Learning Gender-Neutral Word Embeddings

Jieyu Zhao **Yichao Zhou** **Zeyu Li** **Wei Wang** **Kai-Wei Chang**
University of California, Los Angeles
{jyzhao, yz, zyli, weiwang, kwchang}@cs.ucla.edu

Abstract

Word embedding models have become a fundamental component in a wide range of Natural Language Processing (NLP) applications. However, embeddings trained on human-generated corpora have been demonstrated to inherit strong gender stereotypes that reflect social constructs. To address this concern, in this paper, we propose a novel training procedure for learning gender-neutral word embeddings. Our approach aims to preserve gender information in certain dimensions of word vectors while compelling other dimensions to be free of gender influence. Based on the proposed method, we generate a Gender-Neutral variant of GloVe (GN-GloVe). Quantitative and qualitative experiments demonstrate that GN-GloVe successfully isolates gender information without sacrificing the functionality of the embedding model.

1 Introduction

Word embedding models have been designed for representing the meaning of words in a vector space. These models have become a fundamental NLP technique and have been widely used in various applications. However, prior studies show that such models learned from human-generated corpora are often prone to exhibit social biases, such as gender stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017). For example, the word "programmer" is neutral to gender by its definition, but an embedding model trained on a news corpus associates "programmer" closer with "male" than "female".

Such a bias substantially affects downstream applications. Zhao et al. (2018) show that a coreference resolution system is sexist due to the word embedding component used in the system. This concerns the practitioners who use the embedding model to build gender-sensitive applications such as a resume filtering system or a job recommendation system as the automated system may discriminate candidates based on their gender, as reflected by their name. Besides, biased embeddings may implicitly affect downstream applications used in our daily lives. For example, when searching for "computer scientist" using a search engine, as this phrase is closer to male names than female names in the embedding space, a search algorithm using an embedding model in the backbone tends to rank male scientists higher than females', hindering women from being recognized and further exacerbating the gender inequality in the community.

To alleviate gender stereotype in word embeddings, Bolukbasi et al. (2016) propose a post-processing method that projects gender-neutral words to a subspace which is perpendicular to the gender dimension defined by a set of gender-definition words.[1] However, their approach has two limitations. First, the method is essentially a pipeline approach and requires the gender-neutral words to be identified by a classifier before employing the projection. If the classifier makes a mistake, the error will be propagated and affect the performance of the model. Second, their method completely removes gender information from those words which are essential in some domains such as medicine and social science (Back et al., 2010; McFadden et al., 1992).

To overcome these limitations, we propose a learning scheme, Gender-Neutral Global Vectors (GN-GloVe) for training word embedding models with protected attributes (e.g., gender) based on GloVe (Pennington et al., 2014).[2] GN-GloVe represents protected attributes in certain dimen-

[1] Gender-definition words are the words associated with gender by definition (e,g., mother, waitress); the remainder are gender-neutral words.

[2] The code and data are released at https://github.com/uclanlp/gn_glove

sions while neutralizing the others during training. As the information of the protected attribute is restricted in certain dimensions, it can be removed from the embedding easily. By jointly identifying gender-neutral words while learning word vectors, GN-GloVe does not require a separate classifier to identify gender-neutral words; therefore, the error propagation issue is eliminated. The proposed approach is generic and can be incorporated with other word embedding models and be applied in reducing other societal stereotypes.

Our contributions are summarized as follows: 1) To our best knowledge, GN-GloVe is the first method to learn word embeddings with protected attributes; 2) By capturing protected attributes in certain dimensions, our approach ameliorates the interpretability of word representations; 3) Qualitative and quantitative experiments demonstrate that GN-GloVe effectively isolates the protected attributes and preserves the word proximity.

2 Related Work

Word Embeddings Word embeddings serve as a fundamental building block for a broad range of NLP applications (dos Santos and Gatti, 2014; Bahdanau et al., 2014; Zeng et al., 2015) and various approaches (Mikolov et al., 2013b; Pennington et al., 2014; Levy et al., 2015) have been proposed for training the word vectors. Improvements have been made by leveraging semantic lexicons and morphology (Luong et al., 2013; Faruqui et al., 2014), disambiguating multiple senses (Šuster et al., 2016; Arora et al., 2018; Upadhyay et al., 2017), and modeling contextualized information by deep neural networks (Peters et al., 2018). However, none of these works attempts to tackle the problem of stereotypes exhibited in embeddings.

Stereotype Analysis Implicit stereotypes have been observed in applications such as online advertising systems (Sweeney, 2013), web search (Kay et al., 2015), and online reviews (Wallace and Paul, 2016). Besides, Zhao et al. (2017) and Rudinger et al. (2018) show that coreference resolution systems are gender biased. The systems can successfully predict the link between “the president” with male pronoun but fail with the female one. Rudinger et al. (2017) use pointwise mutual information to test the SNLI (Bowman et al., 2015) corpus and demonstrate gender stereotypes as well as varying degrees of racial, religious, and age-based stereotypes in the corpus. A temporal analysis about word embeddings (Garg et al., 2018) captures changes in gender and ethnic stereotypes over time. Researchers attributed such problem partly to the biases in the datasets (Zhao et al., 2017; Yao and Huang, 2017) and word embeddings (Garg et al., 2017; Caliskan et al., 2017) but did not provide constructive solutions.

3 Methodology

In this paper, we take GloVe (Pennington et al., 2014) as the base embedding model and gender as the protected attribute. It is worth noting that our approach is general and can be applied to other embedding models and attributes. Following GloVe (Pennington et al., 2014), we construct a word-to-word co-occurrence matrix X, denoting the frequency of the j-th word appearing in the context of the i-th word as $X_{i,j}$. $w, \tilde{w} \in \mathbb{R}^d$ stand for the embeddings of a center and a context word, respectively, where d is the dimension.

In our embedding model, a word vector w consists of two parts $w = [w^{(a)}; w^{(g)}]$. $w^{(a)} \in \mathbb{R}^{d-k}$ and $w^{(g)} \in \mathbb{R}^k$ stand for neutralized and gendered components respectively, where k is the number of dimensions reserved for gender information.[3] Our proposed gender neutralizing scheme is to reserve the gender feature, known as “protected attribute” into $w^{(g)}$. Therefore, the information encoded in $w^{(a)}$ is independent of gender influence. We use $v_g \in \mathbb{R}^{d-k}$ to denote the direction of gender in the embedding space. We categorize all the vocabulary words into three subsets: male-definition Ω_M, female-definition Ω_F, and gender-neutral Ω_N, based on their definition in WordNet (Miller and Fellbaum, 1998).

Gender Neutral Word Embedding Our minimization objective is designed in accordance with above insights. It contains three components:

$$J = J_G + \lambda_d J_D + \lambda_e J_E, \tag{1}$$

where λ_d and λ_e are hyper-parameters.

The first component J_G is originated from GloVe (Pennington et al., 2014), which captures the word proximity:

$$J_G = \sum_{i,j=1}^{V} f(X_{i,j}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{i,j} \right)^2 .$$

[3] We set $k = 1$ in this paper.