NAACL 2021

# Computational Approaches to Linguistic Code-Switching

## Proceedings of the Fifth Workshop

June 11, 2021

This workshop was sponsored by Facebook



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 978-1-954085-45-9

# Message from the Program Chairs

Bienvenidos to the proceedings of the fifth edition of the workshop on computational approaches for linguistic code-switching (CALCS-2021)! Code-switching is this very interesting phenomenon where multilingual speakers communicate by moving back and forth between the languages they speak when communicating with other multilingual speakers. Code-switching (CSW) is predominantly used in speech but since it also tends to be more prevalent in casual settings, we can observe CSW in genres like social media platforms where interactions tend to be more casual.

However interesting, our current NLP technology is lagging behind in the development of resources and methodologies that can effectively process code-switched language. This is true for even the large multilingual pretrained models such as mBERT and BART. At the same time, the growing adoption of smart devices and automated assistants that rely on speech interfaces, makes it even more pressing that our field addresses CSW language data.

This workshop series brings together experts and practitioners that are currently working on different aspects of CSW with a special focus on motivating tighter collaborations between speech and text researchers. We received 18 regular workshop submissions, of which we accepted 13. But this year we also had a special submission type called "Rising Stars". The goal of the Rising Stars is to allow young scientists that have recently published work in the space of CSW to present this work again to a specialized audience. These submissions are non-archival and are intended to increase visibility of CSW research by young researchers. We received two submissions of this type and we hope to continue this new track in future editions.

Our workshop also aims to motivate new research and energize the community to take on the challenges posed by CSW data. With this in mind, we hosted a new shared task on machine translation in CSW settings colocated with the workshop. This shared task provided two modalities for participation, supervised and unsupervised. For the supervised mode we asked participants to translate English data into Hinglish (Hindi-English). For the unsupervised setting we provided the following language pairs: Spanish-English (Spanglish) to English, English to Spanglish, Modern Standard Arabic-Egyptian Arabic (MSA-EA) to English and English to MSA-EA. The current leaderboard for the task shows 12 individual public system submissions coming out of 5 different teams. The overview of the shared task and the individual system submissions will be presented at the workshop.

The workshop program includes short talks from regular workshop submissions, rising star talks and system description talks. We also have a stellar invited speaker program with talks by Özlem Çetinoğlu, Manish Shrivastava and Ngoc Thang Vu. In addition, the one day program will also feature an exciting panel discussing research challenges unique to Machine Translation in CSW environments. Panelists include: Kalika Bali, Pushpak Bhattacharyya, Marina Fomicheva, Philipp Koehn, and Holger Schwenk.

We would like to thank the NAACL workshop organizers, Bhavana Dalvi, Mamoru Komachi and Michel Galley for their help during the organization of the workshop. We also extend our appreciation to Priscilla Rasmussen for her continuous help in the organization of these events. Last, but not least, we thank the NAACL organizing team for handling the conference organization in such a smooth way, even in the face of the current pandemic.

It would have been great to see everyone face to face in Mexico City, but alas we have another virtual event this year. Nonetheless, we hope that you join us on Friday June 11th and that you enjoy the program we put together.

Let's talk code-switching in June!

The Workshop Organizers

## Workshop Organizers:

Alan W. Black, Carnegie Mellon University (USA)
Mona Diab, Facebook (USA)
Shuguang Chen, University of Houston (USA)
Sunayana Sitaram, Microsoft Research (India)
Thamar Solorio, University of Houston (USA)
Victor Soto, Amazon Alexa AI (USA)
Anirudh Srinivasan, Microsoft Research India (India)
Emre Yilmaz, SRI International (USA)

## Program Committee:

Gustavo Aguilar, University of Houston (USA)
Elena Álvarez Mellado, University of Southern California (USA)
Segun Aroyehun, Instituto Politécnico Nacional (Mexico)
Kalika Bali, Microsoft Research India (India)
Astik Biswas, Oracle (India)
Monojit Choudhury, Microsoft Research India (India)
Amitava Das, Wipro AI Lab (India)
Indranil Dutta, Jadavpur University (India)
Alexander Gelbukh, Insituto Politécnico Nacional (Mexico)
Genta Indra Winata, HKUST (Hong Kong)
Sudipta Kar, Amazon (USA)
Grandee Lee, National University of Singapore (Singapore)
Els Lefever, Ghent University (Belgium)
Constantine Lignos, University of Pennsylvania (USA)
Yang Liu, Amazon (USA)
Manuel Mager, Universität Stuttgart (Germany)
Parth Patwa, Indian Institute of Information Technology Sri City (India)
Sai Krishna Rallabandi, Carnegie Mellon University (USA)
Yihong Theis, Kansas State University (USA)
Van Tung Pham, Nanyang Technological University (Singapore)
Khyathi Raghavi Chandu, Carnegie Mellon University (USA)
Seza Doğruöz, Ghent University (Belgium)

