ACL 2023

Third Workshop on Natural Language Processing for Indigenous Languages of the Americas

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)

July 14, 2023

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

Gold

Google

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 978-1-959429-91-3

Introduction

Welcome to AmericasNLP 2023, the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas!

AmericasNLP aims to...

- ...encourage research on NLP, computational linguistics, corpus linguistics, and speech around the globe to work on Indigenous American languages.
- ...connect researchers and professionals from underrepresented communities and native speakers of endangered languages with the machine learning and NLP communities.
- ...promote research on both neural and non-neural machine learning approaches suitable for low-resource languages.

In 2023, AmericasNLP is being held in Toronto, Canada, on July 14. There will be 3 invited talks, an overview of this year's AmericasNLP shared task, a poster session, and multiple paper as well as shared task system presentations.

We received a total of 33 submissions this year: 22 research papers, 1 extended abstract, 3 previously published papers, and 7 shared task system description papers. 15 research papers were accepted (acceptance rate: 68%) – as well as all extended abstracts, previously published papers, and system description papers. In addition, two Findings of ACL papers will be presented at the workshop.

We would like to extend our gratitude to everyone who helped make AmericasNLP happen: First, we thank our gold sponsor, Google. In addition, AmericasNLP would not be possible without all the work that went into the reviewing process. Thus, we thank the program committee members for committing their time to help us select an excellent technical program. Finally, we thank all the authors who submitted their work to the workshop and all participants who will be at the workshop to exchange their ideas around NLP for Indigenous languages of the Americas!

Manuel Mager, Abteen Ebrahimi, Arturo Oncevay, Enora Rice, Shruti Rijhwani, Alexis Palmer,
and Katharina Kann
AmericasNLP 2023 Organizing Committee

Organizing Committee

Organizing Committee

Mager Manuel, AWS AI Labs, USA
Abteen Ebrahimi, University of Colorado Boulder, USA
Arturo Oncevay, University of Edinburgh, UK
Enora Rice, University of Colorado Boulder, USA
Shruti Rijhwani, Google Research, USA
Alexis Palmer, University of Colorado Boulder, USA
Katharina Kann, University of Colorado Boulder, USA

Program Committee

Program Committee

Eduardo Blanco, University of Arizona
Jie Cao, University of Colorado
Paulo Cavalin, IBM Research - Brazil
Luis Chiruzzo, Universidad de la Republica
Rolando Coto-solano, Dartmouth College
Ruixiang Cui, University of Copenhagen
C.m. Downey, University of Washington
Cristina España-bonet, DFKI GmbH
Luke Gessler, Georgetown University
Héctor Jiménez-salazar, Universidad Autónoma Metropolitana, Cuajimalpa
Kartik Kannapur, Amazon Web Services
Zoey Liu, Department of Linguistics, University of Florida
Arya D. Mccarthy, Johns Hopkins University
Ivan Vladimir Meza Ruiz, Insituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México
Daniela Moctezuma, Centrogeo
Sarah Moeller, University of Colorado
Manuel Montes, INAOE
John E. Ortega, Northeastern University
Shiva Kumar Pentyala, Salesforce AI
Angeles Belem Priego Sanchez, Universidad Autónoma Metropolitana
Amit Sah, Department of Computer Science, South Asian University
Elizabeth Salesky, Johns Hopkins University
Shabnam Tafreshi, UMD:ARLIS
Atnafu Lambebo Tonja, Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)
Ivan Vulić, University of Cambridge
Ekaterina Vylomova, University of Melbourne
Koichiro Watanabe, pluszero, inc
Adam Wiemerslage, University of Colorado Boulder

Keynote Talk: No Language Left Behind: Scaling Human-Centered Machine Translation

