CENTRE@CLEF2019: Overview of the Replicability and Reproducibility Tasks

Nicola Ferro[1], Norbert Fuhr[2], Maria Maistro[1,3], Tetsuya Sakai[4], and Ian Soboroff[5]

[1] University of Padua, Italy
{ferro,maistro}@dei.unipd.it
[2] University of Duisburg-Essen, Germany
norbert.fuhr@uni-due.de
[3] University of Copenhagen, Denmark
mm@di.ku.dk
[4] Waseda University, Japan
tetsuyasakai@acm.org
[5] National Institute of Standards and Technology (NIST), USA
ian.soboroff@nist.gov

Abstract. Reproducibility has become increasingly important for many research areas, among those IR is not an exception and has started to be concerned with reproducibility and its impact on research results. This paper describes our second attempt to propose a lab on reproducibility named CENTRE, held during CLEF 2019. The aim of CENTRE is to run both a replicability and reproducibility challenge across all the major IR evaluation campaigns and to provide the IR community with a venue where previous research results can be explored and discussed. This paper reports the participant results and preliminary considerations on the second edition of CENTRE@CLEF 2019.

1 Introduction

Reproducibility is becoming a primary concern in many areas of science [16, 24] as well as in computer science, as also witnessed by the recent ACM policy on result and artefact review and badging.

Also in *Information Retrieval (IR)* replicability and reproducibility of the experimental results are becoming a more and more central discussion items in the research community [4, 12, 17, 23, 28]. We now commonly find questions about the extent of reproducibility of the reported experiments in the review forms of all the major IR conferences, such as SIGIR, CHIIR, ICTIR and ECIR, as well as journals, such as ACM TOIS. We also witness to the raise of new activities

aimed at verifying the reproducibility of the results: for example, the "Reproducibility Track" at ECIR since 2015 hosts papers which replicate, reproduce and/or generalize previous research results.

Nevertheless, it has been repeatedly shown that best TREC systems still outperform off-the-shelf open source systems [4–6, 22, 23]. This is due to many different factors, among which lack of tuning on a specific collection when using default configuration, but it is also caused by the lack of the specific and advanced components and resources adopted by the best systems.

It has been also shown that additivity is an issue, since adding a component on top of a weak or strong base does not produce the same level of gain [6, 22]. This poses a serious challenge when off-the-shelf open source systems are used as stepping stone to test a new component on top of them, because the gain might appear bigger starting from a weak baseline.

Moreover, besides the problems encountered in replicating/reproducing research, we lack any well established measure to assess and quantify the extent to which something has been replicated/reproduced. In other terms, even if a later researcher can manage to replicate or reproduce an experiment, to which extent can we claim that the experiment is successfully replicated or reproduced? For the replicability task we can compare the original measure score with the score of the replicated run, as done in [15, 14]. However, this can not be done for reproducibility, since the reproduced system is obtained on a different data set and it is not directly comparable with the original system in terms of measure scores.

Finally, both a Dagstuhl Perspectives Workshop [11] and the recent SWIRL III strategic workshop [1] have put on the IR research agenda the need to develop both better explanatory models of IR system performance and new predictive models, able to anticipate the performance of IR systems in new operational conditions.

Overall, the above considerations stress the need and urgency for a systematic approach to reproducibility and generalizability in IR. Therefore, the goal of *CLEF, NTCIR, TREC REproducibility (CENTRE)* at CLEF 2019 is to run a joint CLEF/NTCIR/TREC task on challenging participants:

- to replicate and reproduce best results of best/most interesting systems in previous editions of CLEF/NTCIR/TREC by using standard open source IR systems;
- to contribute back to the community the additional components and resources developed to reproduce the results in order to improve existing open source systems;
- to start exploring the generalizability of our findings and the possibility of predicting IR system performance;
- to investigate possible measures for replicability and reproducibility in IR.

The paper is organized as follows: Section 2 introduces the setup of the lab; Section 3 discusses the participation and the experimental outcomes; and, Section 4 draws some conclusions and outlooks possible future works.

2 Evaluation Lab Setup

2.1 Tasks

Similarly to its previous edition, CENTRE@CLEF 2019 offered the following two tasks:

- *Task 1 - Replicability*: the task focuses on the replicability of selected methods on the same experimental collections;
- *Task 2 - Reproducibility*: the task focuses on the reproducibility of selected methods on different experimental collections;

For *Replicability* and *Reproducibility* we refer to the ACM Artifact Review and Badging definitions[6]:

- *Replicability* (different team, same experimental setup): the measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts. In CENTRE@CLEF 2019 this meant to use the same collections, topics and ground-truth on which the methods and solutions have been developed and evaluated.
- *Reproducibility* (different team, different experimental setup): The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently. In CENTRE@CLEF 2019 this meant to use a different experimental collection, but in the same domain, from those used to originally develop and evaluate a solution.

For Task 1 and Task 2, CENTRE@CLEF 2019 teams up with the *Open-Source IR Replicability Challenge (OSIRRC)* [10] at SIGIR 2019. Therefore, participating groups could consider to submit their runs both to CENTRE@CLEF 2019 and OSIRRC 2019, where the second venue requires to submit the runs as Docker images.

