Word Embeddings: Applications and Evaluation

Recap: Last Class

- Intuitive ideas behind representing words as vectors
- Distributional Hypothesis
- Basic ideas behind TF-IDF weighting
- Basic ideas behind Word2Vec
  - Difference between CBOW and Skip-gram
  - Practical challenges
- *Know where your embeddings came from and how they were made*

This class

- **Applications**
  - How do we use these embeddings for text analysis?
    - Types of questions we can ask (occupational stereotypes, changes over time)
    - Methods for embedding operations

- **Evaluation**
  - How do we know when embeddings actually capture the content we want?

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Measures of Race/Gender Stereotypes

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

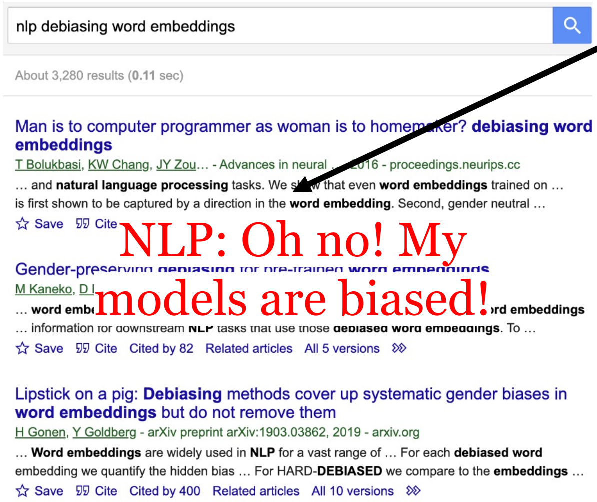
Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

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


nlp debiasing word embeddings

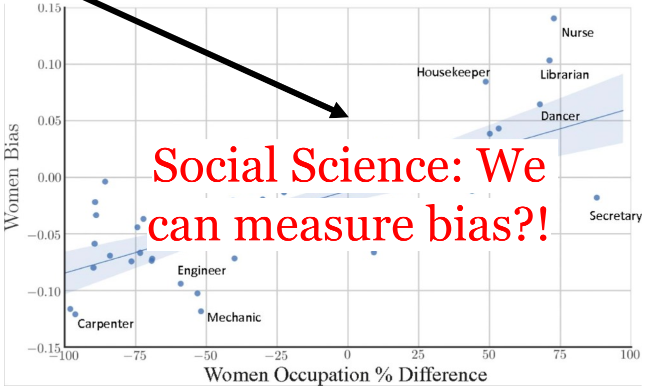
About 3,280 results (0.11 sec)

Man is to computer programmer as woman is to homemaker? **debiasing word embeddings**
T Bolukbasi, KW Chang, JY Zou… - Advances in neural … 2016 - proceedings.neurips.cc
… and **natural language processing** tasks. We show that even **word embeddings** trained on …
is first shown to be captured by a direction in the **word embedding**. Second, gender neutral …
Save Cite

Gender-preserving debiasing for pre-trained word embeddings
M Kaneko, D
… word emb… …rd embeddings
… information for downstream **NLP** tasks that use those **debiased word embeddings**. To …
Save Cite Cited by 82 Related articles All 5 versions

Lipstick on a pig: **Debiasing** methods cover up systematic gender biases in **word embeddings** but do not remove them
H Gonen, Y Goldberg - arXiv preprint arXiv:1903.03862, 2019 - arxiv.org
… **Word embeddings** are widely used in **NLP** for a vast range of … For each **debiased word**
embedding we quantify the hidden bias … For HARD-**DEBIASED** we compare to the **embeddings** …
Save Cite Cited by 400 Related articles All 10 versions

NLP: Oh no! My models are biased!



Social Science: We can measure bias?!

How do we measure similarity between gendered words and stereotype words?

- “Programmer” is more similar to “man”; “homemaker” is more similar to “woman”

