

Towards a Repository for Information Retrieval Runs

Discussion Paper

Elias Bassani¹

¹*Independent*

Abstract

This manuscript discusses our ongoing work on ranxhub, an online repository for sharing pre-computed runs: the ranked lists of documents retrieved for a specific set of queries by a retrieval model. First, we discuss the many advantages and implications that an online repository for sharing runs can bring to the table. Then, we introduce ranxhub and its integration with ranx, a Python library for the evaluation and comparison of Information Retrieval runs, showing its very simple usage.

Keywords

Information Retrieval, Pre-computed Runs, Artifacts Sharing, Evaluation, Comparison, Online Platform

1. Introduction


Offline evaluation and comparison of Information Retrieval systems are fundamental steps in the Information Retrieval research workflow [1, 2]. The introduction of `trec_eval`¹ by the Text Retrieval Conference (TREC) [3] allowed standardizing evaluation measures in Information Retrieval. Recently, the Python library `ranx`² [4, 5] simplified the evaluation and comparison of Information Retrieval runs — the ranked lists of documents retrieved for a specific set of queries by a given system — by providing a user-friendly interface to those functionalities. However, comparing different retrieval models could still require significant efforts, especially in the case of approaches based on Neural Language Models, such as BERT [6]. For example, a checkpoint for a specific baseline model may be missing, reproducibility instructions may not have been shared, or the source code may not be publicly available. In these scenarios, researchers may encounter several issues. First, modern Transformer-based [7] retrieval models require significant hardware resources, which are not always available in academia. Second, the lack of adequate instructions to train/execute a model could significantly slow down comparative evaluation. Finally, when the source code is unavailable, it becomes difficult to reproduce the results of a given research paper because several technical details are generally not included.


To improve this situation, we are building an online platform called `ranxhub`³ for sharing pre-computed runs, allowing for very stress-free comparative evaluations. To our knowledge,

IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

 elias.bssn@gmail.com (E. Bassani)

 0000-0002-0877-7063 (E. Bassani)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹https://github.com/usnistgov/trec_eval

²<https://github.com/AmenRa/ranx>

³<https://amenra.github.io/ranxhub>

there is no other initiative with this specific purpose. To promote an open culture, all the data are available under the very permissive CC BY 4.0 license⁴. In the following, we motivate our work and describe its usage.

2. Motivations

In this section, we introduce the main motivations behind the implementation of `ranxhub`: improving the time effectiveness of research, promoting transparency, and reducing the environmental impact caused by modern Information Retrieval research.

Speed up Research The evaluation and comparison (statistical tests) of new Information Retrieval models w.r.t. the state-of-the-art is an integral part of the research workflow, which often require time-consuming and error-prone activities, such as implementing, training, and executing baseline models to reproduce their results. Across research labs, those activities are usually carried out independently with limited research artifacts sharing, severely affecting the research process time-wise. A public repository providing pre-computed runs and seamlessly integrated into an evaluation library could enable researchers to find appropriate baselines and conduct comparative evaluations in just a few minutes, thus improving the time-effectiveness of their work.

Transparency Transparency is a fundamental principle in Science. Without the openness of research artifacts, it is hard to assess the improvements and findings in a research field. We believe providing tools to support virtuous behavior is as important as promoting research ethics. In this regard, `ranxhub` can provide researchers with a quick and easy way to share the results of their work and demonstrate the trustworthiness of their research papers while gaining visibility. Although sharing well-documented and working source code for training and evaluating a new retrieval model should be the final goal transparency-wise, it also comes with some issues [8] that are not easily solvable (e.g., the availability of hardware resources). In this scenario, `ranxhub` could represent a step forward in the right direction.

Environmental Impact In recent years, Computer Science research has reached a significant environmental impact due to the CO₂ emissions produced by training large Neural Networks [9, 10, 11]. Recent advances in Information Retrieval are tied to a severe increase in CO₂ emissions due to the use of Large Language Models [6, 12] to achieve state-of-the-art results [13]. Since baseline models are trained multiple times across research labs in academia and industry, the overall environmental impact of such activities is much more severe than in the past. However, we could positively influence electricity consumption and pollution by sharing pre-computed runs and relying on them for future research. Specifically, we could train and evaluate once and share the results so that others can benefit from our work with minimal environmental impact.

