Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects

Proceedings of the Workshop

September 10, 2015
Hissar, Bulgaria

Joint Workshop on Language Technology
for Closely Related Languages, Varieties and Dialects
associated with THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING 2015

PROCEEDINGS

Hissar, Bulgaria
10 September 2015

ISBN 978-954-452-031-1

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Introduction

A large number of closely related language varieties and dialects are in daily use, not only as spoken colloquial languages but also in some written media, e.g., in SMS, chats, and social networks. Language resources for these varieties and dialects are sparse and building 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.

Recent interest in language resources and technology for closely related languages, varieties and dialects has led to previous editions of the LT4CloseLang workshop at RANLP2013 and EMNLP2014, and of the VarDial workshop at COLING2014. Both the LT4CloseLang and the VarDial workshops have attracted a lot of research interest, which indicated that there was need for further activities. Thus, this year we decided to join forces between these two workshops and to organize a joint workshop, LT4VarDial, aiming to bring together researchers interested in building language resources for language varieties or dialects and in creating language technology that makes use of language closeness and exploits existing resources in a related language or a language variant.

As part of the workshop, we organized the second edition of the DSL Shared Task on Discriminating between Similar Languages. The first edition was held in conjunction with VarDial, aiming to distinguish between closely related languages and language varieties, thus filling the research gap in fine-grained language identification, which was previously perceived as a solved task. Yet, DSL remains a challenge for state-of-the-art language identification. The attention received from the research community and the feedback provided by the participants of the first edition motivated us to organize this Second DSL Shared Task, where we made two important changes compared to the first edition. First, in order to simulate a real-world language identification scenario, we included in the testing dataset some languages that were not present in the training dataset. Moreover, we included a second test set, where we substituted the named entities with placeholders to make the task more challenging and less dependent on the text topic and domain.

A total of 24 teams subscribed to participate in the shared task, 10 of them submitted official runs, and 8 of the latter also wrote system description papers. These numbers represent a slight increase in participation compared to the 2014 edition, which attracted 22 teams, 8 submissions, and 5 system description papers.

Overall, 12 papers are published in this volume. Nine papers were about the DSL shared task (8 system descriptions and the shared task overview), and three regular workshop papers.

Given the above numbers, we consider the workshop a success, and we take the opportunity to thank the LT4VarDial program committee for their professional and thorough reviews, and the DSL Shared Task participants for the valuable feedback and discussions. We further thank our invited speakers and our panelists for sharing with us their thought-provoking opinions on topics of interest to the workshop.

The workshop organizers: *Preslav Nakov, Marcos Zampieri, Petya Osenova, Liling Tan, Cristina Vertan, Nikola Ljubešić, and Jörg Tiedemann*

Workshop Organizers

Preslav Nakov, Qatar Computing Research Institute, HBKU (Qatar)
Marcos Zampieri, Saarland University and DFKI (Germany)
Petya Osenova, Bulgarian Academy of Sciences (Bulgaria)
Liling Tan, Saarland University (Germany)
Cristina Vertan, University of Hamburg (Germany)
Nikola Ljubešić, University of Zagreb (Croatia)
Jörg Tiedemann, University of Uppsala (Sweden)

DSL Shared Task Organizers

Liling Tan, Saarland University (Germany)
Marcos Zampieri, Saarland University and DFKI (Germany)
Nikola Ljubešić, University of Zagreb (Croatia)
Jörg Tiedemann, University of Uppsala (Sweden)
Preslav Nakov, Qatar Computing Research Institute, HBKU (Qatar)

Program Committee

Željko Agić (University of Copenhagen, Denmark)
Laura Alonso y Alemany (Univeristy of Cordoba, Argentina)
Jorge Baptista (University of Algarve and INESC-ID, Portugal)
Eckhard Bick (University of Southern Denmark)
Francis Bond (Nanyang Technological University, Singapore)
Aoife Cahill (Educational Testing Service, United States)
José Castaño (University of Buenos Aires, Argentina)
Paul Cook (University of New Brunswick, Canada)
Marta Costa-Jussà (Institute for Infocomm Research, Singapore)
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)
Maria Gavrilidou (ILSP, Greece)
Binyam Gebrekidan Gebre (Max Planck Institute for Psycholinguistics, Holland)
Nizar Habash (Columbia University, USA)
Barry Haddow (University of Edinburgh, UK)
Chu-Ren Huang (Hong Kong Polytechnic University, Hong Kong)
Nitin Indurkhya (University of New South Wales, Australia)
Jeremy Jancsary (Nuance Communications, Austria)
Marco Lui (University of Melbourne, Australia)
Vladislav Kuboň (Charles University Prague, Czech Republic)
Shervin Malmasi (Macquarie University, Australia)
Graham Neubig (Nara Institute of Science and Technology, Japan)
John Nerbonne (University of Groningen, Netherlands)
Kemal Oflazer (Carnegie-Mellon University, Qatar)
Maciej Ogrodniczuk (IPAN, Polish Academy of Sciences, Poland)
Santanu Pal (Saarland University, Germany)

