

MPII at the NTCIR-14 CENTRE Task

Andrew Yates

Max Planck Institute for Informatics

ayates@mpi-inf.mpg.de @andrewyates

Overview

MPII participated in all three CENTRE subtasks [4] in order to express support for reproducibility in IR research.

Despite known deviations from the original approaches for all three subtasks, we were able to reproduce the overall behavior for two subtasks. Topicwise reproducibility proved more elusive.

Summary of subtasks:

- T1: replicable **overall** but not **topicwise**
- T2TREC: reproducible **overall**
- T2OPEN: **no significant difference**

Subtask T1: Replicability

Goal: replicate improvement from RMIT's WWW-1 runs, in which SDM (Advanced run) improved over FDM (Baseline). [1]

Known deviations: Indri v5.12 (vs. 5.11); used constant unordered window size of 8; unable to index ClueWeb inlinks.

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)	(0.099e-05)	(3.117-e05)	(0.0532)
(Glass's Δ)	(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 Δ)	(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]
	$p = 0.1211$	$p = 0.0582$	$p = 0.0087$
ΔM^C	0.0806	0.0887	0.0482
ΔM^D	0.0525	0.0488	0.0467
$ER(\Delta M^C, \Delta M^D)$	0.6519	0.5508	0.9689

Subtask T2OPEN

Goal: reproduce improvement from DRMM paper, in which logcount based histograms (LCH) improved over count histograms (CH). Authors hypothesized LCH is important in order for model to learn multiplicative relationships. [2]

Known deviations: new implementation; optimal histogram sizes differed.

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 Δ)	(0.0521)	(0.0521)	(0.0486)

Subtask T2TREC

Goal: reproduce improvement from UDel's TREC WT13 runs, in which axiomatic expansion with an external collection (Advanced run) improved over expansion on the target collection (Baseline run). [5]

Known deviations: did not filter CW12 index; Bing and Google snippets; CW12 B rather than A; lowered beta for A run.

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 Δ)	(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 Δ)	(0.3475)	(0.3201)	(0.3341)

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

	nDCG@10	Q@10	nERR@10
ΔM^D	0.0963	0.0601	0.1536
ΔM^C	0.1023	0.1029	0.1120
$ER(\Delta M^C, \Delta M^D)$	1.0630	1.7116	0.7287

All results tables taken from [4].

[1] Gallagher, L., Mackenzie, J., Benham, R., Chen, R.C., Scholer, F., Culpepper, J.S.: RMIT at the NTCIR-13 We Want Web task. In: NTCIR-13 (2017)

[2] Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (2016).

[3] Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005).

[4] Sakai, T., Ferro, N., Soboroff, I., Zeng, Z., Xiao, P., Maistro, M.: Overview of the NTCIR-14 CENTRE task. In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (2019).

[5] Yang, P., Fang, H.: Evaluating the effectiveness of axiomatic approaches in web track. In: Proceedings of The Twenty-Second Text REtrieval Conference (2013).



MAX-PLANCK-GESELLSCHAFT

mpi max planck institut
informatik