

# Multi-VALUE: A Framework for Cross-Dialectal English NLP

<http://value-nlp.org/>

Caleb Ziems\*



@cjziems

William Held\*



@WilliamBarrHeld

Jingfeng Yang



@JingfengY

Jwala Dhamala



@jwaladhamala

Rahul Gupta



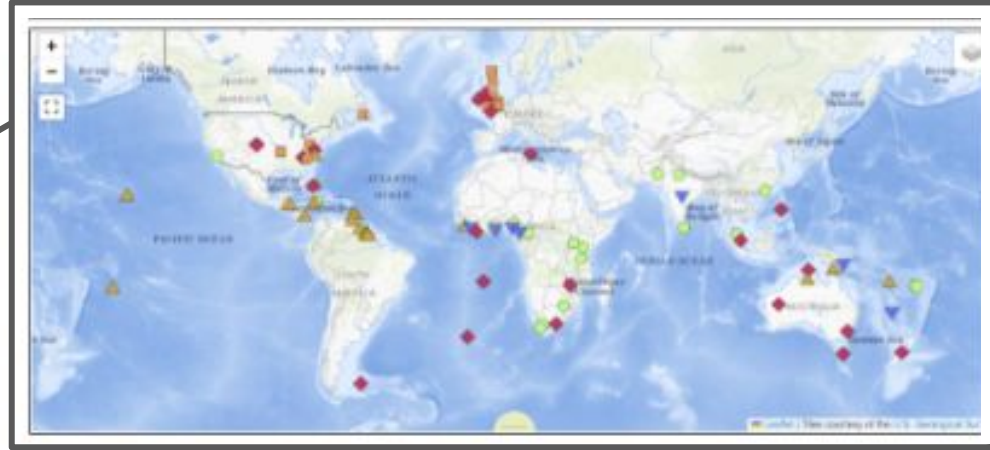
@amazon

Diyi Yang



@Diyi\_Yang

# Linguistic Foundations: *Language Variation*



50 World Englishes Documented

[eWAVE-atlas.org](http://eWAVE-atlas.org)



# Linguistic Foundations:

Aboriginal, Appalachian, Australian, Bahamian, Black South African, Cameroon, Cape Flats, Channel Islands, Chicano, Colloquial American, East Anglican, Falkland Islands, Fiji, Ghanaian, Hong Kong, Indian, Irish, Jamaican, Kenyan, Liberian, Malaysian, Maltese, New Zealand, Newfoundland, Orkney and Shetland, Ozark, Philippine, Pakistani, Scottish, Sri Lankan, St. Helena, Tanzanian, Tristan da Cunha, Urban African American, Ugandan, Welsh, ...

No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	.....

50 World Englishes Documented → 235 Features

eWAVE-atlas.org



No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	.....


# Perturbation Functions:

Aboriginal, Appalachian, Australian, Bahamian, Black South African, Cameroon, Cape Flats, Channel Islands, Chicano, Colloquial American, East Anglican, Falkland Islands, Fiji, Ghanaian, Hong Kong, Indian, Irish, Jamaican, Kenyan, Liberian, Malaysian, Maltese, New Zealand, Newfoundland, Orkney and Shetland, Ozark, Philippine, Pakistani, Scottish, Sri Lankan, St. Helena, Tanzanian, Tristan da Cunha, Urban African American, Ugandan, Welsh, ...



# Multi-VALUE

50 World Englishes Documented → 235 Features

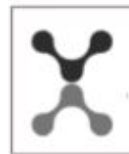
eWAVE-atlas.org											
	<table><tr><th>No.</th><th>Feature Name</th></tr><tr><td>1.</td><td>She/her used for inanimate referents</td></tr><tr><td>2.</td><td>He/him used for inanimate referents</td></tr><tr><td>3.</td><td>Alternative forms for referential (non-dummy) it</td></tr><tr><td>...</td><td>.....</td></tr></table>	No.	Feature Name	1.	She/her used for inanimate referents	2.	He/him used for inanimate referents	3.	Alternative forms for referential (non-dummy) it	...	.....
No.	Feature Name										
1.	She/her used for inanimate referents										
2.	He/him used for inanimate referents										
3.	Alternative forms for referential (non-dummy) it										
...	.....										



Multi-VALUE

# Perturbation Functions:

Aboriginal, Appalachian, Australian, Bahamian, Black South African, Cameroon, Cape Flats, Channel Islands, Chicano, Colloquial American, East Anglican, Falkland Islands, Fiji, Ghanaian, Hong Kong, Indian, Irish, Jamaican, Kenyan, Liberian, Malaysian, Maltese, New Zealand, Newfoundland, Orkney and Shetland, Ozark, Philippine, Pakistani, Scottish, Sri Lankan, St. Helena, Tanzanian, Tristan da Cunha, Urban African American, Ugandan, Welsh, ...



189  
Perturbations



```
def referential_thing(self):  
    # feature 3  
    replace = "the thing"  
    for token in self.tokens:  
        if token.dep_ != "expl" and \  
            token.lower_=="it":  
            self.set_rule( token,  
                           replace )
```

50 World Englishes Documented → 235 Features

eWAVE-atlas.org	
No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	.....



189  
Perturbations



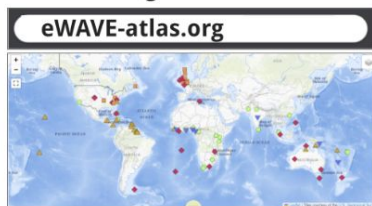
```
def referential_thing(self):  
    # feature 3  
    replace = "the thing"  
    for token in self.tokens:  
        if token.dep_ != "expl" and \  
            token.lower_=="it":  
            self.set_rule( token,  
                           replace )
```

# Speaker Validation:

- 1) Chicano English (29 annotators),
- 2) Colloquial American English (13 annotators),
- 3) Indian English (11 annotators),
- 4) Appalachian English (4 annotators),
- 5) Aboriginal English (4 annotators),
- 6) North of England (3 annotators),
- 7) Ozark English (3 annotators),
- 8) Southeast American Enclave English (3 annotators),
- 9) Urban African American English (1 annotator),
- 10) Black South African English (1 annotator)



50 World Englishes Documented → 235 Features



No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	.....



189  
Perturbations



```
def referential_thing(self):
    # feature 3
    replace = "the thing"
    for token in self.tokens:
        if token.dep_ != "expl" and \
            token.lower == "it":
            self.set_rule(token,
                           replace )
```

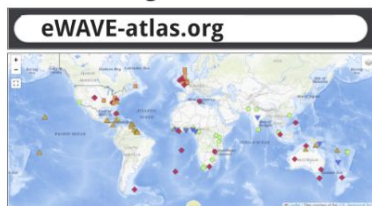


# Gold Standard:

- 1) Chicano English (29 annotators),
- 2) Colloquial American English (13 annotators),
- 3) Indian English (11 annotators),
- 4) Appalachian English (4 annotators),
- 5) Aboriginal English (4 annotators),
- 6) North of England (3 annotators),
- 7) Ozark English (3 annotators),
- 8) Southeast American Enclave English (3 annotators),
- 9) Urban African American English (1 annotator),
- 10) Black South African English (1 annotator)



50 World Englishes Documented → 235 Features



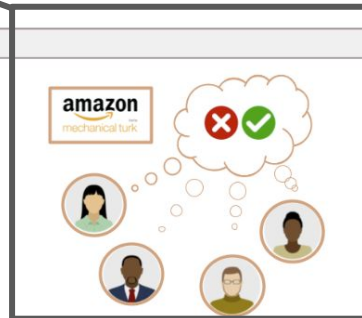
No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	.....



