LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

4th Workshop on
Computational Approaches to Code Switching

PROCEEDINGS

Thamar Solorio, Monojit Choudhury, Kalika Bali, Sunayana Sitaram, Amitava Das, and Mona Diab (eds.)

Proceedings of the LREC 2020
4th Workshop on ComputationalApproaches to Code Switching

Edited by: Thamar Solorio, Monojit Choudhury, Kalika Bali, Sunayana Sitaram,
Amitava Das, and Mona Diab

ISBN: 979-10-95546-66-5
EAN: 9791095546665

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Preface

Welcome to the proceedings of the 4th workshop on Computational Approaches to Linguistic Code Switching (CALCS). 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. Moreover, CS language data is penetrating more traditional formal genres such as newswire in multilingual communities. 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 still presents a new frontier for NLP algorithms on all levels. CS accordingly has been garnering more importance and attention both in academic circles, research labs, and industry. Furthermore, the current pandemic and associated guidelines of physical distancing has created a significant spike in online platform usage in an unprecedented manner. The usage is for social connectivity but even more relevant is for information seeking. This increase in social media usage translates to more CS language usage leading to an even more urgent need for processing.

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 14 submissions, 9 of which were accepted. The papers run the gamut from creation of novel resources (such as a corpus of Spanish newspaper headlines annotated for Anglicisms, to a conversational data set annotated for CS) to modeling papers exploring advanced models (such as multi-task learning for low resource languages, impact of script mixing on modeling, impact of word embeddings on Indonesian-English, multimodal modeling of acoustic and linguistic features for English-IsiZulu, parsing for CS data, efficient grapheme to phoneme conversion) to papers addressing applications such as sentiment analysis and acoustic modeling for speech recognition. Finally, the range of papers cover some novel languages not addressed in previous CALCS workshops such as Indonesian-English, Korean Transliteration, English IsiZulu, Algerian Arabic which code switches with Modern Standard Arabic and French. We would like to thank all authors who submitted their contributions to this workshop. We also thank the program committee members for their help in providing meaningful reviews. Lastly, we thank the LREC 2020 organizers for the opportunity to put together this workshop. See you online/virtually at LREC 2020!

Workshop Organizers
Thamar Solorio, University of Houston (USA)
Monojit Choudhury, Microsoft Research (India)
Kalika Bali, Microsoft Research (India)
Sunayana Sitaram, Microsoft Research (India)
Amitava Das, Wipro AI (India)
Mona Diab, Facebook AI, George Washington University (USA)

Organizers:

Thamar Solorio, University of Houston (USA)
Monojit Choudhury, Microsoft Research (India)
Kalika Bali, Microsoft Research (India)
Sunayana Sitaram, Microsoft Research (India)
Amitava Das, Wipro AI (India)
Mona Diab, Facebook AI, George Washington University (USA)

Program Committee:

Gustavo Aguilar, University of Houston
Barbara Bullock, University of Texas at Austin
Özlem Cetinoglu, University of Stuttgart
Hila Gonen, Bar Ilan University
Sandipan Dandapat, Microsoft
A. Seza Doğruöz, Independent Researcher
William H. Hsu, Kansas State University
Constantine Lingos, Brandeis University
Rupesh Mehta, Microsoft
Joel Moniz, Carnegie Mellon University
Adithya Pratapa, Carnegie Mellon University
Yihong Theis, Kansas State University
Jacqueline Toribio, University of Texas at Austin
Gentra Inda Winata, Hong Kong University of Science and Technology
Dan Garrett, Google

Table of Contents