VarDial 3

Third Workshop on NLP for Similar Languages, Varieties and Dialects

Proceedings of the Workshop

December 12, 2016
Osaka, Japan

ISBN978-4-87974-716-7

Preface

VarDial is a well-established series of workshops, attracting researchers working on a range of topics related to the study of linguistic variation, e.g., on building language resources for language varieties and dialects or in creating language technology and applications that make use of language closeness and exploit existing resources in a related language or a language variant.

The research presented in the two previous editions, namely VarDial'2014, which was co-located with COLING'2014, and LT4VarDial'2015, which was held together with RANLP'2015, focused on topics such as machine translation between closely related languages, adaptation of POS taggers and parsers for similar languages and language varieties, compilation of corpora for language varieties, spelling normalization, and finally discrimination between and identification of similar languages. The latter was also the topic of the DSL shared task, held in conjunction with the workshop.

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.

Examples of closely-related language varieties include the different variants of Spanish in Latin America, the Arabic dialects in North Africa and the Middle East, German in Germany, Austria and Switzerland, French in France and in Belgium, etc. Examples of pairs of related languages include Swedish-Norwegian, Bulgarian-Macedonian, Serbian-Bosnian, Spanish-Catalan, Russian-Ukrainian, Irish-Gaelic Scottish, Malay-Indonesian, Turkish–Azerbaijani, Mandarin-Cantonese, Hindi–Urdu, etc.

This great interest of the community has made possible the third edition of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial'2016), co-located with COLING'2016.

As part of the workshop, we organized the third edition of the Discriminating between Similar Languages (DSL) shared task, which offered an opportunity for researchers and developers to investigate the performance of computational methods to distinguishing between closely-related languages and language varieties, thus bridging an important gap for language identification. For the first time, the DSL task was divided into two sub-tasks: Sub-task 1 focusing on similar languages and language varieties, and Sub-task 2 on Arabic dialect identification.

The third edition of the DSL shared task received a very positive response from the community and a record number of participants. A total of 37 teams subscribed to participate in the DSL shared task, 24 of them submitted official runs, and 20 of the latter also wrote system description papers, which appear in this volume along with a shared task report by the task organizers. These numbers represent a substantial increase in participation compared to the 2014 and 2015 editions of the DSL task.

We further received 13 regular VarDial workshop papers, and we selected nine of them to be presented at the workshop and to appear in this volume.

Given the aforementioned numbers, we consider the workshop a success, and thus we are organizing a fourth edition in 2017, which will be co-located with EACL'2017.

We take the opportunity to thank the VarDial program committee and the additional reviewers for their thorough reviews, and the DSL Shared Task participants, as well as the participants with regular research papers, for the valuable feedback and discussions. We further thank our invited speakers, Mona Diab and Robert Östling, for presenting their interesting work at the workshop.

The organizers: Preslav Nakov, Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Shervin Malmasi

Organisers

Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Marcos Zampieri (University of Cologne, Germany)
Liling Tan (Singapore University of Technology and Design, and Saarland University, Germany)
Nikola Ljubešić (Jožef Stefan Institute, Slovenia, and University of Zagreb, Croatia)
Jörg Tiedemann (University of Helsinki, Finland)
Shervin Malmasi (Harvard Medical School, USA)

DSL Shared Task Organisers

Marcos Zampieri (University of Cologne, Germany)
Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Shervin Malmasi (Harvard Medical School, USA)
Liling Tan (Singapore University of Technology and Design, and Saarland University, Germany)
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)

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 (Institute for Infocomm Research, Singapore)
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 (Macquire 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)
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)
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

Johannes Bjerva (University of Groningen, Netherlands)
Marc Franco Salvador (Polytechnic University of Valencia, Spain)
Aleksander Wawer (Institute of Computer Science, Polish Academy of Sciences, Poland)

Invited Speakers

Mona Diab (George Washington University, USA)
Robert Östling (University of Helsinki, Finland)

Table of Contents

Conference Program

Monday, December 12, 2016

9:00–9:10 ***Opening***

9.10–9.30 *Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task*
Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali and Jörg Tiedemann

9.30–10.00 *Discriminating Similar Languages with Linear SVMs and Neural Networks*
Çağrı Çöltekin and Taraka Rama