189  
Perturbations



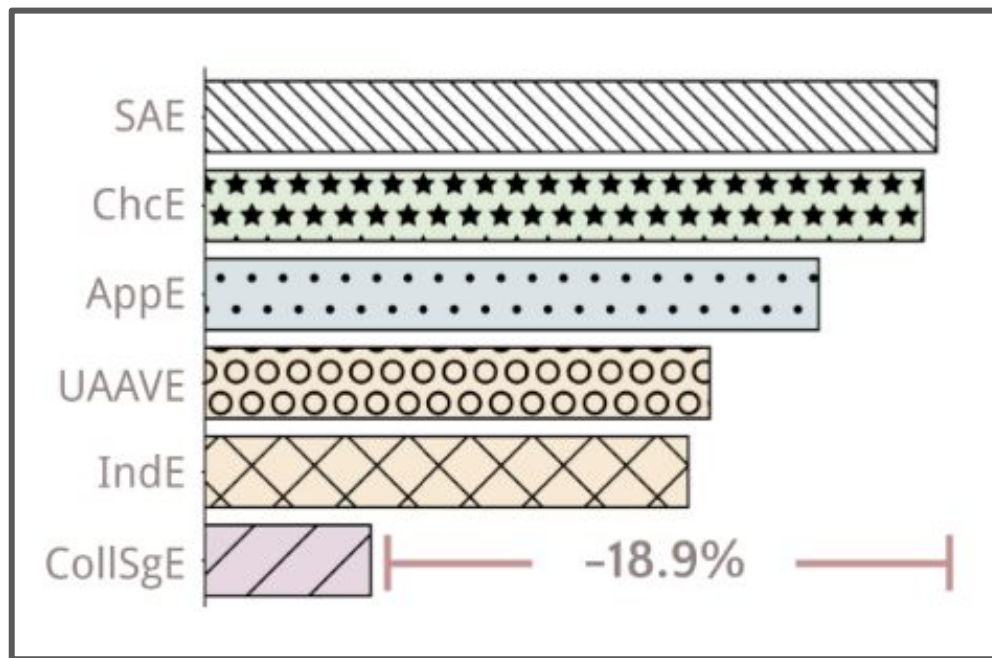
```
def referential_thing(self):  
    # feature 3  
    replace = "the thing"  
    for token in self.tokens:  
        if token.dep_ != "expl" and \  
            token.lower_ == "it":  
            self.set_rule(token,  
                replace )
```





# Stress Testing:

- 1) Chicano English (ChcE)
- 2) Indian English (IndE)
- 3) Appalachian English (AppE)
- 4) Urban African American English (UAAVE)
- 5) Colloquial Singapore English (CollSgE)



50 World Englishes Documented → 235 Features



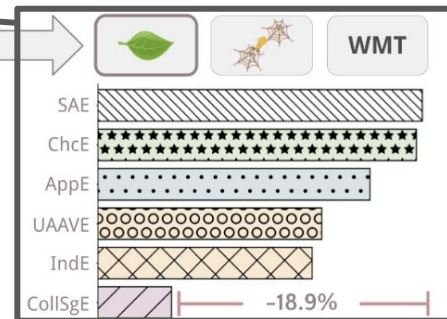
No.	Feature Name
1.	She/her used for inanimate referents
2.	He/him used for inanimate referents
3.	Alternative forms for referential (non-dummy) it
...	.....



189 Perturbations



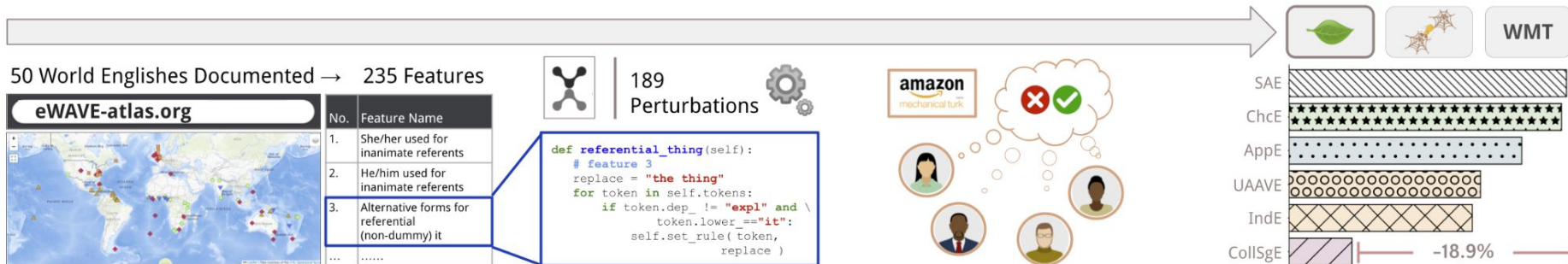
```
def referential_thing(self):
    # feature 3
    replace = "the thing"
    for token in self.tokens:
        if token.dep_ != "expl" and \
           token.lower_ == "it":
            self.set_rule(token,
                           replace)
```





# Contributions:

- 1) rule-based translation system
- 2) gold standard benchmarks
- 3) dialect-robust models



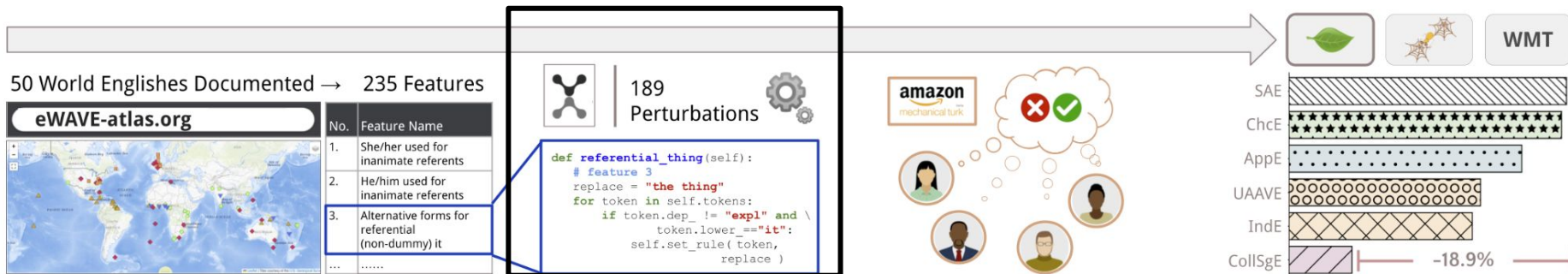
# Contributions:

1) rule-based perturbation system



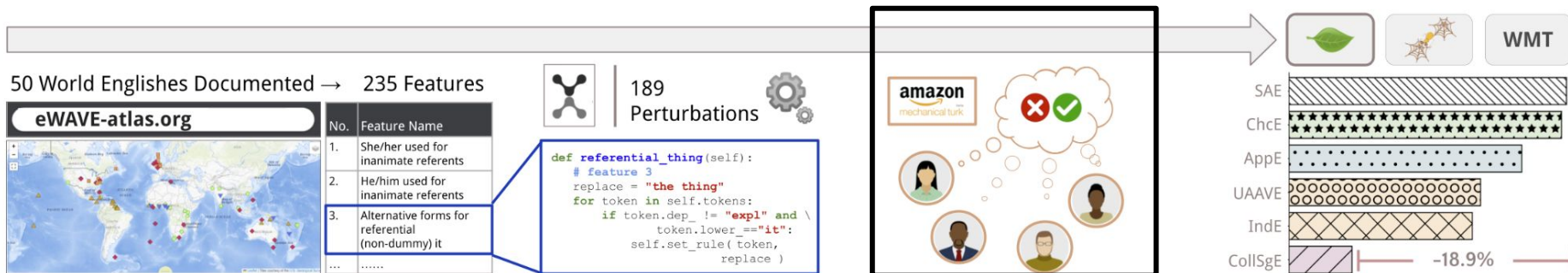
2) gold-standard benchmarks

3) dialect-robust models



# Contributions:

- 1) rule-based perturbation system
- 2) gold-standard benchmark 🤗 Datasets
- 3) dialect-robust models



