The Effect of Inter-Assessor Disagreement on IR System Evaluation: A Case Study with Lancers and Students

Tetsuya Sakai
Waseda University
tetsuyasakai@acm.org

ABSTRACT

This paper reports on a case study on the inter-assessor disagreements in the English NTCIR-13 We Want Web (WWW) collection. For each of our 50 topics, pooled documents were independently judged by three assessors: two "lancers" and one Waseda University student. A lancer is a worker hired through a Japanese part time job matching website, where the hirer is required to rate the quality of the lancer's work upon task completion and therefore the lancer has a reputation to maintain. Nine lancers and five students were hired in total; the hourly pay was the same for all assessors. On the whole, the inter-assessor agreement between two lancers is statistically significantly higher than that between a lancer and a student. We then compared the system rankings and statistical significance test results according to different qrels versions created by changing which asessors to rely on: overall, the outcomes do differ according to the qrels versions, and those that rely on multiple assessors have a higher discriminative power than those that rely on a single assessor. Furthermore, we consider removing topics with relatively low inter-assessor agreements from the original topic set: we thus rank systems using 27 high-agreement topics, after removing 23 low-agreement topics. While the system ranking with the full topic set and that with the high-agreement set are statistically equivalent, the ranking with the high-agreement set and that with the low-agreement set are not. Moreover, the low-agreement set substantially underperforms the full and the high-agreement sets in terms of discriminative power. Hence, from a statistical point of view, our results suggest that a high-agreement topic set is more useful for finding concrete research conclusions than a low-agreement one.

CCS CONCEPTS

•**Information systems → Retrieval effectiveness;**

KEYWORDS

inter-assessor agreement; *p*-values; relevance assessments; statistical significance

1 INTRODUCTION

While IR researchers often view laboratory IR evaluation results as something *objective*, at the core of any laboratory IR experiments lie the relevance assessments, which are the result of *subjective* judgements of documents by a person, or multiple persons, based on a particular (intepretation of an) information need. Hence it is of utmost importance for IR researchers to understand the effects of the subjective nature of the relevance assessment process on the final IR evaluation results.

This paper reports on a case study on the inter-assessor disagreements in a recently-constructed ad hoc web search test collection, namely, the English NTCIR-13 *We Want Web* (WWW) collection [10]. For each of our 50 topics, pooled documents were independently judged by three assessors: two "lancers" and one Waseda University student. A lancer is a worker hired through a Japanese part time job matching website[1], where the hirer is required to rate the quality of the lancer's work upon task completion and therefore the lancer has a reputation to maintain[2]. Nine lancers and five students were hired in total; the hourly pay was the same for all assessors. On the whole, the inter-assessor agreement between two lancers is statistically significantly higher than that between a lancer and a student (Section 3). We then compared the system rankings and statistical significance test results according to different qrels versions created by changing which asessors to rely on: overall, the outcomes do differ according to the qrels versions, and those that rely on multiple assessors have a higher *discriminative power* (i.e., the ability to obtain many statistically significant system pairs [14, 15]) than those that rely on a single assessor (Section 4.1). Furthermore, we consider removing topics with relatively low inter-assessor agreements from the original topic set: we thus rank systems using 27 high-agreement topics, after removing 23 low-agreement topics. While the system ranking with the full topic set and that with the high-agreement set are statistically equivalent, the ranking with the high-agreement set and that with the low-agreement set are not. Moreover, the low-agreement set substantially underperforms the full and the high-agreement sets in terms of discriminative power (Section 4.2). Hence, from a statistical point of view, our results suggest that a high-agreement topic set is more useful for finding concrete research conclusions than a low-agreement one.

2 RELATED WORK/NOVELTY OF OUR WORK

Studies on the effect of inter-assessor (dis)agreement on IR system evaluation have a long history; Bailey *et al.* [2] provides a concise survey on this topic covering the period 1969-2008. More recent work in the literature includes Carterette and Soboroff [4], Webber, Chandar, and Carterette [21], Demeester *et al.* [5] Megorskaya, Kukushkin, and Serdyukov [12], Wang *et al.* [20], Ferrante, Ferro, and Maistro [6], and Maddalena *et al.* [11]. Among these studies, the work of Voorhees [19] from 2000 (or the earlier version reported at SIGIR 1998) is probably one of the most well-known; below, we first

[1] http://www.lancers.jp/ (in Japanese). See also https://www.techinasia.com/lancers-produces-200-million-freelancing-gigs-growing (in English)
[2] The lancer then rates the hirer; therefore the hirer also has a reputation to maintain on the website.

