EMNLP 2021

The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature

Co-located with the 2021 Conference on Empirical Methods in Natural Language Processing EMNLP 2021

Proceedings

November 11, 2021
Punta Cana, Dominican Republic (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 978-1-954085-91-6

Preface

Another virtual workshop… Oh well: maybe next year we will go back to the (new) normal.

Looking at submission numbers, this year's edition of the workshop was a huge success. We have received *really* unusually many submissions (thanks, everyone!). Out of those, we have accepted 22 papers for a 38 % acceptance rate. A round of applause for our wonderful program committee!

The workshop programme consists of brief eight-minute Q&A sessions for ten oral presentations (which you will have watched by then!), and a twelve-poster session during which you will be able to chat with any author you like. Thematically, the papers cover the entire range of "Cultural Heritage, Social Sciences, Humanities and Literature". The programme shows that this area of applied language technology is mature and active.

Last but not least, Sara Tonelli will give a live invited talk. We are grateful, and we look forward to it.

Keep well.

Stefania, Anna, Nils, Stan

`https://sighum.wordpress.com/events/latech-clfl-2021/`

Invited Talk

Dissecting offensive language detection: does it work, and what can we do with it?

Social media messages are often written to attack specific groups of users based on their religion, ethnicity or social status, and they can be particularly threatening to vulnerable users such as teenagers. It is therefore very important to develop reliable, unbiased and robust detection systems to support stakeholders in fighting online hatred. Although state-of-the-art systems yield very good classification results, the problem is far from being solved. In my talk, I will discuss which issues still affect the development of abusive language detection systems, for example the problem of dealing with annotators' disagreement in the creation of training data, and the issues related to contextual information in threads. On the other hand, I will show how the output of offensive language detection systems can be integrated with network-based information to study the behavioral patterns of different types of users, also in relation to misinformation.

About the speaker

Sara Tonelli
Digital Humanties
Fondazione Bruno Kessler
Trento, Italy

Sara Tonelli holds a PhD in Language Sciences from Università Ca' Foscari, Venice. Since 2013 she has been the head of the Digital Humanities research group at Fondazione Bruno Kessler in Trento, Italy. Among many projects in digital humanities, she is currently involved in the H2020 ODEUROPA project (focusing on olfactory information extraction), and in the H2020 PERCEPTIONS project (online perception and migration narratives related to EU). Since January 2021 she is also the scientific coordinator of the KID ACTIONS European project (addressing cyberbullying among children and adolescents). Sara's main research interests are related to temporal and event-based processing of texts, especially in the historical domain, and social media processing, including the detection of abusive language.

Organizers:

Stefania Degaetano-Ortlieb, Department of Language Science and Technology, Universität des Saarlandes
Anna Kazantseva, National Research Council of Canada
Nils Reiter, Institute for Natural Language Processing (IMS), Stuttgart University / Institute for Digital Humanities (IDH), Cologne University
Stan Szpakowicz, School of Electrical Engineering and Computer Science, University of Ottawa

Program Committee:

Syrielle Montariol, INRIA Paris, France
Vivi Nastase, University of Stuttgart, Germany
Borja Navarro Colorado, University of Alicante, Spain
John Nerbonne, University of Freiburg, Germany
Pierre Nugues, Lund University, Sweden
Petya Osenova, Sofia University and IICT-BAS, Bulgaria
Janis Pagel, University of Stuttgart, Germany
Andrew Piper, McGill University, Canada
Petr Plecháč, Institute of Czech Literature of the CAS, Czechia
Thierry Poibeau, CNRS Paris and Lattice, France
Jelena Prokic, Leiden University Centre for Digital Humanities, The Netherlands
Georg Rehm, DFKI, Germany
Martin Reynaert, Tilburg University, Radboud University Nijmegen, The Netherlands
Pablo Ruiz Fabo, Université de Strasbourg, France
Marijn Schraagen, Utrecht University, The Netherlands
Matthew Sims, University of California, Berkeley, United States
Pia Sommerauer, Vrije Universiteit Amsterdam, The Netherlands
Elke Teich, Saarland University, Germany
Laure Thompson, University of Massachusetts Amherst, United States
Ulrich Tiedau, University College London, United Kingdom
Ted Underwood, University of Illinois, Urbana-Champaign, United States
Rob Voigt, Northwestern University, United States
Menno van Zaanen, South African Centre for Digital Language Resources, Potchefstroom, South Africa
Albin Zehe, University of Würzburg, Germany
Heike Zinsmeister, University of Hamburg, Germany

