MOOMIN 2024

Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)

Proceedings of the Workshop

March 21, 2024

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 979-8-89176-084-4

Introduction

Welcome to the 1st Workshop Proceedings on Modular and Open Multilingual NLP (MOOMIN)". The workshop will take place at EACL 2024 in Malta on March 21st.

The MOOMIN workshop's aim is to bring together researchers and NLP practitioners interested in modular approaches to the design of natural language systems. This trend of research is a direct reply to the challenges and opportunities of monolithic large language models: To keep our field sustainable, we need models that are reusable, adaptable, and repurposable. We invited paper submissions on various topics, including Mixture-of-Expert models, modular pre-training of multilingual language and translation models, techniques that leverage adapters and hypernetworks, modular extensions of existing NLP models systems, and especially welcome work focusing on low-resource settings.

We have curated the MOOMIN workshop program to encourage discussions that will lead to valuable insights into the workshop topics. On the day of the workshop, there will be a total of 9 oral presentations of papers that offer innovative approaches and solutions to the challenges of scalability, language coverage, efficiency and re-usability of large language models. Of these 9 presentations, 5 correspond to archival papers published in the workshop proceedings, 1 is a non-archival submission and 3 papers are coming from this years' EACL Findings. The overall acceptance rate of archival submissions was 62.5%. In addition, we also invited two keynote speakers, Edoardo M. Ponti and Angela Fan, whose works have had remarkable impact in the field of modular NLP.

We are grateful to all authors, reviewers, and participants who contributed to the success of this workshop. We would also like to thank the European Research Council and the Research Council of Finland for their support of the workshop through the FoTran project (grant agreement no. 771113) and the GreenNLP project, respectively.

The MOOMIN organizers,

Timothee Mickus, Jörg Tiedemann, Ahmet Üstün, Raúl Vázquez & Ivan Vulić

Program Committee

Program Chairs

Timothee Mickus, University of Helsinki
Jörg Tiedemann, University of Helsinki
Ivan Vulić, University of Cambridge and PolyAI Limited
Raúl Vázquez, University of Helsinki
Ahmet Üstün, Cohere For Ai

Publication Chair

Raúl Vázquez, University of Helsinki

Invited Speakers

Edoardo M. Ponti, University of Edinburgh and University of Cambridge
Angela Fan, Meta AI Research, FAIR

Reviewers

David Ifeoluwa Adelani, University College London
Alan Ansell, University of Cambridge
Lucas Caccia, McGill University
Hande Celikkanat, University of Helsinki
Alexandra Chronopoulou, Ludwig-Maximilians-Universität München
Mathias Creutz, University of Helsinki
Marzieh Fadaee, Cohere For AI
Stig-Arne Grönroos, University of Helsinki
Barry Haddow, University of Edinburgh
Shaoxiong Ji, University of Helsinki
Julia Kreutzer, Cohere for AI
Andrey Kutuzov, University of Oslo
Niki Andreas Loppi, NVIDIA
Kelly Marchisio, Cohere and Cohere
Benjamin Minixhofer, University of Cambridge
Joakim Nivre, Uppsala University
Clifton A Poth, Cohere and Technische Universität Darmstadt
Taido Purason, University of Tartu
Alessandro Raganato, University of Milan - Bicocca
Fabian David Schmidt, Bayerische Julius-Maximilians-Universität Würzburg
Miikka Silfverberg, University of British Columbia
Teemu Vahtola, University of Helsinki

Keynote Talk: Efficiency as an Inductive Bias for Language Learning

Edoardo M. Ponti
University of Edinburgh and University of Cambridge
2024-03-21 09:30:00 – Room: **Room 1**

Abstract: Efficiency in Natural Language Processing is often hailed as a solution to democratise access to AI technology and to make it more environmentally sustainable. In this talk, I emphasise an additional and sometimes neglected advantage of efficiency: namely, providing an inductive bias for language use and acquisition closer to those in humans, where efficiency trade-offs shape the very structure of language. I will start by recapitulating the main aspects of efficiency in deep learning, which are partly interconnected: time, memory, and parameter efficiency. Next, I will explore how efficient designs in state-of-the-art Large Language Models (a) may also act as inductive biases that improve their performance (b). For instance: (1a) Jointly learning to model and segment text allows for merging contiguous groups of token representations in intermediate layers, which reduces time and memory requirements. (1b) In addition, it also leads to learning (possibly reusable and hierarchical) abstractions from raw data, which further increase the model's predictive abilities; (2a) Learning parameter-efficient modules allows for fine-tuning LLMs with limited memory budgets. (2b) In addition, composing these specialised modules through appropriate routing also leads to better generalisation. In particular, I will show how modules can be implemented as highly composable sparse adapters and how routing through modules can be learned automatically. In conclusion, efficient designs of LLMs yield unexpected benefits, such as the ability to learn abstractions, adapt fast, and integrate disparate sources of knowledge.

