COLING 2018

The 27th International Conference on Computational Linguistics

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'2018)

August 20, 2018
Santa Fe, New Mexico, USA

ISBN 978-1-948087-55-1

Introduction

VarDial is now a well-established workshop series that has been attracting researchers working on a wide range of topics related to (diatopic) linguistic variation. VarDial is currently in its 5th edition and since its very first edition it has been co-located with international NLP conferences such as COLING, EACL, and RANLP. This year, VarDial is co-located with COLING in Santa Fe, United States.

In five years we have seen VarDial become the main venue dedicated to research on similar languages, varieties, and dialects within the NLP community and 2018 was a particularly important year for VarDial. For the first time we organized five shared tasks, an all-time record, as part of the second VarDial Evaluation Campaign. The tasks organized this year were: Arabic Dialect Identification (ADI), German Dialect Identification (GDI), Morphosyntactic Tagging of Tweets (MTT), Discriminating between Dutch and Flemish in Subtitles (DSF) and Indo-Aryan Language Identification (ILI). The second edition of the campaign received a very positive response from the community with a total of 54 teams subscribed to participate in the five shared tasks, another all-time record for VarDial. 24 teams submitted official runs to one or more of the five shared tasks, and 22 system description papers appear in this volume along with a shared task report by the task organizers.

We further received 15 regular VarDial workshop papers, and we selected 9 of them to be presented at the workshop. The papers that appear in this volume reflect the wide range of interests related to language variation. This volume includes papers applying NLP methods to perform text normalization, identify false friends in closely-related languages, measure language distance between historical varieties of a pluricentric language, and translate between language varieties.

We take the 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 also thank participants who presented regular research papers, for the valuable feedback and discussions.

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

Workshop Organisers

Marcos Zampieri (University of Wolverhampton, United Kingdom)
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)
Shervin Malmasi (Harvard Medical School, USA)
Ahmed Ali (Qatar Computing Research Institute, HBKU, Qatar)

VarDial Evaluation Campaign Organisers

(ADI) Arabic Dialect Identification

Ahmed Ali (Qatar Computing Research Institute, HBKU, Qatar)
Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Suwon Shon (Massachusetts Institute of Technology, United States)
James Glass (Massachusetts Institute of Technology, United States)

(GDI) German Dialect Identification

Yves Scherrer (University of Helsinki, Finland)
Tanja Samardžić (University of Zurich, Switzerland)

(MTT) Morphosyntactic Tagging of Tweets

Nikola Ljubešić (Jožef Stefan Institute, Slovenia and University of Zagreb, Croatia)
Jörg Tiedemann (University of Helsinki, Finland)

(DFS) Discriminating between Dutch and Flemish in Subtitles

Chris van der Lee (Tilburg University, The Netherlands)
Stef Grondelaers (Radboud University, The Netherlands)
Nelleke Oostdijk (Radboud University, The Netherlands)
Dirk Speelman (University of Leuven, Belgium)
Antal van den Bosch (Meertens Institute and Radboud University, The Netherlands)

(ILI) Indo-Aryan Language Identification

Ritesh Kumar (Bhim Rao Ambedkar University, India)
Bornini Lahiri (Jadavpur University, India)
Mayank Jain (Jawaharlal Nehru University, India)

General Organization

Marcos Zampieri (University of Wolverhampton, United Kingdom)
Shervin Malmasi (Harvard Medical School, USA)

Programme Committee

Željko Agić (IT University of Copenhagen, Denmark)

Cesar Aguilar (Pontifical Catholic University of Chile, Chile)
Laura Alonso y Alemany (University of Cordoba, Argentina)
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)
Mikel L. Forcada (Universitat d'Alacant, Spain)
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)
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, United States)
Yves Scherrer (University of Helsinki, Finland)
Serge Sharoff (University of Leeds, United Kingdom)
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 (Grammarly, United States)
Francis Tyers (Higher School of Economics, Russia)
Duško Vitas (University of Belgrade, Serbia)
Taro Watanabe (Google Inc., Japan)
Pidong Wang (Machine Zone Inc., United States)

Table of Contents

Conference Program

Monday, August 20, 2018

9:00–9:10 ***Opening***

9:10-9:30 *Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign*
Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri and Mayank Jain

