Effect on System Rankings of Further Extending Pools for TREC-COVID Round 1 Submissions

Ellen Voorhees
National Institute of Standards and Technology

June 25, 2020

This paper updates the effect of additional relevance judgments on TREC-COVID Round 1 system rankings by using the judgments from TREC-COVID judgment sets 0.5–2.0 to score Round 1 submissions. The documents judged in Round 2.0 were selected from TREC-COVID Round 2 submissions. This means that judgment set 2.0 is different from the earlier judgment sets just in that it was a different set of runs, but more importantly because many of those runs were feedback runs incorporating prior judgments. It is therefore likely that some of the additional relevant documents are materially different (in terms of content) from the earlier set of relevant documents. The effect on the original system rankings of Round 1 are nonetheless fairly minimal. There are a few runs that do evaluate much better using these new judgments, but Kendall τ scores of system rankings for each of four measures examined (P@5, NDCG@10, bpref, MAP) remain around 0.95.

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.

TREC-COVID Round 2 submissions were scored using the union of Judgment Sets 1.5 and 2.0. Judgment Set 2.0 was created from TREC-COVID Round 2 submissions, using pools to depth 7 (for topics 1–30) of the top priority run from each group as well as from both baseline runs submitted by the "anserini" group. Documents that had been judged in round 1.5 were removed from this set, leaving approximately 4500 documents across the 30 topics.

Table 1 shows the distribution of judgments across topics for the Round 1 qrels, for qrels formed by the union of sets 0.5–1.5, and for the qrels formed from the union of sets 0.5–2.0, restricted to topics 1–30 and including only documents contained in the Round 1 document set. 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.

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

Table 1: Counts of total numbers of judged documents and number of relevant documents per topic in cumulative qrels. All qrels are restricted to documents and topics from Round 1. Percent relevant is the fraction of judged documents that are some form of relevant. Round 1 qrels consist of documents judged in judgment rounds 0.5 and 1.0.

	Round 1 qrels				Qrels 0.5–1.5				Qrels 0.5–2.0			
	Num	Part		%	Num	Part		%	Num	Part		%
Topic	Judged	Rel	Rel	Rel	Judged	Rel	Rel	Rel	Judged	Rel	Rel	Rel
1	323	45	56	0.313	548	71	52	0.285	629	101	101	0.321
2	284	21	26	0.165	471	31	20	0.130	533	39	34	0.137
3	337	66	24	0.267	573	117	102	0.260	652	140	33	0.265
4	357	32	27	0.165	589	51	38	0.154	653	60	42	0.156
5	336	35	96	0.390	548	53	36	0.336	610	56	142	0.325
6	321	80	83	0.508	537	133	106	0.514	594	153	154	0.517
7	275	2	47	0.178	475	2	0	0.124	554	3	71	0.134
8	360	46	30	0.211	610	79	66	0.193	688	89	43	0.192
9	298	25	16	0.138	521	27	4	0.088	608	28	21	0.081
10	191	35	50	0.445	317	46	22	0.375	337	47	75	0.362
11	344	67	5	0.209	576	73	12	0.135	660	75	5	0.121
12	324	76	126	0.623	519	95	38	0.563	587	105	222	0.557
13	373	97	49	0.391	617	156	118	0.358	718	191	67	0.359
14	222	24	5	0.131	373	38	28	0.126	452	43	13	0.124
15	348	45	12	0.164	602	76	62	0.159	738	86	22	0.146
16	340	42	11	0.156	563	72	60	0.160	652	81	21	0.156
17	243	32	45	0.317	385	42	20	0.247	430	45	55	0.233
18	267	79	32	0.416	421	101	44	0.361	456	108	55	0.357
19	301	27	16	0.143	508	39	24	0.120	582	43	28	0.122
20	247	41	25	0.267	412	84	86	0.274	458	105	30	0.295
21	319	15	70	0.266	537	22	14	0.227	600	24	114	0.230
22	259	17	30	0.181	457	25	16	0.136	503	31	37	0.135
23	256	4	22	0.102	429	6	4	0.070	473	14	28	0.089
24	249	14	19	0.133	410	20	12	0.095	457	29	19	0.105
25	308	9	62	0.231	509	11	4	0.171	595	11	92	0.173
26	312	19	106	0.401	501	26	14	0.325	563	30	159	0.336
27	300	30	44	0.247	492	82	104	0.293	538	100	67	0.310
28	180	9	29	0.211	294	11	4	0.136	320	15	29	0.138
29	218	42	58	0.459	355	65	46	0.425	404	73	92	0.408
30	199	39	16	0.276	312	59	40	0.256	343	59	21	0.233

