MS MARCO: Benchmarking Ranking Models in the Large-Data Regime

Nick Craswell
nickcr@microsoft.com
Microsoft
USA

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

Emine Yilmaz
emine.yilmaz@ucl.ac.uk
University College London
United Kingdom

Daniel Campos
dcampos3@illinois.edu
University of Illinois, UC
USA

Jimmy Lin
jimmylin@uwaterloo.ca
University of Waterloo, Microsoft
Canada

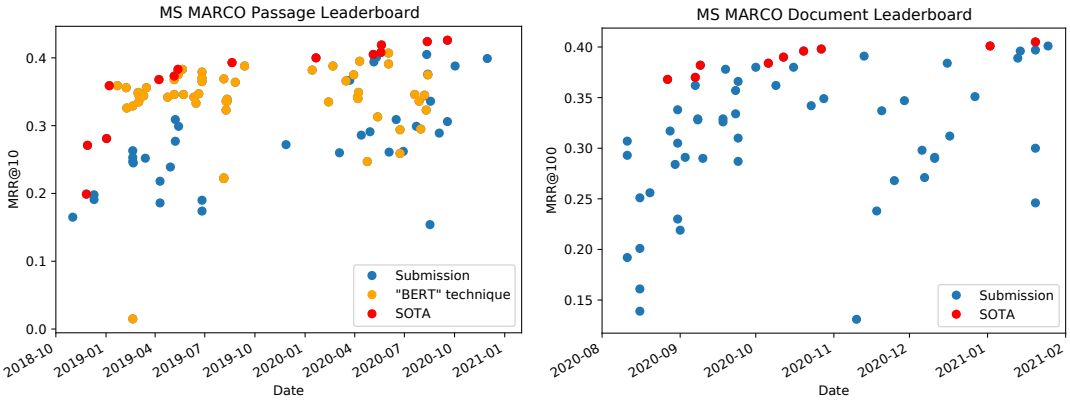

Figure 1: Improvement over time using the MS MARCO passage data (left) and MS MARCO document data (right).

ABSTRACT

Evaluation efforts such as TREC, CLEF, NTCIR and FIRE, alongside public leaderboards such as MS MARCO, are intended to encourage research and track our progress, addressing big questions in our field. However, the goal is not simply to identify which run is "best", achieving the top score. The goal is to move the field forward by developing new robust techniques, that work in many different settings, and are adopted in research and practice. This paper uses the MS MARCO and TREC Deep Learning Track as our case study, comparing it to the case of TREC ad hoc ranking in the 1990s. We show how the design of the evaluation effort can encourage or discourage certain outcomes, raising questions about internal and external validity of results. We provide some analysis of certain pitfalls, and a statement of best practices for avoiding such pitfalls. We summarize the progress of the effort so far, and describe our desired end state of "robust usefulness", along with steps that might be required to get us there.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462804

CCS CONCEPTS

• **Information systems** → **Test collections**; **Retrieval effectiveness**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

IR evaluation; leaderboard; deep learning

ACM Reference Format:
Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462804

1 INTRODUCTION

MS MARCO is a series of datasets, the first of which was released in 2016, aiming to help academic researchers explore information access in the large-data regime [3]. The MS MARCO datasets have been a boon for neural IR researchers to support their exploration

of ever larger and richer models with an insatiable appetite for more (supervised) training data. Over the past few years, the datasets have been used in tasks ranging from keyphrase extraction to question answering to text ranking. Of these tasks, the passage ranking and document ranking tasks have received the most attention from the research community; both are associated with competitive leaderboards[1] and the TREC Deep Learning Track [13–15]. They are standard ad hoc retrieval tasks, with the major difference being the length of the documents that are retrieved: the passage ranking task works with paragraph-length segments of text, while the document ranking task works with full-length web pages.

Figure 1 summarizes both leaderboards, passage on the left and document on the right. The *x*-axes represent time, from the introduction of the leaderboards until early 2021. Each point represents a submission: the *x*-axis plots the date of submission and the *y*-axis plots the official metric (MRR@10 for passage and MRR@100 for document). Circles in red represent the (current and former) state of the art (SOTA) runs, i.e., a top-scoring run on the leaderboard, beginning with the first submission that beat organizer-supplied baselines. On the left panel in Figure 1 for the passage leaderboard, the large jump in the SOTA in January 2019 represents the work of Nogueira and Cho [54], which is the first known application of pretrained transformers to a ranking task. This is considered by many to be a watershed moment in IR, as it ushered in a new era of research dominated by the use of pretrained transformer models. Runs whose description contain the word "BERT" are shown in orange in the left panel. From the multitude of the orange points, we can see the immediate dominance of BERT-based techniques right after its introduction; this is likely even an under-estimate, since there are many ranking models based on pretrained transformer models that do not have BERT in its name (e.g., ELECTRA, T5, etc.). We did not repeat the same coloring in the document leaderboard because, based on our observations, BERT has become so ingrained that its name is nowadays omitted from the model descriptions.

Prior to the advent of the MS MARCO, deep neural methods in IR were largely being benchmarked on proprietary datasets (*e.g.*, [34, 51, 87]), non-English datasets (*e.g.*, [16, 82]), synthetic datasets (*e.g.*, [52, 72]), or under weak supervision settings (*e.g.*, [19, 87]). This made it difficult for the community to compare these emerging methods against each other, as well as against well-tuned traditional IR methods, which led to concerns [41] in the IR community as to whether "real progress" was being made. Subsequently after the release of the MS MARCO dataset, some of these neural methods (*e.g.*, [16, 51]) reproduced their claimed improvements over traditional methods on the public leaderboard. BERT put any remaining concerns to rest, with the initial big jump in effectiveness that we noted, followed by continued upward progress in SOTA, in both leaderboards. The effectiveness of BERT was widely reproduced and shown to be a robust finding, leading Lin [42] to later retract their criticisms.

The MS MARCO datasets have been instrumental in driving this progress because it enabled all researchers (not only those in industry) to examine neural techniques in the large-data regime. The impact of data is shown in Figure 2, taken from Nogueira et al. [56]. The figure shows the effectiveness of BERT-base as a reranker trained with different numbers of training instances (note the log scale in the *x*-axis). Results report means and 95% confidence intervals over five trials. As expected, the larger the data, the better the effectiveness. As pointed out by some researchers [44], to a large extent, the rapid progress made in the IR community would not have been possible without MS MARCO.

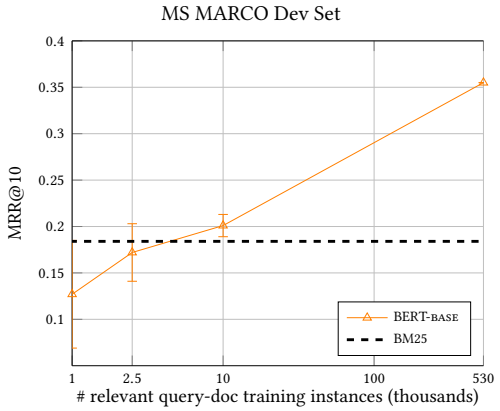

Figure 2: Effectiveness of BERT-base increases as we add more training instances. Showing mean and 95% confidence intervals over five trials. Taken from Nogueira et al. [56].

