SIGMORPHON 2024

The 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics

Volume 1 - Proceedings of the Workshop

June 20, 2024

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 979-8-89176-103-2

Introduction

Welcome to the 21st SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, to be held on June 20, 2024 as part of NAACL in Mexico City, Mexico. The workshop aims to bring together researchers interested in applying computational techniques to problems in morphology, phonology, and phonetics. Our program this year highlights the ongoing investigations into how neural and other learning models process phonology and word structure. We also publish work on a number of new datasets and resources in morphology.

We received 15 submissions, and after a competitive reviewing process, we accepted 9. The workshop is privileged to present two invited talks this year. Jian Zhu (University of British Columbia) and Naomi Feldman (University of Maryland) presented talks at this year's workshop.

Garrett Nicolai, Eleanor Chodroff, Çagri Çöltekin and Fred Mailhot, workshop organization team.

Organizing Committee

Co-Chair

Garrett Nicolai, University of British Columbia
Eleanor Chodroff, University of York
Çagri Çöltekin, University of Tübingen
Fred Mailhot, Dialpad, Inc.

SIGMORPHON Officers

President: Garrett Nicolai, University of British Columbia
Secretary: Miikka Silfverberg, University of British Columbia
At Large: Eleanor Chodroff, University of York
At Large: Çağrı Çöltekin, University of Tübingen
At Large: Fred Mailhot, Dialpad, Inc.

Program Committee

Reviewers

Brian Roark, Google Inc.
Aniello De Santo, University of Utah
Ekaterina Vylomova, University of Melbourne
Michael Ginn, University of Colorado
Kenneth Steimel, Cisco Systems Incorporated
Jelena Prokic, Leiden University
Nizar Habash, New York University Abu Dhabi
Khuyagbaatar Batsuren, National University of Mongolia
Sandra Kübler, Indiana University
Changbing Yang, University of British Columbia
Adam Wiemerslage, University of Colorado Boulder
Kemal Oflazer, Carnegie Mellon University
Sarah Moeller, University of Florida
Cassandra L. Jacobs, University at Buffalo
Rob Malouf, San Diego State University
Nabil Hathout, CLLE, CNRS and Université de Toulouse
Kate McCurdy, University of Edinburgh
Mathilde Hutin, Université Paris-Saclay, CNRS, LIMSI
Micha Elsner, The Ohio State University
Daniel Dakota, Indiana University
Morgan Sonderegger, McGill University
Kristine Yu, University of Massachusetts Amherst
Çağrı Çöltekin, University of Tübingen
Giorgio Magri, Centre National de la Recherche Scientifique
Indranil Dutta, Jadavpur University
Ewan Dunbar, University of Toronto

Keynote Talk

Invited Talk 1

Jian Zhu
University of British Columbia
2024-06-20 09:00:00 –

Abstract: Towards crosslinguistically generalizable speech technologies The diversity of human speech presents a formidable challenge to multilingual speech processing systems. Recently, accumulating evidence indicated that scaling up multilingual data and model parameters can tremendously improve the performance of multilingual speech processing. However, gathering large-scale data from every language in the world is an impossible mission. To tackle this challenge, my research group aims to develop multilingual speech processing systems that generalize to unseen and low-resource languages. Since most, if not all, human speech can be represented by around 150 phonetic symbols and diacritics, I argue that using International Phonetic Alphabet (IPA) as modeling units, rather than orthographic transcriptions, enables speech models to process and recognize sounds in unseen languages. In the past years, leveraging IPA, large-scale multilingual corpora and deep learning, my research team has built a series of massively multilingual speech datasets and technologies including multilingual grapheme-to-phoneme conversion, multilingual keyword spotting, multilingual forced alignment and multilingual phone recognition systems. In this talk, I will introduce our recent works towards crosslinguistically generalizable speech technologies and lessons we learned from working with a diversity of languages.

Bio: Jian Zhu is currently an assistant professor in the Linguistics Department at the University of British Columbia. He is primarily interested in developing multilingual speech and language technologies for low resource and zero resource languages. Trained as both a linguist and an engineer, he combines linguistic theories with data-driven methods in speech processing, natural language processing, network science and machine learning. Before that, he was a post-doctoral research fellow at Blablablab at the School of Information, University of Michigan. He obtained his Ph.D. in Linguistics and Scientific Computing from the Department of Linguistics and the Michigan Institute for Computational Discovery & Engineering at the University of Michigan.

