On the Reliability of Test Collections for Evaluating Systems of Different Types

Emine Yilmaz
emine.yilmaz@ucl.ac.uk
University College London
London, UK

Nick Craswell
nickcr@microsoft.com
Microsoft
Redmond, USA

Bhaskar Mitra
bmitra@microsoft.com
Microsoft, University College London
Montréal, Canada

Daniel Campos
dacamp@microsoft.com
Microsoft, University of Washington
Redmond, USA

ABSTRACT

As deep learning based models are increasingly being used for information retrieval (IR), a major challenge is to ensure the availability of test collections for measuring their quality. Test collections are generated based on pooling results of various retrieval systems, but until recently this did not include deep learning systems. This raises a major challenge for reusable evaluation: Since deep learning based models use external resources (e.g. word embeddings) and advanced representations as opposed to traditional methods that are mainly based on lexical similarity, they may return different types of relevant document that were not identified in the original pooling. If so, test collections constructed using traditional methods are likely to lead to biased and unfair evaluation results for deep learning (neural) systems. This paper uses simulated pooling to test the fairness and reusability of test collections, showing that pooling based on traditional systems only can lead to biased evaluation of deep learning systems.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Retrieval models and ranking**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Test collection, pooling, evaluation, deep learning

1 INTRODUCTION

In recent years, deep neural models achieved state-of-the-art performance on a variety of tasks, and this has happened in a variety of fields ranging from computer vision to information retrieval (IR). It took relatively longer to observe such advances in core IR problems such as ranking [7], especially if we exclude results based on proprietary data—e.g., [8]. Two possible explanations for this delay are related to training data and test data: 1) The lack of large-scale training datasets with tens or hundreds of thousands of queries, since large data would seem to be a requirement based on the experience in other fields, and 2) The lack of test collections to evaluate the quality of neural models in a fair and reliable manner.

The TREC 2019 Deep Learning Track [6] addressed these problems by releasing large-scale training data, as well as by developing reliable and reusable test collections for evaluating the quality of various algorithms (ranging from traditional retrieval models such as BM25 to various neural models). The results of the track showed that when sufficient training data is available, most neural models tend to outperform the traditional retrieval models.

Findings of the track were based on test collections that were created using depth-10 pools of both neural and traditional models. Inclusion of neural runs in pooling is highly unusual, since most test collections were created in the years before neural models had been developed, and even those developed more recently (e.g., [1]) did not have neural models trained on the large labeled datasets that were introduced in TREC 2019.

Even though reusability of test collections for evaluating the quality of unseen systems has been widely studied in literature [4, 5, 9, 10, 13, 13], no previous work has analysed the reusability of test collections when they are created solely using systems of a particular type (e.g. traditional systems based on BM25, language modelling, etc.) towards evaluating the quality of systems that are of different type (e.g. neural systems based on deep learning models), which is the main question we aim to answer in this paper. Our approach is to simulate the earlier test collections, where pooling was with one type of model, to see whether this creates a bias against the other type of model, both when comparing within type and across types.

Our results demonstrate that evaluation results obtained using test collections that are created solely using pools of traditional systems are less reliable in terms of evaluating the quality of neural systems. Our findings suggest that such test collections should be

used wit caution when evaluating the quality of neural systems as they may lead to incorrect conclusions regarding how the quality of a neural model compares with a traditional model, as well as how the neural model compares with another baseline neural model.

2 RELATED WORK

A significant amount of research has been devoted to analysing the fairness and reusability of test collections for retrieval evaluation, where fairness refers to collection being unbiased in its evaluation to different runs that contributed to the construction of the pool and reusability refers to the fairness of the test collection towards evaluating the quality of the runs that did not contribute to the construction of the test collection [13].

Zobel et al. [14] argued that test collections constructed using depth-k pooling [11] tend to be reasonably reusable and tend to be fair towards evaluating the quality of new systems. Various methods have been proposed in order to generate fair and reusable test collections with limited relevance labels [4, 5, 9, 10, 13]. Previous work has shown that when test collections are constructed using pools that are too small compared to the document collection size, the resulting pools could exhibit some bias (in particular, bias towards systems that retrieve documents that contain topic title words) [3].

While most of this previous work analysed the reusability of test collections in terms of their fairness towards evaluating new systems that did not contribute to the pool, none of the previous work analysed the reusability of such collections when they are constructed solely using systems that are of particular type (e.g. traditional systems) but are used to evaluate the quality of systems that are of a different type (e.g. neural systems).

