Gender Bias detection

Suyash Sethia
Jainit Bafna

INDEX

- Introduction on gender bias
- Word embedding system
- Paper on WEAT (word embedding association test)
- Paper on flaws of various de-biasing techniques
- Paper on proposing an efficient de-biasing technique .

Gender and Sex

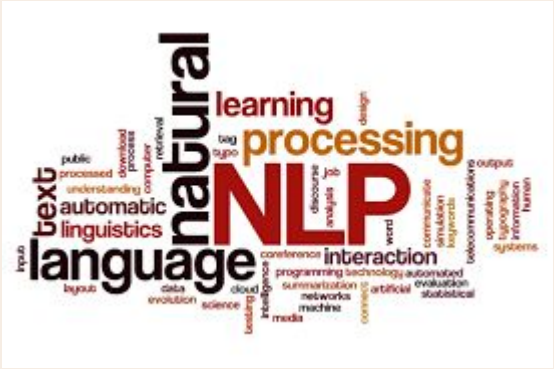
Sentence :- only men can be soldiers . => biased

Sentence :- only men can be fathers . => not-biased

WHAT WHERE HOW ?

Gender bias in NLP tools

- Machine translation
- Article generation
- Word embeddings



Papers

1. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories (Kaytlin Chaloner , Alfredo Maldonado)
2. Understanding Undesirable Word Embedding Associations (Kawin Ethayarajh, David Duvenaud† , Graeme Hirst)
3. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them (Hila Gonen1 and Yoav Goldberg1,2)

Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories

Kaytlin Chaloner
ADAPT Centre, SCSS
Trinity College Dublin
Ireland
chalonek@tcd.ie

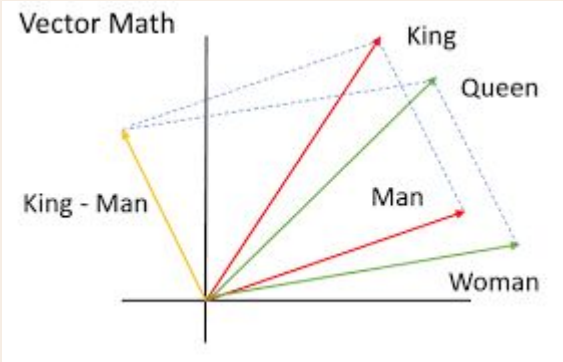
Alfredo Maldonado
ADAPT Centre, SCSS
Trinity College Dublin
Ireland
alfredo.maldonado@adaptcentre.ie

Word embedding system

- The most important feature of word embeddings is that similar words in a semantic sense have a smaller distance (either Euclidean, cosine or other)
- “Ped” and tree will have almost the same word

