# Overview of the NTCIR-14 CENTRE Task

Tetsuya Sakai[1], Nicola Ferro[2], Ian Soboroff[3]
Zhaohao Zeng[1], Peng Xiao[1], and Maria Maistro[2]

[1] Waseda University, Japan tetsuyasakai@acm.org
[2] University of Padua, Italy ferro@dei.unipd.it
[3] NIST, USA ian.soboroff@nist.gov

**Abstract.** CENTRE is the first-ever metatask that operates across the three major information retrieval evaluation venues: CLEF, NTCIR, and TREC. The task had three subtasks: T1 (Replicability), T2TREC (Reproducibility), and T2OPEN (Reproducibility). The T1 subtask examined whether a particular pair of runs from the NTCIR-13 WWW-1 task can be replicated (on the same data). The T2TREC subtask examined whether a particular pair of runs from TREC 2013 Web track can be reproduced on the NTCIR-13 WWW-1 test collection. T2OPEN encouraged participants to reproduce past runs of their own choice on the WWW-1 test collection. Only one team (MPII) participated in CENTRE, but the team participated in all three subtasks. The NTCIR edition of CENTRE focussed on whether the effect of an Advanced run over a Baseline run can be replicated/reproduced. The results of MPII are quite positive for both T1 and T2TREC subtasks in terms of replicating/reproducing the overall effects, as measured by the Effect Ratio.

**Keywords:** CLEF; NTCIR; replicability; reproducibility; TREC

## 1 Introduction

*CENTRE* (*C*LE*F* *N*TCIR *T*REC *RE*produciblity)[4] is the first-ever metatask that operates across the three major information retrieval evaluation venues: CLEF[5], NTCIR[6], and TREC[7]. Its goals are to (1) Examine whether results reported in the IR literature can be replicated or reproduced, and if so, to what extent; and (2) Establish methods for examining replicability and reproducibility. A reported result is *replicable* if a different research group can later obtain a similar result *using the same data*; a reported result is *reproducible* if a different research group can later obtain a similar result *using different data*.

The three CENTRE "editions," namely, CENTRE@CLEF2018 [4], the TREC2018 CENTRE track [12], and the NTCIR-14 CENTRE task, jointly selected the target results to replicate or reproduce, but each had its own task design and goals.

[4] http://www.centre-eval.org/
[5] http://www.clef-initiative.eu/
[6] http://research.nii.ac.jp/ntcir/
[7] https://trec.nist.gov/

Another feature that is common across these three editions is that, unfortunately, each edition had basically only one participating team[8]. For NTCIR-14 CENTRE, Max Planck Institute for Informatics (MPII) submitted runs to all of our subtasks [14] Thank you MPII!

A unique feature of NTCIR-14 CENTRE compared to the other CENTRE editions is that we decided to focus on whether a reported *improvement over a baseline* can be replicated/reproduced rather than whether an absolute performance of a run can be replicated/reproduced. Thus, our evaluation is based on the *difference* between an *A-run* (Advanced run) and a *B-run* (Baseline run).

For both replicability and reproducibility subtasks of NTCIR-14 CENTRE, participating runs were generated on the NTCIR-13 *We Want Web (WWW)* English test collection [7]. Section 2 describes our replicability and reproducibility subtasks, and Section 3 briefly describes the runs submitted by MPII. Section 4 describes how we expanded the NTCIR-13 WWW-1 English qrels based on the pooled results from MPII. Section 5 discusses our results based on the original "WWW-1" qrels, and Section 6 discusses our results based on the new "CENTRE-1" qrels. Finally, Section 7 provides our conclusions.

## 2 Subtasks

### 2.1 T1: Replicability

The T1 subtask is defined as follows. Given a pair of runs (A-run and B-run, where A-run has been reported to outperform B-run) on a test collection $C$, can another research group replicate the improvement on $C$?

Runs submitted to the T1 subtask are evaluated as follows. For an evaluation measure $M$, let $M_j^C(A)$ and $M_j^C(B)$ denote the score of the *original* A-run and that of the *original* B-run for the $j$-th topic of collection $C$ ($1 \leq j \leq n_C$). Similarly, let $M_j^C(A')$ and $M_j^C(B')$ denote the scores for the *replicated* A-run and B-run, respectively. Then, per-topic improvements in the original and replicated experiments are given by

$$\Delta M_j^C = M_j^C(A) - M_j^C(B) \ , \quad \Delta' M_j^C = M_j^C(A') - M_j^C(B') \ . \tag{1}$$

Note that even if the A-run outperforms the B-run on average, the opposite may be true for some topics: that is, per-topic "improvements" may be negative.

To evaluate *how the per-topic improvements are faithfully replicated*, we use two evaluation measures: *Root Mean Squared Error* (RMSE) and *Pearson's correlation coefficient* $r$. RSME for measure $M$ is defined as[9]

$$RSME_M = \sqrt{\frac{1}{n_C}\sum_{j=1}^{n_C}(\Delta' M_j^C - \Delta M_j^C)^2} \ . \tag{2}$$

[8] TREC2018 CENTRE actually had one more team who submitted their results after the official deadline.

[9] We shall omit $C$ from our notations whenever it is clear from the context that $C$ is used throughout the experiment.

