Insights 2022

The Third Workshop on Insights from Negative Results in NLP

Proceedings of the Workshop

May 26, 2022

The Insights organizers gratefully acknowledge the support from the following sponsors.

Silver

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 978-1-955917-40-7

Introduction

Publication of negative results is difficult in most fields, and the current focus on benchmark-driven performance improvement exacerbates this situation and implicitly discourages hypothesis-driven research. As a result, the development of NLP models often devolves into a product of tinkering and tweaking, rather than science. Furthermore, it increases the time, effort, and carbon emissions spent on developing and tuning models, as the researchers have little opportunity to learn from what has already been tried and failed.

Historically, this tendency is hard to combat. ACL 2010 invited negative results as a special type of research paper submissions[1], but received too few submissions and did not continue with it. *The Journal for Interesting Negative Results in NLP and ML*[2] has only produced one issue in 2008.

However, the tide may be turning. Despite the pandemic, the third iteration of the *Workshop on Insights from Negative Results* attracted 43 submissions and 1 from ACL Rolling Reviews.

The workshop maintained roughly the same focus, welcoming many kinds of negative results with the hope that they could yield useful insights and provide a much-needed reality check on the successes of deep learning models in NLP. In particular, we solicited the following types of contributions:

- broadly applicable recommendations for training/fine-tuning, especially if X that didn't work is something that many practitioners would think reasonable to try, and if the demonstration of X's failure is accompanied by some explanation/hypothesis;
- ablation studies of components in previously proposed models, showing that their contributions are different from what was initially reported;
- datasets or probing tasks showing that previous approaches do not generalize to other domains or language phenomena;
- trivial baselines that work suspiciously well for a given task/dataset;
- cross-lingual studies showing that a technique X is only successful for a certain language or language family;
- experiments on (in)stability of the previously published results due to hardware, random initializations, preprocessing pipeline components, etc;
- theoretical arguments and/or proofs for why X should not be expected to work.

In terms of topics/themes, 16 papers from our accepted proceedings discussed "lessons learned in pretraining/training neural architectures/large language models"; 10 discussed "great ideas that didn't work"; 10 papers performed probing tasks and datasets to draw deeper insights or understand reasons for success/failure; 9 dealt with issues of robustness, generalizability, compositionality, and few-shot performance; 2 were on the topic of "analyzing biases, errors, spurious correlations in data/model"; 1 paper focused on issues in replication of research results and 1 paper on the impact of data augmentation. Some submissions fit in more than one category.

We accepted 24 short papers (55.8% acceptance rate) and one paper from ACL Rolling Reviews.

We hope the workshop will continue to contribute to the many reality-check discussions on progress in NLP. If we do not talk about things that do not work, it is harder to see what the biggest problems are and where the community effort is the most needed.

[1] https://mirror.aclweb.org/acl2010/papers.html

[2] http://jinr.site.uottawa.ca/

Organizing Committee

Organizers

Shabnam Tafreshi, University of Maryland: ARLIS, USA
João Sedoc, New York University, USA
Anna Rogers, University of Copenhagen, Denmark
Aleksandr Drozd, RIKEN, Japan
Arjun Reddy Akula, Google AI, USA
Anna Rumshisky, University of Massachusetts Lowell / Amazon Alexa, USA

Program Committee

Program Committee

Ali Seyfi, George Washington University
Alicia Sagae, Amazon
Anil Kumar Nelakanti, Amazon
Arijit Adhikari, Amazon
Ashutosh Modi, IIT Kanpur
Chanjun Park, Korea University
Chen-Yu Lee, Google
Constantine Lignos, Brandeis University
Daniel Cer, Google
Deepika Jindal, Amazon
Djamé Seddah, University Paris-Sorbonne
Edison Marrese-Taylor, National Institute of Advanced Industrial Science and Technology (AIST)
Efsun Kayi, SiteRx
Ekaterina Vylomova, University of Melbourne
Ellie Pavlick, Brown University
Emiel Krahmer, Tilburg University
Emil Vatai, RIKEN
Huda Khayrallah, Microsoft
Machel Reid, The University of Tokyo
Indraneil Paul, Amazon
Jessica Ouyang, University of Texas at Dallas
Joel Mackenzie, University of Queensland
John P. Lalor, University of Notre Dame
Jordan Rodu, University of Virginia, Charlottesville
Kyle Lo, Allen Institute for Artificial Intelligence
Lingjia Deng, Bloomberg
Mahesh Goud Tandarpally, Amazon
Marco Basaldella, Amazon
Maximilian Spliethöver, Universität Paderborn
Michael Gamon, Microsoft Research
Montse Cuadros Oller, Vicomtech
Nada Almarwani, Taibah University
Neha Nayak Kennard, University of Massachusetts, Amherst
Olha Kaminska, Universiteit Gent
Pedro Rodriguez, Facebook
Phu Mon Htut, New York University
Prasanna Parasurama, New York University
Qingqing Cao, University of Washington, Seattle
Raphael Shu, RIKEN
Salvatore Giorgi, University of Pennsylvania
Sawsan Alqahtani, Princess Nourah Bint Abdulrahman University
Shubham Chatterjee, University of New Hampshire, Durham
Sotiris Lamprinidis, Corti
Sven Buechel, Friedrich-Schiller-Universität Jena
Tristan Naumann, Microsoft Research
Udita Patel, Amazon