⁴<https://creativecommons.org/licenses/by/4.0>

3. System Overview

In this section, we overview the main features provided by ranxhub following the platform’s workflow. We first describe the browsing system. Then, we introduce the run cards, the collection of metadata enriching the available runs. Finally, we describe the integration with ranx and how to share a run with the community.

Browsing The browsing process⁵ of ranxhub works as follows: 1) the user chooses a benchmark, such as MSMARCO [14] or the Multi-Domain Benchmark for Personalized Search Evaluation [15], 2) the system shows a table for each test set related to the benchmark (e.g., Dev, TREC DL 2019 [16], TREC DL 2020 [17] for MSMARCO), displaying the available runs, their IDs, and the metric scores used for the specific test set, 3) the user choose a pre-computed run, and 4) the system shows the related run card (described in the following section).

Pre-computed Runs and Run Cards A run comprises the results retrieved by a model for a specific set of queries. Each run is accompanied by a collection of metadata (inspired by ir-metadata [18]) called run card⁶. A run card is organized in four sections: 1) run metadata, 2) model metadata, 3) metric scores, and 4) links to the related resources, such as the model source code, the original paper, and its BibTex. For brevity, we encourage the reader to refer to footnote 6 to get an idea of how a run card is displayed and what metadata it includes.

Integration with ranx We extended ranx, a Python evaluation library, to integrate with ranxhub and allow for downloading pre-computed runs. Thanks to this integration, users can download pre-computed runs with a single function call and perform comparative evaluations very rapidly. Specifically, once the users have chosen the pre-computed baseline run(s), they can easily download and import them with ranx, as exemplified in Listings 1 and 2. In just a few lines of code and a matter of seconds, users can compare multiple runs without the need for implementing, training, and executing the related retrieval models. A working example can be found here⁷. For further details on ranx, please refer to [4, 5].

Sharing To share a pre-computed run with the community, researchers can rely on ranx to pack their runs in the correct format and upload them and their related run cards using our submission form⁸. An empty run card can be found here⁹. Once received, we will upload the run on our server (currently on Amazon AWS S3¹⁰) to make it available to others and assign it a unique identifier. We believe submissions must be moderated to maintain high-quality standards and avoid cluttering. Therefore, we only accept runs related to already published research articles. Runs that replicate published results but not from the authors of the original papers are equally valid. As we do not intend to be a monopoly, all the available runs are

⁵<https://amenra.github.io/ranxhub/browse>

⁶Run card example: <https://amenra.github.io/ranxhub/browse/amdbfpse/cs/bm25>

⁷<https://tinyurl.com/yc639v4y>

⁸<https://forms.gle/fK6wLS83yZeoS1mL8>

⁹<https://github.com/AmenRa/ranxhub/blob/main/files/runcard-empty.yaml>

¹⁰<https://aws.amazon.com/s3/>

```

1  from ranx import Qrels, Run, compare
2
3  # Load qrels and your run
4  qrels = Qrels("path/to/qrels")
5  my_run = Run("path/to/run", name="my_run")
6
7  # Download pre-computed runs from ranxhub
8  bm25_run = Run.from_ranxhub("bm25-run-id")
9  bert_run = Run.from_ranxhub("bert-run-id")
10
11 compare(
12     qrels=qrels,
13     runs=[bm25_run, bert_run, my_run],
14     metrics=["map@100", "mrr@100", "ndcg@10"],
15 )

```

Listing 1: ranx integration with ranxhub.

#	Model	MAP@100	MRR@100	NDCG@10
a	bm25	0.233	0.234	0.239
b	bert	0.366 ^a	0.367 ^a	0.408 ^a
c	my_run	0.405 ^{ab}	0.406 ^{ab}	0.451 ^{ab}

Listing 2: Output of ranx compare method.

exportable using ranx to JSON and TREC-style files. Moreover, as ranx is open source, other libraries can copy-paste part of its code to download data from ranxhub.

