DRDHum24 – Digital Research Data and Human Sciences in the Age of A.I. 2024

Poster presentation

Title: Understanding Document Internal Variation in Eighteenth-Century English Texts

Author: Liina Repo

Abstract

Historical language databases are invaluable resources for linguists and historians, yet their utility is hindered by factors such as complexity, text variety, and lack of register information. Registers, i.e., situationally defined varieties with specific purposes, are important predictors of linguistic variation (Biber 2012). While prior research has predominantly examined registers at the document level, recent studies (Egbert and Gracheva 2023) have revealed that longer documents exhibit features from different registers in different sections due to shifts in e.g. audience or purpose. In this study, we investigate document internal variation in eighteenth-century English texts. Specifically, we 1) explore the impact of different text segments on the classification of registers and 2) inspect the linguistic characteristics associated with the text segments. First, we inspect how different parts of texts (e.g., beginnings vs. middle parts) affect the classification of registers using a BERT-based text classification model. BERT utilizes deep learning to comprehend language context, aiding in the analysis of text segments based on their linguistic characteristics. We fine-tune the model with the Corpus of Founding Era American English (COFEA) that comprises manually register-annotated legal and everyday texts. For testing purposes, we use another subset of COFEA and a small hand-annotated section of the Goldsmiths'-Kress Library of Economic Literature (GKL). To inspect and compare the linguistic characteristics and key features across different text parts and registers, we employ the Stable Attribution Class Explanation (SACX; Rönnqvist et al. 2022) method. SACX provides explanations of text classes in the form of keyword lists that are derived from input attribution (Integrated Gradients; Sundararajan, Taly & Yan 2017) from the BERT model. Our findings suggest varying degrees of influence among different text segments on register classification. Echoing past findings (Laippala et al. 2023), particularly text beginnings seem to produce more reliable classification results. Furthermore, the findings suggest that certain features exhibit connections with specific parts of documents, akin to genre markers (Biber and Conrad 2019), while others demonstrate a more pervasive influence across the document. (327 words)

References:

Biber, D. (2012), 'Register as a predictor of linguistic variation', *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Biber, D. & S. Conrad (2019), *Register, Genre, and Style*, 2nd edn., Cambridge: Cambridge University Press.

Egbert, J. & Gracheva, M. (2023). 'Linguistic variation within registers: granularity in textual units and situational parameters', *Corpus Linguistics and Linguistic Theory*, 19(1), 115–143.

Laippala, V., S. Rönnqvist, M. Oinonen, A.-J. Kyröläinen, A. Salmela, D. Biber, J. Egbert & S. Pyysalo (2023), 'Register Identification from the Unrestricted Open Web Using the Corpus of Online Registers of English', *Language Resources and Evaluation* 57, 1045–1079.

Rönnqvist, S., A.-J. Kyröläinen, A. Myntti, F. Ginter & V. Laippala. (2022), 'Explaining Classes through Stable Word Attributions', *Findings of the Association for Computational Linguistics: ACL 2022*, 1063–1074.

Sundararajan, M., A. Taly & Q. Yan (2017), 'Axiomatic Attribution for Deep Networks', *Proceedings of the 34th International Conference on Machine Learning,* 5109–5118.