EMNLP 2011

DIALECTS2011

Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties

July 31, 2011
Edinburgh, Scotland, UK

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 978-1-937284-17-6 / 1-937284-17-4

Introduction

Language varieties (and specifically dialects) are a primary means of expressing a person's social affiliation and identity. Hence, computer systems that can adapt to the user by displaying a familiar socio-cultural identity are expected to raise the acceptance within certain contexts and target groups dramatically. Although the currently prevailing statistical paradigm has made possible major achievements in many areas of natural language processing, the applicability of the available methods is generally limited to major languages / standard varieties, to the exclusion of dialects or varieties that substantially differ from the standard.

While there are considerable initiatives dealing with the development of language resources for minor languages, and also reliable methods to handle accents of a given language, i.e., for applications like speech synthesis or recognition, the situation for dialects still calls for novel approaches, methods and techniques to overcome or circumvent the problem of data scarcity, but also to enhance and strengthen the standing that language varieties and dialects have in natural language processing technologies, as well as in interaction technologies that build upon the former.

What made us think that a such a workshop would be a fruitful enterprise was our conviction that only joint efforts of researchers with expertise in various disciplines can bring about progress in this field. We therefore aimed in our call to invite and bring together colleagues that deal with topics ranging from machine learning algorithms and active learning, machine translation between language varieties or dialects, speech synthesis and recognition, to issues of orthography, annotation and linguistic modelling.

The 2011 Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties (DIALECTS 2011) is the first workshop to be held on this rather interdisciplinary topic. The workshop received seventeen submissions, out of which six were accepted as oral presentations (long papers) and three as posters (short papers). These papers represent interesting work from almost all the scientific fields that were mentioned in the call as being necessary to contribute to the common goal.

In addition to the submitted papers we are happy to welcome Burr Settles as our invited speaker to give a keynote talk on the topic of using multiple machine learning strategies to facilitate rapid development of NLP tools for new/rare languages/dialects. We hope that this gathering and the proceedings will help to promote and to advance the topic this workshop is centered around. We would like to thank all the authors who submitted their work for consideration. We are also especially grateful to the members of the program committee and the additional reviewers for their insightful and detailed reviews.

Jeremy Jancsary, Friedrich Neubarth, and Harald Trost

Workshop Organizers

Organizers:

Jeremy Jancsary (OFAI, Vienna, Austria)
Friedrich Neubarth (OFAI, Vienna, Austria)
Harald Trost (Medical University Vienna, Austria)

Program Committee:

Gérard Bailly (GIPSA-LAB, CNRS Grenoble, France)
Nick Campbell (CLCS, Trinity College Dublin, Ireland)
Martine Grice (IfL, Phonetik Köln, Germany)
Gholamreza Haffari (BC Cancer Research Center, Vancouver, Canada)
Inmaculada Hernaez Rioja (Univ. of the Basque Country UPV/EHU, Spain)
Philipp Koehn (ILCC, Univ. of Edinburgh, UK)
Michael Pucher (ftw, Vienna, Austria)
Milan Rusko (SAS, Slovak Academy of Sciences, Slovakia)
Kevin Scannell (Dept. of Mathematics and Computer Science, Saint Louis Univ., USA)
Yves Scherrer (LATL, Université de Genève, Switzerland)
Beat Siebenhaar (Institut für Germanistik, Univ. of Leipzig, Germany)

Additional Reviewers:

Johannes Matiasek (OFAI, Vienna, Austria)
Gerard de Melo (Max-Planck-Inst. f. Informatik, Germany/Microsoft Research Cambridge, UK)
Eva Navas Cordón (Univ. of the Basque Country UPV/EHU, Spain)

Invited Speaker:

Burr Settles (Carnegie Mellon University, USA)

Table of Contents

Dialect Translation: Integrating Bayesian Co-segmentation Models with Pivot-based SMT
Michael Paul, Andrew Finch, Paul R. Dixon and Eiichiro Sumita 1

Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation
Wael Salloum and Nizar Habash 10

PaddyWaC: A Minimally-Supervised Web-Corpus of Hiberno-English
Brian Murphy and Egon W. Stemle 22

Syntactic transformations for Swiss German dialects
Yves Scherrer 30

Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus
Mans Hulden, Iñaki Alegria, Izaskun Etxeberria and Montse Maritxalar 39

Modeling of Stylistic Variation in Social Media with Stretchy Patterns
Philip Gianfortoni, David Adamson and Carolyn P. Rosé 49

Adapting Slovak ASR for native Germans speaking Slovak
Štefan Beňuš, Miloš Cerňak, Sakhia Darjaa, Milan Rusko and Marián Trnka 60

Phone set selection for HMM-based dialect speech synthesis
Michael Pucher, Nadja Kerschhofer-Puhalo and Dietmar Schabus 65

WordNet.PT global – Extending WordNet.PT to Portuguese varieties
Palmira Marrafa, Raquel Amaro and Sara Mendes 70

Conference Program

Sunday, July 31, 2011

09:00–09:10 Opening

09:10–10:10 Invited talk by Burr Settles: ”Combining Learning Strategies to Make the Most of Language Resources”

10:10–10:30 Open discussion

10:30–11:00 coffee break

11:00–11:30 *Dialect Translation: Integrating Bayesian Co-segmentation Models with Pivot-based SMT*
Michael Paul, Andrew Finch, Paul R. Dixon and Eiichiro Sumita

11:30–12:00 *Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation*
Wael Salloum and Nizar Habash

12:00–12:30 *PaddyWaC: A Minimally-Supervised Web-Corpus of Hiberno-English*
Brian Murphy and Egon W. Stemle

12:30–14:00 lunch break

14:00–14:40 Poster flash and presentation

14:40–15:10 *Syntactic transformations for Swiss German dialects*
Yves Scherrer

15:10–15:40 *Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus*
Mans Hulden, Iñaki Alegria, Izaskun Etxeberria and Montse Maritxalar

15:40–16:10 coffee break

16:10–16:40 *Modeling of Stylistic Variation in Social Media with Stretchy Patterns*
Philip Gianfortoni, David Adamson and Carolyn P. Rosé

16:40–17:00 Final discussion

Sunday, July 31, 2011 (continued)

Poster presentations:

Adapting Slovak ASR for native Germans speaking Slovak
Štefan Beňuš, Miloš Cerňak, Sakhia Darjaa, Milan Rusko and Marián Trnka

Phone set selection for HMM-based dialect speech synthesis
Michael Pucher, Nadja Kerschhofer-Puhalo and Dietmar Schabus

WordNet.PT global – Extending WordNet.PT to Portuguese varieties
Palmira Marrafa, Raquel Amaro and Sara Mendes