Invited Speaker:

Sara Tonelli, Fondazione Bruno Kessler, Italy

Table of Contents

Workshop Program

Talks, part I

The Early Modern Dutch Mediascape. Detecting Media Mentions in Chronicles Using Word Embeddings and CRF
Alie Lassche and Roser Morante

Batavia asked for advice. Pretrained language models for Named Entity Recognition in historical texts.
Sophie I. Arnoult, Lodewijk Petram and Piek Vossen

Quantifying Contextual Aspects of Inter-annotator Agreement in Intertextuality Research
Enrique Manjavacas Arevalo, Laurence Mellerin and Mike Kestemont

WMDecompose: A Framework for Leveraging the Interpretable Properties of Word Mover's Distance in Sociocultural Analysis
Mikael Brunila and Jack LaViolette

A Mixed-Methods Analysis of Western and Hong Kong–based Reporting on the 2019–2020 Protests
Arya D. McCarthy, James Scharf and Giovanna Maria Dora Dore

Talks, part II

FrameNet-like Annotation of Olfactory Information in Texts
Sara Tonelli and Stefano Menini

End-to-end style-conditioned poetry generation: What does it take to learn from examples alone?
Jörg Wöckener, Thomas Haider, Tristan Miller, The-Khang Nguyen, Thanh Tung Linh Nguyen, Minh Vu Pham, Jonas Belouadi and Steffen Eger

Automating the Detection of Poetic Features: The Limerick as Model Organism
Almas Abdibayev, Yohei Igarashi, Allen Riddell and Daniel Rockmore

Translationese in Russian Literary Texts
Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski and Ruslan Mitkov

Stylometric Literariness Classification: the Case of Stephen King
Andreas van Cranenburgh and Erik Ketzan

Posters

The Multilingual Corpus of Survey Questionnaires Query Interface
Danielly Sorato and Diana Zavala-Rojas

The FairyNet Corpus - Character Networks for German Fairy Tales
David Schmidt, Albin Zehe, Janne Lorenzen, Lisa Sergel, Sebastian Düker, Markus Krug and Frank Puppe

Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language
Thomas Schmidt, Katrin Dennerlein and Christian Wolff

Unsupervised Adverbial Identification in Modern Chinese Literature
Wenxiu Xie, John Lee, Fangqiong Zhan, Xiao Han and Chi-Yin Chow

Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features
Felix Schneider, Björn Barz, Phillip Brandes, Sophie Marshall and Joachim Denzler

BAHP: Benchmark of Assessing Word Embeddings in Historical Portuguese
Zuoyu Tian, Dylan Jarrett, Juan Escalona Torres and Patricia Amaral

The diffusion of scientific terms – tracing individuals' influence in the history of science for English
Yuri Bizzoni, Stefania Degaetano-Ortlieb, Katrin Menzel and Elke Teich

A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek
Pranaydeep Singh, Gorik Rutten and Els Lefever

Zero-Shot Information Extraction to Enhance a Knowledge Graph Describing Silk Textiles
Thomas Schleider and Raphael Troncy

'Tecnologica cosa': Modeling Storyteller Personalities in Boccaccio's 'Decameron'
A. Cooper, Maria Antoniak, Christopher De Sa, Marilyn Migiel and David Mimno

Period Classification in Chinese Historical Texts
Zuoyu Tian and Sandra Kübler