Something wicked this way comes: Analyzing language variation and change with MacBERTh

Lauren Fonteyn (Universiteit Leiden)

Because of great efforts by the corpus-linguistic community, a large number of historical texts have been digitized (and sometimes even syntactically parsed and pos-tagged), which has enabled the automatic retrieval of words/phrases/sentence structures by means of formal queries. The next step in corpus querying, then, would be to move from formal querying to functional-semantic querying, but this has proven a difficult challenge in the past. However, in recent years, it became evident that on-going progress in distributional semantic models, from type-embeddings derived from algorithms like word2vec (Mikolov et al, 2013) to contextualized token-embeddings such as BERT (Devlin et al. 2019), can capture the denotations and connotations of linguistic items. This presentation will focus on MacBERTh, a BERT-based model pre-trained on Early Modern and Late Modern English (3.9B (tokenized) words, time span: 1450-1950; Manjavacas & Fonteyn 2022). This presentation demonstrates how MacBERTh may help researchers to (i) access and (ii) analyse the functional-semantic information encoded in linguistic corpus data in a (semi-)automatic way. More specifically, after surveying the performance of the model on general downstream NLP tasks, I home in on two specific case studies. The first case study homes in on the lexical-semantic changes affecting the scientific terms mass and weight in the Modern English period. The second case study homes in on morphosyntactic variation in English ing-forms.

References

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. In *Proceedings of NAACL-HLT 2019*, 4171–86. Minneapolis, Minnesota, 2019.

Manjavacas, Enrique and Lauren Fonteyn. 2022. Adapting vs Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*. https://hal.inria.fr/hal-03592137/document

Mikolov, Tom s, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. 'Efficient Estimation of Word Representations in Vector Space'. In *1st International Conference on Learning Representations, ICLR 2013*, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, edited by Yoshua Bengio and Yann LeCun, 2013. http://arxiv.org/abs/1301.3781