Adapting FlauBERT for WSD of Eighteenth-Century Aesthetic Terms

Marion Riggs
Rome, Italy
marion.c.riggs@gmail.com

Abstract

This paper argues that domain-adapted pre-trained BERT models show promise as tools for disambiguating eighteenth-century aesthetic terminology. Eighteenth-century aesthetic terms are often highly ambiguous. Pre-trained BERT models have strong word sense disambiguation (WSD) capabilities (Wiedemann, 2019) and, ideally, could be used to analyze this ambiguity. Yet pre-existing models trained on contemporary language may not perform well when it comes to disambiguation of eighteenth-century language or historical languages more broadly. New language models can be, and have been, produced for historical languages (Manjavacas, 2021; Gabay, 2022) in order to overcome these sorts of issues. Alternatively, pre-existing models can be adapted for historical languages via further training on historical corpora (Hosseini, 2021). Such adapted models benefit from the large dataset upon which the original model was trained and do not require the computational resources needed to train a new historical model from scratch.

This paper focuses on the domain adaption of a pre-existing French model, FlauBERT (Le, 2020), for the disambiguation of French eighteenth-century aesthetic terminology. The paper centers on the specific case study of the disambiguation of a single term: "baroque." "Baroque" is an aesthetic term used in various humanistic disciplines (e.g. art, music, literature). It and its cognates in other languages (e.g. "barocco" in Italian and "barock" in German) are highly ambiguous. The historical meaning of "baroque" garnered a substantial amount of scholarly attention across various humanities disciplines in the twentieth century. Nonetheless, the ambiguity of "baroque" in the eighteenth century, the period in which it came to be used as an aesthetic term, has never been fully explored. A FlauBERT model adapted for eighteenth-century French provides a tool with which to explore this ambiguity.

This paper also considers what kind of training data will allow for the optimization of the adaption of FlauBERT for eighteenth-century French and for the specific task of disambiguating "baroque." To explore this, a number of different adapted models are trained using different kinds of corpora. FlauBERT is further trained on (1) a relatively small corpus of eighteenth-century French from Wikisource, (2) a lemmatized version of this Wikisource corpus, (3) a much larger corpus of dirty OCR'd eighteenth-century French from the Bibliothèque nationale de France (BnF), and (4) a portion of the BnF corpus from texts with relatively high OCR rates.

This paper evaluates how well these adapted models perform in relation to each other and in relation to the original FlauBERT model on the specific task of disambiguating "baroque." Clustering patterns of contextualized word embeddings (CWEs) visualized in t-SNE scatterplots are used as the basis of a comparison. FlauBERT and each of the four adapted models are used to generate CWEs of "baroque" from a corpus of eighteenth-century French instances of the word. The corpus is sense-annotated and these annotations are used to colour code the embeddings when they are visualized in the scatterplots. The resulting scatterplots reveal the WSD capabilities of each model and allow for the models to be compared.

The research for this specific case study has broader implications for computational research in the Humanities. Aesthetic terminology is often ambiguous, vague, and semantically unstable. The adaptation of existing BERT models for particular languages in specific periods and the CWEs generated by these models taken together offer a way to investigate and analyze this kind of language in a way that other methodological approaches do not allow for. Moreover, the visualizations of these contextualized embeddings -- and t-SNE scatterplots more specifically -- provide a way to visualize results in a way that can be easily understood by a non-expert audience.

References

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagu'e, Rachel Bawden, Philippe Gambette and Benoît Sagot (2022). From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French. ArXiv, abs/2202.09452.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy (2021). Neural Language Models for Nineteenth-Century English. *Journal of Open Humanities Data*, 7, 22.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, . . . Didier Schwab (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Enrique Manjavacas and Lauren Fonteyn (2021). MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on NLP4DH*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *15th Conference on Natural Language Processing*.