VarDial 2017

Fourth Workshop
on NLP for Similar Languages, Varieties and Dialects
(VarDial'2017)

Proceedings of the Workshop

April 3, 2017
Valencia, Spain

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 978-1-945626-43-2

Preface

VarDial is a well-established series of workshops held annually and co-located with top-tier international NLP conferences. Previous editions of VarDial were VarDial'2014, which was co-located with COLING'2014, LT4VarDial'2015, which was held together with RANLP'2015, and finally VarDial'2016 co-located with COLING'2016. The great interest of the community has made possible the fourth edition of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'2017), co-located with EACL'2017 in Valencia, Spain.

The VarDial series has attracted researchers working on a wide range of topics related to linguistic variation such as building and adapting language resources for language varieties and dialects, creating language technology and applications that make use of language closeness, and exploiting existing resources in a related language or a language variety.

We believe that this is a very timely series of workshops, as research in language variation is much needed in today's multi-lingual world, where several closely-related languages, language varieties, and dialects are in daily use, not only as spoken colloquial language but also in written media, e.g., in SMS, chats, and social networks. Language resources for these varieties and dialects are sparse and extending them could be very labor-intensive. Yet, these efforts can often be reduced by making use of pre-existing resources and tools for related, resource-richer languages.

As part of the workshop, we organized the first VarDial evaluation campaign with four shared tasks: Discriminating between Similar Languages (DSL), Arabic Dialect Identification (ADI), German Dialect Identification (GDI), and Cross-Lingual Parsing (CLP). The campaign received a very positive response from the community. A total of 28 teams subscribed to participate in the four shared tasks, 19 of them submitted official runs, and 15 of the latter also wrote system description papers, which appear in this volume along with a shared task report by the task organizers.

We further received 14 regular VarDial workshop papers, and we selected nine 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. We include papers applying NLP tools to perform dialect analysis, to study mutual intelligibility and diatopic variation in historical corpora, as well as core NLP tasks and applications such as dialect and similar language identification, adaptation of POS taggers, and machine translation between similar languages and dialects.

We take the opportunity to thank the VarDial program committee and the additional reviewers for their thorough reviews. We further thank the VarDial Evaluation Campaign participants, as well as the participants with regular research papers, for the valuable feedback and discussions.

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

Workshop Organisers

Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Marcos Zampieri (University of Cologne, Germany)
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 - Shared Task Organisers

Marcos Zampieri (University of Cologne, Germany)
Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Shervin Malmasi (Harvard Medical School, USA)
Nikola Ljubešić (Jožef Stefan Institute, Slovenia, and University of Zagreb, Croatia)
Jörg Tiedemann (University of Helsinki, Finland)
Ahmed Ali (Qatar Computing Research Institute, HBKU, Qatar)
Yves Scherrer (University of Geneva, Switzerland)
Noëmi Aepli (University of Zürich, Switzerland)

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)
Tim Baldwin (The University of Melbourne, Australia)
Jorge Baptista (University of Algarve and INESC-ID, Portugal)
Eckhard Bick (University of Southern Denmark, 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 (Saarland University and DFKI, Germany)
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)
Binyam Gebrekidan Gebre (Phillips Research, Holland)
Cyril Goutte (National Research Council, Canada)
Nizar Habash (New York University Abu Dhabi, UAE)
Chu-Ren Huang (Hong Kong Polytechnic University, Hong Kong)
Jeremy Jancsary (Nuance Communications, Austria)
Lung-Hao Lee (National Taiwan Normal University, Taiwan)
Marco Lui (Rome2Rio Ltd., Australia)

Teresa Lynn (Dublin City University, Ireland)
John Nerbonne (University of Groningen, Netherlands and University of Freiburg, Germany)
Graham Neubig (Nara Institute of Science and Technology, Japan)
Kemal Oflazer (Carnegie-Mellon University in Qatar, Qatar)
Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences, Poland)
Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
Santanu Pal (Saarland University, Germany)
Reinhard Rapp (University of Mainz, Germany and University of Aix-Marsaille, France)
Paolo Rosso (Polytechnic University of Valencia, Spain)
Fatiha Sadat (Université du Québec à Montréal, Canada)
Tanja Samardžić (University of Zürich, Switzerland)
Felipe Sánchez Martínez (Universitat d'Alacant, Spain)
Kevin Scannell (Saint Louis University, USA)
Yves Scherrer (University of Geneva, Switzerland)
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)
Elke Teich (Saarland University, Germany)
Joel Tetreault (Grammarly, USA)
Francis Tyers (UiT Norgga árktalaš universitehta, Norway)
Duško Vitas (University of Belgrade, Serbia)
Taro Watanabe (Google Inc., Japan)
Pidong Wang (Machine Zone Inc., USA)

