MRL 2021

The 1st Workshop on Multilingual Representation Learning

Proceedings of the Conference

November 11, 2021

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 978-1-954085-96-1

Message from the General Chair

Multilingual representation learning methods have recently been found to be extremely efficient in learning features useful for transfer learning between languages and demonstrating potential in achieving successful adaptation of natural language processing (NLP) models into languages or tasks with little to no training resources. On the other hand, there are many aspects of such models which have the potential for further development and analysis in order to prove their applicability in various context. These contexts include different NLP tasks and also understudied language families, which face important obstacle in achieving practical advances that could improve the state-of-the-art in NLP of various low-resource or underrepresented languages.

The aim of this workshop is to form the first research community in multilingual representation learning providing the rapidly growing number of researchers working on the topic with a means of communication and an opportunity to present their work and exchange ideas. The main objectives of the workshop are constructing and presenting a wide array of multilingual representation learning methods, including their theoretical formulation and analysis, practical aspects such as the application of current state-of-the-art approaches in transfer learning to different tasks or studies on adaptation into previously under-studied context; providing a better understanding on how the language typology may impact the applicability of these methods and motivate the development of novel methods that are more generic or competitive in different languages; and promoting collaborations in developing novel software libraries or benchmarks in implementing or evaluating multilingual models that would accelerate progress in the field.

The proceedings presents the collection of original research contributions submitted by various and greatly diverse parts of the community which advance the conventional approaches to understanding and implementing multi-lingual representation learning methods, immediately extending their applicability as well as proposing previously unexplored ideas. As the organizing committee we are deeply grateful to the engagement and response from the community and hope to continue to foster the collaborative environment the workshop achieves at its first organization.

Message from the Program Chairs

Having been organized for the first time in an emerging area of cutting-edge research, the Workshop on Multilingual Representation Learning has received a large amount of interest from the community, with 56 submissions to the program. 50 of these submissions were long papers describing original research and 6 presented ongoing or previously published work in the form of extended abstracts. Our program committee consisted of 55 experts in the area from all over the world, conducting research in the industry as well as academia. The committee worked throughly on every submission and selected 19 research papers and 5 extended abstracts to be presented at the workshop. We are glad to present our proceedings gathering the products of a great collaborative effort which we hope will inspire many new ideas to be presented at future events of our workshop with increasing capacity in the years to come.

Organizing Committee

Duygu Ataman, New York University
Alexandra Birch, University of Edinburgh
Alexis Conneau, Google
Orhan Firat, Google
Sebastian Ruder, Deepmind
Gozde Gul Sahin, TU Darmstadt

Program Committee

Badr Abdullah, University of Saarland
David Ifeoluwa Adelani, Saarland University
Gustavo Aguliar, Amazon
Roee Aharoni, Google
Orevaoghene Ahia, InstaDeep
Chantal Amrhein, University of Zurich
Timothy Baldwin, University of Melbourne
Ankur Bapna, Google
Chistos Baziotis, University of Edinburgh
Eleftheria Briakou, University of Maryland
Marcel Bollman, Jonkoping University
Iacer Calixto, University of Amsterdam
Mona Diab, Facebook
Sumanth Doddapaneni, IIT Madras
Steffen Eger, TU Darmstadt
Akiko Eriguchi, Microsoft
Xavier Garcia, Google
Ankush Garg, Google
Goran Glavas, University of Mannheim
Hila Gonen, Facebook
Naman Goyal, Facebook
Melvin Johnson, Google
Yova Kementchedjhieva, University of Copenhagen
Simran Khanuja, Google
Christopher Klamm, TU Darmstadt
Sneha Kudugunta, Google
Surafel Melaku Lakew, Amazon
Kaushal Kumar Maurya, Indian Institute of Techonlogy Hyderabad
Stephen Mayhew, Duolingo
Todor Mihaylov, Facebook
Jamshidbek Mirzakhalov, University of South Florida
Ji-Ung Lee, TU Darmstadt
Graham Neubig, Carnegie Mellon University
Dmitry Nikolaev, Stockholm University
Yusuke Oda, LegalForce
Atul Kr. Ojha, National University of Ireland Galway
Arturo Oncevay, University of Edinburgh
Kelechi Ogueji, InstaDeep
Rajaswa Patil, TCS Research
Jonas Pfeiffer, TU Darmstadt
Leonardo Ribeiro, TU Darmstadt
Phillip Rust, University of Copenhagen

