Reducing gender bias in word embeddings

Tuhin Chakraborty (tuhin), Gabrielle Badie (gab47), Brett Rudder (brudder)

Abstract

Word embedding is a popular framework that represents text data as vectors of real numbers. These vectors capture semantics in language, and are used in a variety of natural language processing and machine learning applications. Despite these useful properties, word embeddings derived from ordinary language corpora necessarily exhibit human biases [6]. We measure direct and indirect gender bias for occupation word vectors produced by the GloVe word embedding algorithm [9], then modify this algorithm to produce an embedding with less bias to mitigate amplifying the bias in downstream applications utilizing this embedding.

1 Introduction

Word embeddings represent words as n-dimensional vectors, $\overrightarrow{w} \in \mathbb{R}^n$ as learned from cooccurence data in a large corpus of ordinary language text (news articles, webpages, etc.). A desirable property of these vectors is that they geometrically capture intrinsic relationships between words, making them valuable inputs for applications such as search/result ranking [4] and sentiment analysis [11]. For example, analogies such as ”*man* is to *king* as *woman* is to *queen*”, are captured by the equality $\overrightarrow{man} - \overrightarrow{king} \approx \overrightarrow{woman} - \overrightarrow{queen}$ [7].

Recent research[1] into quantifying and mitigating gender stereotypes in word embeddings focuses on reducing bias in pre-trained vectors [2]. We alternatively mitigate bias by updating the GloVe algorithm itself [9], one of the most popular word embedding frameworks. Our modifications focus on gender bias among gender neutral occupation words (doctor, nurse, programmer, etc.) that would otherwise be considered gender-neutral. We tested our changes using the 1 Billion Word Language Model Benchmark [3] dataset as our input data.

2 Identifying and Measuring Gender Bias

Though there is research discussing discrimination bias in various machine learning results in general, there is little literature focused on quantifying and mitigating discrimination bias in word embeddings. To quanitfy bias, we applied the metrics defined by Bolukbasi, et. al.'s foundational work in this area [1], which is based on the broader definition and analysis of direct and indirect discrimination in data-mining defined by Pedreschi et. al. [8]. Before we consider our metrics for direct and indirect bias, we observe the following equality derived from the word vectors trained by GloVe

$$\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\_programmer} - \overrightarrow{homemaker}$$

Although neither of the terms *computer_programmer* nor *homemaker* are gendered nouns, the geometry indicates that the word *computer_programmer* is more closely related to the

[1] just presented at NIPS this last month!