Besides Task 1 and Task 2, CENTRE@CLEF 2019 offered also a new pilot task:

- *Task 3 - Generalizability*: the task focuses on collection performance prediction and the goal is to rank (sub-)collections on the basis of the expected performance over them.

In details, Task 3 was instatiated as follows:

[6] `https://www.acm.org/publications/policies/artifact-review-badging`

– *Training*: participants need to run plain BM25 and, if they wish, also their own system on the test collection used for TREC 2004 Robust Track (they are allowed to use the corpus, topics and qrels). Participants need to identify features of the corpus and topics that allow them to predict the system score with respect to *Average Precision (AP)*.
– *Validation*: participants can use the test collection used for TREC 2017 Common Core Track (corpus, topics and qrels) to validate their method and determine which set of features represent the best choice for predicting AP score for each system. Note that the TREC 2017 Common Core Track topics are an updated version of the TREC 2004 Robust track topics.
– *Test* (submission): participants need to use the test collection used for TREC 2018 Common Core Track (only corpus and topics). Note that the TREC 2018 Common Core Track topics are a mix of "old" and "new" topics, where old topics were used in TREC 2017 Common Core track. Participants will submit a run for each system (BM25 and their own system) and an additional file (one for each system) including the AP score predicted for each topic. The score predicted can be a single value or a value with the corresponding confidence interval.

2.2 Replicability and Reproducibility Targets

For the previous edition of CENTRE@CLEF 2018 [15, 14] we selected the target runs for replicability and reproducibility among the Ad Hoc tasks in previous editions of CLEF, TREC, and NTCIR. However, even though CENTRE@CLEF 2018 had 17 enrolled teams, eventually only one team managed to submit a run. One of the main issues reported by the participating team is the lack of the external resources exploited in the original paper, which are no longer available [19]. Therefore, for CENTRE@CLEF 2019 we decided to focus on more recent papers submitted at TREC Common Core Track in 2017 and 2018.

To select the target runs from the TREC 2017 and 2018 Common Core Tracks we did not consider the impact of the proposed approaches in terms of number of citations, since both the tracks are recent and the citations received by the submitted papers are not significant. Therefore, we looked at the final ranking of runs reported in the tracks overviews [2, 3] and we chose the best performing runs which exploit open source search systems and do not make use of additional relevance assessments, which are not available to different teams.

Below we list the runs selected as targets of replicability and reproducibility among which the participants can choose. For each run, we specify the corresponding collection for replicability and for reproducibility. For more information, the list also provides references to the papers describing those runs as well as the overviews describing the overall task and collections.

– **Runs**: `WCrobust04` and `WCrobust0405` [18]
  • **Task Type**: TREC 2017 Common Core Track [2]
  • **Replicability**: New York Times Annotated Corpus, with TREC 2017 Common Core Topics

    - **Reproducibility**: TREC Washington Post Corpus, with TREC 2018 Common Core Topics
- **Runs**: `RMITFDA4` and `RMITEXTGIGADA5` [7]
    - **Task Type**: TREC 2018 Common Core Track [3]
    - **Replicability**: TREC Washington Post Corpus, with TREC 2018 Common Core Topics
    - **Reproducibility**: New York Times Annotated Corpus, with TREC 2017 Common Core Topics

Since these runs were not originally thought for being used as targets of a replicability/reproducibility exercise, we contacted the authors of the papers to inform them and ask their consent to use the runs.

The participants in CENTRE@CLEF 2019 were not provided with the corpora necessary to perform the tasks. The following collections were needed to perform the task:

- *The New York Times Annotated Corpus*[7] contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007. The text in this corpus is formatted in News Industry Text Format (NITF), which is an XML specification that provides a standardized representation for the content and structure of discrete news articles. The dataset is available upon payment of a fee.
- *The TREC Washington Post Corpus*[8] contains 608 180 news articles and blog posts from January 2012 through August 2017. The articles are stored in JSON format, and include title, byline, date of publication, kicker (a section header), article text broken into paragraphs, and links to embedded images and multimedia. The dataset is publicly available and free of charge.
- *The TREC 2004 Robust Corpus*[9] corresponds to the set of documents on TREC disks 4 and 5, minus the Congressional Record. This document set contains approximately 528 000 documents. The dataset is available upon payment of a fee.

Finally, Table 1 reports the topics used for the three tasks, with the corresponding number of documents and pool sizes. An example of topic is reported in the Figure 1 for TREC 2018 Common Core Track.

2.3 Evaluation Measures

Task 1 - Replicability: As done in the previous edition of CENTRE [15, 14], the quality of the replicability runs has been evaluated from two points of view:

[7] `https://catalog.ldc.upenn.edu/LDC2008T19`
[8] `https://trec.nist.gov/data/wapost/`
[9] `https://trec.nist.gov/data/qa/T8_QAdata/disks4_5.html`

Table 1. Topics used for the first edition of CENTRE@CLEF 2019 with the number of documents in the pool.

