AmericasNLP 2024

4th Workshop on Natural Language Processing for Indigenous Languages of the Americas

Proceedings of the Workshop

June 21, 2024

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

Platinum



Bronze



©2024 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL) 317 Sidney Baker St. S Suite 400 - 134 Kerrville, TX 78028 USA Tel: +1-855-225-1962 acl@aclweb.org

ISBN 979-8-89176-108-7

Introduction

We would like to welcome you to AmericasNLP 2024, the Fourth Workshop on Natural Language Processing for Indigenous Languages of the Americas!

The main goals of the workshop are to:

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

In 2024, AmericasNLP is being held in Mexico City, Mexico, on June 21. Prior to the workshop two shared tasks were hosted: (1) the Shared Task on Machine Translation into Indigenous Languages and (2) the Shared Task on the Creation of Educational Materials for Indigenous Languages. During the workshop, there will be 2 invited talks, a poster session, and multiple paper and shared task submission presentations.

We received a total of 37 submissions: 21 research papers, 6 previously published papers, and 10 shared task system description papers (across both shared tasks). 16 archival papers were accepted (acceptance rate: 76%) – in addition to the previously published and system description papers.

AmericasNLP would not have been possible without the help of the following: first, we thank our sponsors, Amazon and Aditu, and second, we would like to acknowledge all the time and effort put into the reviewing process, and thank for program committee members for helping us create a high-quality program. Finally, we also thank all the authors who submitted their work to the workshop, the participants of both shared tasks, and everyone who will be at the workshop, both in-person and remote, to exchange and discuss their ideas for improving natural language technologies for Indigenous languages of the Americas!

Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense

AmericasNLP 2024 Organizing Committee

Organizing Committee

Organizing Committe

Manuel Mager, Amazon AWS AI, USA Abteen Ebrahimi, University of Colorado Boulder, USA Shruti Rijhwani, Google Deepmind, USA Arturo Oncevay, Pontificia Universidad Católica del Perú, Peru Luis Chiruzzo, Universidad de la República, Uruguay Robert Pugh, Indiana University, Bloomington, USA Katharina von der Wense, University of Colorado Boulder, USA; Johannes Gutenberg University Mainz, Germany

Program Committee

Program Committee

Eduardo Blanco, University of Arizona Paulo Cavalin, IBM Research - Brazil Raghuveer Chanda, Google Rolando Coto-Solano, Dartmouth College Ruixiang Cui, University of Copenhagen Shridhar Devamane, Global Academy of Technology, Bangalore C.M. Downey, University of Washington Cristina España-Bonet, DFKI GmbH Luke Gessler, University of Colorado, Boulder Kartik Kannapur, Amazon Web Services Satyapriya Krishna, Harvard University Éric Le Ferrand, Boston College Yukyung Lee, korea university Zoey Liu, Department of Linguistics, University of Florida Daniela Moctezuma, Centrogeo Sarah Moeller, University of Florida Remo Nitschke, University of Arizona John E. Ortega, Northeastern University Tanmay Parekh, University of California Los Angeles Angeles Belem Priego Sanchez, Universidad Autónoma Metropolitana Nathaniel Robinson, Johns Hopkins University Amit Sah, Department of Computer Science, South Asian University Shabnam Tafreshi, UMD:ARLIS Atnafu Lambebo Tonja, Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC) Daan Van Esch, Google Research Aniket Vashishtha, Microsoft Research India Raul Vazquez, University of Helsinki Ekaterina Vylomova, University of Melbourne Koichiro Watanabe, pluszero, inc Adam Wiemerslage, University of Colorado Boulder Si Wu, Northeastern University

Keynote Talk

Graham Neubig

Carnegie Mellon University

Bio: My research is concerned with language and its role in human communication. In particular, my long-term research goal is to break down barriers in human-human or human-machine communication through the development of natural language processing (NLP) technologies. This includes the development of technology for machine translation, which helps break down barriers in communication for people who speak different languages, and natural language understanding, which helps computers understand and respond to human language. Within this overall goal of breaking down barriers to human communication, I have focused on several aspects of language that both make it interesting as a scientific subject, and hold potential for the construction of practical systems.

