Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard

Jimmy Lin,[1,2] Daniel Campos,[3] Nick Craswell,[2] Bhaskar Mitra,[2,4] and Emine Yilmaz[4]
[1] University of Waterloo, Canada [2] Microsoft AI & Research, USA
[3] University of Illinois Urbana-Champaign, USA [4] University College London, UK

ABSTRACT

Leaderboards are a ubiquitous part of modern research in applied machine learning. By design, they sort entries into some linear order, where the top-scoring entry is recognized as the "state of the art" (SOTA). Due to the rapid progress being made today, particularly with neural models, the top entry in a leaderboard is replaced with some regularity. These are touted as improvements in the state of the art. Such pronouncements, however, are almost never qualified with significance testing. In the context of the MS MARCO document ranking leaderboard, we pose a specific question: How do we know if a run is *significantly* better than the current SOTA? Against the backdrop of recent IR debates on scale types, our study proposes an evaluation framework that explicitly treats certain outcomes as distinct and avoids aggregating them into a single-point metric. Empirical analysis of SOTA runs from the MS MARCO document ranking leaderboard reveals insights about how one run can be "significantly better" than another that are obscured by the current official evaluation metric (MRR@100).

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**.

KEYWORDS

Significance Testing; Evaluation Metrics

ACM Reference Format:
Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2021. Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard. 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, 5 pages. https://doi.org/10.1145/3404835.3463034

1 INTRODUCTION

Leaderboard rankings and claims of the "state of the art" (SOTA) pervade modern research in applied machine learning, particularly in NLP and IR. There has been much debate in the community on the merits of such activities, compared to alternative uses of the same researcher energy, attention, and resources. Without participating in this debate, this work attempts to address what we view as a technical shortcoming of many, if not most, leaderboards today: the lack of significance testing. Specifically, we wish to answer the question: Does a particular run significantly improve the state of the art? Rapid progress on leaderboards means that the top-scoring run is regularly overtaken and replaced. This is communicated (in papers, blog posts, tweets, etc.) as beating the existing SOTA and achieving a new SOTA. Such pronouncements, however, are rarely qualified with significance tests. We hope to take a small step towards rectifying this.

Our study focuses on the MS MARCO document ranking leaderboard [3–5], and against the backdrop of recent debates about IR evaluation [6, 7, 14], we discover, unsurprisingly, that "it's complicated". Our findings can be summarized as follows:

(1) The existing single-point quality metric (MRR@100) conflates important differences in *how* one run can be "better" than another. Thus, the naive approach of running standard significance tests on the existing metric may lead to questionable results.
(2) To address this issue, we propose an evaluation framework that explicitly tracks outcomes separately, which then permits meaningful aggregation and significance testing. From a qualitative perspective, this framework reveals many insights about differences that are obscured by the existing official metric.
(3) Contributing to recent debates in the IR community on scale types, we find that in our framework, analysis in terms of expected search length (ESL), a ratio scale, and mean reciprocal rank (MRR), an ordinal scale, *can* yield different conclusions.

Our contribution is a novel evaluation framework that compares putative SOTA submissions in a nuanced way that contributes to ongoing debates in the IR community about evaluation methodologies. We find that runs can be "better" in different ways, but these "different ways" cannot be reconciled without appealing to a user model of utility (presently absent in the task definition).

It is worth emphasizing that in this paper, we are asking a very narrow question about entries on a leaderboard and significance testing with respect to a clearly defined metric. There are a number of questions that are outside the scope of inquiry, for example: Is the new SOTA technique practically deployable? Is the improvement in the SOTA meaningful from a user perspective? Might the new SOTA technique encode biases? Etc. While these are all important considerations, they raise orthogonal issues that we do not tackle here. Nevertheless, even for such a narrowly framed question, there is still quite a bit of nuance that is missing in the current discourse.

