Insights 2024

The 5th Workshop on Insights from Negative Results in NLP

Proceedings of the Workshop

June 20, 2024

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)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-102-5

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.

The mission of the workshop on Insights from Negative Results in NLP is to provide a venue for 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 solicit 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;
- demonstration of issues with under-reporting of training details of pre-trained models, including test data contamination and invalid comparisons.

The fifth iteration of the *Workshop on Insights from Negative Results* attracted 28 submissions and 4 from ACL Rolling Reviews. In terms of topics/themes, 4 papers from our accepted proceedings discussed "zero-shot / few-shot learning / low-resource settings"; 1 discussed "cross-modal fine-tuning"; 6 papers examined pre-trained representations / generalization; 1 dealt with tokenization; 6 on the topic of "LLM Reasoning / Alignment / Evaluations / Probing"; 1 on Multi-task Learning. Some submissions fit in more than one category.

We accepted 19 short papers (57.5% acceptance rate).

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.

Organizing Committee

Organizers

Shabnam Tafreshi, AI inQbator at Evernorth Healthcare & UMD
Arjun Reddy Akula, Google DeepMind, USA
João Sedoc, New York University, USA
Anna Rogers, IT University of Copenhagen, Denmark
Aleksandr Drozd, RIKEN, Japan
Anna Rumshisky, University of Massachusetts Lowell / Amazon Alexa, USA

Program Committee

Chairs

Shabnam Tafreshi, AI inQbator at Evernorth Healthcare and UMD
Arjun Akula, Google DeepMind
João Sedoc, New York University
Anna Rogers, IT University of Copenhagen
Aleksandr Drozd, RIKEN Center for Computational Science
Anna Rumshisky, University of Massachusetts Lowell

Program Committee

Wazir Ali, University of Turku
Nihal Balani, Google
Adrian Benton, Google
Shaun Cassini, University of Sheffield
Chung-Chi Chen, National Institute of Advanced Industrial Science and Technology
Young Min Cho, University of Pennsylvania
Tamás Ficsor, University of Szeged
Salvatore Giorgi, University of Pennsylvania
Edward G o w - S m i t h, University of Sheffield
Kazuma Hashimoto, Google Research
Shreya Havaldar, University of Pennsylvania
Marzena Karpinska, University of Massachusetts Amherst
Neha Nayak Kennard, University of Massachusetts Amherst
Anuj Khare, Google LLC
Huda Khayrallah, Microsoft
Saranya Krishnamoorthy, Evernorth Health Services
Gaurav Kumar, Google
Seolhwa Lee, Technical University of Darmstadt
Yifei Li, University of Pennsylvania
Ashutosh Modi, Indian Institute of Technology Kanpur
Tristan Naumann, Microsoft Research
Juan Navarro Horniacek, Google
John E. Ortega, Northeastern University
Chanjun Park, Upstage
Giovanni Puccetti, Scuola Normale Superiore di Pisa
Jitesh Punjabi, Google LLC
Sunny Rai, University of Pennsylvania
Jordan Rodu, University of Virginia
Ayush Singh, Evernorth Health Services Inc.
Maximilian Spliethöver, Leibniz University Hannover
Mahesh Goud Tandarpally, Amazon
Emil Vatai, Riken R-CCS
Shubham Vatsal, New York University

Table of Contents

Program

Tuesday, June 20, 2023

09:15 - 09:30 *Opening Remarks*

09:30 - 10:30 *Oral Session 1*

An Analysis of BPE Vocabulary Trimming in Neural Machine Translation
Marco Cognetta, Tatsuya Hiraoka, Rico Sennrich, Yuval Pinter and Naoaki Okazaki

Pointer-Generator Networks for Low-Resource Machine Translation: Don't Copy That!
Niyati Bafna, Philipp Koehn and David Yarowsky

On the Limits of Multi-modal Meta-Learning with Auxiliary Task Modulation Using Conditional Batch Normalization
Jordi Armengol - Estape, Vincent Michalski, Ramnath Kumar, Pierre - Luc St-Charles, Doina Precup and Samira Ebrahimi Kahou

WINOVIZ: Probing Visual Properties of Objects Under Different States
Woojeong Jin, Tejas Srinivasan, Jesse Thomason and Xiang Ren

10:30 - 11:00 *Coffee*

11:00 - 11:45 *Invited Talk: Marius Mosbach, Analysis Work in NLP: The Good, the Bad and the Ugly*

11:45 - 12:30 *Oral Session 2*

What explains the success of cross-modal fine-tuning with ORCA?
Paloma Garcia De Herreros, Vagrant Gautam, Philipp Slusallek, Dietrich Klakow and Marius Mosbach

I Have an Attention Bridge to Sell You: Generalization Capabilities of Modular Translation Architectures
Timothee Mickus, Raul Vazquez and Joseph Attieh

Knowledge Distillation vs. Pretraining from Scratch under a Fixed (Computation) Budget
Minh Duc Bui, Fabian Schmidt, Goran Glavaš and Katharina Von Der Wense

12:30 - 14:00 *Lunch*

Tuesday, June 20, 2023 (continued)

14:00 - 14:45 *Oral Session 3*

The Paradox of Preference: A Study on LLM Alignment Algorithms and Data Acquisition Methods
Rishikesh Devanathan, Varun Nathan and Ayush Kumar

Can probing classifiers reveal the learning by contact center large language models?: No, it doesn't!
Varun Nathan, Ayush Kumar and Digvijay Ingle

Multi-Task Learning with Adapters for Plausibility Prediction: Bridging the Gap or Falling into the Trenches?
Annerose Eichel and Sabine Schulte Im Walde

14:45 - 15:30 *Invited Talk: Sasha Luccioni, Reproducibility in ML and the Environment: What's the Connection?*

15:30 - 16:00 *Coffee*

16:00 - 17:00 *Poster Session*

17:00 - 17:10 *Closing Remarks*