Motivation

Why did I participate?

● Reproducibility is important! Let’s support it
● Didn’t hurt that I had implementations available

We need incentives to reproduce & to make reproducible
Outline

● Other types of reproducibility
● Subtasks
  ○ T1
  ○ T2TREC
  ○ T2OPEN
● Conclusion
ACM Artifact Review and Badging (OSIRRC ‘19 version)

- Replicability (different team, same experimental setup): an independent group can obtain the same result using the author’s own artifacts.
- Reproducibility (different team, different experimental setup): an independent group can obtain the same result using artifacts which they develop completely independently.

https://www.acm.org/publications/policies/artifact-review-badging
ACM Artifact Review and Badging
(OSIRRC ‘19 version)

Replicability: different team, same experimental setup … same result?
Reproducibility: different team, different experimental setup … same result?

- T1: replication of WWW-1 runs
- T2TREC: reproduction of TREC WT13 run on WWW-1
  Used new implementation (Anserini) by one of runs’ authors.
  Making this replication? (but what about data change?)
- T2OPEN: open-ended reproduction

https://www.acm.org/publications/policies/artifact-review-badging
Outline

● Other types of reproducibility

● Subtasks
  ○ T1
  ○ T2TREC
  ○ T2OPEN

● Conclusion
Subtask T1: Replicability

SDM (A) > FDM (B)?

Obtained details from RMIT's overview paper:

- Indri, Krovetz stemming, keep stopwords
- Spam scores for filtering docs
- MRF params: field weights (title, body, inlink)
- RM3 params: FB docs, FB terms, orig query weight
Subtask T1: Replicability

Metrics

● Topicwise: do same topics perform similarly?
  RMSE & Pearson’s $r$

● Overall: is the mean performance similar?
  Effect Ratio (ER)
## Subtask T1: Replicability

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

<table>
<thead>
<tr>
<th></th>
<th>Mean nDCG@10</th>
<th>Mean Q@10</th>
<th>Mean nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original A: RMIT-E-NU-Own-1</td>
<td>0.6302</td>
<td>0.6548</td>
<td>0.7463</td>
</tr>
<tr>
<td>Original B: RMIT-E-NU-Own-3</td>
<td>0.5493</td>
<td>0.5657</td>
<td>0.6977</td>
</tr>
<tr>
<td>(Paired (t)-test (p)-value)</td>
<td>(9.057e-05)</td>
<td>(2.937e-05)</td>
<td>(0.0519)</td>
</tr>
<tr>
<td>(Glass’s (\Delta))</td>
<td>(0.3358)</td>
<td>(0.3267)</td>
<td>(0.1823)</td>
</tr>
<tr>
<td>CENTRE-1-MPII-T1-A</td>
<td>0.5933</td>
<td>0.5996</td>
<td>0.7412</td>
</tr>
<tr>
<td>CENTRE-1-MPII-T1-B</td>
<td>0.5428</td>
<td>0.5568</td>
<td>0.6937</td>
</tr>
<tr>
<td>(Paired (t)-test (p)-value)</td>
<td>(4.352e-04)</td>
<td>(0.0128)</td>
<td>(0.0126)</td>
</tr>
<tr>
<td>(Glass’s (\Delta))</td>
<td>(0.2017)</td>
<td>(0.1498)</td>
<td>(0.1687)</td>
</tr>
</tbody>
</table>

All results tables taken from NTCIR-14 CENTRE overview paper.
### Subtask T1: Replicability

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

<table>
<thead>
<tr>
<th></th>
<th>nDCG@10</th>
<th>Q@10</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSE</strong>&lt;br&gt; $r$ (95%CI, $p$-value)</td>
<td>0.2256</td>
<td>0.2431</td>
<td>0.2668</td>
</tr>
<tr>
<td></td>
<td>0.1469</td>
<td>0.1797</td>
<td>0.2603</td>
</tr>
<tr>
<td></td>
<td>$[-0.0510, 0.3337]$</td>
<td>$[-0.0174, 0.3633]$</td>
<td>$[0.0673, 0.4345]$</td>
</tr>
<tr>
<td></td>
<td>$p = 0.1446$</td>
<td>$p = 0.0737$</td>
<td>$p = 0.0089$</td>
</tr>
<tr>
<td><strong>$\Delta M^C$</strong></td>
<td>0.0809</td>
<td>0.0891</td>
<td>0.0486</td>
</tr>
<tr>
<td><strong>$\Delta'M^C$</strong></td>
<td>0.0506</td>
<td>0.0428</td>
<td>0.0475</td>
</tr>
<tr>
<td><strong>$ER(\Delta'M^C, \Delta M^C)$</strong></td>
<td>0.6255</td>
<td>0.4800</td>
<td>0.9762</td>
</tr>
</tbody>
</table>

