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
University of Massachusetts Amherst

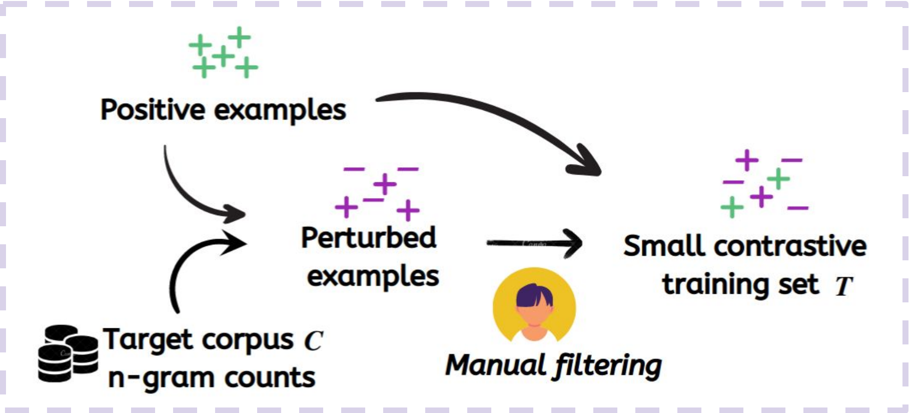
Morphosyntactic Features

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



Positive examples

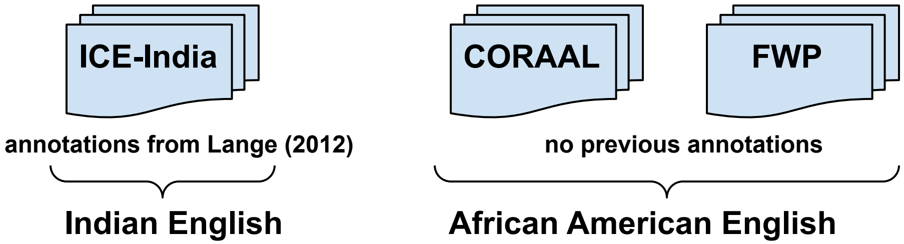
Perturbed examples

Small contrastive training set T

Target corpus C n-gram counts

Manual filtering

Corpora



ICE-India

CORAAL

FWP

annotations from Lange (2012)

no previous annotations

Indian English

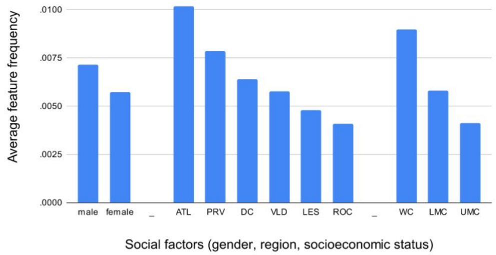
African American English

Intrinsic Evaluation

	ICE-India			CORAAL	FWP
Approach	ROC-AUC	AP	Prec@100	Prec@100	Prec@100
AUTOGEN	68.94	12.63	16.93	-	-
AUTOID	74.90	15.24	17.87	-	-
MANUALGEN	86.83	25.77	31.63	57.88	58.71
AUTOID + MANUALGEN	76.34	19.95	24.30	-	-
CGEDIT	84.92	27.48	32.50	**67.41**	68.00
MANUALGEN + CGEDIT	**88.76**	**29.32**	**35.67**	64.94	**74.35**

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