Defining and Detecting Bias

Bias Definitions

Bias Definitions 1

Measurement modeling is the process of *operationalizing theoretical constructs* and evaluating those operationalizations

Process:

1. Unobservable theoretical construct ($\mathcal{A}$)
2. Operationalization (a)
3. Measurement (â)

[Jacobs and Wallach 2019]

Bias Definitions 1: Measurement modeling

Example:

- $\mathcal{A}$: Socioeconomic Status
- Operationalized by a = i + p
  - i and p are again operationalized theoretical constructs (I: income, p: property)

[Jacobs and Wallach 2019]

Bias Definitions 1: Measurement modeling

Example: Topic models

[Jacobs and Wallach 2019]

Bias Definitions 1: *Measurement modeling*

“[..] many of the fairness-related harms that arise from computational systems emerge from the mismatch between unobservable theoretical constructs and their operationalizations.”

[Jacobs and Wallach 2019]

Bias Definitions 2

1. *Historical ~*

2. *Representation ~*

3. *Measurement ~*

(4. *Aggregation ~*)

(5. *Evaluation ~*)

6. *Deployment ~*

[Suresh and Guttag 2019]

Bias Definitions 3

- "*Stereotypes* are biases that are widely held among a group of people."
- *Gender specific words* (brother, sister, ...)
- *Gender neutral words*
- *Direct Bias* (Compare gender neutral word with pair of Gender specific words)
- *Indirect Bias* (Compare pair of gender neutral words with pair of Gender specific words)

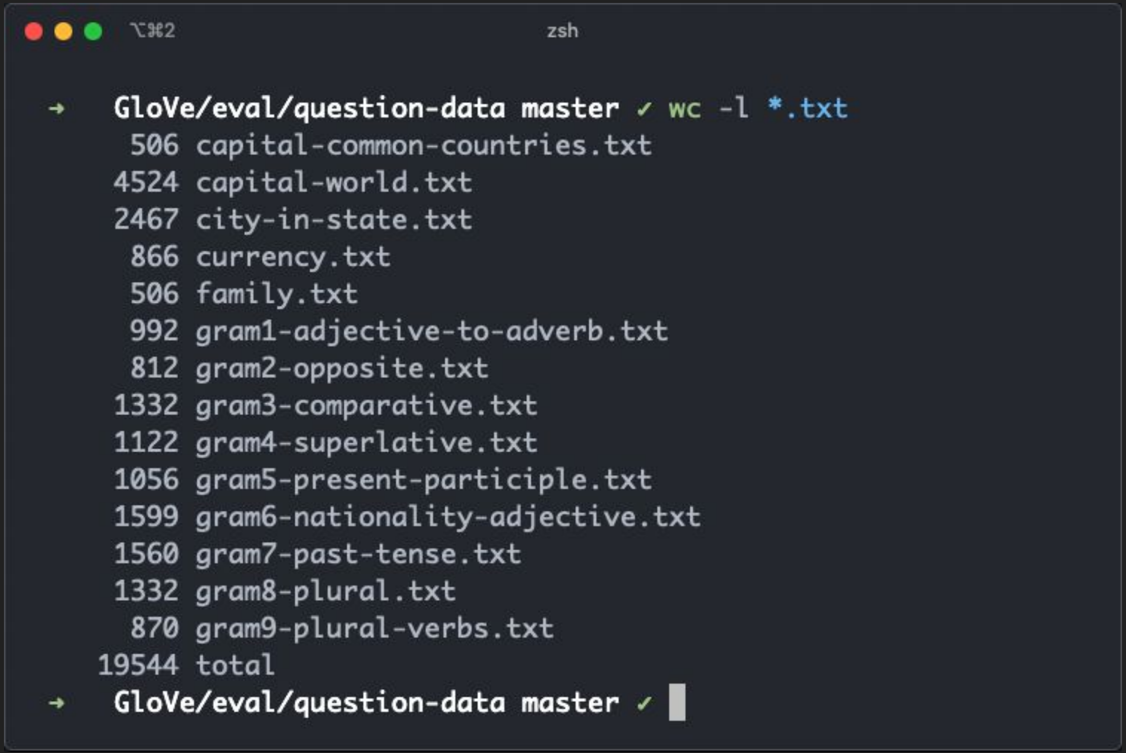
[Bolukbasi et al. 2016]