2 Measures

We evaluated the set of Round 1 runs using each of the four measures Prec@5, NDCG@10, bpref, or MAP, first using the official Round 1 qrels and then using the qrels created from sets 0.5–2.0. Figures 1–4 show the scores of Round 1 runs for each measure and both 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.) A few unjudged runs evaluate much better relative to other runs with the new judgments than with the original set. The largest change in the relative ranking of runs is the RMITBFuseM2 run which rises 33 ranks when using P@5 as the measure (21 ranks by NDCG@10, 7 ranks by map and none for pbref). Similarly, the sab20.1.blind run rises 19 ranks for P@5, 17 for NDCG@10, 4 for MAP and 3 for bpref. The largest change in ranks per measure is 33 for P@5, 21 for NDCG@10, 19 for MAP, and 15 for bpref. Kendall τ scores are nonetheless high, close to 0.95 for each measure.

3 Discussion

While high Kendall τ scores are good, the individual runs with large changes in rank demonstrate that the original qrels does not form as reusable a collection as desired: there exist quality runs that would not be recognized as such using just the Round 1 qrels. This is not particularly surprising since depth-7 pools are shallow pools. Indeed, it is probably more surprising that the original pools are as reusable as they appear to be. TREC-COVID was purposely designed so that judgments for topics would accrete over rounds, and the final test collection's qrels will contain a much better sample of documents than any single round.

The qrels appear to be equally as reusable for measures that are computed over more of the document rankings (such as MAP) as the measures that focus on the very top of the ranking (such as P@5). While this may appear counter-intuitive since deeper measures encounter many more unjudged documents, it is consistent with other findings [1]. The very fact that a measure is computed over many more documents means that it is an inherently more stable measure. When a measure depends on only a very few documents, a change to any one of that small set has a large impact on the score.

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 [2]. 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 fractions greater than 0.33 (see Table 1). For Qrels 0.5–1.5, those same eight topics still have relevant fractions greater than 0.33. For the qrels 0.5–2.0, the topic with relevant fraction closest to 0.33 (topic 5) dips to 0.325, but the remaining seven topics' relevant fraction remains basically unchanged.

References

[1] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 33–40, 2000.

[2] 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.

