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?

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?

Debiasing via Subspace Projection

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

$$M_{doctor,king} = M_{doctor,queen}$$

Debiasing via Subspace Projection

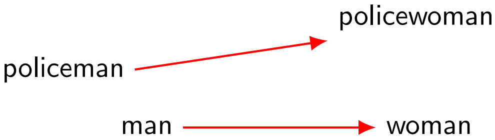
Debiasing Theorem If bias subspace $B = \text{span}(\{\vec{x} - \vec{y} \,|\, (x, y) \in S\})$ for word pairs S, then debiased word vectors $\{w_d\}$ are unbiased wrt S.

- 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.

Lipstick on a Pig?

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 $B \neq \text{span}(\{\vec{x} - \vec{y} \,|\, (x,y) \in S\})$.

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?

WEAT

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"

WEAT

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!

Problem with Using WEAT

You can cherry-pick the attributes to achieve your desired outcome.

Target Word Sets	Attribute Word Sets	Test Stat	p-val	Outcome
{door} vs. {curtain}	{masculine} vs. {feminine}	0.021	0.0	male-assoc.
	{girlish} vs. {boyish}	−0.042	0.5	inconclusive
	{woman} vs. {man}	0.071	0.0	female-assoc.
{dog} vs. {cat}	{masculine} vs. {feminine}	0.063	0.0	male-assoc.
	{actress} vs. {actor}	−0.075	0.5	inconclusive
	{womanly} vs. {manly}	0.001	0.0	female-assoc.

Problem with Using WEAT

You can cherry-pick the attributes to achieve your desired outcome.

Target Word Sets	Attribute Word Sets	Test Stat	p-val	Outcome
{door} vs. {curtain}	{masculine} vs. {feminine}	0.021	0.0	male-assoc.
	{girlish} vs. {boyish}	−0.042	0.5	inconclusive
	{woman} vs. {man}	0.071	0.0	female-assoc.
{dog} vs. {cat}	{masculine} vs. {feminine}	0.063	0.0	male-assoc.
	{actress} vs. {actor}	−0.075	0.5	inconclusive
	{womanly} vs. {manly}	0.001	0.0	female-assoc.

Can we do better?

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'))
- $\vec{b} = \text{principal component}(\{\vec{x} - \vec{y} \mid (x,y) \in S\})$

RIPA

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

Interpreting RIPA

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 $\vec{x}_1 - \vec{y}_1 = \vec{x}_2 - \vec{y}_2$, 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)

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

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)

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

- Δ_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

Word Type	Word	Corpus Genderedness	SGNS Genderedness	Δ
Gender-Appropriate (n = 164)	mom	−0.163	−0.648	0.485
	king	0.058	0.200	0.142
	Avg (abs.)	**0.231**	**0.522**	**0.291**
Gender-Biased (n = 68)	nurse	−0.190	−1.047	0.858
	architect	−0.063	0.162	0.099
	Avg (abs.)	**0.253**	**0.450**	**0.197**
Gender-Neutral (n = 200)	ballpark	0.254	0.050	−0.204
	speed	0.036	−0.005	−0.031
	Avg (abs.)	**0.125**	**0.119**	**−0.006**

Debiasing with Supervision

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

- 'doctor' is gendered by stereotype $\rightarrow$ debias!
- 'king' is gendered by definition $\rightarrow$ don't debias!

Can we debias without such *a priori* knowledge?

Debiasing without Supervision

Our simple approach: create

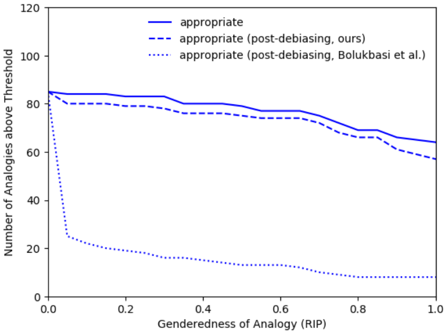
- gender-defining relation vector $\vec{b}^*$ (e.g., $\vec{king} - \vec{queen}$)
- bias-defining relation vector $\vec{b}'$ (e.g., $\vec{doctor} - \vec{midwife}$)

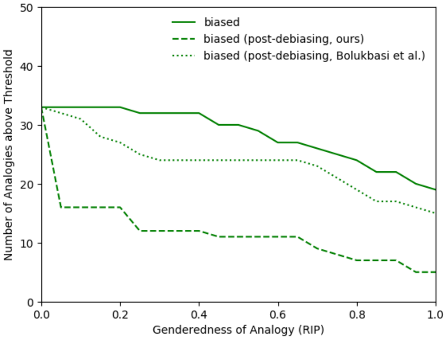
and debias a word w iff

$$|\beta(\vec{w}; \vec{b}^*)| < |\beta(\vec{w}; \vec{b}')|$$

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems, pages 4349–4357.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862.