- We already built embeddings (last class), we just need a measure of distance

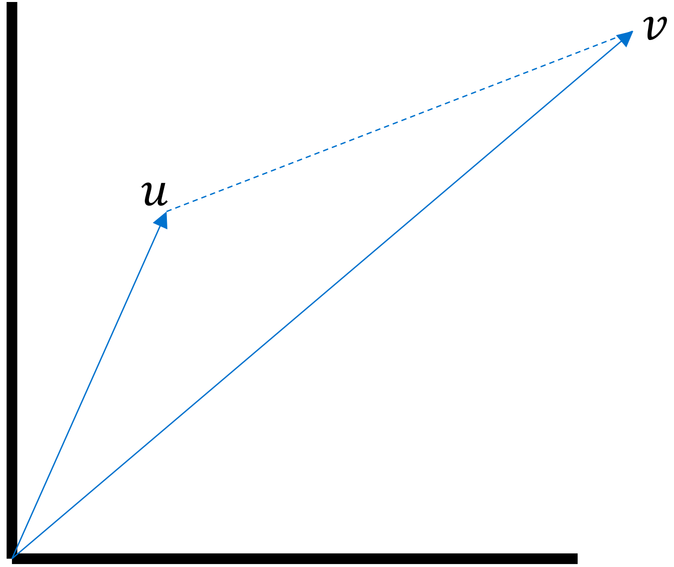
Word Embedding Similarity

- Euclidean distance

$$\sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 \ldots}$$

$$-||u - v||_2$$

- Negate to get a similarity function

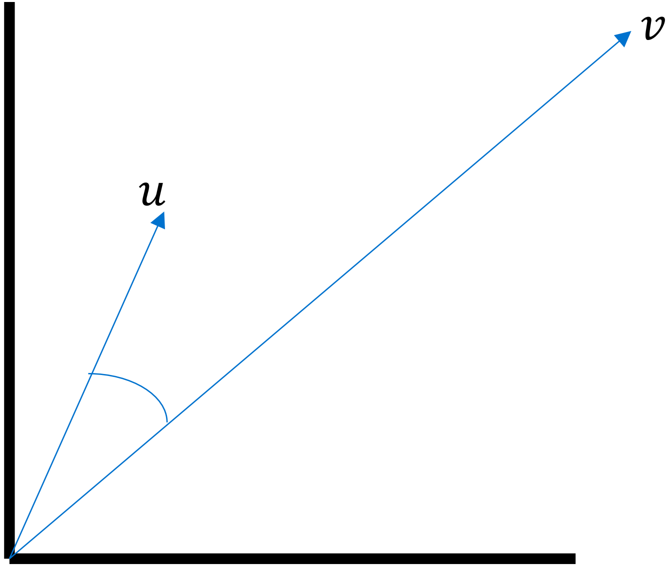


Word Embedding Similarity

- Cosine Similarity

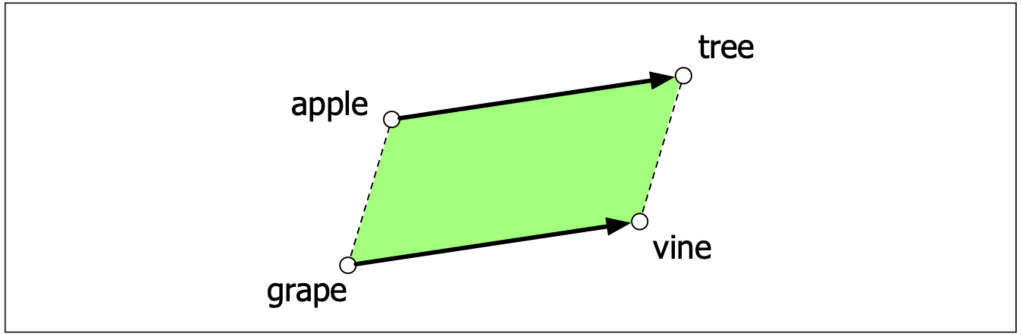
$$\frac{u \cdot v}{||u||||v||}$$

- Recall: Skip-gram objective function
  - $P\left(w_{t+j} \middle| w_t\right) = P(o \mid c) = \frac{\exp(u_o^T v_c)}{\sum_{i=1}^{V} \exp(u_i^T v_c)}$

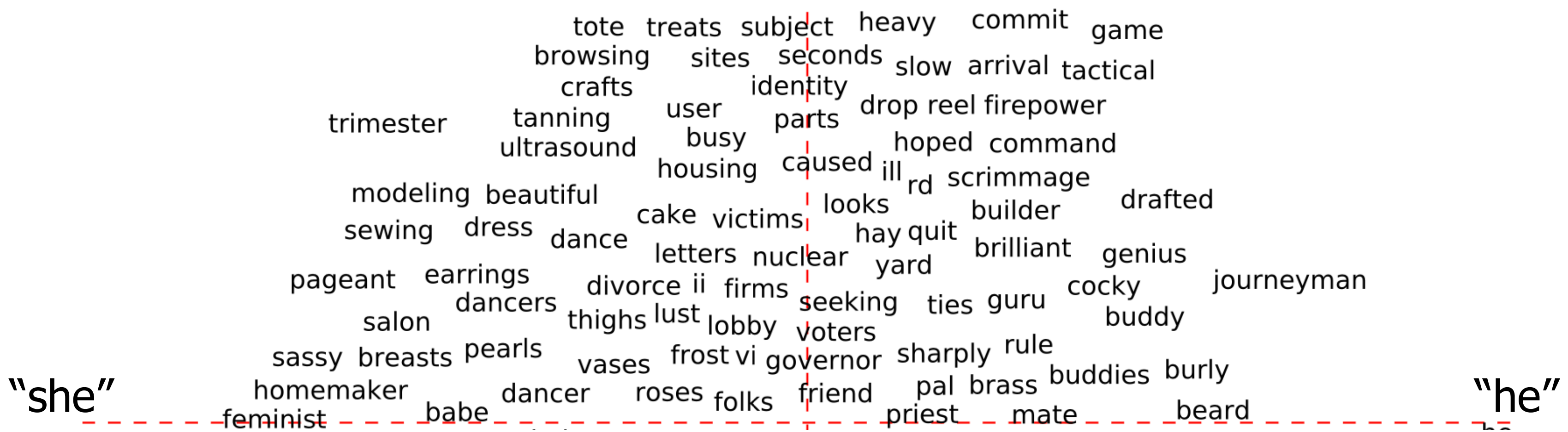


How do we measure similarity between gendered words and stereotype words?

- Vector arithmetic for analogies:
  - “King” – “man” + “woman” = “queen”
  - “computer programmer” – “man” + woman = “homemaker”