Evaluation Campaign	Track	# Topics	Pool Size
TREC 2018	Common Core	50	26 233
TREC 2017	Common Core	50	30 030
TREC 2004	Robust	250	311 410

```
<top>
<num> Number: 321 </num>
<title>
Women in Parliaments
</title>
<desc> Description:
Pertinent documents will reflect the fact that women continue to be poorly represented in parliaments across the world, and the gap in
political power between the sexes is very wide, particularly in the Third World.
</desc>
<narr> Narrative
Pertinent documents relating to this issue will discuss the lack of representation by women, the countries that mandate the inclusion of a
certain percentage of women in their legislatures, decreases if any in female representation in legislatures, and those countries in which
there is no representation of women.
</narr>
</top>
```

Fig. 1. Example of a topic for TREC 2018 Common Core Track.

– *Effectiveness*: how close are the performance scores of the replicated systems to those of the original ones. This is measured using the *Root Mean Square Error (RMSE)* [21] between the new and original measures scores $M(\cdot)$:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\big(M_{orig,i} - M_{replica,i}\big)^2} \tag{1}$$

where T is the total number of topics, $M_{orig,i}$ is the measure score of the original target run on topic t_i and $M_{replica,i}$ is the measure score of the replicated run on topic t. Equation (1) is instantiated with AP, *Normalized Discounted Cumulated Gain (nDCG)* and *Expected Reciprocal Rank (ERR)*.

– *Ranked result lists*: since different result lists may produce the same effectiveness score, we also measure how close are the ranked results list of the replicated systems to those of the original ones. This is measured using Kendall's τ correlation coefficient [20] among the list of retrieved documents for each topic, averaged across all the topics. The Kendall's τ correlation coefficient on a single topic is given by:

$$\begin{aligned} \tau_i\big(orig, replica\big) &= \frac{P - Q}{\sqrt{\big(P + Q + U\big)\big(P + Q + V\big)}} \\ \bar{\tau}\big(orig, replica\big) &= \frac{1}{T}\sum_{i=1}^{T}\tau_i\big(orig, replica\big) \end{aligned} \tag{2}$$

where T is the total number of topics, P is the total number of concordant pairs (document pairs that are ranked in the same order in both vectors) Q the total number of discordant pairs (document pairs that are ranked

in opposite order in the two vectors), U and V are the number of ties, respectively, in the first and in the second ranking.

Note that the definition of Kendall's τ in Equation (2) is originally proposed for permutations of the same set of items, therefore it is not applicable whenever two rankings do not contain the same set of documents. However, for real rankings of systems it is highly likely that two lists do not contain the same set of items, thus we performed some pre-processing with the runs before computing Kendall's τ in Equation (2).

In details, consider a fixed topic t, the original ranking $r_{t,orig}$ and the replicated ranking $r_{t,replica}$. If one of the rankings contains a document that is not retrieved by the other ranking, we define the rank position of that document as zero. For example, if for a document d, $d \in r_{t,orig}$, but $d \notin r_{t,replica}$, then the rank position of d in $r_{t,replica}$ is zero. Whenever the two rankings contains the same set of documents, Equation (2) is not affected by this pre-processing step and the computation of Kendall's *tau* is performed as usual. Furthermore, if two rankings retrieves different documents and place them in the same rank positions, Kendall's tau will still be equal to 1, and the comparison is performed just with respect to the relative order of the documents retrieved by both the rankings.

Task 2 - Reproducibility: Since for the reproducibility runs we do not have an already existing run to compare against, we compare the reproduced run score with respect to a baseline run, to see whether the improvement over the baseline is comparable between the original collection C and the new collection D. In particular we compute the *Effect Ratio (ER)*, which is also exploited in CENTRE@NTCIR 14 [25].

In details, given two runs, we refer to the A-run, as the advanced run, and B-run, as the baseline run, where the A-run has been reported to outperform the B-run on the original test collection C. The intuition behind ER is to evaluate to which extent the improvement on the original collection C is reproduced on a new collection D. For any evaluation measure M, let $M_i^C(A)$ and $M_i^C(B)$ denote the score of the A-run and that of the B-run for the i-th topic of collection C ($1 \leq i \leq T_C$). Similarly, let $M_i^D(A')$ and $M_i^D(B')$ denote the scores for the reproduced A-run and B-run respectively, on the new collection D. Then, ER is computed as follows:

$$\mathrm{ER}(\Delta M^D_{reproduced}, \Delta M^C_{orig}) = \frac{\frac{1}{T_D}\sum_{i=1}^{T_D} \Delta M^D_{i,reproduced}}{\frac{1}{T_C}\sum_{i=1}^{T_C} \Delta M^C_{i,orig}} \tag{3}$$

where $\Delta M^C_{i,orig} = M_i^C(A) - M_i^C(B)$ is the per-topic improvement of the original advanced and baseline runs for the i-th topic on C. Similarly $\Delta M^D_{i,reproduced} = M_i^C(A') - M_i^C(B')$ is the per-topic improvement of the reproduced advanced and baseline runs for the i-th topic on D. Note that the per-topic improvement can be negative, for those topics where the advanced run fails to outperform the baseline run.

Table 2. Path to the submitted runs files in the online repository with their description and the number of assessed topics included in each run.

