How Discriminative Are Your Qrels? How To Study the Statistical Significance of Document Adjudication Methods

David Otero
david.otero.freijeiro@udc.es
Information Retrieval Lab, CITIC
Universidade da Coruña
A Coruña, Spain

Javier Parapar
javier.parapar@udc.es
Information Retrieval Lab, CITIC
Universidade da Coruña
A Coruña, Spain

Nicola Ferro
ferro@dei.unipd.it
University of Padua
Padova, Italy

ABSTRACT

Creating test collections for offline retrieval evaluation requires human effort to judge documents' relevance. This expensive activity motivated much work in developing methods for constructing benchmarks with fewer assessment costs. In this respect, adjudication methods actively decide both which documents and the order in which experts review them, in order to better exploit the assessment budget or to lower it. Researchers evaluate the quality of those methods by measuring the correlation between the known gold ranking of systems under the full collection and the observed ranking of systems under the lower-cost one. This traditional analysis ignores whether and how the low-cost judgements impact on the statistically significant differences among systems with respect to the full collection. We fill this void by proposing a novel methodology to evaluate how the low-cost adjudication methods preserve the pairwise significant differences between systems as the full collection. In other terms, while traditional approaches look for stability in answering the question "is system A better than system B?", our proposed approach looks for stability in answering the question "is system A significantly better than system B?", which is the ultimate questions researchers need to answer to guarantee the generalisability of their results. Among other results, we found that the best methods in terms of ranking of systems correlation do not always match those preserving statistical significance.

CCS CONCEPTS

• **Information systems** → **Relevance assessment**; **Test collections**.

KEYWORDS

Evaluation, Pooling, Adjudication Method, Significance

ACM Reference Format:
David Otero, Javier Parapar, and Nicola Ferro. 2023. How Discriminative Are Your Qrels? How To Study the Statistical Significance of Document Adjudication Methods. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614916

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3614916

1 INTRODUCTION

Information Retrieval (IR) is a field with a strong focus on evaluation [18, 50], whose main purpose is to empirically measure the effectiveness of retrieval systems. Offline batch evaluation allows researchers to perform experiments under controlled conditions and enables the reproducibility of the results. It is based on test collections, which consist of a corpus of documents, topics, and relevance judgements (also called assessments, or *qrels*) [42]. Acquiring the assessments for creating these collections is costly, since human experts have to judge the documents' content and decide which ones are relevant for each topic. The advantage is that once the collections are created, it is straightforward and cheap to conduct as many experiments as needed to evaluate and compare the performance of (new) IR systems [55].

The first and small test collections had complete judgements [18], containing a human assessment for each topic-document pair, thus representing the ideal situation in terms of evaluation quality. However, that exhaustive procedure is only feasible for collections with a very small corpus. Nonetheless, small corpora are not the conditions that operational search systems face. As a consequence, when larger collections arose, there was the need to implement some kind of *sampling* so that assessors would not have to judge the relevance of each document for each topic. However, simple random sampling, the most immediate approach, would not work, since the number of relevant documents for a topic is extremely small compared to the size of the corpus of documents. Thus, a random sample would end up consisting of (almost all) non-relevant documents. The first solution to this problem was the *pooling* technique implemented by TREC [45, 55]. With this technique, assessors only judge a subset of the corpus, the *pool*. For each topic, the pool consists of the union of the top-k documents retrieved by several search systems for that topic. The assessors judge the relevance of the documents in the pool while the rest, i.e. the non-pooled documents, are assumed to be non-relevant. Top-k pooling builds on the assumption that IR systems try to push relevant documents towards the top of the ranking and thus there is a good chance to pool most of the relevant documents for a topic, provided that k is deep enough and the pooled systems are diverse enough. However, the number of judgements that an assessor can perform, i.e. the **budget**, is limited and, therefore, there is a trade-off with the depth k of the pool and the number of pooled systems, since the more they grow, the higher the number of documents in the pool.

Pooling does not guarantee finding all the relevant documents for a topic but, as said, it strives to find a very good share of them. Researchers are interested in comparing systems in order to answer the fundamental question "is system A better than system

B?". Answering this question requires a good estimate of system performance rather than absolute performance scores which, in turn, would demand finding all the relevant documents. Therefore, the quality of a pool is *traditionally* measured on its ability to *fairly rank* systems, i.e. to fairly compare them. This is not limited to the systems which were actually pooled, but it should also hold for systems which were not pooled [59], to ensure the future reusability of a test collection also with new systems.

However, collections kept growing in size, and just judging deep pools over a diverse set of systems stopped to be a practicable approach as well [53]. Therefore, much work has focused on developing alternative methods to better select which documents to pool and judge by performing some sort of *focused sampling*, aimed at picking documents which more probably turn out to be relevant and better employing the assessor budget or allowing for lower budgets at a comparable quality [29, 35]. A method that *actively* decides which document to judge next is called an **adjudication method**. However, alternative prioritisation models may introduce biases or incompleteness in the judgements, hampering the future reusability of an test collection [49].

Therefore, the quality of new adjudication methods is *traditionally* assessed by checking that they rank systems as closely as possible to the full set of judgements of a (good quality) top-k pool, ensuring that they can still properly answer the question "is system A better than system B?". This is quantified by computing the correlation, e.g. Kendall's τ [25, 26], between the ranking of systems produced by an adjudication method and by the full top-k pool. The rationale is that if this correlation is high, one may assume the validity of the new method and aim to use it in the future for building new test collections at a comparable quality but with a lower assessment cost.

However, the question researchers are really interested in is rather "is system A *statistically significantly* better than system B?", since this ensures that observed differences are not due just to the randomness present in the construction process of a collection and, especially, that the found differences would *generalise* better and still hold in operational settings [17, 38]. The problem is that the above correlation measures ignore whether the evaluated systems' statistical significance is preserved.

Let us better explain this problem with an example. Let us assume we have three different IR systems, `Sys1`, `Sys2` and `Sys3`, and that their true ranking, given by the full top-k pool, is (`Sys1`, `Sys2`, `Sys3`). We perform a significance test between all possible pairwise comparisons and we obtain that `Sys1` is significantly better than `Sys2` and `Sys3`, and `Sys2` is also significantly better than `Sys3`. Then, we create a new set of judgements using some adjudication method and repeat the above procedure. Using this new pool, we find the same ranking of systems as when using the full top-k pool, leading to a perfect correlation and concluding that the adjudication method is fully equivalent, but less costly, than the full top-k pool. However, we do not know anything about the significance between systems. If we repeat the same significance test using the new pool instead, we may not find any significant difference between any pair. We may thus conclude that there is no evidence of any system being different from the rest. This would be the opposite conclusion than the one drawn on the full top-k pool, where all the system pairs were significantly different.

In this work, our objectives are two-fold. First, we aim to propose a new approach to evaluate the validity of low-cost adjudication methods, focusing on how they preserve the statistically significant differences between systems. Second, we analyse some state-of-the-art adjudication methods using our new approach to gain new insights about them. In particular, we aim to answer the following research questions: **RQ1** Are the adjudication methods able to preserve the same statistically significant differences as the full top-k pool? **RQ2** When adjudication methods fail to see a real significant difference, do they follow any distinguishable pattern in terms of system position in the ranking? **RQ3** Are the adjudication methods able to preserve the same statistically significant differences as the full top-k pool for new (non-pooled) systems?

The rest of the paper is organised as follows: Section 2 introduces past work; Section 3 explains our methodology; Section 4 and Section 5 report our experiments; and, finally, Section 6 draws conclusions and presents some ideas for future work.

2 RELATED WORK

How to build high-quality experimental collections for retrieval evaluation is still an open research question [13, 53, 56]. Research in adjudication methods looks for ways of prioritising the pooled documents so that the assessors expend their effort in judging relevant documents. In this way, we may only need to judge some of the pooled documents while maintaining the quality of the judgements, thus making more efficient use of the resources.