- Key idea:
  - There is a gender subspace

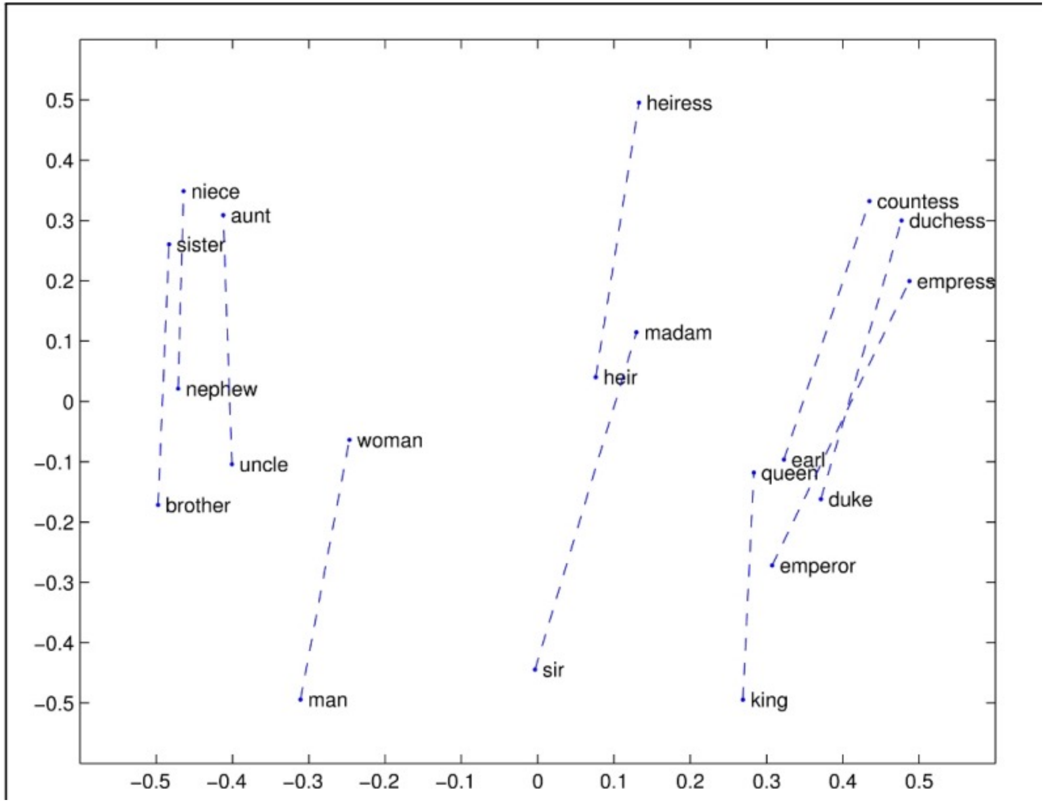


- Disclaimers:
  - Project embeddings onto he-she direction

How do we measure similarity between gendered words and stereotype words?

- “Programmer” is more similar to “man”; “homemaker” is more similar to “woman”
  - “Oh man”
  - “Man the station”
    - “Programmer” co-occurs more often with “man the station” than “homemaker” – not clearly indicate of gender bias

Relational properties of the GloVe vector space (Pennington et al., 2014)

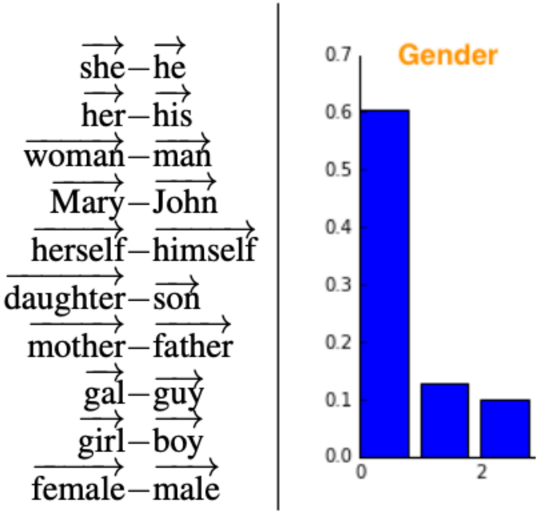


JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Identify gender subspace: Pairs words + PCA

- Principle Component Analysis
  - Identify directions of greatest variance

- First PCA eigenvector explains most of the variance:
  - Consider this component to be the gender (bias) subspace



[In actual formulations, defined gender subspace based on difference from mean of vectors rather than individual vector pairs]

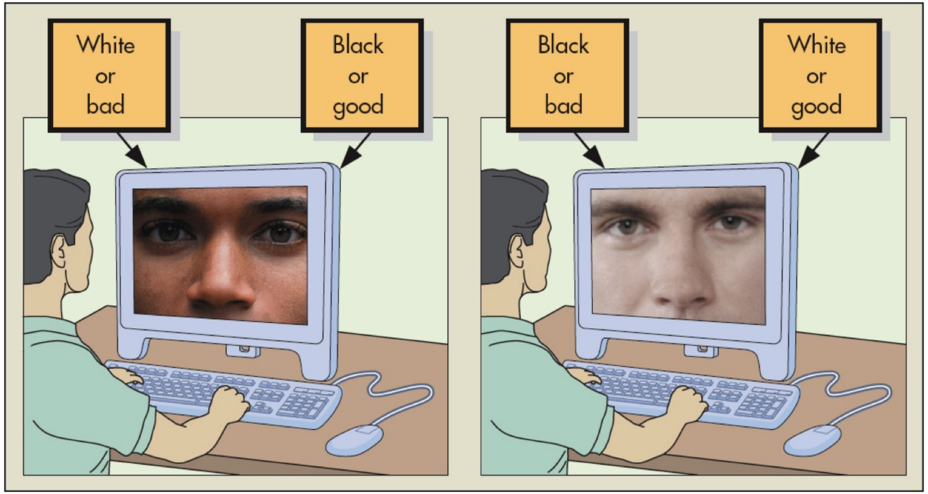
Man is to Computer Programmer as Woman is to Homemaker?

- Is “gender” subspace meaningful?
  - Maybe, but evidence that bias is still recoverable if you try to use it to “debias” embeddings
  - Later work has modified the subspace definition

Gonen, Hila, and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.” NAACL. 2019.
Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. "Understanding Undesirable Word Embedding Associations." ACL. 2019.

