NAACL HLT 2019

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

Proceedings of the Third Workshop

June 7, 2019
Minneapolis, MN, USA

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 978-1-950737-00-0

Preface

Welcome to the third edition of LaTeCH-CLfL—which also is the thirteenth edition of LaTeCH and eighth of CLfL. We have had fun preparing the workshop, and we will be happy if you have fun attending (or at least reading the workshop papers -:). Please visit the website at https://sighum.wordpress.com/events/latech-clfl-2019/ where you will find the workshop presentations, among other things.

The papers cover, as usual, topics which you will not easily find at regular NLP conferences. The authors take on literary texts, including drama and poetry, and more generally literary study; historical texts; ancient or otherwise old languages; government documents; code switching; and more.

Last but certainly not least, we will have an invited talk. Ian Milligan, a historian, has a deep interest in Digital Humanities, and understands the role on Natural Language Processing in his discipline.

It is our pleasant duty to thank the authors: there would be no workshop without you. Nor without the program committee, to whom we are ever so grateful for their thorough and helpful reviews.

Beatrice, Stefania, Nils, Stan, Anna

Invited Talk

Working with Cultural Heritage at Scale: Developing Tools and Platforms to Enable Historians to Explore History in the Age of Abundance

The rise of the Web as a primary source will have deep implications for historians. It will affect our research — how we write and think about the past — and it will change how humanists and social scientists make sense of culture at scale. Scholars are entering an era when there will be more information than ever, left behind by people who rarely entered the historical record before. Web archives, repositories of archived websites dating back to 1996, will fundamentally transform scholarship, requiring a move towards computational methodologies and the digital humanities.

The talk explores this dramatic shift — and what is to be done about it — by arguing that historians will have to understand how to work with textual (and other) data at scale. Historians will soon need to become familiar, at the very least, with NLP techniques. This is not just a marginal problem: the need to explore the big data of the Web (and other digitized repositories) strikes to the core of our discipline.

All Historians Have to Begin to Work with Data

Initial moves towards digital methods have been very promising, as historians begin to study the 1990s. Even so, they will discover sooner than they think that one cannot write most histories of the 1990s or later without reference to web archives. They must be ready, but they are hamstrung. The profession has largely turned away from statistics and from quantitative methodologies more generally; and the web archiving analysis ecosystem is largely based on tools that require a high level of technical expertise. Access to web archives at scale requires, more often than not, fluency with command-line interfaces, access to high-performance computing, and storage at the terabyte scale. Historians need to analyze web archives to write histories, yet that requires skills and infrastructure beyond what one can reasonably expect of them. What, then, can be done?

Tools and Platforms: The Archives Unleashed Project

The talk introduces this problem, and discusses the process of developing tools and platforms to enable historians to explore this "age of abundance". It does so by highlighting the Archives Unleashed Project, an interdisciplinary initiative funded by the Andrew W. Mellon Foundation. The project's goal is to "make petabytes of historical Internet content accessible to scholars and others interested in researching the recent past", and brings together a historian, a computer scientist, and a librarian to lead a team to develop such infrastructure. The project will achieve it in three main ways.

- The Archives Unleashed Toolkit is an open-source platform for analyzing web archives with Apache Spark. It is a scalable toolkit, based upon a process cycle that we have developed; we call it the Filter-Analyze-Aggregate-Visualize cycle. To use the Toolkit, a scholar first filters down a large web (a particular range of dates, a domain, or only pages with certain keywords present); analyzes (finds links, or named entities, sentiment, topics); aggregates (summarizes the output); and visualizes (either through various data tools or tabular data). The Toolkit, based on a command-line interface, is unfortunately very difficult to use.
- The Archives Unleashed Cloud is a web-based front-end for working with the Toolkit. It takes data from the Internet Archive and processes it into formats familiar to researchers: network diagrams,

filtered text files, and other statistical information about a collection. We also provide all of this data for download with a bundled Jupyter Notebook. This allows scholars to use a web-based interface to perform basic data science operations on the data: draw on popular computational linguistics or data science Python libraries to process data and find answers. Suddenly, working with web archives is not so terrifying, and the users have been connected to the mainstream of the Natural Language Processing world.

