Quantifying and Reducing Stereotypes in Word Embeddings

Tolga Bolukbasi[1] TOLGAB@BU.EDU
Kai-Wei Chang[2] KW@KWCHANG.NET
James Zou[2] JAMESYZOU@GMAIL.COM
Venkatesh Saligrama[1] SRV@BU.EDU
Adam Kalai[2] ADAM.KALAI@MICROSOFT.COM

1 Boston University, 8 Saint Mary's Street, Boston, MA
2 Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

Abstract

Machine learning algorithms are optimized to model statistical properties of the training data. If the input data reflects stereotypes and biases of the broader society, then the output of the learning algorithm also captures these stereotypes. In this paper, we initiate the study of gender stereotypes in *word embedding*, a popular framework to represent text data. As their use becomes increasingly common, applications can inadvertently amplify unwanted stereotypes. We show across multiple datasets that the embeddings contain significant gender stereotypes, especially with regard to professions. We created a novel gender analogy task and combined it with crowdsourcing to systematically quantify the gender bias in a given embedding. We developed an efficient algorithm that reduces gender stereotype using just a handful of training examples while preserving the useful geometric properties of the embedding. We evaluated our algorithm on several metrics. While we focus on male/female stereotypes, our framework may be applicable to other types of embedding biases.

1. Introduction

Word embeddings, trained only on word co-occurrence in text corpora, capture rich semantic information about words and their meanings (Mikolov et al., 2013b). Each word (or common phrase) $w \in W$ is encoded as a d-dimensional *word vector* $v_w \in \mathbb{R}^d$. Using simple vector arithmetic, the embeddings are capable of answering *analogy puzzles*. For instance, *man*:*king* :: *woman*:__[1] returns *queen* as the answer, and similarly Japan is returned for *Paris*:*France* :: *Tokyo*:*Japan* (computer-generated answers are underlined). A number of such embeddings have been made publicly available including the popular word2vec (Mikolov et al., 2013a; Mikolov et al.) embedding trained on 3 million words into 300 dimensions, which we refer to here as the w2vNEWS embedding because it was trained on a corpus of text from Google News. These word embeddings have been used in a variety of downstream applications (e.g., document ranking (Nalisnick et al., 2016), sentiment analysis (İrsoy & Cardie, 2014), and question retrieval (Lei et al., 2016)).

While word-embeddings encode semantic information they also exhibit hidden biases inherent in the dataset they are trained on. For instance, word embeddings based on w2vNEWS can return biased solutions to analogy puzzles such as *father*:*doctor* :: *mother*:*nurse* and *man*:*computer programmer* :: *woman*:*homemaker*. Other publicly available embeddings produce similar results exhibiting gender stereotypes. Moreover, the closest word to the query *BLACK MALE* returns *ASSAULTED* while the response to *WHITE MALE* is *ENTITLED TO*. This raises serious concerns about their widespread use.

The prejudices and stereotypes in these embeddings reflect biases implicit in the data on which they were trained. The embedding of a word is typically optimized to predict co-occuring words in the corpus. Therefore, if *mother* and *nurse* frequently co-occur, then the vectors v_{mother} and v_{nurse} also tend to be more similar and encode the gender stereotypes. The use of embeddings in applications can amplify these biases. To illustrate this point, consider Web search where, for example, one recent project has shown that, when carefully combined with existing approaches, word vectors can significantly improve Web page relevance results (Nalisnick et al., 2016) (note that this work is a proof of concept – we do not know which, if any, mainstream search engines presently incorporate word embeddings). Consider a researcher seeking a summer intern to work on a machine learning project on deep learning who searches for, say, "linkedin graduate student machine learning neural networks." Now, a word embedding's semantic knowledge

[1] An *analogy puzzle*, *a*:*b* :: *c*:*d*, involves selecting the most appropriate d given a, b, and c.

2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, New York, NY, USA.

can improve relevance in the sense that a LinkedIn web page containing terms such as "PhD student," "embeddings," and "deep learning," which are related to but different from the query terms, may be ranked highly in the results. However, word embeddings also rank CS research related terms closer to male names than female names. The consequence would be, between two pages that differed in the names Mary and John but were otherwise identical, the search engine would rank John's higher than Mary. In this hypothetical example, the usage of word embedding makes it even harder for women to be recognized as computer scientists and would contribute to widening the existing gender gap in computer science. While we focus on gender bias, specifically male/female, our approach may be applied to other types of biases.