Alternative "Bias" Metric: Word Embedding Association Test (WEAT)

- Origins: Implicit Association test in psychology measures how quickly you associate unpleasant/pleasant stimuli with Black/white (African American/European American) names or faces



Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356.6334 (2017): 183-186.
Greenwald, Anthony G., Debbie E. McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: the implicit association test." *Journal of personality and social psychology* 74.6 (1998): 1464.

WEAT Formulation

- X,Y two sets of target words of equal size
  - X = {programmer, doctor}, Y = {homemaker, nurse}
- A,B the two sets of attribute words
  - A = {man, he}; B = {woman, she}

$$s(X, Y, A, B) = \sum_{x \varepsilon X} s(x, A, B) - \sum_{y \varepsilon Y} s(y, A, B)$$

Where $s(w, A, B) = mean_{a\varepsilon A}\cos(\vec{w}, \vec{a})$ - $mean_{b\varepsilon B}\cos(\vec{w}, \vec{b})$

Paper results

- Using WEAT metrics, bias in embeddings replicates bias found in humans using IAT

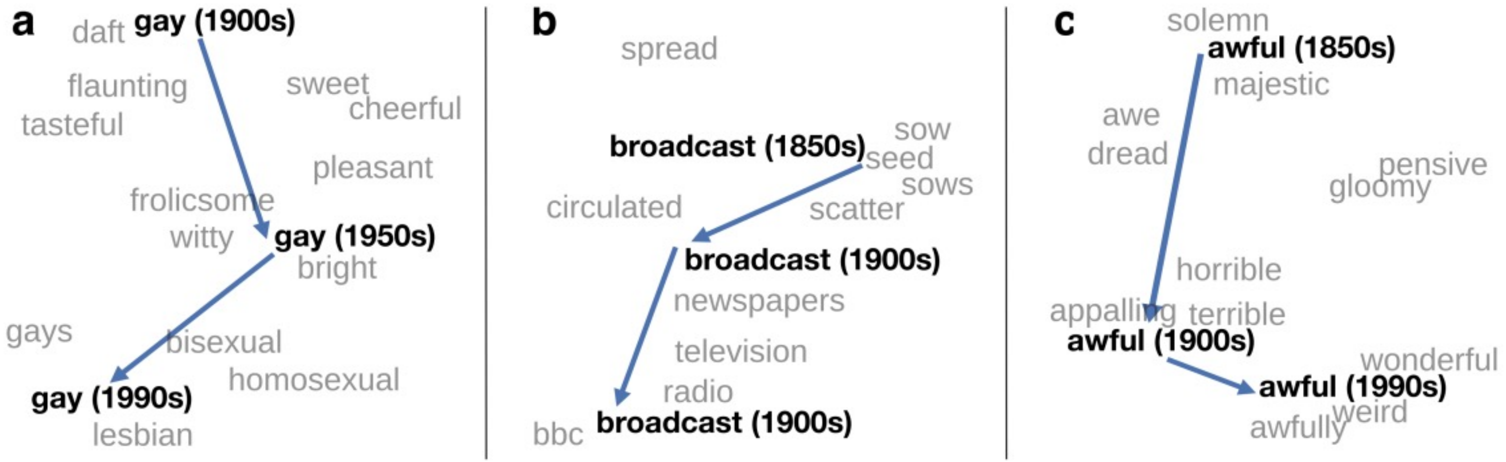
Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	(*5*)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(*5*)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(*5*)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (*5*)	(*7*)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (*9*)	(*7*)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(*9*)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(*9*)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(*10*)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(*23*)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(*9*)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Diachronic Embeddings

Diachronic Embeddings (Sociolinguistics)

- Core question in understanding cultural and language evolution: how do words change meaning over time?



Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

Compute word2vec embeddings for large text corpora divided by decade

Name	Language	Description	Tokens	Years	POS Source
ENGALL	English	Google books (all genres)	8.5×10^{11}	1800-1999	(Davies, 2010)
ENGFIC	English	Fiction from Google books	7.5×10^{10}	1800-1999	(Davies, 2010)
COHA	English	Genre-balanced sample	4.1×10^{8}	1810-2009	(Davies, 2010)
FREALL	French	Google books (all genres)	1.9×10^{11}	1800-1999	(Sagot et al., 2006)
GERALL	German	Google books (all genres)	4.3×10^{10}	1800-1999	(Schneider and Volk, 1998)
CHIALL	Chinese	Google books (all genres)	6.0×10^{10}	1950-1999	(Xue et al., 2005)

- Aggregate data by decades
- Train word embeddings on each decade (skip-gram with negative sampling)
  - Problem! Embedding spaces are not aligned!

Problem: Embedding spaces are not aligned

- Training is a stochastic process conducted on different data sets
    - Our optimization function is about relationship between vectors, not exact values
- We expect relationships between embeddings to be similar for most words (in different decades) but exact learned embedding space may differ