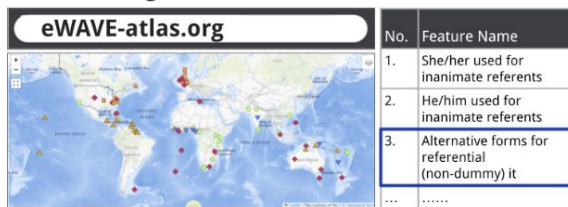
# Contributions:

1) rule-based perturbation system

2) gold-standard benchmarks

3) dialect-robust models 🤗 Transformers

50 World Englishes Documented → 235 Features

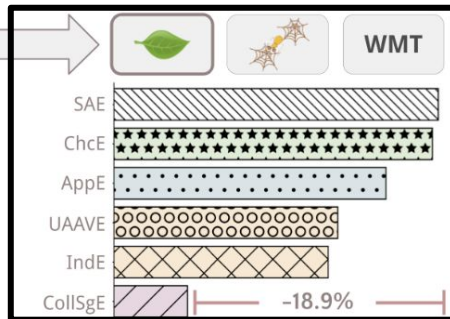


189  
Perturbations



```
def referential_thing(self):  
    # feature 3  
    replace = "the thing"  
    for token in self.tokens:  
        if token.dep_ != "expl" and \  
            token.lower == "it":  
            self.set_rule(token,  
                           replace )
```

amazon  
mechanical turk



# Multi-VALUE Perturbations

# Multi-VALUE Perturbations

John was scolded by his boss

**Feature 153**

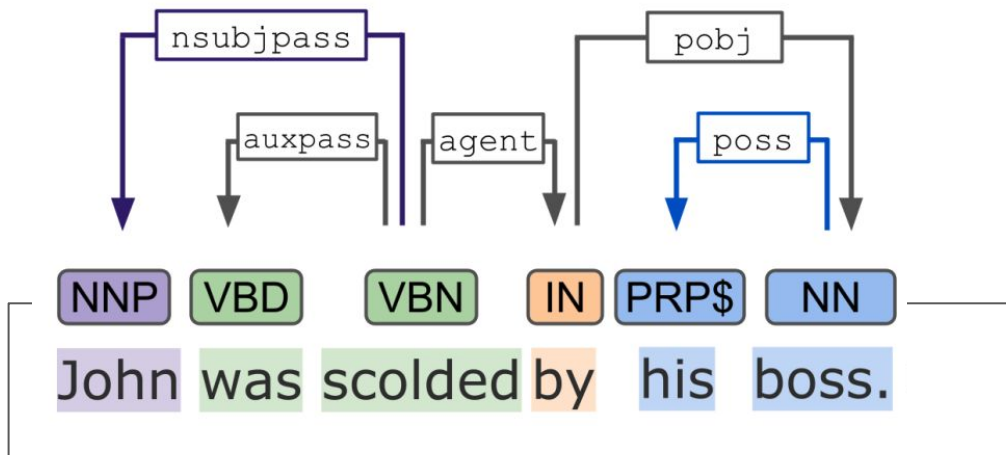


# Multi-VALUE Perturbations

NNP VBD VBN IN PRP\$ NN  
John was scolded by his boss.

