Does Double-Hard Debias Keep its Promises? - A Replication Study

Kristina Kobrock[1], Meike Korsten[1], & Sonja Börgerding[1]

[1] University of Osnabrück

Does Double-Hard Debias Keep its Promises? - A Replication Study

Just edit the text! There are several websites that help with (r)markdown, e.g. https://rmarkdown.rstudio.com/. If you don't want to edit directly, but rather comment use: Cite using: Wang et al. (2020) said that… or simply blablabla (Wang et al., 2020). References must be added to r-references.bib file.

Introduction

- shall we include an abstract? Recent research has shown that word embeddings derived from language corpora inherit human biases. The first seminal study on this topic was by Bolukbasi et al. (2016) who used the word analogy task developed by Mikolov et al. (2013b) to prove that "Man is to Computer Programmer as Woman is to Homemaker" - a clearly gender biased analogical relation derived from a Google News embedding. Caliskan et al. (2017) complement this finding with a large study on human-like semantic biases in Natural Language Processing (NLP) tools. They have shown in more general that human biases as exhibited in psychological studies using, for example, the Implicit Association Test (IAT) are learned by NLP algorithms designed to construct meaningful word representations. According to Zhao et al. (2018a) these biases propagate to downstream tasks. As pre-trained word embeddings are often used for a lot of more complex NLP tasks and architectures of our everyday life, the biased embeddings bear the risk of proliferating and strengthening existing stereotypes in human cultures.

Debiasing Embeddings. But how can the problem of biased embeddings be solved? Several researchers have proposed post-processing techniques or algorithm modifications that promise to "debias" the word embeddings obtained by algorithms like word2vec or GloVe (e.g. Bolukbasi et al., 2016; Kaneko & Bollegala, 2019; Zhao et al., 2018b). Bolukbasi et al. (2016) have developed an algorithm called Hard Debias that is based on the idea of removing (biased) gender directions from the embedding whilst preserving desired gender relations. For example, the encoded gender of the words “king” and “queen” is given by definition. So the relation between “male” and “king” and between “female” and “queen” is desired. On the other hand, words like “nurse” and “doctor” also tend to exhibit relations to a specific gender even though the words do not have a gender-specific meaning but can be used for all genders. The proposed Hard Debias algorithm first identifies a gender subspace of the embedding that best captures the bias. Then a neutralizing step ensures that gender neutral words (like “nurse” and “doctor”) are indeed neutral, i.e. zero, in the gender subspace. An equalizing step is then applied in order to ensure that useful relations apply to words from both genders and are not biased towards one gender anymore. For example, the word “babysit” should be equally distant from both “she” and “he”. Wang et al. (2020) build on Hard Debias in their newly proposed Double Hard Debias algorithm. The main idea is to not only remove the gender direction of the biased embedding, but also the frequency direction as it has been shown that word frequency has a significant impact on the geometry of the embedding space (e.g. Gong et al., 2018; Mu, Bhat, & Viswanath, 2018, more on this later).

Motivation. In this project, we aimed to replicate the debiasing method presented by Wang et al. (2020) and to reproduce their experimental results as it is the most recent paper that proposes a post-processing technique for debiasing algorithms. The reproduction of existing results is not only good practice in science, but it is also essential for gaining a deeper understanding on the methods used and it can help to validate existing results or to shed well-grounded doubt on them. Recent studies of reproducibility in the field of Computer Science (e.g. Collberg, Proebsting, & Warren, 2015) and NLP (e.g. Fokkens et al., 2013; Belz, Agarwal, Shimorina, & Reiter, 2021; Cohen et al., 2018; Mieskes, 2017) explain why reproducibility endeavours are often failing and shed light on the exact nature of the “reproducibility crisis” (term coined by e.g. Baker, 2015) in NLP research. Belz et al. (2021), for example, report that the community’s interest in topics of reproducibility have risen even though reproduction attempts still tend to fail due to problems like missing data, missing code and incomplete documentation (see also Fokkens et al., 2013; Mieskes, Fort, Névéol, Grouin, & Cohen, 2019). This project aims to make a contribution to the increasing body of reproducibility and replication attempts in NLP research.