Embedding despite being in different languages

due to sense similarity



How Word Embeddings are Created

1. Read the text

How Word Embeddings are Created

1. Read the text
2. Preprocess text

How Word Embeddings are Created

1. Read the text
2. Preprocess text
3. Create (x,y) data points

How Word Embeddings are Created

1. Read the text
2. Preprocess text
3. Create (x,y) data points
4. Create one hot encoded (X,Y) matrices

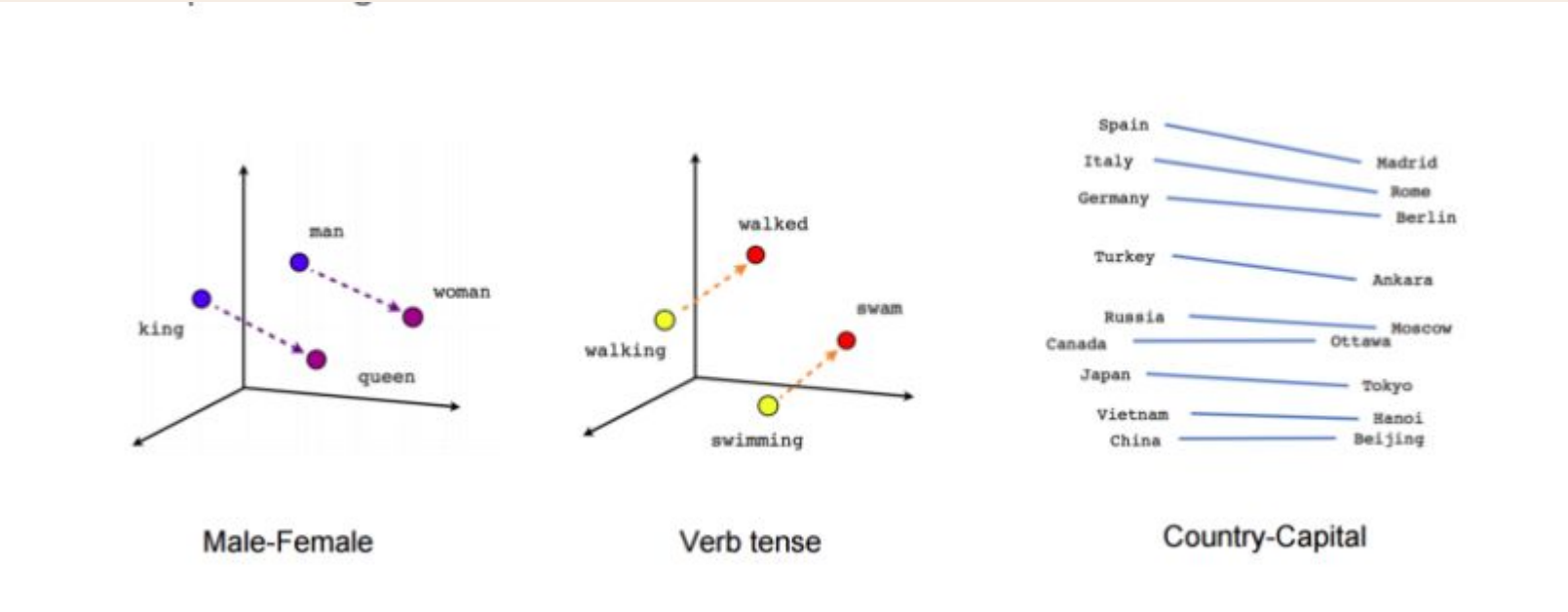
How Word Embeddings are Created

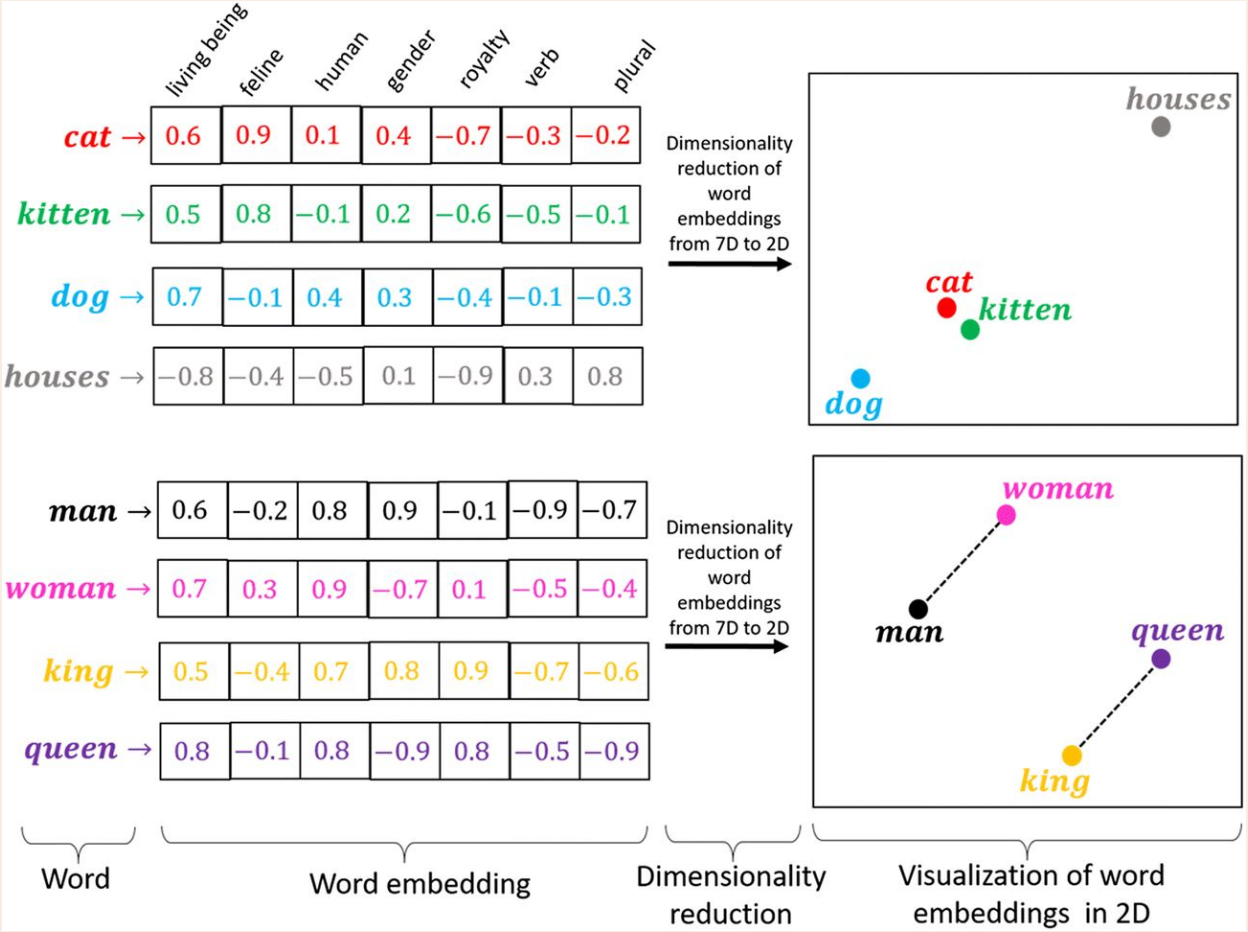
1. Read the text
2. Preprocess text
3. Create (x,y) data points
4. Create one hot encoded (X,Y) matrices
5. Train a neural network