Losada et al. [29] proposed a series of sampling methods based on the multi-armed bandit problem. The multi-armed bandit problem [46, Chapter 2] has been a subject of research for decades in Reinforcement Learning (RL), statistics and other fields. These methods bring ideas from RL to the task of document adjudication for building test collections. They apply Bayesian principles to this problem, formalising the uncertainty associated with reviewing a document from a pooled system. Other works have also explored the development of adjudication methods [11, 27, 32, 35] Section 4 provides further details about the state-of-the-art adjudication methods under experimentation. Adjudication methods have shown remarkable improvements in bringing relevant documents earlier in the pooling process, and indeed they were used to build the collection of the TREC Common Core Track of 2017 [1]. However, the quality of the judgements produced with a limited budget is still an open question [49].

Previous work on adjudicating methods used a series of metrics to evaluate the quality of these algorithms. The commonest is Kendall's τ [25, 26] correlation, which researchers use to measure how well a new adjudication method can induce the gold ranking of systems, i.e. the one on the full top-k pool. Another top-weighted correlation, τ_{AP} [58], is also common. This correlation penalises swaps in higher positions more. In some works [49, 53], they also measure the change in the ranking position of the system that suffers the highest drop as a measure of the reusability of an experimental collection. The problem with all these measures, as we already introduced earlier, is that they ignore the significance between the scores of the systems. If we ignore this, it is meaningless to account for ranking swaps.

In this work, we propose a new methodology to evaluate low-cost adjudication methods that, instead of focusing only on the ranking of the systems, focuses on evaluating how well a method preserves the real pairwise significant differences.

Statistical significance testing is of paramount importance in IR, and studying the properties of significance tests is an active area of research [4, 7, 8, 9, 10, 15, 16, 23, 33, 34, 39, 43, 44, 47, 48, 57]. However, this is out-of-scope for the present work which, instead, focuses on considering the output of a statistical significance test as a way to assess the quality of an adjudication method.

3 METHOD

Let $S = \{s_i\}$, $|S| = n$, be the set of systems under experimentation, and let G be the *gold assessments* (also said gold qrels), i.e. the full top-k pool. Using an effectiveness measure of choice, we compute the per-topic scores for each of the n systems and we perform a statistical test for each pairwise comparison between systems. From this test, we obtain, for each pair of systems s_i and s_j ($i < j \leq n$), a triplet $\langle s_i, s_j, c\rangle$, where $c \in \{>, \gg, <, \ll\}$, denoting the four outcomes we are interested in: s_i is better than s_j ($s_i > s_j$), s_i is *significantly* better than s_j ($s_i \gg s_j$), s_j is better than s_i ($s_i < s_j$), or s_j is *significantly* better than s_i ($s_i \ll s_j$).

Now we use R_G to denote the set of triplets that result from the statistical test performed using the gold qrels. Similarly, we use L to denote the qrels obtained with a low-cost adjudication method ($L \subseteq G$) and R_L to denote the set of triplets that result from the statistical test performed with them. Note that $|R_G| = |R_L| = \frac{n(n-1)}{2}$. Finally, we use T_G to denote the set of comparisons from R_G that are significantly different, that is, the set triplets for which $c \in \{\ll, \gg\}$, and T_L for the significantly different comparisons obtained with the low-cost assessments.

As we already explained, we are interested in studying to what extent the judgements produced by different low-cost adjudication methods preserve the statistically significant differences between systems we observe when using the gold qrels. The idea here is that if the low-cost method is able to preserve such differences, we could confidently use it to build new collections in the future with fewer assessment costs. Thus, we compare how T_G and T_L agree with each other using the measures described in the following section.

3.1 Measures

Kendall's τ. Kendall's τ is the measure traditionally used to evaluate adjudication methods. It computes the correlation between the ranking of systems under the gold qrels setting and the one under the qrels produced with the different adjudication methods.

Given two rankings over the same set of items, Kendall's τ computes how many items are swapped as follows: $\tau = (P - Q)/\binom{n}{2}$, where P is the number of concordant pairs (pairs of systems ranked in the same relative order in both lists), Q is the number of discordant pairs (swapped pairs of systems), and $\binom{n}{2} = \frac{n(n-1)}{2}$ is the number of total pairs, given that we have n items.

Precision and Recall. We consider the *Precision* (P) and *Recall* (R) of the significantly different pairs detected by the low-cost adjudication methods, defined as follows:

$$P = \frac{|T_G \cap T_L|}{|T_L|},\ R = \frac{|T_G \cap T_L|}{|T_G|}$$

where $|T_G \cap T_L|$ is the number of significantly different pairs common to both the gold and adjudication qrles, i.e. the correct ones when assuming the gold qrels detect the "true" differences. Precision indicates how much "noise" is introduced by an adjudication method, meant as additional significant differences not detected by gold qrels; Recall indicates how many of the total possible significant differences are not detected by an adjudication method.

Agreements. We consider an adaptation of a series of agreement measures that have been used in past work [14, 15, 31, 48]. Note that, while Kendall's τ and Precision/Recall focus on ranking of systems (the former) or on matching significantly different pairs (the latter) in isolation, the following agreement measures consider them jointly.

- **Active Agreements (AA)**: the set of consistent outcomes between both methods. This is, $\langle s_i, s_j, \gg\rangle \in T_G$ and $\langle s_i, s_j, \gg\rangle \in T_L$ or $\langle s_i, s_j, \ll\rangle \in T_G$ and $\langle s_i, s_j, \ll\rangle \in T_L$. This is the best possible case, and thus, the larger AA are, the better.
- **Active Disagreements (AD)**: the set of opposite outputs between both methods. This is, $\langle s_i, s_j, \gg\rangle \in T_G$ and $\langle s_i, s_j, \ll\rangle \in T_L$, or $\langle s_i, s_j, \ll\rangle \in T_G$ and $\langle s_i, s_j, \gg\rangle \in T_L$. This is the worst possible case, since it means that both methods reach complete opposite conclusions for a given pair. Thus, the lesser, the better.
- **Mixed Agreements (MA)**: we have four possible options: ❶ $\langle s_i, s_j, \ll\rangle \in T_G$ and $\langle s_i, s_j, <\rangle \in T_L$, or ❷ $\langle s_i, s_j, \gg\rangle \in T_G$ and $\langle s_i, s_j, >\rangle \in T_L$, or ❸ $\langle s_i, s_j, <\rangle \in T_G$ and $\langle s_i, s_j, \ll\rangle \in T_L$, or ❹ $\langle s_i, s_j, >\rangle \in T_G$ and $\langle s_i, s_j, \gg\rangle \in T_L$. We distinguish between MA_G (❶ and ❷), which counts the cases where the adjudication method was not able to see a gold significant difference. Conversely, MA_L (❸ and ❹) counts the cases where a low-cost method sees a significant difference that is not in the gold qrels. Note that $\text{MA}_\text{G} + \text{MA}_\text{L} = \text{MA}$
- **Mixed Disagreements (MD)**: we also have four possible cases here: ❺ $\langle s_i, s_j, \ll\rangle \in T_G$ and $\langle s_i, s_j, >\rangle \in T_L$, or ❻ $\langle s_i, s_j, \gg\rangle \in T_G$ and $\langle s_i, s_j, <\rangle \in T_L$, or ❼ $\langle s_i, s_j, >\rangle \in T_G$ and $\langle s_i, s_j, \ll\rangle \in T_L$, or ❽ $\langle s_i, s_j, <\rangle \in T_G$ and $\langle s_i, s_j, \gg\rangle \in T_L$. Here, as with MA, we also distinguish between MD_G (❺ and ❻) and MD_L (❼ and ❽)

Bias. Analogously to Ferro and Sanderson [15], we also consider the *publication bias*, i.e. the likelihood of a researcher publishing a significant result using an adjudication method when in fact a significance test on the gold qrels would have produced either no significance (MA, MD) or a significant result in the opposite direction (AD). We define it as follows:

$$Bias = 1 - \frac{AA}{AA + AD + MA_L + MD_L}$$

A value of 0% means that every significance detected by an adjudication method leads to the same conclusions (and publication) as those of the gold qrels. Conversely, a value of 100% means that every significance detected by an adjudication method leads to opposite conclusions (and publication) to those of the gold qrels. Thus, the lower the bias, the better. Note that, differently from Ferro and Sanderson [15], we do not consider the whole MA and MD but just MA_L and MD_L, since we are interested only in the

publication bias induced by the adjudication method. This metric tries to measure the situations where a researcher sees a significant outcome under the reduced pools when, in reality, it would be a different conclusion under the gold qrels.

