Effect on System Rankings of Extending Pools in TREC-COVID Round 1

Ellen Voorhees
National Institute of Standards and Technology

May 25, 2020

The compressed time-frame of TREC-COVID limits the number of document relevance judgments that can be obtained during a round. One of the concerns at the conclusion of the first round of TREC-COVID was the relatively shallow pool depth that could be accommodated in the available time [2]. Shallow pools lead to incomplete judgment sets and relatively large uncertainty in systems' evaluation scores, while the measures for which the uncertainty is small, such as P@5, are not necessarily the best measures to understand the target user task. This note looks at the effect on the rankings of Round 1 runs for different evaluation measures when the initial shallow pools are extended, pooling to depth fourteen instead of depth seven. We observe only minor changes in the relative effectiveness of runs for each of the four measures examined.

1 Judgment Sets

The relevance judgment set ("qrels") used to score Round 1 runs is called the Round 1 qrels[1]. This set contains judgments produced prior to the start of Round 1 (called Judgment Set 0.5 and produced by pooling a few retrieval runs created by TREC-COVID organizers) in union with the judgments produced for Round 1. The time constraints for the Round-1-runs judgments limited the number of documents that could be judged to depth-7 pools created from the first priority run from each of 56 participants. The Round 1 qrels contains approximately 8600 judgments across the 30 topics.

Annotators continued to judge documents during the week between the Round 2 kick-off and run submission deadline. These judgments are called Judgment Set 1.5 because the documents were drawn from Round 1 runs but are used to evaluate Round 2 runs. This set was created by pooling the documents at ranks 8–14 (removing any previously judged document) for the same set of first priority runs of Round 1. Judgment Set 1.5 contains 5770 judgments across the 30 topics.

The combination of the Round 1 qrels and Judgment Set 1.5 is not a qrels set for any official TREC-COVID round. However, the combination does allow us to compare scores for the Round 1 runs as computed using depth-7 pools versus depth-14 pools. Table 1 shows the distribution of judgments across topics for the Round 1 qrels, for Judgment Set 1.5, and for the combination we call the Extended qrels. For each set the table gives the number of documents judged, the number of documents judged partially relevant, the number of documents judged fully relevant, and the percentage of judged documents that are some form of relevant.

2 Measures

TREC-COVID Round 1 participants received scores for all of the measures reported by trec_eval, but three measures were emphasized in the score reports[2]: Prec@5, NDCG@10, and bpref. These measures were

[1] https://ir.nist.gov/covidSubmit/data/qrels-rnd1.txt
[2] https://ir.nist.gov/covidSubmit/archive/archive-round1.html

Table 1: Counts of total numbers of judged documents and number of relevant documents per topic. Percent relevant is the fraction of judged documents that are some form of relevant.

	Round 1 qrels				Judgment set 1.5				Extended qrels			
Topic	Num Judged	Part Rel	Rel	% Rel	Num Judged	Part Rel	Rel	% Rel	Num Judged	Part Rel	Rel	% Rel
1	323	45	56	0.313	225	26	29	0.244	548	71	52	0.285
2	284	21	26	0.165	187	10	4	0.075	471	31	20	0.130
3	337	66	24	0.267	236	51	8	0.250	573	117	102	0.260
4	357	32	27	0.165	232	19	13	0.138	589	51	38	0.154
5	336	35	96	0.390	212	18	35	0.250	548	53	36	0.336
6	321	80	83	0.508	216	53	60	0.523	537	133	106	0.514
7	275	2	47	0.178	200	0	10	0.050	475	2	0	0.124
8	360	46	30	0.211	250	33	9	0.168	610	79	66	0.193
9	298	25	16	0.138	223	2	3	0.022	521	27	4	0.088
10	191	35	50	0.445	126	11	23	0.270	317	46	22	0.375
11	344	67	5	0.209	232	6	0	0.026	576	73	12	0.135
12	324	76	126	0.623	195	19	71	0.462	519	95	38	0.563
13	373	97	49	0.391	244	59	16	0.307	617	156	118	0.358
14	222	24	5	0.131	151	14	4	0.119	373	38	28	0.126
15	348	45	12	0.164	254	31	8	0.154	602	76	62	0.159
16	340	42	11	0.156	223	30	7	0.166	563	72	60	0.160
17	243	32	45	0.317	142	10	8	0.127	385	42	20	0.247
18	267	79	32	0.416	154	22	19	0.266	421	101	44	0.361
19	301	27	16	0.143	207	12	6	0.087	508	39	24	0.120
20	247	41	25	0.267	165	43	4	0.285	412	84	86	0.274
21	319	15	70	0.266	218	7	30	0.170	537	22	14	0.227
22	259	17	30	0.181	198	8	7	0.076	457	25	16	0.136
23	256	4	22	0.102	173	2	2	0.023	429	6	4	0.070
24	249	14	19	0.133	161	6	0	0.037	410	20	12	0.095
25	308	9	62	0.231	201	2	14	0.080	509	11	4	0.171
26	312	19	106	0.401	189	7	31	0.201	501	26	14	0.325
27	300	30	44	0.247	192	52	18	0.365	492	82	104	0.293
28	180	9	29	0.211	114	2	0	0.018	294	11	4	0.136
29	218	42	58	0.459	137	23	28	0.372	355	65	46	0.425
30	199	39	16	0.276	113	20	5	0.221	312	59	40	0.256

