Multi-objective Search for Gender-fair and Semantically Correct Word Embeddings (HOP GECCO'23)

Max Hort
Simula Research Laboratory
Oslo, Norway
maxh@simula.no

Rebecca Moussa
University College London
London, United Kingdom
rebecca.moussa.18@ucl.ac.uk

Federica Sarro
University College London
London, United Kingdom
f.sarro@ucl.ac.uk

ABSTRACT

Mitigating algorithmic bias during the development life cycle of AI-enabled software is crucial given that any bias in these algorithms is inherited by the software systems using them. At the Hot-off-the-Press GECCO track, we aim at disseminating our article *Multi-objective search for gender-fair and semantically correct word embeddings. Applied Soft Computing, 2023* [5]. In this work, we exploit multi-objective search to strike an optimal balance between reducing gender bias and improving semantic correctness of word embedding models, which are at the core of many AI-enabled systems. Our results show that, while single-objective search approaches are able to reduce the gender bias of word embeddings, they also reduce their semantic correctness. On the other hand, multi-objective approaches are successful in improving both goals, in contrast to existing work which solely focuses on reducing gender bias. Our results show that multi-objective evolutionary approaches can be successfully exploited to address bias in AI-enable software systems, and we encourage the research community to further explore opportunities in this direction.

CCS CONCEPTS

• **Software and its engineering → Search-based software engineering**.

KEYWORDS

Word Embedding, Optimization, fairness, debiasing

ACM Reference Format:
Max Hort, Rebecca Moussa, and Federica Sarro. 2023. Multi-objective Search for Gender-fair and Semantically Correct Word Embeddings (HOP GECCO'23). In *Genetic and Evolutionary Computation Conference Companion (GECCO '23 Companion), July 15–19, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583133.3595847

1 OVERVIEW

It is crucial for software systems to operate unbiased and to not discriminate against individuals or population groups based on sensitive attributes such as race or gender. Such an unbiased behaviour is especially important for systems based on Artificial Intelligence (AI), which process large amounts of data and learn to make predictions, as these can be difficult to comprehend.

One type of such learning models are based on Natural Language Processing (NLP). These are trained on copious sizes of text to support applications such as sentiment analysis or recommendations. At the foundation of these models lie word embeddings [7], which are a useful tool to represent words numerically, such that they can be easily processed and used by tools. To successfully replicate semantics from human written texts, word embeddings are trained on large amounts of training data, which is time-consuming. Therefore, after a word embedding model is trained, these are often times shared with the public. This provides the benefit of an easy access to trained resources, however it can also lead to the sharing of negative side effects, such as biases. Due to the fact that word embeddings are trained on human-written text, they can learn to replicate human biases that are hidden in the training data. One example for such a bias retained in a word embedding model is provided by Bolukbasi et al. [1]: "man to computer programmer" is the same as "woman to homemaker". This example illustrates a biased relationship of occupations for "man" and "woman" in the embedding space of the word embeddings. Using such embeddings for recommending jobs could risk an unfavourable treatment of female applicants when considering programming related jobs.

To combat the risk and negative effects of bias in language-based systems, several approaches have been proposed to mitigate and remove gender bias from word embeddings [1, 3]. While these approaches have been successful in mitigating gender bias, they do not consider other performance characteristics, such as the semantic correctness of the embeddings (e.g., to what extent do word embeddings agree with human semantics). Therefore, in our work [5] we proposed to tackle the task of debiasing word embeddings as a search-based, multi-objective optimization problem such that we can apply different search approaches to optimize both, semantic correctness and gender bias.

2 OPTIMIZING WORD EMBEDDINGS

Here, we outline the design and formulation of the optimization of word embedding as a search problem [5, 6].

Word Embedding Models: Word embedding models learn to represent words as numerical vectors based on the context they are used (i.e., co-occurrences of a word in a training corpus). These representations of words w have a dimensionality d and can be specified as follows: $\overrightarrow{w} \epsilon \mathbb{R}^d$.

Solution Representation: First, we require a representation which captures the modification of the word vectors. For this purpose, we employ the a solution vector $\overrightarrow{s}$, of the same cardinality as the investigated word embedding model. Existing word vectors $\overrightarrow{w}$

are modified by performing an element-wise vector multiplication: $\overrightarrow{w'} = \overrightarrow{w} \circ \overrightarrow{s}$. This multiplication is applied to each word vector in a word embedding model, and $\overrightarrow{w'}$ represents the resulting, optimized word embeddings.

Initialization: Before starting the search, we need to initialize the solution vector $\overrightarrow{s}$, one time for local search or multiple times for global search. We initialize $\overrightarrow{s}$ by adding a small noise vector to a vector of ones: $\overrightarrow{s} = \overrightarrow{1} + \overrightarrow{noise}$. $\overrightarrow{1}$ is used, as multiplying by it represents the original word embedding model.