On the other hand, $r$ for measure $M$ is defined as

$$r_M = \frac{cov(\Delta' M, \Delta M)}{sd(\Delta' M) sd(\Delta M)} \ , \tag{3}$$

where

$$cov(\Delta' M, \Delta M) = \frac{1}{n_C} \sum_{j=1}^{n_C} (\Delta' M_j^C - \overline{\Delta' M^C})(\Delta M_j^C - \overline{\Delta M^C}) \ , \tag{4}$$

$$sd(\Delta' M) = \sqrt{\frac{1}{n_C} \sum_{j=1}^{n_C} (\Delta' M_j^C - \overline{\Delta' M^C})^2} \ , \tag{5}$$

$$sd(\Delta M) = \sqrt{\frac{1}{n_C} \sum_{j=1}^{n_C} (\Delta M_j^C - \overline{\Delta M^C})^2} \ , \tag{6}$$

with $\overline{\Delta' M^C} = (\sum_{j=1}^{n_C} \Delta' M_j^C)/n_C$ and $\overline{\Delta M^C} = (\sum_{j=1}^{n_C} \Delta M_j^C)/n_C$.

The above two evaluation measures are for discussing topicwise faithfulness of the replicated run pairs. However, claims from comparative IR experiments are usually based on comparing *mean* effectiveness scores. Hence, we also evaluate submitted runs in terms of *Effect Ratio* (ER), which we define as:

$$ER(\Delta' M^C, \Delta M^C) = \frac{\overline{\Delta' M^C}}{\overline{\Delta M^C}} = \frac{\sum_{j=1}^{n_C} \Delta' M_j^C}{\sum_{j=1}^{n_C} \Delta M_j^C} \tag{7}$$

Note that the denominator of ER is the mean improvement in the original experiment, while the numerator is the mean improvement in the replicated experiment. Assuming that the standard deviation for measure $M$ is common across experiments, ER is equivalent to the ratio of *effect sizes* (standardised mean differences, to be precise) [10]: hence the name.

If $ER \leq 0$, 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.

For the NTCIR-14 round, we chose the NTCIR-13 WWW-1 English test collection [7] as $C$ for the following reasons:

- The NTCIR-13 WWW-1 Task was a standard ad hoc IR task which should be easy for participants to tackle;
- The English document collection for WWW-1 was clueweb12-B13, and since clueweb12 is used by several tasks outside NTCIR, this is convenient for us to design not only replicability but also reproducibility subtasks, by considering two different tasks that are reasonably similar. More specifically, T1 can

choose a pair of English runs from WWW-1 and study replicability on the WWW-1 test collection, while T2 can choose a pair of runs from a different task designed on clueweb12 and then study reproducibility on the WWW-1 test collection.

The WWW-1 test collection comprises 100 topics, with graded relevance assessments (relevance levels: $L0$-$L4$). The relevance assessments were constructed from depth-30 pools based on 13 runs from three teams. The measurement depth used at the NTCIR-13 WWW-1 task was 10, as the WWW-1 organisers were interested in the quality of the first search engine result page.

As for the target A-run and B-run from WWW-1, we chose a pair of runs from RMIT [5]. To be more specific, we chose `RMIT-E-NU-Own-1` (*sequential* dependency model: SDM) as the original A-run and `RMIT-E-NU-Own-3` (*full* dependency model: FDM) as the original B-run. Hence, the question is "*Can a replicated experiment confirm that SDM outperforms FDM?*" We chose these runs for the following reasons:

- `RMIT-E-NU-Own-1` was the official top performer at the NTCIR-13 WWW-1 English subtask;
- The techniques used, SDM and FDM, are from a well-cited Metzler-Croft paper from ACM SIGIR 2005 [8], and seem to have stood the test of time;
- RMIT used the publicly-available Indri search engine[10], which should make the replicability challenge relatively easy for participants.

The RMIT team kindly provided their code from their WWW-1 submissions on github[11]. We gave this URL to our only participating team MPII immediately *after* the run submission deadline, so that they had to work without relying on RMIT's original code. All they had before submission were RMIT's WWW-1 participant paper and the SIGIR 2005 paper.

### 2.2 T2TREC: Reproducibility

The T2 subtasks are defined as follows. Given a pair of runs (A-run and B-run, where A-run has been reported to outperform B-run) on a test collection $D$, can another research group replicate the improvement on a different test collection $C$?

Following our previous notations, the *original* per-topic improvements are now denoted by

$$\Delta M_j^D = M_j^D(A) - M_j^D(B) \ , \ \ \Delta' M_j^D = M_j^D(A') - M_j^D(B') \ , \qquad (8)$$

and the corresponding mean differences are $\overline{\Delta' M^D} = (\sum_{j=1}^{n_D} \Delta' M_j^D)/n_D$ and $\overline{\Delta M^D} = (\sum_{j=1}^{n_D} \Delta M_j^D)/n_D$, where $n_D$ is the number of topics in $D$. Then,

[10] `http://www.lemurproject.org/indri/`
[11] `https://github.com/rmit-ir/ntcir13-www`

reproduced runs can also be evaluated using ER as follows:

$$ER(\Delta' M^C, \Delta M^D) = \frac{\overline{\Delta' M^C}}{\overline{\Delta M^D}} = \frac{\frac{1}{n_C}\sum_{j=1}^{n_C} \Delta' M_j^C}{\frac{1}{n_D}\sum_{j=1}^{n_D} \Delta M_j^D} \quad (9)$$

The above ER can then be interpreted in a way similar to the ER for T1.

