Dialect Processing in NLP: African-American English (AAE)

[Blodgett et al., 2016, Blodgett et al., 2018]

Leo Born

HS: Bias
Prof. Dr. Katja Markert
Institute for Computational Linguistics
Ruprecht-Karls-Universität Heidelberg

14.01.2020

Contents

Contents

Dialect vs. *Dialekt*

- In German usage, *Dialekt* has (clear) geographic boundaries and connotations
- However, *dialect* is more general than *Dialekt*
- Usage corresponds to German term *Varietät* of which *Dialekt* is one possible form [Bußmann, 2008]

- What we will deal with represents a *sociolect*

Contents

AAE: Linguistic Characteristics

- Variety of English with phonological, syntactic, semantic, and lexical patterns associated with a subset of African-American communities [Green, 2002]
- Common phonological patterns across AAE variants:
  - Voiced *th* as *d*: *dey*, *dat*, *dis*, *dere*
  - *Derhotacization*: *brotha* (*brother*), *ova* (*over*)
  - Other variations: *wea* (*where*), *sholl* (*sure*), *iont* (*I don't*)

AAE: Linguistic Characteristics

- Common syntactic patterns:
  - Aspect-based:
    - Habitual *be*: *They be running*
    - Future *gone*: *He gone be disappointed*
    - Completive *done*: *They done left*
  - Null copulas: *Where you at?*
  - Null auxiliaries: *If u wit me den u pose to RESPECT ME*

Contents

Bias

- Most NLP tools have been trained with Standard American English (SAE) data:
    - Language identification tools have a hard time detecting AAE as English
    - Parsing accuracy is lower for AAE

$\rightarrow$ Downstream applications such as sentiment or opinion analysis can either under- or misrepresent AAE speakers

Bias

Figure: Examples of AA-aligned tweets.

Bias

- *langid.py*[1] results for previous tweets:

```
>>> aint bout nuffin datz how im coming
('de', -74.3771800994873)
>>> yea u def blessed!!! lolol
('nl', -23.63649320602417)
>>> i aint got nuffin for u hoes i need str8 money
('da', -49.66361713409424)
```

[1] https://github.com/saffsd/langid.py

Bias

- While this is a form of disparate impact, it differs from what we have seen so far:
  - Explicitly linguistic bias
  - Impact is both predicated upon and results in under- and/or misrepresentation of minorities

Contents

TwitterAAE [Blodgett et al., 2016]

- Subset of Twitter messages highly associated with AAE
- Dataset consisting of 830,000 tweets
- Used to validate linguistic phenomena associated with AAE and to investigate disparities in NLP tool performance

- Furthermore, serves as data for subsequent work [Blodgett et al., 2018]

Contents

TwitterAAE: Dataset Construction

- Two-step process:
  1. Find messages on Twitter cross-referenced against US Census demographics data
  2. Topic modeling with demographics as topics

- Prerequisites:
  - Tweets with geodata
  - Tweets were *casual and conversational*:
    - Users with more than 1,000 followers were excluded
    - Retweets were ignored
    - Messages containing more than three hashtags or containing "http", "follow", and "mention" were excluded

TwitterAAE: Dataset Construction

- Tweets from 2013:
  - 59.2 million tweets
  - 2.8 million users
- Each tweet is associated with a US Census blockgroup[2]
- For each blockgroup, race and ethnicity information is used from 2013 Census:
  - % of non-Hispanic white population
  - % of non-Hispanic black population
  - % of Hispanic population
  - % of Asian population

  $\rightarrow$ Each user u gets length-four vector $\pi_u^{(census)}$ by averaging all demographic values of all of u's messages

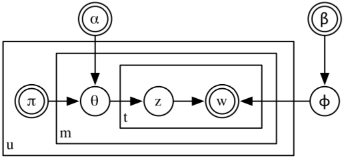
[2] Geographical area containing ca. 600 - 3000 people.

Twitter AAE: Dataset Construction

- Each demographic category is associated a topic via unigram LM over vocabulary[3]
- LDA model over users and messages
  - Allows for multidialectal users
- Posterior probability of a user u using some topic k is fraction of tokens with topic k in all messages by u

[3] Words used by at least 20 users; 191,873 types.

Twitter AAE: Dataset Construction



$$\theta_m \sim Dir(\alpha\pi_u), \;\; \phi \sim Dir(\beta/V)$$
$$z_t \sim \theta_m, \;\; w_z \sim \phi_{z_t}$$

Figure: LDA model for demographic inference.

