How to debias word embeddings.

Stephanie Brandl

Københavns Universitet
TU Berlin

July 22, 2021

Man : Computer Programmer = Woman : Homemaker?

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

In a nutshell:

1. Identify gendered subspace that captures the bias
2. Neutralize and equalize word pairs to be equidistant to neutral words

Bolukbasi et al. [2016]

Projecting words on the *he-she* axis

Occupational stereotypes

Projecting words onto the *he-she* axis to find out how biased they are towards one of the two prononuns.

$$w_{\text{s:he}} = w_{\text{he}} - w_{\text{she}}$$
$$w_{\text{nurse}} = \left(w_{\text{nurse}}^{\top} \cdot w_{\text{s:he}}\right)^{\top} \cdot w_{\text{s:he}}$$

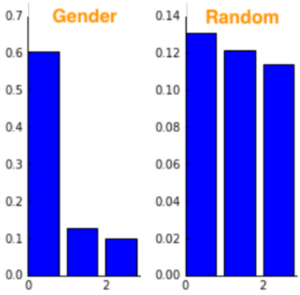
Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Debiasing algorithm

Step 1: Identify gender subspace

For a more robust estimate of the gender subspace, several directions as e.g. *she - he* and *woman - man* are combined. Principal component analysis is applied to ten gender pair difference vectors.

	def.	stereo.
$\overrightarrow{\text{she}}-\overrightarrow{\text{he}}$	92%	89%
$\overrightarrow{\text{her}}-\overrightarrow{\text{his}}$	84%	87%
$\overrightarrow{\text{woman}}-\overrightarrow{\text{man}}$	90%	83%
$\overrightarrow{\text{Mary}}-\overrightarrow{\text{John}}$	75%	87%
$\overrightarrow{\text{herself}}-\overrightarrow{\text{himself}}$	93%	89%
$\overrightarrow{\text{daughter}}-\overrightarrow{\text{son}}$	93%	91%
$\overrightarrow{\text{mother}}-\overrightarrow{\text{father}}$	91%	85%
$\overrightarrow{\text{gal}}-\overrightarrow{\text{guy}}$	85%	85%
$\overrightarrow{\text{girl}}-\overrightarrow{\text{boy}}$	90%	86%
$\overrightarrow{\text{female}}-\overrightarrow{\text{male}}$	84%	75%



Step 2: Neutralize and Equalize

Neutralize: For neutral words/non-gendered words, the gender direction is removed i.e. they are zero in the gender subspace

$$N \subseteq W, \forall w \in N, \tilde{w} := \frac{w - w_B}{\|w - w_B\|}$$

where N is the set of neutral words and W defines the set of all words in the vocabulary, w_B is w projected onto B, the gendered subspace.

Equalize: Pairs of gendered words (e.g. *mother* - *father*) are made equidistant to all neutral words, i.e. word embeddings are centered and then scaled to unit length.

$$\forall E \in \mathcal{E} : \mu_E := \sum_{w \in E} \frac{w}{|E|} \text{ and } \mu_{E_{\perp B}} = \mu_E - \mu_{E_B}$$

$$\forall w \in E : \tilde{w} := \mu_{E_{\perp B}} + \sqrt{1 - \|\mu_{E_{\perp B}}\|^2} \frac{w_B - \mu_{E_B}}{\|w_B - \mu_{E_B}\|}$$

where $\mathcal{E}$ is the set of gendered word pairs.

It’s all in the name

It’s All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution

Rowan Hall Maudslay[1] **Hila Gonen**[2] **Ryan Cotterell**[1] **Simone Teufel**[1]
[1] Department of Computer Science and Technology, University of Cambridge
[2] Department of Computer Science, Bar-Ilan University
{rh635,rdc42,sht25}@cam.ac.uk hilagnn@gmail.com

In a nutshell:

1. Comparison of Word Embedding Debiasing (WED) with Counterfactual Data Augmentation
2. Two add-ons: Counterfactual Data Substitution and Names Intervention

Maudslay et al. [2019]

Naive approach: The original text is transformed and added to the original corpus. For instance gendered word pairs are swapped:

the *woman* cleaned the kitchen → the *man* cleaned the kitchen

The grammar intervention uses Part-of-Speech information to maintain the relation between personal pronoun and possessive determiner:

her teacher was proud of *her* → *his* teacher was proud of *him*

It also prevents swapping of gendered words when they refer to a proper noun, such as

Elizabeth ... *she* ... *queen* would not be changed to
Elizabeth ... *he* ... *king*

Lu et al. [2018]

Instead of duplicating the text which causes peculiar statistical properties such as only even word frequencies, the authors propose *Counterfactual Data Substitution*:

There, text will not be duplicated but substituted with a substitution probability of 0.5 on a per-document basis.