For the NTCIR-14 round, the "new" test collection $C$ is the NTCIR-13 WWW-1 test collection, which uses clueweb12-B13 as the target corpus. As for $D$, the natural choice was a test collection from the TREC web track ad hoc task series that also used clueweb12. We settled on a pair of runs from the University of Delaware [13] from the TREC 2013 Web Track [1]: we chose `UDInfolabWEB2` (selects semantically related terms using *web-based* working sets) as the original A-run, and `UDInfolabWEB1` (selects semantically related terms using *collection-based* workings sets). Hence, the question is "*Do web-based workings sets outperform collection-based ones even on a different test collection?*" It should be noted that while WWW-1 uses clueweb12-B13, which is a subset of the entire clueweb12 corpus, the Delaware runs are Category A runs: they use the entire corpus. Moreover, while the WWW-1 test collection has 100 topics, the TREC 2013 web track test collection has only 50 topics. We chose these runs for the following reasons:

- `UDInfolabWEB2` was one of the official top performers at the TREC 2013 Web Track ad hoc task;
- The techniques originate from a Fang-Zhai paper from ACM SIGIR 2006 [2] and seem to have stood the test of time;
- Delaware also used Indri, which should make the reproducibility relatively easy for participants.

### 2.3 T2OPEN: Reproducibility

In addition to the above two subtasks where the organisers specified what to replicate/reproduce, we also encouraged participants to try their own reproducibility experiments on the WWW-1 test collection: they could choose their own target pair of runs. For this "open" subtask, we leave the evaluation of reproducibility to the participants, and merely provide the effectiveness scores for each submitted run.

## 3 Submitted Runs

Table 1 shows the run names and their descriptions (embedded in their run files) contributed by MPII. For the T2OPEN task, MPII chose to reproduce a pair of runs from a CIKM 2016 paper [6]: for more details about MPII's runs, we refer the reader to their NTCIR-14 participant paper [14].

**Table 1.** Runs submitted by MPII and their descriptions.

| | |
|---|---|
| T1 runs | |
| CENTRE-1-MPII-T1-A | replication of RMIT-E-NU-Own-1 with Indri 5.12, but without inlinks indexed |
| CENTRE-1-MPII-T1-B | replication of RMIT-E-NU-Own-3 with Indri 5.12, without inlinks indexed |
| T2TREC runs | |
| CENTRE-1-MPII-T2TREC-A | Anserini run to reproduce UDInfolabWEB2 but with beta=0.1 |
| CENTRE-1-MPII-T2TREC-B | Anserini run to reproduce UDInfolabWEB1 with beta=0.1 |
| T2OPEN runs | |
| CENTRE-1-MPII-T2OPEN-A | DRMM with LCH normalization, 15 bins, reranking Indri BM25 |
| CENTRE-1-MPII-T2OPEN-B | DRMM with CH normalization, 20 bins, reranking Indri BM25 |

## 4 Additional Relevance Assessments

The original depth-30 pools from the NTCIR-13 WWW-1 task had 22,912 topic-document pairs, all of which were judged to form the official qrels, which we call "WWW-1 qrels." In the relevance assessment phase, pooled documents were ranked by "popularity" [11][12] for each topic and were presented to the assessors: the `NTCIRPOOL` toolkit was used for this purpose[13]. Each topic was judged independently by two assessors; they chose from highly relevant (2 points), relevant (1 point), and others (0 points), and the points were added to form five-point scale ($L0$-$L4$) relevance assessments [7].

From the new CENTRE runs (i.e., the MPII runs), we also created depth-30 pools, which gave us 10,397 topic-document pairs. From this set, we removed all topic-document pairs that were already judged at WWW-1. As a result, we were left with only 2,617 topic-document pairs. That is, these pairs are outside the WWW-1 qrels but within top 30 in at least one of the new CENTRE runs. The unjudged documents covered 99 of the WWW-1 topics: for Topic 0014, no new documents were retrieved above top 30.

Since the above 2,617 unjudged documents may cause underestimation of the MPII runs relative to the original WWW-1 runs, we conducted additional relevance assessments in a way similar to WWW-1: two assessors independently judged all of the above unjudged documents using the assessment interface from WWW-1 [7]. The only difference is that these documents were not sorted by popularity: after removing the topic-document pairs already in the WWW-1 qrels, we simply kept the dictionary-sorted order of the remaining documents,

[12] The first key is the number of runs that returned that document (larger the better), and the second key is the sum of ranks within those runs (smaller the better).

