Text Analysis and Machine Learning for Stylometrics and Stylogenetics

Walter Daelemans
University of Antwerp

Abstract

Automatic Text Categorization, learning to assign documents to specific categories (e.g. in topic assignment or spam filtering), has been an influential application in Natural Language Processing. These systems consist of two components: a first one that constructs representations of documents (mostly bags of words represented as binary or numeric vectors), and a second one that uses standard machine learning techniques to learn mappings between such document vectors and their topics. Recently, this general approach has been put to use for other, more linguistically interesting "stylometric" applications, such as assigning authorship to documents or determining the gender of the author of a document. Such applications need linguistically more sophisticated document representations and provide insight into which linguistic properties of documents are relevant for predicting the (gender of) the author. In my presentation, I will give a brief overview of results in this approach and describe a number of applications of the methodology we are currently investigating in the CNTS research group. For creating linguistically more interesting document representations, we use a memory-based shallow parser that analyzes documents at the levels of morphology, part of speech, phrases, and grammatical relations. More specifically I will describe results on authorship attribution in the context of journalists writing about the same topic (politics). A more challenging task is personality assignment on the basis of text. We constructed a corpus consisting of 145 documents describing the contents of the same documentary, written by 145 different students who also took a personality test. We show which linguistic features correlate with different dimensions of personality and the predictability of personality from these features. Finally, I will describe work on what we dubbed "stylogenetics", stylistic analysis of literary works based on the same general architecture, but using clustering as a machine learning technique rather than supervised learning.