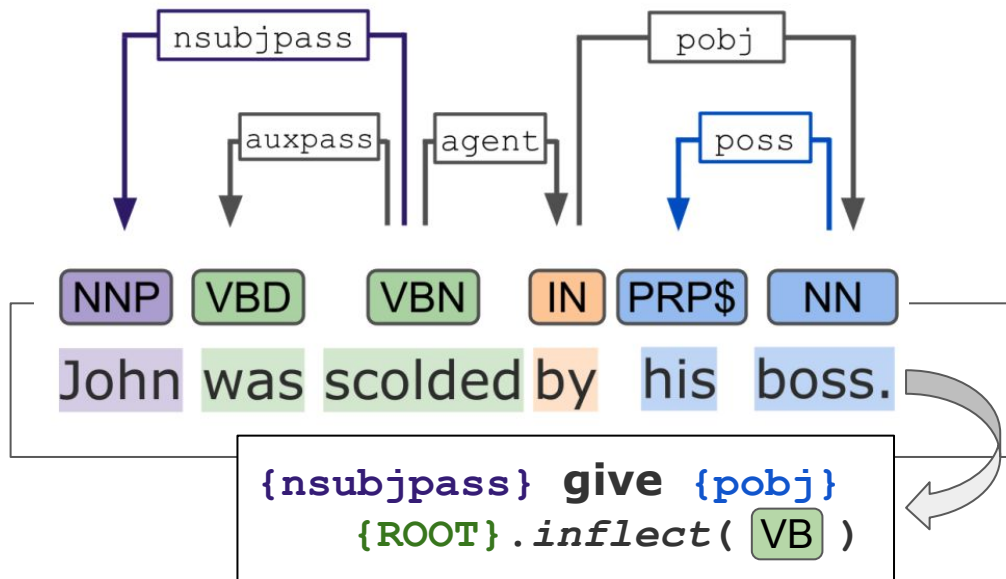
**Feature 153**

# Multi-VALUE Perturbations



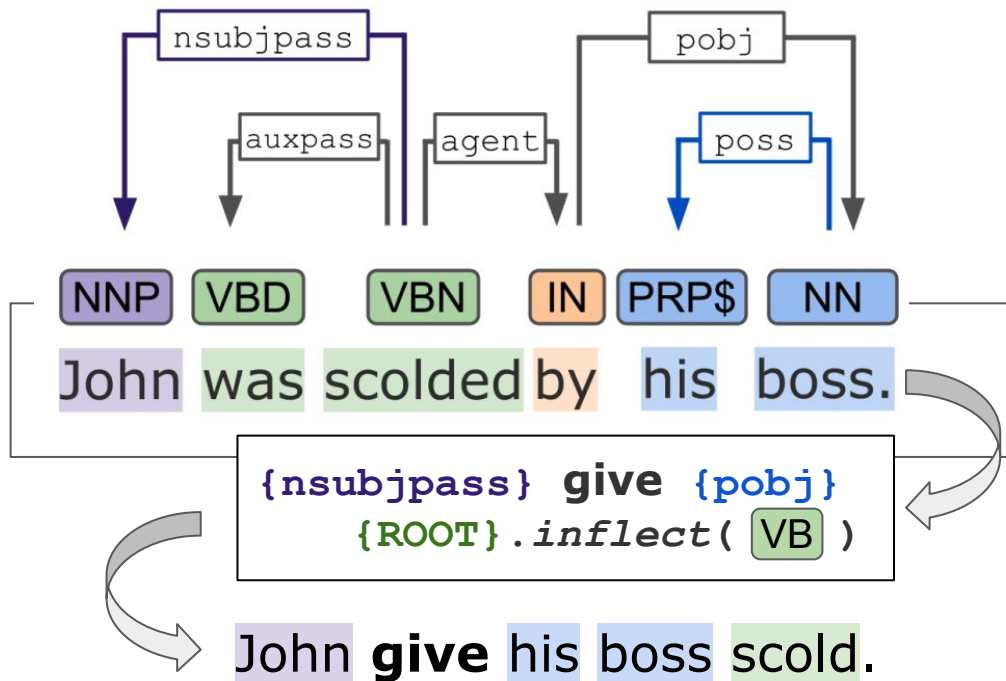
**Feature 153**

# Multi-VALUE Perturbations



**Feature 153**

# Multi-VALUE Perturbations



**Feature 153**

# Multi-VALUE Perturbations

**189 Perturbation Rules!**

Feature 147

Feature 148

Feature 149

Feature 150

**Feature 153**

Feature 154

Feature 155

Feature 156

Feature 157

Feature 158

# Using Multi-VALUE



# Using Multi-VALUE



**Conversational Question Answering: CoQA**



**Semantic Parsing: SPIDER**

**WMT    Machine Translation: WMT-19**

# Using Multi-VALUE

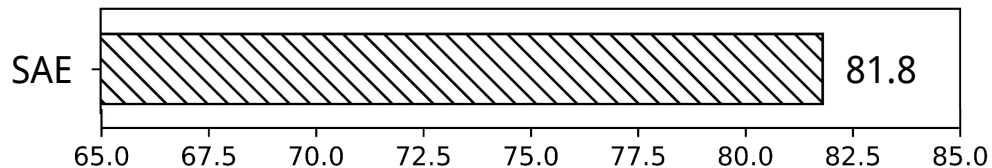


**Conversational Question Answering: CoQA**

# Using Multi-VALUE



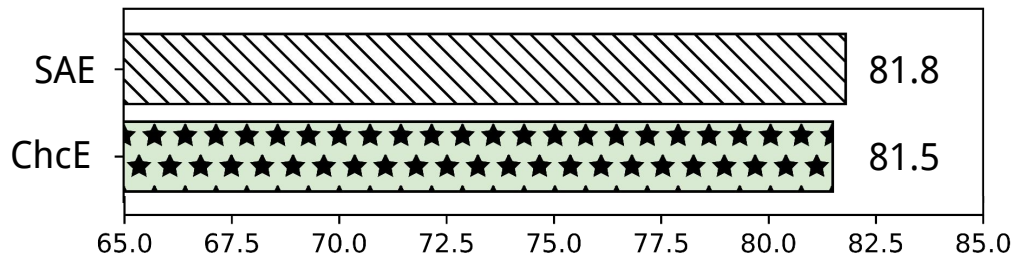
## Conversational Question Answering: CoQA



# Using Multi-VALUE



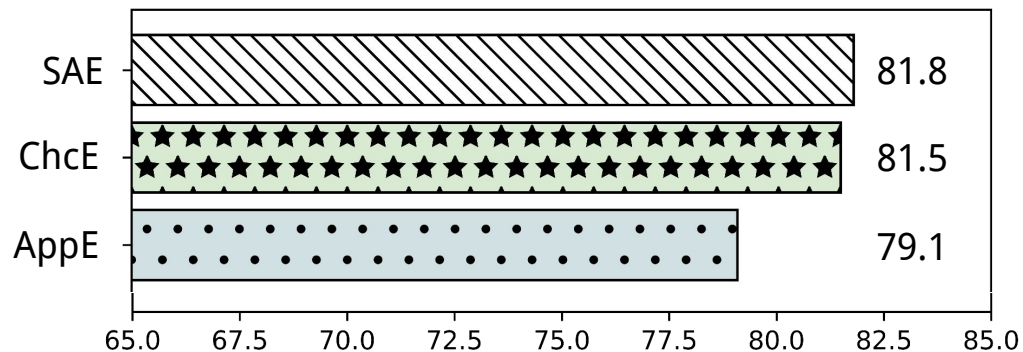
## Conversational Question Answering: CoQA



# Using Multi-VALUE



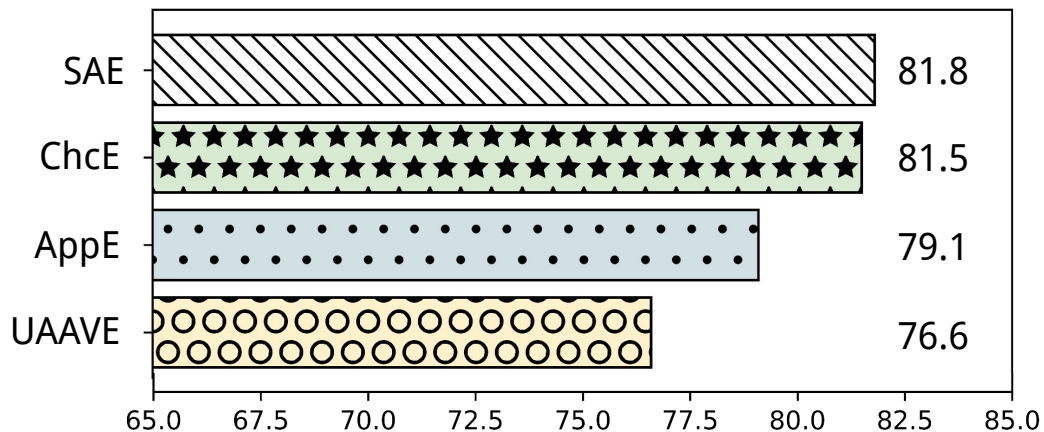
## Conversational Question Answering: CoQA



