MPII at the NTCIR-14 CENTRE Task

Andrew Yates
Max Planck Institute for Informatics

MAX-PLANCK-GESELLSCHAFT

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

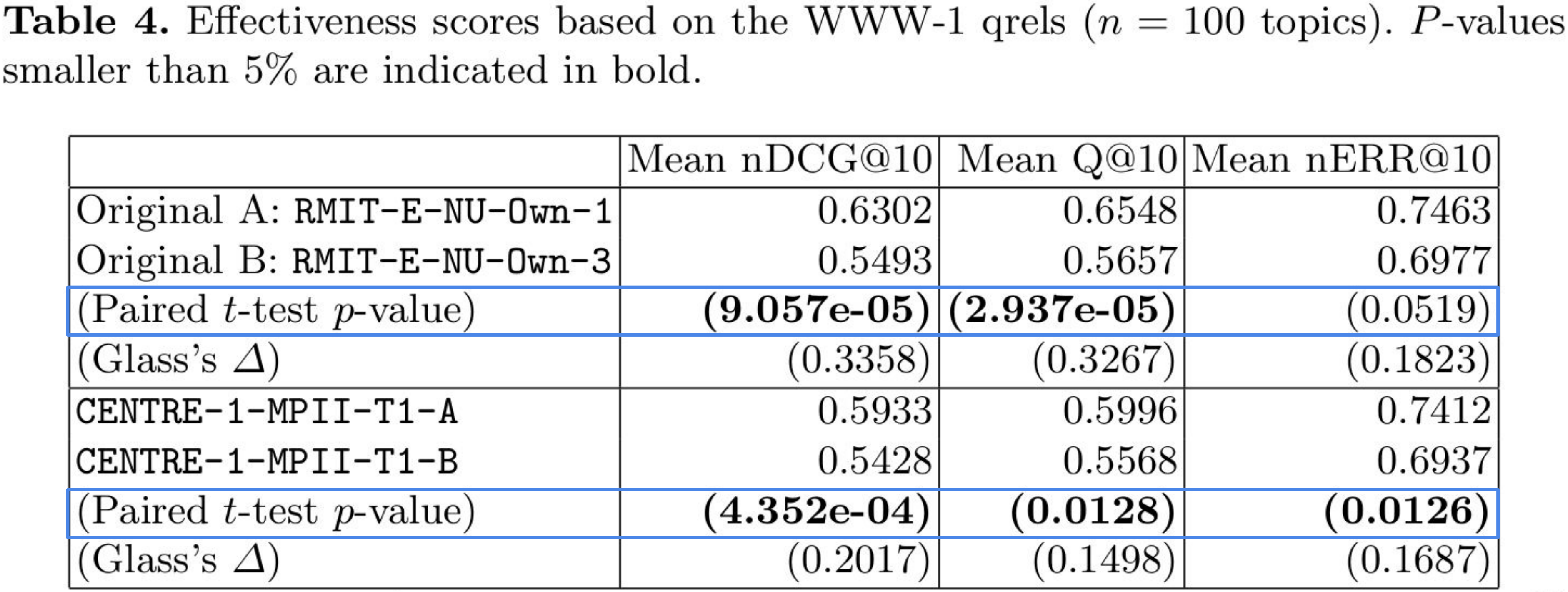

Table 4. Effectiveness scores based on the WWW-1 qrels ($n = 100$ topics). P-values smaller than 5% are indicated in bold.

	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 Δ)	(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 Δ)	(0.2017)	(0.1498)	(0.1687)