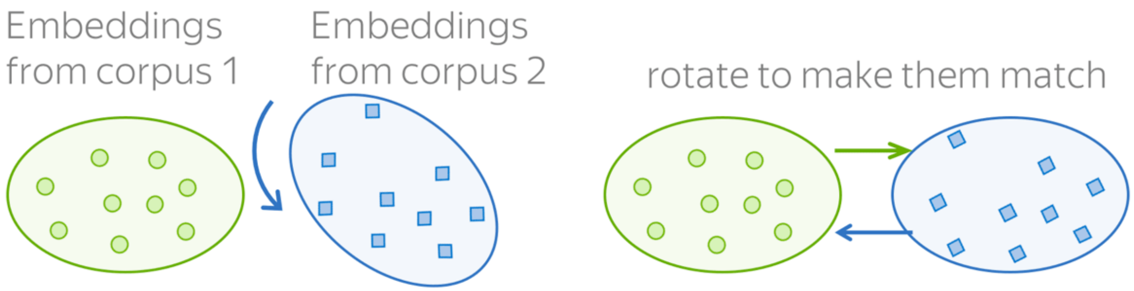


Image credit: https://lena-voita.github.io/nlp_course/word_embeddings.html

Procrustes Alignment Method

Define W_t as the VxD matrix of embeddings for decade/time t.
[V=vocabulary size, D=embedding size]
To align W_{t+1} to W_t, we solve:

$$argmin_{\boxed{Q^T Q=I}} \boxed{||W_{t+1}Q} - W_t||_F$$

$$||A||_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

Constrain that relations between embeddings are preserved in transformation

Find a transformation of W_{t+1}

"Frobenius norm": the transformation must minimize the difference between elements of W_t and W_{t+1}

Procrustes Alignment Method

Define W_t as the VxD matrix of embeddings for decade/time t.

[V=vocabulary size, D=embedding size]

To align W_{t+1} to W_t, we solve:

$$argmin_{Q^TQ=I}||W_{t+1}Q - W_t||_F$$

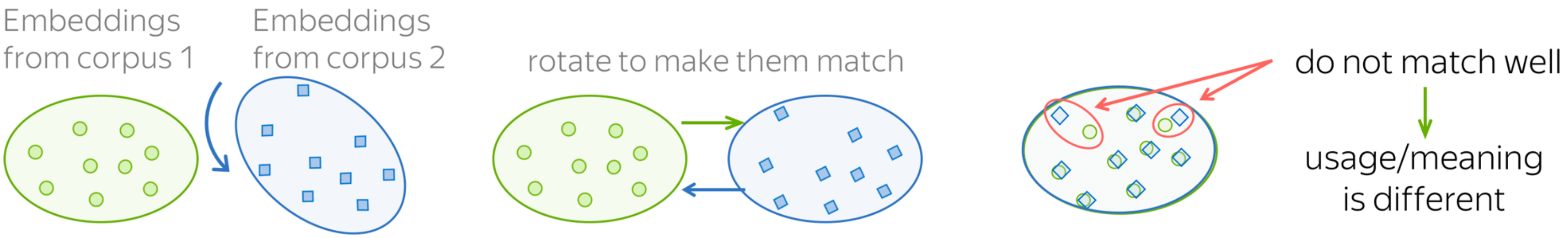
Solution:

- Compute $U\Sigma V^T$= SVD($W_{t+1}^T W_t$)
- $Q = UV^T$

Schönemann, Peter H. "A generalized solution of the orthogonal procrustes problem." *Psychometrika* 31.1 (1966): 1-10.

"**Psychometrika** is a peer-reviewed journal devoted to fostering psychology as a quantitative rational science"

Mismatches after alignment indicate semantic change



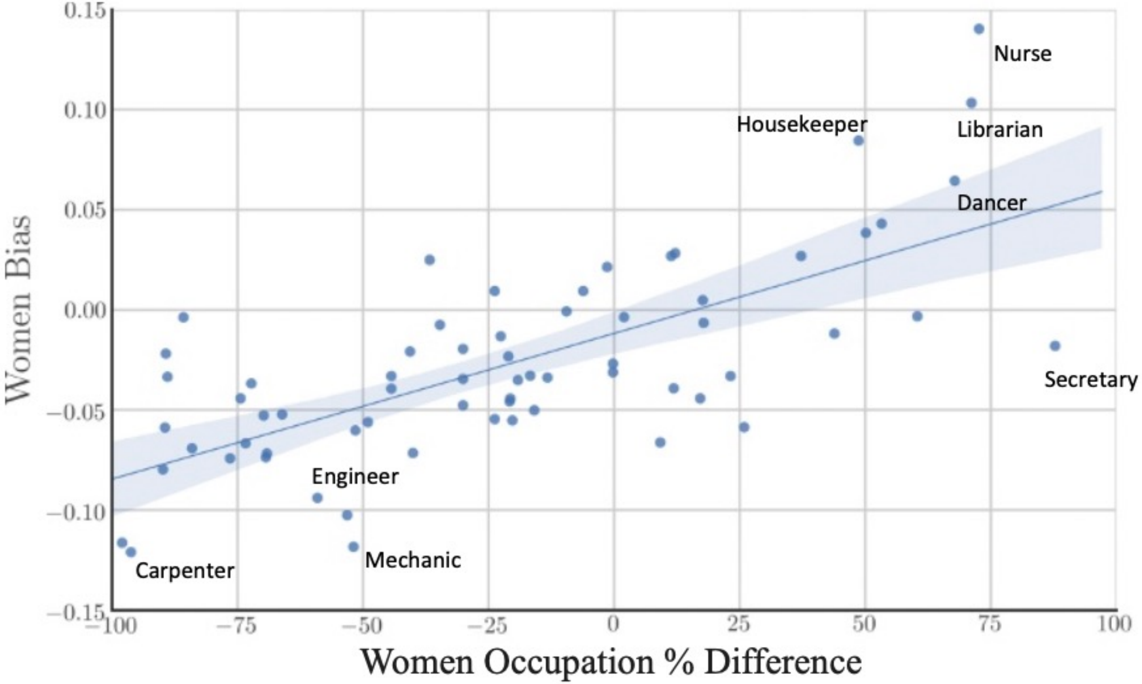
- We can compute distance between embeddings across aligned corpora
- We can also compute similarities between pairs of embeddings (e.g. [“awful”, “majestic”]; [“awful”, “terrible”] without alignment