3.2 Family-Wise Error Rate (FWER)

Performing *multiple comparisons*—in our case between each pair of systems—leads to an increase of the *Type I error*, i.e. incorrectly rejecting the null hypothesis, and inflates the number of significant differences found [20, 22, 37].

The Type I error probability is equal to the significance level α and, as the number of comparisons increases, this probability also does. If we perform k different system comparisons, the probability of correctly accepting the null hypothesis for all of them is equal to $(1-\alpha)^k$. Thus, the probability of committing at least one Type I error is $1-(1-\alpha)^k$. This is the *family-wise error rate* (FWER). If we have, for example, $\alpha = 0.05$ and $k = 6$ comparisons (4 systems, $\frac{4(4-1)}{2} = 6$), this probability would rise to 0.264, which is not acceptable. For this reason, when we perform multiple comparisons, we should employ a technique to adjust the p-values, so that the FWER stays below α. Obviously, this has the side-effect of reducing the *power* of the statistical test and increasing the number of *Type II errors*, i.e. not detecting an actual significant difference.

There are several options to control the FWER in a multiple comparison situation. The Bonferroni correction, for example, is a post-hoc correction where, if we have k different comparisons, we should use $p < \frac{\alpha}{k}$ as our significance level in each pairwise comparison. However, the Bonferroni correction is known to be too conservative and to reduce the power of a test too much, especially when the number of comparisons increases as in our case. Therefore, we employ the randomised version of the Tukey Honestly Significant Difference (HSD) test [8, 37]. This is a nonparametric computer-based generalisation of the common permutation test for handling more than 2 systems. At each permutation, the test perturbs the array of system scores of each topic, and, after this perturbation, computes the difference between the maximum and minimum average system scores. Then the test counts how many times the actual differences between system average performance is greater than the permuted mean to determine if it is *honestly* significant [8]. The Tukey HSD test produces a p-value for each pairwise comparison, which can be compared to the significance level α to decide whether that pair of systems is significantly different or not. Algorithm 1 (adapted from prior work [8, 37]) shows the details of our implementation.

Algorithm 1 Paired Randomised Tukey HSD

Input
X $m \times n$ topic-system scores matrix.
B number of permutations.
Output
P $n \times n$ matrix holding a p-value for each pairwise system comparison.

```
for k ← 1 to B do
    initialise m × n matrix X'
    for each topic t do
        row t of X' ← permutation of values in row t of X
    end for
    d' ← max_i X̄'_i − min_j X̄'_j        ▷ X̄'_i is the mean of column i
    for each pair of systems i, j do
        if d' > |X̄_i − X̄_j| then
            P_{i,j} ← P_{i,j} + 1/B
        end if
    end for
end for
```

4 EXPERIMENTAL SETUP

Collections. We employ the TREC-8 ad hoc collection, known to have a very high-quality pool [54, 56]. It includes 129 system submissions, retrieving 1000 documents for each topic, and 50 topics. Official relevance judgements are based on a pool of depth 100 over 71 out of 129 submitted runs, resulting in 86 830 assessments across all 50 topics. The average pool size per topic is 1736, while the maximum and the minimum are 2992 and 1046, respectively. Additionally, we use the collection from the document ranking task of TREC 2021 Deep Learning track [12], which adopted a shallow pooling approach at depth 10, then enlarged with a method based on active learning. We used only the documents in the top-10 pools as our gold qrels to provide a fairer comparison to the case of TREC-8. It includes 66 runs, retrieving 100 documents for each topic, and 13 058 judgements made by NIST assessors over 57 different topics. The depth-10 pools we used include 6510 judgements, with an average pool size of 114, a maximum of 226 and a minimum of 50.

Adjudication methods. We consider a series of state-of-the-art adjudication methods.

- **top-k pooling.** We adapt the standard method used in TREC to limited-budget situations. When limiting the budget of assessments, we choose a k deep enough to fill that budget. Then, pooled documents are sorted by their document identifier [55].
- **MoveToFront (MTF).** MTF is a dynamic adjudication method proposed by Cormack and colleagues [11] that has been acknowledged as a robust adjudication method [2].
- **MaxMean (MM), MM Non Stationary (MM-NS), Thompson Sampling (TS) and TS Non Stationary (TS-NS).** Bandit-based methods for document adjudication apply bayesian principles to formalise the uncertainty associated with the probabilities of pulling a positive reward (a relevant document) from playing a bandit [28].
- **Hedge.** Hedge is an online learning algorithm adapted for pooling in [3]. A more detailed explanation of applying Hedge for pooling can be found in this article [29].
- **NTCIR top-k prioritization.** Documents in the pool are sorted by the number of runs that contain the document at or above the depth k (the higher the better), ties are solved with the sum of the ranks of that document within the runs (the lower the better) [41].

Other Settings. We used Average Precision (AP) [6] and Normalized Discounted Cumulative Gain (NDCG) [24] as performance measures to score runs. We used $\alpha = 0.05$ as significance level and $B = 1\,000\,000$ permutations in Tukey HSD test. Finally, since MTF, MM, MM-NS, TS, and TS-NS have a stochastic nature, the reported results for those methods are averaged over 50 executions of each.

To ease the reproducibility of the experiments, we release the source code.[1]

5 RESULTS AND DISCUSSION

5.1 RQ1: Preservation of significant differences

In Table 1, we report the Kendall's τ, Precision and Recall, as defined in Section 3, that each adjudication method achieves, while varying the number of assessments per topic. We report the scores for 100 judgements per topic (which is a 6% budget of the original pool), and 300 (17%). All this values were obtained using the pooled systems of the TREC-8 collection, which includes 71 different systems.

Regarding Kendall's τ and consistently with previous findings in the literature, we see almost every method achieves a very high correlation ($\tau > 0.90$) already at a 6% of the original budget. While this means that every method obtains a ranking of systems very similar to the one of the gold qrels, it also makes it very difficult to distinguish among methods. Moreover, we can observe that top-k and NTCIR methods stay behind the rest, leaving room for improvement in developing more efficient adjudication strategies for building new collections in evaluation workshops.

As we mentioned earlier, Kendall's τ does not allow us to know whether the compared algorithms preserve the same statistically significant differences as the gold qrels. Therefore, we study to which extent this effect might hold by using the Precision and Recall measures previously introduced.

We observe that every method obtains Precision and Recall values over 90% in almost all the cases, which is a quite solid result. Moreover, every method is able to mostly preserve the same differences just having a 6% of the original budget. With 300 assessment per topic (17% of the budget), Recall is (almost) 1.00 for most of the methods, indicating that they are able to detect all the significant differences of the gold qrels at less than one third of the cost.

It is also interesting to observe that most of them detect some differences that there were not detected in the gold qrels. Indeed, Precision is lower than 1.00 while Recall is almost 1.00 (all the differences in the gold qrels detected). In other terms, T_L (the set of significant differences detected by the adjudication method) is not a proper subset of T_G (the set of significant differences detected by the gold qrels). A possible explanation might be that, since reduced pools lack some relevant documents, the performance difference of some pair of systems (delta AP/NDCG between the two systems in our case) turns out to be increased with respect to the gold qrels and this makes the pair significantly different on the reduced pool but not on the gold qrels. Since more evaluation on this issue would need more experimentation, due to space restrictions we leave this investigation for future work.

To support a more detailed analysis, in Table 2, we report the raw agreements of each method. The upper half of the table includes the results obtained when using AP for evaluating the runs. In this case, there are a total of 966 gold significant differences ($|T_G| = 966$). The lower half includes the results when using NDCG. In this case, there are a total of 917 gold significant differences ($|T_G| = 917$).

[1] https://github.com/davidoterof/cikm2023

Table 1: Kendall's τ, Precision and Recall (see Section 3) of each adjudication method for a varying number of judgements per topic. 100 and 300 are the budget of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 71 pooled systems of TREC-8. For each column, best values are bolded and worst ones underlined.