How Word Embeddings are Created

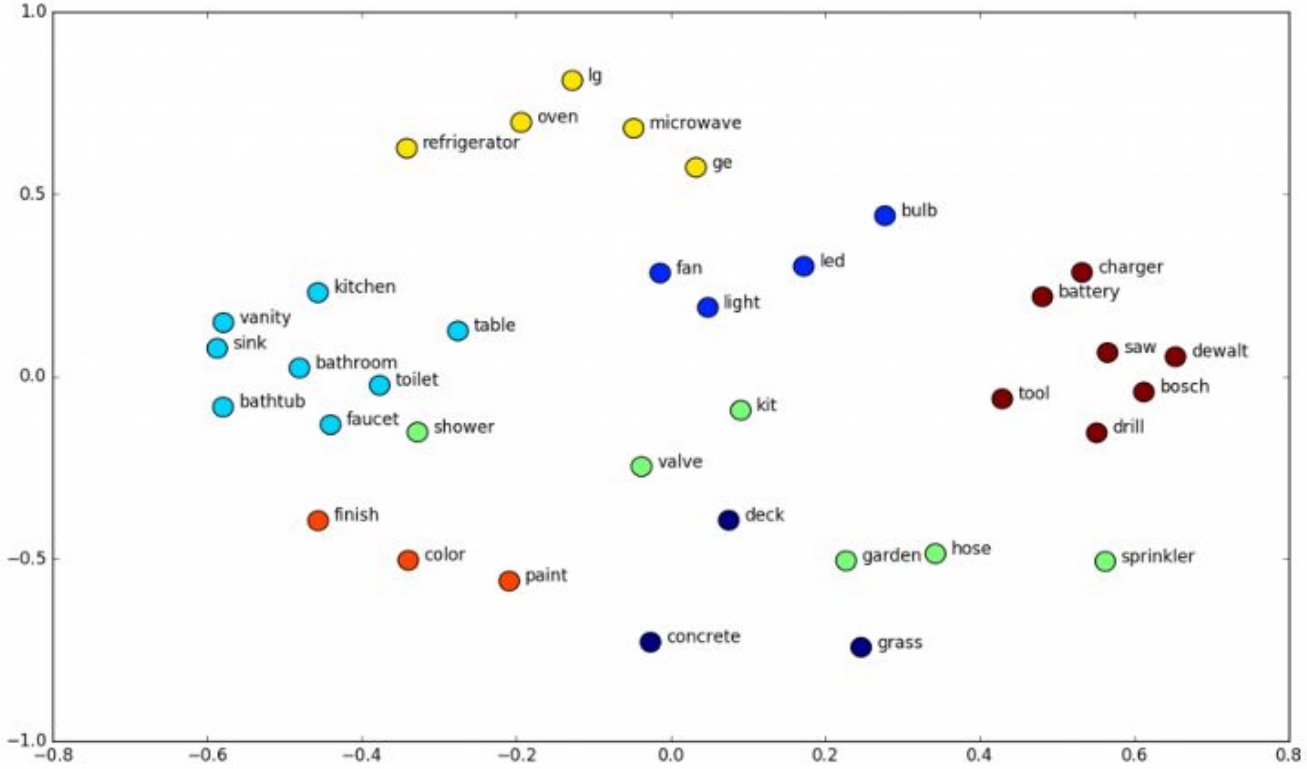
1. Read the text
2. Preprocess text
3. Create (x,y) data points
4. Create one hot encoded (X,Y) matrices
5. Train a neural network
6. Extract the weights from the input layer