[13] `http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html`

**Table 2.** Inter-assessor agreement in terms of quadratic-weighted Cohen's $\kappa$ with 95%CIs. The original WWW-1 qrels, additional assessments, and the final CENTRE-1 qrels.

| | #topic-doc pairs | $\kappa$ with 95%CIs |
|---|---|---|
| WWW-1 | 22,912 | 0.43 [0.42, 0.44] |
| Additional | 2,553 | 0.59 [0.57, 0.61] |
| CENTRE-1 | 25,465 | 0.45 [0.44, 0.46] |

**Table 3.** Raw inter-assessor agreement statistics that were used for computing the $\kappa$ statistics.

| | WWW-1 | | | Additional | | | CENTRE-1 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 0 | 1 | 2 | 0 | 1 | 2 |
| 0 | 8,383 | 2,851 | 794 | 1,438 | 157 | 17 | 9,821 | 3,008 | 811 |
| 1 | 1,900 | 2,579 | 1,134 | 276 | 321 | 73 | 2,176 | 2,900 | 1,207 |
| 2 | 956 | 2,732 | 1,583 | 42 | 142 | 87 | 998 | 2,874 | 1,670 |

under the assumption that document ordering does not matter for patching up an existing qrels file. After the additional relevance assessments, we further removed 64 topic-document pairs as the document contents were not displayed on the relevance assessment interface; the remaining 2,553 relevance assessments were then added to the WWW-1 qrels, giving us a total of 25,465 topic-document pairs. We shall refer to the new qrels as "CENTRE-1 qrels."

Table 2 quantifies the inter-assessor agreement for the WWW-1 qrels, the additional assessments, and the final CENTRE-1 qrels in terms of quadratic-weighted Cohen's $\kappa$: recall that the raw relevance labels were 2, 1, or 0. It should be noted that while the $\kappa$ for the additional assessments represents the agreement between our two new assessors, the two assessors for the WWW-1 qrels were different across topics. Table 3 shows the raw inter-assessor agreement statistics that were used for computing the $\kappa$ statistics.

## 5 Results with the Official WWW-1 Qrels

This section evaluates the submitted runs using the official WWW-1 qrels; recall that the MPII runs did not contribute to the orifinal WWW-1 pools.

Following the official evaluation practice of the NTCIR-13 WWW-1 Task [7], we used the `NTCIREVAL` toolkit[14] to compute *nDCG@10*, *Q@10*, and *nERR@10* [9] with a linear gain value setting (i.e., giving a gain value of 4 to each *L*4-relevant document etc.) for all runs involved.

### 5.1 T1: Replicability Results

The top half of Table 4 shows the mean effectiveness scores of the original A-run and B-run from RMIT, together with information about the differences between

[14] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`

**Table 4.** Effectiveness scores based on the WWW-1 qrels ($n = 100$ topics). $P$-values smaller than 5% are indicated in bold.

| | Mean nDCG@10 | Mean Q@10 | Mean nERR@10 |
|---|---|---|---|
| Original A: `RMIT-E-NU-Own-1` | 0.6302 | 0.6548 | 0.7463 |
| Original B: `RMIT-E-NU-Own-3` | 0.5493 | 0.5657 | 0.6977 |
| (Paired $t$-test $p$-value) | **(9.057e-05)** | **(2.937e-05)** | (0.0519) |
| (Glass's $\Delta$) | (0.3358) | (0.3267) | (0.1823) |
| `CENTRE-1-MPII-T1-A` | 0.5933 | 0.5996 | 0.7412 |
| `CENTRE-1-MPII-T1-B` | 0.5428 | 0.5568 | 0.6937 |
| (Paired $t$-test $p$-value) | **(4.352e-04)** | **(0.0128)** | **(0.0126)** |
| (Glass's $\Delta$) | (0.2017) | (0.1498) | (0.1687) |

**Table 5.** T1 results for MPII based on the WWW-1 qrels. $P$-values smaller than 5% are indicated in bold.

| | nDCG@10 | Q@10 | nERR@10 |
|---|---|---|---|
| RMSE | 0.2256 | 0.2431 | 0.2668 |
| $r$ (95%CI, $p$-value) | 0.1469 | 0.1797 | 0.2603 |
| | [−0.0510, 0.3337] | [−0.0174, 0.3633] | [0.0673, 0.4345] |
| | $p = 0.1446$ | $p = 0.0737$ | $\mathbf{p = 0.0089}$ |
| $\overline{\Delta M^C}$ | 0.0809 | 0.0891 | 0.0486 |
| $\overline{\Delta' M^C}$ | 0.0506 | 0.0428 | 0.0475 |
| $ER(\overline{\Delta' M^C}, \overline{\Delta M^C})$ | 0.6255 | 0.4800 | 0.9762 |

these two runs. The $p$-values show that the A-run is statistically highly significantly better than the B-run in terms of nDCG@10 and Q@10, but not with nERR@10 which puts a heavy weight on the first retrieved relevant document. The effect sizes are measured in terms of Glass's $\Delta$ [10]: this is the standardised mean difference where the standard deviation is computed with the baseline run.

The bottom half of Table 4 shows similar information for the replicated runs of MPII. It can be observed that the replication is quite successful overall, since their A-run does outperform their B-run on average, *and* the difference is statistically significant in terms of all three measures (even with nERR@10). That is, RMIT's original claim in their NTCIR-13 paper "SDM outperforms FDM" is confirmed in MPII's replication experiment. As for the effect sizes, the replicated runs yield slightly smaller Glass's $\Delta$ values relative to the original ones.

Table 5 summarises the results of the T1 subtask in terms of the aforementioned evaluation measures for replicability. The top half of this table discusses *topicwise* replicability in terms of RSME and the Pearson correlation $r$: it can be observed that, *topicwise* replicability is a very difficult problem. Only the $r$ for nERR@10 shows a statisticallly significantly positive correlation ($p = 0.0089$, 95%CI[0.0673, 0.4345]). On the other hand, the bottom half of this table discusses the overall replicability in terms of ER: it can be observed that the ER for nERR@10 is very close to 1 (0.9762), as the mean differences before and after replication are very similar (0.0486 vs. 0.0475). In summary, while MPII's

replicated runs are not very successful when viewed per topic, they are quite successful at the effect size level, especially with nERR@10.

