

DeepLo 2022

**The 3rd Workshop on Deep Learning Approaches for
Low-Resource NLP**

Proceedings of the DeepLo Workshop

July 14, 2022

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

Gold



©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 978-1-955917-97-1

Introduction

The NAACL 2022 Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing (DeepLo) takes place on Thursday, July 22, in Seattle Washington, USA, immediately after the main conference.

Natural Language Processing is being revolutionized by deep learning. However, deep learning requires large amounts of annotated data, and its advantage over traditional statistical methods typically diminishes when such data is not available. Large amounts of annotated data simply do not exist for many low-resource languages. Even for high-resource languages it can be difficult to find linguistically annotated data of sufficient size and quality to allow neural methods to excel; this remains true even as few-shot learning approaches have gained popularity in recent years.

This workshop aims to bring together researchers from the NLP and ML communities who work on learning with neural methods when there is not enough data for those methods to succeed out-of-the-box. Specifically, it will provide attendees with an overview of new and existing approaches from various disciplines, and enable them to distill principles that can be more generally applicable. We will also discuss the main challenges arising in this setting, and outline potential directions for future progress.

Our program covers a broad spectrum of applications and techniques. It is augmented by invited talks from Yulia Tsvetkov, Sebastian Ruder, Graham Neubig, and David Ifeoluwa Adelani.

We would like to thank the members of our Program Committee for their timely and thoughtful reviews.

Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, anyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, Swabha Swayamdipta

Organizing Committee

Organizing Committee Members

Colin Cherry, Google Research
Angela Fan, Facebook AI Research
George Foster, Google Research
Gholamreza (Reza) Haffari, Monash University
Shahram Khadivi, eBay
Nanyun (Violet) Peng, UCLA
Xiang Ren, USC/ISI
Ehsan Shareghi, Monash University
Swabha Swayamdipta, Allen Institute for AI

Program Committee

Invited Speakers

Yulia Tsvetkov, University of Washington
Sebastian Ruder, Google
Graham Neubig, Carnegie Mellon University
David Ifeoluwa Adelani, Saarland University

Reviewers

David Adelani, Saarland University
Emily Allaway, Columbia University
Parnia Bahar, AppTek
Marco Basaldella, Amazon
Leonard Dahlmann, eBay
Haim Dubossarsky, Queen Mary University of London
Kevin Duh, Johns Hopkins University
Markus Freitag, Google
Yingbo Gao, RWTH Aachen University
Thamme Gowda, University of Southern California
Xuanli He, Monash University
Yacine Jernite, Hugging Face
Robin Jia, University of Southern California
Jonathan K., University of Michigan
Zhuang Li, Monash University
Manuel Mager, University of Stuttgart
Evgeny Matusov, AppTek
Zaiqiao Meng, University of Glasgow
Phoebe Mulcaire, University of Washington
Benjamin Muller, INRIA Paris
Arturo Oncevay, University of Edinburgh
Pavel Petrushkov, eBay
Victor Prokhorov, University of Edinburgh
Roi Reichart, Technion
Sebastian Ruder, Google
Partha Talukdar, Google Research India
David Thulke, RWTH Aachen University
Nicola Ueffing, eBay
Thuy-Trang Vu, Monash University
Ivan Vulić, University of Cambridge
Sarah Wiegrefe, Georgia Tech
Zhaofeng Wu, AI2
Minghao Wu, Monash University
Poorya Zareemoodi, Oracle Labs
Jinming Zhao, Monash University
Mengjie Zhao, LMU
Yanpeng Zhao, University of Edinburgh
Wenxuan Zhou, University of Southern California