So what is the state of the field at present? We can summarize as follows: (1) the MS MARCO datasets have enabled large-data exploration of neural models, and (2) from the leaderboards, it appears that progress continues unabated.

But is the "SOTA" progress meaningful? Is MRR a good metric? Are all the top runs tied, with an exhausted leaderboard? Have we seen multiple submission and overfitting? If we change the test data slightly, as a test of external validity, do our findings hold up? Are these easy to deply, with a standard playbook? We describe what is required to make more progress, towards having evaluation with internal validity, external validity and robust usefulness.

2 REQUIREMENTS TO ADVANCE THE STATE OF THE ART

This section outlines some steps that are required to make a valid and useful contribution to the state of the art in ad hoc ranking. Valid because we are sure it is an improvement. Useful because the improvement is easy to deploy in many different real-world applications. We first describe an older improvement, where significantly better rankers such as BM25 were developed using TREC data in the 1990s. We then consider the same criteria for BERT-style rankers using the MS MARCO and TREC Deep Learning Track data. This is a checkpoint on our progress so far, it motivates some of our analysis in this paper and identifies important future work.

2.1 BM25 and TREC data

New data can move the field forward. For example, TREC [76] introduced test collections starting in 1991 led to a new generation of ranking functions. The test collections did not have a large set of training queries, encouraging the development of ranking functions

[1] http://msmarco.org

that work well in a small training data regime. The number of query topics used in each evaluation was 50. Compared to previous evaluation efforts, TREC documents were longer, and they varied in length, writing style, level of editing and vocabulary [32]. By the third year of the effort, this led to the development of new ranking functions that dealt with variation in document length significantly better than previous ranking functions, including Okapi BM25 [61].

Today the "Okapi at TREC-3" paper has 2,420 citations in Google Scholar and searching for that string seems to mostly give papers about information retrieval (checking a few) with an estimated 15,700 results. BM25 was developed just before the appearance of the first Web search engines, but was found to work well on Web documents and was also commonly used in learning to rank data sets, many of which used web data [45]. Papers might use BM25 features in a learning to rank data set without mentioning BM25, but it still had impact. Many real-world information retrieval systems implement BM25 and it has most likely been evaluated on many proprietary data sets, not just with TREC-style evaluation, but also with online tests such as interleaving and A/B tests [33].

Internal validity. With each study, there is a risk that the conclusions we draw are not reliable. Here we focus on statistical and mathematical correctness of the study [7]. A study can be under powered, meaning that we can not draw finer-grained conclusions. For example, we can identify that BM25 is significantly better than a plain tf-idf implementation, but it may be statistically indistinguishable from other modern BM25-like functions.

Multiple testing and selective publication can harm the internal validity of our studies [15]. Statistical significance tests tell us how likely our findings may hold up on a new sample of data from the same distribution. However, if we run multiple tests on the same sample, and selectively report the best outcomes on that sample (and the bad outcomes may be rejected even if reported in a paper), the chances of that result holding on a new sample are reduced.

The best practice for avoiding multiple testing is to avoid reuse of the data, such as an online A/B test, where each new test is on live data, without reuse. Submitting to evaluation efforts such as TREC also avoids reuse, since each year generates a new set of single-shot submissions on a new set of queries. Public leaderboards are a bit worse, since they allow multiple submission to the same dataset. We will discuss methods of reducing the harm and we will analyze the extent of the problem in MS MARCO leaderboards. The most harmful case for multiple testing is with reusable test collections, which allow unlimited iteration on a test set with no public registration of what was done. There have been some claims that the field has a problem with this kind of validity [2] although that paper was not questioning that BM25 was an improvement, but rather questioning whether subsequent studies improved on methods such as BM25 from the 1990s.

If IR metrics are not on an interval scale, as was argued recently by Ferrante et al. [23], Fuhr [29], this is also an internal validity problem. If commonly-used metrics such as Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) are not on an interval scale, then reporting the mean of the metric and doing a statistical test on difference of means is not valid. Many forms of evaluation used on BM25 did calculate mean metrics with t-tests. However, a model that has been very widely deployed such as BM25 has also been tested in online interleaving and A/B tests with large numbers of users, which may not have the same problems. There is also evidence that sufficiently powered online experiments of this sort can agree with a TREC-style NDCG metric [57].

External validity. A study could be internally valid, with statistical tests that indicate how well the results will hold up on a new identically-distributed sample of data, but still lack external validity. Here we focus on slight changes in the data distribution, such as moving to a slightly different document distribution, query distribution or relevance judging scheme. Zobel and Moffat [94] evaluated many BM25-style rankers on six different data distributions (which they called domains), coming from two different document collections, each with title, narrative or full queries. Their finding was that there was no clear best method, noting "success in one domain was a poor predictor for success in another".

The TREC finding from the 1990s, that BM25-like rankers improved on pre-TREC rankers, has good evidence of external validity. BM25 has been tested on many datasets in industry and academia, on public and private datasets, with TREC-style evaluation and presumably with online metrics. It has been selected as a powerful feature many times, by many different machine learned rankers, on many different data distributions. It would be incorrect to say that the improved performance identified in TREC-3 only held up on TREC-3 data or datasets with identical distributions.

Robust usefulness. BM25 is not only valuable in many different settings, but it is useful, robust and easy to deploy. It has a small number of free parameters, but if these are not tuned then the performance is still good. BM25 can be included in an IR system without needing extra training data, without needing a PhD (or PhD student) to carry out finetuning. The chances of BM25 giving very bad results in a new setting are low.

2.2 BERT-style rankers and MS MARCO data

In the case of MS MARCO, the main difference from TREC data is the presence of large training data, with hundreds of thousands of training queries. This encourages the development of rankers that can work well in the large-data regime, such as BERT-style rankers. Have these rankers been evaluated with internal and external validity, in a way that is robustly useful when deployed? Let us assess how far we are from this goal.

Internal validity. Multiple testing is a problem in our field, we discourage multiple submission in several ways. We have experiments in the TREC Deep Learning Track, where there is a single-shot submission each year, which is the gold standard for avoiding data reuse. We then retire the data as a reusable test collection, which is the worst case here, allowing unlimited testing iterations. Tests that do not show a gain may not be written up as papers and/or may not be accepted. We also have a leaderboard, which allows multiple submission, but with limits. First, we limit how frequently each group submits. Second, every submission is public, so we can see which groups seem to be p-hacking and slowly overfitting to the test data through multiple submission. Third, with each submission we have a small number of queries that are not used for the leaderboard metric. We will analyze the extent to which the evaluation on these held out queries diverges from the evaluation using the

queries in the leaderboard, which could happen if participants are iterating on their submissions and using the numbers on the public leaderboard as their guide.

The other threat to internal validity is whether we can find repeatable and valid differences between leaderboard runs. Perhaps the top runs are all statistically indistinguishable, and after a while we should stop the evaluation. Perhaps due to the questions in the field about the interval scale, we shouldn't be using the mean and t-test approach that many papers in the field use. We will analyze the reliability of our leaderboard under different statistical tests and also use bootstrapping to analyze its reliability.