Run Path	Description	# Topics
official/task1/irc_task1_WCrobust04_001	official run, replicating WCrobust04	50
official/task1/irc_task1_WCrobust0405_001	official run, replicating WCrobst0405	33
official/task2/irc_task2_WCrobust04_001	official run, reproducing WCrobust05	25
official/task2/irc_task2_WCrobust0405_001	official run, reproducing WCrobust0405	15
unofficial/complete_topics/task1/irc_task1_WCrobust04_001	unofficial run, replicating WCrobust04	50
unofficial/complete_topics/task1/irc_task2_WCrobust04_001	unofficial run, replicating WCrobust0405	50
unofficial/complete_topics/task2/irc_task2_WCrobust04_001	unofficial run, reproducing WCrobust05	25
unofficial/complete_topics/task2/irc_task2_WCrobust0405_001	unofficial run, reproducing WCrobust0405	25

If ER ≤ 0, that means that the replicated A-run failed to outperform the replicated B-run: the replication is a complete failure. If $0 <$ ER < 1, the replication is somewhat successful, but the effect is smaller compared to the original experiment. If ER $= 1$, the replication is perfect in the sense that the original effect has been recovered as is. If ER > 1, the replication is successful, and the effect is actually larger compared to the original experiment.

Finally, ER in Equation (3) is instantiated with respect to AP, nDCG and ERR. Furthermore, as suggested in [25], ER is computed even for the replicability task, by replacing $\Delta M^D_{i,reproduced}$ with $\Delta M^C_{i,replica}$ in Equation (3).

Task 3 - Generalizability: For the generalizability task we planned to compare the predicted run score with the original run score. This is measured with Mean Absolute Error and RMSE between the predicted and original measures scores, with respect to AP, nDCG and ERR. However, we did not receive any run for the generalizability task, so we did not put in practice this part of the evaluation task.

3 Participation and Outcomes

19 groups registered for participating in CENTRE@CLEF2019, but unfortunately only one group succeeded in submitting two replicability runs and two reproducibility runs. No runs were submitted for the generalizability task.

The team from the University of Applied Science TH Köln [8] replicated and reproduced the runs by Grossman and Cormack [18], i.e. `WCrobust04` and `WCrobust0405`. They could not replicate the runs by Benham et Al. [7] since they do not have access to the Gigaworld dataset[10], which is publicly available upon payment of a fee. The dataset is necessary to perform the external query expansion exploited by the selcted runs from [7].

Eventually, the participating team submitted four official runs and four unofficial runs described in Table 2. The runs and all the code is publicly available online[11].

[10] https://catalog.ldc.upenn.edu/LDC2012T21

[11] https://bitbucket.org/centre_eval/c2019_irc/src/master/

The paper by Grossman and Cormack [18] exploits the principle of automatic routing runs: first, a logistic regression model is trained with the relevance judgments from one or more collections for each topic, then the model is used to predict relevance assessments of documents from a different collection. Both the training and the prediction phases are done on a topic-wise basis.

The routing process represented a challenge for the participating team, which initially submitted a set of four official runs, where some of the topics were missing. For example, the official run `irc_task1_WCrobust0405_001` contains only 33 topics, while the corresponding original run `WCrobust0405` contains all the 50 topics. The participating team could not understand how to derive document rankings for those topics such that no training topics were available for the logistic regression model. For example, when they were attempting to replicate `WCrobust0405`, they exploited as training set the intersection between the topics from TREC 2004 Robust and TREC 2005 Robust. Then, for the prediction phase, only 33 topics from TREC 2017 Common Core were contained in the training set, and no prediction could be performed for the remaining topics. Due to similar issues, the official `irc_task2_WCrobust04_001` and `irc_task2_WCrobust0405_001` contain 25 and 15 topics respectively.

Afterwards, the participating team contacted the authors of the original paper, Grossman and Cormack [18], to understand how to derive rankings even when there are no training topics available. The authors clarified that for `WCrobust0405` the training set contains both the topics from TREC 2004 Robust and TREC 2005 Robust, and when a topic is not contained in TREC 2005 Robust, they used just the TREC 2004 Robust collection as training set. Therefore, the authors submitted four additional unofficial runs, where both `irc_task1_WCrobust04_001` and `irc_task1_WCrobust0405_001` contain all the 50 topics, while the reproduced runs `irc_task2_WCrobust04_001` and `irc_task2_WCrobust04_001` contain 25 topics. Note that some of the topics are missing for the reproduced runs, since no training data is available for 25 out of the 50 topics of TREC 2018 Common Core.

In the following we report the evaluation results for the replicability and reproducibility tasks, both for the official and unofficial submissions.

Table 3 and Table 4 report AP, nDCG and ERR scores for the official replicated runs. As shown by RMSE, the replication task was fairly successful with respect to AP and nDCG, while when ERR is considered, RMSE is greater than 0.2, showing that it is harder to replicate ERR than the other evaluation measures. Indeed, it is well known that ERR is highly sensitive to the position of relevant documents at the very beginning of the ranking, thus even the misplacement of a single relevant documents may cause a significant drop in ERR score.

Furthermore, as the cut-off increases, even RMSE for AP and nDCG increases, showing that the replication is less accurate at lower cut-off levels. On the other side, RMSE for ERR is almost constant when the cut-off increases, showing once more that ERR focuses on the top rank positions rather than considering the whole ranking.

Table 3. Evaluation of the replicability task for the official `WCrobust04` (50 topics): measures scores averaged across the topics and RMSE.