# Using Multi-VALUE



## Conversational Question Answering: CoQA

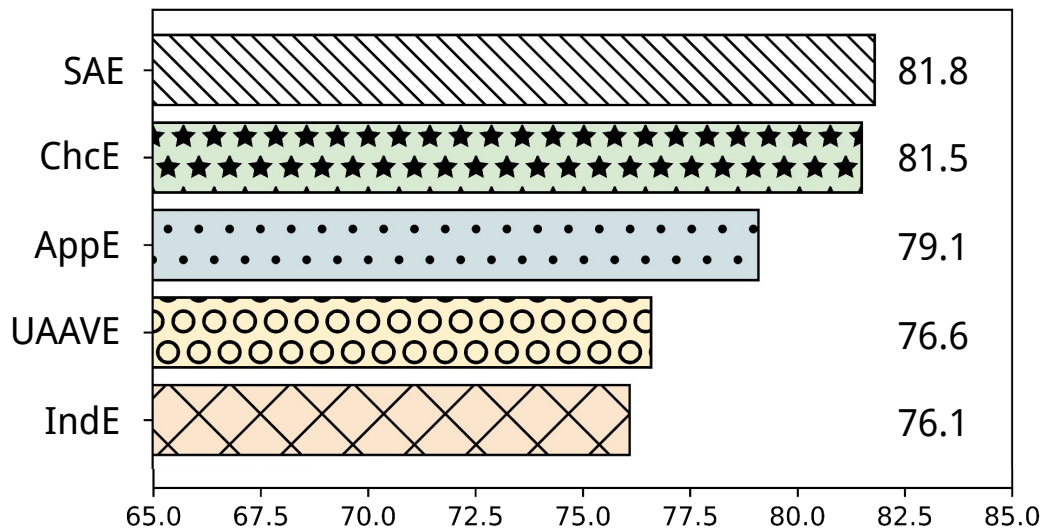




# Using Multi-VALUE



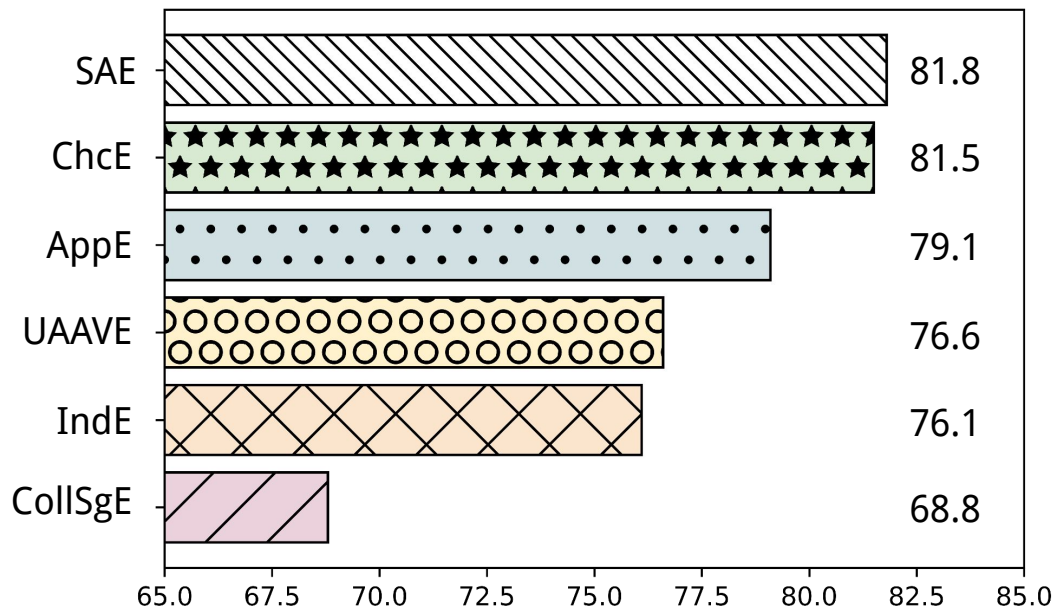
## Conversational Question Answering: CoQA



# Using Multi-VALUE



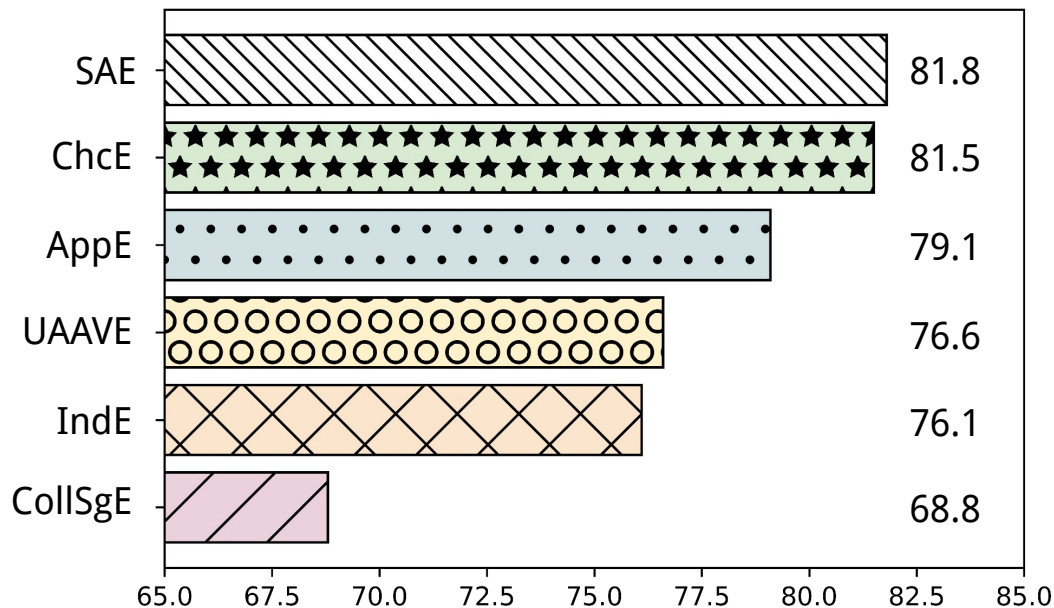
## Conversational Question Answering: CoQA



