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

Tessa Masis¹
she/her/hers

Anissa Neal²
she/her/hers

Lisa Green²
she/her/hers

Brendan O’Connor¹
he/him/his

¹College of Information and Computer Sciences
University of Massachusetts Amherst

²Department 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
Automatic feature detection: our framework

- 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† 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† 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*)

- **Positive**: He on the five dollar bill
- **CGEdit Negative**: on the five dollar bill
- **CGEdit Negative**: was on the five dollar bill
- **CGEdit Negative**: He was on the five dollar bill
Example: human-in-the-loop filtering

<table>
<thead>
<tr>
<th>Perturbed example</th>
</tr>
</thead>
<tbody>
<tr>
<td>He on the last five</td>
</tr>
<tr>
<td>He on the five</td>
</tr>
<tr>
<td>on the other five dollar</td>
</tr>
<tr>
<td>He on the five hundred dollar</td>
</tr>
<tr>
<td>He was on the dollar</td>
</tr>
<tr>
<td>on the five dollar</td>
</tr>
<tr>
<td>the on five dollar</td>
</tr>
<tr>
<td>He and five on the dollar</td>
</tr>
<tr>
<td>He was on the five dollar</td>
</tr>
<tr>
<td>He on the five dollar bill</td>
</tr>
<tr>
<td>He beating on the five dollar</td>
</tr>
<tr>
<td>He on the dollar</td>
</tr>
<tr>
<td>He on the other dollar</td>
</tr>
<tr>
<td>He on five dollar</td>
</tr>
<tr>
<td>He the five dollar</td>
</tr>
<tr>
<td>He on five dollar bill</td>
</tr>
<tr>
<td>was on the five dollar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>He on the five dollar</td>
<td>1</td>
</tr>
<tr>
<td>He on the last five</td>
<td>1</td>
</tr>
<tr>
<td>He on the five</td>
<td>1</td>
</tr>
<tr>
<td>on the other five dollar</td>
<td>0</td>
</tr>
<tr>
<td>He was on the dollar</td>
<td>0</td>
</tr>
<tr>
<td>on the five dollar</td>
<td>0</td>
</tr>
</tbody>
</table>
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

### None

<table>
<thead>
<tr>
<th>IndE Feature</th>
<th>Example utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-initial existential <em>there</em></td>
<td>library facility was not <em>there</em></td>
</tr>
<tr>
<td>Focus <em>itself</em></td>
<td>We are feeling tired now <em>itself</em></td>
</tr>
<tr>
<td>Focus <em>only</em></td>
<td>I like dressing up I told you at the beginning <em>only</em></td>
</tr>
<tr>
<td>Zero copula</td>
<td>Everybody (is) so worried about the exams</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AAE Feature</th>
<th>Example utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero possessive -’s</td>
<td>go over my grandmama(’s) house</td>
</tr>
<tr>
<td>Zero copula</td>
<td>she (is) the folk around here</td>
</tr>
<tr>
<td>Double marked/overregularized</td>
<td>she <em>liked</em>ed me the best</td>
</tr>
<tr>
<td>Habitual <em>be</em></td>
<td>I just <em>be</em> liking the beat</td>
</tr>
</tbody>
</table>
## Intrinsic evaluation

<table>
<thead>
<tr>
<th>Approach</th>
<th>ICE-India</th>
<th>CORAAL</th>
<th>FWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANUALGEN</td>
<td>31.63</td>
<td>57.88</td>
<td>58.71</td>
</tr>
<tr>
<td>CGEdit</td>
<td>32.50</td>
<td>67.41</td>
<td>68.00</td>
</tr>
<tr>
<td>MANUALGEN + CGEdit</td>
<td>35.67</td>
<td>64.94</td>
<td>74.35</td>
</tr>
</tbody>
</table>

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