2 BACKGROUND AND RELATED WORK

The MS MARCO dataset [13] was originally released in 2016 with the aim of helping academic researchers explore information access in the large-data regime, particularly in the context of models based

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

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463034

on neural networks that were known to be data hungry [10, 11]. Initially, the dataset was designed to study question answering on web passages, but it was later adapted into traditional *ad hoc* ranking tasks. Today, the document ranking and passage ranking tasks host competitive leaderboards that attract much attention from researchers around the world.

This paper focuses on the document ranking task, which is a standard *ad hoc* retrieval task over a corpus of 3.2M web pages with URL, title, and body text. The organizers have made available a training set with 367K queries and a development set with 5193 queries; each query has exactly one relevance judgment. There are 5793 evaluation (test) queries; relevance judgments for those queries are withheld from the public. Scores on the evaluation queries can only be obtained by a submission to the leaderboard. The official metric is mean reciprocal rank at a cutoff of 100 (MRR@100).

Of the myriad metrics that have been proposed to evaluate retrieval systems, there are those that make strong claims as to modeling user utility, such as nDCG [8] and RBP [12], and those that do not, say, precision at a fixed cutoff. Specifically in the context of the MS MARCO document ranking task, reciprocal rank (RR) makes at least some plausible claims about utility. At a high level, the metric says that the user only cares about getting a single relevant document (not unrealistic since MS MARCO models question answering "in the wild"), and that utility drops off rapidly as a function of increasing ordinal rank. While the functional form of this dropoff might be a matter of debate, there is strong empirical support for the claim in general, dating back well over a decade. In web search, log analysis (e.g., [1]) as well as eye-tracking experiments (e.g., [9]) have shown that user click probabilities and attention fall rapidly with increasing ordinal rank in the retrieved results.

We believe that at least some of the ongoing controversies about evaluation methodologies in information retrieval stem from confusion on whether a metric is being used simply as a useful proxy for effectiveness (to aid in quantifying model improvements) or is actually making a claim about utility. Thus, in this paper, we are careful to separate the two, and are explicit when making a claim about utility (and appealing to some user model).

The proximate motivation of this study is the recent work of Ferrante et al. [6], who argued that most IR metrics are not interval-scaled and suggested that decades of IR research may be methodologically flawed. We do not have sufficient space to present their detailed arguments, but the crux in our context is that for RR, intervals are not equi-spaced; that is, a difference of 0.1 (let's say) "means" different things at different points on the scale. As a contrast, the standard example of an interval scale is temperature measured in Celsius, where "one degree" means exactly the same thing (i.e., difference in temperature) everywhere. Recognizing that ongoing debates in IR evaluation are by no means settled [14], and without necessarily agreeing with their arguments, we conduct analyses that consider both positions with respect to scale types and permissibility of different operations and statistical analyses.

3 NAIVE FIRST ATTEMPT

A summary of the MS MARCO document ranking leaderboard since its launch in August 2020 (until the end of April, 2021) is shown in Figure 1, where each point represents a run: the x-axis plots the date

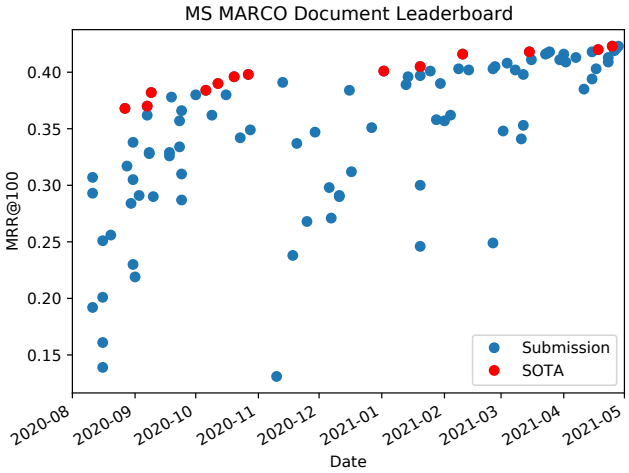

Figure 1: The leaderboard of the MS MARCO document ranking task, showing the effectiveness of runs (MRR@100) on the held-out evaluation set over time. "State of the art" (SOTA) runs are shown in red.

of submission and the y-axis plots the official metric (MRR@100) reported on the leaderboard for the held-out evaluation (test) set. Circles in red represent the (current and former) state-of-the-art (SOTA) runs, i.e., a run that displaced a previous run at the top of the leaderboard, beginning with the first submission that beat the baselines provided by the organizers. Our analysis specifically focused on these SOTA runs. Since the identities of the runs are not germane to our analysis, we simply denote them R_1 (the oldest) to R_{13} (the most recent), arranged chronologically.