Reinhard Rapp (University of Mainz, Germany and University of Aix-Marseille, France)
Laurent Romary (INRIA, France)
Kevin Scanell (Saint Louis University, USA)
Yves Scherrer (University of Geneva, Switzerland)
Serge Sharoff (University of Leeds, United Kingdom)
Kiril Simov (Bulgarian Academy of Sciences, Bulgaria)
Milena Slavcheva (Bulgarian Academy of Sciences)
Marko Tadić (University of Zagreb, Croatia)
Elke Teich (Saarland University, Germany)
Joel Tetreault (Yahoo! Labs, USA)
Francis Tyers (UiT Norgga árktalaš universitehta, Norway)
Duško Vitas (University of Belgrade, Serbia)
Pidong Wang (National University of Singapore, Singapore)
Taro Watanabe (NICT, Japan)

Additional Reviewers

Maja Miličević
Tanja Samardžić

Invited Speakers

Leon Derczynski (Sheffield University, UK)
Eckhard Bick (University of Southern Denmark, Denmark)

Panelists

Marcos Zampieri (Saarland University and DFKI, Germany) – moderator
Cyril Goutte (National Research Council, Canada)
Marc Franco-Salvador (Universitat Politècnica de València, Spain)
Nikola Ljubešić (University of Zagreb, Croatia)
Tanja Samardžić (University of Zurich, Switzerland)

Table of Contents

Workshop Program

Thursday, September 10, 2015

Session 1 (chair: Marcos Zampieri)

9:00–9:30 ***Welcome and Overview of the DSL Shared Task***

Overview of the DSL Shared Task 2015
Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann and Preslav Nakov

9:30–10:30 ****INVITED TALK***: Handling and Mining Linguistic Variation in UGC*
Leon Derczynski

10:30–11:00 *Distributed Representations of Words and Documents for Discriminating Similar Languages*
Marc Franco-Salvador, Paolo Rosso and Francisco Rangel

11:00–11:30 ***Coffee Break***

Session 2 (chair: Petya Osenova)

11:30–12:00 *Joint Bayesian Morphology Learning for Dravidian Languages*
Arun Kumar, Lluís Padró and Antoni Oliver

12:00–12:30 *Use of Transformation-Based Learning in Annotation Pipeline of Igbo, an African Language*
Ikechukwu Onyenwe, Mark Hepple, Chinedu Uchechukwu and Ignatius Ezeani

12:30–14:00 ***Lunch Break***

Thursday, September 10, 2015 (continued)

Session 3 (chair: Nikola Ljubešić)

14:00-15:00 ****INVITED TALK*** WikiTrans: Swedish-Danish Machine Translation in a Constraint Grammar Framework*
Eckhard Bick

15:00–16:00 Poster Session

Language Identification using Classifier Ensembles
Shervin Malmasi and Mark Dras

Discriminating Similar Languages with Token-Based Backoff
Tommi Jauhiainen, Heidi Jauhiainen and Krister Lindén

NLEL UPV Autoritas Participation at Discrimination between Similar Languages (DSL) 2015 Shared Task
Raül Fabra Boluda, Francisco Rangel and Paolo Rosso

Discriminating between Similar Languages Using PPM
Victoria Bobicev

Comparing Approaches to the Identification of Similar Languages
Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa and Josef van Genabith

A Two-level Classifier for Discriminating Similar Languages
Judit Ács, László Grad-Gyenge and Thiago Bruno Rodrigues de Rezende Oliveira

Experiments in Discriminating Similar Languages
Cyril Goutte and Serge Leger

16:00–16:30 ***Coffee Break***

Thursday, September 10, 2015 (continued)

Session 4 (chair: Nikola Ljubešić)

16:30–17:00 *Building Monolingual Word Alignment Corpus for the Greater China Region*
fan xu, Xiongfei Xu, Mingwen Wang and Maoxi Li

17:00–18:00 Panel Discussion: Marcos Zampieri (moderator), Cyril Goutte, Marc Franco-Salvador, Nikola Ljubešić, and Tanja Samardžić (panelists)