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

Tessa Masis[1]
they/them/theirs

Anissa Neal[2]
she/her/hers

Lisa Green[2]
she/her/hers

Brendan O'Connor[1]
he/him/his

[1]College of Information and Computer Sciences
University of Massachusetts Amherst

[2]Department of Linguistics
University of Massachusetts Amherst

Overview

- Task
  - Automatic detection of morphosyntactic features
- Approach
  - Novel method for generating contrast sets
  - Fine-tune large pretrained LM
- Data
  - 3 transcript corpora of nonstandard Englishes
- Results
  - Intrinsic & extrinsic evaluations

What is a morphosyntactic feature?

habitual *be*:	I be out at my bus stop every day.
zero copula:	He on the five dollar bill.
finna:	I'm finna be late.

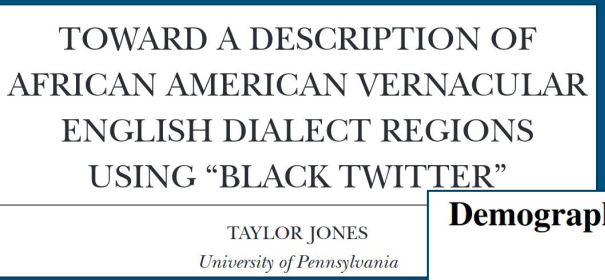
Automatic feature detection

- Task: given textual data, detect specific morphosyntactic features

- Feature detection is useful for linguistic analyses, language ID, etc
- Automatic methods are a valuable alternative to manual annotation

Automatic feature detection

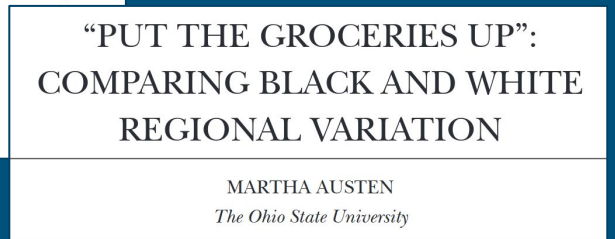
- Accurately detecting morphosyntactic features in nonstandard/low-resource languages or in informal genres (e.g. transcripts, social media) is challenging
  - Variable spellings make keyword searches tricky
  - Regular expressions can't be made for all features
  - Don't have large labeled datasets, so supervised learning -> noisy classifiers


TOWARD A DESCRIPTION OF AFRICAN AMERICAN VERNACULAR ENGLISH DIALECT REGIONS USING "BLACK TWITTER"

TAYLOR JONES
University of Pennsylvania


Demographic Dialectal Variation in Social Media: A Case Study of African-American English

Su Lin Blodgett† Lisa Green* Brendan O'Connor†


"PUT THE GROCERIES UP": COMPARING BLACK AND WHITE REGIONAL VARIATION

MARTHA AUSTEN
The Ohio State University

Automatic feature detection: our framework

Learning to Recognize Dialect Features

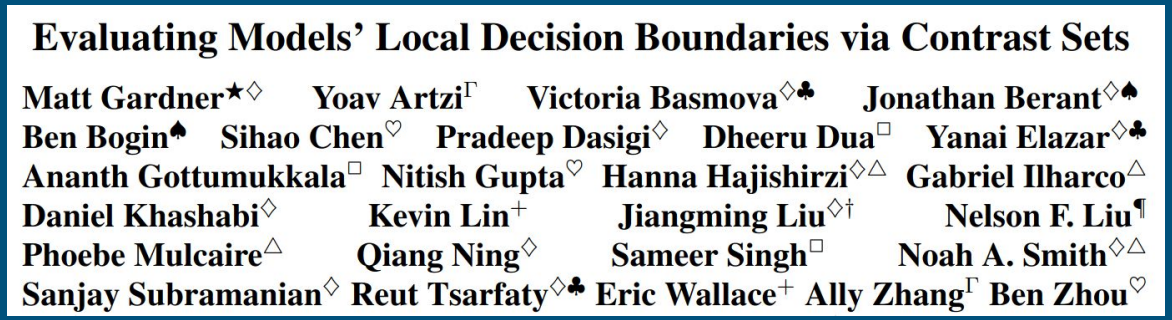
Dorottya Demszky[1]* Devyani Sharma[2] Jonathan H. Clark[3]
Vinodkumar Prabhakaran[3] Jacob Eisenstein[3]
[1]Stanford Linguistics [2]Queen Mary University of London [3]Google Research