Keynote Talk
Invited Talk 2

Naomi Feldman
University of Maryland
2024-06-20 14:30:00 –

Abstract: Modeling speech perception at scale Speech processing is a perfect test case for scaling up cognitive modeling. Recent advances in speech technology provide new tools that can be leveraged to better understand how human listeners perceive speech in naturalistic settings. At the same time, building cognitive models of human speech perception can highlight capabilities that are not yet captured by standard representation learning models in speech technology.

I begin by showing how incorporating unsupervised representation learning into cognitive models of speech perception can impact theories of early language acquisition. Infants' patterns of speech perception have traditionally been interpreted as evidence that they possess certain types of knowledge, such as phonetic categories (like 'r' and 'l') and representations of speech rhythm, but our cognitive modeling results point toward a different interpretation. If correct, this could radically change our view of how phonetic knowledge supports infants' acquisition of words and grammar, and could have broad implications for understanding the challenges associated with learning a new language in adulthood. I then outline ongoing work exploring the mechanisms that could support and, eventually, reproduce human listeners' ability to flexibly adapt to different accents and listening conditions. Together, these studies illustrate how speech representations can be optimized over short and long time scales to support robust speech processing.

This is joint work with Thomas Schatz, Yevgen Matusevych, Ruolan (Leslie) Famularo, Nika Jurov, Ali Aboelata, Grayson Wolf, Xuan-Nga Cao, Herman Kamper, William Idsardi, Emmanuel Dupoux, and Sharon Goldwater.

Bio: Naomi Feldman is an associate professor in the Department of Linguistics and the Institute for Advanced Computer Studies at the University of Maryland, where she is a member and former director of the Computational Linguistics and Information Processing (CLIP) Lab. Her research uses methods from machine learning and automatic speech recognition to formalize questions about how people learn and represent the structure of their language. She primarily uses these methods to study speech representations, modeling the cognitive processes that support learning and perception of speech sounds in the face of highly complex and variable linguistic input. She also computationally characterizes the strategies that facilitate language acquisition more generally, both from the perspective of learners, and from the perspective of clinicians.

Table of Contents

Program

Thursday, June 20, 2024

09:25 - 09:30 *Opening Remarks*

09:30 - 10:30 *Invited Talk 1":Jian Zhu*

10:30 - 11:00 *Break*

11:00 - 12:00 *Session 1*

J-UniMorph":Japanese Morphological Annotation through the Universal Feature Schema
Kosuke Matsuzaki, Masaya Taniguchi, Kentaro Inui and Keisuke Sakaguchi

Ye Olde French":Effect of Old and Middle French on SIGMORPHON-UniMorph Shared Task Data
William Kezerian and Kristine Yu

VeLePa":a Verbal Lexicon of Pame
Borja Herce

12:00 - 13:00 *Lunch*

13:00 - 14:00 *Session 2*

Acoustic barycenters as exemplar production targets
Frederic Mailhot and Cassandra L. Jacobs

Japanese Rule-based Grapheme-to-phoneme Conversion System and Multilingual Named Entity Dataset with International Phonetic Alphabet
Yuhi Matogawa, Yusuke Sakai, Taro Watanabe and Chihiro Taguchi

Decomposing Fusional Morphemes with Vector Embeddings
Michael Ginn and Alexis Palmer

14:00 - 15:00 *Invited Talk 2":Naomi Feldman*

Thursday, June 20, 2024 (continued)

15:00 - 15:30 *Session 3 (ACL Findings)*

Tokenization Matters":Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies
Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter and Rahul Gupta

Low-resource neural machine translation with morphological modeling
Antoine Nzeyimana

15:30 - 16:00 *Break*

16:00 - 17:00 *Session 4*

Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement
Catherine Arnett, Tyler Chang and Sean Trott

The Effect of Model Capacity and Script Diversity on Subword Tokenization for Sorani Kurdish
Ali Salehi and Cassandra L. Jacobs

More than Just Statistical Recurrence":Human and Machine Unsupervised Learning of Māori Word Segmentation across Morphological Processes
Ashvini Varatharaj and Simon Todd