Implementation

- task description
- explanation of model and training choices

Pilot Study: Impact of Word Frequency (Meike).

Datasets and Preliminaries (Sonja).

Hard Debias (Kristina). The authors make use of the Hard Debias algorithm proposed by Bolukbasi et al. (2016). The paper at hand does not give much information on the exact implementation of Hard Debias used. The code uploaded to the authors' Github repository is not well documented, so we were not able to find the exact parts of the algorithm that should refer to the implementation of Hard Debias. That is why we sticked to the original paper from Bolukbasi et al. (2016) in order to re-implement Hard Debias.

The paper describes two steps: First, the gender direction (or, more generally, the subspace) has to be identified. This is achieved with the help of defining sets $D_1, D_2, ..., D_n \subset W$ which consist of *gender specific* words, i.e. words which are associated with a gender by definition like *girl, boy* or *she, he.* These are the words that can help to identify the gender direction by capturing the concept *female, male* in the embedding $\{\vec{w} \in \mathbb{R}^d\}_{w \in W}$. Whereas in some simple implementations of Hard Debias, only one definitional pair might be used, (Bolukbasi et al., 2016) suggest to compute the gender direction B across multiple pairs to more robustly estimate the bias. In our implementation we used the 10 word pairs suggested by Bolukbasi et al. (2016) which were experimentally shown to agree with an intuitive concept of gender. For these 10 pairs the principal components (PCs) are calculated and the bias subspace B is made of the first $k \geq 1$ rows of the decomposition SVD(C). According to Bolukbasi et al. (2016), the first eigenvalue is significantly larger than the rest and so the top PC is hypothesized to capture the gender subspace. So $k = 1$ is chosen and our resulting gender subspace B is thus simply a direction. C is calculated in the following way:

$$C := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i) / |D_i|$$

where $\mu_i := \sum_{w \in D_i} \vec{w} / |D_i|$ are the means of the defining sets $D_1, D_2, ..., D_n \subset W$.

As a second step Hard Debias neutralizes and equalizes the word embeddings. Neutralizing means to transform each word embedding $\vec{w}$ such that every word $w \in N$ has zero projection in the gender subspace. So for each word $w \in N$ in a set of neutral words $N \subseteq W$, we re-embed $\vec{w}$:

$$\vec{w} := (\vec{w} - \vec{w}_B) / ||\vec{w} - \vec{w}_B||$$

. The equalize step ensures that desired analogical properties hold for both female and male words contained in the equality sets $\mathcal{E} = \{E_1, E_2, ..., E_m\}$ where each $E_i \subseteq W$. For example, after debiasing we would like the embeddings of the pair $E = \{grandmother, grandfather\}$ contained in the equality sets to be equidistant from the embedding of *babysit* (Bolukbasi et al., 2016). This is enforced by equating each set of words $E \in \mathcal{E}$ outside of B to their simple average $\nu := \mu - \mu_B$ where $\mu := \sum_{w \in E} w / |E|$ before adjusting vectors so that they are unit length. So for each word $w \in E$, $\vec{w}$ is re-embedded to

$$\vec{w} := \nu + \sqrt{1 - ||\nu||^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||}$$

.

With the help of the original paper, we were successfully able to re-implement Hard Debias.

Double-Hard Debias (Meike).

Evaluation

- stick to the paper

Baselines (Sonja).

Evaluation of Debiasing Performance.

Debiasing in Downstream Applications.

Coreference Resolution (Meike).

- no replication possible, no code provided

Debiasing at Embedding Level.

