NLP4DH 2021

The 2nd International Workshop on Natural Language Processing for Digital Humanities

Proceedings of the Workshop

November 20, 2022

©2022 The Association for Computational Linguistics

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 978-1-955917-75-9

Preface

Textual sources are essential for research in digital humanities. Especially when larger datasets are analyzed, the use of natural language processing (NLP) technologies is essential. However, NLP is still often focused to written standard languages, which customarily differs from specific genres and text types that may interest a digital humanist today. The situation is even more complicated when the research is done on minority languages, or historical and dialectal materials.

Natural language processing has usually a strong computer science focus, which means that methods are developed to cater for higher numerical results and to solve some rather abstract level tasks such as machine translation, poem generation or sentiment analysis. Digital humanities, on the other hand, has usually a strong humanities focus which means that the research questions are typically more concrete, diving deeper to understanding some phenomena rather than solving a problem. Natural language processing also seeks to validate the methods, whereas digital humanities takes the validity of the methods for granted. This is due to the fact that a method is often the end goal in natural language processing, where as a method is just a tool in the digital humanities. The two fields work from very different starting points, and therefore we believe that more venues are needed where scholars from both fields can come together and learn from each other.

We believe that digital humanists recognize the shortcomings of the contemporary natural language processing tools, and the NLP community has already come up with various fully functional solutions. However, these communities would benefit from further communication. For example, model fine tuning and retraining are among useful technologies in NLP that could be applied to efficiently improve the result on these divergent varieties. Similarly work in digital humanities often results in open datasets that could be used to compare different strategies. In this workshop we aimed to foster and initiate wider conversation and sharing of examples of how NLP tools are best leveraged to the research questions that are relevant in humanities.

The Workshop on Natural Language Processing for Digital Humanities (NLP4DH) was organized for the second time in November 20, 2022 with AACL IJCNLP 2022: The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Our workshop received a plethora of submissions, out of which 22 were accepted to be presented in the workshop. We are especially excited about the upcoming special issue in the Journal of Data Mining & Digital Humanities that will feature extended versions of some of the papers accepted in the workshop.

https://rootroo.com

Organizing Committee

- Mika Hämäläinen, University of Helsinki and Rootroo Ltd
- Khalid Alnajjar, University of Helsinki and Rootroo Ltd
- Niko Partanen, University of Helsinki
- Jack Rueter, University of Helsinki
- Thierry Poibeau, École normale supérieure and CNRS

Program Committee

- Iana Atanassova, Université de Bourgogne Franche-Comté
- Yuri Bizzoni, Aarhus University
- Miriam Butt, University of Konstanz
- Won Ik Cho, Seoul National University
- Quan Duong, University of Helsinki
- Hugo Gonçalo Oliveira, University of Coimbra
- Kenichi Iwatsuki, ARIKTTA
- Heiki-Jaan Kaalep, University of Tartu
- Enrique Manjavacas, Leiden University
- Matej Martinc, Jozef Stefan Institute
- Flammie Pirinen, UiT The Arctic University of Norway
- Tyler Shoemaker, University of California, Davis
- Liisa Lotta Tarvainen-Li, Acolad
- Jörg Tiedemann, University of Helsinki
- Jouni Tuominen, Aalto University
- Shuo Zhang, Bose Corporation
- Emily Öhman, Waseda University
- Frederik Arnold, Humboldt-Universität zu Berlin
- Nicolas Gutehrlé, Université de Bourgogne Franche-Comté
- Thibault Clérice, Université PSL
- Aynat Rubinstein, The Hebrew University of Jerusalem
- Lama Alqazlan, University of Warwick
- Gechuan Zhang, University College Dublin
- Moshe Stekel, Ariel University
- Alejandro Sierra-Múnera, University of Potsdam
- Avinash Tulasi, IIIT Delhi