The TREC 2017 Common Core Track showed some evidence of neural runs—*e.g.*, [12]—being more likely to uniquely retrieve a relevant document [13] in comparison to traditional runs. If future neural runs, during reuse of the test collection, also had this property of finding previously unseen relevant results, then the evaluation of those new runs would be unfair, since no new judging is done during reuse. Although this indicates a potential problem, no previous work systematically analysed the reusability of test collections generated using traditional models towards evaluating the quality of such neural models.

3 EXPERIMENTAL ANALYSIS

We analysed the quality of test collections constructed using depth pooling from traditional vs. neural systems in terms of the number of relevant document identified, as well as in terms of the reusability of these pools based on the evaluation results obtained for systems of different types (neural vs. traditional systems). For this purpose, we use the data from The TREC Deep Learning Track [6], details of which are described below.

3.1 Task and datasets

The TREC Deep Learning Track has two tasks: Document retrieval and passage retrieval. Both tasks have large training sets based on human relevance assessments, derived from MS MARCO [2]. The test collection used in the track, which was generated using the depth-10 pools of the participating systems, contains 43 queries.

Judgments were done on a four-point scale: Perfectly relevant, highly relevant, relevant and irrelevant. The track reported both NDCG@10 and MRR metrics, with NDCG@10 being the primary metric used in ranking the systems.

In total 10 groups with a total of 38 runs participated in the document retrieval task and 11 groups with a total of 37 runs participated in the passage retrieval task. For the document retrieval task, out of the 38 runs, 27 of them were based on neural models (models based on deep learning methods or use such models (e.g. BERT) as features)) and 11 of them were based on traditional methods (models that are based on traditional, non-neural methods such as BM25). For the passage retrieval task, 26 runs were based on neural models and 11 of them were based on traditional methods.

Top 10 performing systems in both document ranking and passage ranking tasks were based on neural models. More details about the results of the Deep Learning Track can be found at the overview paper of the track [6].
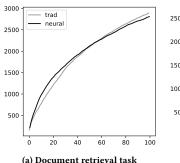
3.2 Experimental Results

3.2.1 Number of Relevant Documents Found. Since the reusability of a test collection highly depends on the number of relevant documents identified, we first analysed the number of relevant documents identified if pools were to be constructed solely using (i) traditional runs vs. (ii) neural runs.

For this purpose, we divided the runs submitted to the Deep Learning Track into two categories: Traditional systems and neural systems. The runs were assigned to the two categories based on the original categorization used by the track, as described in Section 3.1. We then analysed the number of documents identified when test collections are constructed using depth-k pooling by pooling top k results from traditional systems vs. neural systems, for various cutoff values k.

Figure 1 shows the result of this experiment for the document retrieval task (left) and the passage retrieval task (right plot). The x axis in the figures shows the cutoff value k used to constructed the depth-k pools and the y axis shows the number of relevant documents identified using pools constructed via traditional (grey line) vs. neural (black line) models.

It can be seen that for both tasks neural models tend to find more relevant documents at early cutoff levels. For document retrieval task, neural runs seem to be overtaken by the traditional runs as one goes deeper in the ranking whereas for the passage retrieval task neural runs consistently find more relevant results at all cutoffs. Given that most IR metrics tend to be top heavy, these results raise concerns about the reliability of evaluation results in evaluating neural models with pools generated from traditional methods, a commonly faced scenario due to most existing test collections being generated solely using traditional models.

3.2.2 Test Collection Reusability. We then analysed the reusability of test collections generated via pooling top-k results of systems of a particular type for evaluating systems that are of a different type, particularly focusing on traditional vs. neural system types. In particular, we are interested in the question as to whether pools generated using traditional systems can be reliably used to evaluate the quality of neural models, and vice versa.
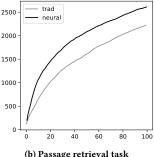
Figure 1: Cumulative count of relevant results at each rank cutoff k. In document retrieval, neural runs find more relevant results at early cutoffs, but then are overtaken by traditional runs at later cutoffs. In passage retrieval, neural runs find more relevant results at all cutoffs.

	MRR			NDCG@10		
Test System:	Trad	Neural	All	Trad	Neural	All
Trad Pool	0.436	-0.12	-0.19	0.772	0.68	0.676
Neural Pool	0.769	0.635	0.842	0.774	0.836	0.852

Table 1: Average Kendall's tau correlations between actual metric and metric values computed using 10 randomly generated traditional (top row) vs. neural pools (bottom row) for document retrieval runs.