External validity. Eventually, if BERT-style rankers are widely adopted, they will be evaluated in many different settings using many different metrics. However, when we first saw good leaderboard results from ML-heavy approaches, we were suspicious that the improvements would only hold true if the training and test data were independent identically distributed (IID) samples from the same distribution. For example, there could be quirks of the MS MARCO sparse labeling that pretrained transformer models can learn, giving good performance on MS MARCO sparse labels in the test set, but the improvements would vanish if we relabeled the data with slightly different judging scheme. In that case, the results would be specific to the setup of our study, lacking external validity. We could only claim a real improvement if we think real users have exactly the same quirks as the MS MARCO labels.

We test this two ways. Firstly, we set up the TREC experiment with a slight data mismatch between the train and test data. Specifically, NIST judging selects queries that have the right level of difficulty (not too easy nor too hard) and instead of roughly one positive result per query as in MS MARCO, TREC judges label many documents per query on a 4-point relevance scale. In the DL track, we found that training on the sparse labels does allow a big improvement on the test set, despite the slightly different data distributions [13, 14]. Second, in the document ranking leaderboard of MS MARCO we included some queries that are not used in the public leaderboard. This allows us to do a private leaderboard analysis, in this case on the 45 TREC 2020 queries, using NIST labels (as well as the sparse labels on the same queries). These are small steps to ensure that the BERT-style rankers will perform well in many applications, these get confirmed over time in industry and academia with tests on many proprietary and public datasets.

Robust usefulness. We survey all the different ways people are using BERT-style rankers. We discuss our concerns about whether we really established a playbook yet, making it easy for a non-PhD to deploy this kind of ranker in a new application in a way that truly works better than previous rankers.

3 MS MARCO LEADERBOARD VALIDITY ANALYSIS

To test the validity of our leaderboard, we first analyze its ability to distinguish different runs using a variety of parametric and non-parametric statistical tests. We also use bootstrapping to analyze the leaderboard stability, which in some cases can indicate that the top-ranked result was lucky (Table 4 of [8]). We also use bootstrapping analysis to test external validity, using our private leaderboard.

Table 1: Passage ranking leaderboard bootstrap analysis.

	Rank under bootstrapping				
Leaderboard run	1	2	3	4	5
1st	72.7%	25.4%	1.9%	0%	0%
2nd	24.2%	62.5%	13.3%	0%	0%
3rd	3.1%	12.1%	83.9%	0.8%	0.1%
4th	0%	0%	0.6%	47.0%	27.1%
5th	0%	0%	0.2%	34.5%	34.0%

Table 2: Document ranking leaderboard bootstrap analysis.

	Rank under bootstrapping				
Leaderboard run	1	2	3	4	5
1st	91.2%	7.4%	1.4%	0%	0%
2nd	6.8%	61.7%	21.1%	8.6%	1.4%
3rd	1.6%	22.7%	36.8%	20.2%	12.2%
4th	0.4%	5.4%	17.7%	27.0%	25.1%
5th	0%	0.5%	15.9%	21.2%	22.9%

These 45 TREC-2020 queries were part of every submitted run, but did not contribute to the public leaderboard numbers. We analyze whether the leaderboard conclusions generalize to these held-out queries, using sparse MS MARCO labels and also using TREC labels. Finally, since we are concerned about multiple submission, we analyze the leaderboard with respect to multiple submissions from the same group, to see if they seem to be benefitting from these submissions and whether their movement on the private leaderboard is different from that on the public leaderboard.

3.1 Public leaderboard stability

We analyze overall leaderboard stability using bootstrapping, similar to previous work by Caruana et al. [8], which avoids running many pairwise statistical tests. For each leaderboard we run 1000 bootstrapping trials, comparing the top-ranked runs, the most recent runs and baseline runs. Each bootstrapping trial samples a queryset of the same size as the original queryset, with replacement.

Our first question is whether the leaderboard's top ranks are stable under bootstrapping. If we saw that many top runs had a similar chance of being top-ranked, we might conclude that the leaderboard is exhausted. It is even possible to find that the top run on the leaderboard was lucky, and some other run has more appearances at the top under bootstrapping [8]. Tables 1 and 2 show the top-5 stability of the passage and document leaderboards, respectively. It is not the case that the top ranks are all tied. The 1st ranked run on each leaderboard never drops below position 3 in any of the 1000 trials. The tables show some indication that lower ranked results are less certain, for example the 5th ranked run has less than 50% chance of being in position five. This can happen when two runs are similar. In the document leaderboard, the 5th and 6th run have expected ranks of 5.1 and 5.4 over 1000 trials.

The overall stability of the document leaderboard under bootstrapping is shown in Figure 3. There are some runs with similar

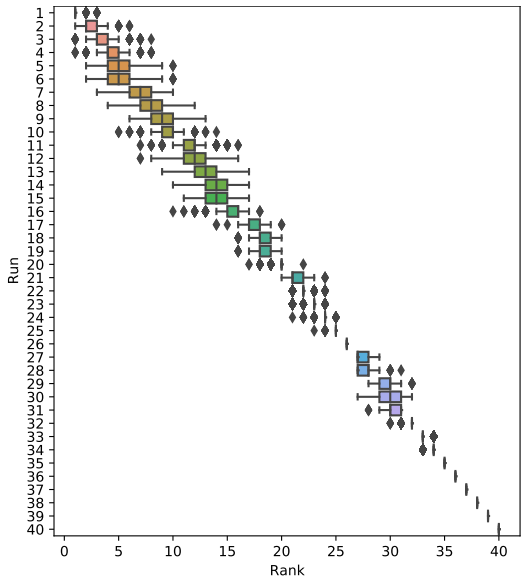

Figure 3: Full results of document leaderboard bootstrap. Runs 1–5 show the same results as Table 2.

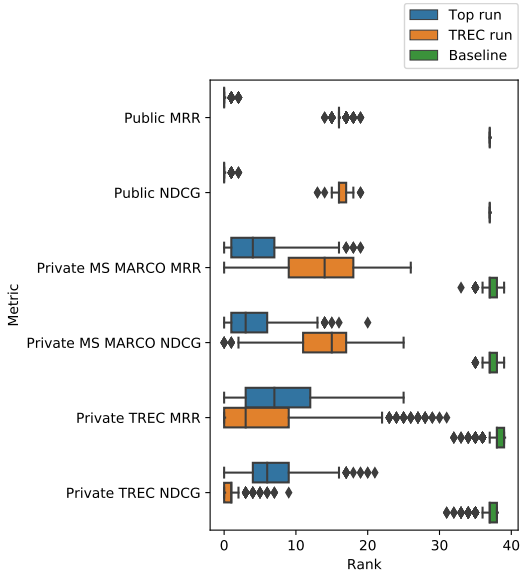

Figure 4: Rank positions of three leaderboard runs under bootstrapping. Metrics are MRR and NDCG@10. The query-sets are the 5,793 Public leaderboard queries and the 45 Private leaderboard queries from TREC-2020. The Private queries can be evaluated with sparse MS MARCO labels or comprehensive TREC labels.

performance lower down in the ranking, having very similar rank distributions. The official baseline at 38th position was ranked 38th in all 1000 bootstrapping trials. Overall it was very unlikely under bootstrapping that a lower-ranked run would overtake a top-ranked run, leading us to conclude that the leaderboard is quite stable. The top-ranked run is not there by luck.

