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

Emine Yilmaz
emine.yilmaz@ucl.ac.uk
University College London, Amazon
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, a major challenge is to ensure the availability of test collections for measuring their quality. Test collections are usually 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 when compared to traditional methods, they may return different types of relevant document that were not identified in the original pooling. If so, test collections constructed using traditional methods could lead to biased and unfair evaluation results for deep learning systems. This paper uses simulated pooling to test the fairness and reusability of test collections, showing that especially when shallow pools (e.g. depth-10 pools) are used, pooling based on traditional systems only may 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

ACM Reference Format:
Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Daniel Campos. 2020. On the Reliability of Test Collections for Evaluating Systems of Different Types. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401317

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401317

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 [4], especially if we exclude results based on proprietary data—e.g., [5]. 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 [3] 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.

While reusability of test collections for evaluating the quality of unseen systems has been widely studied in literature [6, 9], 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 shallow (e.g. depth-10) pools of traditional systems could be less reliable in terms of evaluating the quality of neural systems. 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.

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 a collection being unbiased in its evaluation to different runs than the ones that contributed to the construction of the pool and reusability refers to the fairness of the test collection towards evaluating such runs [9]. Various methods have been proposed in order to generate fair and reusable test collections with limited relevance labels [6, 9].

Zobel et al. [10] argued that test collections constructed using depth-k pooling [7] tend to be reasonably reusable and tend to be fair towards evaluating the quality of new systems. However, 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 [2].

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.*, [8]—being more likely to uniquely retrieve a relevant document [9] 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.

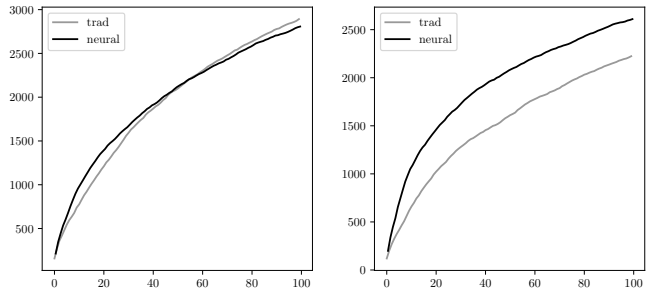
3 EXPERIMENTAL ANALYSIS

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 [3].

3.1 Task and Datasets

The TREC Deep Learning Track 2019 had two tasks: Document retrieval and passage retrieval. Both tasks have large training sets based on human relevance assessments. The test collections used in the track, which were generated using the depth-10 pools of the participating systems, contain 43 queries judged on a four-point scale. 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 in the passage retrieval task. For the document retrieval task, 27 of the runs 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. Top 10 performing systems in both tasks were based on neural models.



(a) Document retrieval task **(b) Passage retrieval task**

Figure 1: Cumulative count of relevant results found via depth-k pooling, for various values of k.

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. We used the same categorization of the runs as the one originally 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 values of 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. This problem could potentially be overcome if very deep pools are used so that pools constructed using traditional models retrieve enough relevant documents (e.g. in the document retrieval task), but there is a chance

	MRR			NDCG@10		
Test System:	Trad	Neural	All	Trad	Neural	All
Trad Pool	0.436	-0.120	-0.190	0.772	0.681	0.676
Neural Pool	0.769	0.635	0.842	0.774	0.836	0.852

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

that this issue could exist even when deeper pools are used (e.g. in the passage retrieval task).

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.

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. Test collections were generating using depth-10 pooling since such pools were used in the Deep Learning Track.

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. 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.

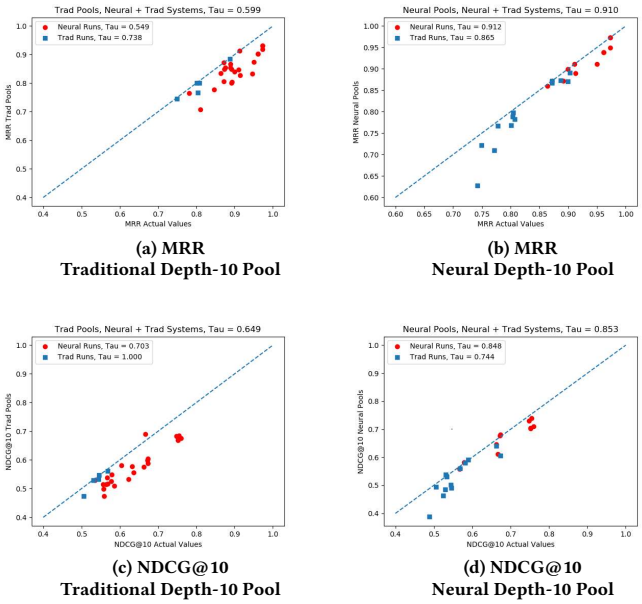
We then used this test collection to evaluate the quality of test systems (neural systems, as well as the traditional systems that did not contribute to the pool). This way we can evaluate the reusability of the test collection constructed with traditional systems in terms of their fairness towards evaluating (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.

