CENTRE@CLEF 2019*

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

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

Abstract. Reproducibility of experimental results has recently become a primary issue in the scientific community at large, and in the information retrieval community as well, where initiatives and incentives to promote and ease reproducibility are arising. In this context, CENTRE is a joint CLEF/TREC/NTCIR lab which aims at raising the attention on this topic and involving the community in a shared reproducibility exercise. In particular, CENTRE focuses on three objectives, e.g. replicability, reproducibility and generalizability, and for each of them a dedicated task is designed. We expect that CENTRE may impact on the validation of some key achievement in IR, help in designing shared protocols for reproducibility, and improve the understanding on generalization across collections and on the additivity issue.

1 Introduction

Reproducibility is becoming a primary concern in many areas of science [12, 17] 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 item in the research community [2, 5, 7, 8, 13, 16, 18]. 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 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 the best TREC systems still outperforms off-the-shelf open source systems [2–4, 15, 16]. This is due to many

* http://www.centre-eval.org/clef2019/

different factors, among which are the lack of tuning on a specific collection when using default configuration, and the lack of specifications about 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 [4, 15]. 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, as also emerged from a recent survey within the SIGIR community [9] 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.

Finally, the other side of reproducibility is the generalizability of the experimental results which plays an important role for future research. Indeed, both a Dagstuhl Perspectives Workshop [6] 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.

This paper is organized as follows: Section 2 presents the objectives and scope of CENTRE, Section 3 describes the tasks proposed at CENTRE@CLEF 2019 and provides details about the measures used to evaluate the submitted runs, finally Section 4 reports some observations and lessons learnt from CENTRE@CLEF 2018, which were useful to design the 2019 edition.

2 Aims and Scope

Overall, the above considerations stress the need and urgency for a systematic approach to reproducibility and generalizability in IR. Therefore, the goal of ***CLE**F **N**TCIR **T**REC **RE**producibility (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 performances.

We targeted evaluation campaigns to run CENTRE since we need third-party-ness with respect to the original developers of a technique, thus the author of the method should not attempt in reproducing it. Moreover, the critical mass

involved in an evaluation campaign is needed for sharing the effort, achieving enough coverage and getting multiple independent checks for the same techniques. Indeed, if a system is reproduced by more than one single group, they can possibly discover more issues concerning a given technique and they can get as close as possible to actually reproducing it. Finally, we need to develop a common and shared protocol for reproducibility, to this end the experimental results and the developed components should be publicly accessible and an evaluation campaign represents one of the best venues to achieve this purpose.

We designed CENTRE as a joint CLEF/NTCIR/TREC task to further promote the possibility for third-party-ness, asking members of a community to reproduce what has been developed in another community. Moreover, we can simultaneously cover almost all the geographical areas, synchronously progressing the IR community at large towards reproducibility and the participants have the possibility to report their results in a globally shared task, at the closest and more convenient venue among CLEF/NTCIR/TREC. Finally, this is also an experiment to understand how a closer cooperation among CLEF/NTCIR/TREC might work.

3 CENTRE@CLEF2019 Tasks

In this edition of the lab, we target three specific objectives, according to the ACM badging terminology, which may need to be slightly adapted to the IR context:

Replicability (different team, same experimental setup): we use the collections, topics and ground-truth on which the methods and solutions have been developed and evaluated.

Reproducibility (different team, different experimental setup): we use a different experimental collection, but in the same domain, from those used to originally develop and evaluate a solution;

Generalizability (different team, different experimental setup): use sub-collections or different collections, but in the same domain.

For each of the aforementioned objectives, we designed a different task. Therefore, CENTRE@CLEF 2019 offers the following three 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 the different experimental collections;
- 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.

3.1 Replicability and Reproducibility

Tasks 1 and 2 are the same tasks as in the CENTRE@CLEF2018 edition[5], targeting selected runs from CLEF/NTCIR/TREC on the same collections for replicability and on different collections for reproducibility. According to the discussion and feedback from attendees at CLEF 2018, we modified and changed the set of the targeted runs with respect to those used during the 2018 edition.

