EMNLP 2016

Second Workshop on Computational Approaches to Code Switching

Proceedings of the Workshop

November 1, 2016
Austin, Texas, USA

Order copies of this and other ACL proceedings from:

Curran Associates
57 Morehouse Lane
Red Hook, New York 12571
USA
Tel: +1-845-758-0400
Fax: +1-845-758-2633
curran@proceedings.com

ISBN 978-1-945626-28-9

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 12 regular workshop submissions of which we accepted nine for publication four of them as workshop talks and five as posters. The accepted workshop submissions cover a wide variety of language combinations from languages such as English, Hindi, Swahili, Mandarin, Dialectical Arabic and Modern Standard Arabic. The majority of the papers focus on social media data such as Twitter, and discussion fora.

Another component of the workshop is the Second Shared Task on Language Identification of CS Data. The shared task focused on social media and included two language pairs: Modern Standard Arabic-Dialectal Arabic and English-Spanish. We received a total of 14 system runs from nine different teams. All teams except one submitted a shared task paper describing their system. All shared task systems will be presented during the workshop poster session and two 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 2016 organizers for the opportunity to put together this workshop.

See you all in Austin, TX at EMNLP 2016!

Workshop co-chairs,

Mona Diab
Pascale Fung
Mahmoud Ghoneim
Julia Hirschberg
Thamar Solorio

Publications & Shared Task Chairs,

Fahad AlGhamdi
Mahmoud Ghoneim
Giovanni Molina

Workshop Co-Chairs:

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

Publications & Shared Task Chairs:

Fahad AlGhamdi, George Washington University
Mahmoud Ghoneim, George Washington University
Giovanni Molina, University of Houston

Program Committee:

Constantine Lignos, University of Pennsylvania
Elabbas Benmamoun, University of Illinois at Urbana-Champaign
Agnes Bolonyia, NC State University
Cecilia Montes-Alcala, Georgia Institute of Technology
Yves Scherre, Université de Genève
Björn Gambäck, Norwegian Universities of Science and Technology
Amitava Das, University of North Texas
Younes Samih, Dusseldorf University
David Vilares, Universidade da Coruña
Sunayana Sitaram, Microsoft Research India
Almeida Jacqueline Toribio, University of Texas at Austin
Fahad AlGhamdi, The George Washington University
Giovanni Molina Ramos, University of Houston
Nicolas Rey Villamizar, University of Houston
Victor Soto, Columbia University
Borja Navarro Colorado, Universidad de Alicante
Rabih Zbib, BBN Technologies
Barbara Bullock, University of Texas at Austin

Invited Speakers:

Monojit Choudhury, Microsoft Research Lab India.
Kalika Bali, Microsoft Research Lab India

Table of Contents

Workshop Program

Tuesday, November 1, 2016

Session 1: Opening Session

08:45–09:00 *Welcome Remarks*

09:00–10:00 *Keynote Talk*
NLP for Code-switching: Why more data is not necessarily the solution
Monojit Choudhury and Kalika Bali

10:00–10:30 *Challenges of Computational Processing of Code-Switching*
Özlem Çetinoğlu, Sarah Schulz and Ngoc Thang Vu

10:30–11:00 **Coffee Break**

Session 2: Workshop Talks

11:00–11:30 *Simple Tools for Exploring Variation in Code-switching for Linguists*
Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock and Almeida Jacqueline Toribio

11:30–12:00 *Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data*
Mario Piergallini, Rouzbeh Shirvani, Gauri S. Gautam and Mohamed Chouikha

12:00–12:30 *Part-of-speech Tagging of Code-mixed Social Media Content: Pipeline, Stacking and Joint Modelling*
Utsab Barman, Joachim Wagner and Jennifer Foster

12:30–14:00 **Lunch**

Session 3: Shared Task

14:00–14:30 *Overview for the Second Shared Task on Language Identification in Code-Switched Data*
Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab and Thamar Solorio

Tuesday, November 1, 2016 (continued)

14:30–15:00 *Multilingual Code-switching Identification via LSTM Recurrent Neural Networks*
Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer and Thamar Solorio

15:00–15:30 *A Neural Model for Language Identification in Code-Switched Tweets*
Aaron Jaech, George Mulcaire, Mari Ostendorf and Noah A. Smith

15:30–16:00 Coffee Break

Session 4: Panel Discussion and Poster Session

16:00–16:45 Panel Discussion

16:45–18:00 Poster Session

SAWT: Sequence Annotation Web Tool
Younes Samih, Wolfgang Maier and Laura Kallmeyer

Accurate Pinyin-English Codeswitched Language Identification
Meng Xuan Xia and Jackie Chi Kit Cheung

Unraveling the English-Bengali Code-Mixing Phenomenon
Arunavha Chanda, Dipankar Das and Chandan Mazumdar

Part-of-speech Tagging of Code-Mixed Social Media Text
Souvick Ghosh, Satanu Ghosh and Dipankar Das

Part of Speech Tagging for Code Switched Data
Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto and Julia Hirschberg

Tuesday, November 1, 2016 (continued)

The George Washington University System for the Code-Switching Workshop Shared Task 2016
Mohamed Al-Badrashiny and Mona Diab

Columbia-Jadavpur submission for EMNLP 2016 Code-Switching Workshop Shared Task: System description
Arunavha Chanda, Dipankar Das and Chandan Mazumdar

The Howard University System Submission for the Shared Task in Language Identification in Spanish-English Codeswitching
Rouzbeh Shirvani, Mario Piergallini, Gauri Shankar Gautam and Mohamed Chouikha

Codeswitching Detection via Lexical Features in Conditional Random Fields
Prajwol Shrestha

Language Identification in Code-Switched Text Using Conditional Random Fields and Babelnet
Utpal Kumar Sikdar and Björn Gambäck

Codeswitching language identification using Subword Information Enriched Word Vectors
Meng Xuan Xia