# Using Multi-VALUE



## Conversational Question Answering: CoQA



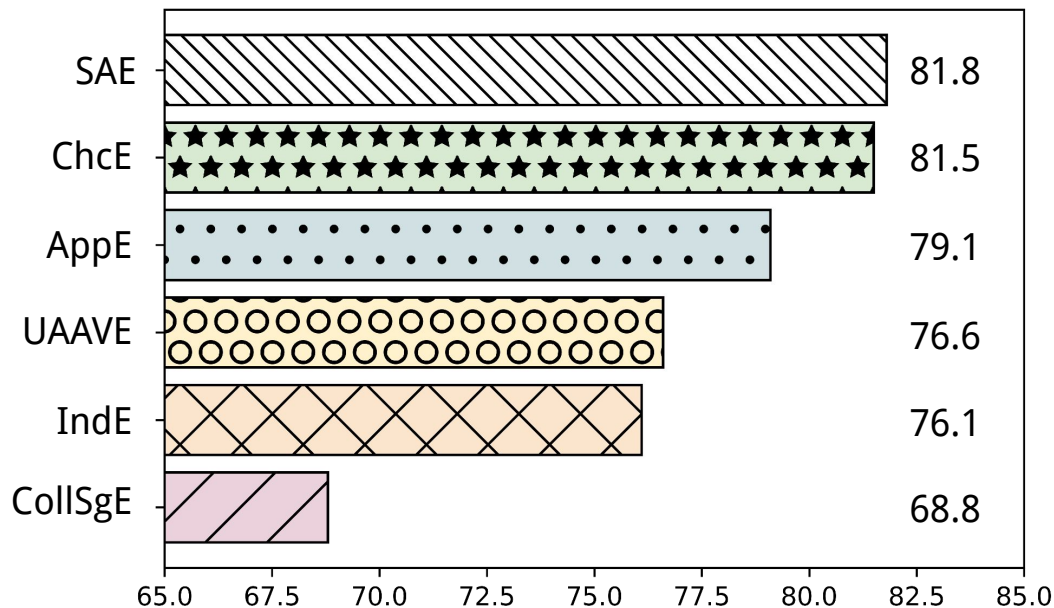
### Qualitative Analysis

- **Cascading Errors**
- Class Errors
- Certain features lead to bigger problems

# Using Multi-VALUE



## Conversational Question Answering: CoQA



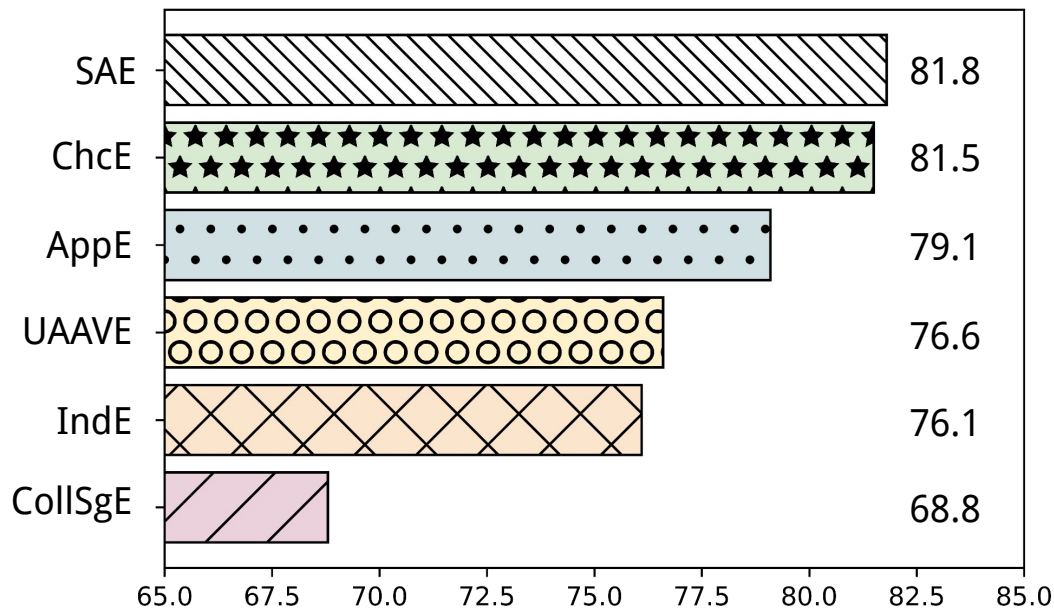
### Qualitative Analysis

- Cascading Errors
- **Class Errors**
- Certain features lead to bigger problems

# Using Multi-VALUE



## Conversational Question Answering: CoQA



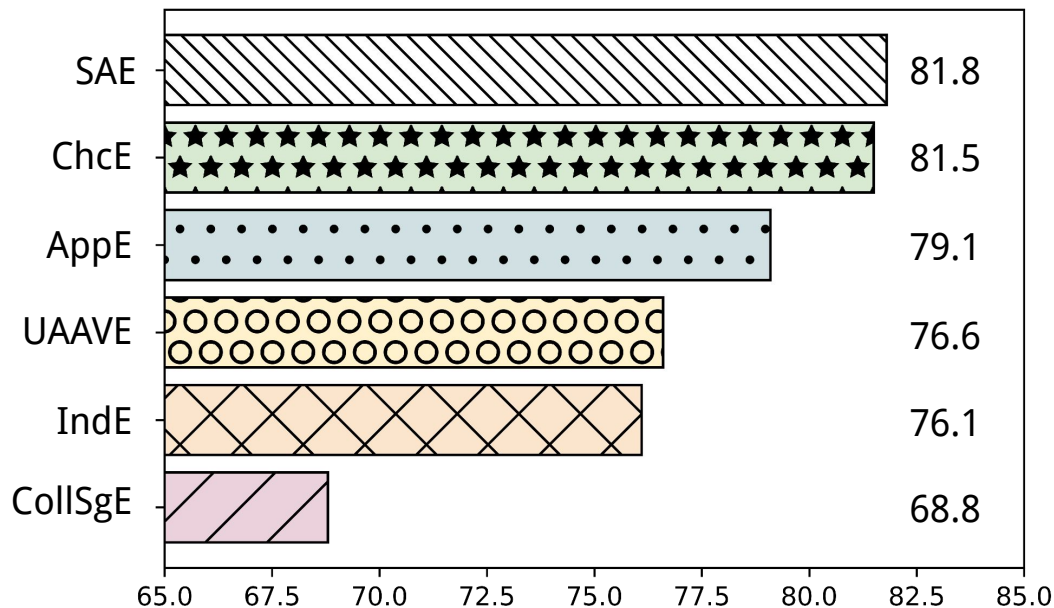
### Qualitative Analysis

- Cascading Errors
- Class Errors
- **Certain features lead to bigger problems**

# Using Multi-VALUE



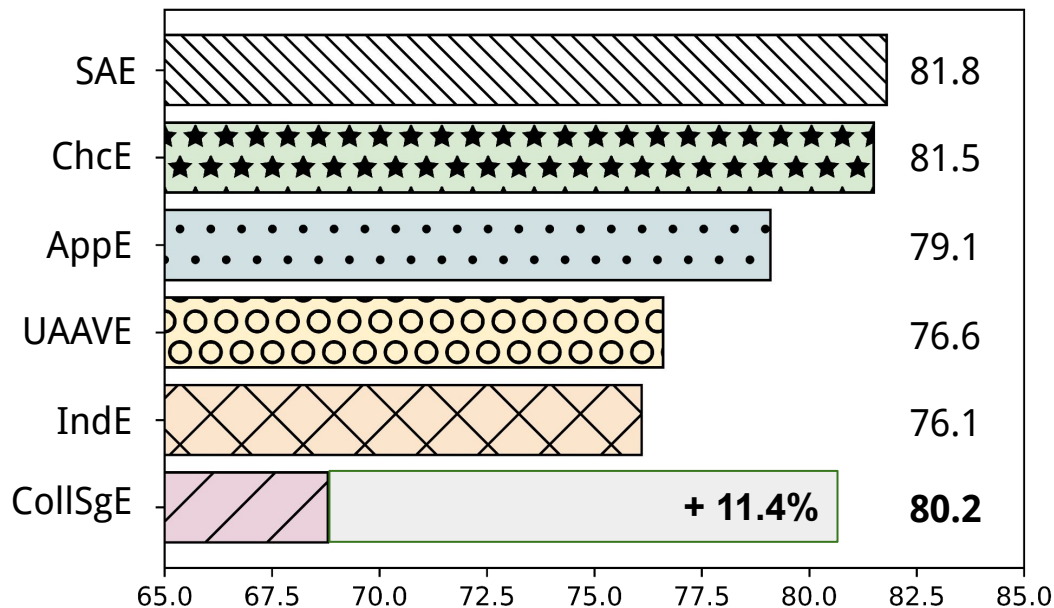
## Conversational Question Answering: CoQA



# Using Multi-VALUE



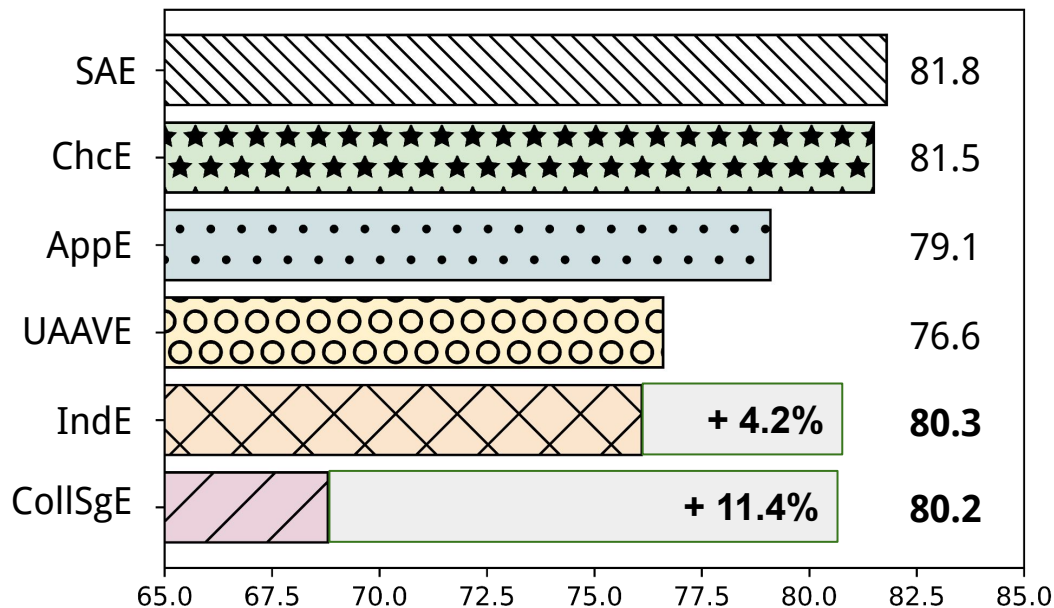
## Conversational Question Answering: CoQA



