NLP4DH 2021

Workshop on Natural Language Processing for Digital Humanities

Proceedings of the Workshop

December 19, 2021

ISBN 978-952-94-5833-2
Rootroo Oy

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 first time in December 19, 2021 with ICON 2021: The 18th International Conference on Natural Language Processing. Our workshop received 42 submissions, out of which 21 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

Program Committee

- Iana Atanassova, Université de Bourgogne Franche-Comté
- Yuri Bizzoni, Aarhus University
- Miriam Butt, University of Konstanz
- Jeremy Bradley, University of Vienna
- Won Ik Cho, Seoul National University
- Stefania Degaetano-Ortlieb, Saarland University
- Quan Duong, University of Helsinki
- Valts Ernštreits, University of Latvia, Livonian Institute
- Luke Gessler, Georgetown University
- Hugo Gonçalo Oliveira, University of Coimbra
- Kenichi Iwatsuki, ARIKTTA
- Maciej Janicki, University of Helsinki
- Heiki-Jaan Kaalep, University of Tartu
- Maximilian Koppatz, Sanoma Media Finland
- Mikko Kurimo, Aalto University
- Leo Leppänen, University of Helsinki
- Enrique Manjavacas Arevalo, University of Leiden
- Matej Martinc, Jozef Stefan Institute
- Flammie Pirinen, UiT The Arctic University of Norway
- Lidia Pivovarova, University of Helsinki
- Tyler Shoemaker, University of California, Davis
- Liisa Lotta Tarvainen-Li, Acolad
- Jörg Tiedemann, University of Helsinki
- Jouni Tuominen, Aalto University

- Linda Wiechetek, UiT The Arctic University of Norway
- Joshua Wilbur, University of Tartu
- Shuo Zhang, Bose Corporation
- Emily Öhman, Waseda University

Table of Contents

Conference Program

Sunday, December 19, 2021

9:45–10:00 ***Opening***

10:00–11:00 **Session 1: Sentiment**

10:00–10:20 *Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences*
Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen and Kristoffer Nielbo

10:20–10:40 *The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research*
Emily Öhman

10:40–11:00 *How Does the Hate Speech Corpus Concern Sociolinguistic Discussions? A Case Study on Korean Online News Comments*
Won Ik Cho and Jihyung Moon

11:00–11:15 **Coffee break**

11:15–12:15 **Session 2: Historical data**

11:15–11:35 *MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)*
Enrique Manjavacas Arevalo and Lauren Fonteyn

11:35–11:55 *Named Entity Recognition for French medieval charters*
Sergio Torres Aguilar and Dominique Stutzmann

11:55–12:15 *Processing M.A. Castrén's Materials: Multilingual Historical Typed and Handwritten Manuscripts*
Niko Partanen, Jack Rueter, Khalid Alnajjar and Mika Hämäläinen

Sunday, December 19, 2021 (continued)

12:15–13:15 Lunch

13:15–14:15 Session 3: Literature

13:15–13:35 *Lotte and Annette: A Framework for Finding and Exploring Key Passages in Literary Works*
Frederik Arnold and Robert Jäschke

13:35–13:55 *Using Referring Expression Generation to Model Literary Style*
Nick Montfort, Ardalan SadeghiKivi, Joanne Yuan and Alan Y. Zhu

13:55–14:15 *The concept of nation in nineteenth-century Greek fiction through computational literary analysis*
Fotini Koidaki, Despina Christou, Katerina Tiktopoulou and Grigorios Tsoumakas

14:15–14:30 Coffee break

14:30–16:00 Session 4: Posters

14:30–16:00 *Logical Layout Analysis Applied to Historical Newspapers*
Nicolas Gutehrlé and Iana Atanassova

14:30–16:00 *"Don't worry, it's just noise'": quantifying the impact of files treated as single textual units when they are really collections*
Thibault Clérice

14:30–16:00 *NLP in the DH pipeline: Transfer-learning to a Chronolect*
Aynat Rubinstein and Avi Shmidman

14:30–16:00 *Using Computational Grounded Theory to Understand Tutors' Experiences in the Gig Economy*
Lama Alqazlan, Rob Procter and Michael Castelle

14:30–16:00 *Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers*
Gechuan Zhang, David Lillis and Paul Nulty

Sunday, December 19, 2021 (continued)

14:30–16:00 *Japanese Beauty Marketing on Social Media: Critical Discourse Analysis Meets NLP*
Emily Öhman and Amy Gracy Metcalfe

14:30–16:00 *Text Zoning of Theater Reviews: How Different are Journalistic from Blogger Reviews?*
Mylene Maignant, Thierry Poibeau and Gaëtan Brison

14:30–16:00 *Word Sense Induction with Attentive Context Clustering*
Moshe Stekel, Amos Azaria and Shai Gordin

14:30–16:00 *Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?*
Baptiste Blouin, Benoit Favre, Jeremy Auguste and Christian Henriot

14:30–16:00 *TFW2V: An Enhanced Document Similarity Method for the Morphologically Rich Finnish Language*
Quan Duong, Mika Hämäläinen and Khalid Alnajjar

14:30–16:00 *Did You Enjoy the Last Supper? An Experimental Study on Cross-Domain NER Models for the Art Domain*
Alejandro Sierra-Múnera and Ralf Krestel

14:30–16:00 *An Exploratory Study on Temporally Evolving Discussion around Covid-19 using Diachronic Word Embeddings*
Avinash Tulasi, Asanobu Kitamoto, Ponnurangam Kumaraguru and Arun Balaji Buduru