EVIA 2017, co-located with NTCIR-13, Tokyo, Japan.

highlight the differences between her work and the present study, since the primary research question of the present study is whether her well-known findings generalise to our new test collection with experimental settings that are quite different from hers in several ways. After that, we also briefly compare the present study with the recent, closely-related work of Maddalena *et al.* [11] from ICTIR 2017.

Voorhees [19] examined the effect of using different qrels versions on ad hoc IR system evaluation. Her experiments used the TREC-4 and TREC-6 data[3]. In particular, in her experiments with the 50 TREC-4 topics, she hired two additional assessors in addition to the primary assessor who created the topic, and discussed the pairwise inter-assessor agreement in terms of *overlap* as well as recall and precision: overlap is defined as the size of the intersection of two relevant sets divided by the size of the union; recall and precison are defined by one of the relevant set as the gold data. However, it was not quite the case that the three assessors judged the same document pool independently: the document sets provided to the additional assessors were created *after* the primary assessment, by mixing both relevant and nonrelevant documents from the primary assessor's judgements. Moreover, all documents judged relevant by the primary assessor but not included in the document set for the additional assessors were counted towards the set intersection when computing the inter-assessor agreement. Her experiments with the TREC-6 experiments relied on a different setting, where University of Waterloo created their own pools and relevance assessments independent of the original pools and assessments. She considered binary relevance only[4], and therefore she considered Average Precision and Recall at 1000 as effectiveness evaluation measures. Her main conclusion was: "*The actual value of the effectiveness measure was affected by the different conditions, but in each case the relative performance of the retrieved runs was almost always the same. These results validate the use of the TREC test collections for comparative retrieval experiments.*"

The present study differs from that of Voorhees in the following aspects at least:

- We use a new English *web search* test collection contructed for the NTCIR-13 WWW task, with depth-30 pools.
- For each of our 50 topics, the same pool was completely independently judged by three assessors. Nine assessors were hired through the lancers website, and an additional five assessors were hired at Waseda University, so that each topic was judged by two lancers and one student.
- We collected *graded* relevance assessments from each assessor: highly relevant (2 points), relevant (1 point), nonrelevant (0) and error (0) for cases where the web pages to judge could not be displayed. When consolidating the multiple assessments, the raw scores were added to form more fine-grained graded relevance data.
- We use graded relevance measures at cutoff 10 (representing the quality of the first search engine result page), namely *nDCG@10*, *Q@10*, and *nERR@10* [17], which are the official measures of the WWW task.
- As our topics were sampled from a query log, none of our assessors are the topic originators (or "primary" [19] or "gold" assessors [2]); The assessors were *not* provided with any information other than the query (e.g., a narrative field [1, 8]): the definition for a highly relevant document was: "*it is* likely *that the user who entered this search query will find this page relevant*"; that for a relevant document was: "*it is* possible *that the user who entered this search query will find this page relevant* " [10].
- We discuss inter-assessor agreeement and system ranking agreement using stastical tools, namely, linear weighted κ with 95%CIs (which, unlike raw overlap measures, takes chance agreement into account [2]) and Kendall's τ with 95%CIs. Moreover, we employ the randomised Tukey HSD test [3, 16] to discuss the discrepancies in statistical significance test results. Furthermore, we consider *removing* topics that appear to be unreliable in terms of inter-assessor agreement.

While the recent work of Maddalena *et al.* [11] addressed several research questions related to inter-assessor agreement, one aspect of their study is closely related to our analysis with high-agreement and low-agreement topic sets. Maddalena *et al.* utilised the TREC 2010 Relevance Feedback track data and exactly five different relevance assessments for each ClueWeb document, and used Krippendorph's α [9] to quantify the inter-assessor agreement. They defined high-agreement and lowagreeement topics based on Krippendorph's α, and reported that high-agreement topics can predict the system ranking with the full topic set more accurately than low-agreement topics. The analysis in the present study differs from the above as discussed below:

- Krippendorph's α disregards which assessments came from which assessors, as it is a measure of the overall reliability of the data. In the present study, where we only have three assessors, we are more interested in the agreement between every pair of assessors and hence utilise Cohen's linear weighted κ. Hence our definition of a high/low-agreement topic differs from that of Maddalena *et al.*: according to our definition, a topic is in high agreement if the κ is statistically significantly positive for every pair of assessors.
- While Maddalena *et al.* discussed the absolute effectiveness scores and system rankings only, the present study discusses statistical significance testing after replacing the full topic set with the high/low-agreement set.
- Maddalena *et al.* focussed on nDCG; we discuss the three aforementioned official measures of the NTCIR-13 WWW task [10].