If we wish to ask if one SOTA run is significantly better than another, an obvious first attempt would be to run some standard statistical test over per-query scores of the official metric (MRR@100). Among the myriad tests available, three stand out: (1) Wilcoxon rank sum test (WRS), a non-parametric test that requires samples to be on an ordinal scale, (2) Wilcoxon signed rank test (WSR), a non-parametric test that requires samples to be on an interval scale, (3) Student's t-test, a parametric test that requires samples to be on an interval scale. As discussed in Ferrante et al. [6], there has been quite a bit of controversy (in IR and beyond) on what tests are permissible for what scale types. Even taking the most stringent position, there is no doubt that reciprocal rank is an ordinal scale. Thus, WRS is unequivocally permissible. With these caveats rendered explicit, let's just run all the tests anyway.

The results of running all three significance tests on pairwise comparisons between R_1 (the earliest SOTA run) and every other subsequent SOTA run $\{R_2 \ldots R_{13}\}$ on the evaluation set are shown in Table 1. Additionally, we compare the three current top runs on the leaderboard. The table reports the absolute differences in effectiveness,[1] along with the raw p-values of the different tests, prior to the application of the Bonferroni correction.

[1]There is an important detail here worth mentioning: the official evaluation script deliberately introduces a metric artifact designed to thwart (simple) attempts at "reverse-engineering" the evaluation set. Thus, the scores reported on the official leaderboard (and plotted in Figure 1) are *not* accurate. This artifact has no impact on the leaderboard rankings, but *does* impact significance testing. All our analyses are based the true MRR@100 scores, after the removal of this artifact. Thus, the absolute score differences may not line up with public leaderboard results.

Runs			WRS	WSR	t-test
A	B	Δ	p-value	p-value	p-value
R_1	R_2	0.0027	0.029393	0.519758	0.722586
R_1	R_3	0.0152	0.109672	0.053127	0.038240
R_1	R_4	0.0179	0.000494	0.008780	0.013607
R_1	R_5	0.0244	2.07E-05	0.000461	0.000785
R_1	R_6	0.0301	1.16E-06	3.21E-05	3.77E-05
R_1	R_7	0.0328	1.16E-07	1.00E-06	4.18E-06
R_1	R_8	0.0360	2.81E-05	3.24E-06	2.50E-06
R_1	R_9	0.0406	4.15E-06	9.06E-08	3.61E-08
R_1	R_{10}	0.0516	7.85E-14	1.92E-12	7.57E-12
R_1	R_{11}	0.0545	7.02E-17	2.14E-13	5.74E-13
R_1	R_{12}	0.0567	5.06E-18	1.52E-14	6.22E-14
R_1	R_{13}	0.0592	3.71E-18	2.07E-15	1.82E-14
R_{11}	R_{13}	0.0046	0.706079	0.493485	0.549086
R_{12}	R_{13}	0.0024	0.949692	0.674093	0.754583

Table 1: Results of running significance tests on SOTA runs: Wilcoxon rank sum test (WRS), Wilcoxon signed rank test (WSR), and the Student's t-test. $R_1 \ldots R_{13}$ are the SOTA runs, arranged chronologically.

We see that based on all three tests, the improvements from successive SOTA runs $R_2, R_3, \ldots$ are not statistically significant until we get to R_5; all subsequent runs thereafter appear to be significantly better (even just focusing on WRS). The absolute difference in MRR@100 between R_1 and R_5 is 2.4 points, which is surprisingly large. Independent of the particulars of any evaluation, the general expectation is that with a large number of queries (over 5K in our case), small significant differences (i.e., small effect sizes) should be detectable. Nevertheless, differences at the current top of the leaderboard are not statistically significant (perhaps not surprising).