Method	MAP/100 (6%)			MAP/300 (17%)			NDCG/100 (6%)			NDCG/300 (17%)		
	τ	P	R	τ	P	R	τ	P	R	τ	P	R
top-k	0.91	0.932	0.888	0.95	0.955	0.955	0.90	**0.975**	0.929	0.94	0.985	0.970
MTF	0.94	0.946	**0.961**	0.97	0.962	0.980	0.91	**0.975**	0.953	**0.96**	0.982	0.985
MM	**0.95**	0.948	0.958	**0.98**	**0.969**	**0.992**	**0.92**	0.942	0.973	**0.96**	0.976	**0.991**
MM-NS	0.93	0.942	0.957	0.97	0.967	0.987	0.90	0.970	0.962	**0.96**	**0.986**	**0.991**
TS	**0.95**	0.947	0.954	**0.98**	**0.969**	0.991	**0.92**	0.940	0.970	**0.96**	0.975	0.990
TS-NS	0.93	0.945	0.949	0.97	0.966	0.983	0.90	0.971	0.960	**0.96**	0.985	**0.991**
Hedge	0.94	**0.955**	0.947	**0.98**	0.968	0.980	**0.91**	0.959	**0.978**	0.95	0.972	0.989
NTCIR	0.83	0.900	0.876	0.96	0.942	0.925	0.81	0.961	0.942	0.93	0.977	0.988

Table 2: Relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size with respect to the full pool. We used the 71 pooled systems of TREC-8. The top-100 full pool includes 4728 relevant documents. There are 2485 pairwise comparisons, of which 966 are significant under the gold qrels with MAP (upper half), and 917 with NDCG (lower half). For each budget and metric, the best values are bolded and the worst ones are underlined.

	Metric	Adjudication method							
		top-k	MTF	MM	MM-NS	TS	TS-NS	Hedge	NTCIR
MAP (966 gold significantly different pairs)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
	AA	858	**929**	926	925	922	917	915	846
	MA_{total}	170	**90**	**90**	98	94	102	94	185
	MA_G	108	**37**	40	41	44	49	51	91
	MA_L	62	52	50	57	50	53	**43**	94
	MD_{total}	**0**	**0**	1	**0**	1	**0**	**0**	29
	MD_G	**0**	**0**	**0**	**0**	**0**	**0**	**0**	29
	MD_L	**0**	**0**	1	**0**	1	**0**	**0**	**0**
	AD	0	0	0	0	0	0	0	0
	Bias	7%	5%	5%	6%	5%	5%	**4%**	10%
Budget per top.: 100 (6%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
	AA	923	**961**	959	954	958	950	947	894
	MA_{total}	86	43	**38**	44	39	50	50	127
	MA_G	43	**5**	7	12	8	16	19	72
	MA_L	43	38	**30**	32	**30**	33	31	55
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
Budget per top.: 300 (17%)	Bias	4%	4%	**3%**	**3%**	**3%**	**3%**	**3%**	6%
NDCG (917 gold significantly different pairs)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
	AA	852	874	893	883	890	881	**897**	864
	MA_{total}	86	65	79	61	83	62	**58**	88
	MA_G	65	43	24	34	27	36	**20**	53
	MA_L	**21**	22	55	27	56	26	38	35
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
Budget per top.: 100 (6%)	Bias	**2%**	**2%**	6%	3%	6%	3%	4%	4%
	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
	AA	890	904	**909**	**909**	**909**	**909**	907	906
	MA_{total}	40	29	30	**20**	31	21	36	32
	MA_G	27	13	**8**	**8**	**8**	**8**	10	11
	MA_L	13	16	22	**12**	22	13	26	21
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
Budget per top.: 300 (17%)	Bias	**1%**	2%	2%	**1%**	2%	**1%**	3%	2%

The AA counts confirm that adjudication methods are more effective than top-k and NTCIR pooling methods in detecting significant pairs in the correct order, especially at lower budgets. They provide further insights about the (almost) 1.00 Recall (see Table 1) we observed for most adjudication methods. Indeed, with AP, the gold qrels detect 966 significantly different pairs and the AA counts is (almost) 966, indicating that the 1.00 Recall is due to significant pairs in the correct order. The same happens for NDCG, where we observe that most methods obtain AA values near 917. In other terms, the slight drop in Kendall's τ observed in Table 1 is not caused by wrongly ordered pairs, even when Recall is 1.00. When it comes to the specific methods, MTF achieves the best AA figures for budgets of 100, 300 when using AP, while under NDCG Hedge works slightly better with lower budgets and bandit-based methods perform the best with a budget of 300.

If we compare the AA counts with the number of relevant documents found by a method (the # rels. row), we observe a somehow unexpected behaviour. One might think that the more relevant documents found, the more AA increases. However, for a budget of 100 judgements per topic, Hedge adjudicated 2170 relevant documents, 485 more than MTF, but the latter one achieves the highest AA with AP; the same happens again for a budget of 300: MTF is not the best one in terms of relevant documents but it is the best in terms of AA. We can observe something similar with NDCG: founding more relevant documents does not necessarily mean more AA. Obviously, having more relevant documents in the pool helps in increasing the number of AA, but these results showcase that it is not the only factor. Overall, these observations suggest that not all the relevant documents are equally discriminative in finding significantly different pairs. Indeed, relevant documents appear at different ranks in the results lists and the same (or even higher) number of relevant documents may contribute differently to the performance score of a run and, in turn, to the significant differences found. So far, research has mostly focused on determining the number of topics needed [5, 40, 43, 51, 52] or on identifying the most discriminative subset of topics [19, 21, 30, 36]. These findings open up the possibility of future research on which are the best relevant documents to more reliably discriminate among systems, an area not well explored yet, to the best of our knowledge.

Almost in every case, no method fails in a mixed or active disagreement, i.e. detecting significant differences when there is a swap. This represents a very important insight from this experiment, since it shows that no method causes a ranking swap between a pair of systems that were originally significantly different. In other terms, the drop in Kendall's τ is not due to swaps between systems that are significantly different on the gold qrels but swaps only happen among not significantly different systems, having a much lower impact.

Let us now consider MA_G and MA_L. The former accounts for significant pairs in the gold qrels which are missed by reduced pools; thus, it helps mainly to explain drops in Recall. The latter accounts for significant pairs in a reduced pool which are not present in the gold qrels; thus, it helps mainly to explain drops in Precision. We can observe that MA_G gets reduced as the budget size increases up to almost 0, with the exception of top-k pooling, Hedge and NTCIR method, consistently with the previous findings in Table 1.

Table 3: Kendall's τ, Precision and Recall (see Section 3) of each adjudication method for a varying number of judgements per topic. 10 and 30 are the budget of judgements per topic. Parentheses indicate the size with respect to the full pool. We used the 66 pooled systems from DL21. For each column, the best values are bolded and the worst ones are underlined.

Method	MAP/10 (9%)			MAP/30 (26%)			NDCG/10 (9%)			NDCG/30 (26%)		
	τ	P	R	τ	P	R	τ	P	R	τ	P	R
top-k	0.46	0.448	0.445	0.69	0.668	0.833	0.61	0.531	0.554	**0.82**	0.723	0.832
MTF	0.49	**0.611**	0.414	0.69	0.687	0.798	0.61	**0.632**	0.534	0.79	0.734	0.808
MM	**0.53**	0.566	0.477	**0.73**	**0.764**	0.778	**0.66**	0.628	0.598	0.81	0.772	0.808
MM-NS	0.50	0.517	0.505	0.70	0.654	0.841	0.64	0.593	0.607	**0.82**	0.725	**0.844**
TS	0.52	0.554	0.489	**0.73**	0.761	0.777	**0.66**	0.624	0.605	**0.82**	**0.780**	0.809
TS-NS	0.50	0.509	0.502	0.69	0.642	0.839	0.63	0.589	0.603	0.81	0.715	0.839
Hedge	0.42	0.430	0.419	0.50	0.558	0.603	0.51	0.521	0.484	0.61	0.657	0.674
NTCIR	0.47	0.423	**0.560**	0.69	0.594	**0.871**	0.59	0.522	**0.621**	0.76	0.669	0.827

Moreover, MA_L is consistently higher than MA_G, explaining the loss in Precision even at very high Recall levels.

When it comes to publication bias, we observe moderate values, from 7% and below, suggesting that all the methods would not lead to draw conclusions severely different from the gold qrels. We can observe that bias quickly decreases as the budget increases and that adjudication methods are more effective than top-k pooling, achieving a bias up to 2-3 times lower than it.

Finally, we can observe that there are not different trends between the two evaluation metrics employed, AP and NDCG. This shows that the results presented here are not an artefact of the metric used, but of the adjudication methods being evaluated.

