Leonard Konle

Universität Würzburg, Germany

email:
leonard.konle@uni-wuerzburg.de

Semantic Zeta: Distinctive word cluster in genre

Keywords: word embeddings / metric learning / genre / authorship attribution

Introduction

Burrow's Zeta (Burrows 2007) is a common method in the field of Digital Humanities to distinguish between two groups of texts by measuring word frequencies in equal sized chunks of data. While its origins are in stylometric and authorship research it is also used to give insights on differences in various other categories like genre or literary epochs. While raw word frequencies seem suitable for authorship attribution, they have shortcomings in identifying words for genres as pointed out in the following example:

Given 100 novels from the genre of western where 80 are written by an author favoring the word "colt" and the rest by a colleague using "revolver" instead. Both words would be good markers to classify authorship, but if the topic of interest is to find distinctive words for western in contrast to love genre there is a good chance of missing the word "revolver" due to its low frequency in the overall western group. If Zeta results are used for a clustering, texts from the "revolver"-author will show as less generic for western.

To resolve this issue a more abstract semantic class of distinctive words is needed. In the example something like "old guns" would be a solution. In the following a method combining Zeta, Word Embeddings and Metric Learning is proposed to meet this requirement.

Method

Texts from both groups are split into test and trainingdata. Than Zeta is applied to determine distinctive words for both groups in the trainingset. Schöch 2018 evaluated different variants of zeta coming up with sd2-zeta as the best choice. These words are transformed into vector representation using a pre-trained embedding. After that unsupervised clustering (Zhang et al. 1996) shrinks the embedded zeta words to more abstract cluster centers. These centers are utilized to bend the embedding space with supervised metric learning. This is performed with Uniform Manifold Approximation (McInnes et al. 2018) and aims at altering the vector space in a way, that cluster centers from both groups are near to other centers of their own class forming a cluster of clusters, while extending the distance between the two classes to a maximum. For the final classification the distance of all token to their nearest cluster center from both groups is calculated. These distances get subtracted from another resulting in a final value for semantic distinctiveness. To determine the quality of distinctiveness scores, classification with linear regression takes these values as input and predicts its group.

Resources

The corpus holds a collection of low brow novels from the genres Love, Medical, Regional, Family, Crime, Horror and Science Fiction (200 each) from various authors. The word embedding is trained on larger corpus containing 20.000 novels with FastText (Bojanowski et al. 2017).

Results

20-fold cross validation over four pairings of genre shows a slight advantage of semantic zeta (.90 f1-score) over sd2-zeta (.85 f1-score). This effect is reversed in authorship attribution leaving sd2-zeta as the prefered choice.