CALCS 2023

Computational Approaches to Linguistic Code-Switching

Proceedings of the Workshop

December 7, 2023

The CALCS organizers gratefully acknowledge the support from the following sponsors.

Gold

Bloomberg

Engineering

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 979-8-89176-053-0

Introduction

Bienvenidos to the proceedings of the sixth edition of the workshop on computational approaches for linguistic code-switching (CALCS-2023)! Code-switching is a common phenomenon in the multilingual communities where multilingual speakers communicate by moving back and forth between the languages they speak when communicating with other multilingual speakers. This year the workshop is being held in Singapore on December 7th, 2023 at EMNLP.

This workshop series brings together experts and practitioners that are currently working on different aspects of code-switching with a special focus on motivating tighter collaborations between speech and text researchers. We received 15 regular workshop submissions, of which we accepted 8 and 1 non-archival. Our workshop also aims to motivate new research and energize the community to take on the challenges posed by code-switching data.

The workshop program includes short talks from regular workshop submissions and keynote speakers. We also have a stellar invited speaker program with a keynote talk by Preethi Jyothi and Haizhou Li. We would like to thank the EMNLP workshop organizers for their help during the organization of the workshop. It would have been great to see everyone face to face in Singapore and we hope that you join us on December 7th and that you enjoy the program we put together.

Let’s talk code-switching in December!

The Workshop Organizers

Organizing Committee

Organizer

Sudipta Kar, Amazon
Genta Indra Winata, Bloomberg
Marina Zhukova, University of California, Santa Barbara
Thamar Solorio, University of Houston and Mohammad bin Zayed University of Artificial Intelligence
Mona Diab, Carnegie Mellon University
Sunayana Sitaram, Microsoft Research
Monojit Choudhury, Microsoft Turing
Kalika Bali, Microsoft Research

Program Committee

Program Committee

A. Seza Doğruöz
Abhinav Arora
Dama Sravani
David Vilares
Elena Álvarez-Mellado
Els Lefever
Holy Lovenia
François Yvon
Ganesh Jawahar
Gustavo Aguilar
Kellen Gillespie
Manuel Mager
Parth Patwa
Salim Sazzed
Segun Aroyehun
Shuguang Chen
Suman Dowlagar
Suraj Maharjan
Tanya Roosta
Vivek Srivastava
Xingzhi Guo
Yihong Theis

Invited Speakers

Haizhou Li, The Chinese University of Hong Kong and National University of Singapore
Preethi Jyothi, IIT Bombay

Keynote Talk: Modeling Code-Switch Languages Using Bilingual Parallel Corpus

Haizhou Li
The Chinese University of Hong Kong, Shenzhen; National University of Sinagapore

Abstract: Language modeling is the technique to estimate the probability of a sequence of words. A bilingual language model is expected to model the sequential dependency for words across languages, which is difficult due to the inherent lack of suitable training data as well as diverse syntactic structure across languages. We propose a bilingual attention language model (BALM) that simultaneously performs language modeling objective with a quasi-translation objective to model both the monolingual as well as the cross-lingual sequential dependency. The attention mechanism learns the bilingual context from a parallel corpus. We will discuss the study of multilingualism in South East Asia and how code-switch language models can be useful for language processing.

Bio: Haizhou Li is the X.Q. Deng Presidential Chair Professor in the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also an Adjunct Professor at the National University of Singapore, Singapore and a Bremen Excellence Chair Professor at the University of Bremen, Germany. Prior to joining CUHK (Shenzhen), Professor Li has taught at Nanyang Technological University and National University of Singapore (2006-2016) in Singapore, University of Eastern Finland (2009) in Finland, and University of New South Wales (2011-2016) in Australia. He was the Principal Scientist and Research Director at the Institute for Infocomm Research (2003-2016), the Agency for Science, Technology and Research, Singapore. Professor Li is an IEEE Fellow, and ISCA Fellow.
He has served as the Editor-in-Chief of IEEE-ACM Transactions on Audio Speech and Language Processing (2015-2018), Associate Editor of Computer Speech and Language (2012-2021), Springer International Journal of Social Robotics (2008-2021), and a Member of IEEE Speech and Language Processing Technical Committee (2013-2015), Awards Board (2021-2023), and Publications Board (2015-2018) of IEEE Signal Processing Society. He was the President of the International Speech Communication Association (ISCA, 2015-2017), the President of Asia Pacific Signal and Information Processing Association (APSIPA, 2015-2016), the President of the Asian Federation of Natural Language Processing (AFNLP, 2017-2018). He was the General Chair of major scientific conferences including ACL 2012, INTERSPEECH 2014, and ICASSP 2022.