Additionally, we run experiments on the TREC Deep Learning (DL) track 2021. We selected this collection as having opposing characteristics to TREC-8. The DL collection adopts a very shallow pooling at just depth 10, representing a quite challenging setting for adjudication methods. We believe that using these two collections helps in supporting the generalizability of the results presented here. Table 3 reports the Kendall's τ, Precision, and Recall, similarly to Table 1 for TREC-8; Table 4 reports the agreement counts, similarly to Table 2 for TREC-8. In general, we observe quite lower and much more varied performance on DL 2021 than on TREC-8.

Kendall's τ is generally low for all the methods with both metrics. In TREC-8, adjudication methods were able to obtain very strong results only with a 17% of the original budget, while in this case no method is able to reach that performance even with a 26%. One important difference is that, while in TREC-8 top-k and NTCIR method were clearly underperforming with respect to the other methods, in DL 2021 Hedge clearly achieves the worst performance.

When it comes to the agreements (Table 3), a notable difference is that, at low budgets (9%), MD appear while they go to (almost) zero for higher budgets. The MD at 9% budget indicate that the drop in Kendall's τ are also due to swaps in the significantly different pairs. The problem concerns more MD_L, i.e. swaps in significant pairs detected by a reduced pool but not the gold qrles, than MD_G, i.e. swaps in significant pairs detected by the gold qrels but not a reduced pool. As a consequence, part of the loss of Precision is due to swaps in the significant pairs a more severe condition than the one causing the loss of Precision in TREC-8. This issue impacts more

Table 4: Relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size with respect to the full pool. We used the 66 pooled systems from DL21. The top-10 pool includes 3541 relevant documents. There are a total of 2145 pairwise comparisons, of which 418 are significant under the gold qrels with MAP (upper half), and 417 with NDCG (lower half). For each budget, the best values are bolded and the worst ones are underlined.

		Metric	Adjudication method							
			top-k	MTF	MM	MM-NS	TS	TS-NS	Hedge	NTCIR
MAP (418 gold significantly different pairs)	Budget per topic: 10 (9%)	# rels.	441	488	489	474	483	469	504	**513**
		AA	186	173	199	211	204	210	175	**234**
		MA_{total}	413	**345**	358	386	361	392	442	464
		MA_G	214	237	212	201	206	203	235	176
		MA_L	199	**108**	146	185	155	189	207	288
		MD_{total}	48	**13**	15	19	18	19	33	39
		MD_G	18	8	7	**6**	8	**6**	8	8
		MD_L	30	**5**	9	13	10	14	25	31
		AD	0	0	0	0	0	0	0	0
		Bias	55%	**39%**	43%	48%	45%	49%	57%	58%
	Budget per top.: 30 (26%)	# rels.	1186	1327	**1359**	1289	1345	1267	1352	1337
		AA	348	334	325	352	325	351	252	**364**
		MA_{total}	243	237	**194**	251	196	262	355	299
		MA_G	70	84	93	66	93	67	161	**54**
		MA_L	173	152	**101**	185	103	194	194	245
		MD_{total}	**0**	**0**	**0**	1	**0**	1	11	4
		MD_G	**0**	**0**	**0**	**0**	**0**	**0**	5	**0**
		MD_L	**0**	**0**	**0**	1	**0**	1	6	4
		AD	0	0	0	0	0	0	0	0
		Bias	33%	31%	**24%**	35%	**24%**	36%	44%	41%
NDCG (417 gold significantly different pairs)	Budget per top.: 10 (9%)	# rels.	441	488	489	474	483	469	504	**513**
		AA	231	223	249	253	252	252	202	**259**
		MA_{total}	376	322	**314**	333	315	337	388	381
		MA_G	184	193	167	164	165	165	215	**158**
		MA_L	192	**129**	146	170	151	172	173	223
		MD_{total}	14	**3**	**3**	4	**3**	5	13	14
		MD_G	2	1	**0**	**0**	**0**	**0**	**0**	0
		MD_L	12	**2**	**2**	4	**2**	5	13	14
		AD	0	0	0	0	0	0	0	0
		Bias	47%	**37%**	**37%**	41%	38%	41%	48%	48%
	Budget per top.: 30 (26%)	# rels.	1186	1327	**1359**	1289	1345	1267	1352	1337
		AA	347	337	337	**352**	338	350	281	345
		MA_{total}	203	203	180	199	**175**	207	283	243
		MA_G	70	80	80	**65**	79	67	136	72
		MA_L	133	122	101	134	**96**	140	147	171
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	28%	27%	23%	27%	**22%**	29%	34%	33%

top-k and NTCIR than the adjudication methods but, overall, low budgets and shallow pools do not lead to reliable enough results.

When it comes to AA, differently from TREC-8, they struggle to get close to the total number of significantly different pairs on the gold qrels. As in the TREC-8 case, an increase in the number of relevant documents found does not necessarily lead to an increase in the AA counts.

On a positive side, AD is always 0, also for DL 2021.

When it comes to MA, we observe two different patterns. Differently from TREC-8, MA_G is always quite high, motivating the general lack of Recall. In addition, MA_L does not substantially decrease as the budget increases, explaining the general lack of Precision.

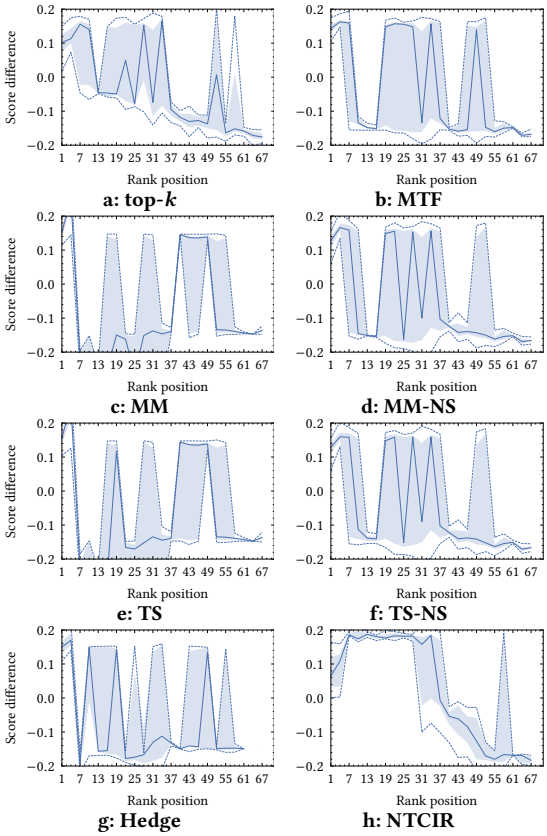

Figure 1: Distribution of MAP differences between systems in MA for a budget of 100 assessments (6%). The x-axis represents the systems sorted by their position in the official ranking. Each data point holds the distribution of 3 systems. The solid line represents the median of the bin. The shaded area is limited by the first and third quartiles of the distribution, i.e. it represents the inter-quartile range. Finally, the dashed lines are the maximum and the minimum. Breaks in the lines mean that there was not any mixed agreement for those systems. We used the 71 pooled systems of TREC-8.

Publication bias is exceedingly high, especially at low budgets, ranging between 25% and 50%. Overall, these high values shed a negative light on the reliability of the conclusions you would draw when using these methods under shallow pool conditions.

5.2 RQ2: How and where the methods fail

We study how and where, in terms of rank positions, the different methods fail in detecting significant differences.

We focus our analysis on the cases of mixed agreements (MA), which have shown to be the main factor for the loss of Precision and Recall. Figure 1 shows the distribution of the score differences in systems pairs which belong to MA with respect to their position in the gold ranking of systems for a budget of 100 assessments (6%). For each MA pair, we compute the difference between the score of the best and the worst system in the pair (under the adjudicated qrels, not the gold ones), recording it with a positive sign for the

best system and a negative one for the worst system. [2] Figure 1 tries to convey information about the distribution of such differences as a series of boxplots would do, but in a more compact and reabable way. The x-axis is the position of each system in the ranking of systems under the gold qrels, and we consider bins of three rank positions to make the figure more readable. For example, the first point in the figure represents the distribution of the mentioned differences for the first three systems in the gold ranking of systems. The solid line represents the median of the bin; the shaded area is limited by the first and third quartiles of the distribution, i.e. it represents the inter-quartile range; finally, the dashed lines are the maximum and the minimum. A break in the lines means that no pair of systems in that range of rank positions is a MA.

