

NLP-Power 2022

The First Workshop on Efficient Benchmarking in NLP

Proceedings of the Workshop

May 26, 2022

The NLP-Power organizers gratefully acknowledge the support from the following sponsors.

In cooperation with



©2022 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-47-6

Introduction

NLP Power! is the workshop on efficient benchmarking in NLP.

Benchmarking has become a standard practice for evaluating upcoming models against one another and human solvers; there are still many unresolved issues and methodological concerns. The main idea of the workshop is to bring together researchers that work on benchmarks for natural language processing (NLP) and discuss how benchmarking can be improved to account for computational efficiency, ethical considerations, user preferences, and out-of-domain robustness. The workshop proceedings present the collection of research contributions on the computational efficiency of model evaluation, transfer learning efficiency estimation, evaluation metrics, robustness and bias assessment, and general best practices in benchmarking for NLP.

This is the first time we have organized a workshop with this particular scope of interest. Our workshop is hosted by the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). Our program committee consisted of experts from all over the world with years of research experience in the industry and academia. The committee worked hard on every submission and selected 12 research papers to be presented at the workshop in the poster and oral sessions. The workshop program also included one ACL Findings paper. Overall, it resulted in 2 oral presentation sessions, which were intermitted by a poster session, three invited talks, and a round table on the problems of canonic benchmark standards.

NLP Power would not be possible without the dedicated intellectual work of the program committee: their peer review and efforts aimed to improve the work have shaped the scientific community, which is now, for the first time, coming forward with a unified workshop mission. We also express our sincere gratitude to the invited speakers: Anna Rumshiski, He He, and Ulises Mejias, for their contribution to the program. We thank the researchers and NLP practitioners for the engagement and responses and hope to continue to provide a platform for fruitful discussions on various topics, ranging from rethinking benchmarking methods to the reproducibility of the leaderboard results.

You can find more details about the workshop on the website: <http://nlp-power.github.io/>.

Tatiana Shavrina, Valentin Malykh, Ekaterina Artemova, Vladislav Mikhailov, Laura Weidinger, Oleg Serikov, and Vitaly Protasov

Organizing Committee

Program Chairs

Tatiana Shavrina, AIRI, SberDevices

Valentin Malykh, Huawei

Ekaterina Artemova, HSE University, Huawei

Vladislav Mikhailov, SberDevices, HSE University

Oleg Serikov, AIRI, HSE University

Vitaly Protasov, AIRI

Program Committee

Senior Program Committee

Jürgen Schmidhuber, Swiss AI Lab IDSIA, USI, SUPSI

Program Committee

Laura Weidinger, DeepMind

Leonid Zhukov, AIRI

Mikhail Burtsev, AIRI

Nitish Hemant Joshi, CILVR / ML2

Richard Yuanzhe Pang, CILVR / ML2

Adaku Uchendu, Penn State University

Ilya Kuznetsov, TU Darmstadt

Anastasia Bonch-Osmolovskaya, HSE University

Andrey Kravchenko, Oxford University

Daniel Karabekyan, HSE University

Preslav Nakov, QCRI

Suresh Manandhar, Wiseyak, USA

Piotr Piękos, DeepMind

Olga Lyashevskaya, Vinogradov IRL RAS, HSE University

Arjun Akula, Google Research

Secondary Reviewers

Tatiana Shavrina, AIRI, SberDevices

Maria Tikhonova, HSE University, SberDevices

Dina Pisarevskaya, QMUL

Invited Speakers

Ulises A. Mejias, SUNY Oswego

Anna Rumshisky, UMASS, Amazon

He He, CILVR / ML2

Table of Contents

<i>Raison d’être of the benchmark dataset: A Survey of Current Practices of Benchmark Dataset Sharing Platforms</i>	
Jaihyun Park and Sullam Jeoung	1
<i>Towards Stronger Adversarial Baselines Through Human-AI Collaboration</i>	
Wencong You and Daniel Lowd	11
<i>Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model</i>	
Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim and Adam Dunn	22
<i>Why only Micro-F1? Class Weighting of Measures for Relation Classification</i>	
David Harbecke, Yuxuan Chen, Leonhard Hennig and Christoph Alt	32
<i>Automatically Discarding Straplines to Improve Data Quality for Abstractive News Summarization</i>	
Amr Keleg, Matthias Lindemann, Danyang Liu, Wanqiu Long and Bonnie L. Webber	42
<i>A global analysis of metrics used for measuring performance in natural language processing</i>	
Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott and Matthias Samwald	52
<i>Beyond Static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages</i>	
Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram and Monojit Choudhury	64
<i>Checking HateCheck: a cross-functional analysis of behaviour-aware learning for hate speech detection</i>	
Pedro Henrique Luz de Araujo and Benjamin Roth	75
<i>Language Invariant Properties in Natural Language Processing</i>	
Federico Bianchi, Debora Nozza and Dirk Hovy	84
<i>DACT-BERT: Differentiable Adaptive Computation Time for an Efficient BERT Inference</i>	
Cristobal Eyzaguirre, Felipe del Rio, Vladimir Araujo and Alvaro Soto	93
<i>Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection</i>	
Giuseppe Attanasio, Debora Nozza, Eliana Pastor and Dirk Hovy	100
<i>Characterizing the Efficiency vs. Accuracy Trade-off for Long-Context NLP Models</i>	
Phyllis Ang, Bhuwan Dhingra and Lisa Wu Wills	113