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

Subtask T1: Replicability

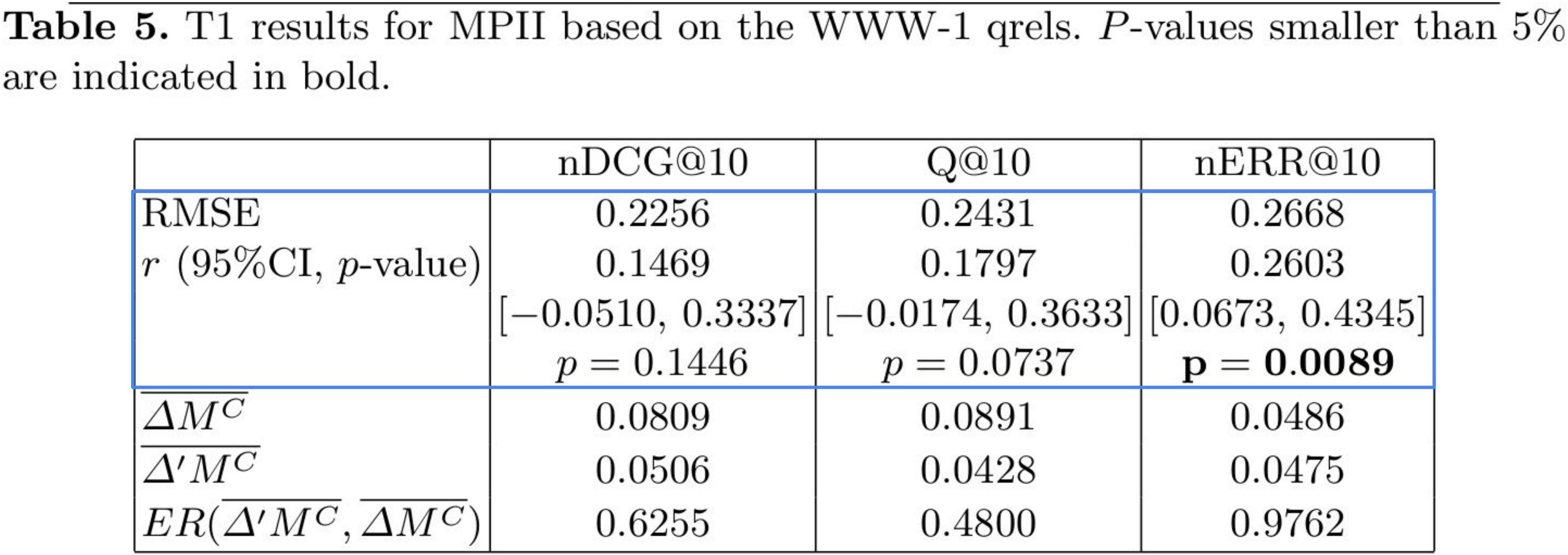

Table 5. T1 results for MPII based on the WWW-1 qrels. *P*-values smaller than 5% are indicated in bold.

	nDCG@10	Q@10	nERR@10
RMSE	0.2256	0.2431	0.2668
r (95%CI, p-value)	0.1469 $[-0.0510, 0.3337]$ $p = 0.1446$	0.1797 $[-0.0174, 0.3633]$ $p = 0.0737$	0.2603 $[0.0673, 0.4345]$ $\mathbf{p = 0.0089}$
$\overline{\Delta M^C}$	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