# Using Multi-VALUE



## Conversational Question Answering: CoQA

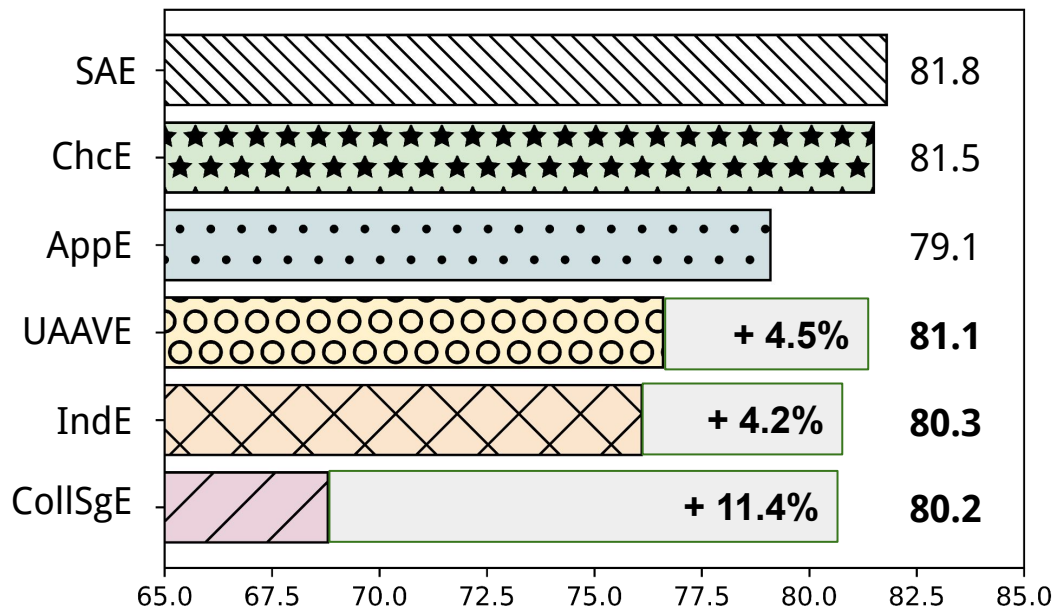




# Using Multi-VALUE



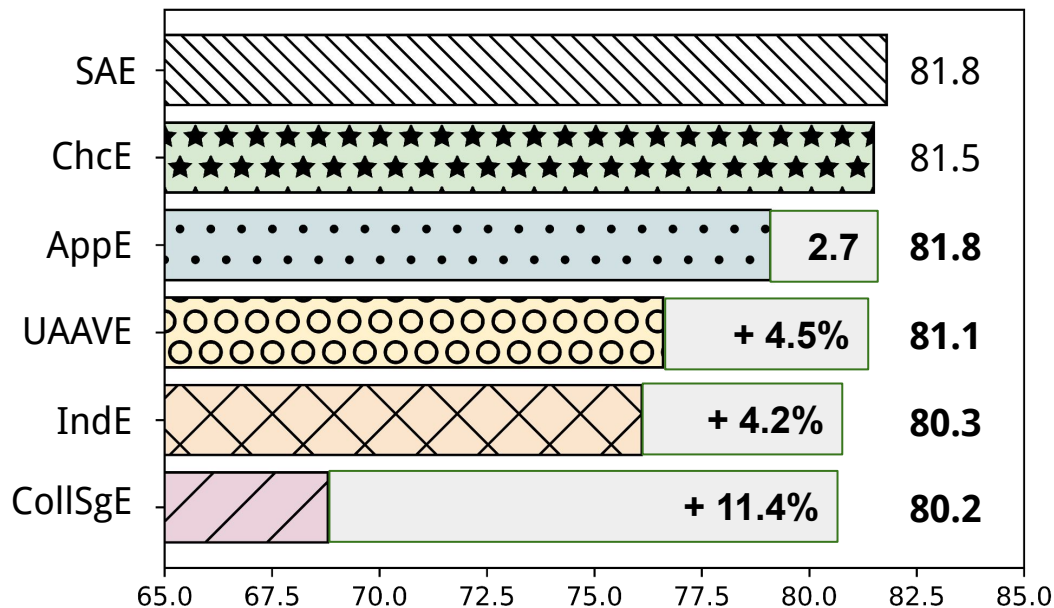
## Conversational Question Answering: CoQA