3.2 Private leaderboard

It is possible for a leaderboard to be stable, as in our bootstrapping analysis, but still have overfitting due to multiple submission. One way of detecting this is to have a private leaderboard, where each of the submissions can be tested on a held-out dataset. If participants are using the public leaderboard to overfit to the test queries, we would see their performance increase on the public query set, and decrease on the private query set.

To allow this sort of testing, we included some additional queries that were run by every participant in the document leaderboard. Here we use the 45 TREC 2020 queries for our analysis. Including our earlier bootstrapping on the Public leaderboard, we now have full bootstrapping analysis with 1000 trials on six alternatives: Metric is MRR or NDCG@10, query set is Public or Private, and relevance labels on the Private queries are sparse MS MARCO labels or comprehensive TREC labels.

Instead of showing the full bootstrapping results for all six combinations, we summarize three key runs and their performance under bootstrapping. Figure 4 shows this analysis. The top run on the public leaderboard has its ranks more spread out on the smaller Private leaderboard queries, since there are 45 rather than 5793 queries. It is overtaken by other runs not only on certain bootstrap trials, but other runs even have a better expected rank. For the MS MARCO labels, the top leaderboard run is ranked fourth in expectation for MRR and third in expectation for NDCG@10. For the TREC labels, the top run is fifth and sixth.

To explain why we saw greater rank for the TREC labels, we note that some runs submitted to the leaderboard were also submitted to TREC and may have used the TREC 2019 labels for training. The TREC run we highlight in the figure is the ranker from University of Waterloo, that achieved the best NDCG@10 at TREC. This could be seen as a lack of external validity, that the top run on MS MARCO labels is not as highly-ranked on TREC labels. It could also be seen as an indication that extra adaptation and training for the target domain is useful. Overall the official baseline run is significantly worse than our top run and TREC run, under all six conditions.

3.3 Multiple submission

To avoid overfitting to our eval queries, the MS MARCO leaderboards have rules limiting multiple submission. We allow each participating group to submit no more than two runs per month, and no more than one run with very small changes such as hyperparameter tuning or random seeds. This makes it more difficult for participants to try minor variations until they get lucky with a higher leaderboard submission. We also track all the submissions, so if a group is submitting many runs we can analyze what they submitted and how often. We believe this makes it much more difficult for participants to overfit, compared to a reusable test collection which allows unlimited iteration with no public record.

We already presented detailed bootstrapping results for the document leaderboard. Now, we consider the institution that submitted the top run, and whether there is a risk of overfitting. On the document leaderboard, grouped by institution, the three institutions

with the most submissions had 12, 11 and 7 submissions. However, the top run came from an institution with only two submissions. Also, on Private MS MARCO evaluation (Figure 4) the top run is still in the top five runs in expectation. If it were only there by overfitting, we do not think it would be in the top five of forty.

There is still a risk of cross-group overfitting, as groups learn about "what works" some of what they learn may be about this sample of test queries. For example, different groups can share code and ideas. They can converge to a common solution, with many groups submitting slight variations, and one group can get lucky. Through this process of code sharing they can also form an ensemble of promising approaches, and since the promising approaches were selected using the evaluation set, this is another form of overfitting. In future we will use the MS MARCO judgments on the private leaderboard to monitor for overfitting.

4 IR METRICS AND THE INTERVAL SCALE

Commonly used information retrieval metrics, including the ones we employed in our leaderboard evaluation such as NDCG [36] and MRR [12], have recently been criticized by Ferrante et al. [24, 25] for not being interval-scale, which would imply that computing their mean values across different queries is not meaningful. Instead, they argue that most IR metrics tend to be in ordinal scale, implying that we should be using the median metric value as opposed to the mean when we aggregate these values across different queries. This ignited a debate in the IR community with Fuhr [28] arguing that it is, therefore, not meaningful to compute the mean of MRR and ERR metrics over multiple relevance topics. Sakai [65] subsequently disagreed citing that this line of reasoning would render many other IR metrics inappropriate and that many of these metrics are practically useful, even if not theoretically justified. More recently, Ferrante et al. [23] have furthered the argument made by Fuhr [28] to point out that indeed for many well-known and commonly used IR metrics it is inappropriate to compute their average.

Because the MS MARCO labels are binary and sparse, we chose to report MRR as our primary metric on the leaderboard. Similarly at TREC, the Deep Learning track has focused on NDCG, NCG [62], MRR, and MAP [93]. So, the validity of these metrics is an important consideration in the context of benchmarking on MS MARCO. Our position in this paper is that Ferrante et al. [23] have raised a valid issue and indeed there is no reason to assume that metrics like MRR, MAP, and NDCG are on an interval scale. However, we do not fully agree with their theoretical argument and recommendations, and present an alternative viewpoint here.

4.1 Preliminaries

The theoretical argument presented by Ferrante et al. [23] is grounded in the *representational theory of measurement* [38] which views measurement as the process of mapping real world entities to numbers such that some entity attributes are represented faithfully as numerical properties. Before we analyze their argument, we define a few preliminary concepts and notations for our reader. We adopt the same notation as Ferrante et al. [23] here for consistency.

Definition 1 (Relational structure)**.** A **relational structure** is an ordered pair $\mathbf{A} = \langle A, R_A \rangle$ where A is a domain set and R_A is a set of relations on A. If the A is a set of entities then we refer to it as an **empirical relational structure**. In contrast, in case of **numerical** or **symbolic relational structure** A is a set of numbers.

Definition 2 (Homomorphism)**.** Given two relational structures $\mathbf{A}_1$ and $\mathbf{A}_2$, a homomorphism $\mathbf{M} : \mathbf{A}_1 \rightarrow \mathbf{A}_2$ is a mapping $\mathbf{M} : \langle M, M_R \rangle$ such that,

i. The function M maps A_1 to $M(A_1) \subseteq A_2$
ii. The function M_R maps R_{A_1} to $M(R_{A_1}) \subseteq R_{A_2}$, such that $\forall r \in R_{A_1}$, r and $M(R_{A_1})$ have the same arity
iii. $\forall r \in R_{A_1}$, if the relation r holds between some elements from the domain set A_1 then the image relation $M(R_{A_1})$ should also hold for the corresponding image elements in A_2.

Note that we use homomorphism instead of isomorphism because M is typically not a one-to-one mapping.

Definition 3 (Measurement)**.** A **measurement (scale)** is the homomorphism $\mathbf{M} : \mathbf{E} \rightarrow \mathbf{N}$ that maps from the empricial relation structure E to the numerical relational structure N. The mapping of an element $e \in E$ to a number $n \in N$ is called a **measure**.