Rico Sennrich, University of Zurich
Aditya Siddhant, Google
Jorg Tiedemann, University of Helsinki
Ahmet Ustun, University of Groningen
Jannis Vamvas, University of Zurich
Clara Vania, Amazon
Ivan Vulic, University of Cambridge
Andreas Waldis, TU Darmstadt
Kexin Wang, TU Darmstadt
Taro Watanabe, NAIST
Genta Indra Winata, Bloomberg
Luke Zettlemoyer, Facebook
Biao Zhang, University of Edinburgh

Table of Contents

Conference Program

Thursday November 11, 2021

10:00–10:30

Language Models are Few-shot Multilingual Learners
Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski and Pascale Fung

10:30–11:00

Learning Contextualised Cross-lingual Word Embeddings and Alignments for Extremely Low-Resource Languages Using Parallel Corpora
Takashi Wada, Tomoharu Iwata, Yuji Matsumoto, Timothy Baldwin and Jey Han Lau

11:00–12:00

Clustering Monolingual Vocabularies to Improve Cross-Lingual Generalization
Riccardo Bassani, Anders Søgaard and Tejaswini Deoskar

Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings
Lisa Woller, Viktor Hangya and Alexander Fraser

Specializing Multilingual Language Models: An Empirical Study
Ethan C. Chau and Noah A. Smith

Learning Cross-lingual Representations for Event Coreference Resolution with Multi-view Alignment and Optimal Transport
Duy Phung, Hieu Minh Tran, Minh Van Nguyen and Thien Huu Nguyen

Multilingual and Multilabel Emotion Recognition using Virtual Adversarial Training
Vikram Gupta

Analyzing the Effects of Reasoning Types on Cross-Lingual Transfer Performance
Karthikeyan K, Aalok Sathe, Somak Aditya and Monojit Choudhury

Identifying the Importance of Content Overlap for Better Cross-lingual Embedding Mappings
Réka Cserháti and Gábor Berend

Thursday November 11, 2021 (continued)

On the Cross-lingual Transferability of Contextualized Sense Embeddings
Kiamehr Rezaee, Daniel Loureiro, Jose Camacho-Collados and Mohammad Taher Pilehvar

Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages
Kelechi Ogueji, Yuxin Zhu and Jimmy Lin

13:45–14:15

Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval
Xinyu Zhang, Xueguang Ma, Peng Shi and Jimmy Lin

14:15–14:45

VisualSem: a high-quality knowledge graph for vision and language
Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania and Iacer Calixto

14:45–15:45

Vyākarana: A Colorless Green Benchmark for Syntactic Evaluation in Indic Languages
Rajaswa Patil, Jasleen Dhillon, Siddhant Mahurkar, Saumitra Kulkarni, Manav Malhotra and Veeky Baths

Improving the Diversity of Unsupervised Paraphrasing with Embedding Outputs
Monisha Jegadeesan, Sachin Kumar, John Wieting and Yulia Tsvetkov

The Effectiveness of Intermediate-Task Training for Code-Switched Natural Language Understanding
Archiki Prasad, Mohammad Ali Rehan, Shreya Pathak and Preethi Jyothi

Shaking Syntactic Trees on the Sesame Street: Multilingual Probing with Controllable Perturbations
Ekaterina Taktasheva, Vladislav Mikhailov and Ekaterina Artemova

Multilingual Code-Switching for Zero-Shot Cross-Lingual Intent Prediction and Slot Filling
Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit and Huzefa Rangwala

Analysis of Zero-Shot Crosslingual Learning between English and Korean for Named Entity Recognition
Jongin Kim, Nayoung Choi, Seunghyun Lim, Jungwhan Kim, Soojin Chung, Hyunsoo Woo, Min Song and Jinho D. Choi

Regularising Fisher Information Improves Cross-lingual Generalisation
Asa Cooper Stickland and Iain Murray

DMix: Distance Constrained Interpolative Mixup
Ramit Sawhney, Megh Thakkar, Shrey Pandit, Debdoot Mukherjee and Lucie Flek

Sequence Mixup for Zero-Shot Cross-Lingual Part-Of-Speech Tagging
Megh Thakkar, Vishwa Shah, Ramit Sawhney and Debdoot Mukherjee

Well-Defined Morphology is Sentence-Level Morphology
Omer Goldman and Reut Tsarfaty

Cross-Lingual Training of Dense Retrievers for Document Retrieval
Peng Shi, Rui Zhang, He Bai and Jimmy Lin