We can see some clear trends among all the evaluated methods. As a general trend for most adjudication methods, the biggest differences occur between MA systems in the middle of the ranking (we see wider areas in the middle of the ranking), whereas we see more narrow distributions in the top-ranked and lowest-ranked methods. This suggests that the MA, and the consequent loss of Precision, happen in a region of moderate impact, since mid-rank systems may receive less interest in any case. Top-k and NTCIR method represent two notable exceptions. Indeed, top-k concentrates most of the score differences in the top ranks; therefore, top-k is not only the less performing method (see Table 1 and Table 2) but it also fails in the most impactful region of the ranking. This is even worse for NTCIR, where the biggest differences (of 0.2 points), are all clustered in the top positions of the ranking.

5.3 RQ3: Evaluation of unseen systems

We investigate the *reusability* of the judgements produced by a low-cost method, i.e. their ability to fairly evaluate unseen systems. Usually, reusability is evaluated by following a *leave-one-group-out* approach. This consists in forming pools leaving one participating group each time and using those pools to evaluate the submissions of the group that was left out. We follow a different approach using the non-pooled systems of TREC-8. [3] To this aim, we performed the same experiments as in the previous sections, but using the non-pooled systems of TREC-8. In this way, we are evaluating systems that did not participate in the constructions of the pools. As commented in Section 4, this collection has been repeatedly acknowledged in the community as a high-quality one to evaluate unseen systems. Thus, we assume that the TREC-8 gold judgements are reusable and, if a low-cost method provides the same significant differences as them, we conclude that it is reusable as well.

Table 5 reports the Kendall's τ, Precision and Recall values of every method, for a varying number of assessments per topic, using the non-pooled systems. On a positive side, Table 5 shows similar trends as Table 1, suggesting that there is not a specific bias against non-pooled systems. On a slightly negative side, we observe that performance in Table 5 are generally slightly lower than those in Table 1, especially at the lowest budget, indicating a bit more loss and some more swaps due to not being pooled.

Table 5: Kendall's τ, Precision and Recall (see Section 3) of each adjudication method for a varying number of judgements per topic. 100 and 300 are the budget of judgements per topic. Parentheses indicate the size with respect to the full pool. We used the 58 non-pooled systems from TREC-8. For each budget, best values are bolded and worst ones underlined.

Method	MAP/100 (6%)			MAP/300 (17%)			NDCG/100 (6%)			NDCG/300 (17%)		
	τ	P	R	τ	P	R	τ	P	R	τ	P	R
top-k	0.82	0.931	0.903	0.91	0.948	0.966	0.83	0.941	0.880	0.90	0.966	0.943
MTF	0.88	0.934	0.933	0.95	0.968	0.988	0.89	0.941	0.916	0.94	0.980	0.968
MM	**0.91**	0.967	**0.942**	**0.97**	0.976	**0.997**	0.92	0.955	**0.946**	**0.97**	**0.983**	**0.979**
MM-NS	0.88	0.948	0.936	0.96	0.966	0.989	0.88	0.952	0.921	0.95	0.978	0.976
TS	**0.91**	0.969	0.940	**0.97**	0.973	0.996	0.92	0.956	0.944	**0.97**	0.979	0.977
TS-NS	0.87	0.945	0.933	0.95	0.966	0.986	0.88	0.952	0.918	0.94	0.979	0.974
Hedge	**0.91**	**0.973**	0.929	0.96	**0.980**	0.982	**0.93**	**0.974**	**0.946**	0.96	0.977	0.977
NTCIR	0.89	0.898	0.931	0.95	0.962	0.984	0.86	0.938	0.911	0.94	0.974	0.977

More in detail, TS, MM and Hedge always have the highest correlation scores and while MM achieves always the best Recall, independently from the budget and the metric. This means that if we were to gather the judgements of a new collection, MM would be the best option in terms of reusability of the collected assessments. As before, top-k and NTCIR method lag behind the other methods in all the cases and for every considered measure. This finding suggests that other alternative methods might be a better option to gather assessments when constructing new experimental collections.

Table 6 reports the agreements for the non-pooled systems, similarly to Table 2 for the pooled ones. [4] The results follow the same trends as with the pooled systems, further supporting the lack of strong biases against non-pooled systems. These scores confirm that alternative adjudication methods are more effective than top-k, which, contrary to what we observed in Table 2, now is clearly the worst method. As before, the more relevant documents found does not necessarily mean the more AA; therefore, not all the relevant documents are equally discriminative also for non-pooled systems.

No method fails in a mixed or active disagreement when evaluating the non-pooled systems. This further supports the fact that most drops in Kendall's τ are due to swaps between systems that are not significantly different under the gold qrels.

When it comes to the publication bias, we observe similar trends as in the case of the pooled systems, even with lower values, indicating that published conclusions would not change also in the case of non-pooled systems.

Finally, we can observe similar trends between the results obtained with AP and those obtained with NDCG, supporting the fact that the results presented here are generalizable in terms of the evaluation of unseen systems, and that they are not an artefact of the evaluation metric used.

6 CONCLUSIONS AND FUTURE WORK

We argued for the need of a more powerful way of evaluating adjudication methods. In particular, while the current approach just focuses on how close two alternative methods rank systems,

[2] For example, if we have the pair of system1 and system2 in mixed agreement, and system1 has the highest score, and their score difference is 0.15 (with the reduced pool). Then, for system1 we record 0.15 and for system2 we save −0.15. The mentioned figure plots the distribution of these differences for each system, according to their position in the ranking induced with the gold qrels.

[3] We do not perform these experiments on the DL21 collection since it does not include non-pooled runs.

[4] Note that the # rels. row is the same as before since the pools are the same, we are only changing the systems we are evaluating.

Table 6: Relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size with respect to the full pool. We used the 58 non-pooled systems from TREC-8. The top-100 full pool includes 4728 relevant documents. There are 1653 pairwise comparisons, of which 509 are significant under the gold qrels with MAP (upper half), and 527 with NDCG (lower half). For each budget, the best values are bolded and the worst ones are underlined.

	Metric	Adjudication method							
		top-k	MTF	MM	MM-NS	TS	TS-NS	Hedge	NTCIR
MAP (509 gold significantly different pairs) Budget per top.: 100 (6%)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
	AA	460	475	**480**	477	479	475	473	474
	MA_{total}	83	68	**45**	58	**45**	62	49	89
	MA_G	49	34	**29**	32	30	34	36	35
	MA_L	34	34	16	26	15	28	**13**	54
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
	Bias	**7%**	**7%**	3%	5%	3%	5%	3%	4%
MAP (509 gold significantly different pairs) Budget per top.: 300 (17%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
	AA	492	503	**508**	504	507	502	500	501
	MA_{total}	44	23	**13**	23	16	25	19	28
	MA_G	17	6	**1**	5	2	7	9	8
	MA_L	27	17	12	18	14	18	**10**	20
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
	Bias	5%	3%	**2%**	3%	3%	3%	**2%**	4%
NDCG (527 gold significantly different pairs) Budget per top.: 100 (6%)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
	AA	464	483	**499**	486	498	484	**499**	480
	MA_{total}	92	74	52	65	52	67	**41**	79
	MA_G	63	44	**28**	41	29	43	**28**	47
	MA_L	29	30	23	24	23	24	**13**	32
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
	Bias	6%	6%	4%	5%	4%	5%	**3%**	6%
NDCG (527 gold significantly different pairs) Budget per top.: 300 (17%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
	AA	497	510	**516**	514	515	514	515	515
	MA_{total}	47	27	**19**	24	23	24	24	26
	MA_G	30	17	**11**	13	12	13	12	12
	MA_L	17	10	**9**	11	11	11	12	14
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
	Bias	3%	**2%**	**2%**	**2%**	**2%**	**2%**	**2%**	3%

quantified by Kendall's τ, we think that we should focus our attention also on how different methods behave with respect to the significantly different pairs of systems detected. Indeed, while the current approach looks for stability in answering the question "is system A better than B?", our proposed method looks for stability in answering the question "is system A significantly better than B?", which is the ultimate questions researchers are interested in to ensure generalizability of results.