3 DATA

The NTCIR-13 WWW English subtask created 100 test topics; 13 runs were submitted from three teams. We acknowledge that this is a clear limitation of the present study: we would have liked a larger number of runs from a larger number of teams. However, we claim that this limitation does not invalidate neither our approach to analysing inter-assessor disagreement nor the actual results on the system ranking and statistical significance. We hope to repeat

[3] The document collections are: disks 2 and 3 for TREC-4; disks 4 and 5 for TREC-6 [8].
[4] The original Waterloo assessments on a tertiary scale, but were collapsed into binary for her analysis.

Table 1: Pairwise inter-assessor agreement: lancer1, lancer2**, and** student**.**

	(a) lancer1 vs. lancer2				(b) lancer1 vs. student				(c) lancer2 vs. student			
		0	1	2		0	1	2		0	1	2
raw scores	0	3991	1354	487	0	3406	1540	886	0	3215	1232	938
	1	947	1260	882	1	1051	1100	938	1	1203	1415	1043
	2	447	1047	799	2	416	787	1090	2	455	780	933
Linear weighted Cohen's κ with 95%CI	0.336 [0.322, 0.351]				0.283 [0.268, 0.298]				0.261 [0.246, 0.276]			
Mean (min/max) per-topic linear weighted Cohen's κ	0.293 (−0.007/0.683)				0.211 (−0.128/0.583)				0.225 (−0.054/0.918)			
#Topics where per-topic κ is not statistically significantly positive	8				15				19			
Binary Cohen's κ with 95%CI	0.424 [0.407, 0.441]				0.309 [0.292, 0.327]				0.314 [0.296, 0.331]			
Binary raw agreement	0.712				0.653				0.659			

Table 2: Number of *Lx*-relevant documents in each qrels.

	*L*0	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	total judged
all3	2,603	1,897	2,135	1,535	1,537	1,035	472	11,214
2lancers	3,991	2,301	2,194	1,929	799	-	-	11,214
lancer1	5,832	3,089	2,293	-	-	-	-	11,214
lancer2	5,385	3,661	2,168	-	-	-	-	11,214
student	4,873	3,427	2,914	-	-	-	-	11,214

the same analysis on a larger set of runs in the next round of the WWW task.

For evaluating the 13 runs submitted to the WWW task, we created a depth-30 pool for each of the 100 topics and this resulted in a total of 22,912 documents to judge. We hired nine lancers who speak English through the lancers website: the job call and the relevance assessment instructions were published on the website in English. None of them had any prior experience in relevance assessments. Topics were assigned at random to the nine assessors so that each topic had two independent judgments from two lancers. The official relevance assessments of the WWW task were formed by consolidating the two lancer scores: since each lancer gave 0, 1, or 2, the final relevance levels were *L*0-*L*4.

For the present study, we focus on a subset of the above test set, which contains 50 topics whose topic IDs are odd numbers. The number of pooled documents for this topic set is 11,214. We then hired five students from the Department of Computer Science and Engineering, Waseda University, to provide a third set of judgments for each of the 50 topics. The instructions given to them were identical to those given to the lancers. The students also did not have any prior experience in relevance assessments. Moreover, lancers and students all received an hourly pay of 1,200 Japanese Yen. However, hiring lancers is more expensive, because we have to pay about 20% to Lancers the company on top of what we pay to the individual lancers. The purpose of collecting the third set of assessments was to compare the lancer-lancer inter-assessor agreement with the lancer-student agreement, which should shed some light on the reliability of the different assessor types. All of the assessors completed the work in about one month.

It should be noted that all of our assessors are "bronze" according to the definition by Bailey *et al.* [2]: they are neither topic originators nor topic experts.

To quantify inter-assessor agreement, we compute Cohen's linear weighted κ for every assessor pair, where, for example, the weight for a $(0, 2)$ disagreement is 2 and that for a $(0, 1)$ or $(1, 2)$ disagreement is 1[5]. It should be noted that κ represents how much agreement there is *beyond chance*.

Table 1 summarises the inter-assessor agreement results. The "raw scores" section shows the 3×3 confusion matrices for each pair of assessors; the counts were summed across topics, although lancer1, lancer2, and student are actually not single persons. The linear weighted κ's were computed based on these matrices. It can be observed that the lancer-lancer κ is statistically significantly higher than the lancer-student κ's, which means that the lancers agree with each other more than they do with students. While the lack of gold data prevents us from concluding that lancers are more reliable than students, it does suggest that lancers are worth hiring if we are looking for high inter-assessor agreement. As we shall see in Section 4.1, the discriminative power results discussed in also support this observation.

We also computed *per-topic* linear weighted κ's so that the assessments of exactly three *individuals* are compared against one another: the mean, minimum and maximum values are also provided in Table 1. It can be observed that the lowest per-topic κ observed is -0.128, for "lancer1 vs. student"; the 95%CI for this instance was $[-0.262, 0.0006]$, suggesting the lack of agreement beyond chance[6].