	Original Run `WCrobust04`	Replicated Run `irc_task1_WCrobust04_001`	RMSE
AP@10	0.0506	0.0564	0.0224
AP@100	0.2252	0.1862	0.0868
AP@1000	0.3821	0.2963	0.1371
nDCG@10	0.1442	0.1503	0.0567
nDCG@100	0.3883	0.3421	0.1110
nDCG@1000	0.6299	0.5418	0.1374
ERR@10	0.5340	0.5663	0.2463
ERR@100	0.5341	0.5693	0.2437
ERR@1000	0.5663	0.5695	0.2436

Table 4. Evaluation of the replicability task for the official `WCrobust0405` (33 topics): measures scores averaged across the topics and RMSE.

	Original Run `WCrobust0405`	Replicated Run `irc_task1_WCrobust0405_001`	RMSE
AP@10	0.0473	0.0491	0.0233
AP@100	0.2541	0.2214	0.0649
AP@1000	0.4428	0.3751	0.1042
nDCG@10	0.1490	0.1511	0.0489
nDCG@100	0.4268	0.3944	0.0814
nDCG@1000	0.6883	0.6237	0.1025
ERR@10	0.6601	0.6756	0.2097
ERR@100	0.6630	0.6777	0.2074
ERR@1000	0.6630	0.6777	0.2074

Table 5. Evaluation of the replicability task for the unofficial WCrobust0405 (50 topics): measures scores averaged across the topics and RMSE.

	Original Run WCrobust0405	Replicated Run irc_task1_WCrobust0405_001	RMSE
AP@10	0.0584	0.0604	0.0209
AP@100	0.2699	0.2244	0.0798
AP@1000	0.4378	0.3534	0.1227
nDCG@10	0.1675	0.1698	0.0484
nDCG@100	0.4480	0.3994	0.1024
nDCG@1000	0.6878	0.6064	0.1279
ERR@10	0.6330	0.6572	0.2106
ERR@100	0.6359	0.6593	0.2095
ERR@1000	0.6360	0.6593	0.2095

Similarly, Table 5 reports AP, nDCG and ERR scores for the unofficial replicated run irc_task1_WCrobust0405_001. Note that the official and unofficial replicated run irc_task1_WCrobust04_001 are identical, therefore the evaluation scores for this unoffical run are the same reported in Table 3 and are omitted in the following.

Again, we can observe that the replication task is more successful for RMSE with respect to AP and nDCG than ERR. Furthermore, RMSE increases as the cut-off increases, meaning that the accuracy of the replicated run decreases as the cut-off level increases.

By comparing the official and unofficial evaluation results for irc_task1_WCrobust0405_001, in Table 4 and Table 5 respectively, we can note that RMSE score are quite similar, showing that the unofficial run is fairly accurate even on the additional topics.

Table 6 reports ER for the replication task with the official runs. We considered WCrobust0405 as advanced run and WCrobust04 as baseline run, therefore the per-topic improvement is computed as WCrobust0405 scores minus WCrobust04 scores for each topic. For the replicated official runs, we needed to select from irc_task1_WCrobust04_001 the 33 topics contained in irc_task1_WCrobust0405_001, otherwise we could not compute the mean per-topic improvement.

ER shows that the replication task is fairly successful for AP, while it is less successful for nDCG and ERR. Furthermore, ER > 1 highlights that the difference between the advanced and the baseline run is more pronounced in the replicated runs than in the original runs. Again, it can be noted that as the cut-off increases, the accuracy of the replicability exercise decreases for AP and nDCG, while it is almost constant for ERR.

Table 6. Evaluation of the replicability task with mean per-topic improvement and *Effect Ratio (ER)* for the official runs (50 topics for original runs and 33 topics for replicated runs).

	ΔM^C_{orig}	$\Delta M^C_{replica}$	ER
AP@10	0.0078	0.0065	0.8333
AP@100	0.0446	0.0576	1.2915
AP@1000	0.0556	0.0866	1.5576
nDCG@10	0.0233	0.0309	1.3262
nDCG@100	0.0597	0.0839	1.4054
nDCG@1000	0.0578	0.0975	1.6869
ERR@10	0.1042	0.1270	1.2188
ERR@100	0.1019	0.1255	1.2316
ERR@1000	0.1019	0.1254	1.2362

Table 7. Evaluation of the replicability task with mean per-topic improvement and *Effect Ratio (ER)* for the unofficial runs (50 topics for original and replicated runs).

	ΔM^C_{orig}	$\Delta M^C_{replica}$	ER
AP@10	0.0078	0.0040	0.5128
AP@100	0.0446	0.0382	0.8565
AP@1000	0.0556	0.0571	1.0270
nDCG@10	0.0233	0.0195	0.8369
nDCG@100	0.0597	0.0573	0.9598
nDCG@1000	0.0578	0.0647	1.1194
ERR@10	0.1042	0.0908	0.8714
ERR@100	0.1019	0.0901	0.8842
ERR@1000	0.1019	0.0899	0.8822

Table 8. Kendall's τ between the original and replicated runs.

