NLP4DH and IWCLUL 2023

The Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages

Proceedings of the Conference

December 1-3, 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 979-8-89176-012-7

Preface

Textual sources are essential for research in digital humanities. Especially when larger datasets are analyzed, the use of natural language processing (We are delighted to welcome participants to the unique and pioneering hybrid event that combines the International Conference on Natural Language Processing for Digital Humanities (NLP4DH) and the International Workshop on Computational Linguistics for Uralic Languages (IWCLUL). This year marks a significant milestone as we bring together two vibrant communities under one umbrella, fostering an interdisciplinary dialogue and collaboration between digital humanities and computational linguistics, with a special focus on Uralic languages and broader linguistic diversity.

The NLP4DH, in its previous edition, highlighted the crucial role of NLP technologies in addressing the specific needs of digital humanists. The focus was on the application of NLP in exploring non-standard languages, dialects, and historical texts, areas that are often overlooked in mainstream NLP research. The event underscored the importance of bridging the gap between the methodological rigor of NLP and the concrete, data-driven inquiries of digital humanities. This year, we continue to emphasize the synergy between these fields, exploring how advanced NLP tools and methods can be fine-tuned and retrained to better serve the nuanced requirements of humanities research.

On the other hand, IWCLUL has been a cornerstone in the study and preservation of Uralic languages, offering insights into traditional language technology resources and modern computational approaches. The previous editions showcased a diverse range of research, from language-specific studies to comparative analyses across the Uralic language family. This year, in conjunction with NLP4DH, IWCLUL aims to extend its reach and impact, exploring how computational linguistics can contribute to the preservation, understanding, and development of Uralic and other minority languages.

The joint event is a testament to our commitment to interdisciplinary research and the recognition of the importance of linguistic diversity in computational studies. We have a rich program that includes high-quality submissions from both NLP4DH and IWCLUL communities. The presentations range from innovative NLP applications in the humanities to cutting-edge computational techniques in Uralic language studies.

We are particularly excited about the potential outcomes of this collaboration. The intersection of digital humanities and computational linguistics, especially in the context of less-researched languages, opens up new avenues for research and application. We anticipate that this event will not only contribute to academic discourse but also pave the way for practical solutions that benefit language communities, researchers, and practitioners alike.

We extend our heartfelt thanks to all contributors, participants, and organizers who have worked tirelessly to make this event a reality. Your enthusiasm and dedication are the driving forces behind this successful collaboration. We look forward to the fruitful discussions, innovative ideas, and new partnerships that will emerge from this unique gathering.

Welcome to the joint NLP4DH and IWCLUL event – a convergence of digital humanities and computational linguistics, celebrating linguistic diversity and interdisciplinary research.

This event is organized in collaboration with SIGUR, ACL Special Interest Group for Uralic Languages.

Organizing Committee

Organizers (NLP4DH)

Mika Hämäläinen, Metropolia University of Applied Sciences
Emily Öhman, Waseda University
Khalid Alnajjar, Rootroo Ltd
So Miyagawa, National Institute for Japanese Language and Linguistics
Yuri Bizzoni, Aarhus University

Organizers (IWCLUL)

Flammie Pirinen, UiT The Arctic University of Norway
Niko Partanen, University of Helsinki
Jack Rueter, University of Helsinki

Program Committee

Reviewers

Aynat Rubinstein, The Hebrew University of Jerusalem
Leo Leppänen, University of Helsinki
Kenichi Iwatsuki, KTTA
Lidia Pivovarova, University of Helsinki
Linda Wiechetek, University of Tromsø
Jouni Tuominen, University of Helsinki
Mikko Kurimo, Aalto University
Balázs Indig, Eötvös Lorand University
Pierre Magistry, Institut National des Langues et Civilisations Orientales
Yoshifumi Kawasaki, The University of Tokyo
Eetu Mäkelä, University of Helsinki
Timofey Arkhangelskiy, Universität Hamburg
Nicolas Gutehrlé, Université de Franche-Comté
Kaisla Kajava, Aalto University
Joshua Wilbur, University of Tartu
Pascale Moreira, School of Communication and Culture
Miikka Silfverberg, University of British Columbia
Francis Tyers, Indiana University
Anna Dmitrieva, University of Helsinki
Somesh Mohapatra, Massachusetts Institute of Technology
Won Ik Cho, Samsung Advanced Institute of Technology
Shuo Zhang, Bose Corp
Pihla Toivanen, University of Helsinki
Antti Kanner, University of Turku
Jeremy Bradley, Universität Vienna
Aatu Liimatta, University of Helsinki
Sijia Ge, University of Colorado Boulder
Michael Rießler, University of Eastern Finland
Irene Russo, Consiglio Nazionale delle Ricerche
Gechuan Zhang, University College Dublin
Maria Antoniak, Allen Institute for Artificial Intelligence
Thomas Schmidt, Universität Regensburg
Juho Pääkkönen, University of Helsinki
Rogier Blokland, Uppsala University
Jenna Kanerva, University of Turku
Katerina Korre, University of Bologna
Mikko Aulamo, University of Helsinki
Mitsunori Ogihara, University of Miami
Miu Takagi, Waseda University
Quan Duong, University of Helsinki
Daniela Teodorescu, University of Alberta
Erkki Mervaala, Finnish Environment Institute
Joachim Scharloth, Waseda University
Dimosthenis Antypas, Cardiff University
Ayana Niwa, Tokyo Institute of Technology
Heiki-Jaan Kaalep, institute of computer science