Table 1 also shows the number of topics for which the per-topic κ's were *not* statistically significantly positive, that is, the 95%CI lower limits were not positive, as exemplified by the above instance. These numbers indicate that the lancer-lancer agreements were statistically significantly positive for $50 - 8 = 42$ topics while the lancer-student agreements were statistically significantly positive for only 35 (31) topics. Again, the lancers agree with each other more than they do with students.

[5] Fleiss' κ [7], designed for more than two assessors, is applicable to *nominal* categories only; the same goes for Randoph's κ_{free} [13]; see also our discussion on Krippendorff's α [9] in Section 2.

[6] It should be noted that negative κ's are not unsual in the context of inter-assessor agreement: for example, according to a figure from Bailey *et al.* [2], when a "gold" assessor (i.e., top originator) was compared with a bronze asessor, a version of κ was in the $[-0.6, -0.4]$ range for one topic, despite the fact that the assessors must have read the *narrative* fields of the TREC Enterprise 2007 test collection [1].

There were 27 topics where all three per-topic κ's were statistically significantly positive; for the remaining 23 topics, at least one per-topic κ was not. We shall refer to the set of the former 27 topics as the high-agreement set and the latter as the low-agreement set. We shall utilise these subsets in Section 4.2.

Also in Table 1, the binary Cohen's κ row shows the κ values after collapsing the 3×3 matrices into 2×2 matrices by treating highly relevant and relevant as just relevant. Again, the lancer-lancer κ is statistically significantly higher than the lancer-student κ's. Finally, the table shows the raw agreement based on the 2×2 confusion matrices: the counts of $(0,0)$ and $(1,1)$ are divided by those of $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. It can be observed that only the lancer-lancer agreement exceeds 70%.

We summed the raw scores of lancer1, lancer2, and student to form a qrels set which we call all3; we also summed the raw scores of lancer1 and lancer2 to form a qrels set which we call 2lancers. Table 2 shows the distribution of documents across the relevance levels. Note that all3 and 2lancers are on 7-point and 5-point scales, respectively, while the others are on a 3-point scale. In this way, we preserve the views of individual assessors instead of collapsing the assessments into binary or to force them to reach a consensus. Note that nDCG@10, Q@10, and nERR@10 can fully utilise the rich relevance assessments. As we shall see in the next section, this approach to combining multiple relevance assessments is beneficial.

For alternatives to simply summing up the raw assessor scores, we refer the reader to Maddalena *et al.* [11] and Sakai [18]: these approaches are beyond the scope of the present study.

4 RESULTS AND DISCUSSIONS

4.1 Different Qrels Versions

The previous section showed that assessors do disagree, and that the lancers agree with each other more than they do with students. This section investigates the effect of inter-assessor disagreement on system ranking and statistical significance through comparisons across the five qrels versions: all3, 2lancers, lancer1, lancer2, and student.

4.1.1 System Ranking. Figure 1 visualises the system rankings and the actual mean effectiveness scores according to the five different qrels, for nDCG@10, Q@10, and nERR@10. In each graph, the runs have been sorted by the all3 scores, and therefore *if* every curve is monotonically decreasing, that means all the qrels versions produce system rankings that are identical to the one based on all3. First, it can be observed that the absolute effectiveness scores differ depending on the qrels version used, just as Voorhees [19] observed with Average Precision and Recall@1000. Second, and more importantly, the five system rankings are *not* identical: for example, in Figure 1(c), the top performing run according to nERR@10 with all3 is only the fifth best run according to the same measure with lancer1. The nDCG@10 and Q@10 curves are relatively consistent across the qrels versions. Table 3 quantifies the above observation in terms of Kendall's τ, with 95%CIs: while the CI upper limits show that all the rankings are *statistiallly* equivalent, the widths of the CIs due to the small sample size (13 runs) suggest that the results should be viewed with caution. For example, the τ for the aforementioned case of all3 vs. lancer1 with nERR@10 is 0.821