Occupation Stereotypes over time

- Three word lists:
  - Words to representing gender
  - Words representing ethnicity (White, Asian, Hispanic; last names)
  - Occupation and adjective words
- Methods:
  - Average vectors in gender/ethnicity group
  - Compute average Euclidean distance between each group vector and each vector in occupation/adjective words
  - Take the difference of these averages between two groups (e.g. are "men" vectors closer to "programmer" than "women" vectors?) as the "relative norm difference" or "embedding bias"

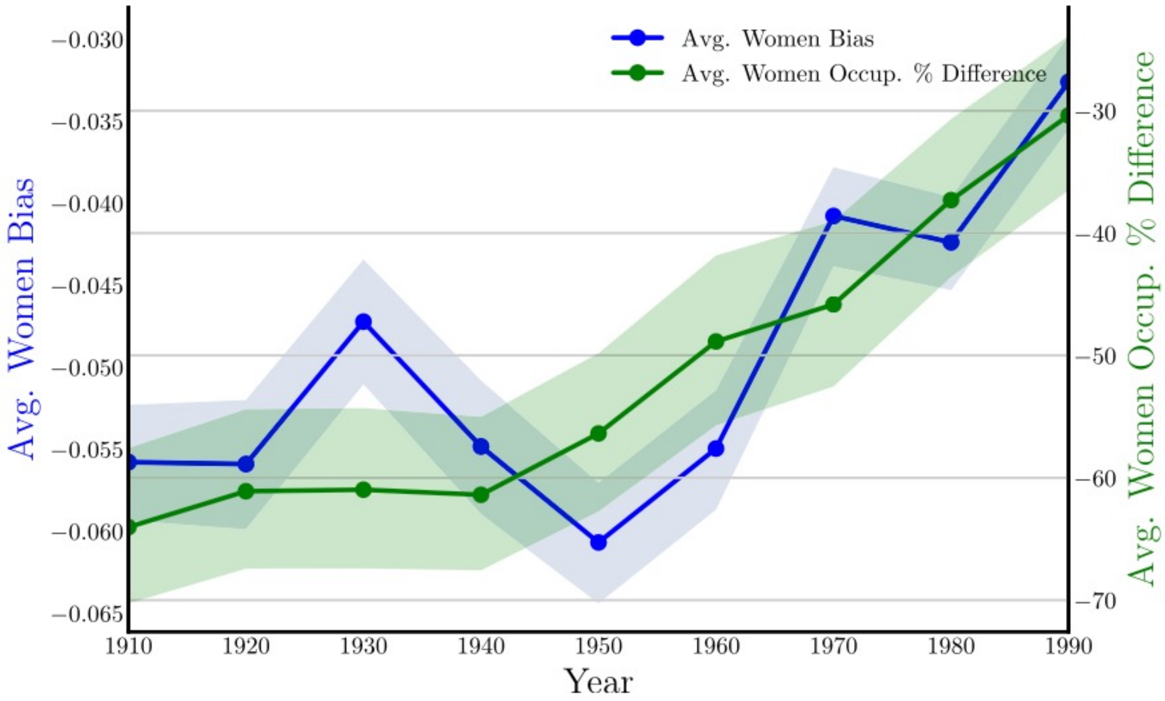
Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes." Proceedings of the National Academy of Sciences 115.16 (2018): E3635-E3644

Validation: comparison with census-reported occupations



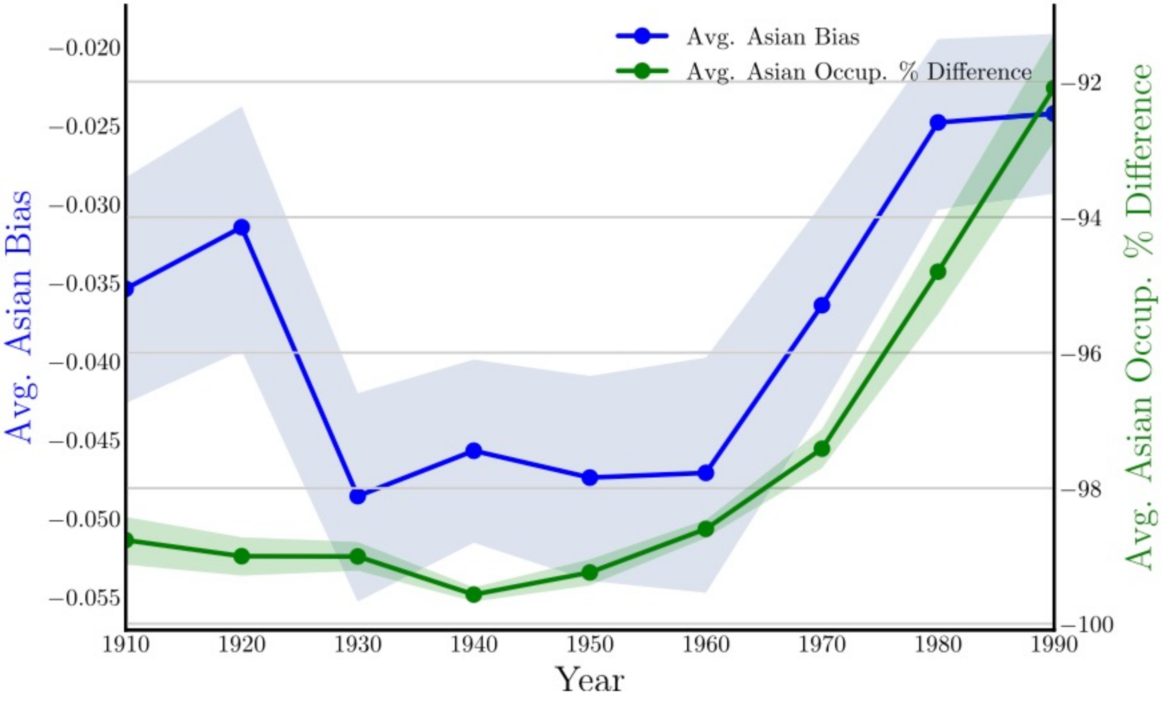
[snapshot of one decade]

