INCORPORATING QUERY TERM INDEPENDENCE ASSUMPTION FOR EFFICIENT RETRIEVAL AND RANKING USING DEEP NEURAL NETWORKS

A PREPRINT

Bhaskar Mitra[*]
Microsoft AI & Research
bmitra@microsoft.com

Corby Rosset[*]
Microsoft AI & Research
corosset@microsoft.com

David Hawking
david.hawking@acm.org

Nick Craswell
Microsoft AI & Research
nickcr@microsoft.com

Fernando Diaz
Microsoft AI & Research
fdiaz@microsoft.com

Emine Yilmaz
Microsoft AI & Research
emine.yilmaz@ucl.ac.uk

July 9, 2019

ABSTRACT

Classical information retrieval (IR) methods, such as query likelihood and BM25, score documents independently *w.r.t.* each query term, and then accumulate the scores. Assuming *query term independence* allows precomputing term-document scores using these models—which can be combined with specialized data structures, such as inverted index, for efficient retrieval. Deep neural IR models, in contrast, compare the whole query to the document and are, therefore, typically employed only for late stage re-ranking. We incorporate query term independence assumption into three state-of-the-art neural IR models: BERT, Duet, and CKNRM—and evaluate their performance on a passage ranking task. Surprisingly, we observe no significant loss in result quality for Duet and CKNRM—and a small degradation in the case of BERT. However, by operating on each query term independently, these otherwise computationally intensive models become amenable to offline precomputation—dramatically reducing the cost of query evaluations employing state-of-the-art neural ranking models. This strategy makes it practical to use deep models for retrieval from large collections—and not restrict their usage to late stage re-ranking.

***K*eywords** Deep learning · Information retrieval · Indexing · Query evaluation

1 Introduction

Many traditional information retrieval (IR) ranking functions—*e.g.*, [Robertson et al., 2009, Ponte and Croft, 1998]—manifest the *query-term independence* property—*i.e.*, the documents can be scored independently *w.r.t.* each query term, and then the scores accumulated. Given a document collection, these term-document scores can be precomputed and combined with specialized IR data structures, such as inverted indexes [Zobel and Moffat, 2006], and clever organization strategies (*e.g.*, impact-ordering [Anh et al., 2001]) to aggressively prune the set of documents that need to be assessed per query. This dramatically speeds up query evaluations enabling fast retrieval from large collections, containing billions of documents.

Recent deep neural architectures—such as BERT [Nogueira and Cho, 2019], Duet [Mitra et al., 2017], and CKNRM [Dai et al., 2018]—have demonstrated state-of-the-art performance on several IR tasks. However, the superior retrieval effectiveness comes at the cost of evaluating deep models with tens of millions to hundreds of millions of parameters at query evaluation time. In practice, this limits the scope of these models to late stage re-ranking. Like traditional IR

[*]Both authors contributed equally to this research.

models, we can incorporate the query term independence assumption into the design of the deep neural model—which would allow offline precomputation of all term-document scores. The query evaluation then involves only their linear combination—alleviating the need to run the computation intensive deep model at query evaluation time. We can further combine these precomputed machine-learned relevance estimates with an inverted index, to retrieve from the full collection. This significantly increases the scope of potential impact of neural methods in the retrieval process. We study this approach in this work.

Of course, by operating independently per query term, the ranking model has access to less information compared to if it has the context of the full query. Therefore, we expect the ranking model to show some loss in retrieval effectiveness under this assumption. However, we trade this off with the expected gains in efficiency of query evaluations and the ability to retrieve, and not just re-rank, using these state-of-the-art deep neural models.

