LT4CloseLang 2014

Proceedings of the EMNLP'2014 Workshop:

Language Technology for Closely Related Languages and Language Variants

October 29, 2014
Doha, Qatar

Production and Manufacturing by
Taberg Media Group AB
Box 94, 562 02 Taberg
Sweden

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 978-1-937284-96-1

Introduction

Recent initiatives in language technology have led to the development of at least minimal language processing toolkits for all EU-official languages as well as for languages with a large number of speakers worldwide such as Chinese and Arabic. This is a big step towards the automatic processing and/or extraction of information, especially from official documents and newspapers, where the standard, literary language is used.

Apart from those official languages, a large number of dialects or closely-related language variants are in daily use, not only as spoken colloquial languages but also in some written media, e.g., in SMS, chats, and social networks. Building language resources and tools for them from scratch is expensive, but the efforts can often be reduced by making use of pre-existing resources and tools for related, resource-richer languages.

Examples of closely-related language variants 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, Dutch in the Netherlands and Flemish 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, and many other.

The workshop aims to bring together researchers interested in building language technology applications that make use of language closeness to exploit existing resources in a related language or a language variant. A previous version of this workshop, organised at RANLP 2013, attracted a lot of research interest, showing the need for further activities.

We received 20 submissions and we selected 11 papers for presentation. The papers cover the following general NLP topics: Parsing, Variety and Adaptation, and Machine Translation.

We would like to thank our reviewers for the professional and in-time reviewing!

Preslav Nakov, Petya Osenova and Cristina Vertan

Program Co-Chairs and Organizers:

Preslav Nakov, Qatar Computing Research Institute
Petya Osenova, Sofia University and Bulgarian Academy of Sciences
Cristina Vertan, University of Hamburg

Program Committee:

Laura Alonso y Alemany (University of Cordoba, Argentina)
César Antonio Aguilar (Pontificia Universidad Católica de Chile, Santiago de Chile, Chile)
José Castaño (University of Buenos Aires, Argentina)
David Chiang (University of Southern California, USA)
Marta Costa-Jussà (Institute for Infocomm Research, Singapore)
Walter Daelemans (University of Antwerp, Belgium)
Kareem Darwish (Qatar Computing Research Institute, Qatar)
Tomaz Erjavec (Jozef Stefan Institute, Slovenia)
Maria Gavrilidou (ILSP, Greece)
Francisco Guzman (Qatar Computing Research Institute, Qatar)
Barry Haddow (University of Edinburgh, UK)
Nizar Habash (Columbia University, USA)
Walther v. Hahn (University of Hamburg,Germany)
Cvetana Krstev (University of Belgrade, Serbia)
Vladislav Kubon (Charles University Prague, Czech Republic)
Thang Luong Minh (Stanford University, USA)
John Nerbonne (University of Groningen, Netherlands)
Graham Neubig (Nara Institute of Science and Technology, Japan)
Kemal Oflazer (Carnegie-Mellon University, Qatar)
Maciej Ogrodniczuk (IPAN, Polish Academy of Sciences, Poland)
Slav Petrov (Google, New York, USA)
Stefan Riezler (University of Heidelberg, Germany)
Laurent Romary (INRIA, France)
Hassan Sajjad (Qatar Computing Research Institute, Qatar)
Kiril Simov (Bulgarian Academy of Sciences)
Milena Slavcheva (Bulgarian Academy of Sciences)
Marco Tadic (University of Zagreb, Croatia)
Jörg Tiedemann (Uppsala University, Sweden)
Dusko Vitas (University of Belgrade, Serbia)
Stephan Vogel (Qatar Computing Research Institute, Qatar)
Pidong Wang (National University of Singapore, Singapore)
Taro Watanabe (NICT, Japan)

Keynote Speakers:

Nizar Habash (New York University Abu Dhabi, UAE)
Slav Petrov (Google, USA)

Panelists:

Houda Bouamor (Carnegie Mellon University, Qatar)
Kareem Darwish (Qatar Computing Research Institute, Qatar)
Vladislav Kubon (Charles University in Prague, Czech Republic)
Wolfgang Maier (University of Dusseldorf, Germany)
Kemal Oflazer (Carnegie Mellon University, Qatar)

Table of Contents

Conference Program

Wednesday, October 29, 2014

Opening Session and Invited Talk 1

8:50–9:00 *Opening Remarks*
The organizers

9:00–10:00 *INVITED TALK 1: Computational Processing of Arabic Dialects*
Nizar Habash

Session 1: Parsing

10:00–10:20 *Learning from a Neighbor: Adapting a Japanese Parser for Korean Through Feature Transfer Learning*
Hiroshi Kanayama, Youngja Park, Yuta Tsuboi and Dongmook Yi

10:20–10:40 *Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets*
Željko Agić, Jörg Tiedemann, Danijela Merkler, Simon Krek, Kaja Dobrovoljc and Sara Moze

10:40–11:00 ***Coffee Break***

Session 2: Variety and Adaptation

11:00–11:20 *Language variety identification in Spanish tweets*
Wolfgang Maier and Carlos Gómez-Rodríguez

11:20–11:40 *Exploiting Language Variants Via Grammar Parsing Having Morphologically Rich Information*
Qaiser Abbas

11:40–12:00 *Adapting Predicate Frames for Urdu PropBanking*
Riyaz Ahmad Bhat, Naman Jain, Ashwini Vaidya, Martha Palmer, Tafseer Ahmed Khan, Dipti Misra Sharma and James Babani

12:00–12:20 *Measuring Language Closeness by Modeling Regularity*
Javad Nouri and Roman Yangarber

12:20–14:00 ***Lunch***

Wednesday, October 29, 2014 (continued)

Invited Talk 2

14:00–15:00 *INVITED TALK 2: Towards Universal Syntactic Processing of Natural Language*
Slav Petrov

Session 3: Machine Translation I

15:00–15:20 *Proper Name Machine Translation from Japanese to Japanese Sign Language*
Taro Miyazaki, Naoto Kato, Seiki Inoue, Shuichi Umeda, Makiko Azuma, Nobuyuki Hiruma and Yuji Nagashima

15:20–15:40 *Exploring cross-language statistical machine translation for closely related South Slavic languages*
Maja Popović and Nikola Ljubešić

15:40–16:00 ***Coffee Break***

Session 4: Machine Translation II

16:00–16:20 *Exploring System Combination approaches for Indo-Aryan MT Systems*
Karan Singla, Anupam Singh, Nishkarsh Shastri, Megha Jhunjhunwala, Srinivas Bangalore and Dipti Misra Sharma

16:20–16:40 *A Comparison of MT Methods for Closely Related Languages: a Case Study on Czech - Slovak Language Pair*
Vladislav Kubon and Jernej Vicic

16:40–17:00 *Handling OOV Words in Dialectal Arabic to English Machine Translation*
Maryam Aminian, Mahmoud Ghoneim and Mona Diab

Wednesday, October 29, 2014 (continued)

Closing Session

17:00–18:00 *Panel*
Panelists: Houda Bouamor, Kareem Darwish, Vladislav Kubon, Wolfgang Maier, Kemal Oflazer

18:00–18:10 *Closing Remarks*
The organizers