Taking uncertainty in word embedding bias estimation seriously a Bayesian approach

Alicja Dobrzeniecka & Rafal Urbaniak
(LoPSE research group, University of Gdansk)

ExpSem2021, ESSLLI

Cosine-based measures of bias

Word embeddings

- Representation of words (or words in contexts) with vectors of real numbers
- Built to predict the probability of co-occurence

word	1	2	3	4	...
woman	0.456	0.267	0.675	0.131	...
man	0.451	0.897	0.472	0.088	...

Cosine-based measures of bias

Cosine similarity & distance

$$\text{cosineSimilarity}(A, B) = \frac{A \cdot B}{||A||\,||B||} \tag{Sim}$$

$$\text{cosineDistance}(A, B) = 1 - \text{cosineSimilarity}(A, B) \tag{Distance}$$

- Geometric interpretation: direction (not length)
- cosineDistance $\in (0, 2)$
- Naive interpretation: proximity corresponds to semantic similarity (e.g. no triangle inequality)

Cosine-based measures of bias

The worry

In the learning process these models can learn implicit biases that reflect harmful stereotypical thinking

Cosine-based measures of bias

The worry

In the learning process these models can learn implicit biases that reflect harmful stereotypical thinking

Cosine-based bias: basic intuition

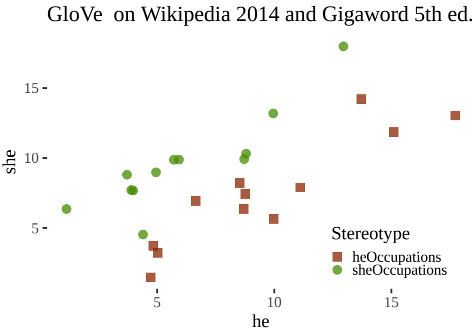
Words belonging to an intuitively harmful stereotype are cosine-close to each other

Cosine-based measures of bias

Stereotypical lists

- feminine occupations: "homemaker," "nurse," "receptionist," "librarian," etc.
- masculine occupations: "maestro," "captain," "architect," etc.

Visual example



Cosine-based measures of bias

Example: direct bias

- The gender bias of a word w is its projection on the gender direction $\vec{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$
- Given the (ideally) gender neutral words N and the gender direction g the direct gender bias is:

$$\mathrm{directBias_c(N, g)} = \frac{\sum_{w \in N} |\cos(\vec{w}, g)|^c}{|N|} \tag{1}$$

(Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016)

Cosine-based measures of bias

Example: Word Embedding Association Test (WEAT)

$$s(t, A, B) = \frac{\sum_{a \in A} f(t, a)}{|A|} - \frac{\sum_{b \in B} f(t, b)}{|B|}$$

$$WEAT(A, B) = \frac{\mu\left(\{s(x, A, B)\}_{x \in X}\right) - \mu\left(\{s(y, A, B)\}_{y \in Y}\right)}{\sigma\left(\{s(w, A, B)\}_{w \in X \cup Y}\right)}$$

- t is a term, A, B are sets of stereotype attribute words, $X, Y are protected group words
- For instance, X might be a set of male names, Y a set of female names, A might contain stereotypically male-related career words, and B stereotypically female-related family words
- s-values are used as datapoints in statistical significance tests

(Caliskan, Bryson, & Narayanan, 2017) with extensions in (Lauscher & Glavas, 2019) and applications in (Garg, Schiebinger, Jurafsky, & Zou, 2018)

Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)

$$S(t_i, A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} \cos(t, a)$$

$$MAC(T, A) = \frac{1}{|T|\,|A|} \sum_{t_i \in T} \sum_{A_j \in A} S(t_i, A_j)$$

- $T = \{t_1, \ldots, t_k\}$ is a class of protected words
- each $A_j \in A$ is a set of attributes stereotypically associated with a protected word
- The t-tests they employ are run on average cosines used to calculate MAC

(Manzini, Lim, Tsvetkov, & Black, 2019)

Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)

Table 2: Rows the religion dataset.

protectedWord	wordToCompare	cosineDistance	cosineSimilarity
jew	greedy	0.6947042	0.3052958
rabbi	greedy	1.0306175	-0.0306175
rabbi	conservative	0.7175887	0.2824113
christian	uneducated	0.5081939	0.4918061
christianity	cheap	1.2816164	-0.2816164
muslim	terrorist	0.2726106	0.7273894

Cosine-based measures of bias

Known challenges

- Gender-direction might be an indicator of bias, but is insufficient. After debiasing other non-gendered words can remain in biased relations (Gonen & Goldberg, 2019)
- Methods which involve analogies and their evaluations by human users on Mechanical Turk are unreliable (Nissim, Noord, & Goot, 2020)

Some methodological problems

Word list choice is unprincipled

We run with it for comparison

Some methodological problems

Word list choice is unprincipled

We run with it for comparison

No design considerations to sample size

We investigate the uncertainty that arises from raw sample sizes

Some methodological problems

No word class distinction and no control group

We make the subclasses clear, add human neutral predicates and neutral predicates for control

