MRL 2024

The 4th Workshop on Multilingual Representation Learning

Proceedings of the Workshop

November 16, 2024

©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-184-1

Organizing Committee

Organizers

David Ifeoluwa Adelani, McGill University, Canada
Duygu Ataman, New York University, USA
Mammad Hajili, Microsoft, USA
Raghav Mantri, New York University, USA
David Stap, University of Amsterdam, Netherlands
Jonne Sälevä, Brandeis University, USA
Francesco Tinner, University of Amsterdam, Netherlands
Abraham Owodunni, Ohio State University, USA

Program Committee

Reviewers

David Ifeoluwa Adelani, Manuel Bobie Amankwatia, Catherine Arnett

Travis M. Bartley, Vishal Bhalla

Jiajing Chen, Xupeng Chen, Olutosoye Taiwo Christian, Zoltan Csaki

Konstantin Dobler, Koel Dutta Chowdhury

Yassine El Kheir, Abdellah El Mekki

Senkang Hu

Yusif Ibrahimov, Mironshoh Inomjonov, Jafar Isbarov

Ainaz Jamshidi, Jiby Mariya Jose

Yixiao Kang, Zhengjian Kang, Christopher Klamm, Hongzhi Kuai

Senyu Li, Yueqian Lin, Weisi Liu

Pranita Yogesh Mahajan, Nimshi Venkat Meripo

Muhammad Amin Nadim, Usman Nawaz

Esther Odunayo Oduntan, Peter Oseghale Ohue, Yewande Ojo

Chester Palen-Michel

Shaibal Saha, Shubham Shukla, Janet Yunchen Sung, Jonne Sälevä

Shaomu Tan, Wenjia Tan, Yihang Tao, Shailja Thakur

Vajratiya Vajrobol, Deepali Verma

Sahil Walia, Azmine Toushik Wasi, Di Wu

Sadia Zaib, Zhehao Zhang, Xufeng Zhao, Huichi Zhou, Ziqi Zhou

Keynote Talk

Invited talk 1

Karen Livescu
TTI at Chicago

2024-11-16 9:10 –

Bio: Karen Livescu is a Professor at the Toyota Technological Institute at Chicago (TTIC). Her research focuses on speech and language processing and related areas of machine learning. She obtained her PhD in Computer Science MIT 2005, working in the Spoken Language Systems group of the Computer Science and Artificial Intelligence Laboratory.

Keynote Talk
Invited talk 2

Hila Gonen
University of Washington
2024-11-16 9:50 –

Bio: Hila Gonen is a postdoctoral Researcher at the Paul G. Allen School of Computer Science & Engineering at the University of Washington working on Natural Language Processing. Her research focuses on two goals: (1) making cutting-edge language technology available and fair across speakers of different languages and users of different socio-demographic groups; (2) developing algorithms and methods for controlling the model's behavior. Prior to joining UW, she completed a Ph.D in Computer Science at the NLP lab at Bar Ilan University.

Keynote Talk
Invited talk 3

Sebastian Ruder
Cohere for AI
2024-11-16 16:00 –

Bio: Sebastian Ruder is a research scientist at Cohere based in Berlin, Germany working on making large language models (LLMs) multilingual. He completed his PhD in Natural Language Processing and Deep Learning at the Insight Research Centre for Data Analytics.

Table of Contents

SambaLingo: Teaching Large Language Models New Languages
Zoltan Csaki, Bo Li, Jonathan Lingjie Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu and Urmish Thakker 1

What an Elegant Bridge: Multilingual LLMs are Biased Similarly in Different Languages
Viktor Mihaylov and Aleksandar Shtedritski 22

Adapting Open-Source Generative Large Language Models for Low-Resource Languages: A Case Study for Turkish
Cagri Toraman 30

An Efficient Approach for Studying Cross-Lingual Transfer in Multilingual Language Models
Fahim Faisal and Antonios Anastasopoulos 45

Are You Sure? Rank Them Again: Repeated Ranking For Better Preference Datasets
Peter Devine 93

Tagengo: A Multilingual Chat Dataset
Peter Devine 106

Language and Task Arithmetic with Parameter-Efficient Layers for Zero-Shot Summarization
Alexandra Chronopoulou, Jonas Pfeiffer, Joshua Maynez, Xinyi Wang, Sebastian Ruder and Priyanka Agrawal 114

Modeling Bilingual Sentence Processing: Evaluating RNN and Transformer Architectures for Cross-Language Structural Priming
Demi Zhang, Bushi Xiao, Chao Gao, Sangpil Youm and Bonnie J Dorr 127

Recipe for Zero-shot POS Tagging: Is It Useful in Realistic Scenarios?
Zeno Vandenbulcke, Lukas Vermeire and Miryam De Lhoneux 137

Gender-specific Machine Translation with Large Language Models
Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe and Marta R. Costa-jussà148