selected because of the incompleteness of the relevance judgments. Prec@5 is exact for judged runs, but is strongly affected by unjudged documents in the unjudged runs. NDCG@10 takes advantage of the different relevance levels and is close-to-exact for judged runs for Round 1 qrels and is exact using the Extended qrels (though like Prec@5 it is strongly affected by unjudged documents in the top ranks). Bpref is a measure designed for incomplete judgments because it is computed only over the judged set; it is a function of the number of known irrelevant documents retrieved before known relevant documents. In addition, we also test Average Precision (AP) since it is is an often-used measure in the literature. AP is a function of the number of relevant documents, so AP scores are affected by incomplete judgments. On the other hand, AP is top-heavy in the sense that highly ranked documents contribute the most to the final score, blunting the impact of incomplete judgments due to shallow pools. Prec@5 scores can only increase when more judgments become available while the three other measures' scores can vary in either direction [1]. In practice, AP scores computed over shallow pools tend to over-estimate the true AP score.

We evaluated the set of Round 1 runs using each of the four measures Prec@5, NDCG@10, bpref, or AP, first using the official Round 1 qrels and then using the Extended qrels. Figures 1–4 show the scores of Round 1 runs for each measure and qrels. In each graph the runs are sorted by decreasing score as computed

using the Round 1 qrels ("orig"). The original score is plotted in red and the score for a run as computed using the Extended qrels ("new") is plotted in blue. (When the two scores for a run are the same, it appears in the graph as a single red dot.) There is some small movement of runs' ranks in each of the plots (as indicated when blue dots are not monotonically decreasing with one another), but the rankings are generally stable for all four measures. The Kendall τ correlation between the rankings produced for the same measure on the two different judgment sets is greater than 0.95 for each measure.

3 Discussion

The stability of the rankings, and in particular the lack of individual runs with large changes in rank, is heartening. But the real question is whether the triple consisting of the April 10 release of CORD-19, the 30 test topics, and the Extended qrels forms a general-purpose, reusable test collection [3]. There is no definitive answer to that question since the only (known) ways of answering it is to find a natural run[3] that is ranked poorly to demonstrate reusability issues or to obtain complete judgments that demonstrate reusability. In the absence of either option, we must weigh the evidence that supports or opposes a conclusion of reusability.

Pool depth and number of documents judged provide evidence for opposite sides. On the one hand, a pool depth of fourteen, while bigger than seven, is still historically small in absolute terms. But the number and diversity of the runs from which the pools were drawn are large. The total number of documents judged per topic, approximately 500, is 1% of the document set of approximately 50,000 documents, a massive percentage judged compared to other TREC collections.

The stability of the system rankings when the qrels were extended is the strongest evidence in support of reusability. The runs that had very little overlap with other runs and ranked close to the bottom using the Round 1 qrels continue to rank close to the bottom. In part this is because those runs continue to have very little overlap with other runs so their new judgments come only from their own contribution to the pools. But those new judgments did not find a cache of relevant documents; the assumption of an unjudged document as an irrelevant document remains reasonable for runs with poor scores using the Round 1 qrels. Those runs that did change in rank are mostly unjudged runs that were relatively good with the Round 1 qrels and scored even better with the newly found relevant documents.