How Word Embeddings are Created




living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Vector space



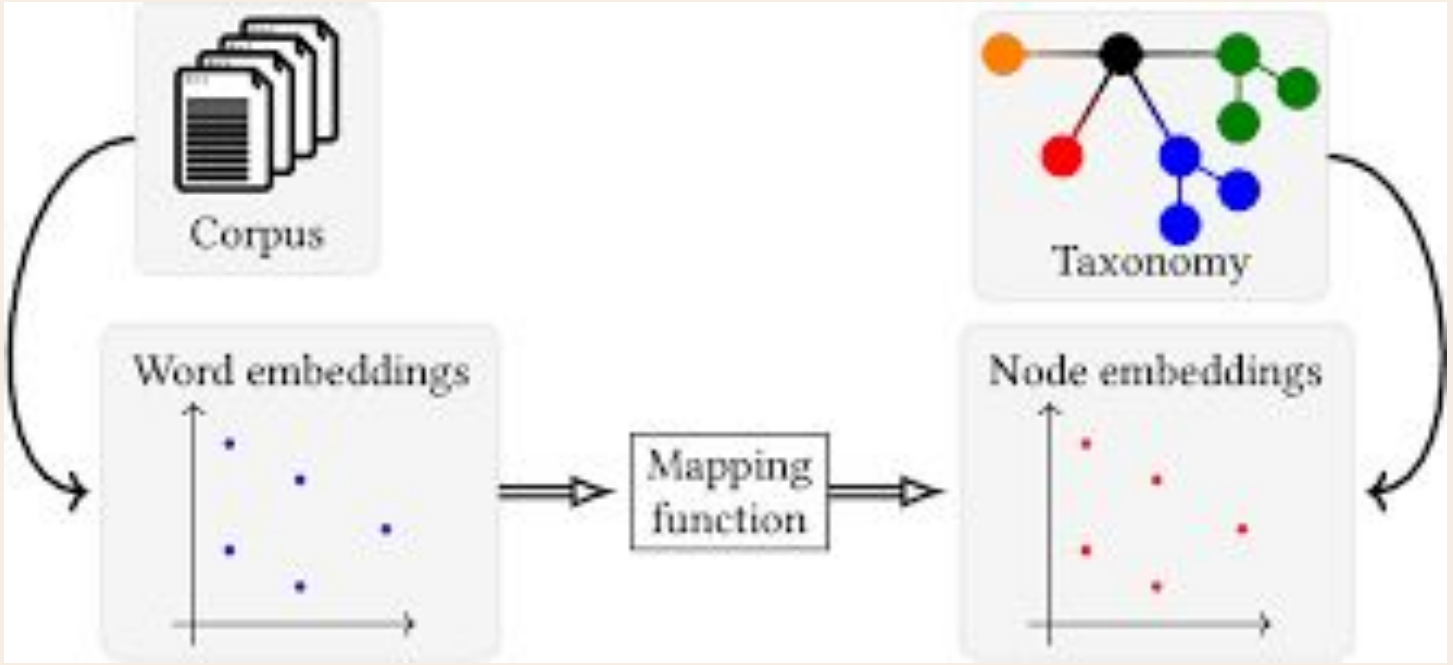
Choice of word embeddings

1. Skip-Gram embed-dings trained on the Google News corpus2 , with a vocabulary of 3M word types (Mikolov et al.,2013)

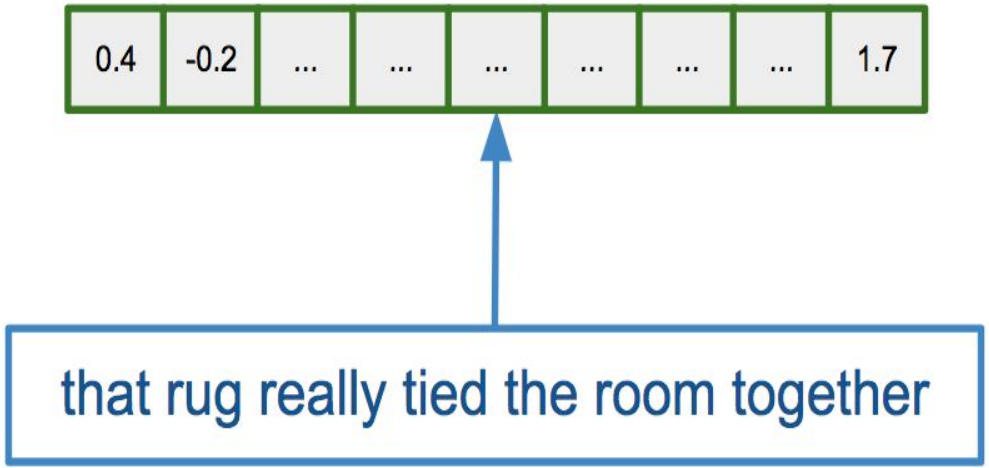
2. Skip-Gram embeddings trained on 400 million Twitter micro-posts3 , with a vocabulary of slightly more than 3M word types (Godin et al.2015)

