COLING 2020

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects,

Co-located with the 28th International Conference on Computational Linguistics COLING'2020

VarDial '2020

December 13, 2020
Barcelona, Spain (Online)

ISBN 978-1-952148-47-7

Preface

These proceedings include the 27 papers presented at the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)[1], co-located with the 28th International Conference on Computational Linguistics (COLING). VarDial and COLING were scheduled to take place in Barcelona, Spain, but both were changed to a virtual format due to the COVID-19 outbreak.

We are glad to see that VarDial keeps growing in popularity, reaching its seventh edition. Moreover, this year, we received an all-time high number of regular submissions —21 papers —, and we accepted 15 of them to be presented at the workshop. These papers deal with various topics related to the processing of diatopic language variation in both text and speech. This volume includes papers on topics such as automatic speech recognition, corpus building, pre-processing, syntactic parsing, language identification, and machine translation, to name a few.

Diversity is innate to VarDial due to its focus on dialects and under-resourced languages. We are happy that the workshop continues to bring together researchers working on different languages, sharing ideas and contributing to advancing the state of the art of NLP for dialects, low-resource languages, and language varieties. This year, we accepted papers dealing with languages such as Armenian, Basque, German, Italian, Kurdish, and Occitan, as well as groups of dialects and low-resource languages from families such as Dravidian, Slavic, and Zaza-Gorani.

As in previous years, together with the workshop, we organized another iteration of the popular VarDial Evaluation Campaign with three shared tasks: Romanian Dialect Identification (RDI), Social Media Variety Geolocation (SMG), and Uralic Language Identification (ULI). These tasks addressed important challenges in dialect and language identification, attracting many teams who submitted runs across the three competitions. Eleven teams prepared system description papers that are included in this volume, along with a report paper summarizing the results and the main findings of the evaluation campaign written by the campaign organizers.

Finally, we would like to take this opportunity to thank the amazing VarDial program committee members for their thorough reviews. They have been playing a very important role in making the VarDial workshop series a success and we are fortunate to have them on board. We further thank the VarDial Evaluation Campaign shared task organizers and the participants for their hard work.

The VarDial workshop organizers:

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

`http://sites.google.com/view/vardial2020/`

[1] https://sites.google.com/view/vardial2020/home

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)

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)
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)
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)
Marko 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)

Invited Speaker:

Barbara Plank (IT University of Copenhagen, Denmark)

Table of Contents

Conference Program

Sunday, December 13, 2020

14:00–14:05 ***Opening Remarks***

14:05–14:20 *A Report on the VarDial Evaluation Campaign 2020*
Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer and Marcos Zampieri

14:30–15:30 *Invited Talk by Barbara Plank*

Oral presentations

16:00–16:15 *ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German*
Iuliia Nigmatulina, Tannon Kew and Tanja Samardzic

16:15–16:30 *LSDC - A comprehensive dataset for Low Saxon Dialect Classification*
Janine Siewert, Yves Scherrer, Martijn Wieling and Jörg Tiedemann

16:30–16:45 *Machine-oriented NMT Adaptation for Zero-shot NLP tasks: Comparing the Usefulness of Close and Distant Languages*
Amirhossein Tebbifakhr, Matteo Negri and Marco Turchi

16:45–17:00 *Character Alignment in Morphologically Complex Translation Sets for Related Languages*
Michael Gasser, Binyam Ephrem Seyoum and Nazareth Amlesom Kifle

Sunday, December 13, 2020 (continued)

17:30-18:30 Poster presentations

Bilingual Lexicon Induction across Orthographically-distinct Under-Resourced Dravidian Languages
Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor and John P. McCrae

Building a Corpus for the Zaza–Gorani Language Family
Sina Ahmadi

Dealing with dialectal variation in the construction of the Basque historical corpus
Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano and Ander Soraluze

Recycling and Comparing Morphological Annotation Models for Armenian Diachronic-Variational Corpus Processing
Chahan Vidal-Gorène, Victoria Khurshudyan and Anaïd Donabédian-Demopoulos

Neural Machine Translation for translating into Croatian and Serbian
Maja Popović, Alberto Poncelas, Marija Brkic and Andy Way

A Tokenization System for the Kurdish Language
Sina Ahmadi

Rediscovering the Slavic Continuum in Representations Emerging from Neural Models of Spoken Language Identification
Badr M. Abdullah, Jacek Kudera, Tania Avgustinova, Bernd Möbius and Dietrich Klakow

A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments
Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade and Jean Sibille

Vulgaris: Analysis of a Corpus for Middle-Age Varieties of Italian Language
Andrea Zugarini, Matteo Tiezzi and Marco Maggini

Towards Augmenting Lexical Resources for Slang and African American English
Alyssa Hwang, William R. Frey and Kathleen McKeown

Uralic Language Identification (ULI) 2020 shared task dataset and the Wanca 2017 corpora
Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen and Krister Lindén

Sunday, December 13, 2020 (continued)

Dialect Identification under Domain Shift: Experiments with Discriminating Romanian and Moldavian
Çağrı Çöltekin

Applying Multilingual and Monolingual Transformer-Based Models for Dialect Identification
Cristian Popa and Vlad Ștefănescu

HeLju@VarDial 2020: Social Media Variety Geolocation with BERT Models
Yves Scherrer and Nikola Ljubešić

A dual-encoding system for dialect classification
Petru Rebeja and Dan Cristea

Experiments in Language Variety Geolocation and Dialect Identification
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

Exploring the Power of Romanian BERT for Dialect Identification
George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel and Traian Rebedea

Combining Deep Learning and String Kernels for the Localization of Swiss German Tweets
Mihaela Gaman and Radu Tudor Ionescu

ZHAW-InIT - Social Media Geolocation at VarDial 2020
Fernando Benites, Manuela Hürlimann, Pius von Däniken and Mark Cieliebak

Discriminating between standard Romanian and Moldavian tweets using filtered character ngrams
Andrea Ceolin and Hong Zhang

Challenges in Neural Language Identification: NRC at VarDial 2020
Gabriel Bernier-Colborne and Cyril Goutte

Geolocation of Tweets with a BiLSTM Regression Model
Piyush Mishra

Sunday, December 13, 2020 (continued)

18:30–19:00 Discussion and Closing