SIGMORPHON 2021

18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology

Proceedings of the Workshop

August 5, 2021
Bangkok, Thailand (online)

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 978-1-954085-62-6

Preface

Welcome to the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, to be held on August 5, 2021 as part of a virtual ACL. 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 models process phonology and morphology, as well as the development of finite-state models for low-resource languages with complex morphology. .

We received 25 submissions, and after a competitive reviewing process, we accepted 14.

The workshop is privileged to present four invited talks this year, all from very respected members of the SIGMORPHON community. Reut Tsarfaty, Kenny Smith, Kristine Yu, and Ekaterina Vylomova all presented talks at this year's workshop.

This year also marks the sixth iteration of the SIGMORPHON Shared Task. Following upon the success of last year's multiple tasks, we again hosted 3 shared tasks:

Task 0:

SIGMORPHON's sixth installment of its inflection generation shared task is divided into two parts: Generalization, and cognitive plausibility.

In the first part, participants designed a model that learned to generate morphological inflections from a lemma and a set of morphosyntactic features of the target form, similar to previous year's tasks. This year, participants learned morphological tendencies on a set of development languages, and then generalized these findings to new languages - without much time to adapt their models to new phenomena.

The second part asks participants to inflect nonce words in the past tense, which are then judged for plausibility by native speakers. This task aims to investigate whether state-of-the-art inflectors are learning in a way that mimics human learners.

Task 1:

The second SIGMORPHON shared task on grapheme-to-phoneme conversion expands on the task from last year, recategorizing data as belonging to one of three different classes: low-resource, medium-resource, and high-resource.

The task saw 23 submissions from 9 participants.

Task 2:

Task 2 continues the effort from the 2020 shared task in unsupervised morphology. Unlike last year's task, which asked participants to implement a complete unsupervised morphology induction pipeline, this year's task concentrates on a single aspect of morphology discovery: paradigm induction. This task asks participants to cluster words into inflectional paradigms, given no more than raw text.

The task saw 14 submissions from 4 teams.

We are grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year. Likewise, we are thankful to the shared task organizers for their hard work in preparing the shared tasks. We are looking forward to a workshop covering a wide range of topics, and we hope for lively discussions.

Garrett Nicolai
Kyle Gorman

Organizing Committee

Garrett Nicolai (University of British Columbia, Canada)
Kyle Gorman (City University of New York, USA)
Ryan Cotterell (ETH Zürich, Switzerland)

Program Committee

Damián Blasi (Harvard University)
Grzegorz Chrupała (Tilburg University)
Jane Chandlee (Haverford College)
Çağrı Çöltekin (University of Tübingen)
Daniel Dakota (Indiana University)
Colin de la Higuera (University of Nantes)
Micha Elsner (The Ohio State University)
Nizar Habash (NYU Abu Dhabi)
Jeffrey Heinz (University of Delaware)
Mans Hulden (University of Colorado)
Adam Jardine (Rutgers University)
Christo Kirov (Google AI)
Greg Kobele (Universität Leipzig)
Grzegorz Kondrak (University of Alberta)
Sandra Kübler (Indiana University)
Adam Lamont (University of Massachusetts Amherst)
Kevin McMullin (University of Ottawa)
Kemal Oflazer (CMU Qatar)
Jeff Parker (Brigham Young University)
Gerald Penn (University of Toronto)
Jelena Prokic (Universiteit Leiden)
Miikka Silfverberg (University of British Columbia)
Kairit Sirts (University of Tartu)
Kenneth Steimel (Indiana University)
Reut Tsarfaty (Bar-Ilan University)
Francis Tyers (Indiana University)
Ekaterina Vylomova (University of Melbourne)
Adina Williams (Facebook AI Research)
Anssi Yli-Jyrä (University of Helsinki)
Kristine Yu (University of Massachusetts)

Task 0 Organizing Committee

Tiago Pimentel (University of Cambridge)
Brian Leonard (Brian Leonard Consulting)
Maria Ryskina (Carnegie Mellon University)
Sabrina Mielke (Johns Hopkins University)
Coleman Haley (Johns Hopkins University)
Eleanor Chodroff (University of York)
Johann-Mattis List (Max Planck Institute)
Adina Williams (Facebook AI Research)
Ryan Cotterell (ETH Zürich)
Ekaterina Vylomova (University of Melbourne)
Ben Ambridge (University of Liverpool)

Task 1 Organizing Committee

To come

Task 2 Organizing Committee

Adam Wiemerslage(University of Colorado Boulder)
Arya McCarthy (Johns Hopkins University)
Alexander Erdmann (Ohio State University)
Manex Agirrezabal (University of Copenhagen)
Garrett Nicolai (University of British Columbia)
Miikka Silfverberg (University of British Columbia)
Mans Hulden (University of Colorado Boulder)
Katharina Kann (University of Colorado Boulder)

Table of Contents

Towards Detection and Remediation of Phonemic Confusion
Francois Roewer-Despres, Arnold Yeung and Ilan Kogan 1

Recursive prosody is not finite-state
Hossep Dolatian, Aniello De Santo and Thomas Graf 11

The Match-Extend serialization algorithm in Multiprecedence
Maxime Papillon 23

Incorporating tone in the calculation of phonotactic probability
James Kirby 32

MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology
Khuyagbaatar Batsuren, Gábor Bella and fausto giunchiglia 39

A Study of Morphological Robustness of Neural Machine Translation
Sai Muralidhar Jayanthi and Adithya Pratapa 49

Sample-efficient Linguistic Generalizations through Program Synthesis: Experiments with Phonology Problems
Saujas Vaduguru, Aalok Sathe, Monojit Choudhury and Dipti Sharma 60