- We run a series of datathons (three to date, as part of the Mellon grant). They bring together domain experts, researchers, and others to work with web archive data at scale and so help lower barriers; connect people interested in the topic and build community; and help develop a body of practice around web archiving collection and analysis practices.

Conclusion

The talk explores ways in which we can help historians move into an age when working with cultural heritage at scale is no longer a "nice to have" but a necessary component of studying periods from the 1990s onwards.

About the speaker

Ian Milligan is an Associate Professor of History at the University of Waterloo, where he teaches Canadian and digital history. He is currently the principal investigator of the Archives Unleashed project, which seeks to make web archives accessible to humanities and social sciences researchers. Ian has published several books: the forthcoming *History in the Age of Abundance? How the Web is Transforming Historical Research* (April 2019), the *SAGE Handbook of Web History* (co-edited with Niels Brügger, 2018), *Exploring Big Historical Data: The Historian's Macroscope* (co-authored with Scott Weingart and Shawn Graham, 2015), and *Rebel Youth: 1960s Labour Unrest, Young workers, and New Leftists in English Canada* (2014). In 2016, Ian was named the Canadian Society for Digital Humanities's recipient of the Outstanding Early Career Award.

Program Committee

JinYeong Bak, KAIST, Republic of Korea
Gosse Bouma, University of Groningen, Netherlands
Paul Buitelaar, Insight Centre for Data Analytics, National University of Ireland Galway, Ireland
Gerard de Melo, Rutgers University, United States
Thierry Declerck, DFKI GmbH, Germany
Stefanie Dipper, Ruhr-University Bochum, Germany
Jacob Eisenstein, Georgia Institute of Technology, United States
Micha Elsner, The Ohio State University, United States
Mark Finlayson, FIU, United States
Serge Heiden, ENS de Lyon, France
Graeme Hirst, University of Toronto, Canada
Mika Hämäläinen, University of Helsinki, Finland
Adam Jatowt, Kyoto University, Japan
Mike Kestemont, University of Antwerp, Belgium
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Stasinos Konstantopoulos, NCSR Demokritos, Greece
John Lee, City University of Hong Kong, Hong Kong
Chaya Liebeskind, Jerusalem College of Technology, Lev Academic Center, Israel
Rada Mihalcea, University of Michigan, United States
Borja Navarro-Colorado, University of Alicante, Spain
John Nerbonne, Albert-Ludwigs Universität Freiburg, Germany
Pierre Nugues, Lund University, Sweden
Petya Osenova, Sofia University and IICT-BAS, Bulgaria
Michael Piotrowski, Université de Lausanne, Switzerland
Andrew Piper, McGill University, Canada
Thierry Poibeau, LATTICE-CNRS, France
Georg Rehm, DFKI, Germany
Martin, Reynaert, Tilburg University, Netherlands
Pablo Ruiz, LINHD, UNED, Spain
Marijn Schraagen, Utrecht University, Netherlands
Eszter Simon, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
Elke Teich, Universität des Saarlandes, Germany
Sara Tonelli, FBK, Italy
Thorsten Trippel, University of Tübingen, Germany
Ted Underwood, Univ of Illinois, United States
Menno van Zaanen, Tilburg University, Netherlands
Kalliopi Zervanou, Eindhoven University of Technology, Netherlands
Heike Zinsmeister, Universität Hamburg, Germany

Invited Speaker

Ian Milligan, Department of History, Faculty of Arts, University of Waterloo, United States

Organizers

Beatrice Alex, School of Informatics, University of Edinburgh
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
Stan Szpakowicz, School of Electrical Engineering and Computer Science, University of Ottawa

Table of Contents

Modeling Word Emotion in Historical Language: Quantity Beats Supposed Stability in Seed Word Selection
Johannes Hellrich, Sven Buechel and Udo Hahn 1

Clustering-Based Article Identification in Historical Newspapers
Martin Riedl, Daniela Betz and Sebastian Padó 12

The Scientization of Literary Study
Stefania Degaetano-Ortlieb and Andrew Piper 18

Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama
Krishnapriya Vishnubhotla, Adam Hammond and Graeme Hirst 29

Automatic Alignment and Annotation Projection for Literary Texts
Uli Steinbach and Ines Rehbein 35

Inferring missing metadata from environmental policy texts
Steven Bethard, Egoitz Laparra, Sophia Wang, Yiyun Zhao, Ragheb Al-Ghezi, Aaron Lien and Laura López-Hoffman 46

