NAACL-HLT 2012

Workshop on Computational Linguistics for Literature

Co-located with

The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

Proceedings of the Workshop

June 8, 2012
Montréal, Canada

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53707
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

ISBN13: 978-1-937284-20-6
ISBN10: 1-937284-20-4

Proceedings of the NAACL Workshop on Computational Linguistics for Literature:

Preface

> "What do you read, my lord?"
> "Words, words, words."

It may well be that our generation is the last to be intimately familiar with the printed book. We live in an age when the percentage of digitized literature increases steadily. Older work comes online thanks in part to scanning initiatives such as Project Gutenberg (`gutenberg.org`), Google Books (`books.google.com`) or Million Book Project (`archive.org/details/millionbooks`). New material is often born digital, and becomes available via e-book stores and through non-traditional outlets such as blogging and self-publishing.

Literature is in many ways distinct from genres usually considered in computational linguistics, such as newspaper prose, unstructured Web pages or speech. That is why the growing availability of online literature presents new opportunities and challenges in language processing. How can automatic methods help readers find new literature on a certain topic, understand a text or a genre, identify the author of an anonymous text, or read a book written in another language? How might language analysis techniques go beyond words to help identify the deeper meaning found in literature, no matter the time, place or culture from which it originates?

The vibrant research field at the intersection of computing and the humanities, known as Digital Humanities, emphasizes the skills in applying computational techniques to data in arts, humanities and social sciences. We have organized this workshop to help nurture a dialogue between Digital Humanities and the computational linguistics community, where the cutting-edge work in text understanding occurs. Our main target audience are computer scientists and linguists interested in literature as a genre of study, especially those well versed in the rigours of language understanding and those with experience in the idiosyncrasies of literary text. Our invited speaker, Inderjeet Mani, sets the tone with a talk entitled "Computing and the Literary Landscape" which frames the field in terms of the low- and high-hanging fruit that we as a community may pursue.

The papers in this volume cover quite a range of research interests, so much so that to group them by topic is a tough nut to crack. There are both *corpus-based studies* and *in-depth treatment* of specific literary works.

Intriguingly, we have *two* papers on the *computational treatment of poetry*. Julian Brooke, Adam Hammond and Graeme Hirst present work on stylistic segmentation of T. S. Eliot's influential *The Waste Land*. Justine Kao and Dan Jurafsky contribute an essay in computational aesthetics: an exploration of what could be seen as contributing to the aesthetic merits of a good poem.

Two papers, by Anders Søgaard and by Michael Bendersky and David Smith, look into what makes certain phrases likely to be quoted. We can but hope to read a book a day, and we would need many lifetimes to barely scratch the surface of the available literature. To be able to extract a salient (quotable!) passage of a longer work gives us at least a fighting chance to stay afloat.

Next come several papers which consider the literary applications of more widely considered challenges in language understanding. Choonkyu Lee and Smaranda Muresan study the *usage of referring expressions* and coreference in literary narrative. Wajdi Zaghouani and Mona Diab discuss the pilot experiments in annotating the Koran for *semantic role labelling*. Apoorv Agarwal, Augusto Corvalan, Jacob Jensen and Owen Rambow put *network analysis* to work in search of insight into character interactions in an abridged version of *Alice in Wonderland*. What social networks did Lewis Carroll anticipate? *Authorship attribution* is the theme of the papers by Bei Yu and by Andreas van Cranenburgh. The former tests a procedure based on function words on novels, essays and blogs. The latter works with fragments of parse trees, and tests this form of stylometry on some twenty books by five celebrated authors. Because literature is global in scope, *machine translation* of literary work should be quite important. Two papers offer two different points of view: Qian Yu, Aurélien Max and François Yvon experiment with aligning literary works available in multiple languages, while Rob Voigt and Dan Jurafsky look at the role of referential cohesion in machine translation of literature.

The changes to a language over time present challenges in processing older texts. A paper by Ann Irvine, Laure Marcellesi and Afra Zomorodian investigates how we might digitize literary work of a certain vintage when tools trained on modern language are not quite adequate for language of two centuries ago. In a different mode, Sophie Kushkuley takes a look at *Harper's Bazaar*: How did people write about fashion trends in the nineteenth century?

Last but not least, a position paper by Antonio Roque presents several problems in literary analysis and discusses how language technology may help solve such problems.