Word Embeddings

Word Embeddings

[Pennington and Socher and Manning 2014]

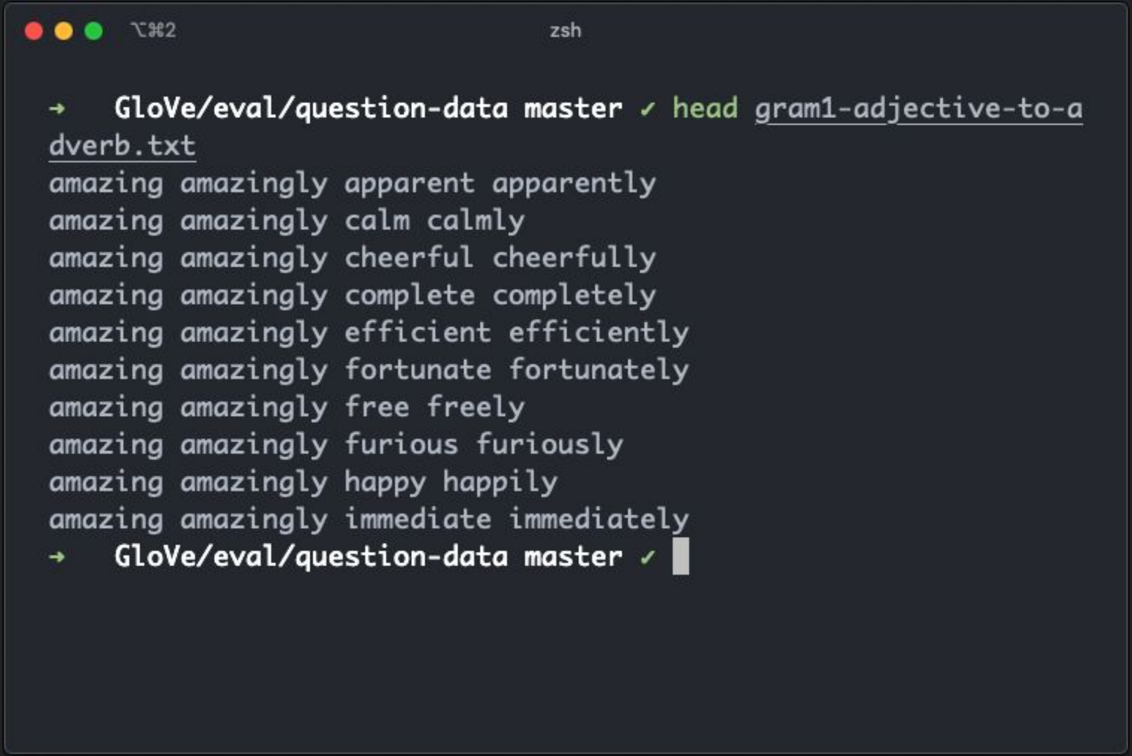
Evaluation of Word Embeddings



```
→  GloVe/eval/question-data master ✓ wc -l *.txt
     506 capital-common-countries.txt
    4524 capital-world.txt
    2467 city-in-state.txt
     866 currency.txt
     506 family.txt
     992 gram1-adjective-to-adverb.txt
     812 gram2-opposite.txt
    1332 gram3-comparative.txt
    1122 gram4-superlative.txt
    1056 gram5-present-participle.txt
    1599 gram6-nationality-adjective.txt
    1560 gram7-past-tense.txt
    1332 gram8-plural.txt
     870 gram9-plural-verbs.txt
   19544 total
→  GloVe/eval/question-data master ✓
```

[Mikolov et al. 2013]