Keynote Talk

Jaime Pérez González

University of California, Santa Barbara

Bio: As a morpho-syntactician, I study word formation and I am especially interested in the interface between morphology, syntax and semantics in human languages. I look at functional factors that shape the use of certain morphological constructions in agglutinative languages. Speakers activate these different domains when using their language, so there must be abstract and formal principles that determine how these patterns are represented in their minds. I investigate these topics in lesser-studied languages, with focus on Mayan languages and Miskitu (Misumalpan language spoken in Nicaragua and Honduras).

Table of Contents

NLP for Language Documentation: Two Reasons for the Gap between Theory and Practice Luke Gessler and Katharina von der Wense 1
Translation systems for low-resource Colombian Indigenous languages, a first step towards cultural preservation Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios and Rubén Man-
rique
Word-level prediction in Plains Cree: First steps Olga Kriukova and Antti Arppe15
Mapping 'when'-clauses in Latin American and Caribbean languages: an experiment in subtoken- based typology Nilo Pedrazzini
Comparing LLM prompting with Cross-lingual transfer performance on Indigenous and Low-resource Brazilian Languages David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian and Atul Kr. Ojha
Analyzing Finetuned Vision Models for Mixtec Codex Interpretation Alexander Webber, Zachary Sayers, Amy Wu, Elizabeth Thorner, Justin Witter, Gabriel Ayoubi and Christan Grant
A New Benchmark for Kalaallisut-Danish Neural Machine Translation Ross Kristensen-Mclachlan and Johanne Nedergård
Morphological Tagging in Bribri Using Universal Dependency FeaturesJessica Karson and Rolando Coto-Solano56
<i>LLM-Assisted Rule Based Machine Translation for Low/No-Resource Languages</i> Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales and Khalil Iskarous
A Concise Survey of OCR for Low-Resource Languages Milind Agarwal and Antonios Anastasopoulos
Unlocking Knowledge with OCR-Driven Document Digitization for Peruvian Indigenous Languages Shadya Sanchez Carrera, Roberto Zariquiey and Arturo Oncevay
Awajun-OP: Multi-domain dataset for Spanish–Awajun Machine Translation Oscar Moreno, Yanua Atamain and Arturo Oncevay
Wav2pos: Exploring syntactic analysis from audio for Highland Puebla NahuatlRobert Pugh, Varun Sreedhar and Francis Tyers121
<i>From Field Linguistics to NLP: Creating a curated dataset in Amuzgo language</i> Antonio Reyes and Hamlet Antonio García
<i>Enenlhet as a case-study to investigate ASR model generalizability for language documentation</i> Éric Le Ferrand, Raina Heaton and Emily Prud'hommeaux
Advancing NMT for Indigenous Languages: A Case Study on Yucatec Mayan and Chol Julio Rangel and Norio Kobayashi

BSC Submission to the AmericasNLP 2024 Shared Task
Javier Garcia Gilabert, Aleix Sant, Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash and Maite Melero
System Description of the NordicsAlps Submission to the AmericasNLP 2024 Machine Translation Shared Task Joseph Attieh, Zachary Hopton, Yves Scherrer and Tanja Samardžić
On the Robustness of Neural Models for Full Sentence Transformation Michael Ginn, Ali Marashian, Bhargav Shandilya, Claire Post, Enora Rice, Juan Vásquez, Marie Mcgregor, Matthew Buchholz, Mans Hulden and Alexis Palmer
The unreasonable effectiveness of large language models for low-resource clause-level morphology: In-context generalization or prior exposure? Coleman Haley 174
A Comparison of Fine-Tuning and In-Context Learning for Clause-Level Morphosyntactic Alternation Jim Su, Justin Ho, George Broadwell, Sarah Moeller and Bonnie Dorr
Experiments in Mamba Sequence Modeling and NLLB-200 Fine-Tuning for Low Resource Multilingual Machine Translation Dan Degenaro and Tom Lupicki
JGU Mainz's Submission to the AmericasNLP 2024 Shared Task on the Creation of Educational Mate- rials for Indigenous Languages Minh Duc Bui and Katharina von der Wense
Applying Linguistic Expertise to LLMs for Educational Material Development in Indigenous Languages Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung and Taro Watanabe
Exploring Very Low-Resource Translation with LLMs: The University of Edinburgh's Submission to AmericasNLP 2024 Translation Task Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow and Alexandra Birch 209
The role of morphosyntactic similarity in generating related sentences Michael Hammond 221

Findings of the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense and Manuel Mager 224

Findings of the AmericasNLP 2024 Shared Task on Machine Translation into Indigenous Languages

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense and Shruti Rijhwani 236