In order to evaluate the reusability of test collections generated using traditional pools towards evaluating the quality of neural models, we randomly split the traditional runs submitted to the TREC Deep Learning Track into two sets. We used the first set of systems to construct the test collection using depth-10 pooling (which we refer to as the *traditional pool*), and we used the second set of systems together with the neural models as test systems, using which we analyse the reusability of the pools generated. Depth-10 pools were used in the pooling process since the original test collection for the Deep Learning Track was generated using depth-10 pooling.

In TREC, most groups tend to submit multiple runs and most of these runs tend to be different variants of the same system, which was also the case for the Deep Learning Track (as described in Section 3.1). In order to avoid having a system in the test set that is very similar to a system used in constructing the pools, if one run from one group is randomly selected to be included in the pool, all the remaining runs from that group are also included in the pool.

We then used this test collection to evaluate the quality of test systems (neural systems, as well as the traditional systems that did not contribute to the pool). This way we can evaluate the reusability of the test collection constructed with traditional systems in terms of their fairness towards evaluating (i) the performance of neural systems within themselves, (ii) the performance of other traditional systems that did not contribute to the pool within themselves, and (iii) the relative performance of neural vs. traditional systems.

Since the test collection for the Deep Learning Track was generated using depth-10 pooling, we also use such depth-10 pools evaluation results obtained by pooling all the systems submitted as our gold standard (which we refer to as the actual metric values).

	MRR			NDCG@10		
Test System:	Trad	Neural	All	Trad	Neural	All
Trad Pool	0.63	0.004	0.0	0.789	0.574	0.612
Neural Pool	0.7	0.81	0.875	0.89	0.874	0.881

Table 2: Average Kendall's tau correlations between actual metric and metric values computed using 10 randomly generated traditional (top row) vs. neural pools (bottom row) for passage retrieval runs.

We then compare the actual metric values with metric values computed using the traditional pool (which we refer to as the estimated metric values). We then compute the Kendall's tau correlation between the actual metric values and estimated metric values when (i) only traditional methods are used as the test systems, (ii) only neural models are used as the test systems, and (iii) systems of both types are used as the test systems. Since NDCG@10 and MRR were two of the primary metrics used in context of the Deep Learning Track, we also focused on these metrics as the primary evaluation metrics.

Since the quality of the pools constructed could be highly affected by the type of runs that are randomly selected for constructing the pools, we repeat this process 10 times to construct 10 random pools generated by using different random splits of traditional runs, and compute the average Kendall's tau correlation results over all the 10 runs.

In order to evaluate the reusability of pools generated using neural models, we repeated the same procedure to by randomly selecting half of the neural runs for constructing the pools (which we refer to as neural pools), which are then used to evaluate the quality of neural models that did not contribute to the pool together with traditional models in a similar way as above.

Table 1 and Table 2 show the average Kendall's tau values obtained over the 10 randomly constructed pools when pools are constructed using half of the traditional systems (upper row) and half of the neural systems (bottom row) for the document retrieval and passage retrieval tasks, respectively. The columns in the tables show the different types of test systems used in evaluation. It can be seen that for both document retrieval and passage retrieval tasks, traditional pools result in poor evaluation results for the neural systems. In fact, traditional pools seem to be worse than neural pools even for evaluating the quality of traditional systems that did not contribute to the pool!

Furthermore, in most cases the Kendall's tau correlation when all the test systems are considered seems to be less than the Kendall's tau scores for traditional systems and neural systems alone. This suggests that such pools are performing very poorly when the pairwise comparisons between the traditional vs. neural systems are considered. Hence, when a neural model is compared with a traditional model using a test collection generated via traditional pools, one might incorrectly infer that the neural model is performing worse than the traditional model.

Note these findings could be partially related to the number and type of runs included in the pools. In our experiments, we were limited by the number and type of runs submitted to the Deep Learning Track. If more runs with more variety were used to create the pools, the resulting pools have the potential to result in more reliable evaluations of neural systems. However, the fact that

(a) MRR
Traditional Depth-10 Pool

(b) MRR
Neural Depth-10 Pool

(c) NDCG@10
Traditional Depth-10 Pool

(d) NDCG@10
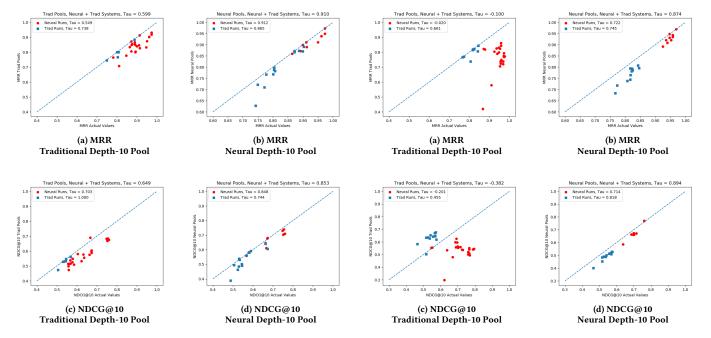**Neural Depth-10 Pool**

Figure 2: MRR (top) and NDCG@10 (bottom) values for document retrieval task, when pools are generated using (left) traditional systems vs. (right) neural systems.