In particular, two valid suggestions were proposed by the participants in CENTRE@CLEF2018. First, we promote a partnership with the ECIR 2020 reproducibility track. To this end we encourage a collaboration among CENTRE participants, who reproduced the same algorithm. The outcome of this collaboration will be a joint paper, summarizing their reproducibility efforts and findings, which can be submitted at the ECIR 2020 reproducibility track. If enough teams will reproduce the same algorithm, the outcome paper will be even strengthened by the different perspectives and strategies adopted in the reproducibility process. We hope that this might represent a reward and a further incentive to participate in CENTRE. Furthermore, from the scheduling point of view, this partnership with ECIR is particularly well timed, since CENTRE deadlines are around May/June, while ECIR is early October. Thus participants will have the possibility to gather during CLEF, in early September, and to jointly finalize their paper.

Second, we select the replication and replicability targets among the best systems submitted at the labs of CLEF 2018. We decided to choose among those labs that will continue with the same task at CLEF 2019. This should motivate prospective participants in developing a baseline, since they would anyway need to do it in order to participate in their preferred lab. Moreover, this should also be useful for lab organizers, since they will be provided with state-of-the-art baselines available for their lab. We have already polled some lab organizers, who gave us their support in this respect.

Therefore, for the replicability and reproducibility activities, we select, among the methods/systems submitted to the CLEF tasks last year, the top performing and most impacting ones. In addition, we select methods/systems from TREC and NTCIR, following the same approach.

Each participating group will be challenged to replicate and/or reproduce one or more of the selected systems by only using standard open source IR systems, like Lucene, Terrier, and others, and they will submit one or more runs, in TREC format, representing the output of their reproduced systems. Participating groups will have to develop and integrate into the open source IR systems all the missing components and resources needed to replicate/reproduce the selected systems and they need to contribute back to open source all the developed components, resources, and configuration via a common repository, e.g. on Bitbucket.

We evaluate the quality of the replicated runs from two points of views: effectiveness and ranking. Effectiveness evaluates how close are the performance

[5] http://www.centre-eval.org/clef2018/

scores of the reproduced systems to those of the original ones. This is measured using the *Root Mean Square Error (RMSE)* between the new and original *Average Precision (AP)* scores as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(AP_{\text{orig},i} - AP_{\text{replica},i}\right)^2} \tag{1}$$

where m is the total number of topics, $AP_{\text{orig},i}$ is the AP score of the original target run on topic t_i and $AP_{\text{replica},i}$ is the AP score of the replicated run on topic t_i.

Since different result lists may produce the same effectiveness score, we also measure how close are the ranked results lists of the replicated systems to those of the original ones. This is measured using the correlation coefficient Kendall's τ between the original and replicated run:

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

where 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), T and U are the number of ties, respectively, in the first and in the second ranking.

Evaluating the quality of the reproduced runs is less straightforward since there is no original run that can be used as a comparison point. Therefore, the idea is to compare the difference with respect to the improvement, in terms of AP, of a baseline run in both collections.

3.2 Generalizability

For the generalizability task, participants needs to rank document collections by the expected performance over them. The task is divided in three phases: training, test, and validation.

During the training phase participants are given topics, ground-truth, and a set of sub-collections (e.g. some newspaper collections from ad-hoc CLEF). They need to work on a selected method (e.g. a specific system as Lucene with BM25, ...) to allow for comparability across participants. Moreover, if they wish, they can also work on their own preferred method. The aim of this phase is to identify features of collections and methods that allow participants to rank and predict collections.

Then, during the test phase, the participants are given different sets of sub-collections (e.g. newspaper from ad-hoc CLEF in a different language) and they have to rank these collections with respect to the mandatory method and their own method.

Finally, the validation phase is conducted after the submission. We provide the topics and the ground-truth on the test sub-collections which are needed to verify how the different methods perform. Note that CLEF topics in different languages are translations one of each other and this should minimize the impact of the topic effect on the prediction. Indeed, generalizing a method through different topics should not be too hard, since topics are related and what differs is just the language used to describe them.

