Capturing Implicit Biases With Positive Operators

J. Bosscher[1,2], *Supervisors:* dr. M.A.F. Lewis[1,2], and dr. K. Schulz[1,2]

[1] University of Amsterdam, Science Park 904, 1012 WX Amsterdam, The Netherlands
[2] Institute for Logic, Language and Computation, Science Park 107, 1098 XG Amsterdam, The Netherlands
j.m.bosscher@uva.nl

Keywords: positive operator · hyponymy · graded hyponymy · distributional semantics · implicit bias · bias · stereotypes · word embeddings

1 Introduction

Modelling words as vectors has been an extremely successful way of representing word meaning in a manner that can be programmed into a computer [5] [7][9]. Word vectors have been shown to be affected by the ideas and beliefs of the humans that generated the corpora they are extracted from [1] [10]. Caliskan et. al. [2] showed that our biases and stereotypical beliefs could also be extracted from these representations. In order to identify implicit biases held by human speakers, the results from a physiological test called the Implicit Association Test (IAT) [4] are used as a benchmark. Caliskan et. al. [2] then compared these results to that of their Natural Language Processing (NLP) version of the IAT called the Word Embedding Association Test (WEAT). Utilizing widely adopted distributional models, mainly focusing on GloVe embeddings [9], they were able to replicate every association documented by the IAT that they tested. This leads them to expect that human biases are in general retrievable from statistical properties of language use.

However, the two test methods used to extract implicit biases from human speakers on the one hand and from corpora on the other differ clearly in methodology. The IAT uses a categorization task between a target concept and attribute. The WEAT uses a similarity measure to test for bias in the corpus. We investigate whether the same biases are present when using a representation for words that allows us to model categorization.

2 Method

In order to use this measure of graded hyponymy we must represent the meaning of words as a collection of their hyponyms. Here we used two different sources of hyponymy: WordNet [3] and Microsoft's Concept Graph [8]. We can then construct the representation of a word by adding together all the positive operators

of a specific word. We use positive operators because they have an ordering to them called the Löwner ordering which can be interpreted as categorization. To build the representations of single vector positive operators we take the outer product of the vector representation of a word with itself, more specifically we use use GloVe embeddings [9] as the source of word vectors in line with Caliskan et. al. [2]. By using this representation we can now define graded categorization using two methods described in Lewis [6] in terms of graded hyponymy (K_E and K_{BA}). The WEAT tests for implicit biases by comparing the similarity of two sets of comparable target words (*e.g., female names vs. male names*) against two sets of opposing attribute words (*e.g., pleasant attributes vs. unpleasant attributes*) with a null hypotheses that states that there is no difference between the similarity of either set of target words with the target concepts. We followed Caliskan et. al. [2] by using the same method but instead of measuring the association in terms of similarity, we measured the association in terms of categorization. The resulting test method for implicit biases in corpora is called the Positive Operator Association Test (POAT).

3 Results

Table 1 shows that the POAT is able to replicate most of the same results of the WEAT. The POAT performs well on non-offensive experiments such as the differential association between *flowers vs. insects* and *pleasant vs. unpleasant.* Additionally, the POAT records stronger stereotyping in the association tested in the last two rows, as well as in the *European-American vs. African-American – pleasant vs. unpleasant* experiment (row 5, Table 1) when using the attributes from the *young vs. old people's names* experiment. Two experiments that were not replicated well by the POAT are those in row 7 and 8. The first one instead shows a reversed association, due to inconsistent hyponymy representation, and the second records a low effect size and a high likelihood of the null hypotheses holding up.

Some target words had very low numbers of hyponyms, which skewed the results. To alleviate this problem, which was the case for most experiments that performed poorly on the POAT in Table 1, we use the same measure as in the regular POAT, but build the positive operators from the single word embeddings for each specific word. The results for these tests are presented in Table 2 and the largest difference in the rows 7 and 8: both now show large positive effect sizes and slightly smaller p-values compared to those of the WEAT. This version of the POAT performs best on all tested IAT findings. The found effect sizes are closer to the IAT effect sizes in 6 out of 8 experiments compared to the WEAT.

Discussion and outlook

In nine out of ten experiments the POAT is able to correctly recognize implicit biases in the word embeddings. Although the POAT was able to recognize the implicit biases, the strength that it recorded was sometimes not comparable to

	Target words	Attribute words			IAT		WEAT		POAT	
			N_T	N_A	d	P	d	P	d	P
1	Flowers vs. insects	Pleasant vs. unpleasant	25×2	25×2	1.35	10^{-8}	1.50	10^{-7}	**1.39**	10^{-6}
2	Musical instruments vs. weapons	Pleasant vs. unpleasant	25×2	25×2	1.66	10^{-10}	**1.53**	10^{-7}	1.47	10^{-7}
3	European-American vs. African-American	Pleasant vs. unpleasant	32×2	25×2	1.17	10^{-5}	**1.41**	10^{-8}	0.89	10^{-3}
4	European-American vs. African-American	Pleasant vs. unpleasant †	16×2	25×2	–	–	1.50	10^{-4}	1.04	10^{-2}
5	European-American vs. African-American	Pleasant vs. unpleasant ‡	16×2	8×2	–	–	1.28	10^{-3}	1.58	10^{-5}
6	Male vs. female names	Career vs. family	8×2	8×2	0.72	$< 10^{-2}$	1.81	10^{-3}	**1.68**	10^{-3}
7	Mental vs. physical disease	Temporary vs. permanent	6×2	7×2	1.01	10^{-2}	**1.38**	10^{-2}	−1.51	10^{-2}
8	Science vs. arts	Male vs. female	8×2	8×2	1.47	10^{-24}	**1.24**	10^{-2}	−0.001	0.50
9	Math vs. arts	Male vs. female	8×2	8×2	0.82	$< 10^{-2}$	**1.06**	10^{-1}	1.25	10^{-2}
10	Young vs. old people's names	Pleasant vs. unpleasant	8×2	8×2	1.42	$< 10^{-2}$	1.21	10^{-2}	**1.29**	10^{-2}