Figure 1 visualises the correlation between the replicated per-topic deltas and the original ones.

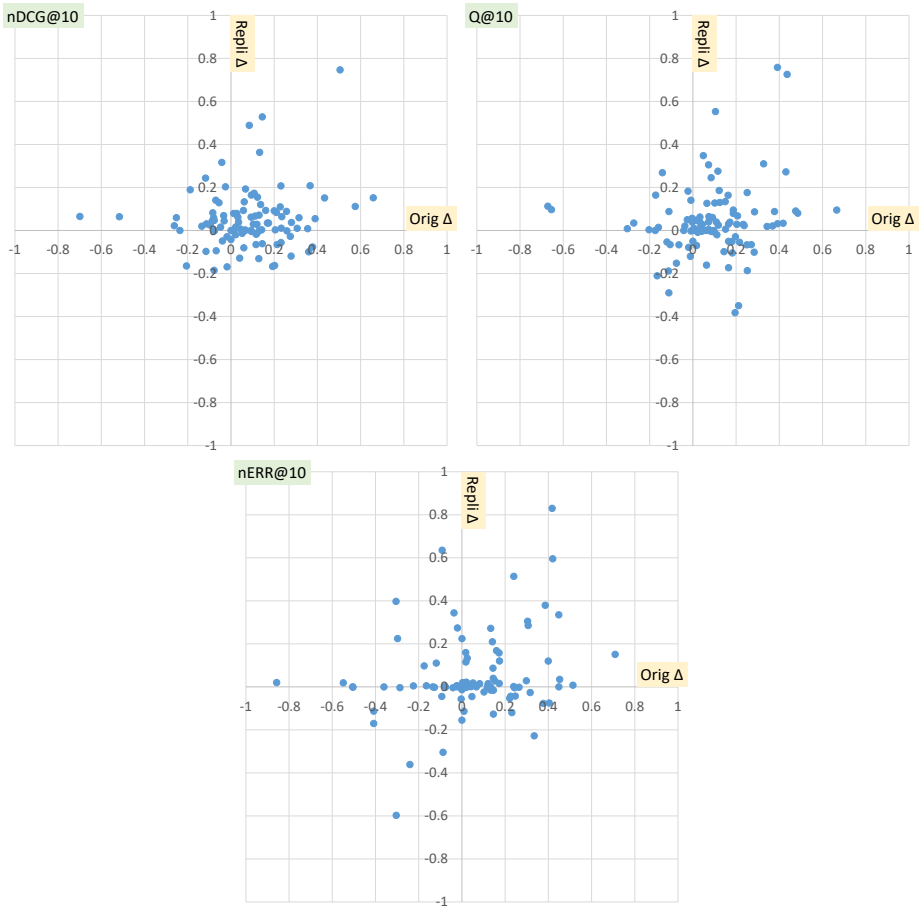


**Fig. 1.** Per-topic deltas of MPII's replicated runs plotted against the corresponding deltas of the original RMIT runs (based on the WWW-1 qrels).

**Table 6.** Effectiveness scores of the TREC Delaware runs ($n = 50$ topics) and those of the T2TREC runs from MPII based on the WWW-1 qrels ($n = 100$ topics). $P$-values smaller than 5% are indicated in bold.

| | Mean nDCG@10 | Mean Q@10 | Mean nERR@10 |
|---|---|---|---|
| `UDInfolabWEB2` | 0.3477 | 0.2937 | 0.4634 |
| `UDInfolabWEB1` | 0.2514 | 0.2336 | 0.3097 |
| (Paired $t$-test $p$-value) | **(0.0023)** | (0.0631) | **(0.0012)** |
| (Glass's $\Delta$) | (0.3834) | (0.2197) | (0.5240) |
| `CENTRE-1-MPII-T2TREC-A` | 0.5019 | 0.4595 | 0.6600 |
| `CENTRE-1-MPII-T2TREC-B` | 0.4271 | 0.3940 | 0.5525 |
| (Paired $t$-test $p$-value) | **(0.0045)** | **(0.0189)** | **(0.0021)** |
| (Glass's $\Delta$) | (0.2478) | (0.2074) | (0.3013) |

**Table 7.** T2TREC results for MPII based on the WWW-1 qrels.

| | nDCG@10 | Q@10 | nERR@10 |
|---|---|---|---|
| $\overline{\Delta M^D}$ | 0.0963 | 0.0601 | 0.1536 |
| $\overline{\Delta' M^C}$ | 0.0748 | 0.0655 | 0.1075 |
| $ER(\overline{\Delta' M^C}, \overline{\Delta M^D})$ | 0.7767 | 1.0893 | 0.6997 |

### 5.2 T2TREC: Reproducibility Results

The top half of Table 6 shows the mean effectiveness scores of the original A-run and B-run from Delaware on the TREC 2013 Web Track Adhoc Task test collection, together with information about the differences between these two runs. Note that these are *not* the official TREC scores; rather, they were re-evaluated using `NTCIREVAL`. The $p$-values show that the A-run is statistically significantly better than the B-run in terms of nDCG@10 and nERR@10, but not with Q@10. As before, Glass's $\Delta$ values are shown beneath the $p$-values.