Table 1

WEAT Test

X	C...F..d	C...F..p	M...A..d	M...A..p	S...A..d	S...A
original GloVe	1.805996	0.0001524	0.6885951	0.0850657	1.1298693	0.01193
Double-Hard-GloVe (Wang et al.)	1.531301	0.0010940	-0.5625492	0.8692854	-0.6502705	0.90275
Double-Hard-GloVe (replication)	1.529532	0.0011171	-0.6895229	0.9160087	-0.9418647	0.97082
GN-Glove	1.821105	0.0001351	-0.2563579	0.6963340	1.0689540	0.01621
GN-Glove(a)	1.755476	0.0002232	0.5030034	0.1584799	0.8797335	0.03953
GP-GloVe	1.805885	0.0001498	1.2085292	0.0078522	1.1064400	0.01319
GP-GN-GloVe	1.797431	0.0001750	-0.0126418	0.5079182	0.8460069	0.04530
Hard-GloVe	1.546646	0.0010009	-0.9826043	0.9751140	-0.5383765	0.85922
Strong-Hard-GloVe	1.546646	0.0009967	-0.9856814	0.9754177	-0.5471175	0.86254

The Word Embeddings Association Test (WEAT) (Sonja).

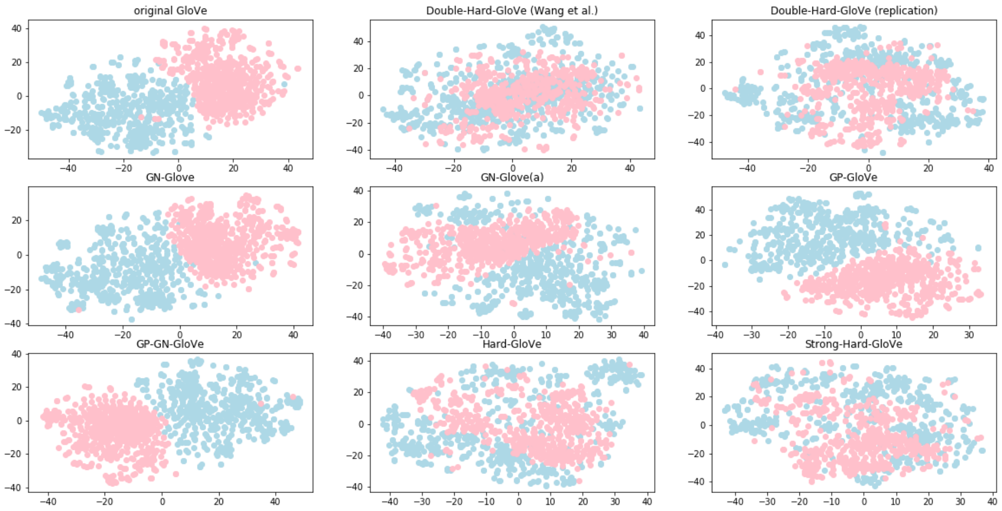
Table 2

Neighborhood Metric Clustering Accuracy

X	Top.100	Top.500	Top.1000
original GloVe	1.000	1.000	1.0000
Double-Hard-GloVe (Wang et al.)	0.545	0.565	0.5800
Double-Hard-GloVe (replication)	0.525	0.506	0.5370
GN-Glove	1.000	0.999	0.9990
GN-Glove(a)	1.000	0.997	0.9835
GP-GloVe	1.000	1.000	1.0000
GP-GN-GloVe	1.000	0.998	0.9980
Hard-GloVe	0.510	0.517	0.5030
Strong-Hard-GloVe	0.510	0.518	0.5025

Neighborhood Metric (Meike).

- discuss that this is shady in the discussion part



Analysis of Retaining Word Semantics (Kristina). One of the most important properties of embeddings is that they represent meaningful word semantics. In this section it is tested whether the debiased embeddings still have this property.

Word Analogy (Kristina). The word analogy task was introduced by Mikolov et al. (2013b). The task is to find the word D such that “A is to B as C is to D”. One example for an unbiased analogy is: “Man is to King as Woman is to Queen” whereas a biased analogy would be: “Man is to Computer Programmer as Woman is to Homemaker” (Bolukbasi et al., 2016). The debiased embeddings are evaluated on two word analogy test sets: the MSR (Mikolov et al., 2013b) and the Google word analogy task (Mikolov et al., 2013a) in order to find out whether they preserve desired unbiased analogies.