Shu Okabe, Univ. Paris-Saclay
Dmitry Nikolaev, University of Stuttgart
Ritwik Bose, The Institute for Human & Machine Cognition
Dongqi Pu, Universität des Saarlandes
Aina Garí, Télécom-Paris
Nils Hjortnaes, Indiana University
László Fejes, Hungarian Research Centre for Linguistics
Ligeti-Nagy Noémi, MTA-PPKE
Tulika Bose, Vivoka
Allison Lahnala, Phillips-Universität Marburg
Gabriel Simmons, University of California
Vilja Hulden, University of Colorado at Boulder
Federico Boschetti, CNR-ILC

Table of Contents

Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K): A Parallel Corpus for Explaining (Im)morality of Actions
Jasper Kyle Catapang and Moses Visperas 1

A Quantitative Discourse Analysis of Asian Workers in the US Historical Newspapers
Jaihyun Park and Ryan Cordell 7

Revisiting Authorship Attribution of Tirant lo Blanc Using Parts of Speech n-grams
Yoshifumi Kawasaki 16

Translation from Historical to Contemporary Japanese Using Japanese T5
Hisao Usui and Kanako Komiya 27

Measuring the distribution of Hume's Scotticisms in the ECCO collection
Iiro Tiihonen, Aatu Liimatta, Lidia Pivovarova, Tanja Säily and Mikko Tolonen 36

Effect of data quality on the automated identification of register features in Eighteenth Century Collections Online
Aatu Liimatta 45

Automated Generation of Multiple-Choice Cloze Questions for Assessing English Vocabulary Using GPT-turbo 3.5
Qiao Wang, Ralph Rose, Naho Orita and Ayaka Sugawara 52

Explicit References to Social Values in Fairy Tales: A Comparison between Three European Cultures
Alba Morollon Diaz-Faes, Carla Murteira and Martin Ruskov 62

The Stylometry of Maoism: Quantifying the Language of Mao Zedong
Maciej Kurzynski 76

Efficient and reliable utilization of automated data collection applied to news on climate change
Erkki Mervaala and Jari Lyytimäki 82

Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts
Baptiste Blouin, Hen-Hsen Huang, Christian Henriot and Cécile Armand 92

Introducing ChatGPT to a researcher's toolkit: An empirical comparison between rule-based and large language model approach in the context of qualitative content analysis of political texts in Finnish
Ilona Kousa 102

Fly, fly little Comet! Exploring Subtoken-Level Metaphorical Patterns in Finnish and Hungarian Texts. New Results from the FiHuComet Corpus.
Tímea Borbála Bajzát 114

Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan
So Miyagawa 120

Understanding Gender Stereotypes in Video Game Character Designs: A Case Study of Honor of Kings
Bingqing Liu, Kyrie Zhixuan Zhou, Danlei Zhu and Jaihyun Park 125

The Great Digital Humanities Disconnect: The Failure of DH Publishing
Emily Öhman, Michael Piotrowski and Mika Hämäläinen 132

Explorative study on verbalizing students' skills with NLP/AI-tool in Digital Living Lab at Laurea UAS, Finland
Asko Mononen 138

Combating Hallucination and Misinformation: Factual Information Generation with Tokenized Generative Transformer
Sourav Das, Sanjay Chatterji and Imon Mukherjee 143

Statistical Measures for Readability Assessment
Mohammed Attia, Younes Samih and Yo Ehara 153

A Question of Confidence: Using OCR Technology for Script analysis
Antonia Karaisl 162

Emil.RuleZ! – An exploratory pilot study of handling a real-life longitudinal email archive
Balázs Indig, Luca Horváth, Dorottya Henrietta Szemigán and Mihály Nagy 172

Banning of ChatGPT from Educational Spaces: A Reddit Perspective
Nicole Miu Takagi 179

Girlbosses, The Red Pill, and the Anomie and Fatale of Gender Online: Analyzing Posts from r/SuicideWatch on Reddit
Elissa Nakajima Wickham 195

Bootstrapping Moksha-Erzya Neural Machine Translation from Rule-Based Apertium
Khalid Alnajjar, Mika Hämäläinen and Jack Rueter 213

Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study
Yuri Bizzoni and Pascale Feldkamp 219

Study on the Domain Adaption of Korean Speech Act using Daily Conversation Dataset and Petition Corpus
Youngsook Song and Won Ik Cho 229

