NAACL HLT 2019

The Workshop on NLP for Similar Languages, Varieties and Dialects

Proceedings of the Sixth Workshop

July 7, 2019
Minneapolis, USA

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

978-1-950737-11-6

Preface

This volume includes the 25 papers presented in the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), which was co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) and was held on June 7, 2019 in Minneapolis, USA.

This is the first time that VarDial is co-located with NAACL and the second time that the workshop is organized in North America. The previous five editions of the workshop were co-located with COLING (in 2014 in Dublin, Ireland; in 2016 in Osaka, Japan; and in 2018 in Santa Fe, USA), with RANLP (in 2015 in Hissar, Bulgaria), and with EACL (in 2017 in Valencia, Spain).

VarDial continues to be the main venue dedicated to the research on similar languages, varieties, and dialects within the CL/NLP community. We are happy to see that VarDial keeps growing, building on the success of the previous editions. This year we received 17 regular workshop submissions, and we accepted 10 papers, which were presented at the workshop. The accepted papers deal with various topics related to language variation such as cross-lingual annotation projection in part-of-speech tagging, machine translation between similar languages and dialects, and the processing of code-switched (or mixed) data, to name a few.

Together with the sixth edition of the workshop, we organized the third edition of the VarDial Evaluation Campaign, which featured five shared tasks. One shared task was a re-run from previous editions, the third German Dialect Identification (GDI), and we had four new tasks: Cross-lingual Morphological Analysis (CMA), Discriminating between Mainland and Taiwan variation of Mandarin Chinese (DMT), Moldavian vs. Romanian Cross-dialect Topic identification (MRC), and Cuneiform Language Identification (CLI). A total of 22 teams submitted official runs to one or more of the five shared tasks, and 14 system description papers appear in this volume along with a shared task report by the evaluation campaign and the task organizers.

Shared tasks have been organized since the workshop's first edition. Most of these tasks were on language and dialect identification, while a few others dealt with NLP tasks such as morphosyntactic tagging and cross-lingual dependency parsing. The focus of the language and dialect identification competitions at VarDial has always been on diatopic variation using synchronic contemporary data. This year, the CLI shared task included historical languages for the first time at VarDial, and it was the most popular shared task of the campaign, which demonstrates the interest of the community in this topic. To further respond to this interest, we included topics related to the diachronic/diatopic variation interplay in the call for papers as topics of interest for VarDial, e.g., philogenetic methods, and historical dialects.

We take this opportunity to thank the VarDial program committee for their thorough reviews. We further thank the VarDial Evaluation Campaign shared task organizers and the participants. Finally, we thank the workshop participants who presented regular research papers, for the valuable feedback and discussions.

The VarDial workshop organizers:

Marcos Zampieri, Preslav Nakov, Shervin Malmasi, Nikola Ljubešić, Jörg Tiedemann, and Ahmed Ali

Organizers:

Marcos Zampieri, University of Wolverhampton (UK)
Preslav Nakov, Qatar Computing Research Institute, HBKU (Qatar)
Shervin Malmasi, Amazon (United States)
Nikola Ljubešić, Jožef Stefan Institute (Slovenia)
Jörg Tiedemann, University of Helsinki (Finland)
Ahmed Ali, Qatar Computing Research Institute, HBKU (Qatar)

Program Committee:

Željko Agić, IT University of Copenhagen (Denmark)
Cesar Aguilar, Pontifical Catholic University of Chile (Chile)
Laura Alonso y Alemany, University of Cordoba (Argentina)
Eric Atwell, University of Leeds (UK)
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 (USA)
David Chiang, University of Notre Dame (USA)
Paul Cook, University of New Brunswick (Canada)
Marta Costa-Jussà, Universitat Politècnica de Catalunya (Spain)
Jon Dehdari, Think Big Analytics (USA)
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)
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)
Barbara Plank, IT University of Copenhagen (Denmark)
Francisco Rangel, Autoritas Consulting (Spain)
Taraka Rama, University of Oslo (Norway)
Reinhard Rapp, University of Mainz (Germany) and University of Aix-Marsaille (France)

Paolo Rosso, Technical University of Valencia (Spain)
Fatiha Sadat, Université du Québec à Montréal, UQAM (Canada)
Tanja Samardžić, University of Zurich (Switzerland)
Felipe Sánchez Martínez, Universitat d'Alacant (Spain)
Kevin Scannell, Saint Louis University (USA)
Yves Scherrer, University of Helsinki (Finland)
Serge Sharoff, University of Leeds (UK)
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, Grammarly (USA)
Francis Tyers, Indiana University (USA)
Taro Watanabe, Google Inc. (Japan)
Pidong Wang, Machine Zone Inc. (USA)

VarDial Evaluation Campaign and Shared Task Organizers:

Marcos Zampieri, University of Wolverhampton (UK) - VarDial Evaluation Campaign
Shervin Malmasi, Amazon (USA) - VarDial Evaluation Campaign
Yves Scherrer, University of Helsinki (Finland) - GDI Shared Task
Tanja Samardžić (University of Zurich, Switzerland) - GDI Shared Task
Francis Tyers Indiana University (USA) - CMA Shared Task
Miikka Silfverberg, University of Helsinki (Finland) - CMA Shared Task
Natalia Klyueva, The Hong Kong Polytechnic University (Hong Kong) - DMT Shared Task
Tung-Le Pan, The Hong Kong Polytechnic University (Hong Kong) - DMT Shared Task
Chu-Ren Huang, The Hong Kong Polytechnic University (Hong Kong) - DMT Shared Task
Radu Ionescu, University of Bucharest (Romania) - MRC Shared Task
Andrei Butnaru, University of Bucharest (Romania) - MRC Shared Task
Tommi Jauhiainen, University of Helsinki (Finland) - CLI Shared Task

Invited Speaker:

David Yarowsky, Johns Hopkins University

Table of Contents

Conference Program

Friday, June 7, 2019

9:15–9:30 ***Opening***

9:30–10:00 *A Report on the Third VarDial Evaluation Campaign*
Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardzic, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru and Tommi Jauhiainen

10:00–10:30 *Improving Cuneiform Language Identification with BERT*
Gabriel Bernier-Colborne, Cyril Goutte and Serge Leger

10:30–11:00 ***Coffee break***

11:00–11:30 *Joint Approach to Deromanization of Code-mixed Texts*
Rashed Rubby Riyadh and Grzegorz Kondrak

11:30–12:00 *Char-RNN for Word Stress Detection in East Slavic Languages*
Ekaterina Chernyak, Maria Ponomareva and Kirill Milintsevich

12:00–12:30 *Modeling Global Syntactic Variation in English Using Dialect Classification*
Jonathan Dunn

12:30–14:00 ***Lunch***

14:00–15:00 ***Invited talk — David Yarowsky (Johns Hopkins University): Massively Multilingual Translingual Knowledge Transfer***

15:00–15:30 *Language Discrimination and Transfer Learning for Similar Languages: Experiments with Feature Combinations and Adaptation*
Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen and Çağrı Çöltekin

15:30–16:00 ***Coffee break***

16:00–17:00 ***Poster Session***

Friday, June 7, 2019 (continued)

Variation between Different Discourse Types: Literate vs. Oral
Katrin Ortmann and Stefanie Dipper

Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)
Thazin Myint Oo, Ye Kyaw Thu and Khin Mar Soe

Language and Dialect Identification of Cuneiform Texts
Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola and Krister Lindén

Leveraging Pretrained Word Embeddings for Part-of-Speech Tagging of Code Switching Data
Fahad AlGhamdi and Mona Diab

Toward a deep dialectological representation of Indo-Aryan
Chundra Cathcart

Naive Bayes and BiLSTM Ensemble for Discriminating between Mainland and Taiwan Variation of Mandarin Chinese
Li Yang and Yang Xiang

BAM: A combination of deep and shallow models for German Dialect Identification.
Andrei M. Butnaru

The R2I_LIS Team Proposes Majority Vote for VarDial's MRC Task
Adrian-Gabriel Chifu

Initial Experiments In Cross-Lingual Morphological Analysis Using Morpheme Segmentation
Vladislav Mikhailov, Lorenzo Tosi, Anastasia Khorosheva and Oleg Serikov

Neural and Linear Pipeline Approaches to Cross-lingual Morphological Analysis
Çağrı Çöltekin and Jeremy Barnes

Ensemble Methods to Distinguish Mainland and Taiwan Chinese
Hai Hu, Wen Li, He Zhou, Zuoyu Tian, Yiwen Zhang and Liang Zou

SC-UPB at the VarDial 2019 Evaluation Campaign: Moldavian vs. Romanian Cross-Dialect Topic Identification
Cristian Onose, Dumitru-Clementin Cercel and Stefan Trausan-Matu

Friday, June 7, 2019 (continued)

Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models
Tommi Jauhiainen, Krister Lindén and Heidi Jauhiainen

Investigating Machine Learning Methods for Language and Dialect Identification of Cuneiform Texts
Ehsan Doostmohammadi and Minoo Nassajian

TwistBytes - Identification of Cuneiform Languages and German Dialects at VarDial 2019
Fernando Benites, Pius von Däniken and Mark Cieliebak

DTeam @ VarDial 2019: Ensemble based on skip-gram and triplet loss neural networks for Moldavian vs. Romanian cross-dialect topic identification
Diana Tudoreanu

Experiments in Cuneiform Language Identification
Gustavo Henrique Paetzold and Marcos Zampieri

17:00–17:30 *Comparing Pipelined and Integrated Approaches to Dialectal Arabic Neural Machine Translation*
Pamela Shapiro and Kevin Duh

17:30–18:00 *Cross-lingual Annotation Projection Is Effective for Neural Part-of-Speech Tagging*
Matthias Huck, Diana Dutka and Alexander Fraser

18:00–18:15 ***Closing Remarks***