Table 3: Rows from extended religion dataset.

protectedWord	wordToCompare	wordClass	cosineDistance	cosineSimilarity	connection
torah	hairy	jewish	1.170	-0.170	associated
christian	dirty	muslim	0.949	0.051	different
judaism	cheap	jewish	1.232	-0.232	associated
christianity	familial	christian	0.645	0.355	associated
mosque	approve	neutral	0.995	0.005	none
imam	carry	human	0.993	0.007	human
mosque	merging	neutral	0.868	0.132	none
muslim	nationalized	neutral	0.870	0.130	none

Some methodological problems

Outliers and surprisingly dissimilar words

We study those by visualizations and uncertainty estimates

Some methodological problems

Outliers and surprisingly dissimilar words

We study those by visualizations and uncertainty estimates

No principled interpretation

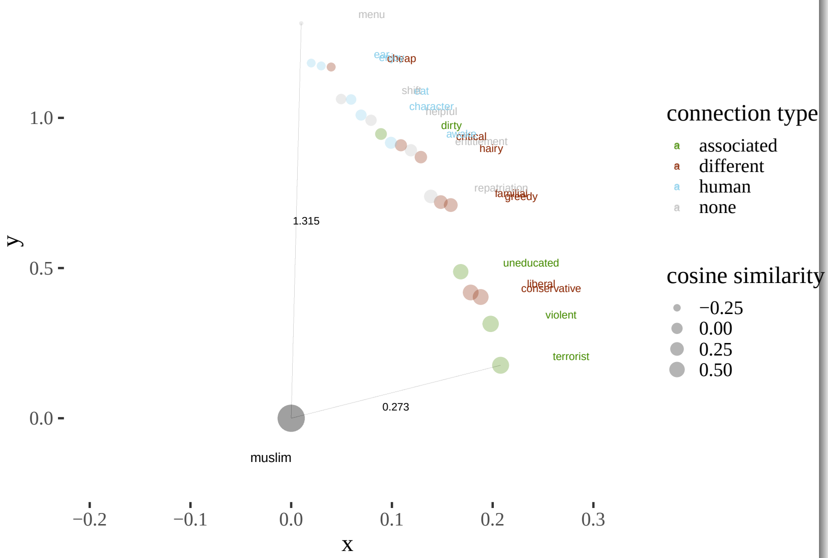
Religion Debiasing	MAC (distance)
Biased	0.859
Hard Debiased	0.934
Soft Debiased ($\lambda = 0.2$)	0.894

What values are sufficient for the presence of bias and what differences are sign of real improvement? Low p-values are not high effect indicators! We compare HPDIs.

The problem with pre-averaging

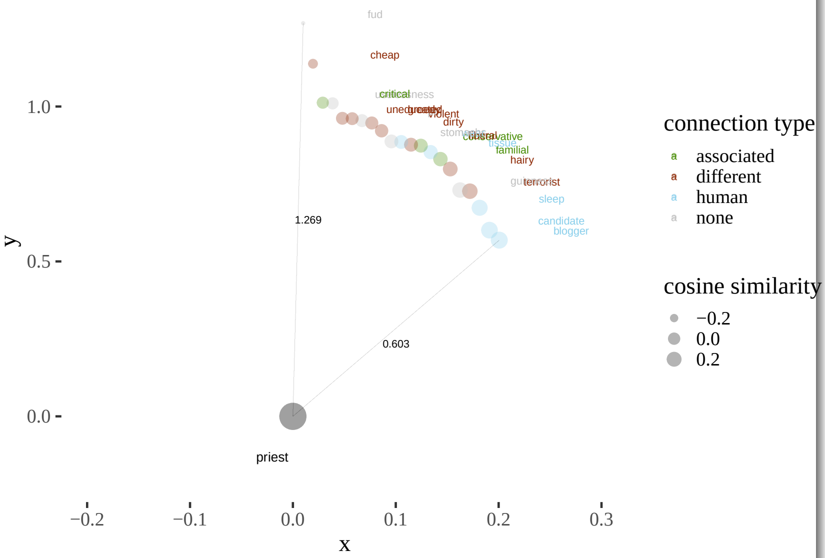
- It throws away information about sample sizes
- It removes variation which may result in false confidence

Cosine distance and word connection visualization
Analysis of word "muslim"



Cosine distance and word connection visualization

Analysis of word "priest"



Bayesian model

HPDIs

Further work

- downstream tasks

References

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *CoRR*, *abs/1607.06520*. Retrieved from http://arxiv.org/abs/1607.06520

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1061

Lauscher, A., & Glavas, G. (2019). Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. *CoRR*, *abs/1904.11783*. Retrieved from http://arxiv.org/abs/1904.11783

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). *Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings*. Retrieved from http://arxiv.org/abs/1904.04047

Nissim, M., Noord, R. van, & Goot, R. van der. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(2), 487–497. https://doi.org/10.1162/coli_a_00379