4. Conclusion

In this manuscript, we discussed our ongoing efforts to provide the community with an online repository for sharing pre-computed retrieval runs, called ranxhub, and the underlying motivations. Specifically, ranxhub could speed up research by avoiding time-consuming and error-prone activities such as implementing, training, and executing baseline models. Moreover, it could promote virtuous behavior and transparency and reduce the environmental impact of modern Information Retrieval research. We described ranxhub’s browsing system, the run cards, and the integration with ranx, a Python library for Information Retrieval evaluation. By leveraging this integration, users can compare the results of multiple systems in just a few lines of code. To conclude, we believe ranxhub could positively impact Information Retrieval research. However, its success can only be determined by a community effort and the will to pursue transparency and improve the research experience of others.

References

- [1] D. Harman, *Information Retrieval Evaluation, Synthesis Lectures on Information Concepts, Retrieval, and Services*, Morgan & Claypool Publishers, 2011.
- [2] M. Sanderson, Test collection based evaluation of information retrieval systems, *Found. Trends Inf. Retr.* 4 (2010) 247–375.
- [3] E. Voorhees, D. Harman, *Experiment and evaluation in information retrieval*, 2005.
- [4] E. Bassani, ranx: A blazing-fast python library for ranking evaluation and comparison, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørnvåg, V. Setty (Eds.), *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 259–264. URL: https://doi.org/10.1007/978-3-030-99739-7_30. doi:10.1007/978-3-030-99739-7_30.
- [5] E. Bassani, L. Romelli, ranx.fuse: A python library for metasearch, in: M. A. Hasan, L. Xiong (Eds.), *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022*, pp. 4808–4812. URL: <https://doi.org/10.1145/3511808.3557207>. doi:10.1145/3511808.3557207.
- [6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, and Short Papers*, Association for Computational Linguistics, 2019. doi:10.18653/v1/n19-1423.
- [7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface’s transformers: State-of-the-art natural language processing, *CoRR abs/1910.03771* (2019). URL: <http://arxiv.org/abs/1910.03771>. arXiv:1910.03771.
- [8] E. M. Voorhees, S. Rajput, I. Soboroff, Promoting repeatability through open runs, in: E. Yilmaz, C. L. A. Clarke (Eds.), *Proceedings of the Seventh International Workshop on Evaluating Information Access, EVIA 2016, a Satellite Workshop of the NTCIR-12 Conference, National Center of Sciences, Tokyo, Japan, June 7, 2016, National Institute of Informatics (NII), 2016*. URL: <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/evia/04-EVIA2016-VoorheesE.pdf>.
- [9] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Association for Computational Linguistics, 2019, pp. 3645–3650. URL: <https://doi.org/10.18653/v1/p19-1355>. doi:10.18653/v1/p19-1355.
- [10] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for modern deep learning research, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020*, pp. 13693–13696. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/7123>.
- [11] D. A. Patterson, J. Gonzalez, Q. V. Le, C. Liang, L. Munguia, D. Rothchild, D. R. So, M. Texier,

- J. Dean, Carbon emissions and large neural network training, CoRR abs/2104.10350 (2021). URL: <https://arxiv.org/abs/2104.10350>. arXiv:2104.10350.
- [12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (2020) 140:1–140:67. URL: <http://jmlr.org/papers/v21/20-074.html>.
- [13] H. Scells, S. Zhuang, G. Zuccon, Reduce, reuse, recycle: Green information retrieval research, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2825–2837. URL: <https://doi.org/10.1145/3477495.3531766>. doi:10.1145/3477495.3531766.
- [14] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, Ms marco: A human generated machine reading comprehension dataset, 2016. URL: <https://arxiv.org/abs/1611.09268>. doi:10.48550/ARXIV.1611.09268.
- [15] E. Bassani, P. Kasela, A. Raganato, G. Pasi, A multi-domain benchmark for personalized search evaluation, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 3822–3827. URL: <https://doi.org/10.1145/3511808.3557536>. doi:10.1145/3511808.3557536.
- [16] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019 deep learning track, CoRR abs/2003.07820 (2020). URL: <https://arxiv.org/abs/2003.07820>. arXiv:2003.07820.
- [17] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 deep learning track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020, volume 1266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2020. URL: <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf>.
- [18] T. Breuer, J. Keller, P. Schaer, ir_metadata: An extensible metadata schema for IR experiments, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3078–3089. URL: <https://doi.org/10.1145/3477495.3531738>. doi:10.1145/3477495.3531738.