Bias in Word Embeddings

Caleb Kaiji Lu
Class 18739 Spring 2019
Agenda

• Recap on word embeddings
• Bias in word embeddings
  • 3 metrics for quantifying embedding stereotypes [Bolukabasi et. al, 2016]
• Debiasing algorithms [Bolukabasi et. al, 2016]
• Embedding as a lens to study history [Garg et al, 2018]
Previously: Word Embedding is a Dictionary

Training data: corpus of text

Context window

The dog is chasing a cat.

\[
\text{find } v\text{'s to } \max \log P(\text{chasing} | \text{dog}) + \log P(\text{cat} | \text{dog})
\]

where \( P(\text{cat} | \text{dog}) \propto \exp(v_{\text{cat}} \cdot v_{\text{dog}}) \)
Previously: Word Embedding

- Word embedding captures relationships among words
  - Semantic relationship: \textit{woman:man::queen:king}
  - Syntactic relationship: \textit{they: their:: he:his}
  - More complicated knowledge-base like relationship:
    - \textit{Beijing:China :: Paris: France}
  - Standard metric to evaluate a word embedding
Word Embeddings Also Capture Bias  [Bolukabasi, 16]

• Man: King :: Woman:Queen
• Paris: France :: Tokyo:Japan

• He:Brother :: She:
• He:Blue :: She
• He:Doctor :: She:
• He:Realist :: She:
• She:Pregnancy :: He:
• She:Baking::He:
• She:Blonde::He:
• He:Computer :: She:
Word Embeddings Also Capture Bias [Bolukabasi, 16]

• He: Computer Programmer :: She: Homemaker
  • Equivalent to having a biased dictionary:

  nurse (ˈnərs)
  1. A woman trained to care for the sick or infirm, especially in a hospital.

  computer programmer (kəmˈpjuːtəˈprəʊɡræmə)
  1. A man who writes programs for the operation of computers, especially as an occupation.
Bias in Downstream Applications: Machine Translation

He
She

ő
ő

ő egy orvos
ő egy nővér

He is a doctor
She is a nurse
Metrics to Quantify Gender bias in WE

• Metric 1: Occupations
  • 327 gender neutral occupations. Project on to she—he direction

Crowdworkers rate each occup. for gender stereotype

\( \text{Corr}(\text{projection}_{\text{she-he}}, \text{crowd rating}) = 0.51 \)
Consistency of embedding stereotype

GloVe trained on web crawl

word2vec trained on Google news

Each dot is an occupation; Spearman = 0.8
Metrics to Quantify Gender bias in WE

• Metric 2: Analogies
  • Automatically generate $he : x :: she : y$ analogies.

$$\min \cos(he - she, x - y) \text{ such that } ||x - y||_2 < \delta$$
Metrics to Quantify Gender bias in WE

• Metric 2: Analogies
  • Automatically generate $he : x :: she : y$ analogies.

$$\min \cos (he - she, x - y) \text{ such that } \|x - y\|_2 < \delta$$
Metrics to Quantify Gender bias in WE

• Metric 3: Indirect Bias
  • Gender stereotype could affect the geometry between words that should be gender-neutral.
  • Project occupations onto softball—football axis.
The Geometry of Gender

[Diagram showing relationships between gender terms: male, he, female, she, his, her, Mary, John, man, woman]
The Geometry of Gender

Gender (he-she) Axis
First Principal Component
Principal Component Analysis

- Principal Components (PC) are orthogonal directions that capture most of the variance in the data.
  - 1\textsuperscript{st} PC – direction of greatest variability in data
  - 2\textsuperscript{nd} PC – Next orthogonal (uncorrelated) direction of greatest variability (remove all variability in first direction, then find next direction of greatest variability)
- And so on...
Geometry of Gender

The top PC seems to capture the gender subspace $B$. 

% of variance explained
Debiasing Algorithm (Hard-debiased)

- Identify words that are gender neutral N and gender-definitional S
- Project away the gender subspace from the gender-neutral words
  - $w := w - w.B$ B is the gender subspace
- Normalize vectors
Identify gender-definitional words

218 gender-definitional words

Linear SVM
Projecting away gender component
Projecting away gender component
Projecting away gender component

“hard debiasing”

299 dimensions
Advanced debiasing (soft debiasing)

• Find a linear transformation $T$ of the gender-neutral words to reduce the gender component while not moving the words too much.

$$ W = \text{matrix of all word vectors.} $$

$$ N = \text{matrix of neutral word vectors.} $$

$$ \min_T \| (TW)^T(TW) - W^TW \|_F^2 + \lambda \| (TN)^T(TB) \|_F^2 $$

- don’t move too much
- minimize gender component
Debiasing results: indirect bias
Debiasing results: indirect bias

Original embedding

softball  pitcher  receptionist  maestro  football

Debiased embedding

softball  pitcher  major leaguer  midfielder  football
Debiasing result analogies

# stereotypic analogies

# appropriate analogies

# analogies generated

# analogies generated
Debiasing result: Appropriate Analogies

- He:King :: She:Queen
- He:Doctor::She:Doctor

<table>
<thead>
<tr>
<th></th>
<th>RG</th>
<th>WS</th>
<th>analogy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>62.3</td>
<td>54.5</td>
<td>57.0</td>
</tr>
<tr>
<td>Hard-debiased</td>
<td>62.4</td>
<td>54.1</td>
<td>57.0</td>
</tr>
</tbody>
</table>
Natural Questions

- Does mitigating bias in word embeddings also mitigate bias in the downstream tasks?
- Does mitigating bias in word embeddings impact the performance of the downstream tasks?
- To be answered in a later lecture
Summary

• Geometry of word embedding captures bias
  • Who’s responsible: data, algorithm or user?
• Can effective debias algorithms for sensitive applications
Thanks!

• References
  • *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings.* NIPS’16