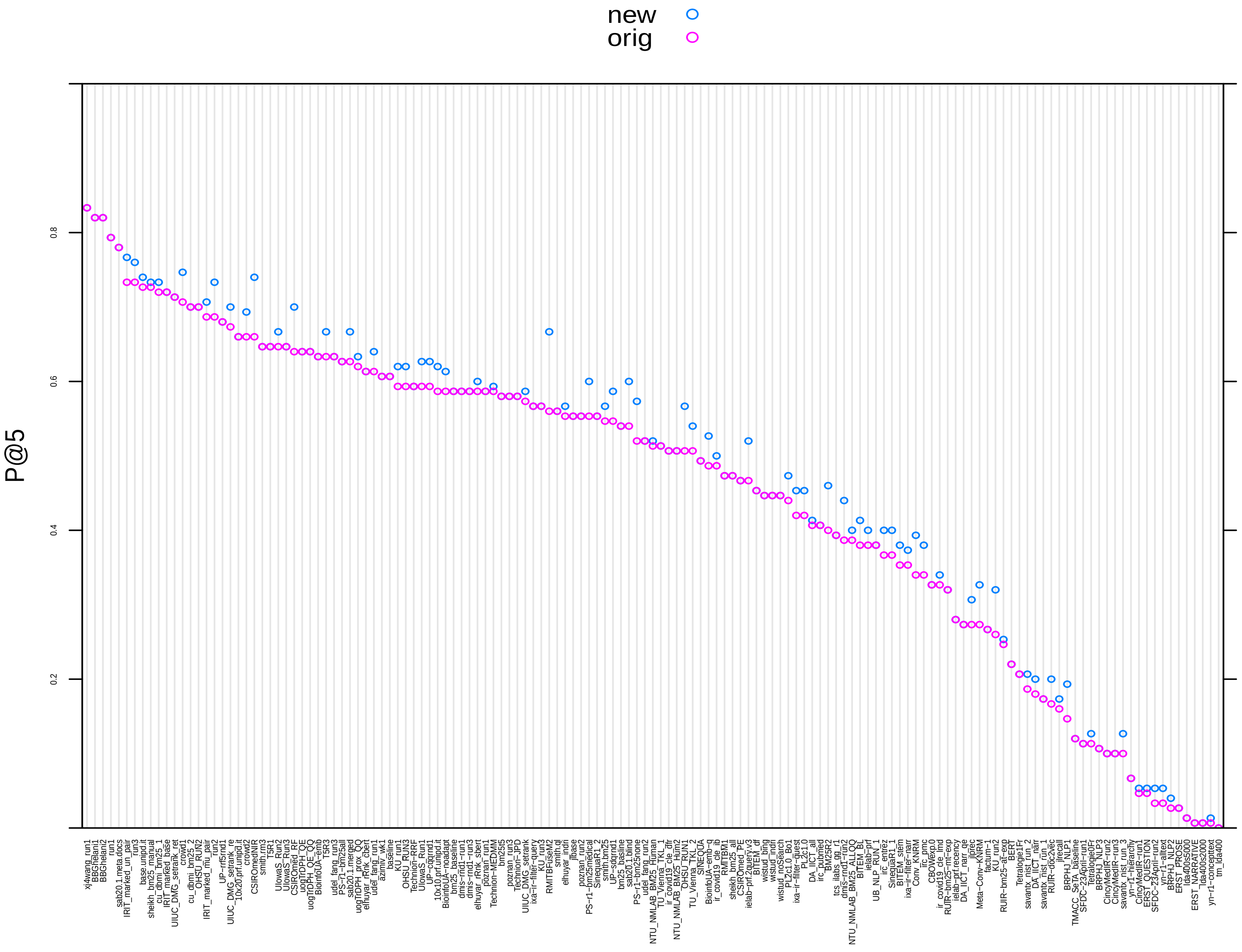

Figure 1: Mean P@5 score per Round 1 run as computed using the Extended (0.5–2.0) qrels ("new" in blue) and the Round 1 qrels ("orig" in red). Runs are sorted by decreasing mean P@5 as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9452. Maximum change in rank is 33 (RMITBFuseM2).