Comparison with census reports over time (gender)



- Blue: bias score from embeddings (more positive indicates stronger association with women)
- Green: % of difference in women and men in the same occupations

Comparison with census reports over time (ethnicity)



JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Adjectives co-occurring with women over time

	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.71	0.68	0.71	0.65	0.67	0.54	0.45	0.45
1920	0.71	1.00	0.71	0.68	0.65	0.66	0.55	0.51	0.46
1930	0.68	0.71	1.00	0.73	0.71	0.70	0.60	0.52	0.53
1940	0.71	0.68	0.73	1.00	0.74	0.69	0.56	0.53	0.51
1950	0.65	0.65	0.71	0.74	1.00	0.71	0.58	0.51	0.49
1960	0.67	0.66	0.70	0.69	0.71	1.00	0.62	0.54	0.50
1970	0.54	0.55	0.60	0.56	0.58	0.62	1.00	0.63	0.56
1980	0.45	0.51	0.52	0.53	0.51	0.54	0.63	1.00	0.62
1990	0.45	0.46	0.53	0.51	0.49	0.50	0.56	0.62	1.00

Height of women's movements in 1960s-70s

- Study how description of women (adjectives) changed over time
- Correlations between distance between women-embeddings and adjective embeddings
- Highest correlations are between adjacent decades
- Weakest correlation is 1960s-1970s corresponding with women's movement

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Break

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Embeddings Evaluation

Evaluation

- We’re using embeddings for analyzing data sets
- How do we know that the embeddings we trained are meaningful?
- How much do decisions like embedding model (word2vec-CBOW, word2vec-skipgram, fasttext), similarity metric, or seed words (man/woman) matter?

Evaluation: Intrinsic Metrics of Embedding Quality

- Test performance on similarity; correlation between an algorithm's word similarity scores and word similarity ratings assigned by humans
  - WordSim-353 (Finkelstein et al., 2002): is ratings from 0 to 10 for 353 noun pairs; for example (plane, car) had an average score of 5.77.
  - SimLex-999 (Hill et al., 2015): more difficult dataset that quantifies similarity (cup, mug) rather than relatedness (cup, coffee), and including both concrete and abstract adjective, noun and verb pairs
  - TOEFL dataset (Landauer and Dumais, 1997): 80 questions, each consisting of a target word with 4 additional word choices; the task is to choose which is the correct synonym
- Data sets that incorporate context, such as sentence-level similarity (Huang et al., 2012; Pilehvar and Camacho-Collados, 2019)
- Analogy tasks (Turney and Littman, 2005)

Evaluation: Extrinsic Metrics of Embedding Quality

- Performance on downstream task when using embeddings in an NLP model
  - Useful for NLP models, less obviously indicative of analysis reliability

- Comparisons with external data
  - Occupation statistics from the census
  - Crowd-sourced annotations of stereotypes (note that we can crowd-source current stereotypes but it’s hard to crowd-source historical ones)

Evaluation: Capacity to capture social variables

- Do word embeddings reflect beliefs about people?
  - E.g. race and gender stereotypes
  - Dimension-level: how well do embeddings capture beliefs about gender relative to race?
  - Belief-level: how well do embeddings capture beliefs about potency (strength) of “children” vs “thugs”?

Methods

- Collect survey data from Amazon Mechanical Turk
  - Limiting assumption, how do we know if the survey data is reliable?

Joseph, Kenneth, and Jonathan Morgan. "When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?." ACL. 2020.

Evaluation: Specific Experimental Design Decisions

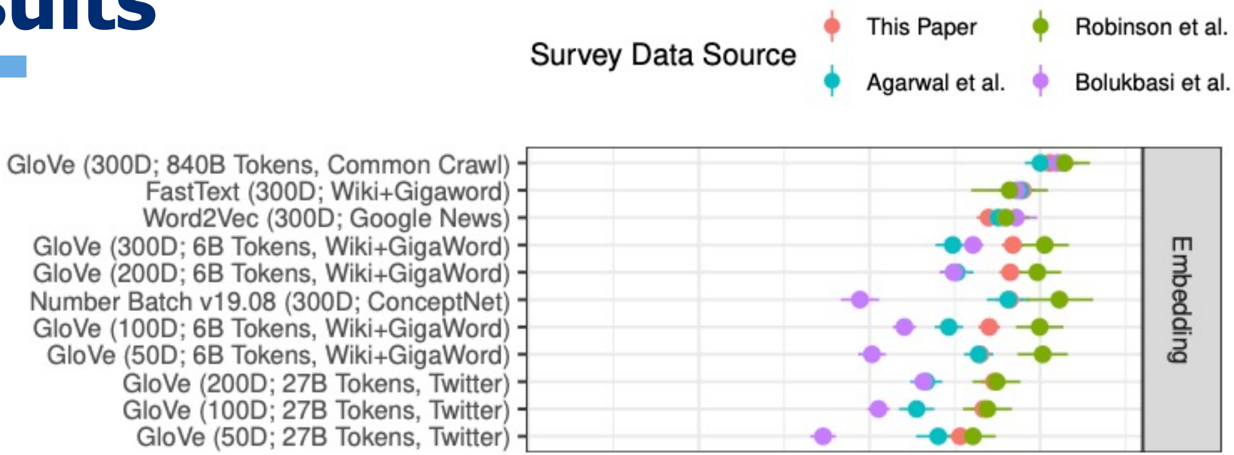
- Corpus/Embedding Selection
- Dimension Selection
  - Dimension-inducing word set
  - Methodology (average embeddings, PCA, etc)
- Word Position Measurement
  - E.g. projection, vector similarity metrics

What approaches work best? How much do these choices matter?