Definition 4 (Difference structure)**.** An empirical relational structure $\mathbf{E} = \langle E, \preceq \rangle$ is a **difference structure** if $\forall a, b \in E$ it defines a **difference** Δ_{ab} and satisfies the following axioms:

i. $\preceq$ is a weak order—*i.e.*, $\preceq$ is a binary relation on $E \times E$ such that $\forall a, b, c \in E$ it satisfies: (a) $a \preceq b$ or $b \preceq a$, and (b) $a \preceq b$ and $b \preceq c \implies a \preceq c$.
ii. $\forall a, b, c, d \in E$, $\Delta_{ab} \preceq \Delta_{cd} \implies \Delta_{dc} \preceq \Delta_{ba}$
iii. $\forall a_1, b_1, c_1, a_2, b_2, c_2 \in E, \Delta_{a_1 b_1} \preceq \Delta_{a_2 b_2}$ and $\Delta_{b_1 c_1} \preceq \Delta_{b_2 c_2} \implies \Delta_{a_1 c_1} \preceq \Delta_{a_2 c_2}$
iv. $\forall a, b, c, d \in E$, if $\Delta_{aa} \preceq \Delta_{cd} \preceq \Delta_{ab}$, then there exists $x, y \in E$ such that $\Delta_{ax} \sim \Delta_{cd} \sim \Delta_{yb}$ (Solvability Condition)

According to the *representation theorem* for difference structures, if there is a difference structure on the empirical set E then there must exist an interval scale M.

4.2 Analysis of argument by Ferrante et al.

Having covered the preliminaries, we now take a closer look at the argument that Ferrante et al. [23] make in their work. They define a domain set over search result page (SERP) states, where each SERP state is a unique rank-ordered list of relevance grades. For example, under this notation a SERP with three documents with a relevant document at rank two and nonrelevant documents at rank one and three corresponds to a SERP state denoted by the tuple $(0, 1, 0)$. For example, if we consider the universe of all SERPs with exactly three documents and binary relevance grades, then the domain set E over all possible SERP states is $S = \{(1, 1, 1), (1, 1, 0), (1, 0, 1), (1, 0, 0), (0, 1, 1), (0, 1, 0), (0, 0, 1), (0, 0, 0)\}$. Ferrante et al. [23] argue that if we can define a difference structure over $\mathbf{S} = \langle S, \preceq \rangle$ then it would imply the existence of a corresponding interval scale. However, for $\mathbf{S}$ to satisfy the Solvability Condition requires the metric to be equi-spaced between any two neighboring items in a partial ordering of the domain set S. For example, Table 3 shows that MRR can take four discrete values $[1.00, 0.50, 0.33, 0.00]$ in context of the same example scenario with SERPs of fixed length three and binary relevance grades. If we consider the four specific SERP states labeled A–D, we observe that the Solvability Condition is violated because the presence of $\Delta_{CD} = 0.17$ implies there should exist $X, Y \in S$,

Table 3: A tabular representation of the domain set S of all SERPs with exactly three results with binary relevance grades and corresponding MRR values.

	A				C		D	B
$s \in S$	(1,1,1)	(1,1,0)	(1,0,1)	(1,0,0)	(0,1,1)	(0,1,0)	(0,0,1)	(0,0,0)
RR	1.00	1.00	1.00	1.00	0.50	0.50	0.33	0.00

such that $\Delta_{AX} = \Delta_{YB} = \Delta_{CD} = 0.17$. The key argument that Ferrante et al. [23] make is that MRR is not interval-scale because values of 0.17 and 0.83 are not realizable under this setting.

However, our position is that relevance metrics like MRR and NDCG are fundamentally not measurements over SERP states, but instead they measure user perceived relevance of the SERPs. Hence, the difference structure should not be applied on the domain set S of all possible SERP states, but instead on the domain set U of all possible user-perceived relevance states. We argue that an appropriate IR metric should be equi-spaced relative to user perception of relevance such that a change in 0.1 in the metric at any point on the scale (*e.g.*, $0.3 \rightarrow 0.4$ *vs.* $0.75 \rightarrow 0.85$) should correspond to same difference in user-perceived relevance. In other words, it is irrelevant if a three-document SERP cannot realize a MRR value of 0.17 as long as we believe that there exists some user-perceived relevance state that corresponds to that value of the metric.

Now, there is no reason to believe that IR metrics without further calibration would be equi-spaced on the scale of user-perceived relevance. We therefore agree with Ferrante et al. [23] that computing mean of many IR metrics may be inappropriate. But the difference between their argument and ours points to different recommendations for addressing these concerns. To remedy the situation, Ferrante et al. [23] propose ranked versions of common IR metrics that are equi-spaced over E. While the mean value of these ranked-metrics may be more meaningful from the viewpoint of *representational theory of measurement* [38], it is possible, if not likely, that it reduces the correspondence of the metric to user-perceived relevance. By our criteria, the better approach is to conduct lab studies and online studies with real users, to understand their preferences and how this reveals their underlying notion of utility, so we can develop metrics that are on an interval scale in user value.

4.3 Reliability of statistical tests

In this section, we analyse the effect of IR metrics not being interval-scale on evaluation outcomes in practice. Apart from the mean not being very meaningful when aggregating metrics that are not in interval scale across different queries, Ferrante et al. [23] have also raised concerns about the reliability of using some of the commonly used significance tests such as t-test or the Wilcoxon Signed Rank, which require that the values are in interval scale. They have instead argued that sign test or the Wilcoxon Rank Sum test should be used with ordinal measurements, such as most top heavy IR metrics.

Aforementioned issues raised by Ferrante et al. [23] could also raise questions regarding the reliability of the evaluation results obtained through leaderboards like MS MARCO. Previous work [35, 67] has indicated that violation of certain assumptions by some significance tests, in particular, the normality assumption for the t-test, do not have a big effect on the conclusions reached using such tests in practice. This raises the question as to how much the metric not being in an interval scale could affect the reliability of evaluation results obtained using different significance tests, or using different aggregation methods (i.e., mean vs. median). To answer this question, we adopt a similar method as the one used by Buckley and Voorhees [6] for evaluating evaluation stability.

Table 4: Agreement rates for different significance tests across 100 different query set splits for different task and metric combinations.

(a) document ranking using MRR

	Sign T.	WX RS	WX SR	t-test	Sign T. (Med.)	WX RS (Med.)	WX SR (Med.)
agree	93.3%	92.1%	91.2%	91.7%	80.2%	81.2%	80.7%
part. agree	3%	3%	3%	3%	16.1%	16.1%	16.1%
disagree	3.7%	4.9%	5.8%	5.3%	3.7%	2.7%	3.2%
perc. signif.	95.0%	79.7%	92.7%	79.8%	95.0%	79.7%	92.7%

(b) passage ranking using MRR

	Sign T.	WX RS	WX SR	t-test	Sign T. (Med.)	WX RS (Med.)	WX SR (Med.)
agree	92.8%	91.6%	90.8%	90.8%	80.6%	81.4%	81.2%
part. agree	3.3%	3.3%	3.3%	3.3%	15.5%	15.5%	15.5%
disagree	3.9%	5.1%	5.9%	5.9%	3.9%	3.1%	3.3%
perc. signif.	94.6%	79.8%	92.8%	79.9%	94.6%	79.8%	92.8%

(c) document ranking using NDCG

	Sign T.	WX RS	WX SR	t-test	Sign T. (Med.)	WX RS (Med.)	WX SR (Med.)
agree	94.9%	92.2%	91.0%	91.8%	85.2%	84.2%	83.3%
part. agree	1.8%	7.7 %	8.4 %	8.1%	1.8%	10.4%	6.2%
disagree	3.2%	0.1%	0.6%	0.1%	13.0	5.4%	10.5%
perc. signif.	97.9%	81.1%	93.3%	81.4%	97.9%	81.1%	93.3%

We divide our query set into two random subsets and for each pair of systems submitted to the leaderboard, we evaluate as to whether the conclusions reached based on the evaluation results obtained using the two subsets agree with each other. We repeat this process 100 times, generating 100 random splits and compute the agreement rates across the different subsets. If the evaluation results are reliable, we expect the results to be robust to the changes in the query sample and hence, the agreement rates should be high.