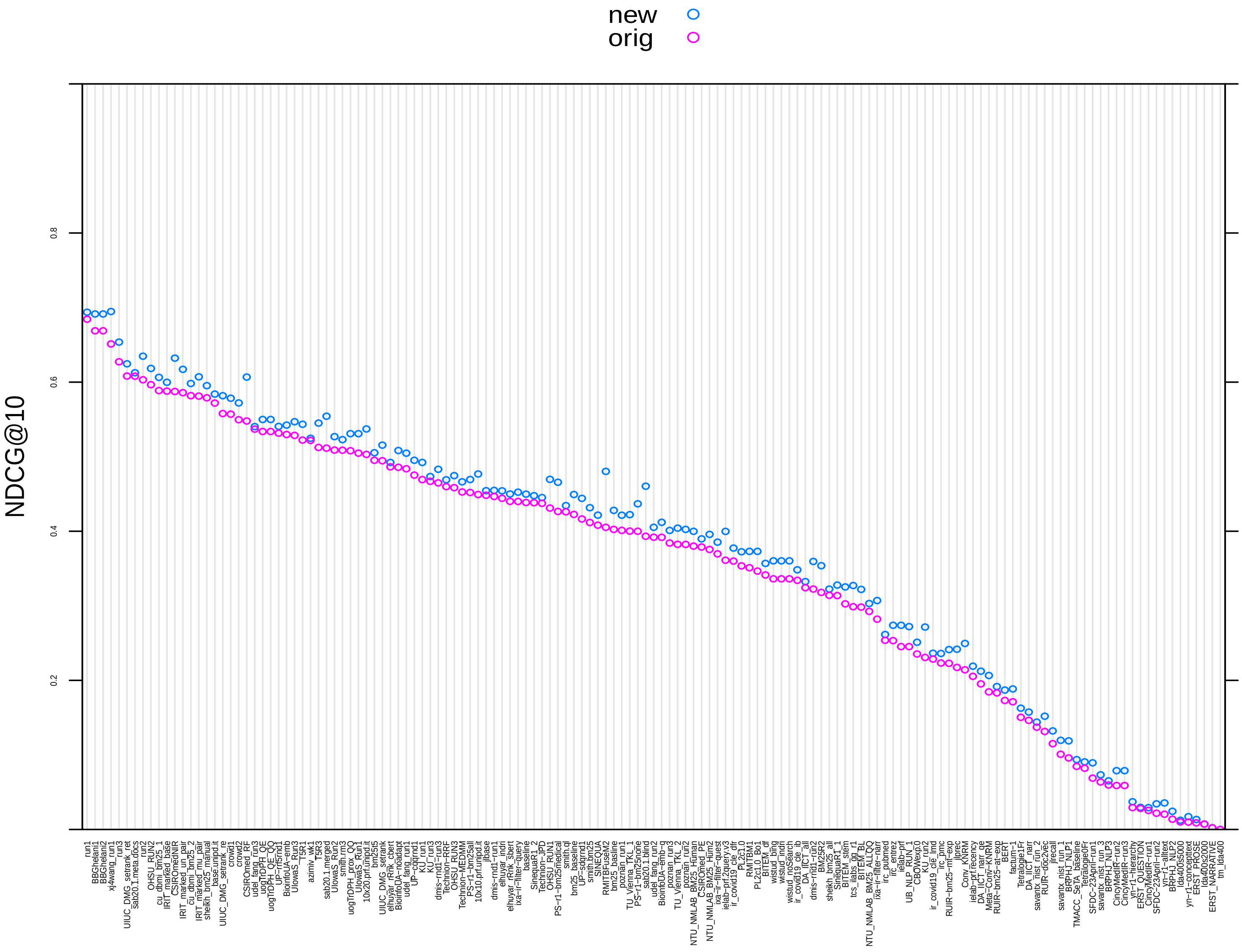

Figure 2: Mean NDCG@10 score per Round 1 run as computed using the Extended (0.5–2.0) qrels ("new" in blue) and the Round 1 qrels ("orig" in red). Runs are sorted by decreasing mean NDCG@10 as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9607. Maximum change in rank is 21 (RMITBFuseM2).