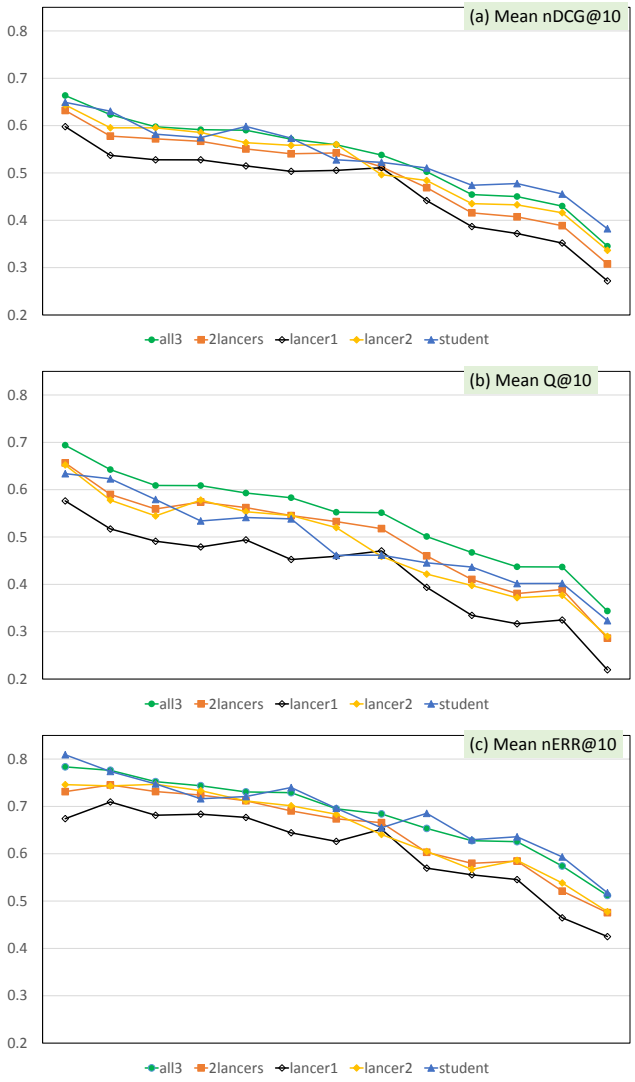

Figure 1: Mean effectiveness scores according to different qrels. The x-axis represents the run ranking according to the all3 qrels.

(95%CI[0.616, 1.077]). The *actual* system swaps that Figure 1 shows are probably more important than these summary statistics.

4.1.2 Statistically Significant Differences across Systems. The next and perhaps more important question is: how do the different qrels versions affect pairwise statistical significance test results? If the researcher is interested in the difference between every system pair, a proper *multiple comparison procedure* should be employed to ensure that the *familywise error rate* is bounded above by the significance criterion α [3]. In this study we use the distribution-free randomised Tukey HSD test using the `Discpower` tool[7], with $B = 10,000$ trials [16]. The input to the tool is a topic-by-run score matrix: in our case, for every combination of evaluation measure and qrels version, we have a 50×13 matrix.

[7] http://research.nii.ac.jp/ntcir/tools/discpower-en.html

Table 3: System ranking consistency in term of Kendall's τ, with 95%CIs, for every pair of qrels (13 runs).

(a) Mean nDCG@10				
	2lancers	lancer1	lancer2	student
all3	0.974 [0.696, 1.048]	0.923 [0.660, 1.032]	0.949 [0.660, 1.032]	0.923 [0.694, 1.035]
2lancers	-	0.949 [0.894, 1.054]	0.974 [0.754, 1.092]	0.897 [0.882, 1.053]
lancer1	-	-	0.923 [0.822, 1.075]	0.846 [0.953, 1.034]
lancer2	-	-	-	0.872 [0.815, 1.069]
(b) Mean Q@10				
	2lancers	lancer1	lancer2	student
all3	0.923 [0.660, 1.018]	0.846 [0.701, 1.028]	0.872 [0.835, 1.049]	0.897 [0.606, 1.019]
2lancers	-	0.872 [0.784, 1.062]	0.949 [0.660, 1.032]	0.872 [0.683, 1.061]
lancer1	-	-	0.821 [0.683, 1.061]	0.897 [0.842, 1.055]
lancer2	-	-	-	0.821 [0.585, 1.056]
(c) Mean nERR@10				
	2lancers	lancer1	lancer2	student
all3	0.923 [0.498, 1.040]	0.821 [0.616, 1.077]	0.923 [0.637, 1.056]	0.872 [0.540, 1.050]
2lancers	-	0.846 [0.784, 1.062]	0.949 [0.585, 1.056]	0.846 [0.784, 1.062]
lancer1	-	-	0.795 [0.673, 1.019]	0.692 [0.822, 1.075]
lancer2	-	-	-	0.846 [0.590, 1.000]

Table 4 shows the results of comparing the outcomes of significance test results at $\alpha = 0.05$: for example, Table 4(a) shows that, in terms of nDCG@10, all3 obtained $2 + 29 = 31$ statistically significantly different run pairs, while 2lancers obtained $29 + 2 = 31$, and that the two qrels versions had 29 pairs in common. Thus the *Statistical Significance Overlap* (SSO) is $29/(2 + 29 + 2) = 87.9\%$. The table shows that the student results disagree relatively often with the others: for example, in Table 4(b), Q@10 with lancer1 have 11 statistically significantly different run pairs that are not statistically significantly different according to the same measure with student, while the opposite is true for five pairs. The two qrels versions have 15 pairs in common and the SSO is only 48.4%. Thus, different qrels versions can lead to different research conclusions.

