Understanding Undesirable Word Embedding Associations

Kawin Ethayarajh, David Duvenaud, Graeme Hirst

University of Toronto

ACL 2019
Background

Do word embedding associations capture social biases?

- Caliskan et al. (2017): According to WEAT, science terms more associated with male attributes; art terms with female ones.
- Bolukbasi et al. (2016): To debias vectors, define a “bias subspace” and subtract from each vector its projection on the subspace.

See also: bias in translation, tagging, etc.
Questions

Undesirable word associations remain poorly understood.

1. Does the subspace projection method provably debias embeddings?
2. Why should WEAT be used to measure word associations?
3. What’s to blame? Training data, the embedding model, or just noise?
Undesirable word associations remain poorly understood.

1. Does the subspace projection method provably debias embeddings?
2. Why should WEAT be used to measure word associations?
3. What’s to blame? Training data, the embedding model, or just noise?
How to define unbiasedness?

- Let $M$ be the word-context matrix the embedding model implicitly factorizes: $WC^T = M$
- Word $w$ is unbiased in $M$ wrt word pairs $S$ iff

$$\forall (x, y) \in S, M_{w, x} = M_{w, y}$$

E.g., ‘doctor’ unbiased wrt {('king', 'queen')} iff

$$M_{\text{doctor, king}} = M_{\text{doctor, queen}}$$
Debiasing via Subspace Projection

**Debiasing Theorem** If bias subspace $B = \text{span}(\{\vec{x} - \vec{y} | (x, y) \in S\})$ for word pairs $S$, then debiased word vectors $\{w_d\}$ are unbiased wrt $S$.

- Can swap $(w, x)$ and $(w, y)$ in reconstructed matrix $W_d C^T = M_d$
- Equivalent to training on a corpus unbiased wrt $S$. 
Gonen and Goldberg (2019): In practice, it is possible to detect gender even after debiasing via subspace projection.

- Why?

  - Debiasing won’t remove all vestiges of gender if either $S$ is non-exhaustive or $B \neq \text{span}(\{\vec{x} - \vec{y} | (x, y) \in S\})$. 

In the diagram:
- policeman to policewoman
- man to woman
- $\vec{x} - \vec{y}$
Questions

Undesirable word associations remain poorly understood.

1. Does the subspace projection method provably debias embeddings?
2. Why should WEAT be used to measure word associations?
3. What’s to blame? Training data, the embedding model, or just noise?
**Word Embedding Association Test:**
Where relatedness is cosine similarity, are words $T_1$ more associated with attributes $X$ than $Y$, relative to $T_2$?

- “flowers” more pleasant than unpleasant, relative to “insects”
- “science” more male then female, relative to “arts”
**Word Embedding Association Test:**
Where relatedness is cosine similarity, are words $T_1$ more associated with attributes $X$ than $Y$, relative to $T_2$?

- “flowers” more pleasant than unpleasant, relative to “insects”
- “science” more male then female, relative to “arts”

User determines composition of these word sets!
You can cherry-pick the attributes to achieve your desired outcome.

<table>
<thead>
<tr>
<th>Target Word Sets</th>
<th>Attribute Word Sets</th>
<th>Test Stat</th>
<th>p-val</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>{door} vs. {curtain}</td>
<td>{masculine} vs. {feminine}</td>
<td>0.021</td>
<td>0.0</td>
<td>male-assoc.</td>
</tr>
<tr>
<td></td>
<td>{girlish} vs. {boyish}</td>
<td>-0.042</td>
<td>0.5</td>
<td>inconclusive</td>
</tr>
<tr>
<td></td>
<td>{woman} vs. {man}</td>
<td>0.071</td>
<td>0.0</td>
<td>female-assoc.</td>
</tr>
<tr>
<td>{dog} vs. {cat}</td>
<td>{masculine} vs. {feminine}</td>
<td>0.063</td>
<td>0.0</td>
<td>male-assoc.</td>
</tr>
<tr>
<td></td>
<td>{actress} vs. {actor}</td>
<td>-0.075</td>
<td>0.5</td>
<td>inconclusive</td>
</tr>
<tr>
<td></td>
<td>{womanly} vs. {manly}</td>
<td>0.001</td>
<td>0.0</td>
<td>female-assoc.</td>
</tr>
</tbody>
</table>
You can cherry-pick the attributes to achieve your desired outcome.

<table>
<thead>
<tr>
<th>Target Word Sets</th>
<th>Attribute Word Sets</th>
<th>Test Stat</th>
<th>p-val</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>{door} vs. {curtain}</td>
<td>{masculine} vs. {feminine}</td>
<td>0.021</td>
<td>0.0</td>
<td>male-assoc.</td>
</tr>
<tr>
<td></td>
<td>{girlish} vs. {boyish}</td>
<td>−0.042</td>
<td>0.5</td>
<td>inconclusive</td>
</tr>
<tr>
<td></td>
<td>{woman} vs. {man}</td>
<td>0.071</td>
<td>0.0</td>
<td>female-assoc.</td>
</tr>
<tr>
<td>{dog} vs. {cat}</td>
<td>{masculine} vs. {feminine}</td>
<td>0.063</td>
<td>0.0</td>
<td>male-assoc.</td>
</tr>
<tr>
<td></td>
<td>{actress} vs. {actor}</td>
<td>−0.075</td>
<td>0.5</td>
<td>inconclusive</td>
</tr>
<tr>
<td></td>
<td>{womanly} vs. {manly}</td>
<td>0.001</td>
<td>0.0</td>
<td>female-assoc.</td>
</tr>
</tbody>
</table>
The relational inner product association (RIPA) of a word \( w \) wrt relation vector \( \vec{b} \):

\[
\beta(\vec{w}; \vec{b}) = \langle \vec{w}, \vec{b} \rangle
\]

where

- word pairs \( S \) define the association (e.g., ('king', 'queen'))
- \( \vec{b} = \text{principal component}(\{\vec{x} - \vec{y} | (x, y) \in S\}) \)
Advantages of RIPA:

- interpretable when embedding model factorizes word-context matrix
- robust to how $\vec{b}$ is defined
- derived from the subspace projection method of debiasing
For noiseless SGNS, where $S = \{(x, y)\}$:

$$\beta_{\text{SGNS}}(\vec{w}; \vec{b}) = \frac{1/\sqrt{\lambda}}{\sqrt{-\text{csPMI}(x, y)} + \alpha \log \frac{p(w|x)}{p(w|y)}}$$

- $\beta(\vec{w}; \vec{b}) \to 0$ the more unrelated $x$ and $y$ are
- $\beta(\vec{w}; \vec{b}) \in [-\|\vec{w}\|, \|\vec{w}\|]$
- if $x_1 - y_1 = x_2 - y_2$, then $\beta(\vec{w}; \vec{b})$ is unchanged
Questions

Undesirable word associations remain poorly understood.

1. Does the subspace projection method provably debias embeddings?
2. Why should WEAT be used to measure word associations?
3. What’s to blame? Training data, the embedding model, or just noise?
Breaking Down Gender Association

- $g$: RIPA (i.e., genderedness in embedding space)

$$
g(w; x, y) = \frac{\langle \vec{w}, \vec{x} - \vec{y} \rangle}{\| \vec{x} - \vec{y} \|}
$$
Breaking Down Gender Association

- $g$: RIPA (i.e., genderedness in embedding space)

\[
g(w; x, y) = \frac{\langle \vec{w}, \vec{x} - \vec{y} \rangle}{\| \vec{x} - \vec{y} \|}
\]

- $\hat{g}$: RIPA for noiseless SGNS (i.e., genderedness in corpus)

\[
\hat{g}(w; x, y) = \frac{1/\sqrt{\lambda}}{\sqrt{-csPMI(x, y) + \alpha}} \log \frac{p(w|x)}{p(w|y)}
\]
Breaking Down Gender Association

- $g$: RIPA (i.e., genderedness in embedding space)
  \[
g(w; x, y) = \frac{\langle \vec{w}, \vec{x} - \vec{y} \rangle}{\|\vec{x} - \vec{y}\|}
\]

- $\hat{g}$: RIPA for noiseless SGNS (i.e., genderedness in corpus)
  \[
  \hat{g}(w; x, y) = \frac{1}{\sqrt{\lambda}} \frac{1}{\sqrt{-\text{csPMI}(x, y) + \alpha}} \log \frac{p(w|x)}{p(w|y)}
  \]

- $\Delta g$: change from corpus $\rightarrow$ embedding space
  \[
  \Delta g(w; S) = \left| \sum_{(x, y) \in S} \frac{g(w; x, y)}{|S|} \right| - \left| \sum_{(x, y) \in S} \frac{\hat{g}(w; x, y)}{|S|} \right|
  \]
## Breaking down Gender Association

<table>
<thead>
<tr>
<th>Word Type</th>
<th>Word</th>
<th>Corpus Genderedness</th>
<th>SGNS Genderedness</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender-Appropriate</td>
<td>mom</td>
<td>−0.163</td>
<td>−0.648</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>king</td>
<td>0.058</td>
<td>0.200</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td><strong>Avg (abs.)</strong></td>
<td><strong>0.231</strong></td>
<td><strong>0.522</strong></td>
<td><strong>0.291</strong></td>
</tr>
<tr>
<td>(n = 164)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender-Biased</td>
<td>nurse</td>
<td>−0.190</td>
<td>−1.047</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>architect</td>
<td>−0.063</td>
<td>0.162</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td><strong>Avg (abs.)</strong></td>
<td><strong>0.253</strong></td>
<td><strong>0.450</strong></td>
<td><strong>0.197</strong></td>
</tr>
<tr>
<td>(n = 68)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender-Neutral</td>
<td>ballpark</td>
<td>0.254</td>
<td>0.050</td>
<td>−0.204</td>
</tr>
<tr>
<td></td>
<td>speed</td>
<td>0.036</td>
<td>−0.005</td>
<td>−0.031</td>
</tr>
<tr>
<td></td>
<td><strong>Avg (abs.)</strong></td>
<td><strong>0.125</strong></td>
<td><strong>0.119</strong></td>
<td><strong>−0.006</strong></td>
</tr>
<tr>
<td>(n = 200)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Debiasing with Supervision

To debias using subspace projection, we need prior knowledge of which words are gender-appropriate.

- ‘doctor’ is gendered by stereotype → debias!
- ‘king’ is gendered by definition → don’t debias!

Can we debias without such a priori knowledge?
Debiasing without Supervision

Our simple approach: create

- gender-defining relation vector $\vec{b}^*$ (e.g., $\vec{king} - \vec{queen}$)
- bias-defining relation vector $\vec{b}'$ (e.g., $\vec{doctor} - \vec{midwife}$)

and debias a word $w$ iff

$$|\beta(\vec{w}; \vec{b}^*)| < |\beta(\vec{w}; \vec{b}')|$$
Debiasing without Supervision

Compared to Bolukbasi et al. (2016), our approach is much better at preserving gender-appropriate analogies and precluding gender-biased ones.
Conclusion

Key findings:

1. The subspace projection method provably debiases word embeddings under certain conditions.

2. WEAT has flaws that cause it to systematically overestimate bias.

3. Only gender-specific and gender-biased words are more gendered in SGNS vector spaces than in the corpus.
References

