DADC 2022

The First Workshop on Dynamic Adversarial Data Collection (DADC)

Proceedings of the Workshop

July 14, 2022

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

Platinum

ML
● Commons

Gold

Meta

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 978-1-955917-94-0

Preface

This volume contains papers from the First Workshop on Dynamic Adversarial Data Collection (DADC), held at NAACL 2022.

Dynamic Adversarial Data Collection (DADC) has been gaining traction in the community as a promising approach to improving data collection practices, model evaluation and performance. DADC allows us to collect human-written data dynamically with models in the loop. Humans can be tasked with finding adversarial examples that fool current state-of-the-art models (SOTA), for example, or they can cooperate with models to find interesting examples. This offers two benefits: it allows us to gauge how good contemporary SOTA methods really are; and it yields data that may be used to train even stronger models by specifically targeting their current weaknesses.

The first workshop on DADC and corresponding shared task focus on three currently under-explored themes: i) understanding how humans can be incentivized to creatively identify and target model weaknesses to increase their chances of fooling the model; ii) how humans and machines can cooperate to produce the most useful data; and iii) how the interaction between humans and machines can further drive performance improvements, both from the perspectives of traditional evaluation metrics as well as those of robustness and fairness.

Organizing Committee

General Chairs

Max Bartolo, University College London
Hannah Rose Kirk, University of Oxford
Pedro Rodriguez, FAIR Labs, Seattle
Katerina Margatina, University of Sheffield
Tristan Thrush, Hugging Face
Robin Jia, University of Southern California

Advisory Committee

Pontus Stenetorp, University College London
Adina Williams, FAIR, NYC
Douwe Kiela, Hugging Face

Program Committee

Program Committee

Giorgos Vernikos, EPFL & HEIG-VD
John P. Lalor, University of Notre Dame
Maharshi Gor, University of Maryland, College Park
Pasquale Minervini, University College London
Paul Rottger, University of Oxford
Shi Feng, University of Maryland, College Park
Unso Eun Seo Jo, Stanford University & HuggingFace

Invited Speakers

Anna Rogers, University of Copenhagen
Sam Bowman, New York University
Jordan Boyd-Graber, University of Maryland
Lora Aroyo, Google
Tongshuang Wu, Carnegie Mellon University

Keynote Talk: What kinds of questions have we been asking? A taxonomy for QA/RC benchmarks

Anna Rogers
University of Copenhagen

Abstract: This talk provides an overview of the current landscape of resources for Question Answering and Reading comprehension, highlighting the current lacunae for future work. I will also present a new taxonomy of “skills” targeted by QA/RC datasets and discuss various ways in which questions may be unanswerable.

Bio: Anna Rogers is an Assistant Professor in the Center for Social Data Science at the University of Copenhagen. She is also a visiting researcher with the RIKEN Center for Computational Science. Anna's main research area is Natural Language Processing, in particular model analysis and evaluation of natural language understanding systems.

Keynote Talk: Why Adversarially-Collected Test Sets Don't Work as Benchmarks

Sam Bowman
New York University

Abstract: Dynamic and/or adversarial data collection can be quite useful as a way of collecting training data for machine-learning models, identifying the conditions under which these models fail, and conducting online head-to-head comparisons between models. However, it is essentially impossible to use these practices to build usable static benchmark datasets for use in evaluating or comparing future new models. I defend this point using a mix of conceptual and empirical points, focusing on the claims (i) that adversarial data collection can skew the distribution of phenomena such as to make it unrepresentative of the intended task, and (ii) that adversarial data collection can arbitrarily shift the rankings of models on its resulting test sets to disfavor systems that are qualitatively similar to the current state of the art.

Bio: Sam Bowman is an Assistant Professor at New York University and a Visiting Researcher (Sabbatical) at Anthropic. His research interests include the study of artificial neural network models for natural language understanding, with a focus on building high-quality training and evaluation data, applying these models to scientific questions in syntax and semantics, and contributing to work on language model alignment and control.

Keynote Talk: Incentives for Experts to Create Adversarial QA and Fact-Checking Examples

Jordan Boyd-Graber
University of Maryland

Abstract: I'll discuss two examples of our work putting experienced writers in front of a retrieval-driven adversarial authoring system: question writing and fact-checking. For question answering, we develop a retrieval-based adversarial authoring platform and create incentives to get people to use our system in the first place, write interesting questions humans can answer, and challenge a QA system. While the best humans lose to computer QA systems on normal questions, computers struggle to answer our adversarial questions. We then turn to fact checking, creating a new game (Fool Me Twice) to solicit difficult-to-verify claims—that can be either true or false—and to test how difficult the claims are both for humans and computers. We argue that the focus on retrieval is important for knowledge-based adversarial examples because it highlights diverse information, prevents frustration in authors, and takes advantage of users' expertise.

Bio: Jordan Boyd-Graber is an Associate Professor in the University of Maryland Computer Science Department (tenure home), Institute of Advanced Computer Studies, iSchool, and Language Science Center. Previously, he was an Assistant Professor at Colorado's Department of Computer Science (tenure granted in 2017). Jordan's research focuses on making machine learning more useful, more interpretable, and able to learn and interact from humans.

