Comparison and Improvement of Bias Mitigation Algorithms for Word Embeddings

María José Zambrano

Department of Computer Science, University of Chile

2022

Table of contents

Background

Problem

Proposal

Methodology

Expected Results

Word Embeddings

- Word embeddings are models that encode the meaning of words in dense vectors, based on the distributional hypothesis [5].
- They are some of the most used models in the Natural Language processing field to represent the human vocabulary.

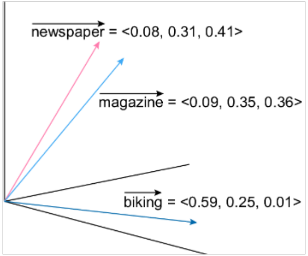

$\overrightarrow{\text{newspaper}}$ = <0.08, 0.31, 0.41>
$\overrightarrow{\text{magazine}}$ = <0.09, 0.35, 0.36>
$\overrightarrow{\text{biking}}$ = <0.59, 0.25, 0.01>

Figure: Example of vectors

Bias in Word Embeddings

- It has been found that some word embedding models learn relationships such as "man" is to "computer programmer" as "woman" is to "homemaker" [3], resulting in unfair representations of the language.
- To address the bias issue several bias mitigation algorithms have been proposed

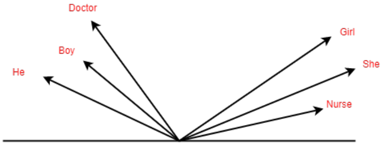


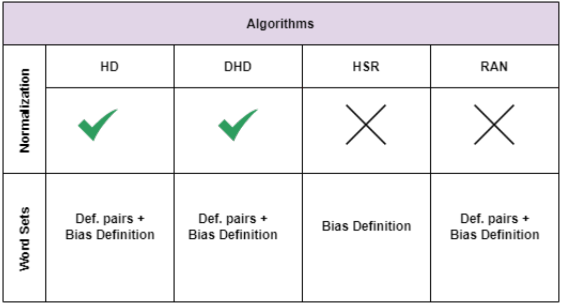
Figure: The vector for "*doctor*" is closer to masculine words and "*nurse*" to feminine words.

Bias in Word Embeddings models

- To address the bias issue:
  - Different metrics have been proposed aiming to quantify the bias in word embedding models.
  - Algorithms that aim to mitigate the bias in word embedding models have been proposed.
- WEFE [2] encapsulates bias measurement metrics and bias mitigation algorithms.

Problem to Address

- There is a lack of systematic comparison of the bias mitigation algorithms
- Comparing them is not a trivial task



	Algorithms			
	HD	DHD	HSR	RAN
Normalization	✓	✓	✗	✗
Word Sets	Def. pairs + Bias Definition	Def. pairs + Bias Definition	Bias Definition	Def. pairs + Bias Definition

Figure: Comparison of the algorithms

Problem to Address

- This makes it unclear which algorithms reduces bias the most.
- Makes it difficult to improve the bias mitigation effect.

This research

For this work we address two research lines:

1. Create a standardize methodology to compare bias mitigation algorithms
2. Combine the algorithms to improve their performance.
   - Using the idea of ensemble methods from classical machine learning

Comparing algorithms

- The algorithms differ in:
  - Word sets they use
  - Pre-operations they perform
- To fairly compare the methods we will eliminate these by standardizing all variables that can affect the bias.

Ensemble methods

Ensembles consists of sets of implemented instances of machine learning algorithms that work together to improve the performance of the overall system [1].

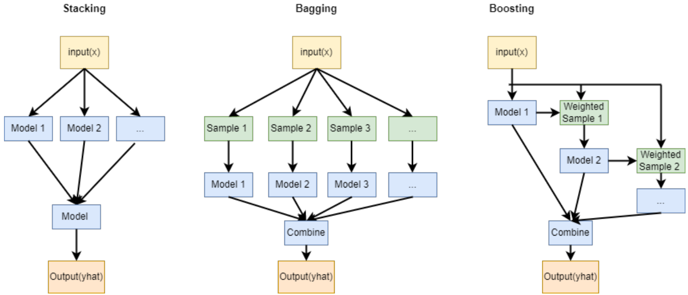
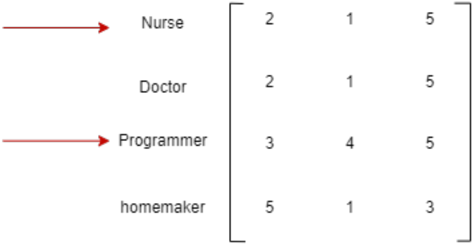


Figure: Types of ensembles [4]

Adapting ensembles

- Instead of sampling training data, sampling words used to perform the debias.

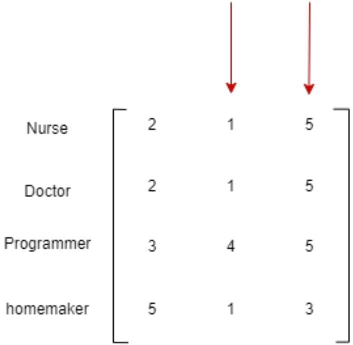


$$\begin{array}{l} \text{Nurse} \\ \text{Doctor} \\ \text{Programmer} \\ \text{homemaker} \end{array} \begin{bmatrix} 2 & 1 & 5 \\ 2 & 1 & 5 \\ 3 & 4 & 5 \\ 5 & 1 & 3 \end{bmatrix}$$

Figure: Apply the debias to some words

Adapting ensembles

- Instead of sampling training data, sampling dimension of the vectors used to perform the debias.



$$\begin{array}{l} & & \downarrow & \downarrow \\ \text{Nurse} & \left[\begin{array}{ccc} 2 & 1 & 5 \\ 2 & 1 & 5 \\ 3 & 4 & 5 \\ 5 & 1 & 3 \end{array}\right. \\ \text{Doctor} \\ \text{Programmer} \\ \text{homemaker} \end{array}$$

$$\begin{array}{l|ccc|} & & \downarrow & \downarrow \\ \text{Nurse} & 2 & 1 & 5 \\ \text{Doctor} & 2 & 1 & 5 \\ \text{Programmer} & 3 & 4 & 5 \\ \text{homemaker} & 5 & 1 & 3 \end{array}$$

Figure: Apply the debias to some of the dimensions of the vectors

Adapting ensembles

- Combining the debiased word vectors of different debias algorithms giving more importance to those that perform better, according to the bias measurement metrics.
- Applying one algorithm after another.

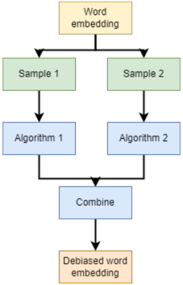


Figure: Combination of bias mitigation algorithms

Contributions

As a result of this research, we expect to contribute by improving bias mitigation methods that have already been proposed by proposing ensemble methods for bias mitigation algorithms.

References I

- Juan Jose García Adeva, Ulises Cerviño Beresi, and Rafael A. Calvo. Accuracy and diversity in ensembles of text categorisers. *CLEI Electron. J.*, 8(2), 2005.

- Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. Wefe: The word embeddings fairness evaluation framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 430–436. International Joint Conferences on Artificial Intelligence Organization, 2020.

- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

- Jason Brownlee. A gentle introduction to ensemble learning algorithms, Apr 2021.

- Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Comparison and Improvement of Bias Mitigation Algorithms for Word Embeddings

María José Zambrano

Department of Computer Science, University of Chile

2022