9:30–10:00 *Encoder-Decoder Methods for Text Normalization*
Massimo Lusetti, Tatyana Ruzsics, Anne Göhring, Tanja Samardžić and Elisabeth Stark

10:00–10:30 *A High Coverage Method for Automatic False Friends Detection for Spanish and Portuguese*
Santiago Castro, Jairo Bonanata and Aiala Rosá

10:30–11:00 ***Coffee break***

11:00–11:30 *Sub-label dependencies for Neural Morphological Tagging – The Joint Submission of University of Colorado and University of Helsinki for VarDial 2018*
Miikka Silfverberg and Senka Drobac

11:30–12:00 *Part of Speech Tagging in Luyia: A Bantu Macrolanguage*
Kenneth Steimel

12:00–12:30 *Tübingen-Oslo Team at the VarDial 2018 Evaluation Campaign: An Analysis of N-gram Features in Language Variety Identification*
Çağrı Çöltekin, Taraka Rama and Verena Blaschke

12:30–14:00 ***Lunch***

14:00–14:30 *Iterative Language Model Adaptation for Indo-Aryan Language Identification*
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

Monday, August 20, 2018 (continued)

14:30–15:40 Invited talk - Timothy Baldwin (University of Melbourne)

Language and the Shifting Sands of Domain, Space and Time (Invited Talk)
Timothy Baldwin

15:40–15:50 ***Closing Remarks***

15:50–16:20 ***Coffee break***

16:20–18:00 Poster Session

UnibucKernel Reloaded: First Place in Arabic Dialect Identification for the Second Year in a Row
Andrei Butnaru and Radu Tudor Ionescu

Varying image description tasks: spoken versus written descriptions
Emiel van Miltenburg, Ruud Koolen and Emiel Krahmer

Transfer Learning for British Sign Language Modelling
Boris Mocialov, Helen Hastie and Graham Turner

Paraphrastic Variance between European and Brazilian Portuguese
Anabela Barreiro and Cristina Mota

Character Level Convolutional Neural Network for Arabic Dialect Identification
Mohamed Ali

Neural Network Architectures for Arabic Dialect Identification
Elise Michon, Minh Quang Pham, Josep Crego and Jean Senellart

HeLI-based Experiments in Discriminating Between Dutch and Flemish Subtitles
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

Monday, August 20, 2018 (continued)

Measuring language distance among historical varieties using perplexity. Application to European Portuguese.
José Ramom Pichel Campos, Pablo Gamallo and Iñaki Alegria

Comparing CRF and LSTM performance on the task of morphosyntactic tagging of non-standard varieties of South Slavic languages
Nikola Ljubešić

Computationally efficient discrimination between language varieties with large feature vectors and regularized classifiers
Adrien Barbaresi

Character Level Convolutional Neural Network for German Dialect Identification
Mohamed Ali

Discriminating between Indo-Aryan Languages Using SVM Ensembles
Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal and Liviu P. Dinu

IIT (BHU) System for Indo-Aryan Language Identification (ILI) at VarDial 2018
Divyanshu Gupta, Gourav Dhakad, Jayprakash Gupta and Anil Kumar Singh

Exploring Classifier Combinations for Language Variety Identification
Tim Kreutz and Walter Daelemans

Identification of Differences between Dutch Language Varieties with the VarDial2018 Dutch-Flemish Subtitle Data
Hans van Halteren and Nelleke Oostdijk

Birzeit Arabic Dialect Identification System for the 2018 VarDial Challenge
Rabee Naser and Abualsoud Hanani

Twist Bytes - German Dialect Identification with Data Mining Optimization
Fernando Benites, Ralf Grubenmann, Pius von Däniken, Dirk von Grünigen, Jan Deriu and Mark Cieliebak

STEVENDU2018's system in VarDial 2018: Discriminating between Dutch and Flemish in Subtitles
Steven Du and Yuan Yuan Wang

Using Neural Transfer Learning for Morpho-syntactic Tagging of South-Slavic Languages Tweets
Sara Meftah, Nasredine Semmar, Fatiha Sadat and Stephan Raaijmakers