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

Subtask T1: Replicability

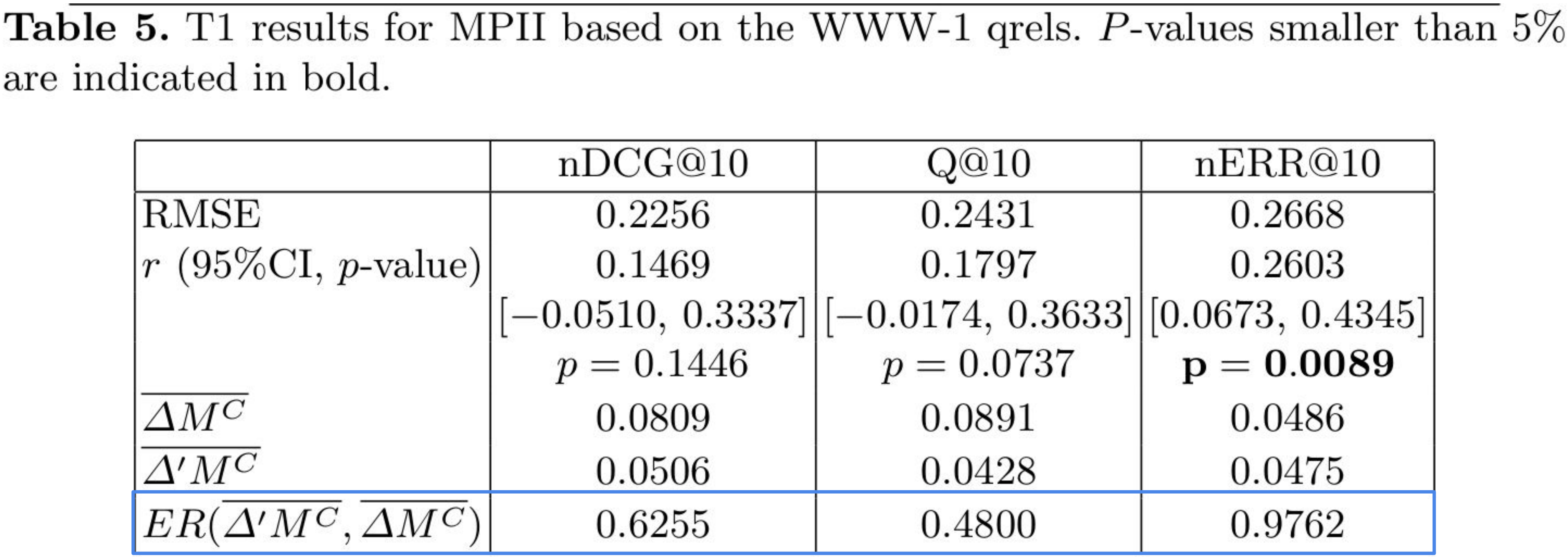

Table 5. T1 results for MPII based on the WWW-1 qrels. *P*-values smaller than 5% are indicated in bold.

	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

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

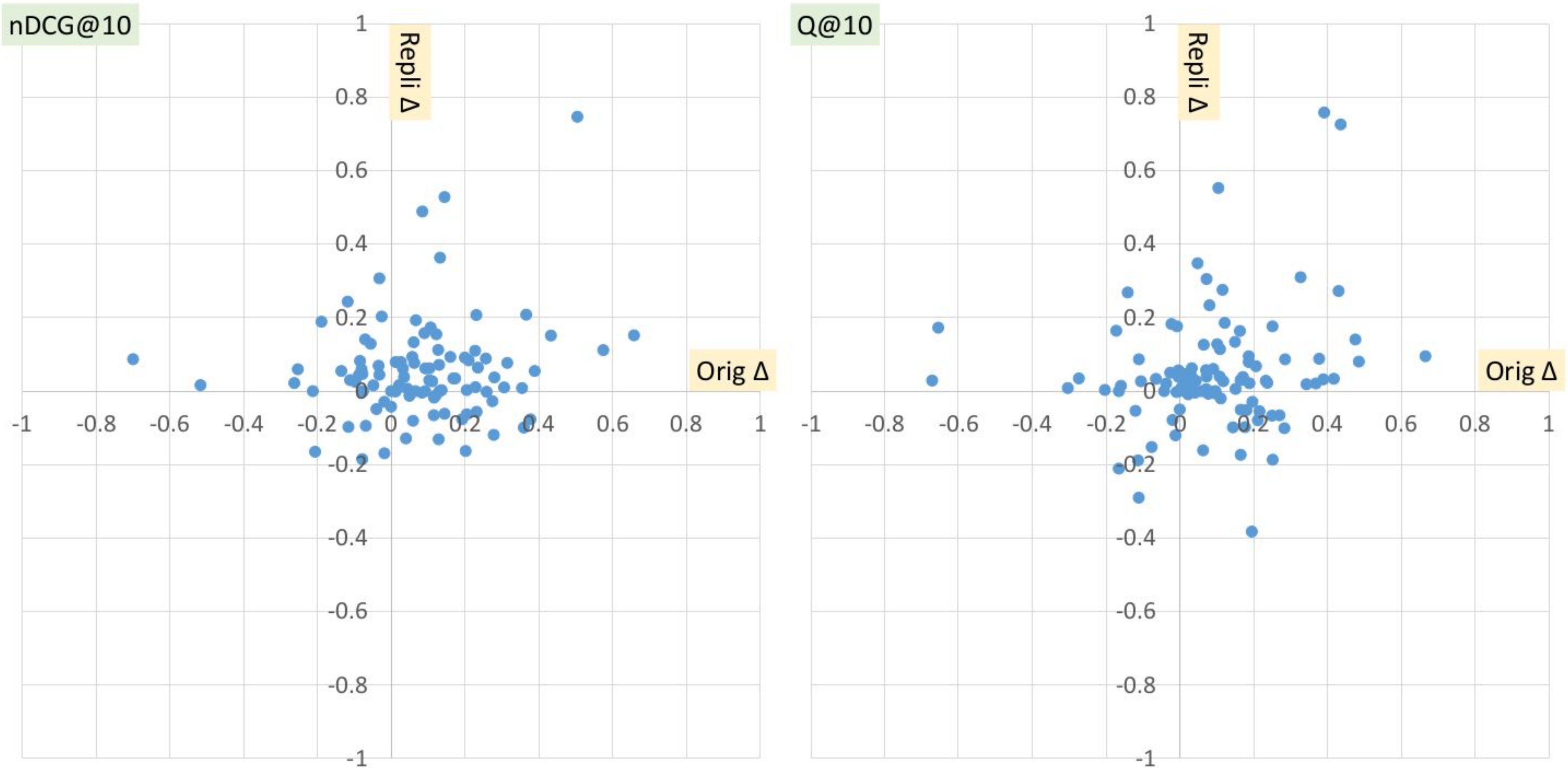

Figure taken from NTCIR-14 CENTRE overview paper.