Keynote Talk: Resource-efficient Computational Models for Code-switched Speech and Text

Preethi Jyothi
IIT Bombay

Abstract: Code-switching, i.e., the linguistic phenomenon of switching between languages within and across sentences, is widely prevalent in multilingual societies. Code-switched inputs pose a serious challenge to existing speech and NLP models. The challenge mainly emerges due to the limited availability of natural code-switched data and the inherent diversity in code-switching. In this talk, we will discuss techniques that aim to effectively address these dual challenges. These techniques will cover how to exploit monolingual speech and text for code-switching, how to generate synthetic and diverse code-switched text to augment real data and how to judiciously use existing real code-switched speech and text in conjunction with other linguistic resources.

Bio: Preethi is an Associate Professor at IIT Bombay. She joined the department in September 2016. Prior to that, she was a Beckman Postdoctoral Fellow at the University of Illinois at Urbana-Champaign. She obtained my Ph.D. from the CSE Department at The Ohio State University in 2013. Her research interests are broadly in the areas of automatic speech recognition and machine learning as applied to speech, and code-switching.

Table of Contents

TongueSwitcher: Fine-Grained Identification of German-English Code-Switching
Igor Sterner and Simone Teufel 1

Towards Real-World Streaming Speech Translation for Code-Switched Speech
Belen Alastruey, Matthias Sperber, Christian Gollan, Dominic Telaar, Tim Ng and Aashish Agarwal 14

Language Preference for Expression of Sentiment for Nepali-English Bilingual Speakers on Social Media
Niraj Pahari and Kazutaka Shimada 23

Text-Derived Language Identity Incorporation for End-to-End Code-Switching Speech Recognition
Qinyi Wang and Haizhou Li 33

Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages
Zheng Xin Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio and Alham Fikri Aji 43

CONFLATOR: Incorporating Switching Point based Rotatory Positional Encodings for Code-Mixed Language Modeling
Mohsin Ali Mohammed, Sai Teja Kandukuri, Neeharika Gupta, Parth Patwa, Anubhab Chatterjee, Vinija Jain, Aman Chadha and Amitava Das 64

Unified Model for Code-Switching Speech Recognition and Language Identification Based on Concatenated Tokenizer
Kunal Dhawan, KDimating Rekesh and Boris Ginsburg 74

Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching
Tolulope Ogunremi, Christopher Manning and Dan Jurafsky 83

Program

Thursday, December 7, 2023

09:10 - 09:00 *Opening Remarks*

09:05 - 10:35 *Paper Oral Presentation 1*

Towards Real-World Streaming Speech Translation for Code-Switched Speech
Belen Alastruey, Matthias Sperber, Christian Gollan, Dominic Telaar, Tim Ng and Aashish Agarwal

Text-Derived Language Identity Incorporation for End-to-End Code-Switching Speech Recognition
Qinyi Wang and Haizhou Li

TongueSwitcher: Fine-Grained Identification of German-English Code-Switching
Igor Sterner and Simone Teufel

CONFLATOR: Incorporating Switching Point based Rotatory Positional Encodings for Code-Mixed Language Modeling
Mohsin Ali Mohammed, Sai Teja Kandukuri, Neeharika Gupta, Parth Patwa, Anubhab Chatterjee, Vinija Jain, Aman Chadha and Amitava Das

Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages
Zheng Xin Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio and Alham Fikri Aji

10:25 - 11:00 *Break*

11:00 - 11:45 *Invited Talk - Preethi Jyothi*

11:45 - 12:15 *Paper Oral Presentation 2*

Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching
Tolulope Ogunremi, Christopher Manning and Dan Jurafsky

Language Preference for Expression of Sentiment for Nepali-English Bilingual Speakers on Social Media
Niraj Pahari and Kazutaka Shimada

12:15 - 12:30 *Findings Paper Code-Switching with Word Senses for Pretraining in Neural Machine Translation*

Thursday, December 7, 2023 (continued)

14:00 - 12:30	*Lunch Break*
14:00 - 12:15	*Paper Oral Presentation 3*
	Unified Model for Code-Switching Speech Recognition and Language Identification Based on Concatenated Tokenizer Kunal Dhawan, KDimating Rekesh and Boris Ginsburg
14:15 - 15:30	*Panel Discussion - Sudipta Kar, Genta Winata, Marina Zhukova*
15:30 - 16:00	*Coffee Break*
16:00 - 16:45	*Invited Talk - Haizhou Li*
16:45 - 16:50	*Best Paper Announcement*
16:50 - 16:55	*Closing Remarks*