To further neutralise the data, the authors propose an explicit treatment of names. With a list of first names from the United Social security Administration (SSA), pairs of names are matched based on name frequency and degree of gender-specificity.

The list of name pairs is added to the gendered word pairs to swap names along with personal pronouns and possessive determiners.

Jordan usually does *his* homework in the late afternoon after soccer practice. $\rightarrow$
Taylor usually does *her* homework in the late afternoon after soccer practice.

method	explanation
none	no debiasing
CDA	naive Counterfactual Data Augmentation
gCDA	CDA with grammar intervention
nCDA	CDA with names intervention
gCDS	CDS with grammar intervention
nCDS	CDS with names intervention
WED40	WED with 40% of variance explained in gender subspace
WED70	WED with 70% of variance explained in gender subspace
nWED70	WED70 with names intervention

Direct bias

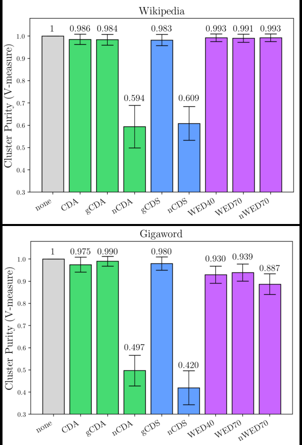
Word Embedding Association Tests (WEAT) measure relative difference between two sets of target words and two sets of attributes (e.g. *female-male*). The distance between word pairs is measured with d (higher - more biased) and a one-sided p-value to decide whether the detected bias is significant.
Three target pairs are applied: *arts-maths*, *arts-sciences*, *careers-family*

Method	Art–Maths d	Art–Maths p	Arts–Sciences d	Arts–Sciences p	Career–Family d	Career–Family p
Gigaword						
none	1.32	$< 10^{-2}$	1.50	$< 10^{-3}$	1.74	$< 10^{-4}$
CDA	0.67	.10	1.05	.02	1.79	$< 10^{-4}$
gCDA	1.16	.01	1.46	$< 10^{-2}$	1.77	$< 10^{-4}$
nCDA	−0.49	.83	0.34	.27	1.45	$< 10^{-3}$
gCDS	0.96	.03	1.31	$< 10^{-2}$	1.78	$< 10^{-4}$
nCDS	−0.19	.63	0.48	.19	1.45	$< 10^{-3}$
WED40	−0.73	.92	0.31	.28	1.24	$< 10^{-2}$
WED70	−0.73	.92	0.30	.29	1.15	$< 10^{-2}$
nWED70	0.30	.47	0.54	.19	0.59	.15
Wikipedia						
none	1.64	$< 10^{-3}$	1.51	$< 10^{-3}$	1.88	$< 10^{-4}$
CDA	1.58	$< 10^{-3}$	1.66	$< 10^{-4}$	1.87	$< 10^{-4}$
gCDA	1.52	$< 10^{-3}$	1.57	$< 10^{-3}$	1.84	$< 10^{-4}$
nCDA	1.06	.02	1.54	$< 10^{-4}$	1.65	$< 10^{-4}$
gCDS	1.45	$< 10^{-3}$	1.53	$< 10^{-3}$	1.87	$< 10^{-4}$
nCDS	1.05	.02	1.37	$< 10^{-3}$	1.65	$< 10^{-4}$
WED40	1.28	$< 10^{-2}$	1.36	$< 10^{-2}$	1.81	$< 10^{-4}$
WED70	1.05	.02	1.24	$< 10^{-2}$	1.67	$< 10^{-3}$
nWED70	−0.46	.52	−0.42	.51	0.85	.05
Nosek et al.	0.82	$< 10^{-2}$	1.47	$< 10^{-24}$	0.72	$< 10^{-2}$

Caliskan et al. [2017]

