ACL 2023

The Tenth Workshop on NLP for Similar Languages, Varieties and Dialects

Proceedings of the Workshop

May 5, 2023

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 978-1-959429-50-0

Preface

These proceedings include the 23 papers presented at the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), co-located with the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Both EACL and VarDial were held in Dubrovnik, Croatia, in a hybrid format, allowing participants to attend on-site or to participate virtually.

This edition marks VarDial's ten-year anniversary. We are pleased to see that the workshop continues to serve the community as the main venue for researchers interested in the computational processing of diatopic language variation. The papers accepted this year address a wide range of topics, such as corpus building, part-of-speech tagging, and machine translation. This volume once again showcases the great linguistic diversity that VarDial embodies, including work on dialects and varieties of many different languages, such as Arabic, Cantonese, Croatian, Finnish, German, Irish, Italian, Mandarin, Occitan, Serbian, and Spanish.

The VarDial evaluation campaign continues to be an essential part of the workshop. In VarDial 2023, three shared tasks were organized: Slot and intent detection for low-resource language varieties (SID4LR), Discriminating Between Similar Languages – True Labels (DSL-TL), and Discriminating Between Similar Languages – Speech (DSL-S). All three tasks were organized for the first time this year. This volume includes the system description papers prepared by the participating teams, as well as a report written by the task organizers summarizing the results and the findings of the evaluation campaign.

Finally, we would like to take this opportunity to thank all the shared task organizers and the participants for their hard work. We further thank the VarDial program committee members for being an important part of the workshop's success over these ten years.

The VarDial workshop organizers:

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri

http://sites.google.com/view/vardial-2023/

Organizing Committee

Organizers

Tommi Jauhiainen, University of Helsinki
Nikola Ljubešić, Jožef Stefan Institute
Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence
Yves Scherrer, University of Helsinki
Jörg Tiedemann, University of Helsinki
Marcos Zampieri, George Mason University

Program Committee

Program Committee

Noëmi Aepli, University of Zurich
Željko Agić, Unity Technologies
César Aguilar, Universidad Veracruzana
Laura Alonso Alemany, Universidad Nacional de Cordoba
Eric Atwell, University of Leeds
Jorge Baptista, University of Algarve
Eckhard Bick, University of Southern Denmark
Johannes Bjerva, Department of Computer Science, Aalborg University
Francis Bond, Palacký University
Aoife Cahill, Dataminr
David Chiang, University of Notre Dame
Paul Cook, University of New Brunswick
Jon Dehdari, Fidelity Investments
Liviu P. Dinu, University of Bucharest
Stefanie Dipper, Ruhr-Universität Bochum
Sascha Diwersy, Université Paul-Valéry Montpellier 3
Mark Dras, Macquarie University
Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute
Pablo Gamallo, CITIUS, University of Santiago de Compostela
Cyril Goutte, National Research Council Canada
Nizar Habash, New York University Abu Dhabi
Chu-ren Huang, The Hong Kong Polytechnic Universiy
Radu Tudor Ionescu, University of Bucharest
Surafel M. Lakew, Amazon.com, Inc
Ekaterina Lapshinova-koltunski, Stiftung Universität Hildesheim
Lung-hao Lee, National Central University
John Nerbonne, Albert-Ludwigs Universität Freiburg
Kai North, George Mason University
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
Petya Osenova, Sofia University St. Kl. Ohridskiand IICT-BAS
Santanu Pal, Wipro
Barbara Plank, LMU Munich
Taraka Rama, Walmart Global Tech
Francisco Manuel Rangel Pardo, Universitat Politècnica de València
Reinhard Rapp, University of Mainz
Paolo Rosso, Universitat Politècnica de València
Rachel Edita Roxas, Ideacorp
Fatiha Sadat, UQAM
Tanja Samardžić, University of Zurich
Kevin Scannell, Saint Louis University
Serge Sharoff, University of Leeds
Miikka Silfverberg, University of British Columbia
Kiril Simov, Artificial Intelligence and Language Technologies Department, IICT, Bulgarian Academy of Sciences
Milena Slavcheva, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences

Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences
Liling Tan, Amazon
Joel Tetreault, Dataminr
Francis Tyers, Indiana University
Rob Van Der Goot, IT University of Copenhagen
Pidong Wang, Google
Taro Watanabe, Nara Institute of Science and Technology
Çağrı Çöltekin, University of Tübingen

Keynote Talk: Bridging the Dialect Gap with Modular Transfer Learning?

Ivan Vulić
University of Cambridge
2023-05-05 14:00:00 –

Table of Contents

Program

Friday, May 5, 2023

09:00 - 09:10 *Opening remarks*

09:10 - 10:30 *Oral presentations*

Findings of the VarDial Evaluation Campaign 2023
Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer and Marcos Zampieri

Two-stage Pipeline for Multilingual Dialect Detection
Ankit Vaidya and Aditya Kane

Fine-Tuning BERT with Character-Level Noise for Zero-Shot Transfer to Dialects and Closely-Related Languages
Aarohi Srivastava and David Chiang

10:30 - 11:00 *Coffee break*

11:00 - 12:15 *Oral presentations*

Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection
Galo Castillo-lópez, Arij Riabi and Djamé Seddah

Optimizing the Size of Subword Vocabularies in Dialect Classification
Vani Kanjirangat, Tanja Samardžić, Ljiljana Dolamic and Fabio Rinaldi

Comparing and Predicting Eye-tracking Data of Mandarin and Cantonese
Junlin Li, Bo Peng, Yu-yin Hsu and Emmanuele Chersoni

12:15 - 12:40 *Poster Boosters I*

12:40 - 14:00 *Lunch break*

14:00 - 14:50 *Keynote Talk by Ivan Vulić: Bridging the Dialect Gap with Modular Transfer Learning?*

Friday, May 5, 2023 (continued)

14:50 - 15:40 *Round Table: VarDial in the Era of Large Language Models*

15:40 - 16:15 *Coffee break*

16:15 - 16:40 *Poster Boosters II*

16:40 - 18:00 *Poster Session*

Murreviikko - A Dialectologically Annotated and Normalized Dataset of Finnish Tweets
Olli Kuparinen

Exploring Enhanced Code-Switched Noising for Pretraining in Neural Machine Translation
Vivek Iyer, Arturo Oncevay and Alexandra Birch

Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages
Verena Blaschke, Hinrich Schütze and Barbara Plank

Temporal Domain Adaptation for Historical Irish
Oksana Dereza, Theodorus Fransen and John P. Mccrae

Variation and Instability in Dialect-Based Embedding Spaces
Jonathan Dunn

PALI: A Language Identification Benchmark for Perso-Arabic Scripts
Sina Ahmadi, Milind Agarwal and Antonios Anastasopoulos

Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora
Taja Kuzman, Peter Rupnik and Nikola Ljubešić

Reconstructing Language History by Using a Phonological Ontology. An Analysis of German Surnames
Hanna Fischer and Robert Engsterhold