Twitter AAE: Dataset Construction

- Correlation of model's posterior demographics' proportions and Census-derived proportions was > 0.8 for all demographics but Asian
- Many Spanish terms ended up in Asian topic

  $\rightarrow$ Uncertainties regarding validity of Asian and Hispanic topics

  $\rightarrow$ [Blodgett et al., 2016] only consider AA and white demographics

Twitter AAE: Dataset Construction

- AA-aligned corpus:
  - All tweets from users whose posterior probability for AA was $> 80\%$
- White-aligned corpus:
  - All tweets from users whose posterior probability for white was $> 80\%$
- Constraint: each user's combined posterior probability of Hispanic and Asian was $< 5\%$

Contents

Data Analysis

- Lexical variations (check against SCOWL dictionary, ca. 630,000 words):
  - For words at least twice as likely to be AA-aligned than white-aligned ($r_{AA}(w) \geq 2$), 79.1% were not in dictionary
  - For words at least twice as likely to be white-aligned than AA-aligned ($r_{white}(w) \geq 2$), 58.2% were not in dictionary
- [Addendum] High values for both might be due to spelling variants common to Twitter[4]

[4] See e.g. http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

Data Analysis

- Phonological variations:
  - 31 variants of SAE words from previous literature were selected
  - For all words, $r_{AA}(w)$ was calculated
  - For 30 out of 31 $r_{AA}(w) \geq 1$[5] and for 13 $r_{AA}(w) \geq 100$

AAE	Ratio	SAE
sholl	1802.49	sure
iont	930.98	I don't
wea	870.45	where
talmbout	809.79	talking about
sumn	520.96	something

Figure: Top five SAE word variations and their AA-alignment ratios.

[5] Exception was *brotha*.

Data Analysis

- Syntactic variations:
  - Sequence of unigrams and POS tags used to extract occurences of three syntactic patterns: habitual *be*, future *gone*, completive *done*
  - All tweets were split into deciles based on posterior AA probability
  - From each decile, 200,000 tweets were sampled to calculate frequency of syntactic patterns

Data Analysis

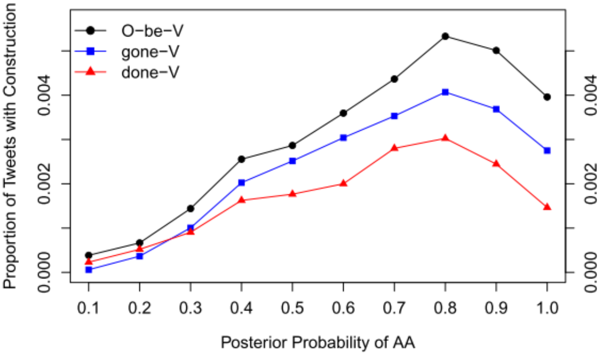


Figure: Frequencies of common AAE syntactic constructions given AA probability.

Data Analysis

Feature	AA Count	WH Count	Example
Dropped copula	44	0	*MY **bestfrienddd mad** at me tho*
Habitual *be*, describing repeated actions	10	0	*fees **be** looking upside my head likee ion kno wat **be** goingg on .* *I kno that clown, u don't **be** around tho*
Dropped possessive marker	5	0	*ATMENTION on Tv...tawkn bout dat **man** gf Twink rude lol can't be calling ppl ugly that's **somebody** child lol...*
Dropped 3rd person singular	5	0	*When a female **owe** you sex you don't even wanna have a conversation with her*
Future *gone*	4	0	*she **gone** dance without da bands lol*
it is instead of *there is*	2	1	***It was** too much goin on in dat mofo .*
Completive *done*	1	0	*damnnn I **done** let alot of time pass by . .*

Figure: Frequencies of common AAE patterns in a sample of 250 AA- und 250 white-aligned tweets.

Contents

Results on Language Identification

- AAE should be classified as English
- Test of *langid.py* and Twitter's identifier whose results are provided in tweet metadata
- From classified "non-English" tweets, 50 per tool-data pair were manually checked
    - Only 3 were really not English

	AAE	White-Aligned
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Figure: Tweets classfied as non-English.

- As messages' posterior AA probability increases, proportion of "non-English" classification rises

Contents

Critique

- Questionable to associate origin of tweet with neighborhood a person supposedly lives in
- No examples of really not-English tweets
- Unclear what the median number of tweets per user is
- Retrieval of orthographic variations only vaguely mentioned
- No examples of OOV words

Contents

Contents

Universal Dependencies [Nivre et al., 2016]

- Designed as a language-independent syntactic annotation framework:
    - Combined several existing frameworks

$\rightarrow$ Dialects can be treated as own languages, therefore previous language-specific frameworks unsuitable

Universal Dependencies [Nivre et al., 2016]

- 40 relations (excerpt):

Core dependents of clausal predicates		
Nominal dep	*Predicate dep*	
nsubj	csubj	
nsubjpass	csubjpass	
dobj	ccomp	xcomp
iobj		
Non-core dependents of clausal predicates		
Nominal dep	*Predicate dep*	*Modifier word*
nmod	advcl	advmod
		neg
Special clausal dependents		
Nominal dep	*Auxiliary*	*Other*
vocative	aux	mark
discourse	auxpass	punct
expl	cop	
Noun dependents		
Nominal dep	*Predicate dep*	*Modifier word*
nummod	acl	amod
appos		det
nmod		neg
Case-marking, prepositions, possessive		
case		
Coordination		
conj	cc	punct