(a) MRR
Traditional Depth-10 Pool

(b) MRR
Neural Depth-10 Pool

(c) NDCG@10
Traditional Depth-10 Pool

(d) NDCG@10
Neural Depth-10 Pool

Figure 3: MRR (top) and NDCG@10 (bottom) values for passage retrieval task, when pools are generated using (left) traditional systems vs. (right) neural systems.

the same exact pools Table 1 Table 2 could lead to very different evaluation results in terms of their reliability when evaluating the quality of traditional vs. neural techniques is highly concerning.

Out results suggest that existing test collections generated using traditional systems should be used with caution when evaluating the quality of neural models as the evaluation results obtained are likely to be unreliable and one might incorrectly infer that the quality of the neural run is worse than a baseline traditional or neural run.

Figure 2 and Figure 3 show how such evaluations look like in detail for a randomly picked pool for the document retrieval and passage retrieval tasks, respectively. The x axis in the plots show the actual metric values when all the systems are used to generate the pools and the y axis shows the estimated metric values computed when half of the traditional (left plots) or neural systems (right plots) are used to generate the pools. The plots also contain line $y = x$ for comparison purposes. The titles in the plots show the Kendall's tau correlation between the actual and the estimated metric values when all systems are considered in the test set. The plots also show the Kendall's tau correlation values within the neural models, as well as within the traditional models in the test set.

It can be seen that pools generated using the traditional pools are particularly unreliable in evaluating neural runs and may have a tendency to underestimate the quality of the neural runs, whereas pools generated via the neural runs tend to be more reliable for evaluating the quality of both traditional and neural systems.

4 CONCLUSION

We analyse the reusability of test collections when they are created solely using systems of a particular type (e.g. traditional systems based on BM25, language modelling, etc.) towards evaluating the quality of systems that are of a different type (e.g. neural systems based on deep learning models or use such models as features).

Our results demonstrate that evaluation results obtained using test collections that are created solely using traditional runs are not very reliable in terms of evaluating the quality of neural systems. In particular, our findings suggest that such test collections should be used with caution when evaluating the quality of neural systems as they may lead to incorrect conclusions regarding how the quality of a neural model compares with a traditional model, as well as how the neural model compares with another baseline neural model.

While our results suggest that pools systematically generated using one system type (e.g. traditional systems) may be problematic when they are used to evaluate the quality of systems of a different type (e.g. neural systems), reusability of test collections generated using neural models towards other (unseen) neural models is a question we would like to analyse further in the future.

REFERENCES
[1] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M Voorhees. 2017. TREC 2017 Common Core Track Overview. In *TREC*.
[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268* (2016).
[3] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the Limits of Pooling for Large Collections. *Inf. Retr.* 10, 6 (Dec. 2007), 491–508. https://doi.org/10.1007/s10791-007-9032-x

[4] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of ACM SIGIR 2006*. 268–275.

[5] Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. 1998. Efficient construction of large test collections. In *Proceedings of ACM SIGIR 1998*. 282–289.

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2019 deep learning track. In *TREC (to appear)*.

[7] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of ACM SIGIR 2017*. 65–74.

[8] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to Match Using Local and Distributed Representations of Text for Web Search. In *Proc. WWW*. 1291–1299.

[9] Mark Sanderson and Hideo Joho. 2004. Forming test collections with no system pooling. In *Proceedings of ACM SIGIR 2004*. 33–40.

[10] Mark Sanderson and Justin Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings of ACM SIGIR 2005*. 162–169.

[11] K Sparck Jones and C Van Rijsbergen. 1975. Report on the Need for and Provision of an Ideal Information Retrieval Test Collection. *Information Retrieval Test Collection* (1975).

[12] Christophe Van Gysel, Dan Li, and Evangelos Kanoulas. 2018. ILPS at TREC 2017 Common Core Track. *arXiv preprint arXiv:1801.10603* (2018).

[13] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections Using Bandit Techniques. In *Proceedings of ACM CIKM 2018*. 407–416.

[14] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of ACM SIGIR 1998*. 307–314.