4 OUTCOMES BREAKDOWN

This section presents our evaluation framework specifically tailored to the MS MARCO document ranking task. We begin with a few (hopefully) uncontroversial claims and from there build an approach to evaluation that explicitly avoids conflating distinct outcomes in a single-point metric. Note that since the official relevance judgments contain only one relevant document per query, the position of that relevant document on the ranked list (or its absence) alone determines the score (the metric, to be defined below) for that query; this nicely sidesteps the challenges with different "recall bases" [6], i.e., queries that have different numbers of relevant documents.

Consider two hypothetical submissions to the MS MARCO document ranking leaderboard, runs A and B, comprising ranked lists over a set of queries Q. For each query $q \in Q$, there are logically the following distinct outcomes that cover all possibilities:

(1) Neither run A nor run B returns the relevant document in the top k. In this case, both runs are equally "bad".
(2) Run A returns the relevant document in the top k, while run B does not; A is thus "better". Vice versa with A and B swapped.
(3) Both runs A and B return the relevant document in the top k, but the document has lower ordinal rank in B, and thus is "better". Vice versa with B and A swapped.

We believe that the above assertions hold regardless of how one might choose to operationalize "bad" and "better". However, to be more precise, let us define a metric in terms of expected search length (ESL), which has a long history in IR research dating back to the 1960s [2]. ESL quantifies how long a user needs to search (more specifically, read the ranked list) before obtaining a relevant document: A relevant document appearing at rank 1 gets a score of 1, rank 2 gets a score of 2, etc. all the way up to rank 100 (in our case). Thus, the lower the score, the better.

Consider a straightforward user model: a patient user who issues a query, reads 100 documents per query to find the relevant document, and then gives up if no relevant document is found. It would be plausible to make the claim that ESL, with respect to this user model, captures utility measured in user time.[2] It is clear that ESL is on a ratio scale (by definition also an interval scale). Against our user model, the following claims would be meaningful with respect to utility (time): a relevant document at rank 4 (4 ESL) costs the user twice as much utility (time) as a relevant document at rank 2 (2 ESL). For case (3) in the list of outcomes above, when comparing run A and run B for a specific query q, we have a good alignment between ESL and utility.

Note, critically, however, that this only applies to case (3) above, when both runs contain the relevant document in their top k lists. For case (2), however, it is unclear how similar statements can be made: i.e., what ESL would we assign for not having a relevant document retrieved? There's nothing in the framework we've presented thus far that would shed light on this without a more refined user model (absent in the current task definition). Note that the official metric MRR@100 *does* encode a specific utility difference between a retrieved document at rank 100 and not retrieving the relevant document (0.01, to be exact), but justifying this value requires appealing to user models and data (e.g., behavior logs) that are beyond the scope of the leaderboard. We argue, instead, that the best way forward is to maintain an explicit separation and breakdown of the different outcomes.

5 APPLICATION TO MS MARCO

Let us apply the framework proposed above to analyze the SOTA runs on the MS MARCO document ranking leaderboard. In our analysis, we compared each of $R_2 \ldots R_{13}$ against R_1, the results of which are shown in Table 2; we additionally compared the current top three runs on the leaderboard, R_{11}–R_{13}. We show the percentage of queries in each outcome—case (1), (2), or (3)—as described in the previous section. Case (2) is broken down into "A wins" and "B wins". For rhetorical convenience, we will use "answered" and "unanswered" for these cases. For case (3), we show the overall percentage, as well as the mean ESL and reciprocal rank (RR) for all queries in that outcome; also presented are p-values from the Wilcoxon signed-rank test and the paired t-test (for each metric). Note that since ESL is on an interval (ratio) scale, these two tests are unequivocally permissible. For RR, the applications of the Wilcoxon signed-rank test and the paired t-test are subjected to the potential

[2]Recognizing that we are making a few simplifying assumptions such as constant document length and fixed reading speed. More realism could be added by, for example, taking into account a more accurate model of reading speed [15], but these refinements are unlikely to change our overall analysis.