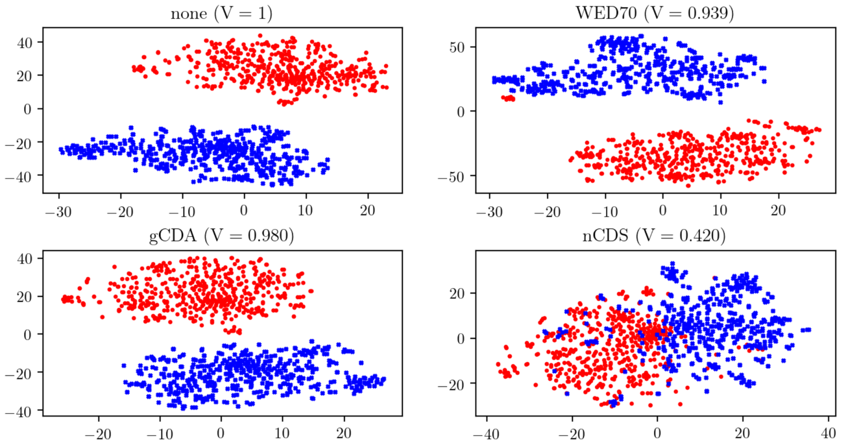
Indirect bias

1. a subspace b_{test} is defined based on 23 word pairs used in the Google Analogy family test subset
2. 1000 most biased words in each corpus are defined as the 500 closest to b_{test} and $-b_{\text{test}}$ in the original embedding space
3. after debiasing, corresponding word embeddings are projected into 2D space (with t-SNE)
4. k-means clustering is applied
5. the cluster's V-measure computed



Lower V-measure means that words are less clustered than before.

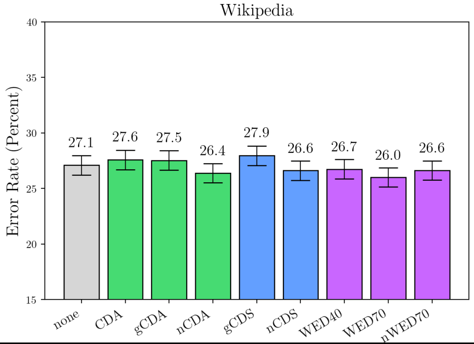
Evaluation Methods: Indirect Bias

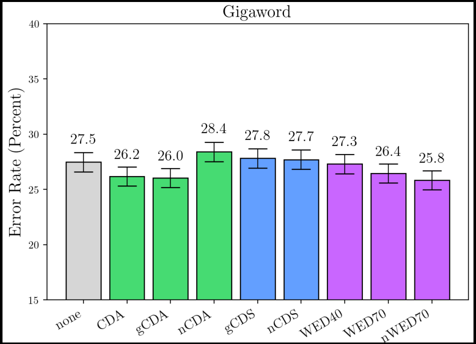


Gonen and Goldberg [2019]

Sentiment Classification

To evaluate how well the debiased embeddings perform on standard downstream tasks, a standard sentiment classification task is applied where the debiased embeddings are used as pretrained word embedding input.

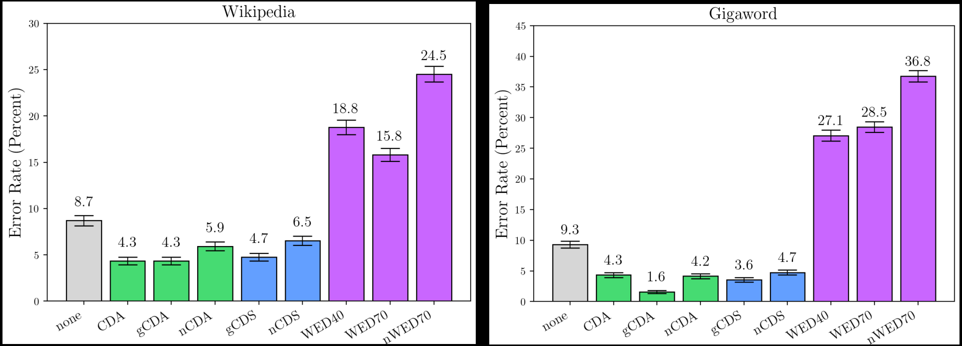




Non-biased gender analogies

The 506 analogies from the *family analogy* subset of the Google Analogy Test set are applied to the debiased word embeddings as

boy:girl :: nephew: ?



Mikolov et al. [2013]

- Word embedding debiasing (WED) mitigates direct bias more succesfully than the other methods and also shows better results in the sentiment classification
- The names intervention clearly mitigates indirect bias much better than all other methods
- Counterfactual data augmentation and Counterfactual data substitution improve the performance on the family analogy tasks while the performance of WED is worse than before debiasing the embeddings

Limitations

- All methods are based on predefined lists of gender words/pairs which for pairs as *manager : manageress* might be problematic
- The main assumption of a gender binary ignores non-binary gender identity
- All presented methods only try to mitigate gender bias.
  What about other biases?
- None of the presented methods can succesfully remove direct and indirect gender bias

References and Further Reading

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

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

Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

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

K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*, 2018.

T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*, 2019.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.