Investigating World Englishes modal verbs using BERT embeddings – Can BERT accelerate World Englishes research?

Jonas Wagner (Universität Bielefeld)

Modal Verbs have been researched in the field of World Englishes for over thirty years now, but while the frequency of each type is simple to calculate and track using well-constructed corpora (such as the International Corpus of English [ICE]), modal verb sense investigation requires extensive manual data exploration and annotation, making it a lengthy and costly affair if one wants to maintain a large sample size. Meanwhile, investigations of (synchronic) semantic differences and (diachronic) semantic shifts have been greatly facilitated by the introduction of word embeddings, which, ostensibly, allow researchers to quantify words' meanings and conduct semantic research en masse. Contextualised embeddings, which can be extracted using pre-trained models like BERT, can even distinguish between different word senses. However, BERT's ability to capture modal verb senses has not yet been established.

The research in this thesis therefore aims to answer three questions:

a. Can BERT embeddings capture modal verb sense?

b. Can differences in modal verb sense usage across World Englishes be detected with BERT?

c. Does BERT perform differently in different varieties of English?

To do this, a series of experiments will be conducted:

1. BERT generated replacements of masked modal verbs in annotated corpora
2. Clustering of modal verbs from annotated corpora by their sense
3. Clustering of modal verbs from different components of ICE
4. Repetition of 1. and 2. on selected manually annotated phrases from ICE components

Initial results point to BERT being able to capture modal verb sense in experiments 1 and 2.