- Generate a small contrast set

- Fine-tune BERT on this contrast set

Automatic feature detection

- Generate a small contrast set
  - A labeled collection of positive and negative examples that are highly similar, where a positive example has the feature/label and a negative example does not



Evaluating Models' Local Decision Boundaries via Contrast Sets

Matt Gardner[★◇] Yoav Artzi[Γ] Victoria Basmova[◇♣] Jonathan Berant[◇♠]
Ben Bogin[♠] Sihao Chen[♡] Pradeep Dasigi[◇] Dheeru Dua[□] Yanai Elazar[◇♣]
Ananth Gottumukkala[□] Nitish Gupta[♡] Hanna Hajishirzi[◇△] Gabriel Ilharco[△]
Daniel Khashabi[◇] Kevin Lin[+] Jiangming Liu[◇†] Nelson F. Liu[¶]
Phoebe Mulcaire[△] Qiang Ning[◇] Sameer Singh[□] Noah A. Smith[◇△]
Sanjay Subramanian[◇] Reut Tsarfaty[◇♣] Eric Wallace[+] Ally Zhang[Γ] Ben Zhou[♡]

Automatic feature detection

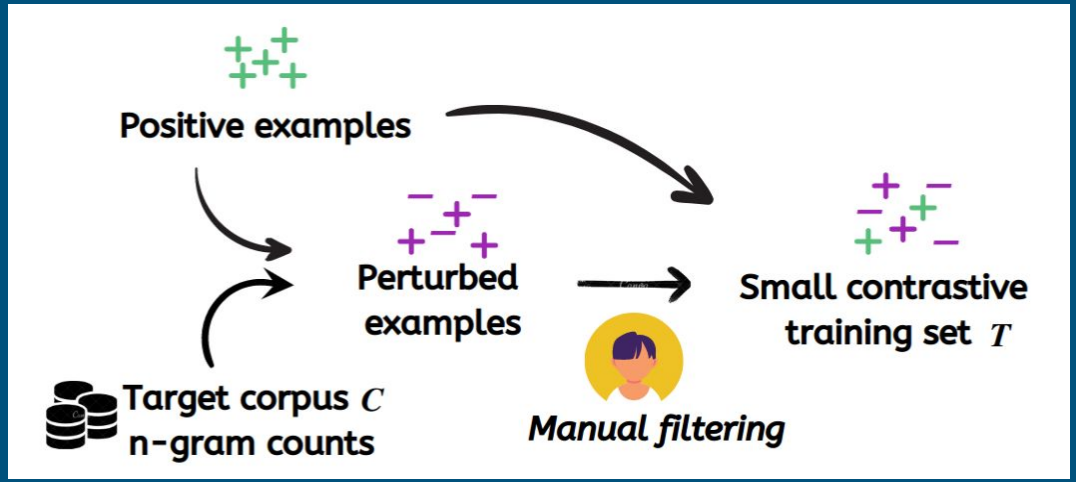
- Generate a small contrast set
  - A labeled collection of positive and negative examples that are highly similar, where a positive example has the feature/label and a negative example does not

I be out at my bus stop every day.

I'm out at my bus stop every day.
I'll be out at my bus stop every day.
I would be out at my bus stop every day.