We evaluate the quality of the rankings and predictions of the generalizability task with *Mean Absolute Error (MAE)*, defined as follows:

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^{n} |AP_{\text{orig},j} - AP_{\text{predict},j}| \tag{3}$$

where n is the number of sub-collections and $AP_{\text{predict},i}$ is the score of the predicted ranking. Furthermore, we use RMSE, as in Equation (1), between the predicted and actual performance on the given collections.

4 Lessons Learnt from CENTRE@CLEF2018

CENTRE has been run for the first time at CLEF 2018 and variants of it are running at TREC 2018 and NTCIR-14 (due June 2019). The CENTRE@CLEF2018 edition [10,11] had 17 registered participants, but only 1 actually submitted results, Technical University of Wien (TUW) [14]. TUW failed to replicate the targeted bilingual run, indeed, AP_{orig} was 0.0667, while AP_{replica} was 0.0030, RMSE computed with Equation (1) was 0.1132 and Kendall's τ in Equation (2) was $-5.69 \cdot 10^{-04}$.

This leads to two observations. First, it indicates that engaging participants is a critical issue and that the community needs to be involved more in reproducibility. Second, replicability, reproducibility, and generalizability are still very hard to achieve, showing once more that reproducibility represents a serious limit for the advancement of research.

These issues were presented during the CENTRE session at CLEF 2018. We discussed with attendees measures for attracting more participation at the task and for lowering their barriers of entry. Thus, for CENTRE@CLEF 2019 we select the target systems among the best systems submitted at CLEF 2018 and we start the partnership with ECIR 2020 reproducibility track.

In addition to these incentives, we are contacting the colleagues who have master courses in IR to consider CENTRE tasks as part of the students assignments they already do. We already had positive feedback and availability from some colleagues.

Finally, we also hope that the new task on generalizability can raise the participation in the lab.

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), 34–90 (June 2018)
2. 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)
3. 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)
4. 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)
5. Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Kekäläinen, J., Rosso, P., Clough, P., Pasi, G., Lioma, C., Mizzaro, S., Di Nunzio, G.M., Hauff, C., Alonso, O., Serdyukov, P., Silvello, G.: Report on ECIR 2016: 38th European Conference on Information Retrieval. SIGIR Forum **50**(1), 12–27 (June 2016)
6. 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.: Manifesto from Dagstuhl Perspectives Workshop 17442 – Building a Predictive Science for Performance of Information Retrieval, Recommender Systems, and Natural Language Processing Applications. Dagstuhl Manifestos, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany **7**(1) (2018)
7. Ferro, N., Fuhr, N., Rauber, A.: Introduction to the Special Issue on Reproducibility in Information Retrieval: Evaluation Campaigns, Collections, and Analyses. ACM Journal of Data and Information Quality (JDIQ) **10**(3), 9:1–9:4 (October 2018)
8. Ferro, N., Fuhr, N., Rauber, A.: Introduction to the Special Issue on Reproducibility in Information Retrieval: Tools and Infrastructures. ACM Journal of Data and Information Quality (JDIQ) **10**(4), 1–4 (2018)
9. Ferro, N., Kelly, D.: SIGIR Initiative to Implement ACM Artifact Review and Badging. SIGIR Forum **52**(1), 4–10 (June 2018)
10. Ferro, N., Maistro, M., Sakai, T., Soboroff, I.: CENTRE@CLEF2018: Overview of the Replicability Task. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2125/ (2018)

11. 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) 11018, Springer, Heidelberg, Germany (2018)
12. 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)
13. Fuhr, N.: Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. SIGIR Forum **51**(3), 32–41 (December 2017)
14. Jungwirth, M., Hanbury, A.: Replicating an Experiment in Cross-lingual Information Retrieval with Explicit Semantic Analysis. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2125/ (2018)
15. 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)
16. 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)
17. 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)
18. 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)