All results tables taken from NTCIR-14 CENTRE overview paper.
Subtask T1: Replicability

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

<table>
<thead>
<tr>
<th></th>
<th>nDCG@10</th>
<th>Q@10</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSE</strong></td>
<td>0.2256</td>
<td>0.2431</td>
<td>0.2668</td>
</tr>
<tr>
<td>r (95%CI, <em>p</em>-value)</td>
<td>0.1469</td>
<td>0.1797</td>
<td>0.2603</td>
</tr>
<tr>
<td>[-0.0510, 0.3337]</td>
<td></td>
<td>[-0.0174, 0.3633]</td>
<td>[0.0673, 0.4345]</td>
</tr>
<tr>
<td></td>
<td>0.1446</td>
<td>0.0737</td>
<td></td>
</tr>
<tr>
<td><strong>ΔMC</strong></td>
<td>0.0809</td>
<td>0.0891</td>
<td>0.0486</td>
</tr>
<tr>
<td><strong>Δ'MC</strong></td>
<td>0.0506</td>
<td>0.0428</td>
<td>0.0475</td>
</tr>
<tr>
<td><strong>ER(Δ'MC, ΔMC)</strong></td>
<td>0.6255</td>
<td>0.4800</td>
<td>0.9762</td>
</tr>
</tbody>
</table>

All results tables taken from NTCIR-14 CENTRE overview paper.
Figure taken from NTCIR-14 CENTRE overview paper.
Subtask T1: Replicability

Why were the topicwise results lower?

- Indri v5.12 (me) vs. v5.11 (RMIT)
- Scaling of unordered window size (fixed 8 vs. 4*n)
- Did not use inlinks field
  - *harvestlinks* ran for 1-2 weeks, then crashed (several times)
  - Possible it was a fault of network storage corpus was on
Subtask T1: Replicability

Is SDM (A) better than FDM (B) on CW12 B13 (C)?

➔ Yes, assuming all parameters are fixed (!)

What if spam filtering changes? Title field weight? ...

We now know I ran Indri (mostly) the way RMIT ran Indri. This doesn’t say much about SDM vs. FDM!
Subtask T1: Replicability

Is SDM (A) better than FDM (B)?

➔ Yes, assuming all parameters are fixed (!)

What if spam filtering changes? Title field weight? ...

We now know I ran Indri (mostly) the way RMIT ran Indri.
This doesn’t say much about SDM vs. FDM!

Where does “consideration of the comprehensiveness of parameter tuning” fit into the reproducibility classification?

Annoying pessimist says: we’re making things worse by reinforcing conclusions that may depend on original work’s poor param choices.

Me: I’m not implying RMIT’s tuning was wrong in any way (& don’t think we’re making situation worse). But how do we consider tuning?
Subtask T1: Replicability

Is SDM (A) better than FDM (B)?

➔ Yes, assuming all parameters are fixed (!)

What if spam filtering changes? Title field weight? ...

We now know I ran Indri (mostly) the way RMIT ran Indri.

This doesn’t say much about SDM vs. FDM!

How do we consider tuning?

One possibility: rather than fixing parameters, report all grid search details in original work & re-run grid search when reproducing?

➔ Replication verifies both chosen params from grid search and model performance
➔ Not always possible (e.g., reasonable param grid too large to confidently search)
➔ Requires specifying train/dev data along with collection C

One alternative: assume chosen params fine?
Subtask T2TREC

Is A better than B on a different collection C?

Obtained details from UDel's overview paper

- Semantic expansion parameters (with F2-LOG)
- Weight given to expansion terms ($\beta$)
Subtask T2TREC

Known differences:

- Assumed Porter stemmer & Lucene tokenization
- Two commercial search engines (vs. 3 unnamed ones)
- CW12 B13 instead of full CW12
- TREC Web Track 2014 data to check correctness
**Subtask T2TREC**

Known differences:

- Assumed Porter stemmer & Lucene tokenization
- Two commercial search engines (vs. 3 unnamed ones)
- CW12 B13 instead of full CW12
- TREC Web Track 2014 data to check correctness

**Dilemma with A run:**

- UDel reported $\beta=1.7$ (term weight)
- On WT14, $\beta=0.1$ better for us
- Reproduce with same params?