Table 1: Effect size (Cohen's d) and p-values for the WEAT and the POAT using the K_E measure and hyponyms derived from WordNet. Each row concerns a different implicit bias documented by the IAT. In each case the first (second) set of target words is found to be more compatible with the first (second) set of attributes words, N_T and N_A indicated the number of target words and attribute words, respectively. Bold values highlight the effect size closest to that of the IAT. † Attributes for this experiment are the same as in *Flowers vs. insects*. ‡Attributes for this experiment are the same as in *Young vs. old people's names*.

	Target words	Attribute words			IAT		WEAT		POAT	
			N_T	N_A	d	P	d	P	d	P
1	Flowers vs. insects	Pleasant vs. unpleasant	25×2	25×2	1.35	10^{-8}	1.50	10^{-7}	**1.30**	10^{-5}
2	Musical instruments vs. weapons	Pleasant vs. unpleasant	25×2	25×2	1.66	10^{-10}	**1.53**	10^{-7}	1.30	10^{-5}
3	European-American vs. African-American	Pleasant vs. unpleasant	32×2	25×2	1.17	10^{-5}	1.41	10^{-8}	**1.29**	10^{-6}
4	European-American vs. African-American	Pleasant vs. unpleasant †	16×2	25×2	–	–	1.50	10^{-4}	1.18	10^{-3}
5	European-American vs. African-American	Pleasant vs. unpleasant ‡	16×2	8×2	–	–	1.28	10^{-3}	1.46	10^{-4}
6	Male vs. female names	Career vs. family	8×2	8×2	0.72	$< 10^{-2}$	1.81	10^{-3}	**1.74**	10^{-3}
7	Mental vs. physical disease	Temporary vs. permanent	6×2	7×2	1.01	10^{-2}	1.38	10^{-2}	**1.26**	10^{-1}
8	Science vs. arts	Male vs. female	8×2	8×2	1.47	10^{-24}	**1.24**	10^{-2}	1.06	10^{-1}
9	Math vs. arts	Male vs. female	8×2	8×2	0.82	$< 10^{-2}$	1.06	10^{-1}	**1.00**	10^{-1}
10	Young vs. old people's names	Pleasant vs. unpleasant	8×2	8×2	1.42	$< 10^{-2}$	1.21	10^{-2}	**1.52**	10^{-2}

Table 2: This table shows the effect size (Cohen's d) and p-values for the WEAT and the POAT using the K_E measure, represented without hyponyms. All other settings are identical to those shown in Table 1

that of the WEAT or IAT due to inconsistent hyponymy representations of the target and concept words (row 7, Table 1). Neither sources of hyponymy we used contained an entry for every word. Nor do they contain all hyponyms of a word and in several cases the entry has zero hyponyms. Therefore, in order to make this method as dependable as possible the problematic word categories must be identified and remedied with some other method of deriving hyponyms. An example word type that for which this issue was prominent are the male and female pronouns that were part of the IAT experiments.

An advantage we found of our approach to detect biases in word meanings of the WEAT is that positive operators fit well inside a compositional framework [6]. This allows us to form phrases and sentences as well as generic sentences. Generic sentences such as "mosquitos carry malaria" express regularities. Using

positive operators gives the potential to assess associations between words and subphrases, such as mosquitos and carry malaria.

Our results shows that the use of the K_E measure as a proxy on single vector positive operators, where the words representations are not built using their hyponyms, outperforms the WEAT on six out of eight replications of IAT findings. This indicates that the use of an asymmetric measure to determine differential association is better at detecting implicit bias in word embeddings than the symmetric distance measure of the WEAT. The next step should be to identify exactly why the POAT performs so well on single vector positive operators.

References

1. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior research methods **39**(3), 510–526 (2007)
2. Caliskan, A., Bryson, J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (4 2017). https://doi.org/10.1126/science.aal4230
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
4. Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology **74**(6), 1464 (1998)
5. Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review **104**(2), 211 (1997)
6. Lewis, M.: Compositional hyponymy with positive operators. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 638–647. INCOMA Ltd., Varna, Bulgaria (Sep 2019). https://doi.org/10.26615/978-954-452-056-4-075, https://www.aclweb.org/anthology/R19-1075
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
8. Nsl, D.I.: Microsoft concept graph: Mining semantic concepts for short text understanding **1**, 262–294 (11 2019)
9. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://www.aclweb.org/anthology/D14-1162
10. Stubbs, M.: Text and corpus analysis: Computer-assisted studies of language and culture. Blackwell Oxford (1996)