Table of Contents

<i>Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua</i> Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas and Hilario Aradiel and Nelsi Melgarejo	1
<i>Improving Distantly Supervised Document-Level Relation Extraction Through Natural Language Inference</i> Clara Vania and Grace Lee and Andrea Pierleoni	14
<i>IDANI: Inference-time Domain Adaptation via Neuron-level Interventions</i> Omer Antverg and Eyal Ben-David and Yonatan Belinkov	21
<i>Generating unlabelled data for a tri-training approach in a low resourced NER task</i> Hugo Boulanger and Thomas Lavergne and Sophie Rosset	30
<i>ANTS: A Framework for Retrieval of Text Segments in Unstructured Documents</i> Brian Chivers, Mason P. Jiang, Wonhee Lee, Amy Ng and Natalya I. Rapstine and Alex Storer	38
<i>Cross-TOP: Zero-Shot Cross-Schema Task-Oriented Parsing</i> Melanie A. Rubino, Nicolas Guenon des mesnards, Uday Shah, Nanjiang Jiang and Weiqi Sun and Konstantine Arkoudas	48
<i>Help from the Neighbors: Estonian Dialect Normalization Using a Finnish Dialect Generator</i> Mika Hämäläinen and Khalid Alnajjar and Tuuli Tuisk	61
<i>Exploring diversity in back translation for low-resource machine translation</i> Laurie Burchell and Alexandra Birch and Kenneth Heafield	67
<i>Punctuation Restoration in Spanish Customer Support Transcripts using Transfer Learning</i> Xiliang Zhu, Shayna Gardiner, David Rossouw and Tere Roldán and Simon Corston-Oliver ..	80
<i>Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese</i> Kurt Micallef, Albert Gatt, Marc Tanti and Lonneke van der Plas and Claudia Borg	90
<i>Building an Event Extractor with Only a Few Examples</i> Pengfei Yu, Zixuan Zhang, Clare Voss and Jonathan May and Heng Ji	102
<i>Task Transfer and Domain Adaptation for Zero-Shot Question Answering</i> Xiang Pan, Alex Sheng, David Shimshoni, Aditya Singhal and Sara Rosenthal and Avirup Sil	110
<i>Let the Model Decide its Curriculum for Multitask Learning</i> Neeraj Varshney and Swaroop Mishra and Chitta Baral	117
<i>AfriTeVA: Extending ?Small Data? Pretraining Approaches to Sequence-to-Sequence Models</i> Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi and Kelechi Ogueji and Jimmy Lin	126
<i>Few-shot Learning for Sumerian Named Entity Recognition</i> Guanghai Wang and Yudong Liu and James Hearne	136
<i>Deep Learning-Based Morphological Segmentation for Indigenous Languages: A Study Case on Innu-Aimun</i> Ngoc Tan Le, Antoine Cadotte, Mathieu Boivin and Fatiha Sadat and Jimena Terraza	146

<i>Clean or Annotate: How to Spend a Limited Data Collection Budget</i>	
Derek Chen and Zhou Yu and Samuel R. Bowman	152
<i>Unsupervised Knowledge Graph Generation Using Semantic Similarity Matching</i>	
Lixian Liu, Amin Omidvar, Zongyang Ma and Ameeta Agrawal and Aijun An	169
<i>FarFetched: Entity-centric Reasoning and Claim Validation for the Greek Language based on Textually Represented Environments</i>	
Dimitris Papadopoulos, Katerina Metropoulou and Nikolaos Papadakis and Nikolaos Matsatsinis	180
<i>Alternative non-BERT model choices for the textual classification in low-resource languages and environments</i>	
Syed Mustavi Maheen, Moshir Rahman Faisal and Md. Rafakat Rahman and Md. Shahriar Karim	192
<i>Generating Complement Data for Aspect Term Extraction with GPT-2</i>	
Amir Pouran Ben Veyseh, Franck Dernoncourt and Bonan Min and Thien Huu Nguyen	203
<i>How to Translate Your Samples and Choose Your Shots? Analyzing Translate-train & Few-shot Cross-lingual Transfer</i>	
Iman Jundi and Gabriella Lapesa	214
<i>Unified NMT models for the Indian subcontinent, transcending script-barriers</i>	
Gokul N.C.	227

Program

Thursday, July 14, 2022

- 08:50 - 09:00 *Opening Remarks*
- 10:00 - 09:45 *Invited talk - Sebastian Ruder (Virtual)*
- 10:00 - 10:30 *Coffee Break*
- 10:30 - 11:15 *Invited talk - David Ifeoluwa Adelani (In-person)*
- 11:15 - 12:15 *Poster session I (Virtual)*
- 13:30 - 12:15 *Lunch Break*
- 13:30 - 14:15 *Invited talk - Yulia Tsvetkov (In-person)*
- 14:15 - 15:15 *Poster session II (In-person) and Poster session III (Virtual)*
- 15:15 - 15:30 *Coffee Break*
- 15:30 - 16:15 *Invited talk - Graham Neubig (In-person)*
- 16:15 - 16:30 *Closing remark*