2022

Challenges & Perspectives in Creating Large Language Models

Proceedings of the Workshop

May 27, 2022

The organizers gratefully acknowledge the support from the following sponsors.

Sponsor

NAVER LABS
Europe

©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 978-1-955917-26-1

Introduction

Two years after the appearance of GPT-3, large language models seem to have taken over NLP. Their capabilities, limitations, societal impact and the potential new applications they unlocked have been discussed and debated at length. A handful of replication studies have been published since then, confirming some of the initial findings and discovering new limitations. This workshop aims to gather researchers and practitioners involved in the creation of these models in order to:

1. Share ideas on the next directions of research in this field, including—but not limited to—grounding, multi-modal models, continuous updates and reasoning capabilities.

2. Share best-practices, brainstorm solutions to identified limitations and discuss challenges, such as infrastructure, data, ethical & legal frameworks, evaluation, training efficiency, etc.

This workshop is organized by the BigScience[1] initiative and will also serve as the closing session of this one year-long initiative aimed at developing a multilingual large language model, which is gathering 1.000+ researchers from more than 60 countries and 250 institutions and research labs. Its goal is to investigate the creation of a large scale dataset and model from a very wide diversity of angles.

[1] https://bigscience.huggingface.co/

Organizing Committee

Organization Committee

Angela Fan, Meta AI
Matthias Gallé, Naver Labs Europe
Suzana Ilić, HuggingFace
Thomas Wolf, HuggingFace

Steering Committee

Yoav Goldberg, Bar Ilan University & Allen Institute for Artificial Intelligence
Percy Lang, Stanford University
Margaret Mitchell, HuggingFace & Ethical AI LLC
Alice Oh, KAIST
Alexander Rush, Cornell University

Program Committee

Program Chairs

Angela Fan, Facebook
Matthias Gallé, Naver Labs Europee
Suzana Ilic, HuggingFace
Thomas Wolf, HuggingFace

Table of Contents

Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora
Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold and Xiang Ren 1

Using ASR-Generated Text for Spoken Language Modeling
Nicolas Hervé, Valentin Pelloin, Benoit Favre, Franck Dary, Antoine Laurent, Sylvain Meignier and Laurent Besacier 17

You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings
Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla and Oskar Van Der Wal 26

Diverse Lottery Tickets Boost Ensemble from a Single Pretrained Model
Sosuke Kobayashi, Shun Kiyono, Jun Suzuki and Kentaro Inui 42

UNIREX: A Unified Learning Framework for Language Model Rationale Extraction
Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren and Hamed Firooz 51

Pipelines for Social Bias Testing of Large Language Models
Debora Nozza, Federico Bianchi and Dirk Hovy 68

Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0
Francesco De Toni, Christopher Akiki, Javier De La Rosa, Clémentine Fourrier, Enrique Manjavacas, Stefan Schweter and Daniel Van Strien 75

A Holistic Assessment of the Carbon Footprint of Noor, a Very Large Arabic Language Model
Imad Lakim, Ebtesam Almazrouei, Ibrahim Abualhaol, Merouane Debbah and Julien Launay 84

GPT-NeoX-20B: An Open-Source Autoregressive Language Model
Sidney Black, Stella Biderman, Eric Hallahan, Quentin Gregory Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Martin Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang and Samuel Weinbach 95

Dataset Debt in Biomedical Language Modeling
Jason Alan Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, Matthias Samwald and Wojciech Kusa 137

Emergent Structures and Training Dynamics in Large Language Models
Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin and Aaron Gokaslan 146

Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned
Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski and Svitlana Volkova 160

Program

Friday, May 27, 2022

12:30 - 11:00	*Poster Session*
14:00 - 15:00	*BigScience*
15:00 - 15:20	*Data Governance*
15:20 - 15:40	*Data*
15:40 - 16:00	*Modeling*
16:00 - 16:20	*Prompt Engineering*
16:20 - 16:40	*Evaluation*