Replicated Run	Original Run	τ@10	τ@100	τ@1000
irc_task1_WCrobust04_001 official	WCrobust04	−0.0222	0.0073	0.0021
irc_task1_WCrobust0405_001 official	WCrobust0405	−0.0034	0.0316	0.0046
irc_task1_WCrobust0405_001 unofficial	WCrobust0405	−0.0107	0.0199	0.0029

```
307 Q0  309412  1   -1  WCrobust04
307 Q0  582044  2   -2  WCrobust04
307 Q0  672305  3   -3  WCrobust04
307 Q0  1438673 4   -4  WCrobust04
307 Q0  733642  5   -5  WCrobust04
307 Q0  377253  6   -6  WCrobust04
307 Q0  284810  7   -7  WCrobust04
307 Q0  1248807 8   -8  WCrobust04
307 Q0  1241952 9   -9  WCrobust04
307 Q0  587044  10  -10 WCrobust04
```

Fig. 2. First 10 rank positions for WCrobust04 for topic 307 form TREC 2017 Common Core Track.

```
307 Q0  733642  1   0.8666604108653863  IRC
307 Q0  309412  2   0.8367380559490579  IRC
307 Q0  241240  3   0.8262481009460336  IRC
307 Q0  1248807 4   0.7941459503093672  IRC
307 Q0  125806  5   0.7714647940416018  IRC
307 Q0  617046  6   0.7704302957730799  IRC
307 Q0  672305  7   0.7682124713790986  IRC
307 Q0  1677923 8   0.7614708545584544  IRC
307 Q0  566174  9   0.7530631567740554  IRC
307 Q0  1620713 10  0.7422523019156065  IRC
```

Fig. 3. First 10 rank positions for irc_task1_WCrobust04_001 for topic 307 form TREC 2017 Common Core Track.

Analogously, Table 7 reports ER for the replication task with the unofficial runs. We considered WCrobust0405 as advanced run and WCrobust04 as baseline run, therefore the per-topic improvement is computed as in Table 6 and the first column is equal. Both the replicated unofficial runs contain the same 50 topics, therefore the per-topic improvement is computed as irc_task1_WCrobust0405_001 scores minus irc_task1_WCrobust04_001 scores for each topic.

When the whole set of 50 topics is considered, the replication is fairly successful with respect to all the measure, with ER ranging between 0.83 and 1.12. The only exception is represented by AP@10, where the replicated runs fails to replicate the per-topic improvements. Again, the accuracy of the replicated runs decreases as the cut-off increases.

Table 8 reports the Kendall's τ correlation between the original and replicated runs, both for the official and unofficial runs. We computed Kendall's τ at different cut-off levels, where we first trimmed the runs at the specified cut-off and subsequently computed Kendall's τ between the trimmed runs.

Table 8 shows that the replication was not successful for any of the runs in terms of Kendall's τ. This means that even if the considered replicated runs were similar to the original runs in terms of placement of relevant and non relevant documents, they actually retrieves different documents.

Figure 2 and Figure 3 shows the first 10 rank positions for WCrobust04 and its replicated version irc_task1_WCrobust04_001, for topic 307 from TREC 2017 Common Core Track. We can observe that even if the runs retrieves a similar set of documents, the relative position of each document is different. For example, document 309412 is at rank position 1 for the original run, but at

Table 9. Evaluation of the reproducibility task with mean per-topic improvement and *Effect Ratio (ER)* for the official runs (50 topics for original runs and 15 topics for the reproduced runs).

	ΔM^{C}_{orig}	$\Delta M^{D}_{reproduced}$	ER
AP@10	0.0078	0.0122	1.5641
AP@100	0.0446	0.0431	0.9664
AP@1000	0.0556	0.0579	1.0414
nDCG@10	0.0233	0.0298	1.2790
nDCG@100	0.0597	0.0767	1.2848
nDCG@1000	0.0578	0.0898	1.5536
ERR@10	0.1042	0.0124	0.1190
ERR@100	0.1019	0.0142	0.1394
ERR@1000	0.1019	0.0135	0.1325

rank position 2 for the replicated run, similary document `733642` is at rank position 1 for the replicated run and at rank position 5 for the original run. Moreover, document `241240` is at rank position 3 for the replicated run, but it does not apper on the first 10 positions for the original run.

Table 8, Figure 2 and Figure 3 highlights how hard is to replicate the exact ranking of documents. Therefore, whenever a replicability task is considered, comparing the evaluation scores with RMSE or ER might not be enough, since these approaches consider just the position of relevant and not relevant documents, and overlook the actual ranking of documents.

Finally, Table 9 reports the mean per-topic improvement and ER for the official runs from the reproducibility task. As done for the replicability task, we considered `WCrobust0405` as advanced run and `WCrobust04` as baseline run on the test collection from TREC 2017 Common Core Track. For the reproduced official runs, we needed to select from `irc_task2_WCrobust04_001` the 15 topics contained in `irc_task2_WCrobust0405_001` from TREC 2018 Common Core Track, otherwise we could not compute the per topic improvement.

Table 9 shows that the reproducibility task is fairly successful with respect to AP, with cut-off 100 and 1000. For nDCG, the improvement of the advanced run over the baseline run is more pronounced for the reproduced runs than for the original runs. Conversely, the improvement of the advanced run over the baseline run is more pronounced for the original runs than for the replicated runs for ERR. Again, ERR is the hardest measure in terms of reproducibility success, with the lowest ER.