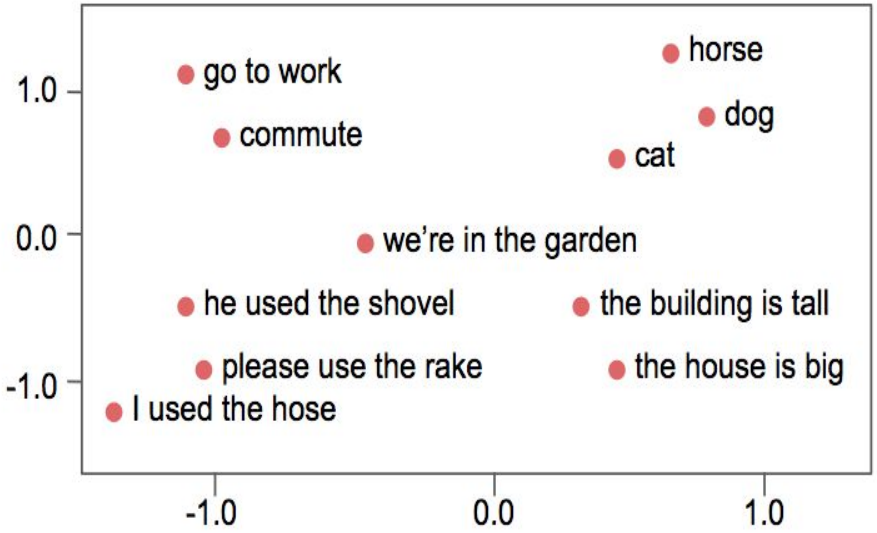
3. Skip-Gram embeddings trained on the PubMed Central Open Access subset (PMC) and PubMed4 , with a vocabulary of about 2.2M word types (Chiu et al., 2016) and trained using two dif- ferent sliding window sizes: 2 and 30 words;

Embeddings represent the corpus



WEAT (word embedding association test)





WEAT

- X = {programmer, engineer, scientist}
  Y = {nurse, teacher, librarian}.

- **Attribute Words :-**
  M = {man, male, he},
  F = {woman, female, she}

Ho = Null hypothesis that there is no difference between X and Y in terms of their relative (cosine) similarity to M and F

$$s(X, Y, M, F) = \sum_{x \in X} s(x, M, F) - \sum_{y \in Y} s(y, M, F) \quad (1)$$

where $s(w, M, F)$ is the **measure of association** between target word w and the attribute words in M and F:

$$s(w, M, F) = \frac{1}{|M|} \sum_{m \in M} \cos(\vec{w}, \vec{m}) - \frac{1}{|F|} \sum_{f \in F} \cos(\vec{w}, \vec{f}) \quad (2)$$

	Attribute words	M	male, man, boy, brother, he, him, his, son, father, uncle, grandfather
		F	female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother
Target words	B1: career vs family	X	executive, management, professional, corporation, salary, office, business, career
		Y	home, parents, children, family, cousins, marriage, wedding, relatives
	B2: maths vs arts	X	math, algebra, geometry, calculus, equations, computation, numbers, addition
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B3: science vs arts	X	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B4: intelligence vs appearance	X	precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative, shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent
		Y	alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong
	B5: strength vs weakness	X	power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, shout, dynamic, winner
		Y	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser

Code snippets

```
def cosine_means_difference(wv, word, male_attrs, female_attrs):
    male_mean = cosine_mean(wv, word, male_attrs)
    female_mean = cosine_mean(wv, word, female_attrs)
    return male_mean - female_mean


def cosine_mean(wv, word, attrs):
    return wv.cosine_similarities(wv[word], [wv[w] for w in attrs]).mean()
```

Results

	Google News		Twitter		PubMed w2		PubMed w30		GAP	
Categories	*p*	*d*	*p*	*d*	*p*	*d*	*p*	*d*	*p*	*d*
B1: career vs family	**0.0012**	1.37	**0.0029**	1.31	0.7947	-0.42	0.0962	0.67	**0.0015**	1.44
B2: maths vs arts	**0.0173**	1.02	0.1035	0.65	0.9996	-1.40	0.9966	-1.20	0.0957	1.04
B3: science vs arts	**0.0044**	1.25	0.0715	0.74	0.9797	-0.98	0.7670	-0.37	0.1434	0.71
B4: intelligence vs appearance	**0.0001**	0.98	0.1003	0.37	0.2653	0.18	0.0848	0.36	0.9988	-0.64
B5: strength vs weakness	**0.0059**	0.89	0.2971	0.20	0.0968	0.48	**0.0237**	0.72	**0.0018**	0.77

Lipstick on a pig

Lipstick on a pig

Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

What is “Biased” According to Bolukbasi et al. ?

What is “Biased” According to Bolukbasi et al. ?

- Bolukbasi et al. (2016b) define the gender of a word *w* by its projection on the “gender direction” :
  $\vec{w}.(\vec{he} - \vec{she})$, assuming all vectors are normalised.

What is “Biased” According to Bolukbasi et al. ?

- Bolukbasi et al. (2016b) define the gender of a word *w* by its projection on the “gender direction” :
  $\vec{w}.(\vec{he} - \vec{she})$, assuming all vectors are normalised.
- The larger a word’s projection is on he − she, the more biased it is. They also quantify the bias in word embeddings using this definition and show that it aligns well with social stereotypes

Debiasing method by

1. Bolukbasi et al. (2016b)

Bolukbasi et al. ‘s method

- It was a post-processing method

Bolukbasi et al. ‘s method

- It was a post-processing method
- Given a word embedding matrix, they make changes to the word vectors in order to reduce the gender bias as much as possible for all words that are not inherently gendered (e.g. mother, brother, queen).