To this end, we considered two measures—namely Precision and Recall—which consider significantly different pairs in isolation, as well as measures—the agreement/disagreement counts—which relate them to swaps in the ranking of systems. We also considered the problem of the publication bias, i.e. the chance of publishing results/conclusions that would not hold or be the opposite when using the full pool instead of a reduced one.

To both validate and to showcase our proposed approach, we conducted a thorough experimentation on TREC-8, a collection renown for its high quality deep pool, and TREC Deep Learning 2021, a collection adopting a very shallow pool. In this way, we have shown that our methodology allows us to obtain insights not possible simply using Kendall's τ.

For example, we found that no active disagreements (AD) and (almost) no mixed disagreements (MD) happen. This means that observed drops in Kendall's τ are mostly due to swaps between not significantly different systems. Therefore, those drops concerns not very interesting system pairs, and it might not be worth to strive for (or to judge a method just by) 1.00 Kendall's τ.

We also found that the number of relevant documents detected by a method does not necessarily increase the number of significantly different pairs detected, suggesting that not all the relevant documents in a pool are equally discriminative. This opens up interesting future investigations on which (relevant) documents would be optimal for a pool while the current focus has been more on determining how many and which topics to sample.

We have shown that drops in Precision and Recall are caused by mixed agreements (MA) which distribute unevenly at different rank positions and, therefore, they have a quite different impact: those happening at mid-to-bottom rank positions are less serious than those happening at the top positions of the ranking.

Finally, we also found that no adjudication methods induces strong biases against non-pooled systems, thus further supporting the use of these methods to construct new test collections for IR evaluation. Previous work evaluated the reusability of bandit-based methods using Kendall's τ and other swap-based measures, and concluded that the collections built with them were less reusable than desirable. With the new evaluation approach we have presented in this paper, we shed some more light on this issue and show that, when focusing on significance between systems, bandit-based method are indeed reusable.

Overall, our approach allowed us to show that existing methods for human assessment adjudication in IR evaluation could preserve most of the true statistical differences between the pairwise comparisons of systems. Besides this, as discussed in detail, our approach allowed us to pinpoint which adjudication method works better in specific conditions, why, and how it is different from other methods. This will thus be a helpful tool and guidance for researchers, when they have to decide which method to choose in their settings.

ACKNOWLEDGMENTS

This work has received support from: (i) project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next GenerationEU); (ii) Programa de Ayudas para la Formación de Profesorado Universitario, grant number FPU20/02659 (Ministerio de Universidades); (iii) project PID2022-137061OB-C21 (Proyectos de Generación de Conocimiento, MCIN); (iv) project ED431-B 2022/33 (Xunta de Galicia/ERDF); (v) CAMEO, PRIN 2022 n. 2022ZLL7MW.

REFERENCES

[1] James Allan, Donna K. Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M. Voorhees. 2017. TREC 2017 common core track overview. In *Proceedings of TREC 2017*. NIST Special Publication 500-324, Gaithersburg, Maryland, USA. https://trec.nist.gov/pubs/trec26/papers/Overview-CC.pdf (cited on p. 2).

[2] Bahadir Altun and Mucahid Kutlu. 2020. Building test collections using bandit techniques: a reproducibility study. In *Proceedings of ACM CIKM 2020*. ACM, New York, NY, USA, 1953–1956. DOI: 10.1145/3340531.3412121 (cited on p. 4).

[3] Javed A. Aslam, Virgil Pavlu, and Robert Savell. 2003. A unified model for metasearch, pooling, and system evaluation. In *Proceedings of ACM CIKM 2003*. ACM, New York, NY, USA, 484–491. DOI: 10.1145/956863.956953 (cited on p. 4).

[4] D. Banks, Paul Over, and N.-F. Zhang. 1999. Blind men and elephants: six approaches to TREC data. *Information Retrieval Journal*, 1, 1, (May 1999), 7–34. DOI: 10.1023/A:1009984519381 (cited on p. 3).

[5] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of ACM SIGIR 2000*. ACM, New York, NY, USA, 33–40. DOI: 10.1145/345508.345543 (cited on p. 6).

[6] Chris Buckley and Ellen M. Voorhees. 2005. Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval*. Ellen M. Voorhees and Donna K. Harman, (Eds.) The MIT Press, 53–78 (cited on p. 4).

[7] Ben Carterette. 2017. But is it statistically significant? statistical significance in ir research, 1995-2014. In *Proceedings of ACM SIGIR 2017*. ACM, New York, NY, USA, 1125–1128. DOI: 10.1145/3077136.3080738 (cited on p. 3).

[8] Ben Carterette. 2012. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems*, 30, 1, (Mar. 2012). DOI: 10.1145/2094072.2094076 (cited on pp. 3, 4).

[9] Gordon V. Cormack and Thomas R. Lynam. 2006. Statistical precision of information retrieval evaluation. In *Proceedings of ACM SIGIR 2006*. ACM, New York, NY, USA, 533–540. DOI: 10.1145/1148170.1148262 (cited on p. 3).

[10] Gordon V. Cormack and Thomas R. Lynam. 2007. Validity and power of t-test for comparing map and gmap. In *Proceedings of ACM SIGIR 2007*. ACM, New York, NY, USA, 753–754. DOI: 10.1145/1277741.1277892 (cited on p. 3).

[11] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. In *Proceedings of ACM SIGIR 1998*. ACM, New York, NY, USA, 282–289. DOI: 10.1145/290941.291009 (cited on pp. 2, 4).

[12] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Overview of the TREC 2021 deep learning track. In *Proceedings of TREC 2021*. NIST Special Publication 500-335, Gaithersburg, Maryland, USA. https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf (cited on p. 4).

[13] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, and Ian Soboroff. 2021. TREC deep learning track: reusable test collections in the large data regime. In *Proceedings of ACM SIGIR 2021*. ACM, New York, NY, USA, 2369–2375. DOI: 10.1145/3404835.3463249 (cited on p. 2).

[14] Guglielmo Faggioli and Nicola Ferro. 2021. System effect estimation by sharding: a comparison between anova approaches to detect significant differences. In *Proceedings of 43rd European Conference on IR Research* (ECIR '21). Springer International Publishing, Cham. ISBN: 978-3-030-72240-1. DOI: 10.1007/978-3-030-72240-1_3 (cited on p. 3).

[15] Nicola Ferro and Mark Sanderson. 2022. How do you test a test? a multifaceted examination of significance tests. In *Proceedings of ACM WSDM 2022*. ACM, New York, NY, USA, 280–288. DOI: 10.1145/3488560.3498406 (cited on p. 3).

[16] Nicola Ferro and Mark Sanderson. 2019. Improving the accuracy of system performance estimation by using shards. In *Proceedings of ACM SIGIR 2019*. ACM, New York, NY, USA, 805–814. DOI: 10.1145/3331184.3338062 (cited on p. 3).

[17] Norbert Fuhr. 2018. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51, 3, (Feb. 2018). DOI: 10.1145/3190580.3190586 (cited on p. 2).

[18] Donna Harman. 2011. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3, 2, (May 2011), 1–119. DOI: 10.2200/S00368ED1V01Y201105ICR019 (cited on p. 1).

[19] Claudia Hauff, Djoerd Hiemstra, Franciska de Jong, and Leif Azzopardi. 2009. Relying on topic subsets for system ranking estimation. In *Proceedings of ACM CIKM 2009*. ACM, New York, NY, USA, 1859–1862. DOI: 10.1145/1645953.1646249 (cited on p. 6).

[20] Yosef Hochberg and Ajit C. Tamhane. 1987. *Multiple Comparison Procedures*. John Wiley & Sons, USA. DOI: 10.1002/9780470316672 (cited on p. 4).

[21] Mehdi Hosseini, Ingemar J. Cox, Natasa Milic-Frayling, Milad Shokouhi, and Emine Yilmaz. 2012. An uncertainty-aware query selection model for evaluation of IR systems. In *Proceedings of ACM SIGIR 2012*. ACM, New York, NY, USA, 901–910. DOI: 10.1145/2348283.2348403 (cited on p. 6).

[22] Jason C. Hsu. 1996. *Multiple Comparisons. Theory and methods*. Chapman and Hall/CRC, USA. DOI: 10.1201/b15074 (cited on p. 4).