Keynote Talk: Data Excellence: Better Data for Better AI

Lora Aroyo
Google

Abstract: The efficacy of machine learning (ML) models depends on both algorithms and data. Training data defines what we want our models to learn, and testing data provides the means by which their empirical progress is measured. Benchmark datasets define the entire world within which models exist and operate, yet research continues to focus on critiquing and improving the algorithmic aspect of the models rather than critiquing and improving the data with which our models operate. If “data is the new oil,” we are still missing work on the refineries by which the data itself could be optimized for more effective use. In this talk, I will discuss data excellence and lessons learned from software engineering to achieve the scare and rigor in assessing data quality.

Bio: Lora Aroyo is Research Scientist at Google Research, NYC, where she works on research for Data Excellence by specifically focussing on metrics and strategies to measure quality of human-labeled data in a reliable and transparent way. Lora is an active member of the Human Computation, User Modeling and Semantic Web communities. She is president of the User Modeling community UM Inc, which serves as a steering committee for the ACM Conference Series “User Modeling, Adaptation and Personalization” (UMAP) sponsored by SIGCHI and SIGWEB. She is also a member of the ACM SIGCHI conferences board. Prior to joining Google, Lora was a computer science professor at the VU University Amsterdam.

Keynote Talk: Model-in-the-loop Data Collection: What Roles does the Model Play?

Tongshuang Wu
Carnegie Mellon University

Abstract: Assistive models have been shown useful for supporting humans in creating challenging datasets, but how exactly do they help? In this talk, I will discuss different roles of assistive models in counterfactual data collection (i.e., perturbing existing text inputs to gain insight into task model decision boundaries), and the characteristics associated with these roles. I will use three examples (CheckList, Polyjuice, Tailor) to demonstrate how our objectives shift when we perturb texts for evaluation, explanation, and improvement, and how that change the corresponding assistive models from enhancing human goals (requiring model controllability) to competing with human bias (requiring careful data reranking). I will conclude by exploring additional roles that these models can play to become more effective.

Bio: Sherry Tongshuang Wu is an Assistant Professor at the Human Computer Interaction Institute at Carnegie Mellon University (CMU HCII), holding a courtesy appointment at the Language Technolgoy Institute (CMU LTI). Sherry's research lies at the intersection of Human-Computer Interaction (HCI) and Natural Language Processing (NLP). She aims to understand and support people coping with imperfect AI models, both when the model is under active development, and after it is deployed for end users.

Table of Contents

Program

Thursday, July 14, 2022

09:00 - 09:10	*Opening Remarks*
09:10 - 09:25	*Collaborative Progress: ML Commons Introduction*
09:25 - 10:00	*Invited Talk 1: Anna Rogers*
10:00 - 10:35	*Invited Talk 2: Jordan Boyd-Graber*
10:35 - 10:50	*Break*
10:50 - 11:10	*Best Paper Talk:* *Overconfidence in the Face of Ambiguity with Adversarial Data* Margaret Li and Julian Michael
11:10 - 11:45	*Invited Talk 3: Sam Bowman*
11:45 - 12:20	*Invited Talk 4: Lora Aroyo*
12:20 - 13:20	*Lunch*
13:20 - 13:55	*Invited Talk 5: Sherry Tongshuang Wu*
13:55 - 14:55	*Panel: The Future of Data Collection*
14:55 - 15:10	*Break*
15:10 - 15:20	*Introduction to the DADC Shared Task: Max Bartolo*
15:20 - 15:40	*Shared Task Winners' Presentations*
15:40 - 16:55	*Poster Session*

Thursday, July 14, 2022 (continued)

Resilience of Named Entity Recognition Models under Adversarial Attack
Sudeshna Das and Jiaul Paik

GreaseVision: Rewriting the Rules of the Interface
Siddhartha Datta, Konrad Kollnig and Nigel Shadbolt

Posthoc Verification and the Fallibility of the Ground Truth
Yifan Ding, Nicholas Botzer and Tim Weninger

Overconfidence in the Face of Ambiguity with Adversarial Data
Margaret Li and Julian Michael

longhorns at DADC 2022: How many linguists does it take to fool a Question Answering model? A systematic approach to adversarial attacks.
Venelin Kovatchev, Trina Chatterjee, Venkata S Govindarajan, Jifan Chen, Eunsol Choi, Gabriella Chronis, Anubrata Das, Katrin Erk, Matthew Lease, Junyi Jessy Li, Yating Wu and Kyle Mahowald

Collecting high-quality adversarial data for machine reading comprehension tasks with humans and models in the loop
Damian Y. Romero Diaz, Magdalena Anioł and John Culnan

Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks
Ruixiang Cui, Daniel Hershcovich and Anders Søgaard

Adversarially Constructed Evaluation Sets Are More Challenging, but May Not Be Fair
Jason Phang, Angelica Chen, William Huang and Samuel R. Bowman

16:55 - 17:00 *Closing Session*