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 T

Is SDM (**A**) b

→ Yes, assu

What if s nt? ...

We now kno n 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 T

Is SDM (**A**) b

➔ Yes, assu

What if s ht? ...

We now kno n Indri.

This doesn't

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 <u>different</u> collection **C**?

Obtained details from UDel's overview paper

- Semantic expansion parameters (with F2-LOG)
- Weight given to expansion terms (**β**)

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 d

- Assur … on
- Two … ed ones)
- CW1…
- TREC … ess

Dilemma with A run:

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

Given new data and changes, set **β**=0.1
(we did not change other params)

Subtask T2TREC

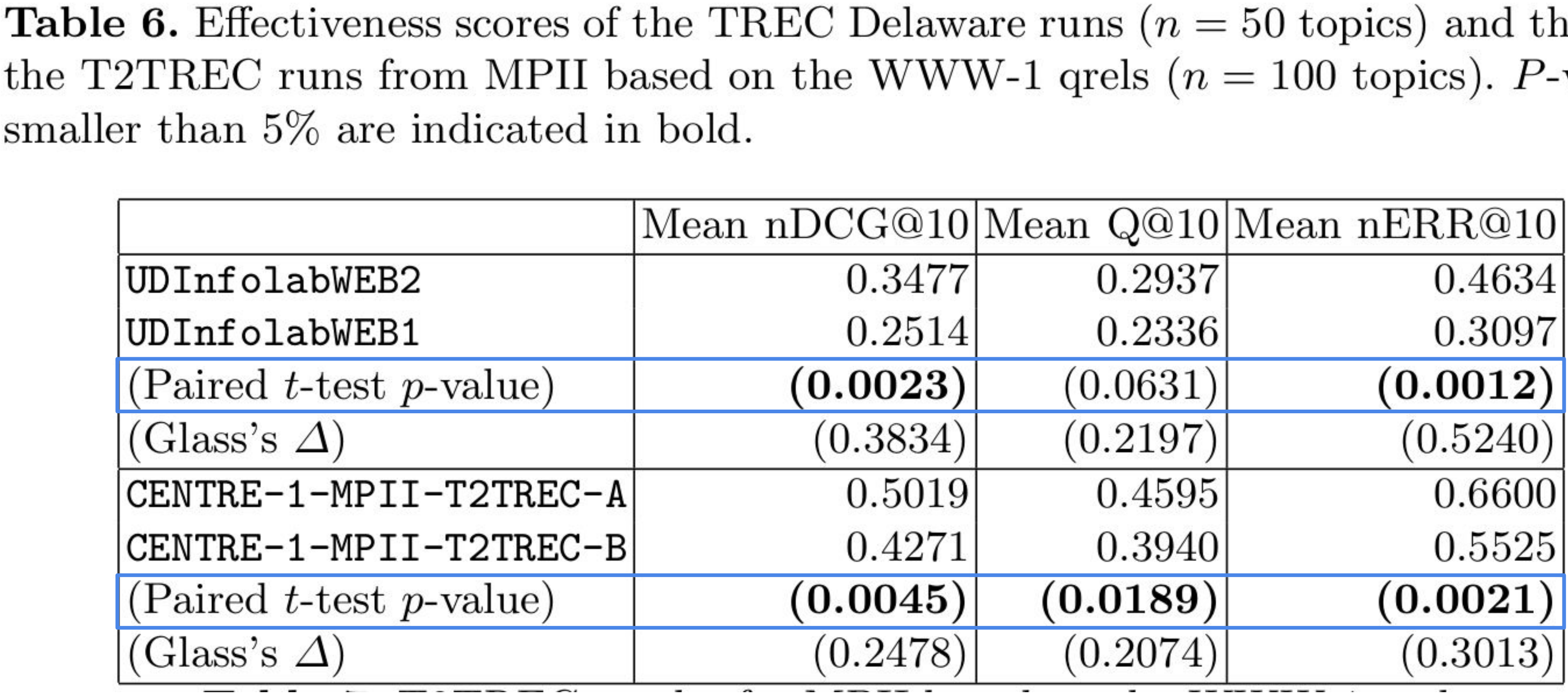

Table 6. Effectiveness scores of the TREC Delaware runs ($n = 50$ topics) and those of the T2TREC runs from MPII based on the WWW-1 qrels ($n = 100$ topics). P-values smaller than 5% are indicated in bold.