[23] David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of ACM SIGIR 1993*. ACM, New York, NY, USA, 329–338. DOI: 10.1145/160688.160758 (cited on p. 3).

[24] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20, 4. DOI: 10.1145/582415.582418 (cited on p. 4).

[25] Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30, 1/2, 81–93. DOI: 10.2307/2332226 (cited on p. 2).

[26] Maurice G. Kendall. 1948. *Rank Correlation Methods*. Chales Griffin and Company Limited (cited on p. 2).

[27] Dan Li and Evangelos Kanoulas. 2017. Active sampling for large-scale information retrieval evaluation. In *Proceedings of ACM CIKM 2017*. ACM, New York, NY, USA, 49–58. DOI: 10.1145/3132847.3133015 (cited on p. 2).

[28] David E. Losada, Javier Parapar, and Álvaro Barreiro. 2016. Feeling lucky?: multi-armed bandits for ordering judgements in pooling-based evaluation. In *Proceedings of ACM SAC 2016*. ACM, New York, NY, USA, 1027–1034. DOI: 10.1145/2851613.2851692 (cited on p. 4).

[29] David E. Losada, Javier Parapar, and Álvaro Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management*, 53, 5, (Sept. 2017), 1005–1025. DOI: 10.1016/j.ipm.2017.04.005 (cited on pp. 2, 4).

[30] Stefano Mizzaro and Stephen Robertson. 2007. Hits hits TREC: exploring IR evaluation results with network analysis. In *Proceedings of ACM SIGIR 2007*. ACM, New York, NY, USA, 479–486. DOI: 10.1145/1277741.1277824 (cited on p. 6).

[31] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In *Proceedings of ADCS 2012*. ACM, New York, NY, USA, 47–54. DOI: 10.1145/2407085.2407092 (cited on p. 3).

[32] Alistair Moffat, William Webber, and Justin Zobel. 2007. Strategic system comparisons via targeted relevance judgments. In *Proceedings of ACM SIGIR 2007*. ACM, New York, NY, USA, 375–382. DOI: 10.1145/1277741.1277806 (cited on p. 2).

[33] Javier Parapar, David E. Losada, and Álvaro Barreiro. 2021. Testing the tests: simulation of rankings to compare statistical significance tests in information retrieval evaluation. In *Proceedings of ACM SAC 2021*. ACM, New York, NY, USA, 655–664. DOI: 10.1145/3412841.3441945 (cited on p. 3).

[34] Javier Parapar, David E. Losada, Manuel A. Presedo-Quindimil, and Álvaro Barreiro. 2020. Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 71, 1, 98–113. DOI: 10.1002/asi.24203 (cited on p. 3).

[35] Md Mustafizur Rahman, Mucahid Kutlu, and Matthew Lease. 2019. Constructing test collections using multi-armed bandits and active learning. In *Proceedings of The World Wide Web Conference 2019*. ACM, New York, NY, USA, 3158–3164. DOI: 10.1145/3308558.3313675 (cited on p. 2).

[36] Kevin Roitero, J. Shane Culpepper, Mark Sanderson, Falk Scholer, and Stefano Mizzaro. 2020. Fewer topics? a million topics? both?! on topics subsets in test collections. *Information Retrieval Journal*, 23, 1, (Feb. 2020), 49–85. DOI: 10.1007/s10791-019-09357-w (cited on p. 6).

[37] Tetsuya Sakai. 2018. *Laboratory Experiments in Information Retrieval. The Information Retrieval Series*. Vol. 40. Springer. ISBN: 978-981-13-1198-7. DOI: 10.1007/978-981-13-1199-4 (cited on p. 4).

[38] Tetsuya Sakai. 2021. On Fuhr's guideline for IR evaluation. *SIGIR Forum*, 54, 1, Article 12, (Feb. 2021). DOI: 10.1145/3451964.3451976 (cited on p. 2).

[39] Tetsuya Sakai. 2016. Statistical significance, power, and sample sizes: a systematic review of sigir and tois, 2006-2015. In *Proceedings of ACM SIGIR 2016*. ACM, New York, NY, USA, 5–14. DOI: 10.1145/2911451.2911492 (cited on p. 3).

[40] Tetsuya Sakai. 2016. Topic set size design. *Information Retrieval Journal*, 19, 3, (June 2016), 256–283. DOI: 10.1007/s10791-015-9273-z (cited on p. 6).

[41] Tetsuya Sakai, N. Kando, Chuan-Jie Lin, Teruko Mitamura, Hideki Shima, Dong-Hong Ji, Kuang-hua Chen, and Eric Nyberg. 2008. Overview of the NTCIR-7 ACLIA IR4QA Task. In *Proceedings of the NTCIR-7* (cited on p. 4).

[42] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4, 4, (June 2010), 247–375. DOI: 10.1561/1500000009 (cited on p. 1).

[43] Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of ACM SIGIR 2005*. ACM, New York, NY, USA, 162–169. DOI: 10.1145/1076034.1076064 (cited on pp. 3, 6).

[44] Jacques Savoy. 1997. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33, 4, (July 1997), 495–512. DOI: 10.1016/S0306-4573(97)00027-7 (cited on p. 3).

[45] K. Spärck Jones and Cornelis J. van Rijsbergen. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. *Computer Laboratory* (cited on p. 1).

[46] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. (Second ed.). *Adaptive computation and Machine Learning*. MIT Press. ISBN: 9780262039246 (cited on p. 2).

[47] Julián Urbano, Harlley Lima, and Alan Hanjalic. 2019. Statistical significance testing in information retrieval: an empirical analysis of type I, type II and type III errors. In *Proceedings of the ACM SIGIR 2019*. ACM, New York, NY, USA, 505–514. DOI: 10.1145/3331184.3331259 (cited on p. 3).

[48] Julián Urbano, Mónica Marrero, and Diego Martín. 2013. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proceedings of ACM SIGIR 2013*. ACM, New York, NY, USA, 925–928. DOI: 10.1145/2484028.2484163 (cited on p. 3).

[49] Ellen M. Voorhees. 2018. On building fair and reusable test collections using bandit techniques. In *Proceedings of ACM CIKM 2018*. ACM, New York, NY, USA, 407–416. DOI: 10.1145/3269206.3271766 (cited on p. 2).

[50] Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *CLEF 2001*. Springer, Berlin, Heidelberg, 355–370. DOI: 10.1007/3-540-45691-0_34 (cited on p. 1).

[51] Ellen M. Voorhees. 2009. Topic set size redux. In *Proceedings of ACM SIGIR 2009*. ACM, New York, NY, USA, 806–807. DOI: 10.1145/1571941.1572138 (cited on p. 6).

[52] Ellen M. Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of ACM SIGIR 2002*. ACM, New York, NY, USA, 316–323. DOI: 10.1145/564376.564432 (cited on p. 6).

[53] Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too many relevants, whither cranfield test collections? In *Proceedings of ACM SIGIR 2022*. ACM, New York, NY, USA, 11 pages. DOI: 10.1145/3477495.3531728 (cited on p. 2).

[54] Ellen M. Voorhees and Donna Harman. 2000. Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of TREC-8*. NIST Special Publication 500-246, Gaithersburg, Maryland, USA, 1–24. DOI: 10.6028/NIST.SP.500-246 (cited on p. 4).

[55] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press. ISBN: 0262220733 (cited on pp. 1, 4).

[56] Ellen M. Voorhees, Ian Soboroff, and Jimmy Lin. 2022. Can old TREC collections reliably evaluate modern neural retrieval models? https://arxiv.org/abs/2201.11086 arXiv: 2201.11086 (cited on pp. 2, 4).

[57] William Webber, Alistair Moffat, and Justin Zobel. 2008. Statistical power in retrieval experimentation. In *Proceedings of ACM CIKM 2008*. ACM, New York, NY, USA, 57100580. DOI: 10.1145/1458082.1458158 (cited on p. 3).

[58] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of ACM SIGIR 2008*. ACM, New York, NY, USA, 587–594. DOI: 10.1145/1390334.1390435 (cited on p. 2).

[59] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of ACM SIGIR 1998*. ACM, New York, NY, USA, 307–314. DOI: 10.1145/290941.291014 (cited on p. 2).