# Table of Contents

# Conference Program

**Friday, June 11, 2021 (Mexico City Time, CDT, GMT -5)**

**8:00–10:30 Morning Session I**

8:00–8:15 *Welcome Remarks*
Thamar Solorio

8:15–9:00 *Invited Talk*
Manish Shrivastava

9:00–9:30 *Lighting Talks*
Thamar Solorio

9:30–9:40 *Political Discourse Analysis: A Case Study of Code Mixing and Code Switching in Political Speeches*
Dama Sravani, Lalitha Kameswari and Radhika Mamidi

9:40–9:50 *Challenges and Limitations with the Metrics Measuring the Complexity of Code-Mixed Text*
Vivek Srivastava and Mayank Singh

9:50–10:00 *Translate and Classify: Improving Sequence Level Classification for English-Hindi Code-Mixed Data*
Devansh Gautam, Kshitij Gupta and Manish Shrivastava

**10:00–10:30** ***Break I***

**Friday, June 11, 2021 (Mexico City Time, CDT, GMT -5) (continued)**

**10:30–13:00 Morning Session II**

10:30–10:40 *Shared Task Overview*
Thamar Solorio

10:40–10:50 *Gated Convolutional Sequence to Sequence Based Learning for English-Hingilsh Code-Switched Machine Translation.*
Suman Dowlagar and Radhika Mamidi

10:50–11:00 *IITP-MT at CALCS2021: English to Hinglish Neural Machine Translation using Unsupervised Synthetic Code-Mixed Parallel Corpus*
Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal and Pushpak Bhattacharyya

11:00–11:10 *Exploring Text-to-Text Transformers for English to Hinglish Machine Translation with Synthetic Code-Mixing*
Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed and Laks Lakshmanan, V.S.

11:10–11:20 *CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences*
Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava and Ponnurangam Kumaraguru

11:20–11:30 *Investigating Code-Mixed Modern Standard Arabic-Egyptian to English Machine Translation*
El Moatez Billah Nagoudi, AbdelRahim Elmadany and Muhammad Abdul-Mageed

11:30–12:15 *Invited Talk*
Özlem Çetinoğlu

**12:15–13:00** ***Lunch Break***

**Friday, June 11, 2021 (Mexico City Time, CDT, GMT -5) (continued)**

**13:00–15:30 Afternoon Session I**

13:00–13:10 *Much Gracias: Semi-supervised Code-switch Detection for Spanish-English: How far can we get?*
Dana-Maria Iliescu, Rasmus Grand, Sara Qirko and Rob van der Goot

13:10–13:20 *A Language-aware Approach to Code-switched Morphological Tagging*
Şaziye Betül Özateş and Özlem Çetinoğlu

13:20–13:30 *Can You Traducir This? Machine Translation for Code-Switched Input*
Jitao Xu and François Yvon

13:30–13:40 *On the logistical difficulties and findings of Jopara Sentiment Analysis*
Marvin Agüero-Torales, David Vilares and Antonio López-Herrera

13:40–15:00 *Panel Discussion Moderated by Mona Diab*
Panelists: Kalika Bali, Pushpak Bhattacharyya, Marina Fomicheva, Philipp Koehn, Holger Schwenk

**15:00–15:30** ***Midday Short Break***

**15:30–16:45 Afternoon Session II**

15:30–16:15 *Invited Talk*
Ngoc Thang Vu

**16:15–16:45** ***Evening Break***

**Friday, June 11, 2021 (Mexico City Time, CDT, GMT -5) (continued)**

**16:45–18:15 Evening Session**

16:45–16:55 *Unsupervised Self-Training for Sentiment Analysis of Code-Switched Data*
Akshat Gupta, Sargam Menghani, Sai Krishna Rallabandi and Alan W Black

16:55–17:05 *CodemixedNLP: An Extensible and Open NLP Toolkit for Code-Mixing*
Sai Muralidhar Jayanthi, Kavya Nerella, Khyathi Raghavi Chandu and Alan W Black

17:05–17:15 *Normalization and Back-Transliteration for Code-Switched Data*
Dwija Parikh and Thamar Solorio

17:15–17:25 *Abusive content detection in transliterated Bengali-English social media corpus*
Salim Sazzed

17:25–17:35 *Developing ASR for Indonesian-English Bilingual Language Teaching*
Zara Maxwelll-Smith and Ben Foley

17:35–17:45 *Transliteration for Low-Resource Code-Switching Texts: Building an Automatic Cyrillic-to-Latin Converter for Tatar*
Chihiro Taguchi, Yusuke Sakai and Taro Watanabe

17:45–17:55 *Code-Mixing on Sesame Street: Dawn of the Adversarial Polyglots*
Samson Tan and Shafiq Joty

17:55–18:05 *Are Multilingual Models Effective in Code-Switching?*
Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto and Pascale Fung

**18:05–18:15** ***Closing Remarks***