EACL 2021

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects

Co-located with the 16th European Chapter of the Association for Computational Linguistics (EACL)

April 20, 2021

©2021 The 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 978-1-954085-12-1

Preface

These proceedings include the 16 papers presented at the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), which was co-located with the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL). VarDial and EACL were originally scheduled to take place in Kiev, Ukraine, but the COVID-19 situation eventually forced a switch to a virtual format.

We are happy to see that after many years, VarDial keeps serving the community as an important venue for researchers interested in the computational processing of diatopic language variation. The papers accepted this year deal with a wide variety of topics such as language identification, part-of-speech tagging, machine translation, distributional semantic models, and corpus similarity.

As in previous years, together with the workshop, we organized another iteration of the popular VarDial Evaluation Campaign with four shared tasks: Dravidian Language Identification (DLI), Romanian Dialect Identification (RDI), Social Media Variety Geolocation (SMG), and Uralic Language Identification (ULI). Each of these tasks addressed an important challenge in dialect and language identification. Eight teams prepared system description papers that are included in this volume, along with a report paper written by the campaign organizers summarizing the results and the main findings of the evaluation campaign.

Finally, we would like to take this opportunity to thank the VarDial Evaluation Campaign shared task organizers and the task participants for their hard work. We further thank our amazing VarDial program committee members for their thorough reviews. They have been a very important part of the workshop's success in these eight years.

The VarDial workshop organizers:

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

http://sites.google.com/view/vardial2021/

Organizers:

Marcos Zampieri - Rochester Institute of Technology (USA)
Preslav Nakov - Qatar Computing Research Institute, HBKU (Qatar)
Nikola Ljubešić - Jožef Stefan Institute (Slovenia) and University of Zagreb (Croatia)
Jörg Tiedemann - University of Helsinki (Finland)
Yves Scherrer - University of Helsinki (Finland)
Tommi Jauhiainen - University of Helsinki (Finland)

Program Committee:

Željko Agić (Corti, Denmark)
Cesar Aguilar (Pontifical Catholic University of Chile, Chile)
Laura Alonso y Alemany (University of Cordoba, Argentina)
Eric Atwell (University of Leeds, United Kingdom)
Jorge Baptista (University of Algarve and INESC-ID, Portugal)
Eckhard Bick (University of Southern Denmark, Denmark)
Johannes Bjerva (University of Copenhagen, Denmark)
Francis Bond (Nanyang Technological University, Singapore)
Aoife Cahill (Educational Testing Service, United States)
David Chiang (University of Notre Dame, United States)
Paul Cook (University of New Brunswick, Canada)
Çağrı Çöltekin (University of Tübingen)
Marta Costa-Jussà (Universitat Politècnica de Catalunya, Spain)
Jon Dehdari (Think Big Analytics, United States)
Liviu Dinu (University of Bucharest, Romania)
Stefanie Dipper (Ruhr University Bochum, Germany)
Sascha Diwersy (University of Montpellier, France)
Mark Dras (Macquarie University, Australia)
Tomaž Erjavec (Jožef Stefan Institute, Slovenia)
Pablo Gamallo (University of Santiago de Compostela, Spain)
Binyam Gebrekidan Gebre (Phillips Research, The Netherlands)
Cyril Goutte (National Research Council, Canada)
Nizar Habash (New York University Abu Dhabi, UAE)
Chu-Ren Huang (Hong Kong Polytechnic University, Hong Kong)
Radu Ionescu (University of Bucharest, Romania)
Jeremy Jancsary (Nuance Communications, Austria)
Tommi Jauhiainen (University of Helsinki, Finland)
Surafel Melaku Lakew (FBK , Italy)
Ekaterina Lapshinova-Koltunski (Saarland University, Germany)
Lung-Hao Lee (National Taiwan Normal University, Taiwan)
John Nerbonne (University of Groningen, Netherlands and University of Freiburg, Germany)
Kemal Oflazer (Carnegie-Mellon University in Qatar, Qatar)
Maciej Ogrodniczuk (IPAN, Polish Academy of Sciences, Poland)
Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
Santanu Pal (Saarland University, Germany)
Francisco Rangel (Autoritas Consulting, Spain)
Taraka Rama (University of North Texas, United States)