The MSR word analogy dataset contains 8000 syntactic questions in the form presented above. The missing word D is computed by maximizing the cosine similarity between D and C - A + B. The evaluation metric is the percentage of correctly answered questions (see Wang et al., 2020).

The Google word analogy dataset contains 19.544 (**Total**) questions, 8.869 of which are semantic (**Sem**) and 10.675 are syntactic (**Syn**) questions.

Concept Categorization (Sonja).

Discussion

- analysis of results and evaluation of performance evaluation
- ablation studies (not applicable)
- discuss the results and what could be (partly) replicated and what not

Conclusion

Table 3

MSR Word Analogy Task

X	MSR
original GloVe	0.5440213
Double-Hard-GloVe (Wang et al.)	0.4240000
Double-Hard-GloVe (replication)	0.6212121
GN-Glove	0.5171990
GN-Glove(a)	0.5071663
GP-GloVe	0.5161753
GP-GN-GloVe	0.5198608
Hard-GloVe	0.6257166
Strong-Hard-GloVe	0.6214169

References

Baker, M. (2015). Reproducibility crisis: Blame it on the antibodies. *Nature*, *521*(7552), 274–276. https://doi.org/https://doi.org/10.1038/521274a

Belz, A., Agarwal, S., Shimorina, A., & Reiter, E. (2021). *A systematic review of reproducibility research in natural language processing.* Retrieved from http://arxiv.org/abs/2103.07929

Bolukbasi, T., Chang, K., 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

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T., Hargraves, O., Goss, F., . . . Hunter,

Table 4

Google Word Analogy Task

X	Sem	Syn	Total
original GloVe	0.8014096	0.5569170	0.6389759
Double-Hard-GloVe (Wang et al.)	0.7135838	0.4799063	0.5667098
Double-Hard-GloVe (replication)	0.8020315	0.6655147	0.7113338
GN-Glove	0.7597430	0.5466541	0.6181730
GN-Glove(a)	0.7589138	0.5425699	0.6151812
GP-GloVe	0.8192371	0.5541942	0.6431504
GP-GN-GloVe	0.7574627	0.5667609	0.6307660
Hard-GloVe	0.7999585	0.6610116	0.7076463
Strong-Hard-GloVe	0.7674129	0.6573463	0.6942879

L. E. (2018). Three dimensions of reproducibility in natural language processing. *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1025

Collberg, C., Proebsting, T., & Warren, A. M. (2015). *Repeatability and benefaction in computer systems research.* studie. Retrieved from http://reproducibility.cs.arizona.edu/v2/RepeatabilityTR.pdf

Fokkens, A., Erp, M. van, Postma, M., Pedersen, T., Vossen, P., & Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1691–1701. Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P13-1166

Gong, C., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2018). FRAGE: Frequency-agnostic word representation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2018/file/e555ebe0ce426f7f9b2bef0706315e0c-Paper.pdf

Kaneko, M., & Bollegala, D. (2019). Gender-preserving debiasing for pre-trained word embeddings. *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1641–1650. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1160

Mieskes, M. (2017). A quantitative study of data in the NLP community. *Proceedings of the first ACL workshop on ethics in natural language processing*, 23–29. Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1603

Mieskes, M., Fort, K., Névéol, A., Grouin, C., & Cohen, K. B. (2019). NLP Community Perspectives on Replicability. *Recent Advances in Natural Language Processing*. Varna, Bulgaria. Retrieved from https://hal.archives-ouvertes.fr/hal-02282794

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies*, 746–751. Atlanta, Georgia: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N13-1090

Mu, J., Bhat, S., & Viswanath, P. (2018). *All-but-the-top: Simple and effective postprocessing for word representations*. Retrieved from http://arxiv.org/abs/1702.01417

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2020). Double-hard debias: Tailoring word embeddings for gender bias mitigation. *Association for computational linguistics (acl)*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018a, January). *Gender bias in coreference resolution: Evaluation and debiasing methods*. 15–20. https://doi.org/10.18653/v1/N18-2003

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018b). Learning gender-neutral word embeddings. *Proceedings of the 2018 conference on empirical methods in natural language processing*, 4847–4853. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1521