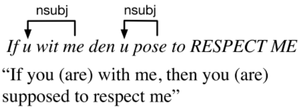
Contents

Annotating AAE with UD

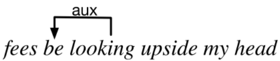
- Data:
  - 500 tweets sampled from TwitterAAE
  - 250 AA-aligned and 250 white-aligned tweets
  - Manual annotation of tweets by two annotators

Annotating AAE with UD

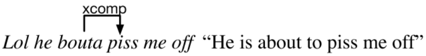
- Null copulas and null auxiliaries:
    - Simply omit *cop* and *aux* edges



- Habitual *be*, future *gone*, completive *done*:
    - Handled as verbal auxiliaries → *aux* edge to main verb gets added



- Verbal contractions (e.g. *about to* → *bouta*):
    - UD handles similar SAE constructions (*want to*) as main verbs, so do the same here



Contents

Models

- Dependency parsing:
  - UDPipe [Straka et al., 2016]
  - Deep Biaffine [Dozat et al., 2017]
- POS tagging:
  - UDPipe's internal POS tagger (*Morpho-Tagger*)
  - ARK POS Tagger [Owoputi et al., 2013]
- Word embeddings:
  - 200-dimensional *word2vec* [Mikolov et al., 2013] embeddings trained on TwitterAAE

Data settings

- Cross-domain and in-domain scenarios
- In-domain scenario:
  - UDPipe with ARK POS tagger, Twitter embeddings
  - 2-fold cross-validation, random 250/250 train/test splits
  - Twitter-only vs. Twitter+UDT

Results: In-Domain Training

Model	LAS
(10) UDPipe, Twitter embeddings	62.2
(11) + UDT	70.3

- Fairly acceptable results given the small dataset
- Even though UDT is non-Twitter data, inclusion increases performance

Data settings

- Cross-domain scenario (train on UDT, test on TwitterAAE):
  1. Re-train UDPipe parser both with in-house POS tagger as well as ARK tagger results
  2. Add synthetic data
     - Insertion of e.g. @-mentions, emoticons, hashtags
     - Insertion of AAE constructions that are infrequent in UDT (e.g. collapsing *about to* -> *bouta*; replacing *will* with *gone*; deleting copulae)
  3. Compare pre-trained with custom word embeddings

Results: Cross-Domain Settings

Model	LAS
(1) UDPipe, Morpho-Tagger, UDT	50.5
(2) + Twitter embeddings	53.9
(3) + synthetic, Twitter embeddings	58.9
(4) UDPipe, ARK Tagger, UDT	53.3
(5) + Twitter embeddings	58.6
(6) + synthetic, Twitter embeddings	64.3
Deep Biaffine, UDT	
(7) + CoNLL MAE embeddings	62.3
(8) + Twitter embeddings	63.7
(9) + synthetic, Twitter embeddings	65.0

- ARK tagger outperforms Morpho-Tagger
- Larger improvements when using Twitter embeddings and synthetic data
    - However, synthetic data improvement might be due to increased training size

Results: AAE/SAE disparities

Model	AA LAS	WH LAS	Gap
(1) UDPipe, Morpho-Tagger	43.0	57.0	14.0
(2) + Twitter embeddings	45.5	61.2	15.7
(3) + synthetic, Twitter embeddings	50.7	66.2	15.5
(4) UDPipe, ARK Tagger	50.2	56.1	5.9
(5) + Twitter embeddings	54.1	62.5	8.4
(6) + synthetic, Twitter embeddings	59.9	68.1	8.2
Deep Biaffine, ARK Tagger			
(7) + CoNLL MAE embeddings	56.1	67.7	11.6
(8) + Twitter embeddings	58.7	66.7	8.0
(9) + synthetic, Twitter embeddings	59.9	70.8	10.9

- Performance gap between AA- and white-aligned tweets
- ARK tagger raises AA performance and reduces gap
- Adding synthetic data and Twitter embeddings boosts performance but increases gap

Contents

Critique

- Highly important topic and motivation
- Showed that current NLP tools fail on dialects

- However, no clear implications as to potential consequences

Contents

Discussion

- What future perspectives do you see in this work?
- What do you think about the dataset construction?

Bibliography I

Blodgett, S. L., Green, L., and O'Connor, B. (2016).
Demographic Dialectal Variation in Social Media: A Case Study of African-American English.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Blodgett, S. L., Wei, J. T. Z., and O'Connor, B. (2018).
Twitter universal dependency parsing for African-American and mainstream American English.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1415–1425.

Bußmann, H. (2008).
Lexikon der Sprachwissenschaft.
Kröner, Stuttgart, 4. edition.

Bibliography II

Dozat, T., Qi, P., and Manning, C. D. (2017).
Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task.
In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Green, L. J. (2002).
African American English.
Cambridge University Press.

Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013).
Efficient Estimation of Word Representations in Vector Space.
In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

Bibliography III

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016).
Universal Dependencies v1: A Multilingual Treebank Collection.
Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, pages 1659–1666.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013).
Improved Part-of-Speech Tagging for Online Conversational Text.
In *Proceedings ofNAACL-HLT 2013*, pages 380–390.

Straka, M., Hajič, J., and Straková, J. (2016).
UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and parsing.
In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4290–4297.