The bottom half of Table 6 shows similar information for the reproduced runs of MPII. It can be observed that the reproduction is quite successful overall, since their A-run does outperform their B-run on average, *and* the difference is statistically significant in terms of all three measures (even with Q@10). That is, Delaware's original claim in their TREC 2013 paper "web-based working sets outperform collection-based working sets" is confirmed in MPII's reproduction experiment. As for the effect sizes, the reproduced runs yield slightly smaller Glass's $\Delta$ values relative to the original ones, especially in terms of nDCG@10 and nERR@10.

Table 7 summarises the results of the T2TREC subtask in terms ER, the measure for reproducibility. It can be observed that the ER for Q@10 is very close to 1 (1.0893), as the mean differences before and after reproduction are very similar (0.0601 vs. 0.0655). In summary, MPII's reproduced runs are quite successful, especially with Q@10.

**Table 8.** Effectiveness scores of the T2OPEN runs from MPII based on the WWW-1 qrels ($n = 100$ topics).

| | Mean nDCG | Mean Q | Mean nERR |
|---|---|---|---|
| CENTRE-1-MPII-T2OPEN-A | 0.5279 | 0.5349 | 0.6587 |
| CENTRE-1-MPII-T2OPEN-B | 0.5147 | 0.5198 | 0.6449 |
| (Paired $t$-test $p$-value) | (0.4591) | (0.4678) | (0.6018) |
| (Glass's $\Delta$) | (0.0515) | (0.0519) | (0.0472) |

### 5.3 T2OPEN: Reproducibility Results

Table 8 shows the mean effectiveness scores of the T2OPEN reproduced runs from MPII, together with information about the differences between these two runs. It can be observed that none of the differences between the two OPEN runs are statistically significant. As the choice of target runs to reproduce was left to the participating team, we leave the discussion of the T2OPEN results to MPII's participant paper.

## 6 Results with the New CENTRE-1 Qrels

This section evaluates the submitted runs using the new CENTRE-1 qrels, to ensure that the MPII runs are evaluated fairly in comparison with the WWW-1 runs.

### 6.1 T1: Replicability Results

Table 9 shows the effectiveness scores of the original and replicated A-run and B-run as well as the $p$-values and effect sizes in a way similar to Table 4. Table 10 summarises the results of the T1 subtask in a way similar to Table 5. By comparing Tables 9 and 10 (i.e., results based on the CENTRE-1 qrels) with Tables 4 and 5 ((i.e., results based on the WWW-1 qrels), it can be observed that the absolute effectiveness scores, the significance test results, and the effect sizes are generally very similar regardless of the qrels. That is, the impact of adding 2,553 new topic-document pairs to the WWW-1 qrels is small for the replication experiment.

Figure 2 visualises the correlation between the replicated per-topic deltas and the original ones.

**Table 9.** Effectiveness scores based on the CENTRE-1 qrels ($n = 100$ topics). $P$-values smaller than 5% are indicated in bold.

| | Mean nDCG@10 | Mean Q@10 | Mean nERR@10 |
|---|---|---|---|
| Original A: `RMIT-E-NU-Own-1` | 0.6250 | 0.6503 | 0.7436 |
| Original B: `RMIT-E-NU-Own-3` | 0.5444 | 0.5616 | 0.6954 |
| (Paired $t$-test $p$-value) | **(9.099e-05)** | **(3.117-e05)** | (0.0532) |
| (Glass's $\Delta$) | (0.3388) | (0.3284) | (0.1809) |
| `CENTRE-1-MPII-T1-A` | 0.5909 | 0.6026 | 0.7385 |
| `CENTRE-1-MPII-T1-B` | 0.5384 | 0.5538 | 0.6917 |
| (Paired $t$-test $p$-value) | **(0.0002)** | **(0.0036)** | **(0.0133)** |
| (Glass's $\Delta$) | (0.2119) | (0.1723) | (0.1660) |

**Table 10.** T1 results for MPII based on the CENTRE-1 qrels. $P$-values smaller than 5% are indicated in bold.

| | nDCG@10 | Q@10 | nERR@10 |
|---|---|---|---|
| RMSE | 0.2230 | 0.2377 | 0.2657 |
| $r$ (95%CI, $p$-value) | 0.1560 | 0.1900 | 0.2610 |
| | [−0.0416, 0.3420] | [−0.0066, 0.3726] | [0.0680, 0.4351] |
| | $p = 0.1211$ | $p = 0.0582$ | $\mathbf{p = 0.0087}$ |
| $\overline{\Delta M^C}$ | 0.0806 | 0.0887 | 0.0482 |
| $\overline{\Delta' M^C}$ | 0.0525 | 0.0488 | 0.0467 |
| $ER(\overline{\Delta' M^C}, \overline{\Delta M^C})$ | 0.6519 | 0.5508 | 0.9689 |

