Multilingual NLP as Interface

David Bamman*[1], Quinn Dombrowski*†[2], Natalia Ermolaev*[3], Andrew Janco*[4], Toma Tasovac*[5], Melanie Walsh*[6], and David Lassner*‡[7]

[1]UC Berkeley – United States
[2]Stanford University – United States
[3]Princeton University – United States
[4]Haverford College – United States
[5]DARIAH – Germany
[6]Cornell University – United States
[7]TU Berlin – Germany

Abstract of the answer to the call for proposals

The extreme focus on modern English in much of the natural language processing (NLP) community has led to a chasm between what is computationally possible for English and, for some languages, the feasibility of using computational methods at all. Even within the sphere of modern English, one
encounters a performance gap when applying state-of-the-art algorithms to literature; models trained on news corpora and Wikipedia lose some efficacy when applied to such different kinds of text (Bamman et al 2019). While they typically lack a graphical user interface, NLP models and packages serve as interfaces to text, enabling scholars to do some things, but not others, depending on how they were created, and the nature and quality of their training data. This panel features three talks by scholars working to create new NLP tools and pedagogical materials that address the needs of humanities scholars who work with languages other than English – in effect, building better interfaces for a wider range of computational scholarship.

New Languages for NLP

New Languages for NLP, a collaboration between DARIAH and cooperating partner institution Princeton University, funded by the National Endowment for the Humanities in the United States, is holding three workshops to instruct teams of humanities scholars in developing or improving NLP models for their research languages. Materials for these workshops will be revised based on participant feedback, and published on DARIAH Campus. The workshop series includes ten languages, some of which currently have few or no usable resources for computational text analysis: Classical Arabic, Classical Chinese, Kanbun, Kannada, Ottoman Turkish, Quechua, 19th century Russian, Tigrinya, Yiddish, and Yoruba. The first workshop, to be held in June 2021, will focus on the fundamentals of annotating texts based on the kinds of research questions you hope to address with the resulting model. By the time the DARIAH Annual Event is held in September, the teams will have made

*Speaker
†Contact person: qad@stanford.edu
‡Contact person: davidlassner@googlemail.com

significant progress on annotation and preparations will be well underway for the second workshop on model training, to be held in January 2022. In this presentation, we will report on the results so far, as well as the challenges and opportunities encountered along the way.

Designing Multilingual Teaching Materials for Cultural Analytics

In early 2021, I released an open-source textbook, Introduction to Cultural Analytics & Python, which

introduces Python programming to students and scholars who are interested in studying cultural materials like books, screenplays, and newspapers. Though the textbook originally focused on English language texts and U.S. literature and history, I noticed that it was being used and read by scholars from around the world who were interested in studying other languages and national contexts. This reception prompted me to think about how to redesign the textbook to better support multilingual cultural analytics research, a task that I have begun in collaboration with Quinn Dombrowski. In this paper, I will discuss some of the challenges and insights gleaned from our work so far. Most broadly, I will argue that designing multilingual teaching materials for cultural analytics demands collaboration between different scholars in order to overcome gaps in domain knowledge about specific languages.

Multilingual Book NLP

BookNLP (Bamman et al. 2014) is a natural language processing pipeline for reasoning about the

linguistic structure of text in books, specifically designed for works of fiction. In addition to its pipeline of part-of-speech tagging, named entity recognition, and coreference resolution, BookNLP identifies the characters in a literary text, and represents them through the actions they participate in, the objects they possess, their attributes, and dialogue. The availability of this tool has driven much work in the computational humanities, especially surrounding character (Underwood et al. 2018; Kraicer and Piper 2018; Cheng 2020}. At the same time, however, BookNLP has one major limitation: it currently only supports texts written in English. In this talk, I will describe our efforts to expand BookNLP to support literature in Spanish, Japanese, Russian and German, and create a blueprint for others to develop it for additional languages in the future.

Bibliography

David Bamman, Sejal Popat and Sheng Shen (2019), "An Annotated Dataset of Literary Entities,"
NAACL 2019.

Keywords: multilingual, pedagogy, NLP