The strongest evidence opposing a conclusion of reusability is the percentage of judged documents that are relevant for a sizeable fraction of the topics. Historically, when more than a third of the judged documents are relevant for a topic, it is highly likely that many more relevant documents that have not yet been identified remain in the collection [3]. Having fewer than one third relevant is not a guarantee that the collection is stable, but more than a third has been strong evidence that it is not. For the Round 1 qrels, more than a quarter of the topics (8/30) have relevant percentages greater than 0.33 (see Table 1). For the Extended qrels, those same eight topics still have relevant percentages greater than one third, though the percentages are smaller than for the Round 1 qrels.

Researchers can easily detect the presence of unjudged documents in their own runs and decide how to proceed from there. If the runs to be compared have similar numbers of unjudged documents, and especially if it is a small number of unjudged documents, then comparisons will be stable for a majority of measures. When the number of unjudged is skewed, it is best to take precautions such as using incompleteness-tolerant measures or requiring larger differences in scores before concluding that runs are actually different.

References

[1] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1), 2008.

[3] A natural run is a run that is the output of a well-defined search process. Given a qrels, it will always be possible to construct a document ranking that is likely to score poorly when the qrels are extended (by ranking the currently unjudged document most similar to each known relevant document in the top ranks, for example.) But such a construction does not mimic any real search process, which would almost certainly highly rank some of the known relevants. We are concerned with the behavior of real search processes only.

[2] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. TREC-COVID: Constructing a pandemic information retrieval test collection, 2020.

[3] Ellen M. Voorhees. On building fair and reusable test collections using bandit techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 407–416, 2018.