Stylometric Classification of Ancient Greek Literary Texts by Genre
Efthimios Gianitsos, Thomas Bolt, Pramit Chaudhuri and Joseph Dexter 52

A framework for streamlined statistical prediction using topic models
Vanessa Glenny, Jonathan Tuke, Nigel Bean and Lewis Mitchell 61

Revisiting NMT for Normalization of Early English Letters
Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann and Eetu Mäkelä 71

Graph convolutional networks for exploring authorship hypotheses
Tom Lippincott 76

Semantics and Homothetic Clustering of Hafez Poetry
Arya Rahgozar and Diana Inkpen 82

Computational Linguistics Applications for Multimedia Services
Kyeongmin Rim, Kelley Lynch and James Pustejovsky 91

Correcting Whitespace Errors in Digitized Historical Texts
Sandeep Soni, Lauren Klein and Jacob Eisenstein 98

On the Feasibility of Automated Detection of Allusive Text Reuse
Enrique Manjavacas, Brian Long and Mike Kestemont 104

The limits of Spanglish?
Barbara Bullock, Wally Guzman and Almeida Jacqueline Toribio 115

Sign Clustering and Topic Extraction in Proto-Elamite
Logan Born, Kate Kelley, Nishant Kambhatla, Carolyn Chen and Anoop Sarkar 122

Conference Program

Friday, June 7, 2019

08:55–10:30 ***Session 1***

08:55–09:00 ***Welcome***

09:00–09:30 *Modeling Word Emotion in Historical Language: Quantity Beats Supposed Stability in Seed Word Selection*
Johannes Hellrich, Sven Buechel and Udo Hahn

09:30–10:00 *Clustering-Based Article Identification in Historical Newspapers*
Martin Riedl, Daniela Betz and Sebastian Padó

10:00–10:30 ***Poster Teasers***

11:00–12:30 ***Session 2***

11:00–11:30 *The Scientization of Literary Study*
Stefania Degaetano-Ortlieb and Andrew Piper

11:30–12:00 *Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama*
Krishnapriya Vishnubhotla, Adam Hammond and Graeme Hirst

12:00–12:30 *Automatic Alignment and Annotation Projection for Literary Texts*
Uli Steinbach and Ines Rehbein

14:00–15:00 ***Invited Talk***

14:00–15:00 *Working with Cultural Heritage at Scale: Developing Tools and Platforms to Enable Historians to Explore History in the Age of Abundance*
Ian Milligan

15:00–15:30 ***Poster Session***

Friday, June 7, 2019 (continued)

Inferring missing metadata from environmental policy texts
Steven Bethard, Egoitz Laparra, Sophia Wang, Yiyun Zhao, Ragheb Al-Ghezi, Aaron Lien and Laura López-Hoffman

Stylometric Classification of Ancient Greek Literary Texts by Genre
Efthimios Gianitsos, Thomas Bolt, Pramit Chaudhuri and Joseph Dexter

A framework for streamlined statistical prediction using topic models
Vanessa Glenny, Jonathan Tuke, Nigel Bean and Lewis Mitchell

Revisiting NMT for Normalization of Early English Letters
Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann and Eetu Mäkelä

Graph convolutional networks for exploring authorship hypotheses
Tom Lippincott

Semantics and Homothetic Clustering of Hafez Poetry
Arya Rahgozar and Diana Inkpen

Computational Linguistics Applications for Multimedia Services
Kyeongmin Rim, Kelley Lynch and James Pustejovsky

Correcting Whitespace Errors in Digitized Historical Texts
Sandeep Soni, Lauren Klein and Jacob Eisenstein

16:00–17:35 ***Session 3***

16:00–16:30 *On the Feasibility of Automated Detection of Allusive Text Reuse*
Enrique Manjavacas, Brian Long and Mike Kestemont

16:30–17:00 *The limits of Spanglish?*
Barbara Bullock, Wally Guzman and Almeida Jacqueline Toribio

17:00–17:30 *Sign Clustering and Topic Extraction in Proto-Elamite*
Logan Born, Kate Kelley, Nishant Kambhatla, Carolyn Chen and Anoop Sarkar

Friday, June 7, 2019 (continued)

17:30–17:35 ***Closing***