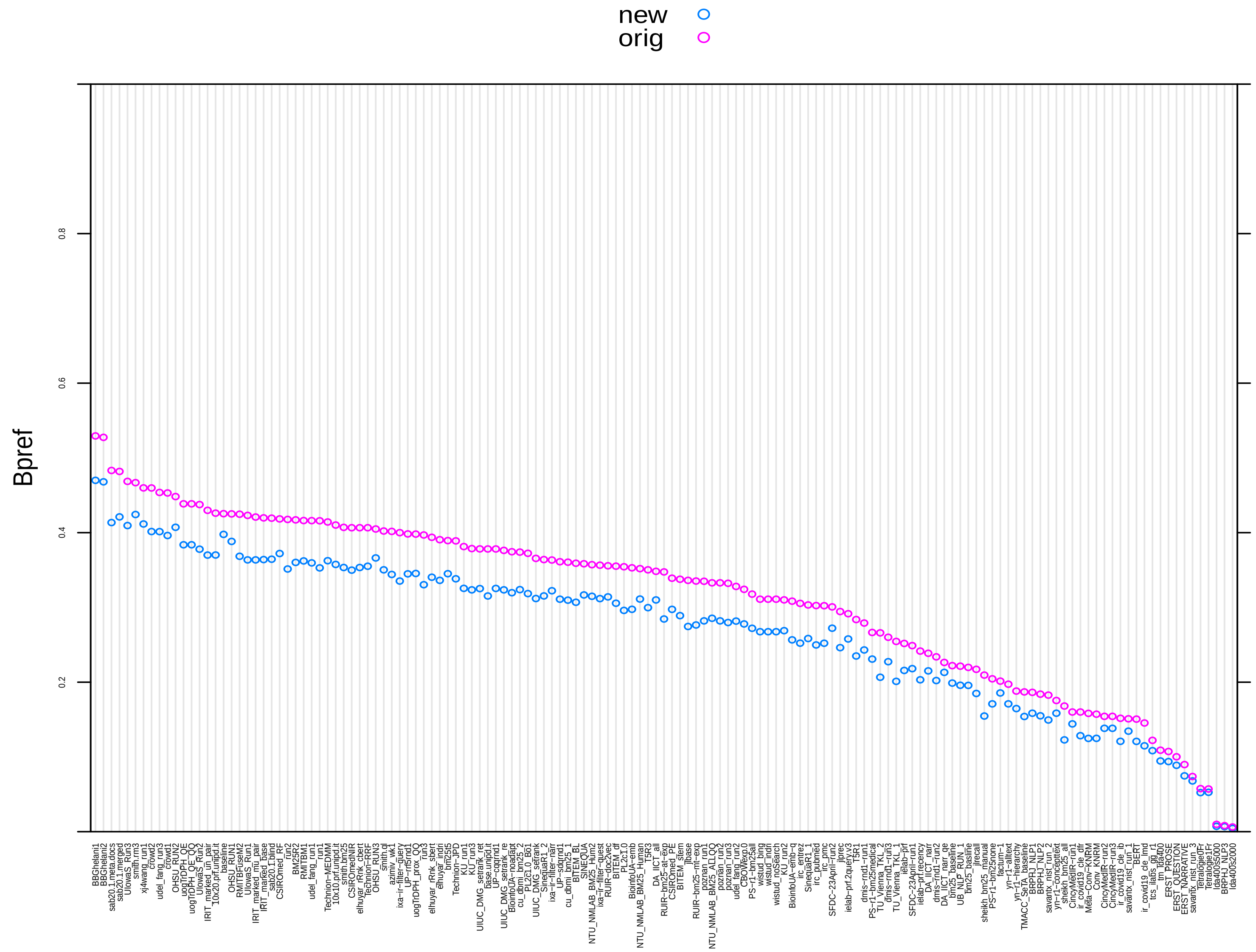

Figure 3: Mean bpref score per Round 1 run as computed using the Extended (0.5–2.0) qrels (“new” in blue) and the Round 1 qrels (“orig” in red). Runs are sorted by decreasing mean bpref as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9552. Maximum change in rank is 15 (OHSU_RUNS3).