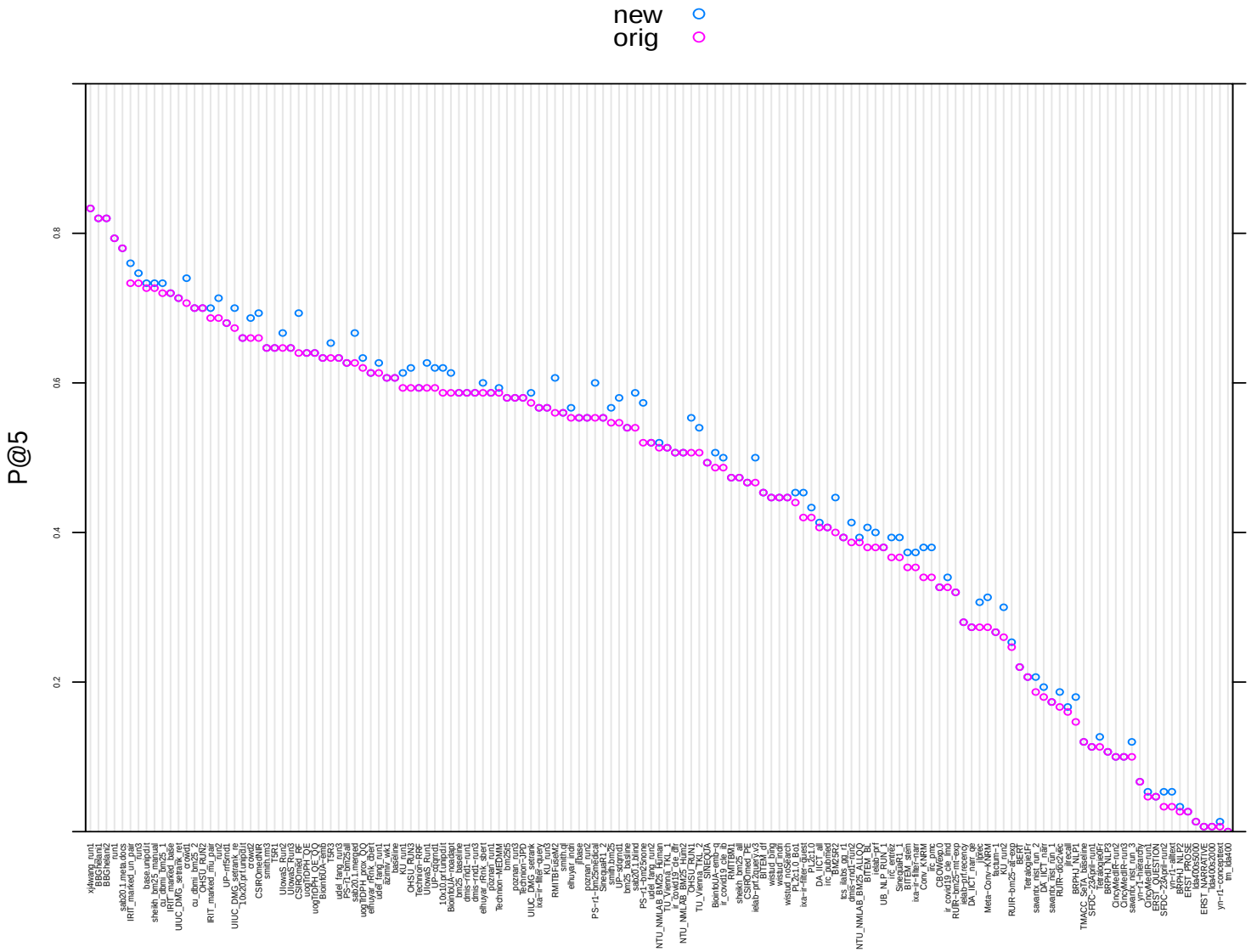

Figure 1: Mean P@5 score per Round 1 run as computed using the Extended qrels ("new" in blue) and the Round 1 qrels ("orig" in red). Runs are sorted by decreasing mean P@5 as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9604.