Runs			(1)	(2)		(3)								
A	B	Δ	All	A wins	B wins	All	A ESL	B ESL	WSR p-value	t-test p-value	A RR	B RR	WSR p-value	t-test p-value
R_1	R_2	0.0027	2%	9%	16%	73%	6.99	9.14	3.22E-12	1.60E-07	0.4812	0.4468	6.75E-05	2.66E-05
R_1	R_3	0.0152	4%	15%	14%	67%	7.15	5.99	4.62E-06	1.34E-05	0.4857	0.5107	0.004032	0.003871
R_1	R_4	0.0179	3%	10%	15%	72%	7.06	8.13	0.000249	0.004918	0.4838	0.4775	0.448624	0.368583
R_1	R_5	0.0244	3%	11%	15%	71%	7.13	7.27	0.609428	0.813828	0.4833	0.4861	0.730242	0.765044
R_1	R_6	0.0301	3%	10%	15%	71%	7.15	7.19	0.886547	0.381946	0.4829	0.4910	0.331145	0.364266
R_1	R_7	0.0328	4%	10%	15%	72%	7.27	6.98	0.279188	0.022938	0.4799	0.4979	0.024974	0.020952
R_1	R_8	0.0360	3%	15%	15%	67%	7.06	5.71	7.94E-08	8.68E-08	0.4829	0.5273	8.97E-07	7.59E-07
R_1	R_9	0.0406	4%	15%	14%	67%	7.15	5.15	2.03E-16	4.33E-17	0.4857	0.5404	2.49E-10	1.85E-10
R_1	R_{10}	0.0516	3%	12%	15%	70%	7.08	5.88	4.79E-06	1.46E-06	0.4829	0.5204	1.60E-05	1.10E-05
R_1	R_{11}	0.0545	2%	10%	16%	72%	7.13	6.51	2.07E-03	1.61E-02	0.4818	0.5112	7.01E-04	6.72E-04
R_1	R_{12}	0.0567	2%	10%	16%	72%	7.13	6.36	2.71E-04	2.82E-03	0.4818	0.5146	1.33E-04	1.43E-04
R_1	R_{13}	0.0592	2%	10%	16%	72%	7.26	7.31	0.017467	0.866769	0.4777	0.5079	4.87E-04	5.08E-04
R_{11}	R_{13}	0.0046	1%	10%	11%	78%	6.66	7.28	0.534906	0.024494	0.5092	0.5129	0.662144	0.668119
R_{12}	R_{13}	0.0024	1%	10%	11%	78%	6.56	7.28	0.328843	0.009061	0.5110	0.5131	0.788919	0.808087

Table 2: Analysis of SOTA runs from the MS MARCO document ranking leaderboard, broken into distinct outcomes.

objections raised by Ferrante et al. [6] regarding scale types. In all cases, we report raw p-values, prior to Bonferroni correction.

This case study reveals interesting insights that are completely hidden if we simply reported the means of per-query reciprocal ranks and ran significance tests on them, as in Section 3. In terms of ESL, we highlight a number of interesting observations:

