Analyzing the semantic evolution of biases in French news articles using word embeddings

Louis Escouflaire[1,2]
Antonin Descampe[1]
Grégoire Lits[1]
Cédrick Fairon[2]

[1] ORM, UCLouvain
[2] CENTAL, UCLouvain

In this contribution, we investigate the possible use of word embeddings models for the sociolinguistic analysis of semantic change regarding different types of biases in a longitudinal corpus of journalistic articles.

Word embeddings are computed through a machine learning algorithm which was first introduced with the *word2vec* model (Mikolov et al., 2013). This method can be applied to a large corpus of text to represent each word as a multidimensional vector based on the lexical contexts of all appearances of the given word in the corpus. By training on the windows of occurrence of the words, the model can capture subtle semantic features carried by those words in the corpus, largely outdoing previous NLP methods of co-occurrence analysis. Through dimensional reduction, embeddings can also be visualized in a two- or three-dimensional space to get a more qualitative view at the different ways in which the words in the corpus interact with each other.

Among such semantic features and relationships, word embeddings have also been found to incorporate a wide variety of biases and stereotypes that are present in the corpus from which they are modeled (Stoltz & Taylor, 2021). Because word embeddings were initially developed as tools for NLP applications such as automated translation or recommendation algorithms (and are largely used in the field), a lot of the research on word embeddings has focused on exploring ways to debias the models (Bolukbasi et al., 2016). Keeping them from reproducing the stereotypes on which they were trained has been made a priority since their rise in popularity. However, recent work has shed the light on the opportunities provided by word embeddings for measuring the presence and importance of biases in text corpora and for detecting hidden stereotypes. For example, this method can help for analyzing the evolution of gender bias over time by training multiple word embedding models on data collected from successive periods of time (Garg et al., 2018).

We trained the *word2vec* algorithm on multiple chunks of the RTBF Corpus (Escouflaire et al., 2023), which contains 750.000 news articles published between 2008 and 2021 by the Belgian French-speaking public media RTBF (Radio-Télévision Belge Francophone). We choose to use the embeddings provided by *word2vec* and not the contextualized embeddings of more recent transformer models such as BERT (Devlin et al., 2019), because those models require very large amounts of data and are usually pretrained on datasets which contain prior biases that may alter our results. Exploring a variety of approaches, we analyze the evolution of two types of biases over 14 years of press articles: gender bias and evaluative bias. We compute mathematical operations on the word vectors to compute the distance between a selection of

words related to those biases and observe how distances between vectors evolve over time. To account for the inherent influence of grammatical gender (inherent to languages such as French) on word embeddings, we develop an approach derived from Bolukbasi et al. (2016) for building a balanced vector subspace representing a given bias axis. Several words representing typical occupations or relevant concepts can then be projected into this vector subspace to quantify how much they are affected by that bias. We first evaluate the relevance of these tools for bias measurement by running a few semantic tests using lists of inherently biased words. Then, applying this technique, we analyze the evolution of gender stereotypes for multiple words standing for occupations and social groups in Belgian French articles between 2008 and 2021, a period particularly characterized by the influence of the feminist revolution led by the #metoo movement. We show that the gender bias of some words representing occupations, such as *minister* and *pilot*, has significantly changed before and after the emergence of #metoo. Projecting the same words in a vector subspace representing evaluative bias (*good* vs. *bad*), we also examine the stereotype shifts over the period covered by the RTBF Corpus. Finally, we look for significant correlations between results concerning gender bias and evaluative bias and confirm the existence of relationships between these two types of stereotypes. Dimensionally reduced vector spaces (using t-SNE) are used to visualize our results and allow for a better understanding of our interpretations.

Our work shows the efficiency of word embeddings for analyzing gender and evaluative stereotypes, even when applied to languages in which grammatical gender is central, such as French. We also highlight the other types of bias subspaces that could be investigated using this method and present its limitations. Eventually, we explore the potential research perspectives of this approach in the field of journalism studies, with an overview of the different contributions of word embeddings to sociolinguistic analysis of biases and stereotypes in news media content.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems, 29*.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115(16)*, 3635-3644.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Stoltz, D. S., & Taylor, M. A. (2021). Cultural cartography with word embeddings. *Poetics, 88*, 101567. https://doi.org/10.1016/j.poetic.2021.101567