Jina-ColBERT-v2: A General-Purpose Multilingual Late Interaction Retriever
Han Xiao, Bo Wang and Rohan Jha 159

Cross-Lingual Named Entity Recognition for Low-Resource Languages: A Hindi-Nepali Case Study Using Multilingual BERT Models
Dipendra Yadav, Sumaiya Suravee, Tobias Strauß and Kristina Yordanova 167

Parameter-efficient Adaptation of Multilingual Multimodal Models for Low-resource ASR
Abhishek Gupta, Amruta Parulekar, Sameep Chattopadhyay and Preethi Jyothi 175

Towards Cross-Linguistic Semantic Grounding using Dictionary Graph Analysis
Ethan Eschrich and Zoey Liu 186

Vikhr: Constructing a State-of-the-art Bilingual Open-Source Instruction-Following Large Language Model for Russian
Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev and Artem Shelmanov
189

Mitigating the Linguistic Gap with Phonemic Representations for Robust Cross-lingual Transfer
Haeji Jung, Changdae Oh, Jooeon Kang, Jimin Sohn, Kyungwoo Song, Jinkyu Kim and David R Mortensen 200

Leveraging Adapters for Improved Cross-lingual Transfer for Low-Resource Creole MT
Marcell Richard Fekete, Ernests Lavrinovics, Nathaniel Romney Robinson, Heather Lent, Raj Dabre and Johannes Bjerva 212

Evaluating Multilingual Long-Context Models for Retrieval and Reasoning
Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel and Russell Scheinberg 216

Community OSCAR: A Community Effort for Multilingual Web Data
Manuel Brack, Malte Ostendorff, Pedro Ortiz Suarez, José Javier Saiz, Iñaki Lacunza Castilla, Jorge Palomar-Giner, Alexander Shvets, Patrick Schramowski, Georg Rehm, Marta Villegas and Kristian Kersting 232

Leveraging LLMs for Translating and Classifying Mental Health Data
Konstantinos Skianis, A. Seza Doğruöz and John Pavlopoulos 236

Bridging the Bosphorus: Advancing Turkish Large Language Models through Strategies for Low-Resource Language Adaptation and Benchmarking
Emre Can Acikgoz, Mete Erdogan and Deniz Yuret 242

Unsupervised Text Representation Learning via Instruction-Tuning for Zero-Shot Dense Retrieval
Qiuhai Zeng, Zimeng Qiu, Dae Yon Hwang, Xin He and William M. Campbell 269

Language Bias in Multilingual Information Retrieval: The Nature of the Beast and Mitigation Methods
Jinrui Yang, Fan Jiang and Timothy Baldwin 280

Representational Isomorphism and Alignment of Multilingual Large Language Models
Di Wu, Yibin Lei, Andrew Yates and Christof Monz 293

Generalization Measures for Zero-Shot Cross-Lingual Transfer
Saksham Bassi, Duygu Ataman and Kyunghyun Cho 298

Detecting and Translating Language Ambiguity with Multilingual LLMs
Behrang Mehrparvar and Sandro Pezzelle 310

MLT-DR: Multi-Lingual/Task Demonstration Retrieval
An Attempt towards Generalized Retriever for In-Context Learning
Kazuma Hashimoto, Arjun Reddy Akula, Karthik Raman and Michael Bendersky 324

McGill NLP Group Submission to the MRL 2024 Shared Task: Ensembling Enhances Effectiveness of Multilingual Small LMs
Senyu Li, Hao Yu, Jessica Ojo and David Ifeoluwa Adelani 346

CUNI and LMU Submission to the MRL 2024 Shared Task on Multi-lingual Multi-task Information Retrieval
Katharina Hämmerl, Andrei-Alexandru Manea, Gianluca Vico, Jindřich Helcl and Jindřich Libovický 357

Findings of the 2nd Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2024
Francesco Tinner, Raghav Mantri, Mammad Hajili, Chiamaka Chukwuneke, Dylan Massey, Benjamin A. Ajibade, Bilge Deniz Kocak, Abolade Dawud, Jonathan Atala, Hale Sirin, Kayode Olaleye, Anar Rzayev, David Adelani and Duygu Ataman 365

Program

Saturday, November 16, 2024

09:00 - 09:10	*Opening Remarks*
09:10 - 09:50	*Invited Talk by Karen Livescu*
09:50 - 10:30	*Invited Talk by Hila Gonen*
10:30 - 11:00	*Coffee Break*
11:00 - 12:30	*Poster Session*
12:30 - 14:00	*Lunch Break*
14:00 - 14:15	*Findings Paper*
14:15 - 14:30	*Winning Team Presentation*
14:30 - 15:00	*Best Paper*
15:00 - 15:30	*Honorable Mentions*
15:30 - 16:00	*Coffee Break*
16:00 - 16:50	*Invited Talk by Sebastian Ruder*
16:50 - 17:00	*Closing Remarks*