# Using Multi-VALUE



## Conversational Question Answering: CoQA



# TADA: Task-Agnostic Dialect Adapters for English

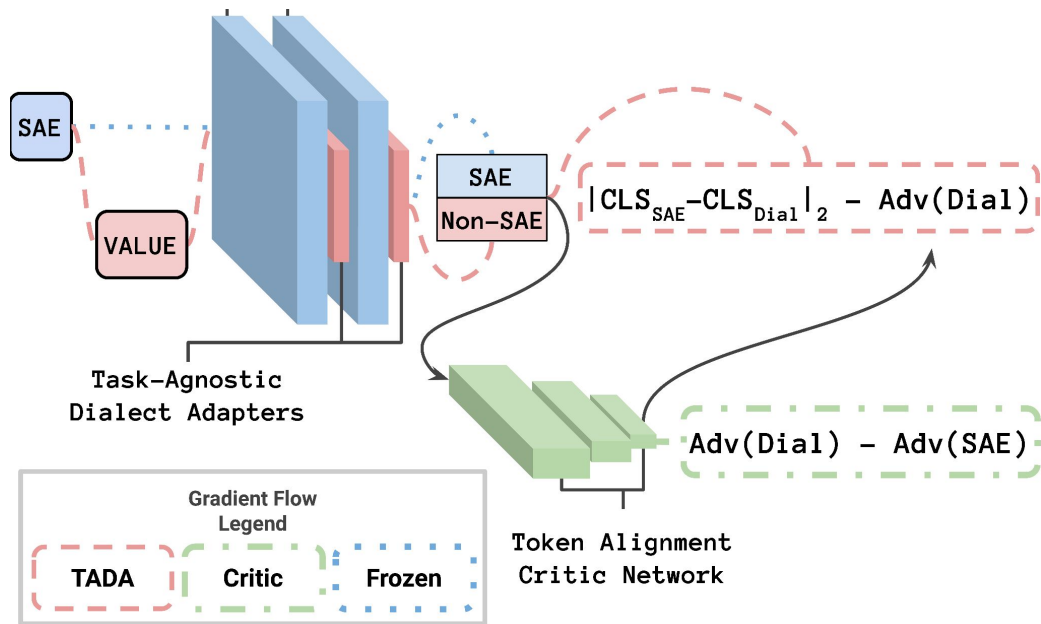
William Held 

Caleb Ziems 

Diyi Yang 



Georgia Institute of Technology,  Stanford University

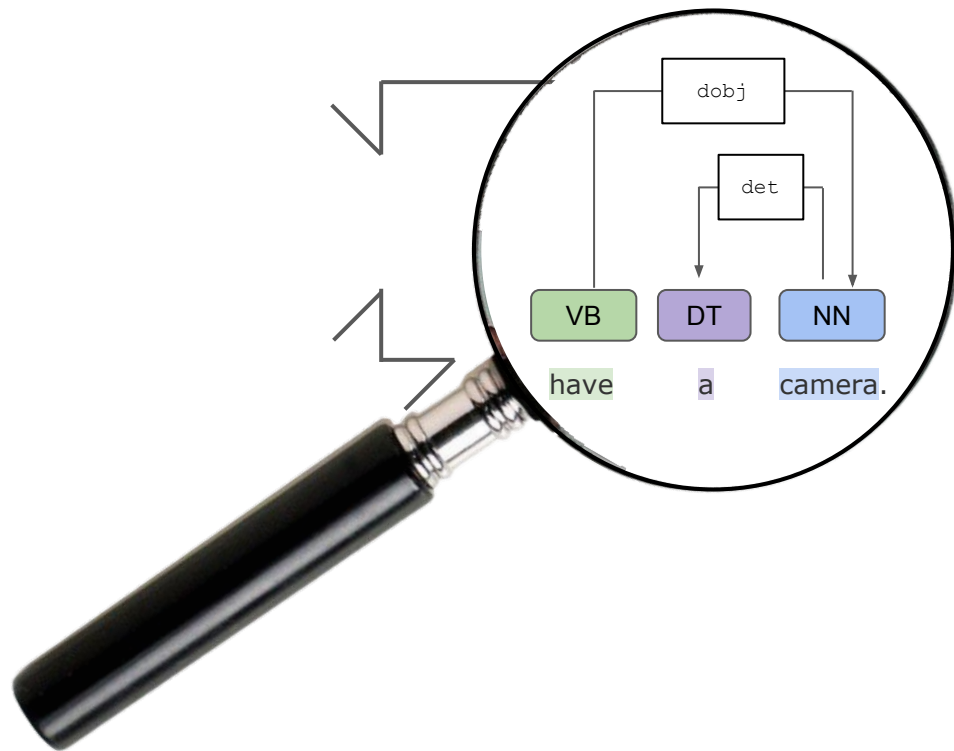


# Why Multi-VALUE

# Multi-VALUE: A Framework for Cross-Dialectal English NLP

## Advantages:

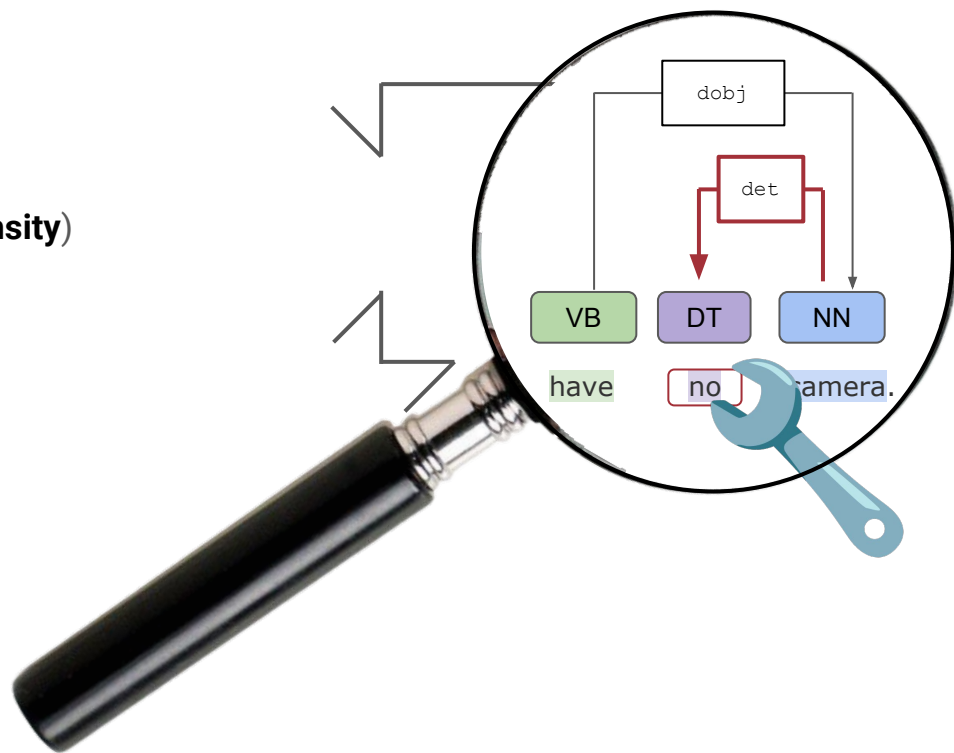
1. **Interpretable** (not **black-box**)



# Multi-VALUE: A Framework for Cross-Dialectal English NLP

## Advantages:

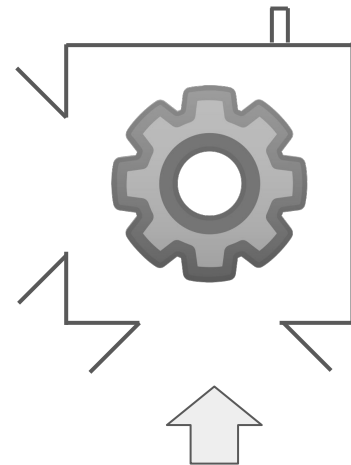
1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)



# Multi-VALUE: A Framework for Cross-Dialectal English NLP

## Advantages:

1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)
3. **Scalable** (**mix + match** datasets)



 **GLUE** **SQuAD**

 **SuperGLUE**

**CoQA**   
A Conversational Question Answering Challenge

**WinoGrande**  


# Multi-VALUE: A Framework for Cross-Dialectal English NLP

## Advantages:

1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)
3. **Scalable** (**mix + match** datasets)
4. **Responsible** (**speaker-validated**)

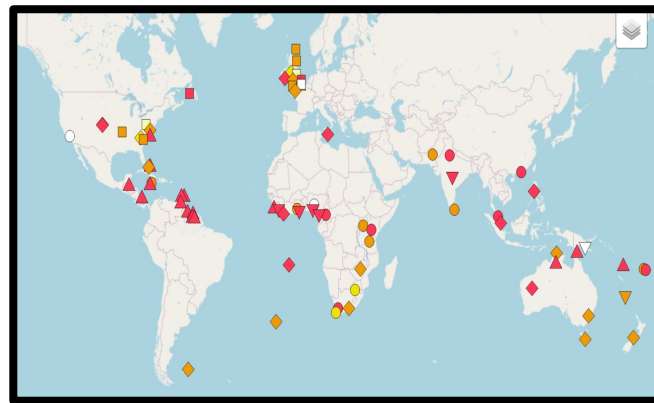




# Multi-VALUE: A Framework for Cross-Dialectal English NLP

## Advantages:

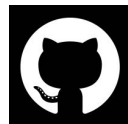
1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)
3. **Scalable** (**mix + match** datasets)
4. **Responsible** (**speaker-validated**)
5. **Generalizable** (truly **cross-dialectal** findings)



# Open Source Hub for Dialect Tools

Drive Adoption & Usage Through:

- Multi-VALUE Transformation package



- Gold Multi-VALUE Benchmarks on



**Datasets**

- Pre-trained Multi-VALUE Models on



**Transformers**

# Multi-VALUE: A Framework for Cross-Dialectal English NLP

<http://value-nlp.org/>

Caleb Ziems



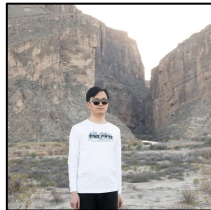
@cjziems

William Held



@WilliamBarrHeld

Jingfeng Yang



@JingfengY

Jwala Dhamala



@jwaladhamala

Rahul Gupta



@amazon

Diyi Yang



@Diyi\_Yang