Furthermore, when the cut-off increases, the accuracy of the reproducibility exercise increases for AP, while it decreases for nDCG and remains almost constant for ERR.

Table 10. Evaluation of the reproducibility task with mean per-topic improvement and *Effect Ratio (ER)* for the unofficial runs (50 topics for original runs and 25 topics for the reproduced runs).

	ΔM^{C}_{orig}	$\Delta M^{D}_{reproduced}$	ER
AP@10	0.0078	0.0065	0.8333
AP@100	0.0446	0.0241	0.5404
AP@1000	0.0556	0.0336	0.6043
nDCG@10	0.0233	0.0155	0.6652
nDCG@100	0.0597	0.0426	0.7136
nDCG@1000	0.0578	0.0509	0.8806
ERR@10	0.1042	0.0004	0.0038
ERR@100	0.1019	0.0033	0.0324
ERR@1000	0.1019	0.0029	0.0285

Similarly, Table 10 reports the mean per-topic improvement and ER for the unofficial runs from the reproducibility task. We considered `WCrobust0405` as advanced run and `WCrobust04` as baseline run on TREC 2017 Common Core, therefore the per-topic improvement is computed as in Table 9 and the first column is equal. Both the reproduced unofficial runs contain the same 25 topics, therefore the per-topic improvement is computed as `irc_task2_WCrobust-0405_001` scores minus `irc_task2_WCrobust04_001` scores for each topic.

The best reproducibility results are obtained with respect to AP@10 and nDCG@1000, thus the effect of the advanced run over the baseline run is better reproduced at the beginning of the ranking for AP, and when the whole ranked list is considered, for nDCG. Again, ERR is the hardest measure to be reproduced, indeed it has the lowest ER score for each cut-off level.

4 Conclusions and Future Work

This paper reports the results on the second edition of CENTRE@CLEF2019. A total of 19 participants enrolled in the lab, however just one group managed to submit two replicability runs and two reproducibility runs. As reported in Section 3, the participating team could not reproduce the runs from Benham et Al. [7], due to the lack of the Gigaworld dataset, but they managed to replicate and reproduce the runs from Grossman and Cormack [18]. More details regarding the implementation are described in their paper [8].

The experimental results show that the replicated runs are fairly successful with respect to AP and nDCG, while the lowest replicability results are obtained with respect to ERR. As ERR mainly focuses on the beginning of the ranking, misplacing even a single relevant document can deteriorate ERR score and have a great impact on the replicability evaluation scores.

Moreover, whenever replicability is considered, RMSE and ER are not enough to evaluate the replicated runs. Indeed, they only account for the position of relevant and not relevant documents by considering the similarity between the original scores and the replicated scores, and they overlook the actual ranking of documents. When the runs are evaluated with Kendall's τ to account for the actual position of the documents in the ranking, the experiments show that the replicability is not successful at all, with Kendall's τ values close to 0. This confirms that, even if it is possible to achieve similar scores in terms of IR evaluation measures, it is challenging to replicate the same documents ranking.

When it comes to reproducibility, there are no well-established evaluation measures to determine to which extent a system can be reproduced. Therefore, we compute ER, firstly exploited in [25], which focuses on the reproduction of the improvement of an advanced run over a baseline run. The experiments show that reproducibility was fairly successful in terms of AP@10 and nDCG@1000, while, similarly to the replicability task, ERR is the hardest measure in terms of reproducibility success.

Finally, as reported in [14, 15], the lack of participation is a signal that the IR community is somehow overlooking replicability and reproducibility issues. As it also emerged from a recent survey within the SIGIR community [13], while there is a very positive attitude towards reproducibility and it is considered very important from a scientific point of view, there are many obstacles to it such as the effort required to put it into practice, the lack of rewards for achieving it, the possible barriers for new and inexperienced groups, and, last but not least, the (somehow optimistic) researcher's perception that their own research is already reproducible.

For the next edition of the lab we are planning to propose some changes in the lab organization to increase the interest and participation of the research community. First, we will target for more popular systems to be replicated and reproduced, moreover we will consider other tasks than the AdHoc, as for example the medical or other popular domains.

References

1. Allan, J., Arguello, J., Azzopardi, L., Bailey, P., Baldwin, T., Balog, K., Bast, H., Belkin, N., Berberich, K., von Billerbeck, B., Callan, J., Capra, R., Carman, M., Carterette, B., Clarke, C.L.A., Collins-Thompson, K., Craswell, N., Croft, W.B., Culpepper, J.S., Dalton, J., Demartini, G., Diaz, F., Dietz, L., Dumais, S., Eickhoff, C., Ferro, N., Fuhr, N., Geva, S., Hauff, C., Hawking, D., Joho, H., Jones, G.J.F., Kamps, J., Kando, N., Kelly, D., Kim, J., Kiseleva, J., Liu, Y., Lu, X., Mizzaro, S., Moffat, A., Nie, J.Y., Olteanu, A., Ounis, I., Radlinski, F., de Rijke, M., Sanderson, M., Scholer, F., Sitbon, L., Smucker, M.D., Soboroff, I., Spina, D., Suel, T., Thom, J., Thomas, P., Trotman, A., Voorhees, E.M., de Vries, A.P., Yilmaz, E., Zuccon, G.: Research Frontiers in Information Retrieval – Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). SIGIR Forum 52(1) (June 2018)
2. Allan, J., Harman, D.K., Kanoulas, E., Li, D., Van Gysel, C., Voorhees, E.M.: TREC 2017 Common Core Track Overview. In: Voorhees and Ellis [26]