Neighbor Creation: To explore the search space, one needs to create neighbors to allow the modification of the previously initialized vector $\overrightarrow{s}$. For this purpose, we consider two modification operators: 1) adding a small noise value to a single element of $\overrightarrow{s}$; 2) adding a small uniform noise vector to $\overrightarrow{s}$.

Fitness Functions: We measure two characteristics to determine the fitness of word embedding modifications $\overrightarrow{s}$: gender bias, semantic correctness. We measure gender bias according to the Word Embedding Association Tests (WEATs) proposed by Caliskan et al. [2]. In total, Caliskan et al. [2] provided ten different sets (scenarios) to compute different kinds of biases according to target and attribute sets. Three of these sets are concerned with gender bias and are used in proceeding experiments (WEAT 6, 7, 8). The intuition behind WEAT is that for example, a set of "male" attribute words should have the same similarity to science-related target words as a set of "female" attribute words would in the embedding space (e.g., the similarity of $\overrightarrow{he}$ to $\overrightarrow{physics}$ should be identical to the similarity $\overrightarrow{she}$ to $\overrightarrow{physics}$). We measure the semantic correctness of word embeddings with the word similarity method [4]. Based on a list of word pairs and an associated similarity score determined by humans, the word pair similarities are measured according to the respective word embeddings. Semantic correctness is then determined by the Spearman's ρ rank correlation coefficient [8] of human judged similarity and the ones provided by the word embedding model.

Computational Search: To optimize word embeddings, we apply four different search approaches, three single-objective methods (Hill Climbing (HC), Tabu Search (TS), Genetic Algorithms (GA)) and one multi-objective optimization approach (NSGA-II). Additionally, we use a random baseline, as well as a comparison against two existing debiasing methods for word embedding models: Hard Debiasing (HD) [1] and Linear Projection (LP) [3].

3 SUMMARY OF EXPERIMENTAL RESULTS

Our experiments are conducted on a Word2Vec (W2V) model, pretrained on news articles. Here, we present the results of our optimization of the W2V model, which provides 300-dimensional vectors, with regards to gender bias and semantic correctness.

Single-Objective Optimization of Gender Bias: At first, we investigated the ability of single-objective optimization approaches with regards to their ability to reduce gender bias. For this purpose, we trained HC, TS and GA on each of the three WEAT sets to find a transformation of the word embeddings to minimize gender bias. The performance is then evaluated on the other two WEAT sets. For each of the WEAT sets used for testing, we were able to find statistically significant reductions of gender bias. In particular, GAs have been the best performing approach for reducing gender bias.

Effect on Semantic Correctness: While we showed that HS, TS and GAs are able to reduce gender bias, such an improvement often times comes at the cost of a reduced accuracy. To quantify such a detoriation in performance, we measured whether the semantic correctness, according to the semantic similarity test with the MEN dataset, decreases after reducing bias. Our results confirm this concern, given that the semantic correctness is reduced in all cases with a statistical significance.

Multi-Objective Optimization: After verifying the ability of search approaches to reduce gender bias of word embeddings and the associated reduction of semantic correctness, we applied four search approaches to optimize both objectives in a multi-objective scenario. In particular, we applied the three single-objective approaches to optimize a weighted sum of the two objectives, as well as NSGA-II, a multi-objective optimization approach. Moreover, we compared our approaches to two existing methods solely focused on reducing gender bias (HD and LD).

By using multi-objective optimizations, we were able to improve both, gender bias and semantic correctness, of word embeddings. The approach with the highest semantic correctness is NSGA-II, while the lowest bias was achieved by HD, with a constant level of semantic correctness.

4 CONCLUSIONS AND FUTURE WORK

While previous work only provided the engineer with a single solution, the use of multi-objective approaches enables them to explore the trade-offs between two important competing objectives (accuracy and fairness) among a rich set of equally viable solutions to the problem at hand. This opens up a rich avenue for future work. Our proposal can be further explored for other pre-trained word embedding models and semantic evaluation measures. Besides, it can be used to address the reduction of additional bias types (e.g., race, age) or take other objectives into account, such as the performance of word embedding models on downstream tasks (e.g., sentiment analysis).

ACKNOWLEDGEMENTS

This research is supported by the ERC grant no. 741278 (EPIC).

REFERENCES

[1] T. Bolukbasi, K. Chang, J.Y. Zou, V. Saligrama, and A.T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.
[2] A. Caliskan, J.J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
[3] S. Dev and J. Phillips. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR, 2019.
[4] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. Problems with evaluation of word embeddings using word similarity tasks. *preprint arXiv:1605.02276*, 2016.
[5] M. Hort, R. Moussa, and F. Sarro. Multi-objective search for gender-fair and semantically correct word embeddings. *Applied Soft Computing*, 133, 2023. doi: https://doi.org/10.1016/j.asoc.2022.109916.
[6] Max Hort and Federica Sarro. Optimising word embeddings with search-based approaches. In *Procs. of GECCO Companion*, GECCO. ACM, 2020.
[7] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
[8] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101, jan 1904.