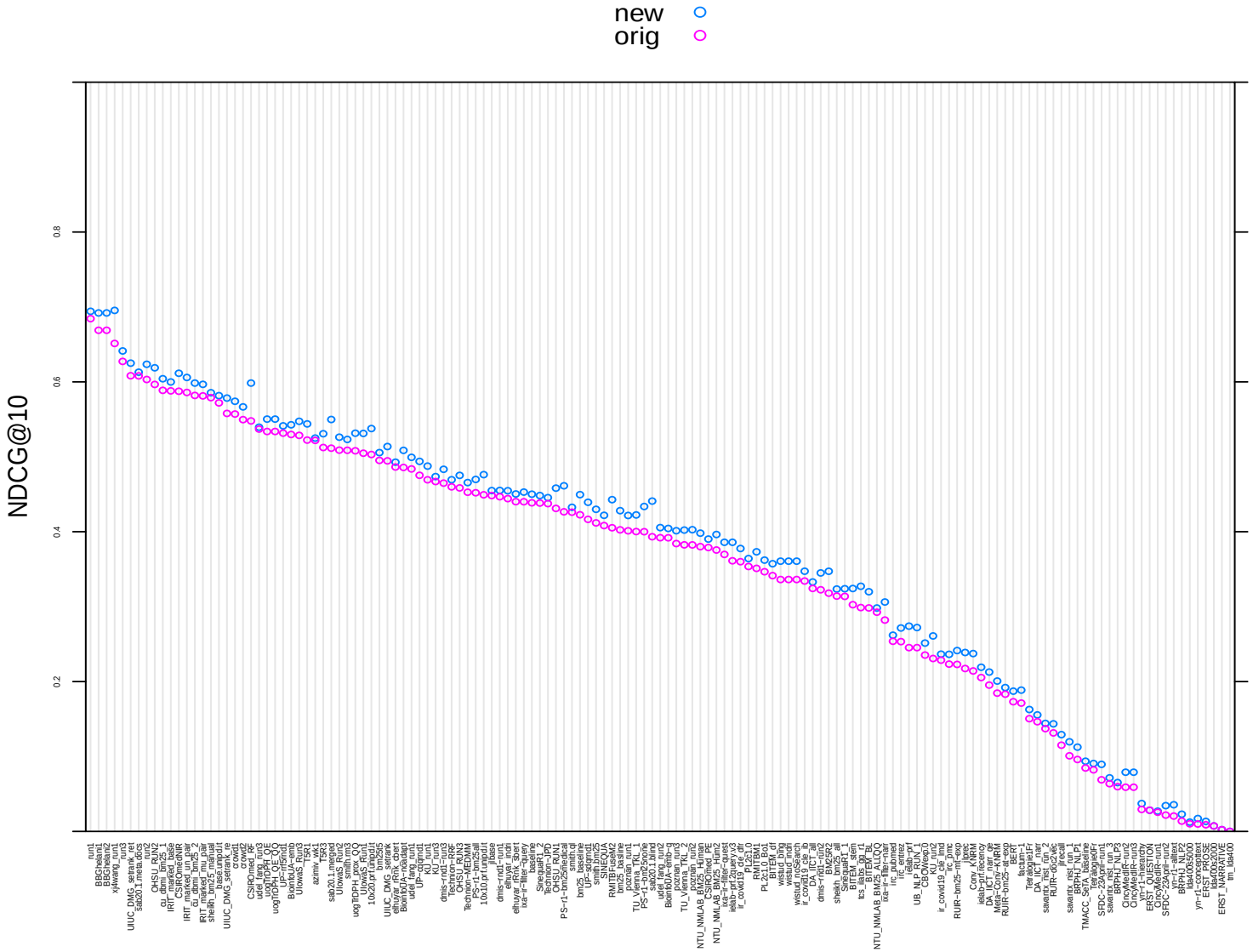

Figure 2: Mean NDCG@10 score per Round 1 run as computed using the Extended qrels ("new" in blue) and the Round 1 qrels ("orig" in red). Runs are sorted by decreasing mean NDCG@10 as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9703.

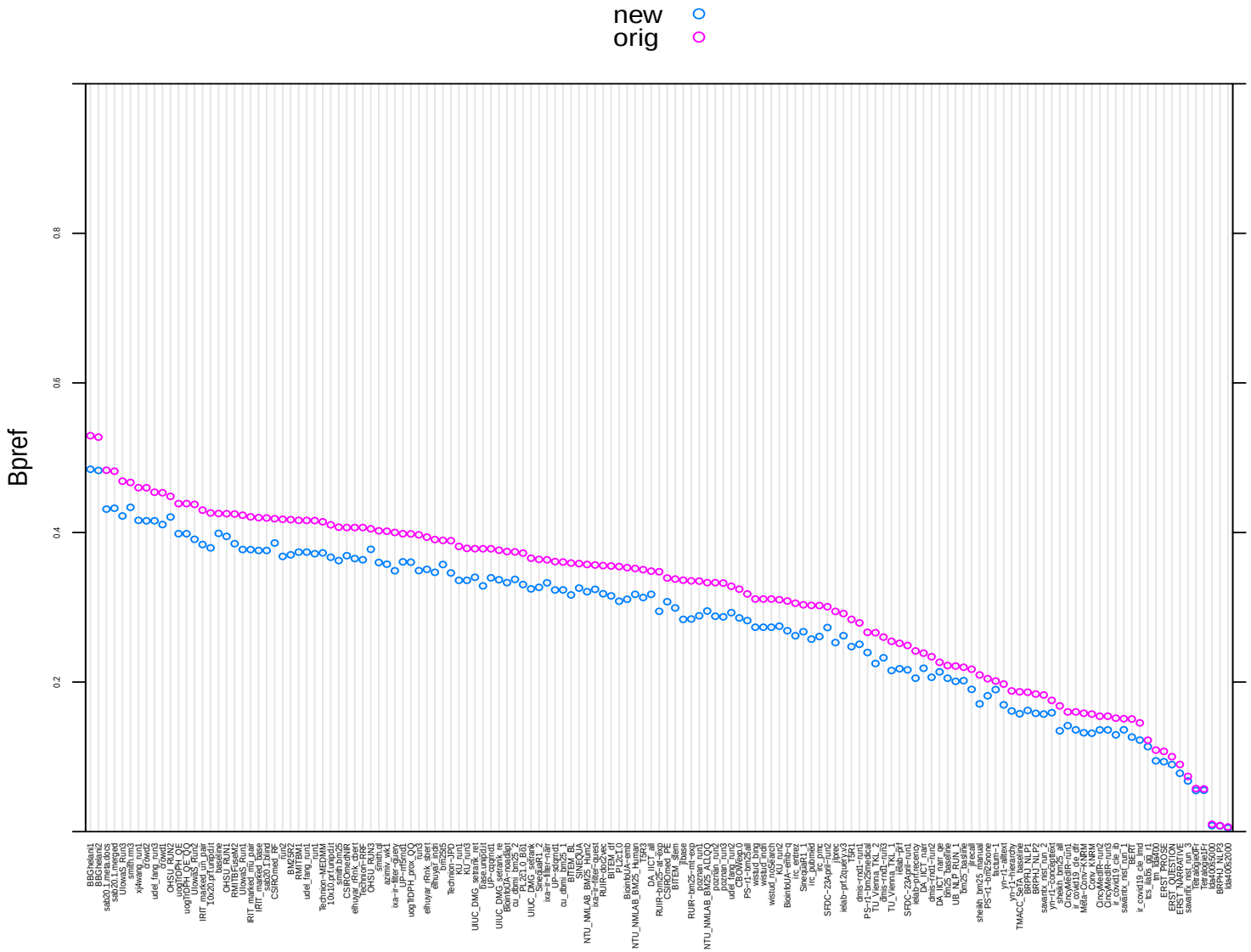

Figure 3: Mean bpref score per Round 1 run as computed using the Extended qrels ("new" in blue) and the Round 1 qrels ("orig" in red). Runs are sorted by decreasing mean bpref as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9638.

