COLING

Volume 29 (2022), No. 19

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2022)

The 29th International Conference on Computational Linguistics

October 12 - 17, 2022
Gyeongju, Republic of Korea

ISSN 2951-2093

Preface

These proceedings include the 13 papers presented at the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), co-located with the 29th International Conference on Computational Linguistics (COLING). Both COLING and VarDial were held in Gyeongju, South Korea, in a hybrid format, allowing all participants to either be present on-site or join virtually.

VarDial has now reached its ninth edition and continues serving 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 NLP tasks such as corpus building, part-of-speech tagging and machine translation, but also address more theoretical questions related to micro-scale variation, cognate detection, mutual intelligibility and dialectometry. We are happy to see such a diverse set of research papers advancing the state of the art of NLP for dialects, low-resource languages, and language varieties.

As in previous years, the evaluation campaign continues to be an essential part of the VarDial workshop. This year, three shared tasks were proposed: Identification of Languages and Dialects of Italy (ITDI), French Cross-Domain Dialect Identification (FDI), and Dialectal Extractive Question Answering (DialQA). All three tasks address important issues in dialect and language identification. This volume includes five system description papers prepared by the participating teams, as well as a report summarizing the results and findings of the evaluation campaign.

Finally, we would like to take this opportunity to thank the shared task organizers and the participants of the evaluation campaign 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 the past years.

The VarDial workshop organizers:

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

`http://sites.google.com/view/vardial-2022/`

Organizers:

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

Program Committee:

Željko Agić (Corti, Denmark)
César Aguilar (Universidad Veracruzana, Mexico)
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)
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)
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)
Surafel Melaku Lakew (FBK , Italy)
Ekaterina Lapshinova-Koltunski (Saarland University, Germany)
Lung-Hao Lee (National Central 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)
Barbara Plank (LMU Munich, Germany and ITU Copenhagen, Denmark)
Taraka Rama (University of North Texas, United States)
Francisco Rangel (Autoritas Consulting, Spain)
Reinhard Rapp (University of Mainz, Germany and University of Aix-Marseille, 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 British Columbia, Canada)
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)
Pidong Wang (Google Inc., United States)
Taro Watanabe (Google Inc., Japan)

Table of Contents

Conference Program

Sunday, October 20, 2022

10:00–10:10 ***Opening Session***

10:10–10:30 *Findings of the VarDial Evaluation Campaign 2022*
Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu and Yves Scherrer

10:30–10:45 *Transfer Learning Improves French Cross-Domain Dialect Identification: NRC @ VarDial 2022*
Gabriel Bernier-Colborne, Serge Leger and Cyril Goutte

10:45–11:00 *Is Encoder-Decoder Transformer the Shiny Hammer?*
Nat Gillin

11:00–11:30 ***Coffee break***

11:30–11:45 *Italian Language and Dialect Identification and Regional French Variety Detection using Adaptive Naive Bayes*
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

11:45–12:00 *The Curious Case of Logistic Regression for Italian Languages and Dialects Identification*
Giacomo Camposampiero, Quynh Anh Nguyen and Francesco Di Stefano

12:00–12:15 *Neural Networks for Cross-domain Language Identification. Phlyers @Vardial 2022*
Andrea Ceolin

12:15–13:15 ***Invited Talk by Tanja Samardžić (University of Zurich)***

13:15–14:30 ***Lunch break***

Phonetic, Semantic, and Articulatory Features in Assamese-Bengali Cognate Detection
14:30–15:00 Abhijnan Nath, Rahul Ghosh and Nikhil Krishnaswamy

15:00-15:15 *Annotating Norwegian language varieties on Twitter for Part-of-speech*
Petter Mæhlum, Andre Kåsen, Samia Touileb and Jeremy Barnes

15:15–15:45 *OcWikiDisc: a Corpus of Wikipedia Talk Pages in Occitan*
Aleksandra Miletic and Yves Scherrer

Sunday, October 20, 2022

15:45–16:00 *Low-Resource Neural Machine Translation: A Case Study of Cantonese*
Evelyn Kai-Yan Liu

16:00–16:30 ***Coffee break***

16:30–17:00 *Mapping Phonology to Semantics: A Computational Model of Cross-Lingual Spoken-Word Recognition*
Iuliia Zaitova, Badr Abdullah and Dietrich Klakow

17:00–17:15 *Social Context and User Profiles of Linguistic Variation on a Micro Scale*
Olga Kellert and Nicholas Hill Matlis

17:15–17:45 *dialectR: Doing Dialectometry in R*
Ryan Soh-Eun Shim and John Nerbonne

17:45–18:45 ***Invited Talk by Dong Nguyen (Utrecht University)***

18:45–19:00 ***Closing Remarks***