In this preliminary study, we incorporate the query term independence assumption into three state-of-the-art neural ranking models—BERT [Nogueira and Cho, 2019], Duet [Mitra et al., 2017], and CKNRM [Dai et al., 2018]—and evaluate their effectiveness on the MS MARCO passage ranking task [Bajaj et al., 2016]. We surprisingly find that the two of the models suffer no statistically significant adverse affect *w.r.t.* ranking effectiveness on this task under the query term independence assumption. While the performance of BERT degrades under the strong query term independence assumption—the drop in MRR is reasonably small and the model maintains a significant performance gap compared to other non-BERT based approaches. We conclude that at least for a certain class of existing neural IR models, incorporating query term independence assumption may result in significant efficiency gains in query evaluation at minimal (or no) cost to retrieval effectiveness.

2 Related work

Several neural IR methods—*e.g.*, [Ganguly et al., 2015, Kenter and De Rijke, 2015, Nalisnick et al., 2016, Guo et al., 2016]—already operate under query term independence assumption. However, recent performance breakthroughs on many IR tasks have been achieved by neural models [Hu et al., 2014, Pang et al., 2016, Mitra et al., 2017, Dai et al., 2018, Nogueira and Cho, 2019] that learn latent representations of the query or inspect interaction patterns between query and document terms. In this work, we demonstrate the potential to incorporate query term independence assumption in these recent representation learning and interaction focused models.

Some neural IR models [Huang et al., 2013, Gao et al., 2011] learn low dimensional dense vector representations of query and document that can be computed independently during inference. These models are also amenable to precomputation of document representations—and fast retrieval using approximate nearest neighbor search [Aumüller et al., 2017, Boytsov et al., 2016]. An alternative involves learning higher dimensional but sparse representations of query and document [Salakhutdinov and Hinton, 2009, Zamani et al., 2018a] that can also be employed for fast lookup. However, these approaches—where the document representation is computed independently of the query—do not allow for interactions between the query term and document representations. Early interaction between query and document representation is important to many neural architectures [Hu et al., 2014, Pang et al., 2016, Mitra et al., 2017, Dai et al., 2018, Nogueira and Cho, 2019]. The approach proposed in this study allows for interactions between individual query terms and documents.

Finally, we refer the reader to [Mitra and Craswell, 2018] for a more general survey of neural methods for IR tasks.

3 Neural Ranking Models with Query Term Independence Assumption

IR functions that assume query term independence observe the following general form:

$$S_{q,d} = \sum_{t \in q} s_{t,d} \tag{1}$$

Where, $s \in \mathbb{R}_{\geq 0}^{|V| \times |C|}$ is the set of positive real-valued scores as estimated by the relevance model corresponding to documents $d \in C$ in collection C *w.r.t.* to terms $t \in V$ in vocabulary V—and $S_{q,d}$ denotes the aggregated score of document d *w.r.t.* to query q. For example, in case of BM25 [Robertson et al., 2009]:

$$s_{t,d} = \text{idf}_t \cdot \frac{\text{tf}_{td} \cdot (k_1 + 1)}{\text{tf}_{td} + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{avgdl}\right)} \tag{2}$$

Where, tf and idf denote term-frequency and inverse document frequency, respectively—and k_1 and b are the free parameters of the BM25 model.

Deep neural models for ranking, in contrast, do not typically assume query term independence. Instead, they learn complex matching functions to compare the candidate document to the full query. The parameters of such a model ϕ is typically learned discriminatively by minimizing a loss function of the following form:

$$\mathcal{L} = \mathbb{E}_{q\sim\theta_q,\, d_+\sim\theta_{d_+}, d_-\sim\theta_{d_-}} [\ell(\Delta_{q,d_+,d_-})] \tag{3}$$

$$\text{where,} \quad \Delta_{q,d_+,d_-} = \phi_{q,d_+} - \phi_{q,d_-} \tag{4}$$

We use d_+ and d_- to denote a pair of relevant and non-relevant documents, respectively, *w.r.t.* query q. The instance loss ℓ in Equation 3 can take different forms—*e.g.*, ranknet [Burges et al., 2005] or hinge [Herbrich et al., 2000] loss.

$$\ell_{\text{ranknet}}(\Delta_{q,d_+,d_-}) = \log(1 + e^{-\sigma\cdot\Delta_{q,d_+,d_-}}) \tag{5}$$

$$\ell_{\text{hinge}}(\Delta_{q,d_+,d_-}) = \max\{0, \epsilon - \Delta_{q,d_+,d_-}\} \tag{6}$$

