Literature Review

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

Man is to Computer Programmer is one of the first and seminal works in the area of debiasing word vectors. The researchers define direct bias as the sum of the absolute cosine similarity between the gender neutral words and a learned gender direction g; g is roughly learned by combining several distances such as the differences between the vectors of she and he, woman and man, etc. Indirect bias is measured roughly by measuring the contribution of gender to each word vector; the gender component is defined as the similarity between two word vectors. To debias the pairs that should be gender neutral, the researchers subtracted the projection onto the gender direction of these gender neutral word pairs.

We choose to implement the work done in this paper since it serves as a good baseline to compare both other work in this area and our extensions to. Furthermore, the methodology used in this paper is relatively straightforward, so implementing the work done here will serve as a good stepping stone to extending their research.

Semantic Derived Automatically From Language Corpora Contain Human-Like Biases

In this paper, researchers define another metric called Word-Embedding Association Test (WEAT) to measure bias in word embeddings (specifically GloVe). The researchers use cosine similarity between a pair of vectors as analogous to reaction time in the Implicit Association Test (IAT). The researchers find that there were implicit gender-occupation biases, and suggest that there are historic biases contained within text corpora. The researchers highlight the importance of understanding bias and prejudice in making algorithms and other technologies that can impact the lives of others.

Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes

In this paper, researchers analyze different word embeddings, including word2vec and COHA, trained over 100 years of text corpora to understand how gender and ethnic stereotypes have changed over time. To do so, they measure the strength of the association between neutral words and a group. Their metric for measuring bias is the relative norm difference, which is computed by first finding the representative word vector by taking the average of the vectors for each word in the given gender / ethnic group, and then the average Euclidean distance between each representative group vector and each vector of the neutral words. Ordinary least-squares regression is used to measure associations. The researchers found that bias generally moved in the direction they expected; for instance, the researchers found that gender bias in occupations moved toward 0 from the 1950s to the 1990s, which is in line with the gender movement.

Lipstick on a Pig

Hila Gonen and Yoav Goldberg show that the work done in Man is to Computer Programmer and Mitigating Unwanted Biases with Adversarial Learning only superficially removes the debiasing, as the "bias ... is ingrained much more deeply in the embeddings space." Gonen and Goldberg do so by clustering the most biased words in the original vocabulary using k-means, and find that the clusters still align with gender with a high accuracy. Thus, Gonen and Goldberg posit that existing metrics that have been explored in research may be good at measuring bias, but are not necessarily effective as a baseline for comparison for debiasing. We include this paper since it highlights some of the shortcomings in current debiasing research.

Works Cited

1. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., & Kalai, A.T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS.
2. Caliskan, A., Bryson, J.J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356, 183-186.
3. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J.Y. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceedings of the National Academy of Sciences of the United States of America, 115 16, E3635-E3644.
4. Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. CoRR, abs/1903.03862.