10.00–10.30 *LSTM Autoencoders for Dialect Analysis*
Taraka Rama and Çağrı Çöltekin

10.30–11.00 *The GW/LT3 VarDial 2016 Shared Task System for Dialects and Similar Languages Detection*
Ayah Zirikly, Bart Desmet and Mona Diab

11.00–12.00 **Invited talk 1**

Processing Dialectal Arabic: Exploiting Variability and Similarity to Overcome Challenges and Discover Opportunities
Mona Diab

12.00–14.00 ***Lunch***

14.00–14.30 *Language Related Issues for Machine Translation between Closely Related South Slavic Languages*
Maja Popović, Mihael Arcan and Filip Klubička

14.30–15.00 *Romanized Berber and Romanized Arabic Automatic Language Identification Using Machine Learning*
Wafia Adouane, Nasredine Semmar and Richard Johansson

Monday, December 12, 2016 (continued)

15.00–16.00 Invited talk 2

How Many Languages Can a Language Model Model?
Robert Östling

16.00–16.30 ***Coffee break***

16.30–18.00 Poster Session

Automatic Detection of Arabicized Berber and Arabic Varieties
Wafia Adouane, Nasredine Semmar, Richard Johansson and Victoria Bobicev

Automatic Verification and Augmentation of Multilingual Lexicons
Maryam Aminian, Mohamed Al-Badrashiny and Mona Diab

Faster Decoding for Subword Level Phrase-based SMT between Related Languages
Anoop Kunchukuttan and Pushpak Bhattacharyya

Subdialectal Differences in Sorani Kurdish
Shervin Malmasi

Enlarging Scarce In-domain English-Croatian Corpus for SMT of MOOCs Using Serbian
Maja Popović, Kostadin Cholakov, Valia Kordoni and Nikola Ljubešić

Arabic Dialect Identification in Speech Transcripts
Shervin Malmasi and Marcos Zampieri

DSL Shared Task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation–Maximization and Chunk-based Language Model
Ondřej Herman, Vit Suchomel, Vít Baisa and Pavel Rychlý

Byte-based Language Identification with Deep Convolutional Networks
Johannes Bjerva

Classifying ASR Transcriptions According to Arabic Dialect
Abualsoud Hanani, Aziz Qaroush and Stephen Taylor

Monday, December 12, 2016 (continued)

UnibucKernel: An Approach for Arabic Dialect Identification Based on Multiple String Kernels
Radu Tudor Ionescu and Marius Popescu

A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects
Yonatan Belinkov and James Glass

HeLI, a Word-Based Backoff Method for Language Identification
Tommi Jauhiainen, Krister Lindén and Heidi Jauhiainen

ASIREM Participation at the Discriminating Similar Languages Shared Task 2016
Wafia Adouane, Nasredine Semmar and Richard Johansson

Comparing Two Basic Methods for Discriminating Between Similar Languages and Varieties
Pablo Gamallo, Iñaki Alegria, José Ramom Pichel and Manex Agirrezabal

Advances in Ngram-based Discrimination of Similar Languages
Cyril Goutte and Serge Léger

Discrimination between Similar Languages, Varieties and Dialects using CNN- and LSTM-based Deep Neural Networks
Chinnappa Guggilla

Language and Dialect Discrimination Using Compression-Inspired Language Models
Paul McNamee

Arabic Language WEKA-Based Dialect Classifier for Arabic Automatic Speech Recognition Transcripts
Areej Alshutayri, Eric Atwell, Abdulrahman Alosaimy, James Dickins, Michael Ingleby and Janet Watson

An Unsupervised Morphological Criterion for Discriminating Similar Languages
Adrien Barbaresi

QCRI @ DSL 2016: Spoken Arabic Dialect Identification Using Textual Features
Mohamed Eldesouki, Fahim Dalvi, Hassan Sajjad and Kareem Darwish

Tuning Bayes Baseline for Dialect Detection
Hector-Hugo Franco-Penya and Liliana Mamani Sanchez

Vanilla Classifiers for Distinguishing between Similar Languages
Sergiu Nisioi, Alina Maria Ciobanu and Liviu P. Dinu