Table 5 shows the number of statistically significantly different run pairs (i.e., *discriminative power* [14, 15]) deduced from Table 4. It can be observed that combining multiple assessors' labels and thereby having fine-grained relevance levels can result in high discriminative power, and also that student underperforms the others in terms of discriminative power. Thus, it appears that student is not only *different* from the two lancers: they fail to provide many significantly different pairs.

4.2 Using Reliable Topics Only

In Section 3, we defined a high-agreement set containing 27 topics and a low-agreement set containg 23 topics. A high-agreement topic is one for which every assessor pair "statistically agreed," in the sense that the 95%CI lower limit of the per-topic κ was positive. A low-agreement topic is one for which at least one assessor pair did not show any agreement beyond chance, and therefore deemed unreliable. While to the best of our knowledge this kind of close analysis of inter-assessor agreement is rarely done prior to evaluating the submitted runs, removing such topics at an early stage may be a useful practice for ensuring test collection reliability. Hence, in this section, we focus on the all3 qrels, and compare the evaluation outcomes when the full topic set (50 topics) is replaced with just the high-agreement set or even just the low-agreement set. The fact that the high-agreement and low-agreement sets are similar in sample size is highly convenient for comparing them in terms of discriminative power.

4.2.1 System Ranking. Figure 2 visualises the system rankings and the actual mean effectiveness scores according to the three topic sets, for nDCG@10, Q@10, and nERR@10. Again, in each graph, the runs have been sorted by the all3 scores (mean over 50 topics). Table 6 compares the system rankings in terms of Kendall's τ with 95%CIs. The values in bold indicate the cases where the two rankings are statistically not equivalent. It can be observed that while the system rankings by the full set and the high-agreement set are statistically equivalent, those by the full set and the low-agreement set are not. Thus, the properties of the high-agreement topics appear to be dominant in the full topic set.

Table 4: Statistical Significance Overlap between two qrels versions ($\alpha = 0.05$).

(a) Mean nDCG@10				
	2lancers	lancer1	lancer2	student
all3	2/29/2 (87.9%)	4/27/3 (79.4%)	5/26/0 (83.9%)	11/20/0 (64.5%)
2lancers	-	3/28/2 (84.8%)	6/25/1 (78.1%)	12/19/1 (59.4%)
lancer1	-	-	5/25/1 (80.6%)	12/18/2 (56.2%)
lancer2	-	-	-	7/19/1 (70.4%)
(b) Mean Q@10				
	2lancers	lancer1	lancer2	student
all3	3/26/3 (81.2%)	6/23/7 (71.9%)	3/26/2 (83.9%)	11/18/2 (58.1%)
2lancers	-	5/24/2 (77.4%)	4/25/3 (78.1%)	14/15/5 (44.1%)
lancer1	-	-	4/22/6 (68.8%)	11/15/5 (48.4%)
lancer2	-	-	-	11/17/3 (54.8%)
(c) Mean nERR@10				
	2lancers	lancer1	lancer2	student
all3	3/16/1 (80.0%)	5/14/1 (70.0%)	5/14/2 (66.7%)	7/12/0 (63.2%)
2lancers	-	3/14/1 (77.8%)	4/13/3 (65.0%)	8/9/3 (45.0%)
lancer1	-	-	3/12/4 (63.2%)	6/9/3 (50.0%)
lancer2	-	-	-	6/10/2 (55.6%)

Table 5: Number of significantly different run pairs deduced from Table 4 ($\alpha = 0.05$).

	(a) Mean nDCG@10	(b) Mean Q@10	(c) Mean nERR@10
all3	31	29	19
2lancers	31	29	17
lancer1	30	26	15
lancer2	26	28	16
student	20	20	12

4.2.2 Statistically Significant Differences across Systems. Table 7 compares the outcomes of statistical significance test results (Randomised Tukey HSD with $B = 10,000$ trials) across the three topic sets in a way similar to Table 4. Note that the two subsets are inherently less discriminative than the full set as the sample sizes are about half that of the full set. It can be observed that the set of statistically signifantly different pairs according to the high-agreement (low-agreement) set is always a subset of the set of statistically signifantly different pairs according to the full topic set. More interestingly, the set of statistically signifantly different pairs according to the low-agreement set is *almost* a subet of the set of statistically signifantly different pairs according to the high-agreement set: for example, Table 7(b) shows that there is only one system pair for which the low-agreement set obtained a statistically significant difference while the high-agreement set did not in terms of Q@10.

