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

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

[1] University of Osnabrück

Author Note

The authors made the following contributions. Kristina Kobrock: müssen wir natürlich nicht ausfüllen; Meike Korsten: . . . ; Sonja Börgerding: . . . .

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. (2013) 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; 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

Datasets and Preliminaries.

Hard Debias.

Double-Hard Debias.

Evaluation

- stick to the paper

Baselines.

Evaluation of Debiasing Performance.

Debiasing in Downstream Applications.

Coreference Resolution.

- no replication possible, no code provided

Debiasing at Embedding Level.

The Word Embeddings Association Test (WEAT).

78 ***Neighborhood Metric.***

79 • discuss that this is shady in the discussion part

80 **Analysis of Retaining Word Semantics**

81 ***Word Analogy.***

82 ***Concept Categorization.***

83 **Discussion**

84 • analysis of results and evaluation of performance evaluation

85 • ablation studies (not applicable)

86 • discuss the results and what could be (partly) replicated and what not

87 **Conclusion**

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

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., Yih, W.-t., & Zweig, G. (2013). 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