Additional Reviewers

Yves Bestgen (Université Catholique de Louvain, Belgium)
Johannes Bjerva (University of Groningen, Netherlands)
Alina Maria Ciobanu (University of Bucharest, Romania)
Çağrı Çöltekin (University of Tübingen, Germany)
Marcelo Criscuolo (University of São Paulo, Brazil)
Abualsoud Hanani (Birzeit University, Palestine)
Pablo Gamallo (University of Santiago de Compostela, Spain)
Helena Gómez-Adorno (Instituto Politécnico Nacional, Mexico)
Radu Tudor Ionescu (University of Bucharest, Romania)
Tommi Jauhiainen (University of Helsinki, Finland)
Martin Kroon (University of Groningen, Netherlands)
Peter Makavorov (University of Zürich, Switzerland)
David Marecek (Charles University Prague, Czech Republic)
Ilia Markov (Instituto Politecnico Nacional, Mexico)
Maria Medvedeva (Saarland University, Germany)
Sergiu Nisioi (University of Bucharest, Romania)
Stephen Taylor (Fitchburg State University, USA)
Zdeněk Žabokrtský (Charles University Prague, Czech Republic)
Daniel Zeman (Charles University Prague, Czech Republic)

Table of Contents

Conference Program

Monday, April 3, 2017

9:30–9:40 ***Opening***

09:40–10:00 *Findings of the VarDial Evaluation Campaign 2017*
Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer and Noëmi Aepli

10:00–10:30 *Dialectometric analysis of language variation in Twitter*
Gonzalo Donoso and David Sanchez

10:30–11:00 *Computational analysis of Gondi dialects*
Taraka Rama, Çağrı Çöltekin and Pavel Sofroniev

11.00–11.30 ***Coffee break***

11:30–12:00 *Investigating Diatopic Variation in a Historical Corpus*
Stefanie Dipper and Sandra Waldenberger

12.00–13.00 **Invited talk - Paolo Rosso (Polytechnic University of Valencia, Spain)**

Author Profiling at PAN: from Age and Gender Identification to Language Variety Identification (invited talk)
Paolo Rosso

13.00–14.30 ***Lunch***

14.30–16.00 Poster Session

The similarity and Mutual Intelligibility between Amharic and Tigrigna Varieties
Tekabe Legesse Feleke

Why Catalan-Spanish Neural Machine Translation? Analysis, comparison and combination with standard Rule and Phrase-based technologies
Marta R. Costa-jussà

Kurdish Interdialect Machine Translation
Hossein Hassani

Twitter Language Identification Of Similar Languages And Dialects Without Ground Truth
Jennifer Williams and Charlie Dagli

Multi-source morphosyntactic tagging for spoken Rusyn
Yves Scherrer and Achim Rabus

Identifying dialects with textual and acoustic cues
Abualsoud Hanani, Aziz Qaroush and Stephen Taylor

Evaluating HeLI with Non-Linear Mappings
Tommi Jauhiainen, Krister Lindén and Heidi Jauhiainen

A Perplexity-Based Method for Similar Languages Discrimination
Pablo Gamallo, Jose Ramom Pichel and Iñaki Alegria

Improving the Character Ngram Model for the DSL Task with BM25 Weighting and Less Frequently Used Feature Sets
Yves Bestgen

Discriminating between Similar Languages with Word-level Convolutional Neural Networks
Marcelo Criscuolo and Sandra Maria Aluisio

Cross-lingual dependency parsing for closely related languages - Helsinki's submission to VarDial 2017
Jörg Tiedemann

Discriminating between Similar Languages Using a Combination of Typed and Untyped Character N-grams and Words
Helena Gomez, Ilia Markov, Jorge Baptista, Grigori Sidorov and David Pinto

Monday, April 3, 2017 (continued)

Tübingen system in VarDial 2017 shared task: experiments with language identification and cross-lingual parsing
Çağrı Çöltekin and Taraka Rama

When Sparse Traditional Models Outperform Dense Neural Networks: the Curious Case of Discriminating between Similar Languages
Maria Medvedeva, Martin Kroon and Barbara Plank

German Dialect Identification in Interview Transcriptions
Shervin Malmasi and Marcos Zampieri

CLUZH at VarDial GDI 2017: Testing a Variety of Machine Learning Tools for the Classification of Swiss German Dialects
Simon Clematide and Peter Makarov

Arabic Dialect Identification Using iVectors and ASR Transcripts
Shervin Malmasi and Marcos Zampieri

Discriminating between Similar Languages using Weighted Subword Features
Adrien Barbaresi

16.00–16.30 ***Coffee break***

16:30–17:00 *Exploring Lexical and Syntactic Features for Language Variety Identification*
Chris van der Lee and Antal van den Bosch

17:00–17:30 *Learning to Identify Arabic and German Dialects using Multiple Kernels*
Radu Tudor Ionescu and Andrei Butnaru

17:30–18:00 *Slavic Forest, Norwegian Wood*
Rudolf Rosa, Daniel Zeman, David Mareček and Zdeněk Žabokrtský

18.00 ***Closing remarks***