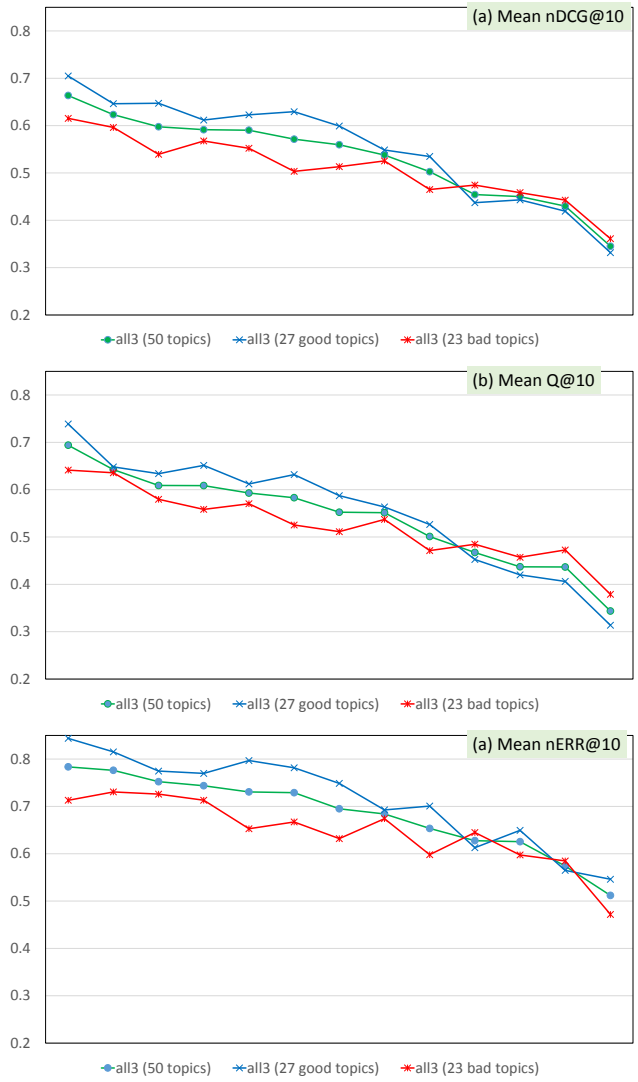

Figure 2: Mean effectiveness scores according to different topic sets. The x-axis represents the run ranking according to all3 **(50 topics).**

Table 8 shows the number of statistically significantly different pairs for each condition based on Table 7. Again, it can be observed that the high-agreement set is substantially more discriminative than the low-agreement set, despite the fact that the sample sizes are similar. Thus, the results suggest that, from a statistical point of view, a high-agreement topic set is more useful for finding concrete research conclusions than a low-agreement one.

5 CONCLUSIONS AND FUTURE WORK

This paper reported on a case study involving only 13 runs contributed from only three teams. Hence we do not claim that our finding will generalise; we merely hope to apply the same methodology to test collections that will be created for the future rounds of the WWW task and possibly even other tasks. Our main findings using the English NTCIR-13 WWW test collection are as follows:

Table 6: System ranking consistency in term of Kendall's τ, with 95%CIs, for every pair of topic sets (13 runs).

(a) Mean nDCG@10		
	all3 (27 high-agreement topics)	all3 (23 low-agreement topics)
all3 (50 topics)	0.872 [0.696, 1.048]	0.846 [0.660, 1.032]
all3 (27 high-agreement topics)	-	**0.718 [0.450, 0.986]**
(b) Mean Q@10		
	all3 (27 high-agreement topics)	all3 (23 low-agreement topics)
all3 (50 topics)	0.923 [0.784, 1.062]	0.846 [0.673, 1.019]
all3 (27 high-agreement topics)	-	**0.769 [0.566, 0.972]**
(c) Mean nERR@10		
	all3 (27 high-agreement topics)	all3 (23 low-agreement topics)
all3 (50 topics)	0.846 [0.660, 1.032]	**0.769 [0.545, 0.994]**
all3 (27 high-agreement topics)	-	**0.615 [0.319, 0.912]**

Table 7: Statistical Significance Overlap between two topic sets ($\alpha = 0.05$).

(a) Mean nDCG@10		
	all3 (27 high-agreement topics)	all3 (23 low-agreement topics)
all3 (50 topics)	4/27/0 (87.1%)	22/9/0 (29.0%)
all3 (27 high-agreement topics)	-	18/9/0 (33.3%)
(b) Mean Q@10		
	all3 (27 high-agreement topics)	all3 (23 low-agreement topics)
all3 (50 topics)	4/25/0 (86.2%)	19/10/0 (34.5%)
all3 (27 high-agreement topics)	-	16/9/1 (34.6%)
(c) Mean nERR@10		
	all3 (27 high-agreement topics)	all3 (23 low-agreement topics)
all3 (50 topics)	4/15/0 (78.9%)	13/6/0 (31.6%)
all3 (27 high-agreement topics)	-	10/5/1 (31.2%)

Table 8: Number of significantly different run pairs deduced from Table 7 ($\alpha = 0.05$).