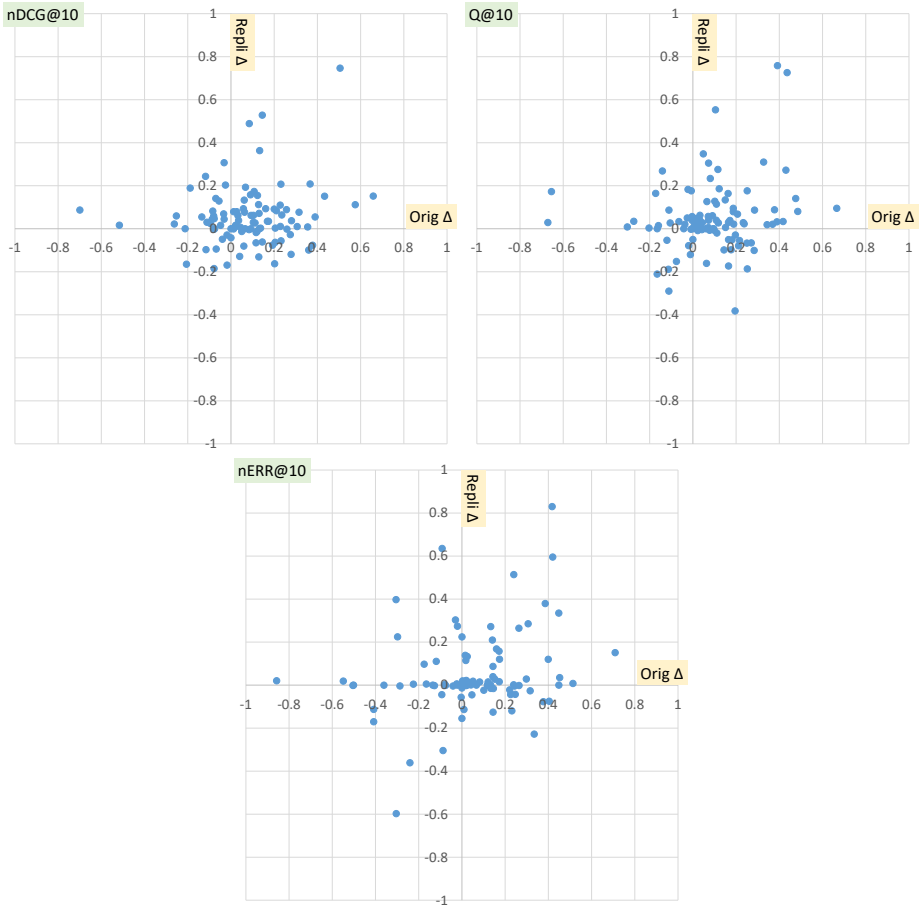


**Fig. 2.** Per-topic deltas of MPII's replicated runs plotted against the corresponding deltas of the original RMIT runs (based on the CENTRE-1 qrels).

**Table 11.** Effectiveness scores of the TREC Delaware runs ($n = 50$ topics) and those of the T2TREC runs from MPII based on the CENTRE-1 qrels ($n = 100$ topics).

| | Mean nDCG | Mean Q | Mean nERR |
|---|---|---|---|
| `UDInfolabWEB2` | 0.3477 | 0.2937 | 0.4634 |
| `UDInfolabWEB1` | 0.2514 | 0.2336 | 0.3097 |
| (Paired $t$-test $p$-value) | **(0.0023)** | (0.0631) | **(0.0012)** |
| (Glass's $\Delta$) | (0.3834) | (0.2197) | (0.5240) |
| `CENTRE-1-MPII-T2TREC-A` | 0.5800 | 0.5837 | 0.7159 |
| `CENTRE-1-MPII-T2TREC-B` | 0.4777 | 0.4808 | 0.6039 |
| (Paired $t$-test $p$-value) | **(2.107e-06)** | **(7.705e-06)** | **(0.0003)** |
| (Glass's $\Delta$) | (0.3475) | (0.3201) | (0.3341) |

**Table 12.** T2TREC results for MPII based on the CENTRE-1 qrels.

| | nDCG@10 | Q@10 | nERR@10 |
|---|---|---|---|
| $\overline{\Delta M^D}$ | 0.0963 | 0.0601 | 0.1536 |
| $\overline{\Delta' M^C}$ | 0.1023 | 0.1029 | 0.1120 |
| $ER(\overline{\Delta' M^C}, \overline{\Delta M^D})$ | 1.0630 | 1.7116 | 0.7287 |

### 6.2 T2TREC: Reproducibility Results

Table 11 shows the effectiveness scores of the original and reproduced A-run and B-run as well as the $p$-values and effect sizes in a way similar to Table 6. Note that the TREC run scores have just been copied from Table 6. Table 12 summarises the results of the T2TREC subtask in a way similar to Table 7. By comparing Tables 11 and 12 (i.e., results based on the CENTRE-1 qrels) with Tables 6 and 7 (i.e., results based on the WWW-1 qrels), it can be observed that the reproduced results look substantially more successful with the new CENTRE-1 qrels: the absolute scores have improved (e.g., from 0.5019 to 0.5800 in Mean nDCG) and hence the effect sizes are now larger; the $p$-values are now much smaller (e.g., down from 0.0045 to 2.107e-06 for nDCG), and the ERs have also improved (e.g., up from 0.7767 to 1.0630 for nDCG, indicating almost perfect effect size reproduction). This shows that the reproduced MPII runs were indeed underestimated using the original WWW-1 qrels and that patching it up was worthwhile, although this was not clear from the *replicability* results.