We anticipate a lively and wide-ranging discussion between the authors of these diverse contributions. We hope that this workshop, and others like it, will galvanize the area of literary analysis within computational linguistics. Literature is a carrier of our culture and its history, so advances in the application of natural language processing to literature will help unlock and explore the insights found within.

We owe a word of thanks to the many individuals who made this workshop possible. First and foremost, we thank the authors. The number of submissions we received shows that the field of computational linguistics for literature is doing very well indeed. We are deeply indebted to the reviewers for their hard work. They enabled us to select an exciting program and to give valuable feedback to the authors. We also thank Mona Diab and Colin Cherry for their incessant help with the logistics. Finally, many thanks to Google Inc. for their financial support.

Enjoy the workshop!

Anna, David, Stan, and Rada

Organizers:

David K. Elson (Google)
Anna Kazantseva (University of Ottawa)
Rada Mihalcea (University of North Texas)
Stan Szpakowicz (University of Ottawa)

Program Committee:

Cecilia Ovesdotter Alm (Rochester Institute of Technology)
Nicholas Dames (Columbia University)
Hal Daumé III (University of Maryland)
Anna Feldman (Montclair State University)
Mark Finlayson (MIT)
Pablo Gervás (Universidad Complutense de Madrid)
Roxana Girju (University of Illinois at Urbana-Champaign)
Amit Goyal (University of Maryland)
Katherine Havasi (MIT Media Lab)
Matthew Jockers (Stanford University)
James Lester (North Carolina State University)
Inderjeet Mani (Children's Organization of Southeast Asia)
Kathy McKeown (Columbia University)
Saif Mohammad (National Research Council, Canada)
Vivi Nastase (HITS gGmbH)
Rebecca Passonneau (Columbia University)
Livia Polanyi (LDM Associates)
Owen Rambow (Columbia University)
Michaela Regneri (Saarland University)
Reid Swanson (University of California, Santa Cruz)
Marilyn Walker (University of California, Santa Cruz)
Janice Wiebe (University of Pittsburgh)

Invited Speaker:

Inderjeet Mani (Children's Organization of Southeast Asia)

Table of Contents

Workshop Schedule

9:00-9:03 Welcome

9:03-10:00 Invited Talk
Computing and the Literary Landscape
Inderjeet Mani

Abstract

I begin by distinguishing literary from non-literary narrative, and then go onto to describe a framework for narrative computing involving environments for authoring and interacting with literary artifacts as well as searching, analyzing, and translating them. These artifacts concern events whose characters act and react based on their beliefs. I focus here on computational issues related to four facets of narrative structure: story embedding, accessibility relations, time, and plot. I conclude with some recommendations for research strategies.

10:00-10:30 Session A
Computational Analysis of Referring Expressions in Narratives of Picture Books
Choonkyu Lee, Smaranda Muresan and Karin Stromswold

10:30-11:00 Break

11:00-12:30 Session B
A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry
Justine Kao and Dan Jurafsky
Towards a Literary Machine Translation: The Role of Referential Cohesion
Rob Voigt and Dan Jurafsky
Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features
Julian Brooke, Adam Hammond and Graeme Hirst

12:30-2:00 Lunch

2:00-3:00 Session C1
Aligning Bilingual Literary Works: a Pilot Study
Qian Yu, Aurélien Max and François Yvon
Function Words for Chinese Authorship Attribution
Bei Yu

3:00-3:30 Session C2
Poster teasers

3:30-4:00 Break

4:00-5:00 Session D

Posters

Mining wisdom
Anders Søgaard
Literary authorship attribution with phrase-structure fragments
Andreas van Cranenburgh
Digitizing 18th-Century French Literature: Comparing transcription methods for a critical edition text
Ann Irvine, Laure Marcellesi and Afra Zomorodian
A Dictionary of Wisdom and Wit: Learning to Extract Quotable Phrases
Michael Bendersky and David Smith
A Pilot PropBank Annotation for Quranic Arabic
Wajdi Zaghouani, Abdelati Hawwari and Mona Diab
Trend Analysis in Harper's Bazaar
Sophie Kushkuley
Social Network Analysis of Alice in Wonderland
Apoorv Agarwal, Augusto Corvalan, Jacob Jensen and Owen Rambow
Towards a computational approach to literary text analysis
Antonio Roque

5:00-5:55 Session E
Open discussion

5:55-6:00 Farewell