- Comparing R_1 vs. R_2, the overall MRR@100 scores are quite close, but the runs appear to be very different. Looking at the case (3) breakdowns, we see that R_2 has a higher ESL than R_1, and this difference is (highly) statistically significant. From this perspective, R_2 is worse than R_1. However, R_2 answered more queries that went unanswered in R_1 than the other way around. A similar observation can be made in R_1 vs. R_4. When focusing only on ranking, case (3), R_4 is significantly worse than R_1, but R_4 compensates by answering more queries that went unanswered in R_1, leading to a higher score in terms of MRR@100.
- Consider R_1 vs. R_3: contrary to the above examples, we see that R_3 significantly improves ranking, case (3), but has slightly more unanswered queries compared to the baseline. This also leads to an overall improvement in terms of MRR@100.
- Consider R_1 and R_8, the prevalence of case (2): there are equal percentages of cases where the query was answered by one run but not the other. However, looking at case (3), we see that R_8 obtains a statistically significant reduction in ESL. That is, R_8 is better than R_1 because it does a better job ranking.
- Another interesting observation relates to R_1, R_5, R_6, R_7, which are runs from the same team. Comparing R_1 to $\{R_5, R_6, R_7\}$, differences in ESL are not statistically significant, case (3). That is, the runs are comparable when it comes to ranking. The differences in MRR@100 come primarily from case (2), where $\{R_5, R_6, R_7\}$ have fewer unanswered queries overall.
- Looking at the current top of the leaderboard: Consider R_1 vs. R_{13}, where the latter has substantially higher MRR@100, but from case (3), it is unclear if R_{13} does a significantly better job at ranking. Instead, the higher effectiveness appears to come from answering more questions. Note that R_{11} and R_{12} are from the same team, so we can set aside an analysis of R_{11}. Looking at R_{12} vs. R_{13}, we see that the current SOTA run is actually worse than the second-best run at ranking, from case (3).

Examining differences between ESL and RR, in some cases they lead to opposite conclusions—in fact, at the current top of the leaderboard. Comparing R_{12} vs. R_{13}: while the latter answers a few more questions, the key difference lies in case (3) outcomes. According to RR, R_{13} is a tiny bit better than R_{12} (*n.s.*), but R_{12} appears to produce better rankings, by ~0.7 rank positions on average (sig. on t-test, but not WSR). Based on ESL, an argument can be made that R_{12} is the SOTA. Indeed, different metrics can give rise to different run orderings, so this is not a surprising finding.

Back to the original question that we set out to answer: Is "this" SOTA run better than "that" SOTA run? We might say that run B is better than run A if run B wins in terms of case (2) *and* has a smaller ESL for case (3), or alternatively, higher MRR. We might further claim that run B is significantly better than run A if the improvements in both outcomes are statistically significant: for case (3), significant testing as we have performed above, and for case (2), perhaps the binomial test (we have done this analysis, but lack the space to share results). Alternatively, we might adopt a less stringent definition, akin to the Hippocratic Oath (i.e., "do no harm"): a run can be considered "significantly better" if it significantly increases answered questions without significantly increasing the ESL *or* that it significantly decreases ESL without significantly increasing unanswered questions. This proposal has the advantage in providing two concrete facets of "goodness" that researchers can independently tackle while still being amenable to a linear sort order for populating a leaderboard.

6 CONCLUDING THOUGHTS

How might we generalize our framework to encompass other retrieval tasks and possibly beyond? We see at least one challenge that limits broader applicability: our approach depends on having only one relevant document per query, since this property is necessary to separate the outcomes. While this restriction is not unrealistic for some tasks (e.g., question answering), there clearly needs to be more work before our approach can be generalized.

REFERENCES

[1] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. 2006. Learning User Interaction Models for Predicting Web Search Result Preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*. Seattle, Washington, 3–10.

[2] William S. Cooper. 1968. Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. *Journal of the American Society for Information Science* 19, 1 (1968), 31–40.

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2020 Deep Learning Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference Proceedings (TREC 2020)*.

[4] 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 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*.

[5] 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. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*.

[6] 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).

[7] 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.

[8] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulative Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[9] Thorsten Joachims, Laura Granka, Bing Pang, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*. Salvador, Brazil, 154–161.

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

[11] Jimmy Lin. 2019. The Neural Hype, Justified! A Recantation. *SIGIR Forum* 53, 2 (2019), 88–93.

[12] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems* 27, 1 (2008), Article 2.

[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268v1* (2016).

[14] Tetsuya Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1 (2020), Article No. 12.

[15] Mark D. Smucker and Charles L. A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. In *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*. Portland, Oregon, 95–104.