Readability and Complexity: Diachronic Evolution of Literary Language Across 9000 Novels
Pascale Feldkamp, Yuri Bizzoni, Ida Marie S. Lassen, Mads Rosendahl Thomsen and Kristoffer Nielbo 235

Bridging the Gap: Demonstrating the Applicability of Linguistic Analysis Tools in Digital Musicology
Sebastian Oliver Eck 248

MITRA-zh: An efficient, open machine translation solution for Buddhist Chinese
Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody and Kurt Keutzer 266

Comparison on Heterosexual and Homosexual Woman's Lonely Heart Ads in Taiwan: Taking AllTogether and Lesbian Board on PTT Web Forum as Examples
Yu-Hsuan Lin 278

Program

Friday, December 1, 2023

09:30 - 10:00 *Opening and Lightning Talks*

10:00 - 10:30 *Comparison on Heterosexual and Homosexual Woman's Lonely Heart Ads in Taiwan: Taking AllTogether and Lesbian Board on PTT Web Forum as Examples*

10:30 - 11:00 *Automated Generation of Multiple-Choice Cloze Questions for Assessing English Vocabulary Using GPT-turbo 3.5*

11:00 - 11:30 *Banning of ChatGPT from Educational Spaces: A Reddit Perspective*

11:30 - 12:00 *Bootstrapping Moksha-Erzya Neural Machine Translation from Rule-Based Apertium*

12:00 - 13:00 *Lunch*

13:00 - 13:30 *A Question of Confidence: Using OCR Technology for Script analysis*

13:30 - 14:00 *An efficient, open machine translation solution for Buddhist Chinese*

14:00 - 14:30 *Translation from Historical to Contemporary Japanese Using Japanese T5*

14:30 - 14:45 *Coffee break*

14:45 - 15:15 *Fly, fly little Comet! Exploring Subtoken-Level Metaphorical Patterns in Finnish and Hungarian Texts. New Results from the FiHuComet Corpus.*

15:15 - 15:45 *Emil.RuleZ! – An exploratory pilot study of handling a real-life longitudinal email archive*

15:45 - 16:15 *Introducing ChatGPT to a researcher's toolkit: An empirical comparison between rule-based and large language model approach in the context of qualitative content analysis of political texts in Finnish*

16:15 - 16:45 *Efficient and reliable utilization of automated data collection applied to news on climate change*

Saturday, December 2, 2023

09:30 - 10:30 *Keynote: Kyo Kageura*

10:30 - 11:00 *The Stylometry of Maoism: Quantifying the Language of Mao Zedong*

11:00 - 11:30 *Explorative study on verbalizing students' skills with NLP/AI-tool in Digital Living Lab at Laurea UAS, Finland*

11:30 - 12:00 *The Great Digital Humanities Disconnect: The Failure of DH Publishing*

12:00 - 13:00 *Lunch*

13:00 - 13:30 *Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K): A Parallel Corpus for Explaining (Im)morality of Actions*

13:30 - 14:00 *Understanding Gender Stereotypes in Video Game Character Designs: A Case Study of Honor of Kings*

14:00 - 14:30 *Girlbosses, The Red Pill, and the Anomie and Fatale of Gender Online: Analyzing Posts from r/SuicideWatch on Reddit*

14:30 - 14:45 *Coffee break*

14:45 - 15:15 *Revisiting Authorship Attribution of Tirant lo Blanc Using Parts of Speech n-grams*

15:15 - 15:45 *Explicit References to Social Values in Fairy Tales: A Comparison between Three European Cultures*

15:45 - 16:15 *Readability and Complexity: Diachronic Evolution of Literary Language Across 9000 Novels*

16:15 - 16:30 *Coffee break*

16:30 - 17:30 *SIGUR Business Meeting*

Sunday, December 3, 2023

09:30 - 10:00 *Keynote: Maria Antonia*

10:30 - 11:00 *Statistical Measures for Readability Assessment*

11:00 - 11:30 *Study on the Domain Adaption of Korean Speech Act using Daily Conversation Dataset and Petition Corpus*

11:30 - 12:00 *Bridging the Gap: Demonstrating the Applicability of Linguistic Analysis Tools in Digital Musicology*

12:00 - 13:00 *Lunch*

13:00 - 13:30 *Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan*

13:30 - 14:00 *Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts*

14:00 - 14:30 *Combating Hallucination and Misinformation: Factual Information Generation with Tokenized Generative Transformer*

14:30 - 14:45 *Coffee break*

14:45 - 15:15 *A Quantitative Discourse Analysis of Asian Workers in the US Historical Newspapers*

15:15 - 15:45 *Measuring the distribution of Hume's Scotticisms in the ECCO collection*

15:45 - 16:15 *Effect of data quality on the automated identification of register features in Eighteenth Century Collections Online*

16:15 - 16:45 *Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study*