Bio: Edoardo M. Ponti is a Lecturer (Assistant Professor) in Natural Language Processing at the University of Edinburgh, where he is part of the Institute for Language, Cognition, and Computation (ILCC), and an Affiliated Lecturer at the University of Cambridge. Previously, he was a visiting postdoctoral scholar at Stanford University and a postdoctoral fellow at Mila and McGill University in Montreal. In 2021, he obtained a PhD in computational linguistics from the University of Cambridge, St John's College. His main research foci are modular deep learning, sample-efficient learning, faithful text generation, computational typology and multilingual NLP. His research earned him a Google Research Faculty Award and 2 Best Paper Awards at EMNLP 2021 and RepL4NLP 2019. He is a board member and co-founder of SIGTYP, the ACL special interest group for computational typology, and a scholar of the European Lab for Learning and Intelligent Systems (ELLIS). He is a (terrible) violinist, football player, and an aspiring practitioner of heroic viticulture.

Keynote Talk: No Language Left Behind - Scaling Human-Centered Machine Translation

Angela Fan
Meta AI Research, FAIR
2024-03-21 16:00:00 – Room: **Room 2**

Abstract: Driven by the goal of eradicating language barriers on a global scale, machine translation has solidified itself as a key focus of artificial intelligence research today. However, such efforts have coalesced around a small subset of languages, leaving behind the vast majority of mostly low-resource languages. What does it take to break the 200 language barrier while ensuring safe, high-quality results, all while keeping ethical considerations in mind? In this talk, I introduce No Language Left Behind, an initiative to break language barriers for low-resource languages. In No Language Left Behind, we took on the low-resource language translation challenge by first contextualizing the need for translation support through exploratory interviews with native speakers. Then, we created datasets and models aimed at narrowing the performance gap between low and high-resource languages. We proposed multiple architectural and training improvements to counteract overfitting while training on thousands of tasks. Critically, we evaluated the performance of over 40,000 different translation directions using a human-translated benchmark, Flores-200, and combined human evaluation with a novel toxicity benchmark covering all languages in Flores-200 to assess translation safety. Our model achieves an improvement of 44% BLEU relative to the previous state-of-the-art, laying important groundwork towards realizing a universal translation system in an open-source manner.

Bio: Angela is a research scientist at Meta AI Research in New York, focusing on research in text generation. Currently, Angela works on language modeling and developing the line AI Agents Meta products. Recent research projects include No Language Left Behind, Universal Speech Translation for Unwritten Languages, and Llama2.

Table of Contents

Program

Thursday, March 21, 2024

09:15 - 09:30 *Opening Remarks*

09:30 - 10:30 *Keynote 1: Edoardo M. Ponti*

10:30 - 11:00 *Coffee Break*

11:00 - 12:20 *Session 1: Efficient use of Adapters*

The Impact of Language Adapters in Cross-Lingual Transfer for NLU
Jenny Kunz and Oskar Holmström

Modular Adaptation of Multilingual Encoders to Written Swiss German Dialect
Jannis Vamvas, Noëmi Aepli and Rico Sennrich

Less is Fed More: Sparsity Reduces Feature Distortion in Federated Learning
Aashiq Muhamed, Harshita Diddee and Abhinav Rao

Toward the Modular Training of Controlled Paraphrase Adapters
Teemu Vahtola and Mathias Creutz

12:20 - 14:00 *Lunch Break*

14:00 - 15:00 *Session 2: Selection and weighting of modules*

Mixing and Matching: Combining Independently Trained Translation Model Components
Taido Purason, Andre Tättar and Mark Fishel

Sequence Shortening for Context-Aware Machine Translation
Paweł Maka, Yusuf Can Semerci, Jan Scholtes and Gerasimos Spanakis

What the Weight?! A Unified Framework for Zero-Shot Knowledge Composition
Carolin Holtermann, Markus Frohmann, Navid Rekabsaz and Anne Lauscher

Thursday, March 21, 2024 (continued)

15:00 - 15:40 *Session 3: Tuning LLMs*

Soft Prompt Tuning for Cross-Lingual Transfer: When Less is More
Fred Philippy, Siwen Guo, Shohreh Haddadan, Cedric Lothritz, Jacques Klein and Tegawendé F. Bissyandé

Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca
Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow and Kenneth Heafield

15:40 - 16:00 *Coffee Break*

16:00 - 17:00 *Keynote 2: Angela Fan*

17:00 - 17:20 *Closing Remarks*