When we compare the evaluation results across the two subsets, we use the following definition for agreement, partial agreement, and disagreement. Evaluation results in the two subsets:

- **Agree** with each other if the two subsets agree as to which system is better, and the difference is: (i) statistically significant according to both subsets, or (ii) not significantly different according to both subsets,
- **Partially agree** with each other if (i) the two subsets agree as to which system is better, and the difference is significant according the one subset but not significant according to the other, or (ii) the two subsets disagree as to which system is better, but the difference is not statistically significant according to both sides,

- **Disagree** with each other if the two subsets disagree as to which system is better, and the difference is: (i) statistically significant according to both subsets, or (ii) statistically significant according the one subset but not significant according to the other

We report the results of this experiment using MRR as the metric for the document and passage ranking tasks in Table 4a and 4b, respectively. Each column in the tables shows the agreement rates obtained when a different significance test is used in evaluation, focusing on the sign test (Sign T), Wilcoxon Rank-Sum test (WX RS), Wilcoxon Signed-Rank test (WX SR), and t-test as the significance tests. Since previous work has argued for using the median instead of the mean when the metrics are in ordinal scale, we also report the agreement rates for the significance tests that do not have the interval scale requirement (sign test and Wilcoxon Rank-Sum test), when median is used to compute the aggregate performance across different queries. While the Wilcoxon Signed-Rank test does require the metrics to be interval-scale Ferrante et al. [23], since the null hypothesis for this test is that the median (as opposed to the mean) of the differences is zero, we also report the results for this test when median is used for aggregation. The last three columns in the tables show the agreement rates when median (Med.) is used for aggregation instead of the mean. As seen in the tables, when mean is used for aggregation, the agreement rates for all four significance tests are above 90%. This is true both for significance tests that require interval measurements (t-test and Wilcoxon Signed-Rank test) and also for those that can be used with ordinal measurements (Sign Test and Wilcoxon Rank-Sum test).

One potential reason for the high agreement rates could be caused by a test not being very powerful and hence mostly predicting differences as not statistically significant. Hence, we also report the fraction of pairs of systems that were deemed as significantly different by at least one of the two split sets using a particular significance test, which is reported in the last row of the tables. It can be seen that the agreement rates are not really correlated with the percentage of pairs a test identifies as significantly different.

When median is used instead of the mean, agreement rates drop significantly and consistently across the three significance tests. Our results suggest that, even though the most commonly used IR metrics are not on interval scale, reliability of evaluation results obtained are not widely affected by this. In fact, unlike what was recommended before, using mean instead of the median seems to result in more reliable evaluation results, possibly caused by mean being a more discriminatory statistic than the median. Our results seem consistent across both document and passage ranking tasks.

In Table 4c we show the results for the document ranking task when NDCG@10 is used as the evaluation metric. As expected, NDCG@10 results in higher agreement rates consistently for all significance tests when compared to MRR, even though the difference is not very big. Similar results were observed for the passage ranking task. Overall, our results suggest that even though most commonly used IR metrics such as MRR and NDCG@10 may not be in interval scale, evaluation results obtained in practice seem not to be highly affected and results obtained using benchmarks such as MS MARCO seem to be mostly reliable.

5 ON TRANSFER LEARNING FROM MS MARCO TO OTHER IR BENCHMARKS

The primary motivation behind curating the MS MARCO ranking datasets was to answer the question "How much better can our IR systems be if we had access to millions of positively labeled query-document pairs?" It is exciting to witness the large jumps in performance metrics on this benchmark from the development of new ranking models that can adequately leverage the provided large training datasets. However, if the benefits of MS MARCO's large training data is limited to its own test sets and access to domain-specific large training datasets is only limited to large for-profit private institutions—*e.g.*, major commercial search engines—then the creation of such benchmarks only serves to outsource research and development of models to the academic community that ironically the academic community then cannot operationalize for their own scenarios. To avoid this undesirable dynamic, it is important to also study whether the large training dataset from MS MARCO can bring about meaningful improvements from transfer learning to other IR benchmarks and tasks.

As noted earlier, a successful application of transfer learning from the MS MARCO dataset has been for the TREC Deep Learning track. An initial test set of 200 queries is sampled from the MS MARCO distribution, but then NIST selects a subset of queries to judge which are neither too difficult nor too easy, then apply a 4-point labeling scheme to results pooled from submitted runs. As Craswell et al. [13, 14] have reported, several pretraining-based deep models finetuned on the MS MARCO training data achieve significant improvements over traditional IR methods in this setting.

Transfer learning from MS MARCO to other ad hoc retrieval benchmarks have also been attempted with promising early success. Yilmaz et al. [85] finetune a BERT-based [21] model on MS MARCO, TREC CAR [22] and TREC Microblog [43] datasets and evaluate them on three TREC newswire collections: Robust04 [74], Core17 [1], and Core18. They find that finetuning on MS MARCO alone achieves mixed results on these benchmarks, but finetuning on MS MARCO followed by further finetuning on the TREC Microblog dataset achieves state-of-the-art performance on all three test sets. Since then, the combination of finetuning on MS MARCO followed by on TREC Microblog dataset has also achieved state-of-the-art results on the English subtask of the NTCIR15 WWW-3 task [66, 68]. Recently, Nogueira et al. [55] adapted T5 [58], a pretrained sequence-to-sequence model, by finetuning only on MS MARCO to significantly improve over the previous state-of-the-art results reported by Yilmaz et al. [85] on Robust04. Similar strategies of finetuning on MS MARCO and evaluating on Robust04, GOV2 [10], and ClueWeb [11] have been employed in other recent studies [30, 37, 40, 91, 92], sometimes in combination with weak supervision [71, 90]. Additionally, Ma et al. [47] have employed the document collection in MS MARCO for pretraining before evaluating on these other standard IR benchmarks.

