ACL 2017

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

Proceedings of the Workshop

August 4, 2017
Vancouver, Canada

ISBN 978-1-945626-58-6

Introduction

LaTeCH-CLfL 2017 continues the tradition of two separate yet not dissimilar events. It is both the 11th Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities, and the 6th Workshop on Computational Linguistics for Literature—held jointly for the first time, with beneficial effects. We have been able to cast the net more widely. We received more, and more varied, submissions. Nine long papers, five short papers and a position paper will appear at the workshop, a 58% acceptance rate. We also had the advantage of two program committees from past years helping us select the best papers. We are ever so grateful to all those attentive, thorough and helpful reviewers. Sure enough, we thank all authors for the hard work they invested in their submissions.

Our distinguished invited speaker, Andrew Piper, is a perfect match for our joint workshop: he applies tools and techniques of data science to literature as well as to culture. He will introduce new work on the process of characterization: how writers construct animate entities on the page. This contributes to a better understanding of the specific nature of literary characters as linguistic entities.

The papers accepted this year cover an intriguing variety of topics. First off, we have a few papers which deal with poetry, each tackling a very different problem. Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz and Chris Tanasescu identify a specific type of literary metaphor in poems. They rely on a combination of statistical analysis and rules. Pablo Ruiz Fabo, Clara Martínez Cantón, Thierry Poibeau and Elena González-Blanco seek to discover enjambment: places in a poem where a syntactic unit is split across two lines. They apply their method to a diachronic corpus of Spanish sonnets, and analyze the results across four centuries. Christopher Hench works on medieval German poetry. He uses syllabification to analyze soundscapes and thus to shed light on how those primarily oral poems may have sounded.

A lot of interesting work revolves around the computational analysis of prose. We have papers which present tools for scholars in Digital Humanities, and more specific studies of certain phenomena or of particular novels. Andre Blessing, Nora Echelmeyer, Markus John and Nils Reiter present an end-to-end environment intended to help analyze relations between entities in a document in a principled way. Evgeny Kim, Sebastian Padó and Roman Klinger adopt lexicon-based methods to the study of the emotional trajectory of novels, and compare their findings across five genres. Stefania Degaetano-Ortlieb and Elke Teich outline a generic data-driven method of tracking intra-textual variation, showing how information-theoretic measures allow the detection of both topical and stylistic patterns of variation.

Liviu Dinu and Ana Sabina Uban verify if characters of a given novel are believable, using methods established in the authorship attribution community. They present the preliminary results for the novel Les Liaisons Dangereuses. Conor Kelleher and Mark Keane describe an experiment in distant reading applied to a post-modern novel with non-linear structure, Wittgenstein's Mistress by David Markson. The paper contrasts the analysis which arises from the distant read with David Foster Wallace's "manual" analysis.

A good portion of our workshop is devoted to historical, low-resource or non-standard languages. Amrith Krishna, Pavankumar Satuluri and Pawan Goyal write about challenges of working with Sanskrit manuscripts. They release a dataset for the segmentation of Sanskrit words. Nina Seemann, Marie-Luis Merten, Michaela Geierhos, Doris Tophinke and Eyke Hüllermeier share the experience of annotating texts in Middle Low German. It turns out that the process is fraught with uncertainties; the Authors discuss them and describe lessons learned.

Next, we have a paper by Maria Sukhareva, Francesco Fuscagni, Johannes Daxenberger, Susanne Görke, Doris Prechel and Iryna Gurevych. In their experiments, they apply distant supervision to the building of a part-of-speech tagger for Hittite. Unsurprisingly, no annotated corpora exist for this ancient language.

Émilie Pagé-Perron, Maria Sukhareva (yes!), Ilya Khait and Christian Chiarcos are no less ambitious. They describe experiments in machine translation of Sumerian texts of an administrative or legal nature. The aim is to make those texts available to a wider audience. Géraldine Walther and Benoît Sagot talk about a productive synergy between fully manual and semi-automatic process when building a corpus of Romansh Tuatschin, a dialect of one of the official languages in southwestern Switzerland.

