Taking uncertainty in word embedding bias estimation seriously: a bayesian approach

Alicja Dobrzeniecka and Rafal Urbaniak

1 Cosine-based measures of bias

fix bibliography

Modern Natural Language Processing (NLP) models are used to complete various tasks such as providing email filters, smart assistants, search results, language translations, text analytics and so on. All of them need as an input words represented by means of numbers which is accomplished with word embeddings. It seems that in the learning process these models can learn implicit biases that reflect harmful stereotypical thinking. One of the sources of bias in NLP can be located in the way the word embeddings are made. There is a considerable amount of literature available on the topic of bias detection and mitigation in NLP models.

One of the first measures in the discussion has been developed by Bolukbasi, Chang, Zou, Saligrama, & Kalai (2016). There, the gender bias of a word w is understood as its projection on the gender direction $\vec{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$ (the gender direction is the top principal compontent of ten gender pair difference vectors). The underlying idea is that no bias is present if non-explicitly gendered words are in equal distance to both elements in all explicitly gender pairs. Given the (ideally) gender neutral words N and the gender direction g the direct gender bias is defined as the average distance of the words in N from g (c is a parameter determining how strict we want to be):

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

The use of projections has been ciriticized for instance by Gonen & Goldberg (2019), who point out that while gender-direction might be an indicator of bias, it is only one possible manifestation of it, and reducing a projection of words might be insufficient. For instance, "math" and "delicate" might be in equal distance to both explicitly gendered words while being closer to quite different stereotypical attribute words. Further, the authors point out that most word pairs preserve similarity under debiasing meant to minimize projection-based bias.[1]

To measure bias in word embeddings, Caliskan, Bryson, & Narayanan (2017) proposed the Word Embedding Association Test (WEAT). The idea is that the measure of biases between two sets of target words, X and Y, (we call them protected words) should be quantified in terms of the cosine similarity between the protected words and attribute words coming from two sets of stereotype attribute words, A and B (we'll call them attributes). 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. WEAT is a modification of the Implicit Association Test (IAT) (Nosek, Banaji, & Greenwald, 2002) used in psychology and uses almost the same word sets, allowing for a *prima facie* sensible comparison with bias in humans. If the person's attitude towards given pair of concept is to be interpreted as neutral, there should be no noticeable task completion time difference, and the final value from the formula should be around 0. Let f be a similarity measure (usually, cosine similarity). The association difference for a term t is:

$$s(t,A,B) = \frac{\sum_{a \in A} f(t,a)}{|A|} - \frac{\sum_{b \in B} f(t,b)}{|B|} \tag{2}$$

[1] Bolukbasi et al. (2016) use also another method which involves analogies and their evaluations by human users on Mechanical Turk. It is discussed and criticized in (Nissim, Noord, & Goot, 2020).

then, the association difference between A a B is:

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B) \tag{3}$$

$s(X,Y,A,B)$ is the statistic used in the significance test, and the p-value obtained by bootstrapping: it is the frequency of $s(X_i,Y_i,A,B) > s(X,Y,A,B)$ for all equally sized partitions X_i, Y_i of $X \cup Y$. The effect size is computed by normalizing the difference in means as follows:

$$bias(A,B) = \frac{\mu(\{s(x,A,B)\}_{x \in X}) - \mu(\{s(y,A,B)\}_{y \in Y})}{\sigma(\{s(w,A,B)\}_{w \in X \cup Y})} \tag{4}$$

Caliskan et al. (2017) show that significant biases—thus measured— similar to the ones discovered by IAT can be discovered in word embeddings. Lauscher & Glavas (2019) extended the methodology to a multilingual and cross-lingual setting, arguing that using Euclidean distance instead of similarity does not make much difference, while the bias effects vary greatly across embedding models (interestingly, with social media-text trained embeddings being less biased than those based on Wikipedia).

A similar methodology is employed by Garg, Schiebinger, Jurafsky, & Zou (2018), who employ word embeddings trained on corpora from different decades to study the shifts in various biases. For instance, to compute the occupational embeddings bias for women the authors first compute the average vector of vector embeddings of words that represent women (e.g. "she," "female"), then calculate the Euclidean distance between this mean vector and words for occupations. Then they take the mean of these distances and subtract from it the analogously obtained mean for the average vector of vector embeddings of words that represent men. Formally they take the relative norm distance between X and Y to be:

$$\text{relative norm distance} = \sum_{v_m \in M} ||v_m - v_X||_2 - |v_m - v_Y||_2 \tag{5}$$

where the norm used is Euclidean, and v_X and x_Y are average vectors for sets X and Y respectively.

Manzini, Lim, Tsvetkov, & Black (2019) modify WEAT to a multi-class setting, introducing Mean Average Cosine similarity as a measure of bias (in fact, in the paper they report distances rather than similarities). Let $T = \{t_1, \ldots, t_k\}$ be a class of protected word embeddings, and let each $A_j \in A$ be a set of attributes stereotypically associated with a protected word). Then:

$$S(t_i,A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} \cos(t,a) \tag{6}$$

$$MAC(T,A) = \frac{1}{|T|\,|A|} \sum_{t_i \in T} \sum_{A_j \in A} S(t_i,A_j) \tag{7}$$

That is, for each protected word $T and each attribute class, they first take the mean for this protected word and all attributes in a given attribute class, and then take the mean of thus obtained means for all the protected words. The t-tests they employ are run on average cosines used to calculate MAC.

References

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

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101