Angela Fan
Meta AI Research
2023-07-14 09:15:00 – Room: **TBD**

Abstract: Driven by the goal of eradicating language barriers on a global scale, machine translation has solidified itself as a key focus of artificial intelligence research today. However, such efforts have coalesced around a small subset of languages, leaving behind the vast majority of mostly low-resource languages. What does it take to break the 200 language barrier while ensuring safe, high-quality results, all while keeping ethical considerations in mind? In this talk, I introduce No Language Left Behind, an initiative to break language barriers for low-resource languages. In No Language Left Behind, we took on the low-resource language translation challenge by first contextualizing the need for translation support through exploratory interviews with native speakers. Then, we created datasets and models aimed at narrowing the performance gap between low and high-resource languages. We proposed multiple architectural and training improvements to counteract overfitting while training on thousands of tasks. Critically, we evaluated the performance of over 40,000 different translation directions using a human-translated benchmark, Flores-200, and combined human evaluation with a novel toxicity benchmark covering all languages in Flores-200 to assess translation safety. Our model achieves an improvement of 44% BLEU relative to the previous state-of-the-art, laying important groundwork towards realizing a universal translation system in an open-source manner.

Bio: Angela is a research scientist at Meta AI Research in New York, focusing on research in text generation. Currently, Angela works on language modeling. Recent projects include No Language Left Behind (https://ai.facebook.com/research/no-language-left-behind/) and Universal Speech Translation for Unwritten Languages (https://ai.facebook.com/blog/ai-translation-hokkien/). Before translation, Angela previously focused on research in on-device models for NLP and computer vision and text generation.

Keynote Talk: From fieldwork to ”data” - A behind-the-scenes look from Brazilian Amazonia

Kristine Stenzel
Federal University of Rio de Janeiro / University of Colorado Boulder
2023-07-14 11:00:00 – Room: **TBD**

Abstract: This talk offers an overview of one linguist's experience in language documentation with two indigenous groups in the northwest Amazon. Based on over twenty years of fieldwork, it aims to provide broader perspective on what goes into the collection, organization, and annotation of "data" from endangered or low-resource languages.

Bio: Kristine Stenzel was an Associate Professor of Linguistics at the Federal University of Rio de Janeiro, Brazil from 2009-2022 and is currently at the University of Colorado as Coordinator of the Computational Linguistics, Analytics, Search, and Informatics Professional Master's Program. She has conducted research with the Kotiria and Wa'ikhana language communities since 2000, receiving grants from NSF, NEH, ELDP, as well as CNPq and CAPES in Brazil. Her scientific contributions include *A Reference Grammar of Kotiria* and publications in English and in Portuguese on diverse topics in phonology, morphosyntax, discourse, multilingualism, contact phenomena, and language documentation. She has developed language maintenance and revitalization materials for the Kotiria and Wa'ikhana, including practical orthographies, pedagogical publications, documentary films, and audiovisual archives (ELAR, open access).

Keynote Talk: From doctoral thesis to the classroom: The case of San Juan Quiahije Chatino

Emiliana Cruz Cruz
CIESAS-CDMX
2023-07-14 16:00:00 – Room: **TBD**

Abstract: In this presentation I will address an issue that is very important to us as speakers of indigenous languages: how to ensure that linguistic studies on indigenous languages reach the hands of the speakers of these languages. Over the last 20 years, the Chatino Language Documentation Project (CLDP) has resulted in seven doctoral theses in the three Chatino languages, all written in English. For the Eastern San Juan Quiahije Chatino, there are four doctoral theses. The theses are of great importance for the speakers. However, generating pedagogical products based on these doctoral theses has been a slow process. It is not just a translation issue, as CLDP linguists have tried to make teaching materials out of their research. So, what are the challenges when we are dealing with a "well-studied" Chatino language? In this talk I will present some reflections around this question based on a project in the municipality of Quiahije.

Bio: I am a linguistic anthropologist and assistant professor at CIESAS-DF. I primarily work on language treatment and revitalization, with a focus on the Chatino language of Oaxaca, Mexico; cultural identity and maintenance through language programs and curriculum development; orthography development; and, finally, issues related to sovereignty and decolonization. One aspect of my research is the application of anthropological methods in the documentation of naturally occurring discourse in indigenous languages. An essential contribution of the anthropological perspective is the recognition of the crucial role to be played by native speaker linguists in all phases of research. My linguistic work centers on the complex tonal structure of Chatino languages, and I developed the San Juan Quiahije variety's alphabet. An important result of this project has been the creation of pedagogical materials that will enable members of the Chatino community to preserve their language and cultural integrity. I am a native speaker of Chatino and founder of The Chatino Language Documentation Project, a team of linguists which aims to document and revitalize Chatino languages.

Table of Contents