Bolukbasi et al. ‘s method

- It was a post-processing method
- Given a word embedding matrix, they make changes to the word vectors in order to reduce the gender bias as much as possible for all words that are not inherently gendered (e.g. mother, brother, queen).
- They do that by zeroing the gender projection of each word on a predefined gender direction.

Bolukbasi et al. ‘s method

- It was a post-processing method
- Given a word embedding matrix, they make changes to the word vectors in order to reduce the gender bias as much as possible for all words that are not inherently gendered (e.g. mother, brother, queen).
- They do that by zeroing the gender projection of each word on a predefined gender direction.
- However, as we show in this work, while the gender-direction is a great indicator of bias, it is only an indicator and not the complete manifestation of this bias

Debiasing method by

1. Bolukbasi et al. (2016b)
2. Zhao et al. (2018)

Zhao el al. ‘s method

- Took a different approach and suggested to train debiased word embeddings from scratch.

Zhao el al. ‘s method

- Take a different approach and suggest to train debiased word embeddings from scratch.
- Instead of debiasing existing word vectors, they alter the loss of the GloVe model , aiming to concentrate most of the gender information in the last coordinate of each vector.

Zhao el al. ‘s method

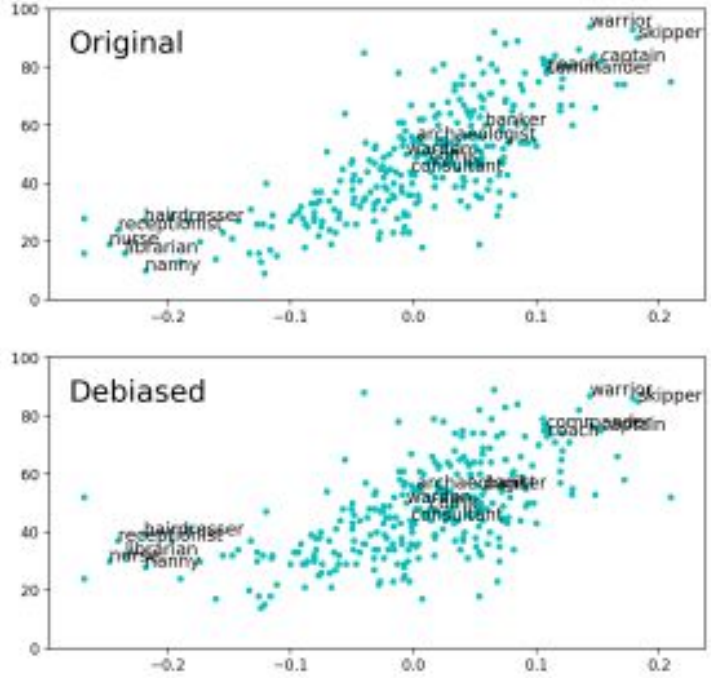
- Take a different approach and suggest to train debiased word embeddings from scratch.
- Instead of debiasing existing word vectors, they alter the loss of the GloVe model , aiming to concentrate most of the gender information in the last coordinate of each vector.
- This way, one can later use the word representations excluding the gender coordinate.

Zhao el al. ‘s method

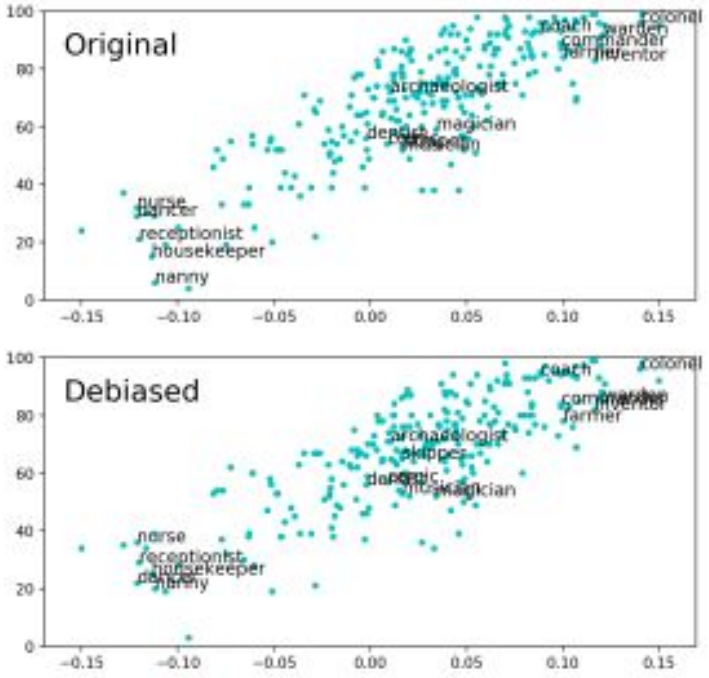
- Take a different approach and suggest to train debiased word embeddings from scratch.
- Instead of debiasing existing word vectors, they alter the loss of the GloVe model , aiming to concentrate most of the gender information in the last coordinate of each vector.
- This way, one can later use the word representations excluding the gender coordinate.
- They do that by using two groups of male/female seed words, and encouraging words that belong to different groups to differ in their last coordinate.