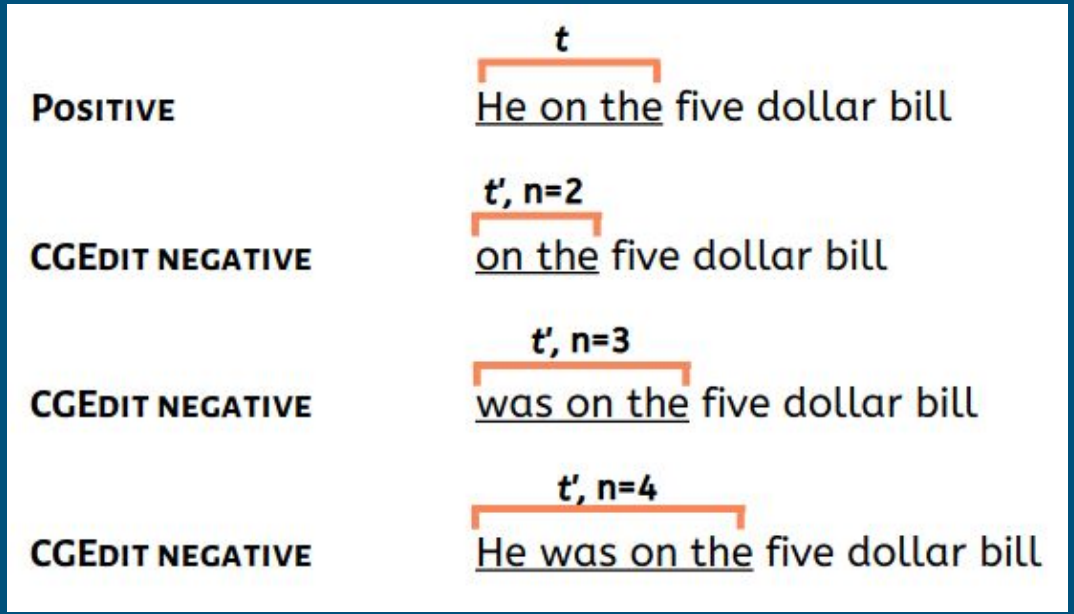
Generate contrast sets: CGEdit

- Input:
  - Seed set of positive examples
  - Target corpus n-gram counts
- Method:
  - Corpus-guided edits
  - Human-in-the-loop filtering
- Output:
  - Morphosyntactically contrastive training data



Example: corpus-guided edits

Feature: zero copula (omission of a copula i.e. *is, are*)



Example: human-in-the-loop filtering

Perturbed example
He on the last five
He on the five
on the other five dollar
He on the five hundred dollar
He was on the dollar
on the five dollar
the on five dollar
He and five on the dollar
He was on the five dollar
He on the five dollar bill
He beating on the five dollar
He on the dollar
He on the other dollar
He on five dollar
He the five dollar
He on five dollar bill
was on the five dollar

Manual filtering

Example	Label
He on the five dollar	1
He on the last five	1
He on the five	1
on the other five dollar	0
He was on the dollar	0
on the five dollar	0

Automatic feature detection

- Generate a small contrast set

- Fine-tune BERT on this contrast set, where each head is a binary classifier for a single feature

Data

Indian English (IndE) corpora:

- **ICE-India**: International Corpus of English India subcorpus
  - 1990 - 1993

African American English (AAE) corpora:

- **CORAAL:** Corpus of Regional African American Language
  - 1968 - 2017
- **FWP:** Slave Narratives from the Federal Writers' Project
  - 1936 - 1938

Feature lists

IndE Feature	Example utterance
Non-initial existential *there*	library facility was not there
Focus *itself*	We are feeling tired now itself
Focus *only*	I like dressing up I told you at the beginning only
Zero copula	Everybody (is) so worried about the exams

AAE Feature	Example utterance
Zero possessive -*'s*	go over my grandmama('s) house
Zero copula	she (is) the folk around here
Double marked/overregularized	she likeded me the best
Habitual *be*	I just be liking the beat

