VALUE: Understanding Dialect Disparity in NLU

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, Diyi Yang

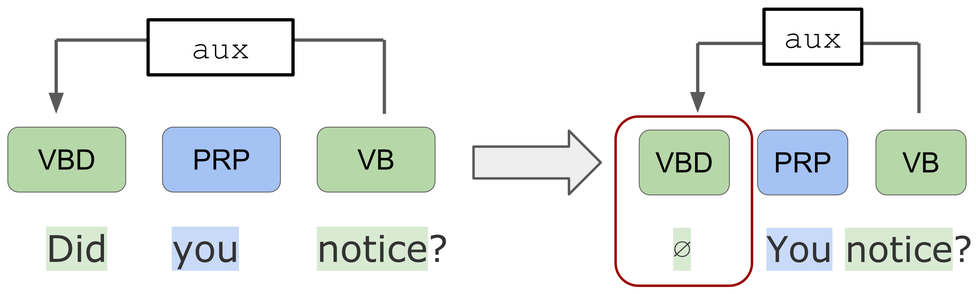
Georgia Institute of Technology
cziems@gatech.edu

GT-SALT/value

1. Introduction

- **Goal:** Measure and understand **dialect disparity** in NLU systems
- **Contributions:**
  - **Dialect Transformations:** 11 SAE → AAVE transformation rules
  - **Validation:** Robust validation of synthetic transformations + gold data from native speakers
  - **VALUE:** AAVE benchmark for 7 NLU tasks
  - **Benchmark Evaluation:** Experiments with RoBERTA baselines
  - **Dialect-Specific Analysis:** Demonstrate specific challenges of grammatical features
- **Advantages:**
  1. **Interpretable** (not **black-box**)
  2. **Flexible** (tunable **feature-density**)
  3. **Scalable** (**mix + match** datasets)
  4. **Responsible** (**participatory design**)

2. Dialect Transformations



1. Train word2vec on: **TwitterAAE** dataset
2. Linguistic code axis: $\mathbf{c} = \sum_{(\mathbf{x_i}, \mathbf{y_i}) \in S} \frac{\mathbf{x_i} - \mathbf{y_i}}{|S|}$
3. Rank candidate word pairs by: $\cos(\mathbf{c}, \mathbf{w_i} - \mathbf{w_j})$
4. Hand-filter any semantically unequal words

SAE	AAVE
arguing	beefing, beefin, arguin
anymore	nomore, nomo
classy	fly
rad	dope
screaming	screamin, yellin, hollerin
these	dese, dem
with	wit

3. VALUE Statistics

Dataset	# data	aux	been	dey/it	got	lexical	neg cncrd	null gen	null relcl	uninflect
CoLA	1,063	15%	6%	2%	2%	51%	4%	3%	3%	17%
MNLI	9,682	20%	9%	4%	5%	69%	4%	11%	10%	23%
QNLI	5,725	42%	2%	1%	3%	50%	1%	10%	4%	17%
QQP	390,690	2%	3%	63%	3%	59%	1%	3%	3%	13%
RTE	3,029	40%	36%	3%	5%	81%	4%	28%	25	40%
SST-2	1,821	25%	5%	3%	4%	64%	4%	14%	15%	39%
STS-B	1,894	~0	32%	2%	3%	2%	9%	4%	2%	5%
WNLI	146	36%	38%	3%	16%	90%	1%	37%	12%	33%

4. Validation

- **Validation Goal:** confirm that our rules are aligned with real AAVE speakers' grammars
- **Gold Standard:** if the transformation is unacceptable, annotators provide an alternative "translation"

Sentence (1): can't nothing good happen
Sentence (2): nothing good can happen

Gold Set

Task.	#Gold
MNLI	656
QNLI	663
QQP	669
RTE	192
SST-2	151
STS-B	264
WNLI	285

Validation Results

Transf.	Acc
aux	96.6
been	95.4
dey/it	91.4
gonna	95.4
got	96.2
neg cncrd	95.9
null gen	95.0
null relcl	94.1
uninflect	97.1

5. Benchmarking

- Base RoBERTa performance drops on full AAVE set
- In-domain training helps models start to move towards closing the performance gap
- We need more dialect-robust NLU systems

Train ⬇️ Test ➡️	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
CoLA: GLUE	**56.3**	**55.6**	-
MNLI: GLUE	**83.6**	**82.5**	**82.1**
QNLI: GLUE	**92.8**	**91.4**	**91.2**
RTE: GLUE	**66.4**	**67.8**	**67.6**
SST-2: GLUE	**94.6**	**92.4**	**92.0**
STS-B: GLUE	**89.4**	**88.5**	**88.2**
QQP: GLUE	**90.9**	**89.5**	**89.2**

6. Conclusion

Limitations

1. VALUE should not be considered *natural* AAVE
   → Exaggerated feature density [stress test]
2. 🗣️ Speech ≠ orthography ✏️
3. Synthetic test does not prove real-world readiness
4. Misuse: hateful speech and appropriation

Future Work

1. **Extend Scope:** Consider other tasks
2. **Extend Impact:** Reach other dialects
3. **Build:** Dialect-Aware NLP