On Measuring Social Biases in Sentence Encoders

Chandler May[1] **Alex Wang**[2] **Shikha Bordia**[2]
Samuel R. Bowman[2] **Rachel Rudinger**[1]
[1]Johns Hopkins University [2]New York University
{cjmay,rudinger}@jhu.edu {alexwang,sb6416,bowman}@nyu.edu

Abstract

The *Word Embedding Association Test* shows that GloVe and word2vec word embeddings exhibit human-like implicit biases based on gender, race, and other social constructs (Caliskan et al., 2017). Meanwhile, research on learning reusable text representations has begun to explore sentence-level texts, with some sentence encoders seeing enthusiastic adoption. Accordingly, we extend the Word Embedding Association Test to measure bias in sentence encoders. We then test several sentence encoders, including state-of-the-art methods such as ELMo and BERT, for the social biases studied in prior work and two important biases that are difficult or impossible to test at the word level. We observe mixed results including suspicious patterns of sensitivity that suggest the test's assumptions may not hold in general. We conclude by proposing directions for future work on measuring bias in sentence encoders.

1 Introduction

Word embeddings quickly achieved wide adoption in natural language processing (NLP), precipitating the development of efficient, word-level neural models of human language. However, prominent word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) encode systematic biases against women and black people (Bolukbasi et al., 2016; Garg et al., 2018, i.a.), implicating many NLP systems in scaling up social injustice. We investigate whether sentence encoders, which extend the word embedding approach to sentences, are similarly biased.[1]

The previously developed Word Embedding Association Test (WEAT; Caliskan et al., 2017) measures bias in word embeddings by comparing two sets of target-concept words to two sets of attribute words. We propose a simple generalization of WEAT to phrases and sentences: the Sentence Encoder Association Test (SEAT). We apply SEAT to sentences generated by inserting individual words from Caliskan et al.'s tests into simple templates such as "This is a[n] <word>."

To demonstrate the new potential of a sentence-level approach and advance the discourse on bias in NLP, we also introduce tests of two biases that are less amenable to word-level representation: the *angry black woman* stereotype (Collins, 2004; Madison, 2009; Harris-Perry, 2011; hooks, 2015; Gillespie, 2016) and a *double bind* on women in professional settings (Heilman et al., 2004).

The use of sentence-level contexts also facilitates testing the impact of different experimental designs. For example, several of Caliskan et al.'s tests rely on given names associated with European American and African American people or rely on terms referring to women and men as groups (such as "woman" and "man"). We explore the effect of using given names versus group terms by creating alternate versions of several bias tests that swap the two. This is not generally feasible with WEAT, as categories like *African Americans* lack common single-word group terms.

We find varying evidence of human-like bias in sentence encoders using SEAT. Sentence-to-vector encoders largely exhibit the angry black woman stereotype and Caliskan biases, and to a lesser degree the double bind biases. Recent sentence encoders such as BERT (Devlin et al., 2018) display limited evidence of the tested biases. However, while SEAT can confirm the existence of bias, negative results do not indicate the model is bias-free. Furthermore, discrepancies in the results suggest that the confirmed biases may not generalize beyond the specific words and sentences in our test data, and in particular that cosine similarity may not be a suitable measure of representational similarity in recent models, indicating a need for alternate bias detection techniques.

[1] While encoder training data may contain perspectives from outside the U.S., we focus on biases in U.S. contexts.

Target Concepts	**Attributes**
European American names: Adam, Harry, Nancy, Ellen, Alan, Paul, Katie, …	*Pleasant*: love, cheer, miracle, peace, friend, happy, …
African American names: Jamel, Lavar, Lavon, Tia, Latisha, Malika, …	*Unpleasant*: ugly, evil, abuse, murder, assault, rotten, …

Table 1: Subsets of target concepts and attributes from Caliskan Test 3. Concept and attribute names are in italics. The test compares the strength of association between the two target concepts and two attributes, where all four are represented as sets of words.

Target Concepts	**Attributes**
European American names: “This is Katie.”, “This is Adam.” “Adam is there.”, …	*Pleasant*: “There is love.”, “That is happy.”, “This is a friend.”, …
African American names: “Jamel is here.”, “That is Tia.”, “Tia is a person.”, …	*Unpleasant*: “This is evil.”, “They are evil.”, “That can kill.”, …

Table 2: Subsets of target concepts and attributes from the bleached sentence version of Caliskan Test 3.

2 Methods

The Word Embedding Association Test WEAT imitates the human implicit association test (Greenwald et al., 1998) for word embeddings, measuring the association between two sets of target concepts and two sets of attributes. Let X and Y be equal-size sets of target concept embeddings and let A and B be sets of attribute embeddings. The test statistic is a difference between sums over the respective target concepts,

$$s(X, Y, A, B) = \left[\textstyle\sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)\right],$$

where each addend is the difference between mean cosine similarities of the respective attributes,

$$s(w, A, B) = \left[\mathrm{mean}_{a \in A} \cos(w, a) - \mathrm{mean}_{b \in B} \cos(w, b)\right]$$

A permutation test on $s(X, Y, A, B)$ is used to compute the significance of the association between (A, B) and (X, Y),

$$p = \Pr\left[s(X_i, Y_i, A, B) > s(X, Y, A, B)\right],$$

where the probability is computed over the space of partitions (X_i, Y_i) of $X \cup Y$ such that X_i and Y_i are of equal size, and a normalized difference of means of $s(w, A, B)$ is used to measure the magnitude of the association (the effect size; Caliskan et al., 2017),

$$d = \frac{\mathrm{mean}_{x \in X} s(x, A, B) - \mathrm{mean}_{y \in Y} s(y, A, B)}{\mathrm{std\_dev}_{w \in X \cup Y} s(w, A, B)}.$$

Controlling for significance, a larger effect size reflects a more severe bias. We detail our implementations in the supplement.

The Sentence Encoder Association Test SEAT compares sets of sentences, rather than sets of words, by applying WEAT to the vector representation of a sentence. Because SEAT operates on fixed-sized vectors and some encoders produce variable-length vector sequences, we use pooling as needed to aggregate outputs into a fixed-sized vector. We can view WEAT as a special case of SEAT in which the sentence is a single word. In fact, the original WEAT tests have been run on the Universal Sentence Encoder (Cer et al., 2018).

To extend a word-level test to sentence contexts, we slot each word into each of several semantically bleached sentence templates such as “This is <word>.”, “<word> is here.”, “This will <word>.”, and “<word> are things.”. These templates make heavy use of deixis and are designed to convey little specific meaning beyond that of the terms inserted into them.[2] For example, the word version of Caliskan Test 3 is illustrated in Table 1 and the sentence version is illustrated in Table 2. We choose this design to focus on the associations a sentence encoder makes with a given term rather than those it happens to make with the contexts of that term that are prevalent in the training data; a similar design was used in a recent sentiment analysis evaluation corpus stratified by race and gender (Kiritchenko and Mohammad, 2018). To facilitate future work, we publicly release code for SEAT and all of our experiments.[3]

3 Biases Tested

Caliskan Tests We first test whether the sentence encoders reproduce the same biases that word embedding models exhibited in Caliskan et al. (2017). These biases correspond to past social psychology studies of implicit associations in human subjects.[4] We apply both the original

[2] See the supplement for further details and examples.

[3] http://github.com/W4ngatang/sent-bias

[4] See Greenwald et al. (2009) for a review of this work.