	(a) Mean nDCG@10	(b) Mean (b) Q@10	(c) Mean (c) nERR@10
all3 (50 topics)	31	29	19
high-agreement (27 topics)	27	25	15
low-agreement (23 topics)	9	10	6

- Lancer-lancer inter-assessor agreements are statistically significantly higher than lancer-student agreements. The student qrels is less discriminative than the lancers qrels. While the lack of gold data prevents us from concluding which type of assessors is more reliable, these results suggest that hiring lancers has some merit despite the extra cost.
- Different qrels versions based on different (combinations of) assessors can lead to somewhat different system rankings and statistical significance test results. Combining multiple assessors' labels to form fine-grained relevance levels is beneficial in terms of discriminative power.
- Removing 23 low-agreement topics (in terms of inter-assessor agreement) from the full set of 50 topics prior to evaluating runs did not have a major impact on the evaluation results, as the properties of the 27 high-agreement topics are dominant in the full set. However, replacing the high-agreement set with the low-agreement set resulted in a statistically significantly different system ranking, and substantially lower discriminative power. Hence, from a statistical point of view, our results suggest that a high-agreement topic set is more useful for finding concrete research conclusions than a low-agreement one.

ACKNOWLEDGEMENTS

I thank the PLY team (Peng Xiao, Lingtao Li, Yimeng Fan) of my laboratory for developing the PLY relevance assessment tool and collecting the assessments. I also thank the NTCIR-13 WWW task organisers and participants for making this study possible.

REFERENCES

[1] Peter Bailey, Nick Craswell adn Arjen P. de Vries, and Ian Soboroff. 2008. Overview of the TREC 2007 Enterprise Track. In *Proceedings of TREC 2007*.
[2] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter?. In *Proceedings of ACM SIGIR 2008*. 667–674.
[3] Ben Carterette. 2012. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS* 30, 1 (2012).
[4] Ben Carterette and Ian Soboroff. 2010. The Effect of Assessor Errors on IR System Evaluation. In *Procceedings of ACM SIGIR 2010*. 539–546.
[5] Thomas Demeester, Robin Aly, Djoerd Hiemstra, Dong Nguyen, Dolf Trieschnigg, and Chris Develder. 2014. Exploiting User Disagreement for Web Search Evaluation: an Experimental Approach. In *Proceedings of ACM WSDM 2014*. 33–42.
[6] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2017. AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. *ACM TOIS* 36, 2 (2017).
[7] Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin* 76, 5 (1971), 378–382.
[8] Donna K. Harman. 2005. The TREC Ad Hoc Experiments. In *TREC: Experiment and Evaluation in Information Retrieval*, Ellen M. Voorhees and Donna K. Harman (Eds.). The MIT Press, Chapter 4.

[9] Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology (Third Edition)*. Sage Publications.
[10] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the NTCIR-13 WWW Task. In *Proceedings of NTCIR-13*.
[11] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering Assessor Agreement in IR Evaluation. In *Proceedings of ACM ICTIR 2017*. 75–82.
[12] Olga Megorskaya, Vladimir Kukushkin, and Pavel Serdyukov. 2015. On the Relation between Asessor's Agreement and Accuracy in Gamified Relevance Assessment. In *Proceedings of ACM SIGIR 2015*. 605–614.
[13] Justus J. Randolph. 2005. Free-Marginal Multirater Kappa (Multirater κ_{free}): An Alternative to Fleiss' Fixed Marginal Multirater Kappa. In *Joensuu Learning and Instruction Symposium 2005*.
[14] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics based on the Bootstrap. In *Proceedings of ACM SIGIR 2006*. 525–532.
[15] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of ACM SIGIR 2007*. 71–78.
[16] Tetsuya Sakai. 2012. Evaluation with Informational and Navigational Intents. In *Proceedings of WWW 2012*. 499–508.
[17] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*. 116–163.
[18] Tetsuya Sakai. 2017. Unanimity-Aware Gain for Highly Subjective Assessments. In *Proceedings of EVIA 2017*.
[19] Ellen Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management* (2000), 697–716.
[20] Yulu Wang, Garrick Sherman, Jimmy Lin, and Miles Efron. 2015. Assessor Differences and User Preferences in Tweet Timeline Generation. In *Proceedings of ACM SIGIR 2015*. 615–624.
[21] William Webber, Praveen Chandar, and Ben Carterette. 2012. Alternative Assessor Disagreement and Retrieval Depth. In *Proceedings of ACM CIKM 2012*. 125–134.