COLING 2014

The 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects

Proceedings of the Workshop

August 23, 2014
Dublin, Ireland

ISBN 978-1-873769-39-3
Proceedings of VarDial: Applying NLP Tools to Similar Languages, Varieties and Dialects
Marcos Zampieri, Liling Tan, Nikola Ljubešić and Jörg Tiedemann (eds.)

Introduction

The interest in language resources and computational models for the study of similar languages, varieties and dialects has been growing substantially in the last few years. The first edition of the Workshop on Applying NLP tools to similar languages, varieties and dialects (VarDial) confirms the interest in the topic.

Within the NLP community, the impact of language variation in the development of language resources and NLP applications has been explored in recent years with experiments in different directions. For example, automatic classification or identification of closely related languages such as in Huang and Lee (2008) and Tiedemann and Ljubešić (2012); corpus-driven studies focusing on lexical variation between varieties such as the one by Piersman et al. (2010) or Ljubešić and Fišer (2013); and finally, the adaptation of language models in the context of machine translation such as in Nakov and Tiedemann (2012).

Together with the VarDial workshop we organized the Discriminating between Similar Languages (DSL) shared task. Discriminating between similar languages and language varieties is one of the bottlenecks of state-of-the-art language identification and it has been topic of a number of papers published in the last years. The DSL shared task provided a dataset to evaluate system's performance on discriminating 13 different languages in 6 groups of languages.

The 18 papers that appear in this volume deal with different NLP tasks and applications such as parsing, morphological analysis, part-of-speech tagging, language identification and speech recognition. The VarDial workshop received 18 submissions and 12 of them are published in this volume. The DSL shared task received 22 inscriptions and 8 final submissions. Five system description papers plus the DSL shared task report appear in this volume.

We take this opportunity to thank the VarDial program committee who thoroughly reviewed all papers; the DSL shared task participants for valuable feedback and discussions; and the COLING organizers for their support, specially Jennifer Foster who replied promptly to all our inquiries.

Marcos, Liling, Nikola and Jörg
VarDial Organizers

Organizers

Marcos Zampieri, Saarland University, Germany
Liling Tan, Saarland University, Germany
Nikola Ljubešić, University of Zagreb, Croatia
Jörg Tiedemann, Uppsala University, Sweden

Program Committee

Željko Agić, University of Potsdam, Germany
Jorge Baptista, University of Algarve and INESC-ID, Portugal
Francis Bond, Nanyang Technological University, Singapore
Aoife Cahill, Educational Testing Service, USA
Paul Cook, University of Melboune, Australia
Liviu Dinu, University of Bucarest, Romania
Stefanie Dipper, Ruhr University Bochum, Germany
Sascha Diwersy, University of Cologne, Germany
Tomaž Erjavec, Jozef Stefan Institute, Slovenia
Mikel L. Forcada, Universitat d'Alacant, Spain
Binyam Gebrekidan Gebre, Max Planck Institute for Psycholinguistics, Holland
Nitin Indurkhya, University of New South Wales, Australia
Jeremy Jancsary, Nuance Communications, Austria
Marco Lui, University of Melbourne, Australia
Preslav Nakov, Qatar Computing Research Institute, Qatar
Santanu Pal, Saarland University, Germany
Sebastian Padó, University of Stuttgart, Germany
Reinhard Rapp, University of Mainz, Germany and University of Aix-Marsaille, France
Felipe Sánchez Martínez, University of Alicante, Spain
Kevin Scanell, Saint Louis University, USA
Yves Scherrer, University of Geneva, Switzerland
Serge Sharoff, Leeds University, United Kingdom
Kiril Simov, Bulgarian Academy of Sciences, Bulgaria
Elke Teich, Saarland University, Germany
Joel Tetreault, Yahoo! Labs, USA
Francis Tyers, UiT Norgga Árktalaš Universitehta, Norway
Cristina Vertan, University of Hamburg, Germany
Torsten Zesch, University of Duisburg-Essen, Germany

Table of Contents

Conference Program

Saturday, August 23, 2014

9:15–9:30	Opening Remarks
09:30–10:00	*Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations* Chu-Ren Huang, Jingxia Lin, Menghan JIANG and Hongzhi Xu
10:00–10:30	*Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic* Maria Sukhareva and Christian Chiarcos
10:30–11:00	*Pos-tagging different varieties of Occitan with single-dialect resources* Marianne Vergez-Couret and Assaf Urieli
11:00–11:30	Coffee Break
11:30–12:00	*Unsupervised adaptation of supervised part-of-speech taggers for closely related languages* Yves Scherrer
12:00–12:30	*Morphological Disambiguation and Text Normalization for Southern Quechua Varieties* Annette Rios Gonzales and Richard Alexander Castro Mamani
12:30–14:00	Lunch
14:00–14:30	*The Varitext platform and the Corpus des variétés nationales du français (CoVaNa-FR) as resources for the study of French from a pluricentric perspective* Sascha Diwersy
14:30–15:00	*A Report on the DSL Shared Task 2014* Marcos Zampieri, Liling Tan, Nikola Ljubešić and Jörg Tiedemann
15:00–15:30	Coffee Break

Saturday, August 23, 2014 (continued)

Poster Session

15:30–17:00 *Employing Phonetic Speech Recognition for Language and Dialect Specific Search*
Corey Miller, Rachel Strong, Evan Jones and Mark Vinson

15:30–17:00 *Part-of-Speech Tag Disambiguation by Cross-Linguistic Majority Vote*
Noëmi Aepli, Ruprecht von Waldenfels and Tanja Samardžić

15:30–17:00 *Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging*
Nora Hollenstein and Noëmi Aepli

15:30–17:00 *Automatically building a Tunisian Lexicon for Deverbal Nouns*
Ahmed Hamdi, Nuria Gala and Alexis Nasr

15:30–17:00 *Statistical Morph Analyzer (SMA++) for Indian Languages*
Saikrishna Srirampur, Ravi Chandibhamar and Radhika Mamidi

15:30–17:00 *Improved Sentence-Level Arabic Dialect Classification*
Christoph Tillmann, Saab Mansour and Yaser Al-Onaizan

15:30–17:00 *Using Maximum Entropy Models to Discriminate between Similar Languages and Varieties*
Jordi Porta and José-Luis Sancho

15:30–17:00 *Exploring Methods and Resources for Discriminating Similar Languages*
Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook and Timothy Baldwin

15:30–17:00 *The NRC System for Discriminating Similar Languages*
Cyril Goutte, Serge Léger and Marine Carpuat

15:30–17:00 *Experiments in Sentence Language Identification with Groups of Similar Languages*
Ben King, Dragomir Radev and Steven Abney

15:30–17:00 *A Simple Baseline for Discriminating Similar Languages*
Matthew Purver