranking models (Mitra and Craswell, 2019), but 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 log data only available to a few organizations.

One important recent innovation is the use of neural models that make heavy use of pretraining (Peters et al., 2018; Radford et al., 2018), culminating in BERT (Devlin et al., 2018), the most popular example of this approach today. Researchers have applied BERT to a broad range of NLP tasks and reported impressive gains. Most relevant to document retrieval, BERTserini (Yang et al., 2019) integrates passage retrieval using the open-source Anserini IR toolkit with a BERT-based reader to achieve large gains

over the previous state of the art in identifying answer spans from a large Wikipedia corpus.

Given the successes in applying BERT to question answering and the similarities between QA and document retrieval, we naturally wondered: Would it be possible to apply BERT to improve document retrieval as well? In short, the answer is yes. Adapting BERT for document retrieval requires overcoming the challenges associated with long documents, both during training and inference. We present a simple yet effective approach, based on the same BERTserini framework, that applies inference over individual sentences in a document and then combines sentence scores into document scores.

Our approach is evaluated on standard *ad hoc* retrieval test collections from the TREC Microblog Tracks (2011–2014) and the TREC 2004 Robust Track. We report the highest average precision on these datasets for neural approaches that we are aware of. The contribution of our work is, to our knowledge, the first successful application of BERT to *ad hoc* document retrieval, yielding state of the art results.

## 2 Background and Related Work

In *ad hoc* document retrieval, the system is given a short query  $q$  and the task is to produce the best ranking of documents in a corpus, according to some standard metric such as average precision (AP). Mitra and Craswell (2019) provide a recent overview of many of these models, to which we refer interested readers in lieu of a detailed literature review due to space considerations.

However, there are aspects of the task worth discussing. Researchers have understood for a few years now that *relevance matching* and *semantic matching* (for example, paraphrase detection, natural language inference, etc.) are different tasks,

<sup>\*</sup> equal contribution

despite shared common characteristics (Guo et al., 2016). The first task has a heavier dependence on exact match (i.e., “one-hot”) signals, whereas the second task generally requires models to more accurately capture semantics. Question answering has elements of both, but nevertheless remains a different task from document retrieval. Due to these task differences, neural models for document ranking, for example, DRMM (Guo et al., 2016), are quite different architecturally from neural models for capturing similarity; see, for example, the survey of Lan and Xu (2018).

Another salient fact is that documents can be longer than the length of input texts that BERT was designed for. This creates a problem during training because relevance judgments are annotations on documents, not on individual sentences or passages. Typically, within a relevant document, only a few passages are relevant, but such fine-grained annotations are not available in most test collections. Thus, it is unclear how exactly one would fine-tune BERT given (only) existing document-level relevance judgments. In this paper, we sidestep the training challenge completely and present a simple approach to aggregating sentence-level scores during inference.

### 3 Searching Social Media Posts

Despite the task mismatch between QA and *ad hoc* document retrieval, our working hypothesis is that BERT can be fine-tuned to capture relevance matching, as long as we can provide appropriate training data. To begin, we tackled microblog retrieval—searching short social media posts—where document length does not pose an issue. Fortunately, test collections from the TREC Microblog Tracks (Lin et al., 2014), from 2011 to 2014, provide data for exactly this task.

As with BERTserini, we adopted a simple architecture that uses the Anserini IR toolkit<sup>1</sup> for initial retrieval, followed by inference using a BERT model. Building on best practice, query likelihood (QL) with RM3 relevance feedback (Abdul-Jaleel et al., 2004) provides the initial ranking to depth 1000. The texts of the retrieved documents (posts) are then fed into a BERT classifier, and the BERT scores are combined with the retrieval scores via linear interpolation. We used the BERT-Base model (uncased, 12-layer, 768-hidden, 12-heads, 110M pa-

rameters) described in Devlin et al. (2018). As input, we concatenated the query  $Q$  and the document  $D$  into a text sequence  $[[CLS], Q, [SEP], D, [SEP]]$ , and then padded each text sequence in a mini-batch to  $N$  tokens, where  $N$  is the maximum length in the batch. Following Nogueira and Cho (2019), BERT is used for binary classification (i.e., relevance) by taking the [CLS] vector as input to a single layer neural network.