3. Allan, J., Harman, D.K., Kanoulas, E., Li, D., Van Gysel, C., Voorhees, E.M.: TREC 2018 Common Core Track Overview. In: Voorhees and Ellis [27]
4. Arguello, J., Crane, M., Diaz, F., Lin, J., Trotman, A.: Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). SIGIR Forum 49(2), 107–116 (December 2015)
5. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Has Adhoc Retrieval Improved Since 1994? In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009). pp. 692–693. ACM Press, New York, USA (2009)
6. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In: Cheung, D.W.L., Song, I.Y., Chu, W.W., Hu, X., Lin, J.J. (eds.) Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009). pp. 601–610. ACM Press, New York, USA (2009)
7. Benham, R., Gallagher, L., Mackenzie, J., Liu, B., Lu, X., Scholer, F., Moffat, A., Culpepper, J.S.: RMIT at the 2018 TREC CORE Track. In: Voorhees and Ellis [27]
8. Breuer, T., Schaer, P.: Replicability and Reproducibility of Automatic Routing Runs . In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2019)
9. Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): CLEF 2018 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073 (2018)
10. Clancy, R., Ferro, N., Hauff, C., Lin, J., Sakai, T., Wu, Z.Z.: The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). In: Chevalier, M., Gaussier, É., Piwowarski, B., Maarek, Y., Nie, J.Y., Scholer, F. (eds.) Proc. 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019) (2019)
11. Ferro, N., Fuhr, N., Grefenstette, G., Konstan, J.A., Castells, P., Daly, E.M., Declerck, T., Ekstrand, M.D., Geyer, W., Gonzalo, J., Kuflik, T., Lindén, K., Magnini, B., Nie, J.Y., Perego, R., Shapira, B., Soboroff, I., Tintarev, N., Verspoor, K., Willemsen, M.C., Zobel, J.: The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction. SIGIR Forum 52(1) (June 2018)
12. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". SIGIR Forum 50(1), 68–82 (June 2016)
13. Ferro, N., Kelly, D.: SIGIR Initiative to Implement ACM Artifact Review and Badging. SIGIR Forum 52(1) (June 2018)
14. Ferro, N., Maistro, M., Sakai, T., Soboroff, I.: CENTRE@CLEF2018: Overview of the Replicability Task. In: Cappellato et al. [9]
15. Ferro, N., Maistro, M., Sakai, T., Soboroff, I.: Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018). pp. 239–246. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (2018)
16. Freire, J., Fuhr, N., Rauber, A. (eds.): Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. Dagstuhl Reports, Volume 6, Number 1, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany (2016)

17. Fuhr, N.: Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. SIGIR Forum 51(3), 32–41 (December 2017)
18. Grossman, M.R., Cormack, G.V.: MRG_UWaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track. In: Voorhees and Ellis [26]
19. Jungwirth, M., Hanbury, A.: Replicating an Experiment in Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Cappellato et al. [9]
20. Kendall, M.G.: Rank correlation methods. Griffin, Oxford, England (1948)
21. Kenney, J.F., Keeping, E.S.: Mathematics of Statistics – Part One. D. Van Nostrand Company, Princeton, USA, 3rd edn. (1954)
22. Kharazmi, S., Scholer, F., Vallet, D., Sanderson, M.: Examining Additivity and Weak Baselines. ACM Transactions on Information Systems (TOIS) 34(4), 23:1–23:18 (June 2016)
23. Lin, J., Crane, M., Trotman, A., Callan, J., Chattopadhyaya, I., Foley, J., Ingersoll, G., Macdonald, C., Vigna, S.: Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016). pp. 357–368. Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany (2016)
24. Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P.A.: A manifesto for reproducible science. Nature Human Behaviour 1, 0021:1–0021:9 (January 2017)
25. Sakai, T., Ferro, N., Soboroff, I., Zeng, Z., Xiao, P., Maistro, M.: Overview of the NTCIR-14 CENTRE Task. In: Ishita, E., Kando, N., Kato, M.P., Liu, Y. (eds.) Proc. 14th NTCIR Conference on Evaluation of Information Access Technologies. pp. 494–509. National Institute of Informatics, Tokyo, Japan (2019)
26. Voorhees, E.M., Ellis, A. (eds.): The Twenty-Sixth Text REtrieval Conference Proceedings (TREC 2017). National Institute of Standards and Technology (NIST), Special Publication 500-324, Washington, USA (2018)
27. Voorhees, E.M., Ellis, A. (eds.): The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018). National Institute of Standards and Technology (NIST), Washington, USA (2019)
28. Zobel, J., Webber, W., Sanderson, M., Moffat, A.: Principles for Robust Evaluation Infrastructure. In: Agosti, M., Ferro, N., Thanos, C. (eds.) Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011). pp. 3–6. ACM Press, New York, USA (2011)