MPII at the NTCIR-14 CENTRE Task

Andrew Yates

Max Planck Institute for Informatics, Saarbrücken, Germany ayates@mpi-inf.mpg.de

Abstract. The MPII team participated in the T1, T2TREC, and T2OPEN subtasks of the NTCIR-14 CENTRE Task. This report describes our approaches, the known ways in which our approaches differed from the runs being reproduced, and the success of our reproductions. While our T1 replication and T2TREC reproduction were successful from an overall perspective, the per-topic results were mixed, and our T2OPEN reproduction was inconclusive. We discuss several factors that may have contributed to these outcomes.

Team Name. MPII

Subtasks. T1 T2TREC T2Open

Keywords: replicaton · reproduction · Web search · ad-hoc retrieval

1 Introduction

The MPII team participated in all three subtasks (T1, T2TREC, and T2Open) of the NTCIR-14 CENTRE task. [4] This report describes our approach, including the systems used, assumptions made, and known deviations from the target runs, and discusses the performance of our runs as compared to the target runs. As described in the task overview [4], the success of a replication or reproduction is evaluated by comparing the per-topic performance differences between an **A**dvanced and **B**aseline run between a pair of target runs (i.e., the original runs being reproduced) and a pair of reproduction runs (i.e., the runs produced by the task participant). Given these per-topic differences, RMSE and Pearson's correlation coefficient r are used to evaluate *topicwise reproducibility*: the similarity between the per-topic performances of the target runs and the reproduction runs. An effectiveness ratio (ER) is used to evaluate the *overall reproducibility* of the target runs.

2 Subtask T1: Replicability

As described in the NTCIR-14 CENTRE overview [4], subtask T1 evaluated the extent to which an A run and B run from RMIT's NTCIR-13 WWW submission [1] could be replicated. The A run used a sequential dependency model (SDM) with title, body, and inlink fields, while the B run used a full dependency model (FDM) with only the body field. [3] Both approaches used RM3 query expansion.

We used Indri v5.12 to replicate RMIT's runs, which were originally produced using v5.11.[1] The ClueWeb12-B13 corpus was preprocessed using Indri with the Krovetz stemmer and without stopword removal. We first describe the additional algorithm details obtained from RMIT's overview paper [1] before describing the known ways in which our replication runs differed. For both the A and B runs, documents with a spam score less than 70 were removed and Dirichlet smoothing was used with $\mu = 2000$. For the A run, a weight of 0.20 was given to the title field, 0.05 to the inlink field, and 0.75 to the combination of body fields. For both the A and B runs, the body fields were given a weight of 0.8 for unigram matches, 0.1 for ordered matches, and 0.1 for unordered window matches. Indri RM3 is set to use 10 feedback documents for the A run and 20 for the B run, to add 50 terms for the A run and 10 for the B run, and to give the original query a weight of 0.6 for the A run and 0.8 for the B run.

As shown in Tables 4 and 5 of the CENTRE overview, our replication runs were more successful in terms of overall replicability (as measured by ER) than they were in terms of topicwise replicability (as measured by RMSE and r), with only nERR@10 having a significant topicwise correlation. We are aware of two algorithmic differences between the target and replicated runs, which may be related to our negative topicwise replicability results. First, we considered an unordered window size of 8 for all queries, whereas RMIT's runs increased this window size with the query length. Second, we were unable to successfully index inlinks with Indri and thus did not include the inlinks field.

3 Subtask T2TREC: Reproducibility

Subtask T2TREC considered the question of whether the improvement from a TREC 2013 Web Track B run to A run could be reproduced on the NTCIR-13 WWW-1 test collection. The selected runs were from the University of Delaware's submission [5], in which the F2-LOG axiomatic retrieval function was used with query expansion. The document collection was used for query expansion with the B run, and an external document collection made up of snippets from three Web search engines was used as the A run. We used Anserini's F2LOG implementation with semantic expansion[2] and obtained the semantic expansion's parameter values from UDel's overview paper: $R = 20$, $N = 19$, $M = 20$, and $K = 1000$. The overview paper additionally states that the weight given to expansion terms (β) was 0.1 for the B run and 1.7 for the A run.

