computel 2023

Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages

Proceedings of the Workshop

March 5-6, 2023

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 None

Introduction

We are excited to welcome you to the 6th Workshop on the Use of Computational Methods in the Study of Endangered Languages. This year it is going to be held as a virtual event from March 5-6, 2023, and co-located with the 8th International Conference on Language Documentation and Conservation (ICLDC).

As the name implies, this is the sixth workshop held on the topic—the first meeting was co-located with the ACL main conference in Baltimore, Maryland in 2014 and the second, third, and fourth ones in 2017, 2019, and 2021 were co-located with the 5th, 6th, and 7th editions of the International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawaii at Mānoa. The fifth iteration of the workshop was held in 2022 alongside the 60th Association of Computational Linguistics (ACL) conference in Dublin, Ireland. This is the fourth time this workshop has been co-located with the ICLDC, and it enhances the conference's 2023 theme of justice. The workshop covers a wide range of topics relevant to the study and documentation of endangered languages, ranging from technical papers on working systems and applications, to reports on community activities with supporting computational components. This year, the workshop held a special session centering work on justice and equity in language technology as a part of ICLDC 8.

The purpose of the workshop is to bring together computational researchers, documentary linguists, and people involved with community efforts of language documentation and revitalization to take part in both formal and informal exchanges on how to integrate rapidly evolving language processing methods and tools into efforts of language description, documentation, and revitalization. The organizers are pleased with the range of papers, many of which highlight the importance of interdisciplinary work and interaction between the various communities that the workshop is aimed towards.

We received 24 submissions as papers or extended abstracts. After a thorough review process, 17 submissions were selected to be published in the ACL Anthology. In addition to these papers, we had four oral presentations that made up our special session on justice.

The Organizing Committee would like to thank the Program Committee for their thoughtful review of the submissions. We are also grateful to the Social Sciences and Humanities Research Council (SSHRC) of Canada for supporting the workshop through their Partnership Grant 895-2019-1012. We would moreover want to acknowledge the support of the organizers of ICLDC 8.

Organizing Committee

Organizers

Atticus Harrigan, University of Alberta, Canada
Aditi Chaudhary, Google Research, USA
Shruti Rijhwani, Google Research, USA
Sarah Moeller, University of Florida, USA
Antti Arppe, University of Alberta, Canada
Alexis Palmer, University of Colorado Boulder, USA
Ryan Henke, University of Wisconsin-Madison, USA
Daisy Rosenblum, The University of British Columbia, Canada

Program Committee

Program Committee

Aditi Chaudhary, Google Research, USA
Atticus Harrigan, University of Alberta
Shruti Rijhwani, Google Research, USA
Daisy Rosenblum, The University of British Columbia
Ryan Henke, University of Wisconsin-Madison
Robert Forkel, Max Planck Institute for Evolutionary Anthropology
Anna Kazantseva, National Research Council Canada
Antti Arppe, University of Alberta
Vera Ferreira, Centro Interdisciplinar de Documentação Linguística e Social
Daan van Esch, Google
Jeff Good, University at Buffalo
Matthew Kelley, University of Washington
Alexis Palmer, University of Colorado Boulder
Michael Maxwell, University of Maryland
Jordan Lachler, University of Alberta
Emily M. Bender, University of Washington
František Kratochvíl, Palacký University
Martin Benjamin, Kamusi Project International
Olivia Sammons, First Nations University of Canada
Gary Simons, SIL International
Yves Scherrer, University of Helsinki
Conor Snoek, University of Lethbridge
Steven Bird, Charles Darwin University
Roland Kuhn, National Research Council Canada
Andrea Berez-Kroeker, University of Hawaii at Mānoa Department of Linguistics
Richard Sproat, Google
Fei Xia, University of Washington
Kevin Scannell, Saint Louis University
Harald Hammarström, MPI Nijmegen
Paul Trilsbeek, Max Planck Institute for Psycholinguistics
Atticus Harrigan, University of Alberta
Olga Lovick, University of Saskatchewan
Sarah Moeller, University of Florida
Lori Levin, Carnegie Mellon University
W N Martin, University of Virginia
Dorothee Beermann, Norwegian University of Science and Technology

Table of Contents

Program

Sunday, March 5, 2023

10:30 - 10:40 *Day-1 Welcome + Introduction*

10:40 - 13:30 *Special Session Justice through Technology*

16:00 - 17:45 *Session 1*

User-Centric Evaluation of OCR Systems for Kwak'wala
Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos and Graham Neubig

Towards a finite-state morphological analyser for San Mateo Huave
Francis M. Tyers and Samuel Herrera Castro

A morphological analyzer for Huasteca Nahuatl
Ana Tona, Guillaume Thomas and Ewan Dunbar

16:45 - 17:15 *Day-1 Break*

17:00 - 17:30 *Session 2*

LSDT a Dependency Treebank of Lombard Sinti
Marco Forlano and Luca Brigada Villa

Towards Universal Dependencies in Cook Islands Māori
Sarah Karnes, Rolando Coto-Solano and Sally Akevai Nicholas

Monday, March 6, 2023

09:00 - 09:45 *Session 1*

Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions
Nay San, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell and Dan Jurafsky

Studying the impact of language model size for low-resource ASR
Zoey Liu, Justin Spence and Emily Prud'Hommeaux

Speech-to-text recognition for multilingual spoken data in language documentation
Lorena Martín Rodríguez and Christopher Cox

09:45 - 10:00 *Break*

10:00 - 10:45 *Session 2*

FileLingR An R Script validation tool for depositors and users of digital language collections
Irene Yi and Claire Bowern

Investigating Speaker Diarization of Endangered Language Data
Gina-Anne Levow

From Raw Data to Acoustic Analysis A Roadmap for Acquaviva Collecroce
Simon Gonzalez

10:45 - 11:45 *Midday Break*

11:45 - 14:15 *Session 3*

A Survey of Computational Infrastructure to Help Preserve and Revitalize Bodwéwadmimwen
Robert Lewis

12:30 - 13:15 *Break*

Monday, March 6, 2023 (continued)

12:30 - 13:15 *Session 4*

Challenges and Issue of Gender Bias in Under-Represented Languages An Empirical Study on Inuktitut-English NMT
Ngoc Tan Le, Oussama Hansal and Fatiha Sadat

A text-to-speech synthesis system for Border Lakes Ojibwe
Christopher Hammerly, Sonja Fougère, Giancarlo Sierra, Scott Parkhill, Harrison Porteous and Chad Quinn

Text normalization for low-resource languages the case of Ligurian
Stefano Lusito, Edoardo Ferrante and Jean Maillard

13:15 - 13:30 *Day-2 Break*

13:30 - 14:15 *Session 5*

Morphological Data Generation from FLEx
Shengyu Liao, Beth Bryson and Sarah Moeller

Using LARA to rescue a legacy Pitjantjatjara course
Manny Rayner and Sasha Wilmoth

Applications of classification trees for endangered language description Finite verb morphology in Kolyma Yukaghir
Albert Ventayol-Boada

14:15 - 15:15 *SIGEL Business Meeting*