Processing non-canonical or noisy text: fortuitous data to the rescue

Barbara Plank
CLCG, University of Groningen
Netherlands
b.plank@rug.nl

Abstract

Real world data differs radically from the benchmark corpora we use in NLP, resulting in large performance drops. The reason for this problem is obvious: NLP models are trained on limited samples from canonical varieties considered standard. However, there are many dimensions, e.g., sociodemographic, language, genre, sentence type, etc. on which texts can differ from the standard. The solution is not obvious: we cannot control for all factors, and it is not clear how to best go beyond the current practice of training on homogeneous data from a single domain and language.

In this talk, I review the notion of canonicity, and how it shapes our community's approach to language. I argue for the use of fortuitous data. Fortuitous data is data out there that just waits to be harvested. It includes data which is in plain sight, but is often neglected, and more distant sources like behavioral data, which first need to be refined. They provide additional contexts and a myriad of opportunities to build more adaptive language technology, some of which I will explore in this talk.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 2nd Workshop on Noisy User-generated Text,
page 1, Osaka, Japan, December 11 2016.