NAACL HLT 2018

Generalization in the Age of Deep Learning

Proceedings of the Workshop

June 5, 2018
New Orleans, Louisiana

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 978-1-948087-16-2

Deep learning has brought a wealth of state-of-the-art results and new capabilities. Although methods have achieved near human-level performance on many benchmarks, numerous recent studies imply that these benchmarks only weakly test their intended purpose, and that simple examples produced either by human or machine, cause systems to fail spectacularly. For example, a recently released textual entailment demo was criticized on social media for predicting that "John killed Mary" entails "Mary killed John" with 92% confidence. Such surprising failures combined with the inability to interpret state-of-the-art models have eroded confidence in our systems, and while these systems are not perfect, the real flaw lies with our benchmarks that do not adequately measure a model's ability to generalize, and are thus easily gameable.

This workshop provides a venue for exploring new approaches for measuring and enforcing generalization in models. We have solicited work in the following areas:

1. Analysis of existing models and their failings.
2. Creation of new evaluation paradigms, e.g. zero-shot learning, Winnograd schema, and datasets that avoid explicit types of gamification.
3. Modeling advances such as regularization, compositionality, interpretability, inductive bias, multi-task learning, and other methods that promote generalization.

Our goals are similar in spirit to those of the recent "Build it Break it" shared task. However, we propose going beyond identifying areas of weakness (i.e. "breaking" existing systems), and discussing evaluations that rigorously test generalization as well as modeling techniques for enforcing it.

We received eight archival submissions and seven cross submission, accepting five archival papers and all cross submission. Predominately papers covered the first two stated goals of workshop, with the majority identifying flaws in either methods or data. Of the papers proposing new evaluations, many explored using synthetic data. The papers will be presented as posters at the workshop and we are excited to see what discussions they generate. In addition to twelve papers that will be presented we are equally excited for talks from Sam Bowman, Yejin Choi, Percy Liang, Ndapa Nakashole, Devi Parikh, and Dan Roth. Finally , we would also like to thank Yejin, Devi and Dan for helping through service on the steering committee.

– Yonatan, Omer, Mark

Organizers:

Yonatan Bisk	University of Washington
Omer Levy	University of Washington
Mark Yatskar	University of Washington

Steering:

Yejin Choi	University of Washington
Dan Roth	University of Pennsylvania
Devi Parikh	Georgia Tech / Facebook AI Research

Program Committee:

Jacob Andreas	UC Berkeley
Antoine Bosselut	U Washington
Kai-Wei Chang	UCLA
Christos Christodoulopoulos	Amazon, Inc
Greg Durrett	UT Austin
Maxwell Forbes	U Washington
Spandana Gella	Edinburgh U
Luheng He	U Washington
Srinivasan Iyer	U Washington
Mohit Iyyer	UMass Amherst
Robin Jia	Stanford U
Ioannis Konstas	Heriot-Watt U
Jonathan Kummerfeld	U Michigan
Alice Lai	UIUC
Mike Lewis	FAIR
Tal Linzen	JHU
Vicente Ordonez	U Virginia
Siva Reddy	Stanford U
Alan Ritter	Ohio State U
Rajhans Samdani	Spoke
Sameer Singh	UC Irvine
Alane Suhr	Cornell U
Chen-Tse Tsai	U Pennsylvania
Shyam Upadhyay	U Pennsylvania
Andreas Vlachos	U Sheffield

Table of Contents

Towards Inference-Oriented Reading Comprehension: ParallelQA
Soumya Wadhwa, Varsha Embar, Matthias Grabmair and Eric Nyberg 1

Commonsense mining as knowledge base completion? A study on the impact of novelty
Stanislaw Jastrzebski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio and Jackie Cheung 8

Deep learning evaluation using deep linguistic processing
Alexander Kuhnle and Ann Copestake 17

The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models
Noah Weber, Leena Shekhar and Niranjan Balasubramanian 24

Extrapolation in NLP
Jeff Mitchell, Pontus Stenetorp, Pasquale Minervini and Sebastian Riedel 28

Conference Program

June 5th

9:00–9:15 **Welcome**

9:15–9:50 **Yejin Choi**

9:50–10:25 **Dan Roth**

10:25–10:35 **Break**

10:35–11:10 **Percy Liang**

11:10–11:45 **Ndapa Nakashole**

11:45–12:20 **Hal Daume III**

12:20–13:30 **Lunch**

13:30–14:30 **Poster Session**

Towards Inference-Oriented Reading Comprehension: ParallelQA
Soumya Wadhwa, Varsha Embar, Matthias Grabmair and Eric Nyberg

Commonsense mining as knowledge base completion? A study on the impact of novelty
Stanislaw Jastrzebski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio and Jackie Cheung

Deep learning evaluation using deep linguistic processing
Alexander Kuhnle and Ann Copestake

The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models
Noah Weber, Leena Shekhar and Niranjan Balasubramanian

June 5th (continued)

Extrapolation in NLP
Jeff Mitchell, Pontus Stenetorp, Pasquale Minervini and Sebastian Riedel

14:45–15:20 Sam Bowman

15:20–15:55 Devi Parikh

15:55–16:10 Break

16:10–17:10 Panel

17:10–17:15 Closing