An interesting implication of the large size of the MS MARCO training dataset is that it allows for further filtering to generate new domain-specific training datasets that may be adequately large to finetune deep models specializing in a given domain. This is particularly interesting when due to time sensitivity or resource constraints it is infeasible to curate a domain-specific training dataset

from scratch. Such a scenario emerged in 2020, when in response to the COVID-19 pandemic, the body of academic literature on Coronavirus grew significantly which in turn posed a difficult challenge for the information retrieval community to quickly devise better methods for searching over this growing scientific corpus. This prompted the creation of the Covid-19 Open Research Dataset (CORD-19) [81] and the TREC-COVID [60, 73] benchmarking effort on one hand, and a flurry of new research and development of IR systems specializing on this task [9, 69, 80] on the other. In particular, MacAvaney et al. [48, 49] created Med-MARCO, a subset of the MS MARCO dataset that are related to medical questions. Subsequently, several groups benchmarking on TREC-COVID employed this subset for model training [46, 83, 88, 89], while others explored finetuning on the full MS MARCO for this task [5, 40, 46, 53, 63, 71]. In a meta-analysis of participating runs in the TREC-COVID challenge, Chen and Hersh [9] found the use of MS MARCO dataset for finetuning to be associated with higher retrieval performance. Similar to Med-MARCO [48, 49], Hamzei et al. [31] studies place-related subset of the MS MARCO dataset. Another interesting case study in this context is the application of MS MARCO to conversational search where it has been useful for both creation of new benchmarks [17, 18, 59] and model training [26, 39, 50, 70, 77–79, 84, 86]. The adoption of MS MARCO in so many transfer learning settings is encouraging, and while it may be premature to draw parallels between its impact on the IR community and what ImageNet [20, 64] did for computer vision research, the current trends definitely bode well for MS MARCO's potential role in the future of IR research.

6 ROBUST USEFULNESS AND EXTERNALITIES

For rankers based on pretrained transformers to become a standard solution in research and industry, we need to show that they can be easily be deployed in new settings. Section 5 indicated that models can work well in a new target domain, but this may involve domain-specific data and multiple stages of finetuning. Future research could work on developing a "play book" for ranker deployment, with the goal of simplifying the process and decreasing the chances of problems or failure. This could include development of self-tuning rankers that can learn from the corpus and/or usage logs when deployed. It could also include the development of a general-purpose ranker, that works reasonably well in a new application with no additional finetuning.

Considering issues of deployment raises the common adage "you can't improve what you don't measure". Data sets and evaluation efforts have an incentive structure, that encourages work towards exactly what is measured, creating blind spots in other areas. MS MARCO not only serves to compare existing IR methods but also plays the Pied Piper guiding a significant section of the community down specific lanes of research. As the curator of such benchmarks, it is therefore crucial that we critically reflect on where we are going and also importantly where we are choosing not to invest.

For example, the availability of a large training dataset directly incentivizes new methods that can take advantage of millions of labeled query-document pairs. Excitement in the large-data work means we may see too few submissions in our benchmarking of methods in the small-data regime, even though such approaches have advantages of efficiency and robustness. Similarly, MS MARCO is English-only, reducing our likelihood of seeing related advances in non-English and cross-language IR.

Both the MS MARCO leaderboard and the TREC Deep Learning track focuses singularly on measuring the relevance quality of retrieved documents and passages, without any consideration for other critical aspects such as efficiency or cost of deployment. This could for example lead the community towards building new models that are frustratingly hard for others, with limited compute resources, to further optimize or deploy. That could again create a divide between what we focus on as a community and what is practically useful. The scenario may be even more serious if we were to consider the potential social harms—specifically on those who belong to historically marginalized communities—and ecological costs of large language models [4], exactly the type of technology that MS MARCO and TREC Deep Learning track may encourage us to work on. As we develop these new benchmarks, the responsibility rests squarely on our own shoulders to think broadly and have open and inclusive conversations about the impact of leading a large section of our community down a given path.

7 CONCLUSION

The MS MARCO leaderboards and TREC Deep Learning track have led to several new ranking approaches, and we have considered multiple aspects of the validity of such studies and usefulness of such approaches. Our bootstrapping analysis showed that the leaderboards are quite stable, meaning that we can be highly confident that we are distinguishing between new methods and our baseline. Since a stable leaderboard can still have overfitting through multiple submission, we limit submissions per group and we have a private leaderboard that we monitor. We also have two evaluation efforts with different labeling schemes, where we use both schemes in both efforts to see if our results are robust. Given potential problems with multiple testing on reusable test collections, we agree with recent SIGIR keynotes [27, 75] that the gold standard is to submit to evaluation efforts such as TREC, and we also found value from having an associated leaderboard with appropriate safeguards.

Since IR metrics may not be interval-scale, and our leaderboard uses one of the most-criticized of these metrics (MRR), we analyzed our leaderboard also using NDCG, both in the bootstrapping analysis and in a test focused on the reliability of statistical tests. We found similar results using MRR and NDCG. Using a variety of statistical tests, some of which do not assume an interval scale, we found a reasonable level of reliability in results. We also argued that there is probably a big gap between true user-percieved utility and current IR metrics, so we suggest work in closing this gap.

We noted that there has been a lot of progress in adapting MS MARCO to other applications, in some cases with multi-stage finetuning. Although this is promising, it suggests that we can not simply deploy a ranker in a new domain without significant data collection and machine learning work. To truly make the new rankers robust and useful, we should make them easier to deploy. We also note a variety of blind spots in our evaluation efforts, suggesting new directions for data and evaluation efforts 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] Timothy G Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 601–610.

[3] 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:1611.09268v3* (2016).

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of FAccT 2021*.

[5] Michael Bendersky, Honglei Zhuang, Ji Ma, Shuguang Han, Keith Hall, and Ryan McDonald. 2020. RRF102: Meeting the TREC-COVID challenge with a 100+ runs ensemble. *arXiv preprint arXiv:2010.00200* (2020).

[6] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, 33–40.

[7] Ben Carterette. 2015. The Best Published Result is Random: Sequential Testing and Its Effect on Reported Effectiveness. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 747–750.

[8] Rich Caruana, Thorsten Joachims, and Lars Backstrom. 2004. KDD-Cup 2004: results and analysis. *ACM SIGKDD Explorations Newsletter* 6, 2 (2004), 95–108.

[9] Jimmy Chen and William Hersh. 2020. A Comparative Analysis of System Features Used in the TREC-COVID Information Retrieval Challenge. *medRxiv* (2020).

[10] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track.. In *TREC*, Vol. 4. 74.

[11] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the trec 2009 web track*. Technical Report. Waterloo Univ (Ontario).

[12] Nick Craswell. 2009. Mean Reciprocal Rank. *Encyclopedia of database systems* 1703 (2009).

[13] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2019 deep learning track.

[14] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track.

[15] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen Voorhees, and Ian Soboroff. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. ACM (to appear).

[16] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. 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*. 126–134.

[17] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624* (2020).

[18] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1985–1988.

[19] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 65–74.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

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

[22] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC complex answer retrieval overview. In *Proceedings of TREC*.

[23] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. 2021. Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *arXiv preprint arXiv:2101.02668* (2021).

[24] Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2017. Are IR evaluation measures on an interval scale?. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 67–74.

[25] Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2018. A general theory of IR evaluation measures. *IEEE Transactions on Knowledge and Data Engineering* 31, 3 (2018), 409–422.

[26] Rafael Ferreira, Mariana Leite, David Semedo, and Joao Magalhaes. 2021. Open-Domain Conversational Search Assistant with Transformers. *arXiv preprint arXiv:2101.08197* (2021).

[27] Norbert Fuhr. [n.d.]. SIGIR Keynote: Proof By Experimentation? Towards Better IR Research. ([n. d.]).

