EMNLP 2014

First Workshop on Computational Approaches to Code Switching

Proceedings of the Workshop

October 25, 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

Code-switching (CS) is the phenomenon by which multilingual speakers switch back and forth between their common languages in written or spoken communication. CS is pervasive in informal text communications such as news groups, tweets, blogs, and other social media of multilingual communities. Such genres are increasingly being studied as rich sources of social, commercial and political information. Apart from the informal genre challenge associated with such data within a single language processing scenario, the CS phenomenon adds another significant layer of complexity to the processing of the data. Efficiently and robustly processing CS data presents a new frontier for our NLP algorithms on all levels. The goal of this workshop is to bring together researchers interested in exploring these new frontiers, discussing state of the art research in CS, and identifying the next steps in this fascinating research area.

The workshop program includes exciting papers discussing new approaches for CS data and the development of linguistic resources needed to process and study CS. We received a total of 17 regular workshop submissions of which we accepted eight for publication (47% acceptance rate), five of them as workshop talks and three as posters. The accepted workshop submissions cover a wide variety of language combinations from languages such as English, Hindi, Bengali, Turkish, Dutch, German, Italian, Romansh, Mandarin, Dialectical Arabic and Modern Standard Arabic. Although most papers focus on some kind of social media data, there is also work on more formal genres, such as that from the Canadian Hansard.

Another component of the workshop is the First Shared Task on Language Identification of CS Data. The shared task focused on social media and included four language pairs: Mandarin-English, Modern Standard Arabic-Dialectal Arabic, Nepali-English, and Spanish-English. We received a total of 42 system runs from seven different teams. Each team submitted a shared task paper describing their system. All shared task systems will be presented during the workshop poster session and a subset of them will also present a talk.

We would like to thank all authors who submitted their contributions to this workshop and all shared task participants for taking on the challenge of language identification in code switched data. We also thank the program committee members for their help in providing meaningful reviews. Lastly, we thank the EMNLP 2014 organizers for the opportunity to put together this workshop.

See you all in Qatar, see ypu al in Qatar at EMNLP 2014!
Workshop co-chairs,

Mona Diab
Julia Hirschberg
Pascale Fung
Thamar Solorio

Workshop Co-Chairs:

Mona Diab, George Washington University
Julia Hirschberg, Columbia University
Pascale Fung, Hong Kong University of Science and Technology
Thamar Solorio, University of Houston

Program Committee:

Steven Abney, University of Michigan
Laura Alonso i Alemany, Universidad Nacional de Córdoba
Elabbas Benmamoun, University of Illinois at Urbana-Champaign
Steven Bethard, University of Alabama at Birmingham
Rakesh Bhatt, University of Illinois at Urbana-Champaign
Agnes Bolonyia, NC State University
Barbara Bullock, University of Texas at Austin
Amitava Das, University of North Texas
Suzanne Dikker, New York University
Björn Gambäck, Norwegian Universities of Science and Technology
Nizar Habash, Columbia University
Aravind Joshi, University of Pennsylvania
Ben King, University of Michigan
Constantine Lignos, University of Pennsylvania
Yang Liu, University of Texas at Dallas
Suraj Maharjan, University of Alabama at Birmingham
Mitchell P. Marcus, University of Pennsylvania
Cecilia Montes-Alcala, Georgia Institute of Technology
Raymond Mooney, University of Texas at Austin
Borja Navarro Colorado, Universidad de Alicante
Owen Rambow, Columbia University
Yves Scherrer, Université de Genève
Chilin Shih, University of Illinois at Urbana-Champaign
Jacqueline Toribio, University of Texas at Austin
Rabih Zbib, BBN Technologies

Table of Contents

Workshop Program

Saturday, October 25, 2014

Session 1: Workshop talks

09:00–09:10 *Welcome Remarks*
The organizers

09:10–09:30 *Foreign Words and the Automatic Processing of Arabic Social Media Text Written in Roman Script*
Ramy Eskander, Mohamed Al-Badrashiny, Nizar Habash and Owen Rambow

09:30–09:50 *Code Mixing: A Challenge for Language Identification in the Language of Social Media*
Utsab Barman, Amitava Das, Joachim Wagner and Jennifer Foster

09:50–10:10 *Detecting Code-Switching in a Multilingual Alpine Heritage Corpus*
Martin Volk and Simon Clematide

10:10–10:30 *Exploration of the Impact of Maximum Entropy in Recurrent Neural Network Language Models for Code-Switching Speech*
Ngoc Thang Vu and Tanja Schultz

10:30–11:00 ***Coffee Break***

Session 2: Workshop Talks and Shared Task Systems

11:00–11:20 *Predicting Code-switching in Multilingual Communication for Immigrant Communities*
Evangelos Papalexakis, Dong Nguyen and A. Seza Doğruöz

11:20–11:40 *Twitter Users #CodeSwitch Hashtags! #MoltoImportante #wow*
David Jurgens, Stefan Dimitrov and Derek Ruths

11:40–11:50 *Overview for the First Shared Task on Language Identification in Code-Switched Data*
Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang and Pascale Fung

11:50–12:10 *Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System*
Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali and Monojit Choudhury

Saturday, October 25, 2014 (continued)

12:10–12:30 *The CMU Submission for the Shared Task on Language Identification in Code-Switched Data*
Chu-Cheng Lin, Waleed Ammar, Lori Levin and Chris Dyer

12:30–14:00 ***Lunch break***

Session 3: Shared Task and Next Steps

14:00–14:20 *Language Identification in Code-Switching Scenario*
Naman Jain and Riyaz Ahmad Bhat

14:20–14:40 *AIDA: Identifying Code Switching in Informal Arabic Text*
Heba Elfardy, Mohamed Al-Badrashiny and Mona Diab

14:40–15:00 *The IUCL+ System: Word-Level Language Identification via Extended Markov Models*
Levi King, Eric Baucom, Timur Gilmanov, Sandra Kübler, Dan Whyatt, Wolfgang Maier and Paul Rodrigues

15:00–15:30 *Panel Discussion: Next Steps in CS Research*
Group Discussion

15:30–16:00 ***Coffee Break (Posters set up time)***

Session 4: Poster Session

16:00–17:30 *Workshop and Shared Task Posters*
Multiple presenters

Mixed Language and Code-Switching in the Canadian Hansard
Marine Carpuat

"I am borrowing ya mixing ?" An Analysis of English-Hindi Code Mixing in Facebook
Kalika Bali, Jatin Sharma, Monojit Choudhury and Yogarshi Vyas

DCU-UVT: Word-Level Language Classification with Code-Mixed Data
Utsab Barman, Joachim Wagner, Grzegorz Chrupała and Jennifer Foster

Saturday, October 25, 2014 (continued)