Valentin Barriere, Joint Research Center
Wasi Uddin Ahmad, Amazon
Wazir Ali, ILMA University Karachi
Xutan Peng, University of Sheffield
Yash Parag Butala, Indian Institute of Technology Kharagpur
Yev V Perevodchikov, Amazon

Invited Speakers

Barbara Plank, IT University of Copenhagen
Tal Linzen, New York University

Keynote Talk: Power, Uncertainty and the Null

Tal Linzen
IT University of Copenhagen, Denmark

Bio: Tal Linzen is an Assistant Professor of Linguistics and Data Science at New York University and a Research Scientist at Google. Before moving to NYU in 2020, he was a faculty member at Johns Hopkins University, a postdoctoral researcher at the École Normale Supérieure in Paris, and a PhD student at NYU. At NYU, Tal directs the Computational Psycholinguistics Lab, which develops computational models of human language comprehension and acquisition, as well as methods for interpreting and evaluating neural network models for language technologies.

Keynote Talk: Off the Beaten Track: To Turn "Failures" into Signal and Insights

Barbara Plank
IT University of Copenhagen, Denmark

Bio: Barbara Plank is Chair (Professor) of AI and Computational Linguistics at LMU Munich, with a part-time affiliation at the IT University of Copenhagen. Her research focuses on various aspects of NLP and include learning under sample selection bias (domain adaptation, transfer learning), annotation bias (human disagreements and human uncertainty), learning from beyond the text, and in general learning under limited supervision. Barbara is the recipient of a 2019 Sapere Aude Research Leader grant and an Amazon Research Award. Barbara is on the advisory board of the European Association for Computational Linguistics, publicity director of the Association for Computational Linguistics and since 2022 president of the Northern European Association for Language Technology.

Table of Contents

Program

Thursday, May 26, 2022

08:45 - 09:00 *Opening Remarks*

09:00 - 10:00 *Invited Talk: Barbara Plank*

10:30 - 11:00 *Coffee Break*

11:00 - 11:30 *Thematic Session 1: Linguistically Informed Analysis*

Do Dependency Relations Help in the Task of Stance Detection?
Alessandra Teresa Cignarella, Cristina Bosco and Paolo Rosso

BPE beyond Word Boundary: How NOT to use Multi Word Expressions in Neural Machine Translation
Dipesh Kumar and Avijit Thawani

Challenges in including extra-linguistic context in pre-trained language models
Ionut Teodor Sorodoc, Laura Aina and Gemma Boleda

11:30 - 12:00 *Thematic Session 2: Transformers*

How Much Do Modifications to Transformer Language Models Affect Their Ability to Learn Linguistic Knowledge?
Simeng Sun, Brian Dillon and Mohit Iyyer

Pathologies of Pre-trained Language Models in Few-shot Fine-tuning
Hanjie Chen, Guoqing Zheng, Ahmed Hassan Awadallah and Yangfeng Ji

On Isotropy Calibration of Transformer Models
Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide and Roger Wattenhofer

12:00 - 12:30 *Thematic Session 3: Towards Better Data*

Do Data-based Curricula Work?
Maxim K. Surkov, Vladislav D. Mosin and Ivan P. Yamshchikov

Clustering Examples in Multi-Dataset Benchmarks with Item Response Theory
Pedro Rodriguez, Phu Mon Htut, John P. Lalor and João Sedoc

Thursday, May 26, 2022 (continued)

On the Impact of Data Augmentation on Downstream Performance in Natural Language Processing
Itsuki Okimura, Machel Reid, Makoto Kawano and Yutaka Matsuo

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Panel Discussion*

15:00 - 15:30 *Coffee Break*

10:00 - 10:30 *Thematic Session 4: Improving Evaluation Practices*

Replicability under Near-Perfect Conditions – A Case-Study from Automatic Summarization
Margot Mieskes

On the Limits of Evaluating Embodied Agent Model Generalization Using Validation Sets
Hyounghun Kim, Aishwarya Padmakumar, Di Jin, Mohit Bansal and Dilek Hakkani-Tur

16:00 - 17:00 *Invited Talk: Tal Linzen*

17:00 - 18:00 *Poster Session*

18:00 - 18:10 *Closing Remarks*