Given a neural ranking model ϕ, we define Φ—the corresponding model under the query term independence assumption—as:

$$\Phi_{q,d} = \sum_{t\in q} \phi_{t,d} \tag{7}$$

The new model Φ preserves the same architecture as ϕ but estimates the relevance of a document independently *w.r.t.* each query term. The parameters of Φ are learned using the modified loss:

$$\mathcal{L} = \mathbb{E}_{q\sim\theta_q,\, d_+\sim\theta_{d_+}, d_-\sim\theta_{d_-}} [\ell(\delta_{q,d_+,d_-})] \tag{8}$$

$$\text{where,} \quad \delta_{q,d_+,d_-} = \sum_{t\in q} \phi_{t,d_+} - \phi_{t,d_-} \tag{9}$$

Given collection C and vocabulary V, we precompute $\phi_{t,d}$ for all $t \in V$ and $d \in C$. In practice, the total number of combinations of t and d may be large but we can enforce additional constraints on which $\langle t, d\rangle$ pairs to evaluate, and assume no contributions from remaining pairs. During query evaluation, we can lookup the precomputed score $\phi_{t,d}$ without dedicating any additional time and resource to evaluate the deep ranking model. We employ an inverted index, in combination with the precomputed scores, to perform retrieval from the full collection using the learned relevance function Φ. We note that several IR data structures assume that $\phi_{t,d}$ be always positive which may not hold for any arbitrary neural architecture. But this can be addressed by applying a rectified linear unit activation on the model's output. The remainder of this paper describes our empirical study and summarizes our findings.

4 Experiments

4.1 Task description

We study the effect of the query term independence assumption on deep neural IR models in the context of the MS MARCO passage ranking task [Bajaj et al., 2016]. We find this ranking task to be suitable for this study for several reasons. Firstly, with one million question queries sampled from Bing's search logs, 8.8 million passages extracted from web documents, and 400,000 positively labeled query-passage pairs for training, it is one of the few large datasets available today for benchmarking deep neural IR methods. Secondly, the challenge leaderboard[2]—with 18 entries as of March 3, 2019—is a useful catalog of approaches that show state-of-the-art performance on this task. Conveniently, several of these high-performing models include public implementations for the ease of reproducibility.

The MS MARCO passage ranking task comprises of one thousand passages per query that the IR model, being evaluated, should re-rank. Corresponding to every query, one or few passages have been annotated by human editors as

[2] `http://www.msmarco.org/leaders.aspx`

Table 1: Comparing ranking effectiveness of BERT, Duet, and CKNRM with the query independence assumption (denoted as "Term ind.") with their original counterparts (denoted as "Full"). The difference between the median MRR for "full" and "term ind." models are *not* statistically significant based on a student's t-test ($p < 0.05$) for Duet and CKNRM. The difference in MRR is statistically significant based on a student's t-test ($p < 0.05$) for BERT (single run). The BM25 baseline (single run) is included for reference.

Model	MRR@10		
	Mean	(± Std. dev)	Median
BERT			
Full	0.356		0.356
Term ind.	**0.333**		**0.333**
Duet			
Full	0.239	(±0.002)	0.240
Term ind.	0.244	(±0.002)	0.244
CKNRM			
Full	0.223	(±0.004)	0.224
Term ind.	0.222	(±0.005)	0.221
BM25	0.167		0.167

containing the answer relevant to the query. The rank list produced by the model is evaluated using the mean reciprocal rank (MRR) metric against the ground truth annotations. We use the MS MARCO training dataset to train all baseline and treatment models, and report their performance on the publicly available development set which we consider—and hereafter refer to—as the test set for our experiments. This test set contains about seven thousand queries which we posit is sufficient for reliable hypothesis testing.