Test collections from the TREC Microblog Tracks were used for fine-tuning the BERT model, using cross-entropy loss. For evaluation on each year’s dataset, we used the remaining years for fine tuning, e.g., tuning on 2011–2013 data, testing on 2014 data. From the training data, we sampled 10% for validation. We fine-tuned BERT with a learning rate of  $3 \times 10^{-6}$  for 10 epochs. The interpolation weight between the BERT scores and the retrieval scores was tuned on the validation data. We only used as training examples the social media posts that appear in our initial ranking (i.e., as opposed to all available relevance judgments). There are a total of 225 topics (50, 60, 60, 55) in the four datasets, which yields 225,000 examples (unjudged posts are treated as not relevant).

Experimental results are shown in Table 1, where we present average precision (AP) and precision at rank 30 (P30), the two official metrics of the evaluation (Ounis et al., 2011). The first two blocks of the table are copied from Rao et al. (2019), who compared bag-of-words baselines (QL and RM3) to several popular neural ranking models as well as MP-HCNN, the model they introduced. Results for all the neural models include interpolation with the original document scores. Rao et al. (2019) demonstrated that previous neural models are not suitable for ranking short social media posts, and are no better than the RM3 baseline in many cases. In contrast, MP-HCNN was explicitly designed with characteristics of tweets in mind: it significantly outperforms previous neural ranking models (see original paper for comparisons, not repeated here). We also copied results from Shi et al. (2018), who reported even higher effectiveness than MP-HCNN.

These results represent, to our knowledge, the most comprehensive summary of search effectiveness measured on the TREC Microblog datasets. Note that for these comparisons we leave aside many non-neural approaches that take advantage of learning-to-ranking techniques over manually-

<sup>1</sup><http://anserini.io/>

Model	2011		2012		2013		2014	
	AP	P30	AP	P30	AP	P30	AP	P30
QL	0.3576	0.4000	0.2091	0.3311	0.2532	0.4450	0.3924	0.6182
RM3	0.3824	0.4211	0.2342	0.3452	0.2766	0.4733	0.4480	0.6339
DRMM (Guo et al., 2016)	0.3477	0.4034	0.2213	0.3537	0.2639	0.4772	0.4042	0.6139
DUET (Mitra et al., 2017)	0.3576	0.4000	0.2243	0.3644	0.2779	0.4878	0.4219	0.6467
K-NRM (Xiong et al., 2017)	0.3576	0.4000	0.2277	0.3520	0.2721	0.4756	0.4137	0.6358
PACRR (Hui et al., 2017)	0.3810	0.4286	0.2311	0.3576	0.2803	0.4944	0.4140	0.6358
MP-HCNN (Rao et al., 2019)	0.4043	0.4293	0.2460	0.3791	0.2896	0.5294	0.4420	0.6394
BiCNN (Shi et al., 2018)	0.4293	0.4728	0.2621	0.4147	0.2990	0.5367	0.4563	0.6806
BERT	<b>0.4697</b>	<b>0.5040</b>	<b>0.3073</b>	<b>0.4356</b>	<b>0.3357</b>	<b>0.5656</b>	<b>0.5176</b>	<b>0.7006</b>

Table 1: Results on test collections from the TREC Microblog Tracks, comparing BERT with selected neural ranking models. The first two blocks of the table contain results copied from Rao et al. (2019).

engineered features, as we do not believe they form a fair basis of comparison. In general, such approaches also take advantage of non-textual features (e.g., social signals), and these additional signals (naturally) allow them to beat approaches that use only the text of the social media posts (like all the models discussed here).

The final row of Table 1 reports results using our simple BERT-based technique, showing quite substantial and consistent improvements over previous results. Since we have directly copied results from previous papers, we did not conduct significance tests.