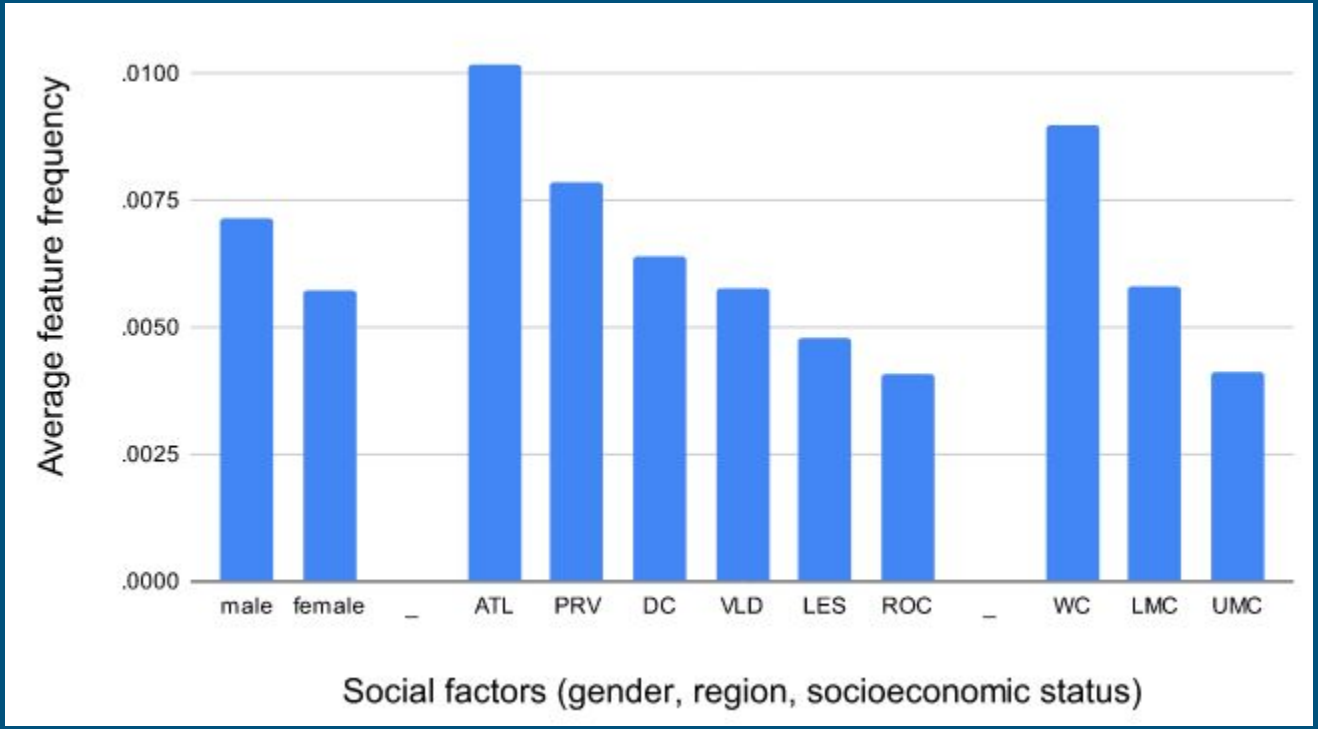
Complete feature lists for both IndE and AAE in Appendix A

Intrinsic evaluation

Approach	ICE-India	CORAAL	FWP
MANUALGEN	31.63	57.88	58.71
CGEDIT	32.50	**67.41**	68.00
MANUALGEN + CGEDIT	**35.67**	64.94	**74.35**

Table 1: Precision@100 in percentages for feature detection on all three corpora. Results are averages over all features (10 in ICE-India, 17 in CORAAL and FWP). Reported scores for ICE-India are averaged from three runs with different random seeds. Best scores are bolded.

Extrinsic evaluation



Summary

- Generate morphosyntactically diverse contrast sets via CGEdit method using simple corpus-guided edits

- Improves feature detection by up to 16 points in Prec@100 scores

- Extended prior findings on CORAAL to externally validate utility for linguistic research

Thank you!

Slides and paper available at tmasis.github.io/

This material is based upon work supported by the National Science Foundation under grant BCS-2042939. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Tessa Masis
tmasis@cs.umass.edu

Anissa Neal
anneal@linguist.umass.edu

Lisa Green
lgreen@linguist.umass.edu

Brendan O'Connor
brenocon@cs.umass.edu