Corpus-Guided Contrast Sets for Morphosyntactic Feature Detection in Low-Resource English Varieties

Tessa Masis, Anissa Neal, Lisa Green, Brendan O’Connor

Paper, models, data: https://github.com/slanglab/CGEdit

Morphosyntactic Features

I just be liking the beat → contains habitual be feature

Goal: given a list of features F, for each f ∈ F identify utterances which contain f

Trained Feature Detectors

Approach: broadly following Demszky et al. (2021), we fine-tune BERT on contrast sets generated via proposed CGEdit method

Corpora

ICE-India

CORAAL

FWP

Annotations from Lange (2012)

no previous annotations

Indian English

African American English

Intrinsic Evaluation

Extrinsic Evaluation

Confirmed + extended three sociolinguistic studies on CORAAL, which used manual feature annotation to examine if feature use aligned with social factors

Summary & Future Work

- Generate morphosyntactically diverse contrast sets via simple corpus-guided edits
- Feature detection improves by 16 points in Prec@100 scores by fine-tuning on corpus-guided contrast sets
- Extended prior findings on CORAAL to externally validate use for linguistic research
- Ongoing project (Masis et al., NWAV50) uses this method to analyze regional variation of feature use