Reinhard Rapp (University of Mainz, Germany and University of Aix-Marsaille, France)
Paolo Rosso (Technical University of Valencia, Spain)
Rachel Edita O. Roxas (National University, Phillipines)
Fatiha Sadat (Université du Québec à Montréal (UQAM), Canada)
Tanja Samardžić (University of Zurich, Switzerland)
Kevin Scannell (Saint Louis University, United States)
Yves Scherrer (University of Helsinki, Finland)
Serge Sharoff (University of Leeds, United Kingdom)
Miikka Silfverberg (University of Helsinki, Finland)
Kiril Simov (Bulgarian Academy of Sciences, Bulgaria)
Milena Slavcheva (Bulgarian Academy of Sciences, Bulgaria)
Marco Tadić (University of Zagreb, Croatia)
Liling Tan (Rakuten Institute of Technology, Singapore)
Joel Tetreault (Dataminr, United States)
Francis Tyers (Indiana University, United States)
Taro Watanabe (Google Inc., Japan)
Pidong Wang (Google Inc., United States)

Table of Contents

Conference Program

Tuesday, April 20, 2021

10:30–10:40 ***Opening Session***

10:40–11:00 *Findings of the VarDial Evaluation Campaign 2021*
Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer and Marcos Zampieri

11:00–11:30 ***Coffee break***

11:30–12:30 ***Invited Talk by Jack Grieve***

Oral Presentations

12:30–12:45 *Hierarchical Transformer for Multilingual Machine Translation*
Albina Khusainova, Adil Khan, Adín Ramírez Rivera and Vitaly Romanov

12:45–13:00 *Regression Analysis of Lexical and Morpho-Syntactic Properties of Kiezdeutsch*
Diego Frassinelli, Gabriella Lapesa, Reem Alatrash, Dominik Schlechtweg and Sabine Schulte im Walde

13:00–13:15 *Representations of Language Varieties Are Reliable Given Corpus Similarity Measures*
Jonathan Dunn

13:15–13:30 *Whit's the Richt Pairt o Speech: PoS tagging for Scots*
Harm Lameris and Sara Stymne

13:30–15:00 ***Lunch break***

Tuesday, April 20, 2021 (continued)

15:00–16:00 Poster Session

Efficient Unsupervised NMT for Related Languages with Cross-Lingual Language Models and Fidelity Objectives
Rami Aly, Andrew Caines and Paula Buttery

Fine-tuning Distributional Semantic Models for Closely-Related Languages
Kushagra Bhatia, Divyanshu Aggarwal and Ashwini Vaidya

Discriminating Between Similar Nordic Languages
René Haas and Leon Derczynski

Naive Bayes-based Experiments in Romanian Dialect Identification
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

UnibucKernel: Geolocating Swiss German Jodels Using Ensemble Learning
Gaman Mihaela, Sebastian Cojocariu and Radu Tudor Ionescu

Optimizing a Supervised Classifier for a Difficult Language Identification Problem
Yves Bestgen

Comparing the Performance of CNNs and Shallow Models for Language Identification
Andrea Ceolin

Dialect Identification through Adversarial Learning and Knowledge Distillation on Romanian BERT
George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel and Traian Rebedea

Comparing Approaches to Dravidian Language Identification
Tommi Jauhiainen, Tharindu Ranasinghe and Marcos Zampieri

N-gram and Neural Models for Uralic Language Identification: NRC at VarDial 2021
Gabriel Bernier-Colborne, Serge Leger and Cyril Goutte

Social Media Variety Geolocation with geoBERT
Yves Scherrer and Nikola Ljubešić

Tuesday, April 20, 2021 (continued)

16:00–17:00 *Invited Talk by Katharina Kann*

17:00–17:15 *Closing Remarks*