#### 4 Searching Newswire Articles

Results on the microblog test collections confirm our working hypothesis that BERT can be fine-tuned to capture document relevance, at least for short social media posts. In other words, task differences between QA and document retrieval do not appear to hinder BERT’s adaptability. Having demonstrated this, we turn our attention to longer documents. For this, we take advantage of the test collection from the TREC 2004 Robust Track (Voorhees, 2004), which comprises 250 topics over a newswire corpus. We selected this collection for a couple of reasons: it is the largest newswire collection we know of in terms of training data, and Lin (2018) provides well-tuned baselines that support fair comparisons to recent neural ranking models.

Given the success of BERT on microblogs, one simple idea is to apply inference over each sentence in a candidate document, select the one with the highest score, and then combine that with the original document score (with linear interpo-

lation). One rationale for this approach comes from Zhang et al. (2018b,a), who found that the “best” sentence or paragraph in a document provides a good proxy for document relevance. This is also consistent with a long thread of work in information retrieval that leverages passage retrieval techniques for document ranking (Callan, 1994; Clarke et al., 2000; Liu and Croft, 2002).

Generalizing, we could consider the top  $n$  scoring sentences as follows:

$$\text{Score}_d = a \cdot S_{\text{doc}} + (1 - a) \cdot \sum_{i=1}^n w_i \cdot S_i$$

where  $S_{\text{doc}}$  is the original document score and  $S_i$  is the  $i$ -th top scoring sentence according to BERT. The hyperparameters  $a$  and  $w_i$  can be tuned via cross-validation.

Sentence-level inference seems like a reasonable initial attempt at adapting BERT to document retrieval, but what about fine-tuning? As previously discussed, the issue is that we lack sentence-level relevance judgments. Since our efforts represent an initial exploration, we simply sidestep this challenge (for now) and fine tune on existing sentence-level datasets. Specifically, we used: (1) the microblog data from the previous section and (2) the union of the TrecQA (Yao et al., 2013) and WikiQA (Yang et al., 2015) datasets. This sets up an interesting contrast: the first dataset captures the document retrieval task but on a different domain, while the second dataset captures a different task but on corpora that are much closer to newswire. It is an empirical question as to which source is more effective.

To support a fair comparison, we adopted the same experimental procedure as Lin (2018). He

described two separate data conditions: one based on two-fold cross-validation to compare against “Paper 1” and one based on five-fold cross-validation to compare against “Paper 2”.<sup>2</sup> The exact fold settings are provided online, which ensures a fair comparison.<sup>3</sup> In our implementation, documents are first cleaned by stripping all tags and then segmenting the text into sentences using NLTK. If the input to BERT is longer than 512 tokens (BERT’s maximum limit), we further split sentences into fixed sized chunks. Across the 250 topics, each document averages 43 sentences, with 27 tokens per sentence.

In our experiments, we considered up to the top four sentences. For up to three sentences,  $a$  and  $w_i$  are tuned via exhaustive grid search in the following range:  $a \in [0, 1]$ ,  $w_1 = 1$  (fixed),  $w_2 \in [0, 1]$ , and  $w_3 \in [0, 1]$ , all with step size 0.1. In the four-sentence condition, to reduce the search space, we started with the best three-sentence parameters and explored  $w_4 \in [0, 1]$  with step size 0.1, along with neighboring regions in  $a$ ,  $w_2$ , and  $w_3$ . We selected the parameters with the highest AP score on the training folds.

Results of our experiments are shown in Table 2, divided into two blocks: Paper 1 on the top and Paper 2 on the bottom. The effectiveness of the two papers are directly copied from Lin (2018); all other results are our own runs. The paper aggregation site “Papers With Code” places Lin’s result as the state of the art on Robust04 as of this writing.<sup>4</sup> As a point of comparison, in the most recent survey of neural ranking models by Guo et al. (2019), the best AP on Robust04 is in the 0.29 range, consistent with the above site. Therefore, we are quite confident that we are evaluating against competitive models. In the results table, “FT” indicates the dataset used for fine-tuning and  $nS$  indicates inference using the top  $n$  scoring sentences of the document. We find that the learned  $w_4$  value is zero, indicating that additional sentences do not help beyond the top three (at least according to our tuning procedure); thus, 4S results are omitted from the table. Interestingly,

<sup>2</sup>Since Lin’s article is critical of neural methods, he anonymized the neural approaches but mentioned that they come from articles published in late 2018 and are representative of the most recent advances in neural approaches to document retrieval.