Design Choices

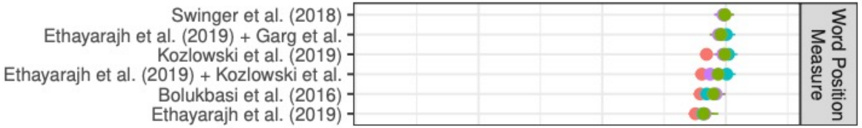
Measure	Normalized?	Position Measure	Direction-Specification	Multiclass
Ethayarajh et al. (2019)	N	$\frac{\langle w,b \rangle}{\|\|b\|\|}$	Same as Bolukbasi et al. (2016)	N
Kozlowski et al. (2019)	Y	$\frac{\langle w,b \rangle}{\|\|b\|\|\|\|w\|\|}$	$\sum_{p_i \in P} \frac{p_{i,l} - p_{i,r}}{\|\|P\|\|}$	N
Bolukbasi et al. (2016)	Y	$\frac{\langle w,b \rangle}{\|\|b\|\|\|\|w\|\|}$	$SVD\left(\mathrm{c}\left(p_{i,j} - \mu_{p_{ij}} \quad p_i \in P\right)\right)$	N
Swinger et al. (2019)	Y	$\mathrm{avg}_{p_i \in P} \frac{\langle w,p_{i,l} \rangle}{\|\|w\|\|\|\|p_{i,l}\|\|} - \mathrm{avg}_{p_i \in P} \frac{\langle w,p_{i,r} \rangle}{\|\|w\|\|\|\|p_{i,r}\|\|}$	N/A	Y
Garg et al. (2018)	Y	$\|\|w - b_r\|\| - \|\|w - b_l\|\|$	$b_l := \sum_{p_i \in p_r} \frac{p_i}{\|\|P\|\|}$	Y

Results



- [Generally embedding results do correlate with survey results]
- Selection of embedding model can decrease correlation with survey results
- Less variation for 300D embeddings
- No embedding model is universally the best

Results

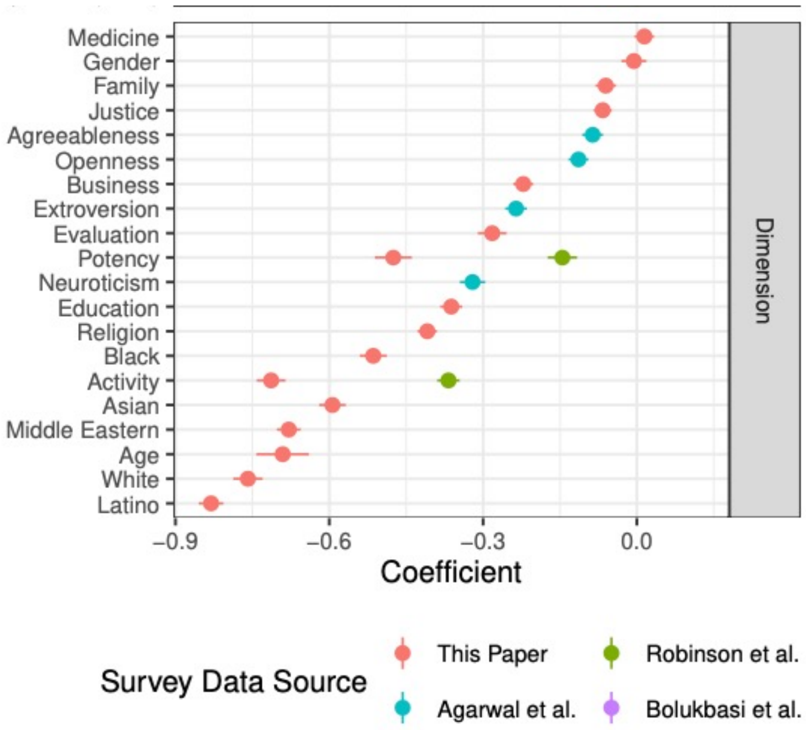


- Selection of dimension-inducing words doesn't really matter (though you could make a particularly bad choice) [Note that other work has found more variance]
- Choice of position measure (e.g. similarity metric) has almost no effect

Antoniak, Maria, and David Mimno. "Bad seeds: Evaluating lexical methods for bias measurement." ACL 2021.

Results

- Correlations for some dimensions (e.g. gender) are much stronger than for others (e.g. race)!



Recap

- Example applications:
  - Measuring bias (gender bias / occupational stereotypes)
  - Measuring change in word meanings over time
  - Measuring stereotypes over time
- Embedding manipulation:
  - Cosine similarity, Euclidean distance
  - Gender subspace
  - Averaging keywords
- Evaluations:
  - Analogy tasks, similarity benchmarks, extrinsic metrics
  - Comparisons with hand-curated analyses or benchmarks
  - Comparisons with survey or crowd-sourced data

References

- Jurafsky&Martin 6.11-6.13 https://web.stanford.edu/~jurafsky/slp3/6.pdf
- Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems* 29 (2016).
- Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *ACL*. 2016.
- Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115.16 (2018): E3635-E3644.
- Joseph, Kenneth, and Jonathan Morgan. "When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?." *ACL*. 2020.