[28] Norbert Fuhr. 2018. Some common mistakes in IR evaluation, and how they can be avoided. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 32–41.

[29] Norbert Fuhr. 2020. Proof by Experimentation? Towards Better IR Research. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2.

[30] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2020. Modularized transfomer-based ranking framework. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4180–4190.

[31] Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko. 2019. Place questions and human-generated answers: A data analysis approach. In *International Conference on Geographic Information Science*. Springer, 3–19.

[32] Donna Harman. 1993. Overview of the first TREC conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 36–47.

[33] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online evaluation for information retrieval. *Foundations and trends in information retrieval* 10, 1 (2016), 1–117.

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

[35] David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 329–338.

[36] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM TOIS* 20, 4 (2002), 422–446.

[37] Jyun-Yu Jiang, Chenyan Xiong, Chia-Jung Lee, and Wei Wang. 2020. Long Document Ranking with Query-Directed Sparse Transformer. *arXiv preprint arXiv:2010.12683* (2020).

[38] David H Krantz, Patrick Suppes, and R Duncan Luce. 2006. *Foundations of Measurement: Additive and polynomial representations*. Vol. 1. Courier Corporation.

[39] Vaibhav Kumar and Jamie Callan. 2020. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 3971–3980.

[40] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage representation aggregation for document reranking. *arXiv preprint arXiv:2008.09093* (2020).

[41] Jimmy Lin. 2018. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (2018), 40–51.

[42] Jimmy Lin. 2019. The neural hype, justified! A recantation. In *SIGIR Forum*, Vol. 53. 88–93.

[43] Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. *Overview of the trec-2014 microblog track*. Technical Report. MARYLAND UNIV COLLEGE PARK.

[44] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. arXiv:2010.06467 [cs.IR]

[45] Tie-Yan Liu. 2011. Learning to rank for information retrieval. (2011).

[46] Zhenghao Liu, Chenyan Xiong, Zhuyun Dai, Si Sun, Maosong Sun, and Zhiyuan Liu. 2020. Adapting Open Domain Fact Extraction and Verification to COVID-FACT through In-Domain Language Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2395–2400.

[47] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2020. PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. *arXiv preprint arXiv:2010.10137* (2020).

[48] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. SLEDGE: A simple yet effective baseline for coronavirus scientific knowledge search. *arXiv preprint arXiv:2005.02365* (2020).

[49] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search. *arXiv preprint arXiv:2010.05987* (2020).

[50] Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2020. Topic Propagation in Conversational Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2057–2060.

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

[52] Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. 2017. Benchmark for complex answer retrieval. ACM, 293–296.

[53] Vincent Nguyen, Maciek Rybinski, Sarvnaz Karimi, and Zhenchang Xing. 2020. Pandemic Literature Search: Finding Information on COVID-19. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*. 92–97.

[54] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[55] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713* (2020).

[56] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[57] Filip Radlinski and Nick Craswell. 2010. Comparing the sensitivity of information retrieval metrics. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 667–674.

[58] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[59] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with search engines. *arXiv preprint arXiv:2004.14162* (2020).

[60] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association* 27, 9 (2020), 1431–1436.

[61] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[62] Corby Rosset, Damien Jose, Gargi Ghosh, Bhaskar Mitra, and Saurabh Tiwary. 2018. Optimizing query evaluations using reinforcement learning for web search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1193–1196.

[63] Yastil Rughbeer, Anban W Pillay, and Edgar Jembere. 2021. Dataset Selection for Transfer Learning in Information Retrieval. In *Southern African Conference for Artificial Intelligence Research*. Springer, 53–65.

[64] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[65] Tetsuya Sakai. 2020. On Fuhr's Guideline for IR Evaluation. In *SIGIR Forum*, Vol. 54. p14.

[66] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, et al. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. *Proceedings of NTCIR-15. to appear* (2020).

[67] Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 162–169.

[68] Kohei Shinden, Atsuki Maruta, and Makoto P Kato. [n.d.]. KASYS at the NTCIR-15 WWW-3 Task. ([n. d.]).

[69] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Deep Learning applications for COVID-19. *Journal of Big Data* 8, 1 (2021).

[70] Vasileios Stamatis, Leif Azzopardi, and Alan Wilson. 2019. VES Team at TREC Conversational Assistance Track (CAsT) 2019.. In *TREC*.

[71] Si Sun, Yingzhuo Qian, Zhenghao Liu, Chenyan Xiong, Kaitao Zhang, Jie Bao, Zhiyuan Liu, and Paul Bennett. 2020. Meta Adaptive Neural Ranking with Contrastive Synthetic Supervision. *arXiv preprint arXiv:2012.14862* (2020).

[72] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2016. Learning latent vector spaces for product search. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 165–174.

[73] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: constructing a pandemic information retrieval test collection. *arXiv preprint arXiv:2005.04474* (2020).

[74] Ellen M Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Trec*.

[75] Ellen M Voorhees. 2020. Coopetition in IR Research. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[76] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge, MA.

[77] Nikos Voskarides, Dan Li, Andreas Panteli, and Pengjie Ren. 2019. ILPS at TREC 2019 Conversational Assistant Track.. In *TREC*.

[78] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 921–930.

[79] Disen Wang and Hui Fang. 2019. Exploring Query Reformulation for Conversational Information Seeking.. In *TREC*.

[80] Lucy Lu Wang and Kyle Lo. 2020. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics* (2020).

[81] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv* (2020).

[82] 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*. 55–64.

[83] Chenyan Xiong, Zhenghao Liu, Si Sun, Zhuyun Dai, Kaitao Zhang, Shi Yu, Zhiyuan Liu, Hoifung Poon, Jianfeng Gao, and Paul Bennett. 2020. CMT in TREC-COVID Round 2: Mitigating the Generalization Gaps from Web to Special Domain Search. *arXiv preprint arXiv:2011.01580* (2020).

[84] Jheng-Hong Yang, Sheng-Chieh Lin, Chuan-Ju Wang, Jimmy Lin, and Ming-Feng Tsai. 2019. Query and Answer Expansion from Conversation History.. In *TREC*.

[85] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3481–3487.

[86] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1933–1936.

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

[88] Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv preprint arXiv:2004.05125* (2020).

[89] Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, et al. 2020. Covidex: Neural ranking models and keyword search infrastructure for the covid-19 open research dataset. *arXiv preprint arXiv:2007.07846* (2020).

[90] Kaitao Zhang, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2020. Selective weak supervision for neural information retrieval. In *Proceedings of The Web Conference 2020*. 474–485.

[91] Xinyu Zhang, Andrew Yates, and Jimmy Lin. 2020. A Little Bit Is Worse Than None: Ranking with Limited Training Data. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. 107–112.

[92] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: contextualized query expansion for document re-ranking. *arXiv preprint arXiv:2009.07258* (2020).

[93] Mu Zhu. 2004. Recall, Precision and Average Precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo* 2 (2004), 30.

[94] Justin Zobel and Alistair Moffat. 1998. Exploring the similarity space. In *Acm Sigir Forum*, Vol. 32. ACM New York, NY, USA, 18–34.