Findings of the SIGMORPHON 2021 Shared Task on Unsupervised Morphological Paradigm Clustering
Adam Wiemerslage, Arya D. McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden and Katharina Kann 72

Adaptor Grammars for Unsupervised Paradigm Clustering
Kate McCurdy, Sharon Goldwater and Adam Lopez 82

Orthographic vs. Semantic Representations for Unsupervised Morphological Paradigm Clustering
E. Margaret Perkoff, Josh Daniels and Alexis Palmer 90

Unsupervised Paradigm Clustering Using Transformation Rules
Changbing Yang, Garrett Nicolai and Miikka Silfverberg 98

Paradigm Clustering with Weighted Edit Distance
Andrew Gerlach, Adam Wiemerslage and Katharina Kann 107

Results of the Second SIGMORPHON Shared Task on Multilingual Grapheme-to-Phoneme Conversion
Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor and Winnie Yan 115

Data augmentation for low-resource grapheme-to-phoneme mapping
Michael Hammond 126

Linguistic Knowledge in Multilingual Grapheme-to-Phoneme Conversion
Roger Yu-Hsiang Lo and Garrett Nicolai 131

Avengers, Ensemble! Benefits of ensembling in grapheme-to-phoneme prediction
Vasundhara Gautam, Wang Yau Li, Zafarullah Mahmood, Frederic Mailhot, Shreekantha Nadig, Riqiang WANG and Nathan Zhang 141

CLUZH at SIGMORPHON 2021 Shared Task on Multilingual Grapheme-to-Phoneme Conversion: Variations on a Baseline
Simon Clematide and Peter Makarov 148

What transfers in morphological inflection? Experiments with analogical models
Micha Elsner 154

Simple induction of (deterministic) probabilistic finite-state automata for phonotactics by stochastic gradient descent
Huteng Dai and Richard Futrell 167

Recognizing Reduplicated Forms: Finite-State Buffered Machines
Yang Wang 177

An FST morphological analyzer for the Gitksan language
Clarissa Forbes, Garrett Nicolai and Miikka Silfverberg 188

Comparative Error Analysis in Neural and Finite-state Models for Unsupervised Character-level Transduction
Maria Ryskina, Eduard Hovy, Taylor Berg-Kirkpatrick and Matthew R. Gormley 198

Finite-state Model of Shupamem Reduplication
Magdalena Markowska, Jeffrey Heinz and Owen Rambow 212

Improved pronunciation prediction accuracy using morphology
Dravyansh Sharma, Saumya Sahai, Neha Chaudhari and Antoine Bruguier 222

Workshop Program

Due to the ongoing pandemic, and the virtual nature of the workshop, the papers will be presented asynchronously, with designated question periods.

Towards Detection and Remediation of Phonemic Confusion
Francois Roewer-Despres, Arnold Yeung and Ilan Kogan

Recursive prosody is not finite-state
Hossep Dolatian, Aniello De Santo and Thomas Graf

The Match-Extend serialization algorithm in Multiprecedence
Maxime Papillon

What transfers in morphological inflection? Experiments with analogical models
Micha Elsner

Simple induction of (deterministic) probabilistic finite-state automata for phonotactics by stochastic gradient descent
Huteng Dai and Richard Futrell

Incorporating tone in the calculation of phonotactic probability
James Kirby

Recognizing Reduplicated Forms: Finite-State Buffered Machines
Yang Wang

An FST morphological analyzer for the Gitksan language
Clarissa Forbes, Garrett Nicolai and Miikka Silfverberg

MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology
Khuyagbaatar Batsuren, Gábor Bella and Fausto Giunchiglia

Comparative Error Analysis in Neural and Finite-state Models for Unsupervised Character-level Transduction
Maria Ryskina, Eduard Hovy, Taylor Berg-Kirkpatrick and Matthew R. Gormley

Finite-state Model of Shupamem Reduplication
Magdalena Markowska, Jeffrey Heinz and Owen Rambow

A Study of Morphological Robustness of Neural Machine Translation
Sai Muralidhar Jayanthi and Adithya Pratapa

No Day Set (continued)

Sample-efficient Linguistic Generalizations through Program Synthesis: Experiments with Phonology Problems
Saujas Vaduguru, Aalok Sathe, Monojit Choudhury and Dipti Sharma

Improved pronunciation prediction accuracy using morphology
Dravyansh Sharma, Saumya Sahai, Neha Chaudhari and Antoine Bruguier

Data augmentation for low-resource grapheme-to-phoneme mapping
Michael Hammond

Linguistic Knowledge in Multilingual Grapheme-to-Phoneme Conversion
Roger Yu-Hsiang Lo and Garrett Nicolai

Avengers, Ensemble! Benefits of ensembling in grapheme-to-phoneme prediction
Vasundhara Gautam, Wang Yau Li, Zafarullah Mahmood, Frederic Mailhot, Shreekantha Nadig, Riqiang WANG and Nathan Zhang

CLUZH at SIGMORPHON 2021 Shared Task on Multilingual Grapheme-to-Phoneme Conversion: Variations on a Baseline
Simon Clematide and Peter Makarov

Findings of the SIGMORPHON 2021 Shared Task on Unsupervised Morphological Paradigm Clustering
Adam Wiemerslage, Arya D. McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden and Katharina Kann

Orthographic vs. Semantic Representations for Unsupervised Morphological Paradigm Clustering
E. Margaret Perkoff, Josh Daniels and Alexis Palmer

Unsupervised Paradigm Clustering Using Transformation Rules
Changbing Yang, Garrett Nicolai and Miikka Silfverberg

Paradigm Clustering with Weighted Edit Distance
Andrew Gerlach, Adam Wiemerslage and Katharina Kann

Adaptor Grammars for Unsupervised Paradigm Clustering
Kate McCurdy, Sharon Goldwater and Adam Lopez