Note that the thousand passages per query were originally retrieved using BM25 from a collection that is provided as part of the MS MARCO dataset. This allows us to also use this dataset in a retrieval setting—in addition to the re-ranking setting used for the official challenge. We take advantage of this in our study.

4.2 Baseline models

We begin by identifying models listed on the MS MARCO leaderboard that can serve as baselines for our work. We only consider the models with public implementations. We find that a number of top performing entries—*e.g.*, [Nogueira and Cho, 2019]—are based on recently released large scale language model called BERT [Devlin et al., 2018]. The BERT based entries are followed in ranking by the Duet [Mitra et al., 2017] and the Convolutional Kernel-based Neural Ranking Model (CKNRM) [Dai et al., 2018]. Therefore, we limit this study to BERT, Duet, and CKNRM.

BERT Nogueira and Cho [2019] report state-of-the-art retrieval performance on the MS MARCO passage re-ranking task by fine tuning BERT [Devlin et al., 2018] pretrained models. In this study, we reproduce the results from their paper corresponding to the BERT Base model and use it as our baseline. Under the term independence assumption, we evaluate the BERT model once per query term—wherein we input the query term as sentence A and the passage as sentence B.

Duet The Duet [Mitra et al., 2017] model estimates the relevance of a passage to a query by a combination of (i) examining the patterns of exact matches of query terms in the passage, and (ii) computing similarity between learned latent representations of query and passage. Duet has previously demonstrated state-of-the-art performance on TREC CAR [Nanni et al., 2017] and is an official baseline for the MS MARCO challenge. The particular implementation of Duet listed on the leaderboard includes modifications[3] to the original model [Mitra and Craswell, 2019]. We use this provided implementation for our study. Besides evaluating the model once per query term, no additional changes were necessary to its architecture under the query term independence assumption.

CKNRM The CKNRM model combines kernel pooling based soft matching [Xiong et al., 2017] with a convolutional architecture for comparing n-grams. CKNRM uses kernel pooling to extract ranking signals from interaction matrices of query and passage n-grams. Under the query term independence assumption, the model considers one

[3] `https://github.com/dfcf93/MSMARCO/blob/master/Ranking/Baselines/Duet.ipynb`

Table 2: Comparing Duet (with query term independence assumption) and BM25 under the full retrieval settings on a subset of MS MARCO dev queries. The differences in recall and MRR between Duet (term ind.) and BM25 are statistically significant according to student's t-test ($p < 0.01$).

Model	Recall@1000	MRR@10
BM25	0.80	0.169
Duet (term ind.)	**0.85**	**0.218**

query term at a time—and therefore we only consider the interactions between the query unigrams and passage n-grams. We base our study on the public implementation[4] of this model.

For all models we re-use the published hyperparameter values and other settings from the MS MARCO website.

5 Results

Table 1 compares the BERT, the Duet, and the CKNRM models trained under the query term independence assumption to their original counterparts on the passage re-ranking task. We train and evaluate the Duet and the CKNRM based models five and eight times, respectively, using different random seeds—and report mean and median MRR. For the BERT based models, due to long training time we only report results based on a single training and evaluation run. As table 1 shows, we observe no statistically significant difference in effectiveness from incorporating the query term independence assumptions in either Duet or CKNRM. The query term independent BERT model performs slightly worse than its original counterpart on MRR but the performance is still superior to other non-BERT based approaches listed on the public leaderboard.

We posit that models with query term independence assumption—even when slightly less effective compared to their full counterparts—are likely to retrieve better candidate sets for re-ranking. To substantiate this claim, we conduct a small-scale retrieval experiment based on a random sample of 395 queries from the test set. We use the Duet model with the query term independence assumption to precompute the term-passage scores constrained to (i) the term appears at least once in the passage, and (ii) the term does not appear in more than 5% of the passage collection. Table 2 compares Duet and BM25 on their effectiveness as a first stage retrieval method in a potential telescoping setting [Matveeva et al., 2006]. We observe a 6.25% improvement in recall@1000 from Duet over the BM25 baseline. To perform similar retrieval from the full collection using the full Duet model, unlike its query-term-independent counterpart, is prohibitive because it involves evaluating the model on every passage in the collection against every incoming query.

