Editorial: Computational linguistics and literature

Szpakowicz, Stan; Feldman, Anna; Kazantseva, Anna

Publisher's version / Version de l'éditeur:
https://doi.org/10.3389/fdigh.2018.00024
Frontiers in Digital Humanities, 5, 2018-09-27

frontiers in Digital Humanities

EDITORIAL
published: 27 September 2018
doi: 10.3389/fdigh.2018.00024

Editorial: Computational Linguistics and Literature

Stan Szpakowicz[1]*, *Anna Feldman*[2] *and Anna Kazantseva*[3]

[1] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada, [2] Department of Linguistics and Department of Computer Science, Montclair State University, Montclair, NJ, United States, [3] National Research Council Canada (NRC-CNRC), Ottawa, ON, Canada

Keywords: computational linguisitics, literary data, machine learning, machine translation, post-editing, rhetorical figures, themes and motifs in poetry, Middle High German epic poetry

Editorial on the Research Topic

Computational Linguistics and Literature

Computational Linguistics—or, more technically, Natural Language Processing—has made great strides in the past several years. Machine learning (ML) is the technology of choice; deep learning, in particular, has pushed the envelope. These methods work best in narrow domains, given vast amounts of data and asked for information rather than for interpretation. Literary data are not limited by topic, and they are hardly ever plentiful enough. That is why work on such data has been, as yet, on the periphery of research and development in Computational Linguistics. This Research Topic aims to bring the processing of literary data to the attention of a broader audience.

The early versions of the articles we present here have been published at ACL workshops on Computational Linguistics for Literature held in 2015 and 2016[1]. The papers offer a representative enough sample of the research brought to those workshops.

Toral et al. tackle a weakness of an ostensibly successful major application, Machine Translation (MT). The raw results—especially long machine-translated texts, novels above all—are so error-ridden that substantial post-editing is required. Here is the big question: is MT even worthwhile in such cases? In an instructive experiment with a chapter of a novel, the authors show that the overall cost of translation from scratch significantly outstrips the cost of post-editing. They also show that neural MT beats the more traditional phrase-based statistical MT in terms of productivity gains.

Dubremetz and Nivre look at three rhetorical figures in which words are repeated to a strong stylistic effect: chiasmus, epanaphora, and epiphora[2]. While it is trivial to find repeated lexical material, and indeed such repetition is not infrequent, the actual rhetorical devices are very rare and impossible to recognize automatically without sufficient training data for a ML system. The authors run a multi-faceted experiment with literary texts, scientific texts and quotations. They show that various devices predominate in various types of texts. For example, and perhaps a little surprisingly, the seemingly very artistic chiasmus is more likely to appear in scientific texts[3].

Navarro-Colorado processes a large corpus of Golden Age Spanish sonnets. He uses latent Dirichlet allocation (LDA), a ML algorithm widely accepted as good at determining topics in a text. In addition to a few fairly technical findings, the paper shows that LDA recognizes either a theme in a sonnet, or a poetic motif. The immediate conclusions apply to texts of the same type—moderately short well-structured poems—but the paper can spur similar investigation on similar literary data.

[1] A previous special issue builds upon selected papers from the workshops held in 2012, 2013, and 2014.
[2] See the paper for definitions.
[3] The paper actually studies a type of chiasmus called antimetabole, *e.g.*, "eat to live, not live to eat".

OPEN ACCESS

Edited and reviewed by:
Jean-Gabriel Ganascia, Université Pierre et Marie Curie, France

***Correspondence:**
Stan Szpakowicz
szpak44@gmail.com

Specialty section:
This article was submitted to Digital Literary Studies, a section of the journal Frontiers in Digital Humanities

Received: *23 August 2018*
Accepted: *06 September 2018*
Published: *27 September 2018*

Citation:
Szpakowicz S, Feldman A and Kazantseva A (2018) Editorial: Computational Linguistics and Literature. Front. Digit. Humanit. 5:24. doi: 10.3389/fdigh.2018.00024

Hench and Estes work with highly specialized data: Middle High German epic poetry. We will not steal the authors' thunder; they do a great job of presenting this esoteric matter in a clear and entertaining fashion. In a nutshell, they examine stress patterns in verses, and study how medieval poets used such patterns for semantic and stylistic purposes. Once again, ML is at play. The authors rely on a conditional random field, a method useful in analyzing sequential data.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This Research Topic owes everything to the authors and the reviewers—our thanks to all. We also thank Cecilia Ovesdotter Alm for an initial review of one of the papers.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.