Given new data and changes, set $\beta=0.1$ (we did not change other params)
## Subtask T2TREC

**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.

<table>
<thead>
<tr>
<th></th>
<th>Mean nDCG@10</th>
<th>Mean Q@10</th>
<th>Mean nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDInfolabWEB2</td>
<td>0.3477</td>
<td>0.2937</td>
<td>0.4634</td>
</tr>
<tr>
<td>UDInfolabWEB1</td>
<td>0.2514</td>
<td>0.2336</td>
<td>0.3097</td>
</tr>
<tr>
<td>(Paired t-test p-value)</td>
<td>(0.0023)</td>
<td>(0.0631)</td>
<td>(0.0012)</td>
</tr>
<tr>
<td>(Glass’s Δ)</td>
<td>(0.3834)</td>
<td>(0.2197)</td>
<td>(0.5240)</td>
</tr>
<tr>
<td>CENTRE-1-MPII-T2TREC-A</td>
<td>0.5019</td>
<td>0.4595</td>
<td>0.6600</td>
</tr>
<tr>
<td>CENTRE-1-MPII-T2TREC-B</td>
<td>0.4271</td>
<td>0.3940</td>
<td>0.5525</td>
</tr>
<tr>
<td>(Paired t-test p-value)</td>
<td>(0.0045)</td>
<td>(0.0189)</td>
<td>(0.0021)</td>
</tr>
<tr>
<td>(Glass’s Δ)</td>
<td>(0.2478)</td>
<td>(0.2074)</td>
<td>(0.3013)</td>
</tr>
</tbody>
</table>

All results tables taken from NTCIR-14 CENTRE overview paper.
Subtask T2TREC

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

<table>
<thead>
<tr>
<th></th>
<th>nDCG@10</th>
<th>Q@10</th>
<th>nERR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta M^D$</td>
<td>0.0963</td>
<td>0.0601</td>
<td>0.1536</td>
</tr>
<tr>
<td>$\frac{\Delta' M^C}{\Delta M^C}$</td>
<td>0.0748</td>
<td>0.0655</td>
<td>0.1075</td>
</tr>
<tr>
<td>$ER(\Delta' M^C, \Delta M^D)$</td>
<td>0.7767</td>
<td>1.0893</td>
<td>0.6997</td>
</tr>
</tbody>
</table>

All results tables taken from NTCIR-14 CENTRE overview paper.
Subtask T2TREC

Is $A$ better than $B$ on a different collection $C$?

➔ Yes, assuming parameter choices $P$ are fixed

Better than replication situation:
We observed $A > B$ (given $P$) on two collections
(but different $P$ might still change this)
Subtask T2OPEN

Is **A** better than **B** on a different collection **C**?

- Variants of DRMM neural model for both A and B
- DRMM’s input is a histogram of (query, doc term) embedding similarities for each query term
- Taking log of histogram (A) was better across datasets, metrics, and TREC title vs. description queries

A Deep Relevance Matching Model for Ad-hoc Retrieval.
Jiafeng Guo, Yixing Fan, Qingyao Ai, W. Bruce Croft. CIKM 2016.
Subtask T2OPEN

Is DRMM with LCH better on a different collection $\mathbf{C}$?

- Implemented DRMM & checked against other code
- Trained on TREC WT2009-2013 & validated on WT14
- Tuned hyperparameters

A Deep Relevance Matching Model for Ad-hoc Retrieval.
Jiafeng Guo, Yixing Fan, Qingyao Ai, W. Bruce Croft. CIKM 2016.
Subtask T2OPEN

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

<table>
<thead>
<tr>
<th></th>
<th>Mean nDCG</th>
<th>Mean Q</th>
<th>Mean nERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENTRE-1-MPII-T2OPEN-A</td>
<td>0.5279</td>
<td>0.5349</td>
<td>0.6587</td>
</tr>
<tr>
<td>CENTRE-1-MPII-T2OPEN-B</td>
<td>0.5147</td>
<td>0.5198</td>
<td>0.6449</td>
</tr>
<tr>
<td>(Paired t-test $p$-value)</td>
<td>(0.4591)</td>
<td>(0.4678)</td>
<td>(0.6018)</td>
</tr>
<tr>
<td>(Glass’s $\Delta$)</td>
<td>(0.0515)</td>
<td>(0.0519)</td>
<td>(0.0472)</td>
</tr>
</tbody>
</table>

High $p$-value. Tuning differences? Dataset? Just a small effect?

All results tables taken from NTCIR-14 CENTRE overview paper.
Conclusion

- Successful overall reproductions for T1 and T2TREC
- Can reproducibility incentives be stronger?
- When we replicate, how best to deal with tuning? Ignore? Report grid search? Do we fix train/dev then?
- Faithfulness to original setup sometimes conflicts with using best parameters (given specific training/dev set)
Conclusion

- Successful overall reproductions for T1 and T2TREC
- Can reproducibility incentives be stronger?
- When we replicate, how best to deal with tuning?
  Ignore? Report grid search? Do we fix train/dev then?
- Faithfulness to original setup sometimes conflicts with using best parameters (given specific training/dev set)

Thanks!