Remaining bias after using debiasing methods

- Word similarity is still same
- The definition of biasness is not essentially correct ie it is very naive .
- Just the vector projection on (man vector - woman vector) is removed which does not account for complete de-bias .



(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.



(b) The plots for GN-GLOVE embedding, before (top) and after (bottom) debiasing.

Conclusion

Research questions

Our focus is on mainly three questions

- Does debiasing via subspace projection method provably debias embeddings ?
- Why should WEAT be used to measure word associations ?
- What's to blame , why we still have gender bias in data , is the word embedding model , training data ?

Defining unbiasedness

- Let M be the word-context matrix the embedding model implicitly factorizes $WC^T = M$
- Word w **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

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

Lipstick on a pig

- If we define a bias subspace using S and use it to debias ‘*w*’, we can only say definitively that w is unbiased with respect to S . where S contains {‘male’, ‘female’}

Lipstick on a pig

- If we define a bias subspace using S and use it to debias ‘*w*’, we can only say definitively that w is unbiased with respect to S . where S contains {‘m$\vec{a}$le’, ‘fem$\vec{a}$le’}
- We cannot claim,for example, that *‘w’* is also unbiased with respect to {(‘police$\vec{w}$oman’, ‘policew$\vec{o}$man’)}, because it is possible that

  $\vec{police man} - \vec{police woman} \neq \vec{man} - \vec{woman}$

Problem with WEAT

Target Word Sets	Attribute Word Sets	Test Statistic	Effect Size	p-value	Outcome (WEAT)
{door} vs. {curtain}	{masculine} vs. {feminine}	0.021	2.0	0.0	more male-associated
	{girlish} vs. {boyish}	−0.042	−2.0	0.5	inconclusive
	{woman} vs. {man}	0.071	2.0	0.0	more female-associated
{dog} vs. {cat}	{masculine} vs. {feminine}	0.063	2.0	0.0	more male-associated
	{actress} vs. {actor}	−0.075	−2.0	0.5	inconclusive
	{womanly} vs. {manly}	0.001	2.0	0.0	more female-associated
{bowtie} vs. {corsage}	{masculine} vs. {feminine}	0.017	2.0	0.0	more male-associated
	{woman} vs. {masculine}	−0.071	−2.0	0.5	inconclusive
	{girly} vs. {masculine}	0.054	2.0	0.0	more female-associated

Table 1: By contriving the male and female attribute words, we can easily manipulate WEAT to claim that a given target word is more female-biased or male-biased than another. For example, in the top row, $\vec{door}$ is more male-associated than $\vec{curtain}$ when the attribute words are 'masculine' and 'feminine', but it is more female-associated when the attribute words are 'woman' and 'man'. In both cases, the associations are highly statistically significant.

New Method RIPA

- RIPA stands for Relational Inner Product Association
- The RIPA of a word *w* wrt to relational vector :

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

where

- Word pairs S define the association (e.g. {‘king’ , ‘queen’})
- b =principal component ({x-y | (x,y) ∈ S })

Advantages of RIPA

- Interpretable when embedding model factorizes word-context matrix

Advantages of RIPA

- Interpretable when embedding model factorizes word-context matrix

- Robust to how *b* is defined

Advantages of RIPA

- Interpretable when embedding model factorizes word-context matrix
- Robust to how *b* is defined
- Derived from the subspace projection method of debiasing

What Cause Bias ,Training Data or Model?

Word Type	Word	Genderedness in Corpus	Genderedness in Embedding Space	Change (abs.)
Gender-Appropriate (n = 164)	mom	−0.163	−0.648	0.485
	dad	0.125	0.217	0.092
	queen	−0.365	−0.826	0.462
	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
	doctor	−0.135	−0.059	−0.077
	housekeeper	−0.132	−0.927	0.795
	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
	calf	−0.039	0.027	−0.012
	hormonal	−0.326	−0.551	0.225
	speed	0.036	−0.005	−0.031
	Avg (abs.)	**0.125**	**0.119**	**−0.006**

Table 2: On average, SGNS makes gender-appropriate words (e.g., 'queen') and gender-biased words (e.g., 'nurse') *more* gendered in the embedding space than they are in the training corpus. As seen in the last column (in bold), the average change in absolute genderedness is 0.291 and 0.197 respectively ($p < 0.001$ for both). For gender-neutral words, the average change is only -0.006 ($p = 0.84$): SGNS does not make them any more gendered.