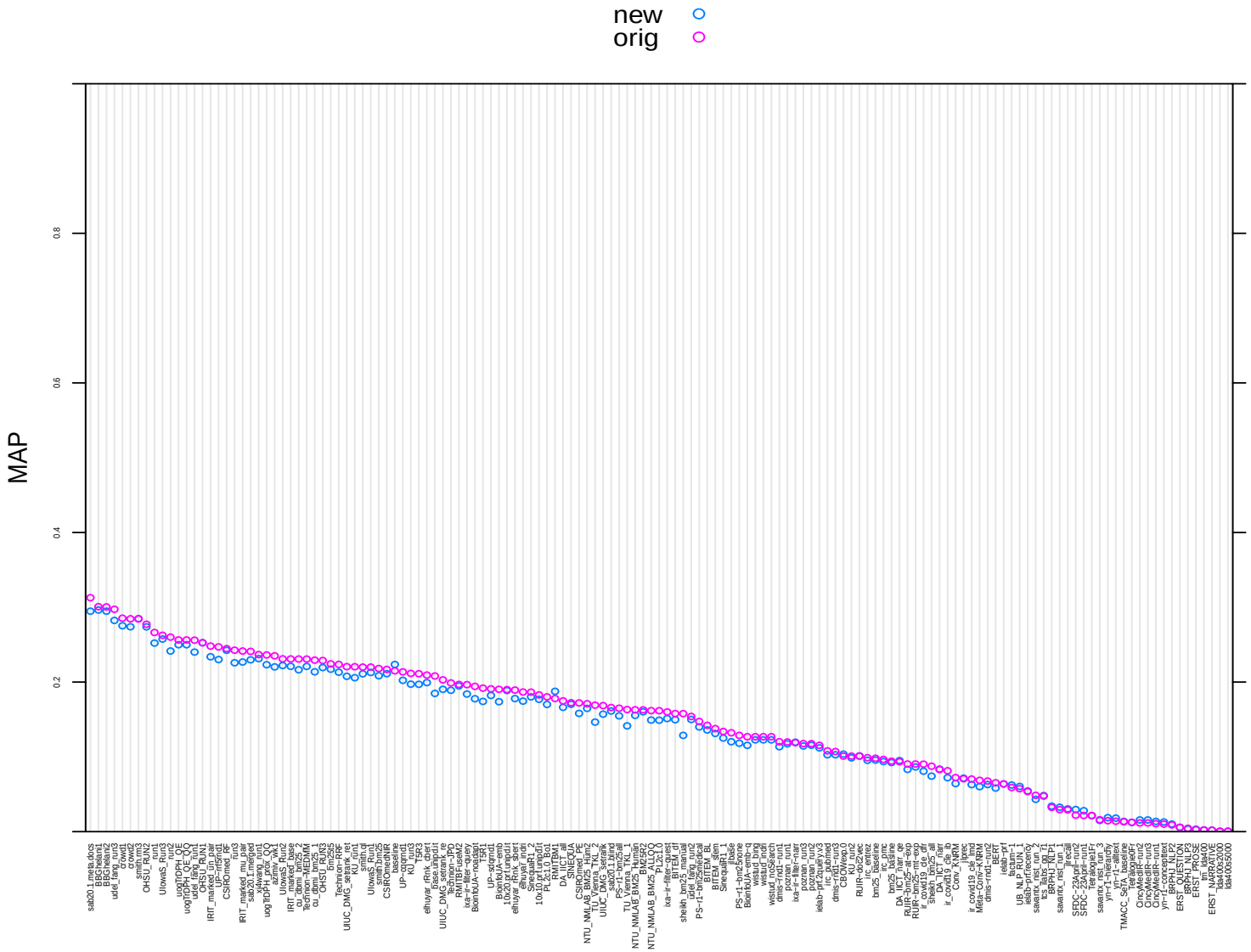

Figure 4: MAP score per Round 1 run as computed using the Extended qrels ("new" in blue) and the Round 1 qrels ("orig" in red). Runs are sorted by decreasing MAP as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9636.