<sup>3</sup><https://github.com/castorini/Anserini/blob/master/docs/experiments-forum2018.md>

<sup>4</sup><https://paperswithcode.com/sota/ad-hoc-information-retrieval-trec-robust>

Model	AP	P20
Paper 1 (two fold)	0.2971	0.3948
BM25+RM3	0.2987	0.3871
1S: BERT FT(QA)	0.3014	0.3928
2S: BERT FT(QA)	0.3003	0.3948
3S: BERT FT(QA)	0.3003	0.3948
1S: BERT FT(Microblog)	0.3241	0.4217
2S: BERT FT(Microblog)	0.3240	0.4209
3S: BERT FT(Microblog)	<b>0.3244</b>	<b>0.4219</b>
Paper 2 (five fold)	0.272	0.386
BM25+RM3	0.3033	0.3974
1S: BERT FT(QA)	0.3102	0.4068
2S: BERT FT(QA)	0.3090	0.4064
3S: BERT FT(QA)	0.3090	0.4064
1S: BERT FT(Microblog)	0.3266	0.4245
2S: BERT FT(Microblog)	<b>0.3278</b>	0.4267
3S: BERT FT(Microblog)	<b>0.3278</b>	<b>0.4287</b>

Table 2: Results on Robust04. FT indicates the dataset used for fine tuning;  $nS$  indicates inference using the top  $n$  scoring sentences of the document.

we find that fine-tuning BERT on microblog data is more effective than QA data, suggesting that task (QA vs. relevance matching) is more important than document genre (tweets vs. newswire). Cognizant of the potential dangers of repeated hypothesis testing, we probed the statistical significance of one five-fold setting, BM25+RM3 vs. “3S: BERT FT (Microblog)”. According to a paired  $t$ -test, the differences are statistically significant ( $p < 10^{-7}$ ).

As a summary, we see that a well-tuned BM25+RM3 baseline already outperforms neural ranking approaches (which was Lin’s original point). Our simple BERT-based reranker yields further significant improvements.

## 5 Conclusions

In this preliminary study, we have adapted BERT for document retrieval in the most obvious manner, via sentence-level inference and simple score aggregation. Results show substantial improvements in both ranking social media posts and newswire documents—to our knowledge, the highest AP scores reported on the TREC Microblog and Robust04 datasets for neural approaches that we are aware of (although the literature does report non-neural approaches that are even better, for both tasks). We readily con-

cede that our techniques are quite simple and that there are many obvious next steps. In particular, we simply sidestepped the issue of not having sentence-level relevance judgments, although there are some obvious distant supervision techniques to “project” relevance labels down to the sentence level that should be explored. We are actively pursuing these and other directions.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## References

- Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Donald Metzler, Mark D. Smucker, Trevor Strohman, Howard Turtle, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*.
- 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 ’94, pages 302–310.
- Charles Clarke, Gordon Cormack, and Elizabeth Tudhope. 2000. Relevance ranking for one to three term queries. *Information Processing and Management*, 36:291–311.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.
- Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM ’16, pages 55–64.
- Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *arXiv:1903.06902v1*.
- Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A position-aware neural IR model for relevance matching. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1049–1058.
- Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902.
- 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)*.
- Xiaoyong Liu and W. Bruce Croft. 2002. Passage retrieval based on language models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM ’02, pages 375–382.
- Bhaskar Mitra and Nick Craswell. 2019. An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 13(1):1–126.
- Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW ’17, pages 1291–1299.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv:1901.04085*.
- Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. 2011. Overview of the TREC-2011 Microblog Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*.
- 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.
- 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. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*.
- Peng Shi, Jinfeng Rao, and Jimmy Lin. 2018. Simple attention-based representation learning for ranking short social media posts. *arxiv:1811.01013*.
- Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 52–69.



- 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 the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 55–64.
- Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. *arXiv:1902.01718*.
- Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.
- Xuchen Yao, Benjamin Van Durme, Chris Callison-burch, and Peter Clark. 2013. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 858–867.
- 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 '18*, pages 187–196.
- 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*.