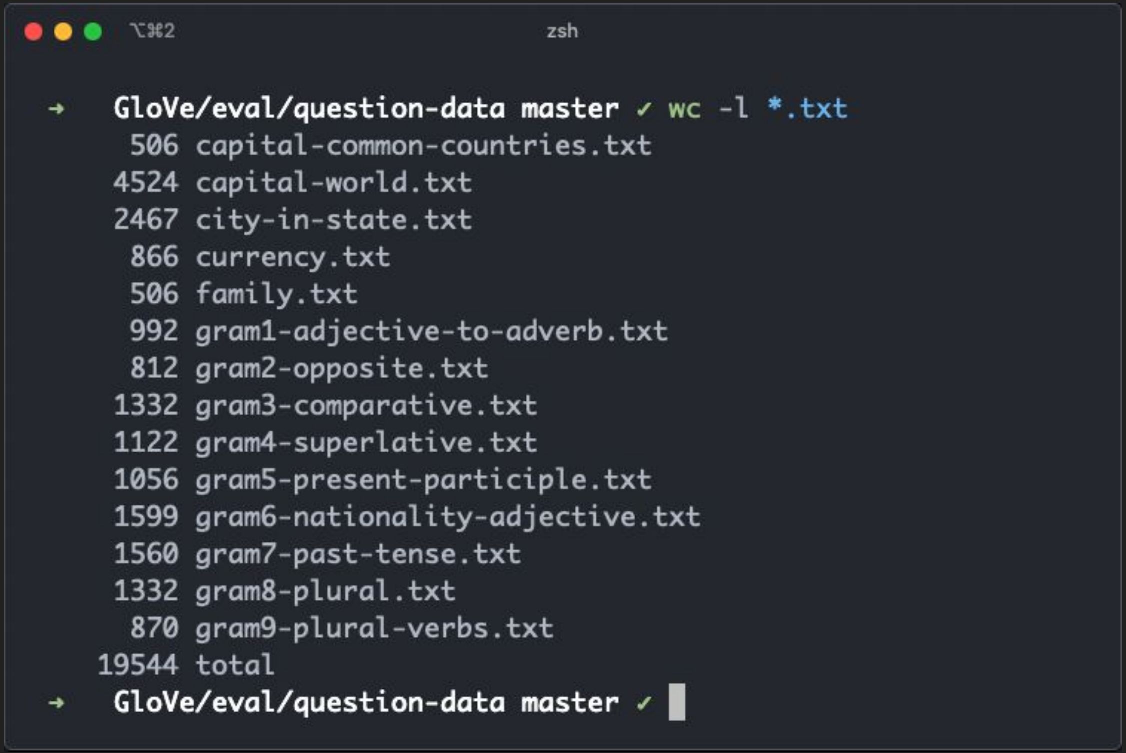
Evaluation of Word Embeddings



```
➜  GloVe/eval/question-data master ✓ head gram1-adjective-to-a
dverb.txt
amazing amazingly apparent apparently
amazing amazingly calm calmly
amazing amazingly cheerful cheerfully
amazing amazingly complete completely
amazing amazingly efficient efficiently
amazing amazingly fortunate fortunately
amazing amazingly free freely
amazing amazingly furious furiously
amazing amazingly happy happily
amazing amazingly immediate immediately
➜  GloVe/eval/question-data master ✓
```

[Mikolov et al. 2013]

Evaluation of Word Embeddings



```
→  GloVe/eval/question-data master ✓ wc -l *.txt
     506 capital-common-countries.txt
    4524 capital-world.txt
    2467 city-in-state.txt
     866 currency.txt
     506 family.txt
     992 gram1-adjective-to-adverb.txt
     812 gram2-opposite.txt
    1332 gram3-comparative.txt
    1122 gram4-superlative.txt
    1056 gram5-present-participle.txt
    1599 gram6-nationality-adjective.txt
    1560 gram7-past-tense.txt
    1332 gram8-plural.txt
     870 gram9-plural-verbs.txt
   19544 total
→  GloVe/eval/question-data master ✓ 
```

[Mikolov et al. 2013]

Bias in Word Embeddings

Gender Subspace

- Create a list of gender specific words
- All other words are seen as gender neutral words
- The *gender subspace* is the difference between male and female gender specific words
- For direct bias, compare a query word with the gender subspace
- For indirect bias, compare the distance between a pair of words with the distance of the words without the gender subspace

[Bolukbasi et al. 2016]

**Reporting Bias*

[Bolukbasi et al. 2016]

Downstream Tasks

Word Embeddings can be seen as an intermediate measurement

- Test individual samples and identify feature contribution
- Test for Representation Bias with latent feature

[Sap et al. 2019]

Critique