RIPA model is Unsupervised

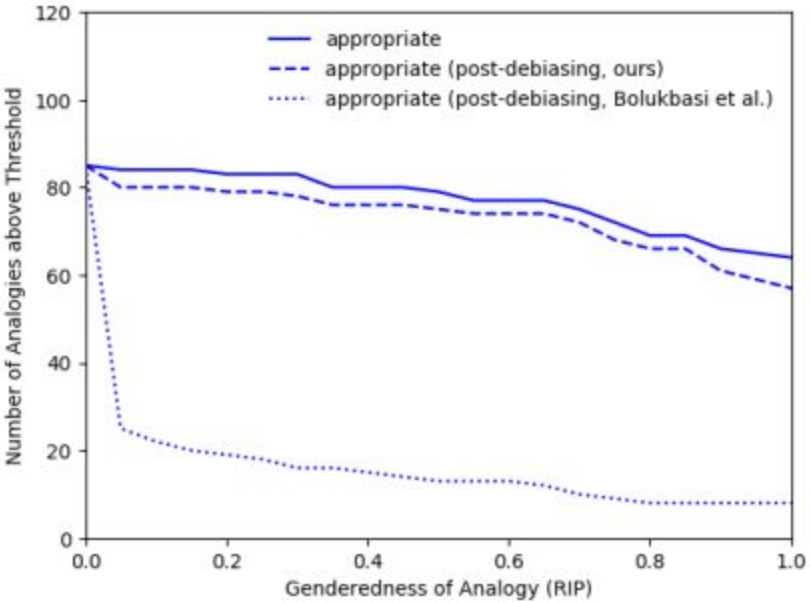
- So unlike the method introduced by Bolukbasi, the algorithm automatically decides which words should be debiased.

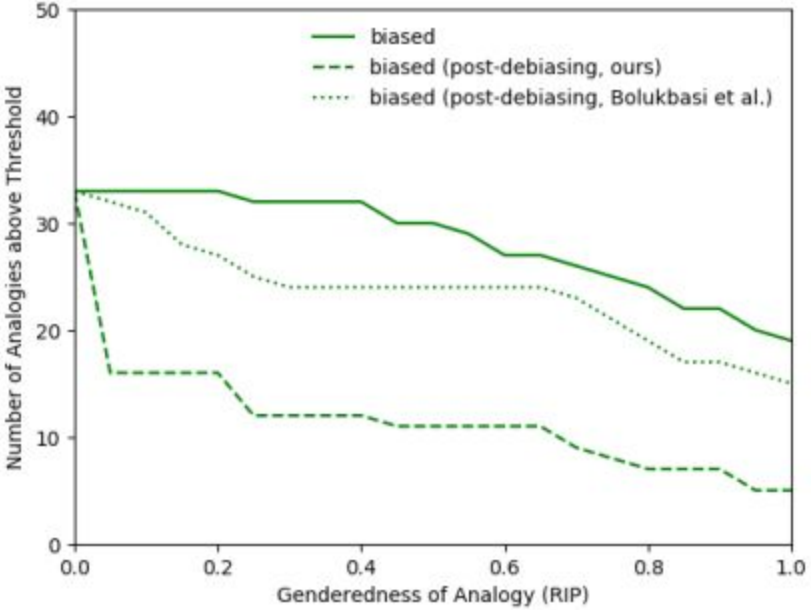
RIPA model is Unsupervised

- So unlike the method introduced by Bolukbasi, the algorithm automatically decides which words should be debiased.
- We propose an unsupervised method for finding gender-appropriate words. We first create a gender-defining relation vector $b*$ by taking the first principal component of gender-defining difference vectors such as man− woman.

RIPA model is Unsupervised

- So unlike the method introduced by Bolukbasi, the algorithm automatically decides which words should be debiased.
- We propose an unsupervised method for finding gender-appropriate words. We first create a gender-defining relation vector $b*$ by taking the first principal component of gender-defining difference vectors such as man− woman.
- Using difference vectors from biased analogies, such as doctor - midwife we then create a bias-defining relation vector b the same way. We then debias a word w using the subspace projection method iff it satisfies $|\beta(w;b*)| < |\beta(w;b0)|$his simple condition is sufficient to preserve almost all gender-appropriate analogies while precluding most gender-biased ones.


appropriate
appropriate (post-debiasing, ours)
appropriate (post-debiasing, Bolukbasi et al.)
Number of Analogies above Threshold
Genderedness of Analogy (RIP)
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0


biased
biased (post-debiasing, ours)
biased (post-debiasing, Bolukbasi et al.)
Number of Analogies above Threshold
Genderedness of Analogy (RIP)
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0

Biblography

- Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories - Chaloner & Maldonado, 2019.
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.
- Understanding Undesirable Word Embedding Associations (Ethayarajh et al., ACL 2019)
- Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai- Wei Chang. 2018. Learning gender-neutral word embeddings. In Proceedings of EMNLP, pages 4847–4853.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.