### 6.3 T2OPEN: Reproducibility Results

Table 13 is a table that updates Table 8, by replacing the WWW1 qrels with the new CENTRE1 qrels. It can be observed that none of the differences between the two OPEN runs are statistically significant. Moreover, the numbers in the two tables are very similar. We leave the discussion of the T2OPEN results to MPII's participant paper.

**Table 13.** Effectiveness scores of the T2OPEN runs from MPII based on the CENTRE-1 qrels ($n = 100$ topics).

| | Mean nDCG | Mean Q | Mean nERR |
|---|---|---|---|
| CENTRE-1-MPII-T2OPEN-A | 0.5237 | 0.5311 | 0.6558 |
| CENTRE-1-MPII-T2OPEN-B | 0.5104 | 0.5160 | 0.6416 |
| (Paired $t$-test $p$-value) | (0.4557) | (0.4680) | (0.5919) |
| (Glass's $\Delta$) | (0.0521) | (0.0521) | (0.0486) |

## 7 Conclusions

Based on the results with the WWW-1 qrels (whose pool files did not consider the new MPII runs), MPII is quite successful in both T1 and T2TREC subtasks in terms of effect ratio (ER), which compares the replicated mean difference against the original mean difference. The replicated and reproduced A-runs statistically significantly outperformed the corresponding B-runs in terms of all three evaluation measures, confirming the claims in the original papers: thus, the results from RMIT were replicated successfully and the TREC results from Delaware were reproduced successfully. On the other hand, for the T1 task, MPII was not successful at replicating the *topicwise* delta; this is clearly a much harder problem to tackle.

MPII's replicability and open-reproducibility results based on the CENTRE-1 qrels (which includes 2,553 topic-document pairs newly contributed by the MPII runs) are very similar to those based on the WWW-1 qrels. On the other hand, our TREC-reproducibility results based on the CENTRE-1 qrels look substantially more successful than those based on the WWW-1 qrels, demonstrating that the additional relevance assessments were worthwhile.

Given the fact that we had only one participating team at NTCIR-14 CENTRE, we propose that we run CENTRE at NTCIR-15 as a subtask of the Third We Want Web (WWW) task (WWW-3). Thus we can continute to work on the clueweb12-13B corpus, and select target runs to replicate from top-performing NTCIR-14 WWW-2 runs. MPII's replication and reproduction efforts at NTCIR-14 CENTRE were quite successful in terms of ER: will we observe similar successes for different target runs, with different CENTRE participants?

There will be another round of CENTRE at CLEF 2019 [3], but unfortunately there will be no CENTRE track at TREC2019. However, CLEF, NTCIR, and TREC will continue the collaboration, and CENTRE will hopefully continue to stand for *C*L*E*F *N*TCIR *T*REC *RE*produciblity.

## References


1. Collins-Thompson, K., Bennett, P., Diaz, F., Clarke, C.L., Voorhees, E.M.: TREC 2013 web track overview (2014)
2. Fang, H., Zhai, C.: Semantic term matching in axiomatic approaches to information retrieval. In: Proceedings of ACM SIGIR 2006. pp. 115–122 (2006)
3. Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Soboroff, I.: CENTRE@CLEF 2019. In: Proceedings of ECIR 2019 Part II (LNCS 11438). pp. 283–290 (2019)
4. Ferro, N., Maistro, M., Sakai, T., Soboroff, I.: Overview of CENTRE@CLEF 2018: a first tale in the systematic reproducibility realm. In: Proceedings of CLEF 2018 (LNCS 11018). pp. 239–246 (2018)
5. Gallagher, L., Mackenzie, J., Benham, R., Chen, R.C., Scholer, F., Culpepper, J.: RMIT at the NTCIR-13 we want web task. In: Proceedings of NTCIR-13. pp. 402–406 (2017), http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/02-NTCIR13-WWW-GallagherL.pdf
6. Guo, J., Fan, Y., Ai, Q., Croft, W.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of ACM CIKM 2016. pp. 55–64 (2016)
7. Luo, C., Sakai, T., Liu, Y., Dou, Z., Xiong, C., Xu, J.: Overview of the NTCIR-13 we want web task. In: Proceedings of NTCIR-13. pp. 394–401 (2017), http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-NTCIR13-OV-WWW-LuoC.pdf
8. Metzler, D., Croft, W.: A markov random field model for term dependencies. pp. 472–479 (2005)
9. Sakai, T.: Metrics, statistics, tests. In: PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173). pp. 116–163 (2014)
10. Sakai, T.: Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power. Springer (2018), https://link.springer.com/book/10.1007/978-981-13-1199-4
11. Sakai, T., Kando, N.: Are popular documents more likely to be relevant? a dive into the ACLIA IR4QA pools. In: Proceedings of EVIA 2008. pp. 8–9 (2008)
12. Soboroff, I., Ferro, N., Sakai, T.: Overview of the TREC 2018 CENTRE track. In: Proceedings of TREC 2018 (2019)
13. Yang, P., Fang, H.: Evaluating the effectiveness of axiomatic approaches in web track. In: Proceedings of TREC 2013 (2014)
14. Yates, A.: MPII at the NTCIR-14 CENTRE task. In: Online Proceedings of NTCIR-14 (2019)