6 Discussion and conclusion

The emergence of compute intensive ranking models, such as BERT, motivates rethinking how these models should be evaluated in large scale IR systems. The approach proposed in this paper moves the burden of model evaluation from the query evaluation stage to the document indexing stage. This may have further consequences on computational efficiency by allowing batched model evaluation that more effectively leverages GPU (or TPU) parallelization.

This preliminary study is based on three state-of-the-art deep neural models on a public passage ranking benchmark. The original design of all three models—BERT, Duet, and CKNRM—emphasize on early interactions between query and passage representations. However, we observe that limiting the interactions to passage and individual query terms has reasonably small impact on their effectiveness. These results are promising as they support the possibility of dramatically speeding up query evaluation for some deep neural models, and even employing them to retrieve from the full collection. The ability to retrieve—and not just re-rank—using deep models has significant implications for neural IR research. Any loss in retrieval effectiveness due to incorporating strong query term independence assumptions may be further recovered by additional stages of re-ranking in a telescoping approach [Matveeva et al., 2006].

This study is focused on the passage ranking task. The trade-off between effectiveness and efficiency may be different for document retrieval and other IR tasks. Traditional IR methods in more complex retrieval settings—*e.g.*, when the document is represented by multiple fields [Robertson et al., 2004]—also observe the query term independence assumption. So, studying the query term independence assumption in the context of corresponding neural models—*e.g.*, [Zamani et al., 2018b]—may also be appropriate. We note these as important future directions for our research.

[4] `https://github.com/thunlp/Kernel-Based-Neural-Ranking-Models`

The findings from this study may also be interpreted as pointing to a gap in our current state-of-the-art neural IR models that do not take adequate advantage of term proximity signals for matching. This is another finding that may hold interesting clues for IR researchers who want to extract more retrieval effectiveness from deep neural methods.

References

Vo Ngoc Anh, Owen de Kretser, and Alistair Moffat. Vector-space ranking with effective early termination. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 35–42. ACM, 2001.

Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. In *International Conference on Similarity Search and Applications*, pages 34–49. Springer, 2017.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Leonid Boytsov, David Novak, Yury Malkov, and Eric Nyberg. Off the beaten path: Let's replace term-based retrieval with k-nn search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1099–1108. ACM, 2016.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134. ACM, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth JF Jones. Word embedding based generalized language model for information retrieval. In *Proc. SIGIR*, pages 795–798. ACM, 2015.

Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. Clickthrough-based latent semantic models for web search. In *Proc. SIGIR*, pages 675–684. ACM, 2011.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proc. CIKM*, pages 55–64. ACM, 2016.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*, 2000.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Proc. NIPS*, pages 2042–2050, 2014.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proc. CIKM*, pages 2333–2338. ACM, 2013.

Tom Kenter and Maarten De Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1411–1420. ACM, 2015.

Irina Matveeva, Chris Burges, Timo Burkard, Andy Laucius, and Leon Wong. High accuracy retrieval with multiple nested ranker. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 437–444. ACM, 2006.

Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval (to appear)*, 2018.

Bhaskar Mitra and Nick Craswell. An updated duet model for passage re-ranking. *arXiv preprint arXiv:1903.07666*, 2019.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proc. WWW*, pages 1291–1299, 2017.

Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In *Proc. WWW*, 2016.

Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. Benchmark for complex answer retrieval. In *Proc. ICTIR*. ACM, 2017.

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

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *Proc. AAAI*, 2016.

Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proc. SIGIR*, pages 275–281. ACM, 1998.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM, 2004.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50 (7):969–978, 2009.

Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 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*, pages 55–64. ACM, 2017.

Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proc. CIKM*, pages 497–506. ACM, 2018a.

Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. Neural ranking models with multiple document fields. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 700–708. ACM, 2018b.

Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM computing surveys (CSUR)*, 38(2):6, 2006.