	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 Δ)	(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 Δ)	(0.2478)	(0.2074)	(0.3013)

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

Subtask T2TREC

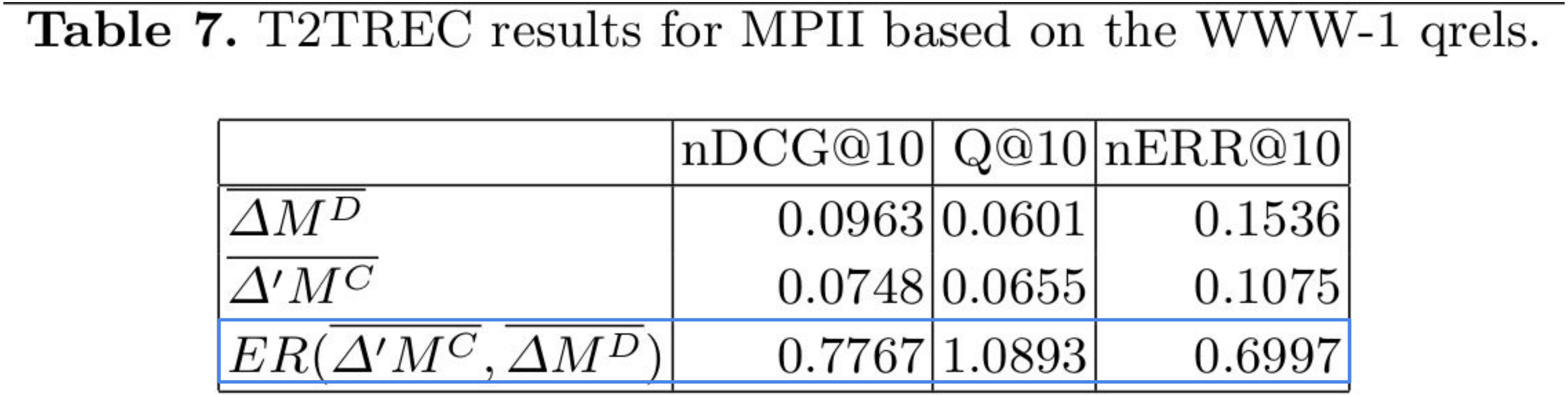

Table 7. T2TREC results for MPII based on the WWW-1 qrels.

	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

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 **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

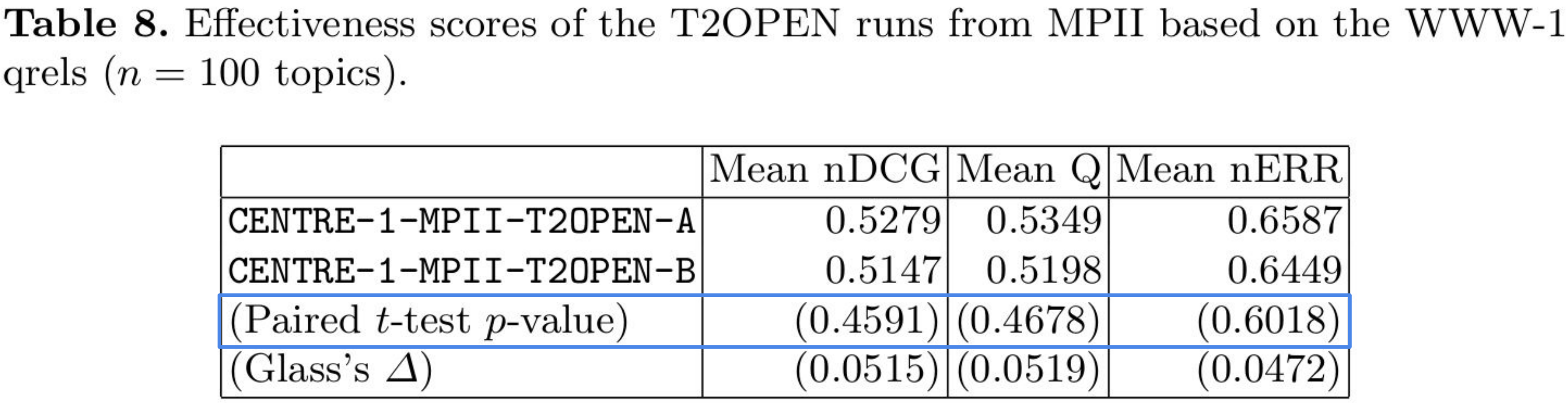

Table 8. Effectiveness scores of the T2OPEN runs from MPII based on the WWW-1 qrels ($n = 100$ topics).

	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 Δ)	(0.0515)	(0.0519)	(0.0472)

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!