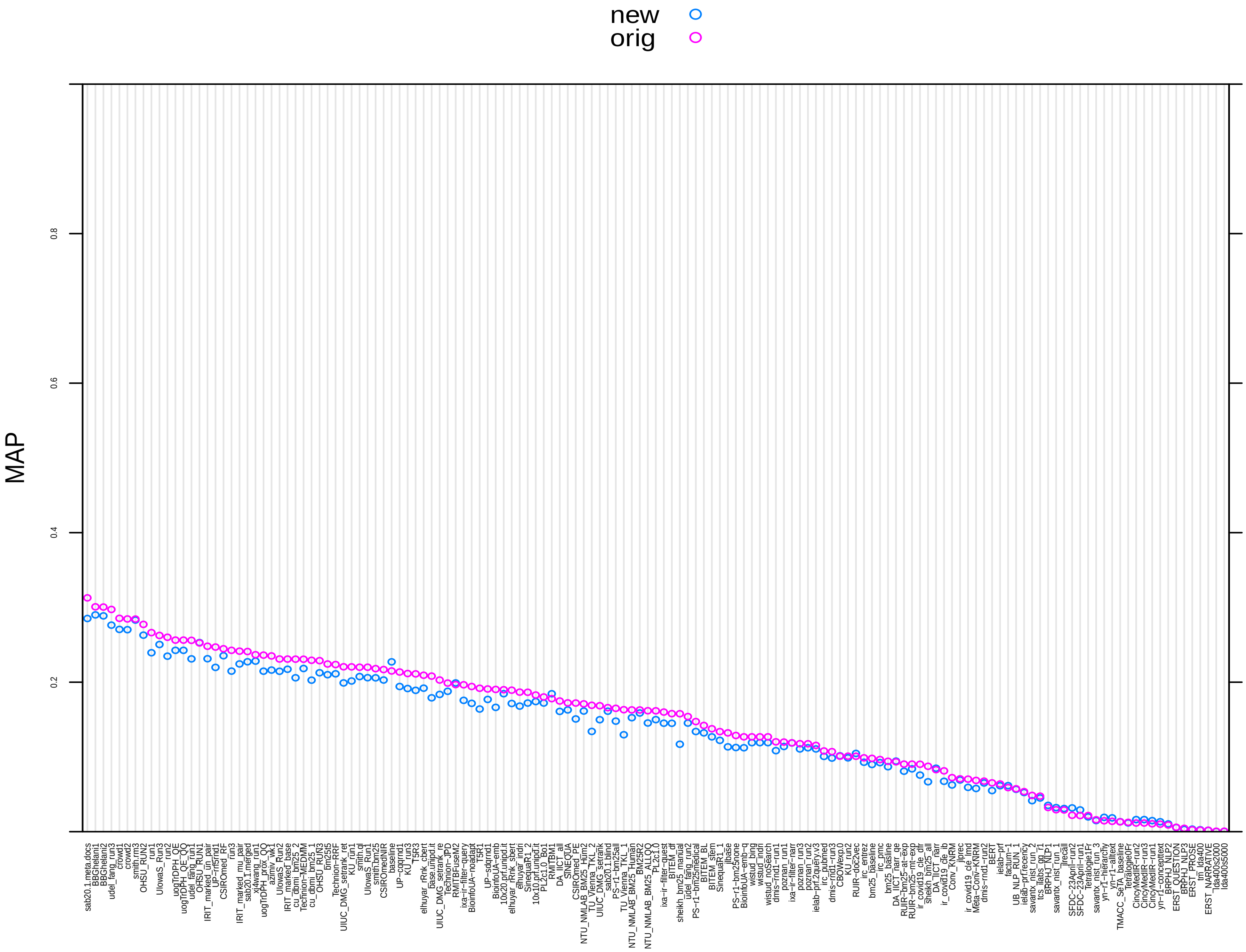

Figure 4: MAP score per Round 1 run as computed using the Extended (0.5–2.0) qrels (“new” in blue) and the Round 1 qrels (“orig” in red). Runs are sorted by decreasing MAP as computed using the Round 1 qrels. The Kendall τ of the run rankings using mean score from different qrels is 0.9505. Maximum change in rank is 19 (baseline).