Table of Contents

Conference Program

Sunday, November 20, 2022

17:00–17:00 Workshop opening

17:00–18:30 Poster session 1

17:00–18:30 *A Stylometric Analysis of Amadís de Gaula and Sergas de Esplandián*
Yoshifumi Kawasaki

17:00–18:30 *Computational Exploration of the Origin of Mood in Literary Texts*
Emily Öhman and Riikka H. Rossi

17:00–18:30 *Sentiment is all you need to win US Presidential elections*
Sovesh Mohapatra and Somesh Mohapatra

17:00–18:30 *Interactive Analysis and Visualisation of Annotated Collocations in Spanish (AVAnCES)*
Simon Gonzalez

17:00–18:30 *Fractality of sentiment arcs for literary quality assessment: The case of Nobel laureates*
Yuri Bizzoni, Kristoffer Laigaard Nielbo and Mads Rosendahl Thomsen

17:00–18:30 *Style Classification of Rabbinic Literature for Detection of Lost Midrash Tanhuma Material*
Solomon Tannor, Nachum Dershowitz and Moshe Lavee

17:00–18:30 *Use the Metadata, Luke! – An Experimental Joint Metadata Search and N-gram Trend Viewer for Personal Web Archives*
Balázs Indig, Zsófia Sárközi-Lindner and Mihály Nagy

Sunday, November 20, 2022 (continued)

18:30–19:30 Lunch break

19:30–21:00 Poster session 2

19:30–21:00 *MALM: Mixing Augmented Language Modeling for Zero-Shot Machine Translation*
Kshitij Gupta

19:30–21:00 *ParsSimpleQA: The Persian Simple Question Answering Dataset and System over Knowledge Graph*
Hamed Babaei Giglou, Niloufar Beyranvand, Reza Moradi, Amir Mohammad Salehoof and Saeed Bibak

19:30–21:00 *Enhancing Digital History – Event discovery via Topic Modeling and Change Detection*
King Ip Lin and Sabrina Peng

19:30–21:00 *A Parallel Corpus and Dictionary for Amis-Mandarin Translation*
Francis Zheng, Edison Marrese-Taylor and Yutaka Matsuko

19:30–21:00 *Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers*
Nilo Pedrazzini and Barbara McGillivray

19:30–21:00 *Optimizing the weighted sequence alignment algorithm for large-scale text similarity computation*
Maciej Janicki

19:30–21:00 *Domain-specific Evaluation of Word Embeddings for Philosophical Text using Direct Intrinsic Evaluation*
Goya van Boven and Jelke Bloem

Sunday, November 20, 2022 (continued)

21:00–21:30 Coffee break

21:30–23:00 Poster session 3

21:30–23:00 *Towards Bootstrapping a Chatbot on Industrial Heritage through Term and Relation Extraction*
Mihael Arcan, Rory O'Halloran, Cécile Robin and Paul Buitelaar

21:30–23:00 *Non-Parametric Word Sense Disambiguation for Historical Languages*
Enrique Manjavacas Arevalo and Lauren Fonteyn

21:30–23:00 *Introducing a Large Corpus of Tokenized Classical Chinese Poems of Tang and Song Dynasties*
Chao-Lin Liu, Ti-Yong Zheng, Kuan-Chun Chen and Meng-Han Chung

21:30–23:00 *Creative Text-to-Image Generation: Suggestions for a Benchmark*
Irene Russo

21:30–23:00 *The predictability of literary translation*
Andrew Piper and Matt Erlin

21:30–23:00 *Emotion Conditioned Creative Dialog Generation*
Khalid Alnajjar and Mika Hämäläinen

21:30–23:00 *Integration of Named Entity Recognition and Sentence Segmentation on Ancient Chinese based on Siku-BERT*
Sijia Ge

21:30–23:00 *(Re-)Digitizing Ngôo Siú-lé's Mandarin – Taiwanese Dictionary*
Pierre Magistry and Afala Phaxay