Two more papers complete this palette of topics. Maria Pia di Buono proposes an ontology-based method of extracting nominal compounds in the domain of cultural heritage. Maciej Ogrodniczuk and Mateusz Kopeć explore modern political discourse in the context of Twitter. They present a series of experiments in on-the-fly analysis of the lexical, topical and visual aspects of political tweets.

There you have it. Welcome to our workshop, and by all means have fun.

Beatrice, Stefania, Anna, Anna, Nils and Stan

Invited Talk

Title: Characterization

Speaker: Andrew Piper

Abstract

Characters are some of the most important, and most beloved, elements of literature. From Ishmael to Mrs. Dalloway to Gregor Samsa, literary characters are woven into the fabric of culture. And yet until recently, little work has been done to understand the specific nature of characters as linguistic entities. This talk will introduce new work by our lab that aims to address this process of characterization – of how writers construct animate entities on the page. It will present a new character feature tool designed to allow researchers to study a variety of qualities surrounding the construction of character as well as a new study where it has been implemented.

About the speaker

Andrew Piper is Professor and William Dawson Scholar in the Department of Languages, Literatures, and Cultures at McGill University. His work explores the application of computational approaches to the study of literature and culture. He is the director of .txtLAB,[1] a digital humanities laboratory at McGill, as well as leader of the international partnership grant, "NovelTM: Text Mining the Novel",[2] which brings together 21 partners across North America to undertake the first large-scale quantitative and cross-cultural study of the novel. He is the author most recently of *Book Was There: Reading in Electronic Times* (Chicago 2012) and is currently completing a new book entitled *Enumerations: The Quantities of Literature*.

[1] http://txtlab.org/
[2] http://novel-tm.ca/

Program Committee:

Cecilia Ovesdotter Alm, Rochester Institute of Technology, USA
JinYeong Bak, KAIST, Republic of Korea
Gosse Bouma, University of Groningen, Netherlands
Julian Brooke, University of Melbourne, Australia
Paul Buitelaar, National University of Ireland, Galway, Ireland
Thierry Declerck, Deutsche Forschungszentrum für Künstliche Intelligenz GmbH, Germany
Stefanie Dipper, Ruhr-University, Bochum, Germany
Jacob Eisenstein, Georgia Institute of Technology, United States
Micha Elsner, Ohio State University, United States
Stefan Evert, Erlangen-Nürnberg University, Germany
Mark Finlayson, Florida International University, United States
Antske Fokkens, Vrije Universiteit Amsterdam, Netherlands
Pablo Gervás Gómez-Navarro, Universidad Complutense de Madrid, Spain
Serge Heiden, École normale supérieure de Lyon, France
Iris Hendrickx, Radboud University, Nijmegen, Netherlands
Aurélie Herbelot, University of Trento, Italy
Gerhard Heyer, University of Leipzig, Germany
Graeme Hirst, University of Toronto, Canada
Eero Hyvönen, University of Helsinki, Finland
Amy Isard, University of Edinburgh, United Kingdom
Adam Jatowt, Kyoto University, Japan
Vangelis Karkaletsis, National Centre of Scientific Research "Demokritos", Greece
Mike Kestemont, University of Antwerp, Belgium
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Stasinos Konstantopoulos, National Centre of Scientific Research "Demokritos", Greece
Jonas Kuhn, Stuttgart University, Germany
John Lee, City University of Hong Kong, Hong Kong
Chaya Liebeskind, Jerusalem College of Technology, Israel
Clare Llewellyn, University of Edinburgh, United Kingdom
Barbara McGillivray, Alan Turing Institute/University of Cambridge, United Kingdom
Gerard de Melo, Tsinghua University, China
Rada Mihalcea, University of Michigan, United States
Borja Navarro Colorado, University of Alicante, Spain
John Nerbonne, Groningen University, Netherlands
Dong-Phuong Nguyen, University of Twente, Netherlands
Pierre Nugues, Lund University, Sweden
Mick O'Donnel, Universidad Autónoma de Madrid, Spain
Petya Osenova, Sofia University and IICT-BAS, Bulgaria
Michael Piotrowski, Leibniz Institute of European History, Germany
Livia Polanyi, LDM Associates, United States
Georg Rehm, DFKI, Germany
Martin Reynaert, Tilburg University, Radboud University Nijmegen, Netherlands
Marijn Schraagen, Utrecht University, Netherlands
Sarah Schulz, University of Stuttgart, Germany
Eszter Simon, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary

Caroline Sporleder, Goettingen University, Germany
Jannik Strötgen, Max-Planck-Institut für Informatik, Germany
Reid Swanson, University of California, Santa Cruz, United States
Elke Teich, Saarland University, Germany
Mariët Theune, University of Twente, Netherlands
Sara Tonelli, FBK, Trento, Italy
Thorsten Trippel, University of Tübingen, Germany
Menno van Zaanen, Tilburg University, Netherlands
Heike Zinsmeister, University of Hamburg, Germany

Invited Speaker:

Andrew Piper, McGill University, Canada

Organizers:

Beatrice Alex, School of Informatics, University of Edinburgh
Stefania Degaetano-Ortlieb, Department of Language Science and Technology, Universität des Saarlandes
Anna Feldman, Department of Linguistics & Department of Computer Science, Montclair State University
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

Conference Program

August 4th, 2017

09:00–10:00 Session 1

09:00–09:30 *Metaphor Detection in a Poetry Corpus*
Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz and Chris Tanasescu

09:30–10:00 *Machine Translation and Automated Analysis of the Sumerian Language*
Émilie Pagé-Perron, Maria Sukhareva, Ilya Khait and Christian Chiarcos

10:00–10:30 Poster Teaser Talks

11:00–12:30 Session 2

11:00–11:30 *Investigating the Relationship between Literary Genres and Emotional Plot Development*
Evgeny Kim, Sebastian Padó and Roman Klinger

11:30–12:00 *Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets*
Pablo Ruiz, Clara Martínez Cantón, Thierry Poibeau and Elena González-Blanco

12:00–12:30 *Plotting Markson's "Mistress"*
Conor Kelleher and Mark Keane

August 4th, 2017 (continued)

13:30–14:00 SIGHUM Business Meeting

14:00–15:00 Invited Talk

14:00–15:00 *Characterization*
Andrew Piper

15:00–16:00 Poster Session

15:00–16:00 *Annotation Challenges for Reconstructing the Structural Elaboration of Middle Low German*
Nina Seemann, Marie-Luis Merten, Michaela Geierhos, Doris Tophinke and Eyke Hüllermeier

15:00–16:00 *Phonological Soundscapes in Medieval Poetry*
Christopher Hench

15:00–16:00 *An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis*
Andre Blessing, Nora Echelmeyer, Markus John and Nils Reiter

15:00–16:00 *Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns*
Stefania Degaetano-Ortlieb and Elke Teich

15:00–16:00 *Finding a Character's Voice: Stylome Classification on Literary Characters*
Liviu P. Dinu and Ana Sabina Uban

15:00–16:00 *An Ontology-Based Method for Extracting and Classifying Domain-Specific Compositional Nominal Compounds*
Maria Pia di Buono

15:00–16:00 *Speeding up corpus development for linguistic research: language documentation and acquisition in Romansh Tuatschin*
Géraldine Walther and Benoît Sagot

August 4th, 2017 (continued)

16:00–17:30 Session 4

16:00–16:30 *Distantly Supervised POS Tagging of Low-Resource Languages under Extreme Data Sparsity: The Case of Hittite*
Maria Sukhareva, Francesco Fuscagni, Johannes Daxenberger, Susanne Görke, Doris Prechel and Iryna Gurevych

16:30–17:00 *A Dataset for Sanskrit Word Segmentation*
Amrith Krishna, Pavan Kumar Satuluri and Pawan Goyal

17:00–17:30 *Lexical Correction of Polish Twitter Political Data*
Maciej Ogrodniczuk and Mateusz Kopeć