We use evaluation results obtained using depth-10 pools of all submitted systems as our gold standard (which we refer to as the actual metric values). We then compare the rankings of systems obtained using actual metric values with rankings obtained using metric values computed using the traditional pool (which we refer to as the estimated metric values) by computing the Kendall's τ correlation between these metrics 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. Similar to the Deep Learning Track, we focus on NDCG@10 and MRR 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, the aforementioned process was repeated 10 times to construct 10 random pools generated by using different random splits of traditional runs and compute the average Kendall's τ correlation across the 10 randomly generated pools.

	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 τ correlations between actual and estimated metric values computed using traditional (top row) vs. neural pools (bottom row) for passage retrieval runs.



(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 (left) traditional (right) neural pools are used.

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 τ values over the 10 randomly generated 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 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 τ correlation when all the test systems are considered seems to be less than the Kendall's τ 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 that 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 the same exact pools in 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.

Our 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 could be unreliable and one might incorrectly infer that the quality of the neural run is worse than a baseline traditional or neural run.

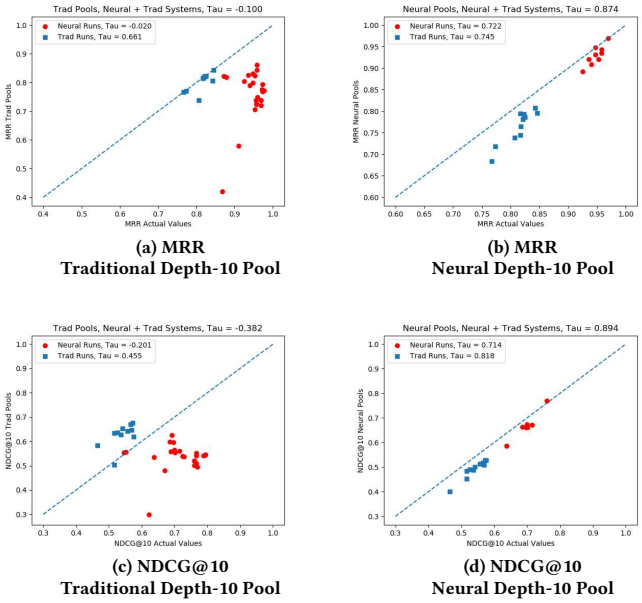
Figures 2 and 3 show how such evaluations look like in detail for a randomly picked pool for the document and passage retrieval tasks, respectively. The x axis in the plots show the actual metric values the y axis shows the estimated metric values computed when half of the traditional (left plots) or 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 τ 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 τ correlation values within the neural models, as well as within the traditional models in the test set. It can be seen that traditional pools could be particularly unreliable in evaluating neural runs and may have a tendency to underestimate the quality of the neural runs, whereas neural pools tend to be more reliable for evaluating the quality of both traditional and neural systems.

4 CONCLUSIONS

We 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 a different type (e.g. neural models based on deep learning or systems that use such models as features).

Our results demonstrate that when test collections are generated using shallow depth pooling (e.g., depth-10 pooling), evaluation results obtained using test collections that are created solely using traditional runs may not be 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.

The results presented in this paper mainly focus on test collections generated via shallow pools, in particular depth-10 pools. One important research questions for the future is to analyse whether the problems presented in this paper would still be valid when deeper pools are used for test collection construction. Due to the limited availability of test collections that contain both neural and traditional runs, in this paper we have been restricted by one collection with a limited number of system. In the future we would like to also analyse test collection reusability when a larger variety of systems are available to construct the pools.



(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, wwhen (left) traditional (right) neural pools are used.

5 ACKNOWLEDGEMENTS

This work was partially funded by the EPSRC grant titled "Task Based Information Retrieval", reference number EP/P024289/1.

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] 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.
[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2019 deep learning track. In *TREC*.
[4] 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.
[5] 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.
[6] Mark Sanderson and Justin Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings of ACM SIGIR 2005*. 162–169.
[7] 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).
[8] Christophe Van Gysel, Dan Li, and Evangelos Kanoulas. 2018. ILPS at TREC 2017 Common Core Track. *arXiv preprint arXiv:1801.10603* (2018).
[9] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections Using Bandit Techniques. In *Proceedings of ACM CIKM 2018*. 407–416.
[10] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of ACM SIGIR 1998*. 307–314.