We are aware of several assumptions and deviations made in our reproduction runs. In the preprocessing step, we assumed the Porter stemmer was used and that stopwords were removed. We do not build a filtered ClueWeb12 index as described in section 3 of UDel's report [5]. To build the external document collection used with the A run, we collected the top 100 snippets using both Google and Bing, whereas UDel used three unnamed Web search engines. We

[1] https://github.com/rmit-ir/ntcir13-www

[2] https://github.com/castorini/Anserini/commit/10255e0f15c8caca94f8d5376a2c7c9ad1f5b5fd

used ClueWeb12 B13 rather than Category A as the expansion document collection with the B run. Finally, we set $\beta = 0.1$ for the A run after observing that higher values of β degraded performance on the TREC WT14 collection. Given the difference in our external document collection and the task's goal of reproducing the target runs on a new document collection, we opted to deviate from the target A run's β value rather than using the original value that performed worse in our WT14 evaluation.

As shown in Tables 6 and 7 of the CENTRE overview, our reproduction was successful in terms of both topicwise reproducibility and overall reproducibility. In the topicwise case, Table 6 indicates that our A run significantly improves on our B run across all three metrics. Similarly, in terms of overall reproducibility, the effectiveness ratio is close to 1 for Q@10 (1.0893) and reasonably high for nDCG@10 and nERR@10 (0.7767 and 0.6997, respectively).

4 Subtask T2Open: Reproducibility

Subtask T2Open allowed participants to choose their own pair of runs to reproduce on the WWW-1 test collection. We opted to compare two runs from the recent DRMM neural re-ranking model [2], which represents query-document pairs as histograms of embedding similarities. The paper proposes using "logcount-based histograms" (LCH) over normalized histograms (i.e., divided by document length) or histograms based on raw counts (CH). The LCH histograms perform approximately 8% better than the CH histograms in terms of MAP on both the test collections considered, though the improvements in nDCG are more modest (2.4% and 4.6%). The authors hypothesize that "the good performance of LCH-based models indicates that deep neural networks can benefit from input signals with reduced range and nonlinear transformation useful for learning multiplicative relationships." No significance tests were conducted between these run pairs, however. Using DRMM with LCH for our A run and with CH for our B run, we attempted to test whether a similar improvement could be obtained on the WWW-1 test collection and whether this improvement might be significant.

In more detail, the DRMM $LCH \times IDF$ variant served as our A run and the DRMM $CH \times IDF$ variant served as our B run. We implemented DRMM using the MatchZoo implementation as a reference.[3] As described in the DRMM paper [2], we used a hidden layer of size 5 in the matching network. We trained the model using hinge loss on the TREC 2009-2013 Web Track data for 150 iterations consisting of 4096 instances with a batch size of 64. We used the 2014 Web Track (WT14) for validation in order to tune the histogram size. While the original work suggests using histograms of size 30, we found size 20 to perform better with LCH and size 15 to perform better with CH on WT14.

As shown in Table 8 of the CENTRE overview, the A and B run performed similarly in terms of all three metrics, with $p > 0.45$ in all three cases and only a 2.6% improvement in terms of nDCG. Despite the fact that DRMM's authors

[3] https://github.com/NTMC-Community/MatchZoo

saw reasonable MAP improvements on two test collections with LCH, we are unable to support the conclusion that the LCH method is better. While there are many possible reasons for this failure, we note that our experiments on WT14 indicate that different histogram sizes are optimal for the two methods. It may be that the performance difference observed in the DRMM paper is related to the fact that the same fixed histogram size was used for both methods. We leave further investigation of the differences between the two methods for future work.

5 Conclusions

In this report we described MPII's participation in the NTCIR-14 CENTRE task. MPII submited one run to each of the three subtasks. While our overall reproduction results were successful for T1 and T2TREC, we did not observe any statistically significant differences in T2Open. Our per-topic reproduction was only successful for T2TREC, illustrating the difficulty of reproducing per-topic performance differences.

6 Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal GPU used in this research.

References

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. pp. 55–64. CIKM '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2983323.2983769, http://doi.acm.org/10.1145/2983323.2983769
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. pp. 472–479. SIGIR '05, ACM, New York, NY, USA (2005). https://doi.org/10.1145/1076034.1076115, http://doi.acm.org/10.1145/1076034.1076115
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, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013 (2013)