We propose two methods to systematically quantify the gender bias in a set of word embeddings. First, we quantify how words, such as those corresponding to professions, are distributed along the direction between embeddings of *he* and *she*. Second, we design an algorithm for generating analogy pairs from an embedding given two seed words and we use crowdworkers to quantify whether these embedding analogies reflect stereotypes. Some analogies reflect stereotypes such as *he*:*janitor* :: *she*:*housekeeper* and *he*:*alcoholism* :: *she*:*eating disorders*. Finally, others may provoke interesting discussions such as *he*:*realist* :: *she*:*feminist* and *he*:*injured* :: *she*:*victim*.

Since biases are cultural, we enlist U.S.-based crowdworkers to identify analogies to judge whether analogies: (a) reflect *stereotypes* (to understand biases), or (b) are nonsensical (to ensure accuracy). We first establish that biases indeed exist in the embeddings. We then show that, surprisingly, information to distinguish stereotypical associations like female:homemaker from definitional associations like female:sister can often be removed. We propose an approach that, given an embedding and only a handful of words, can reduce the amount of bias present in that embedding without significantly reducing its performance on other benchmarks.

Contributions. **(1)** We initiate the study of stereotypes and biases in word embeddings. Our work follows a large body of literature on bias in language, but word embeddings are of specific interest because they are commonly used in machine learning and they have simple geometric structures that can be quantified mathematically. **(2)** We develop two metrics to quantify gender stereotypes in word embeddings based on words associated with professions together with automatically generated analogies which are then scored by the crowd. **(3)** We develop a new algorithm that reduces gender stereotypes in the embedding using only a handful of training examples while preserving useful properties of the embedding.

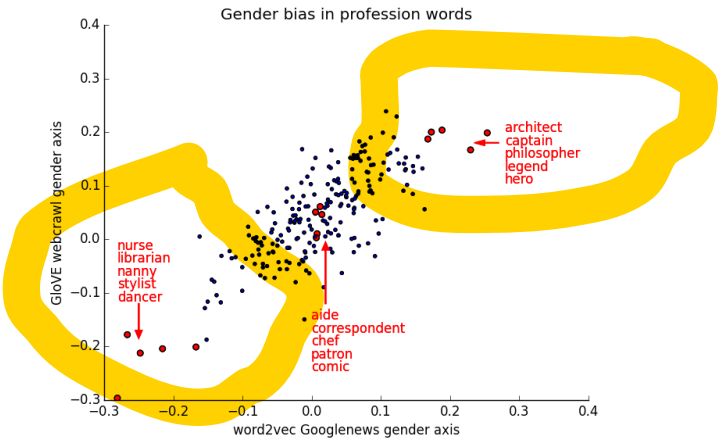

Figure 1. Comparison of gender bias of profession words across two embeddings: word2vec trained on Googlenews and GloVe trained web-crawl texts. The x and y axes show projections onto the *he-she* direction in the two embeddings. Each dot is one of 249 common profession words. Words closest to *he*, closest to *she*, and in between the two are colored in red and shown in the plot.

Prior work. The body of prior work on bias in language and prejudice in machine learning algorithms is too large to fully cover here. We note that gender stereotypes have been shown to develop in children as young as two years old (Turner & Gervai, 1995). Statistical analyses of language have shown interesting contrasts between language used to describe men and women, e.g., in recommendation letters (Schmader et al., 2007). A number of online systems have been shown to exhibit various biases, such as racial discrimination in the ads presented to users (Sweeney, 2013). Approaches to modify classification algorithms to define and achieve various notions of fairness have been described in a number of works, see, e.g., (Barocas & Selbst, 2014; Dwork et al., 2012) and a recent survey (Zliobaite, 2015).

2. Implicit stereotypes in word embedding

Stereotyped words. A simple approach to explore how gender stereotypes manifest in embeddings is to quantify which words are closer to *he* versus *she* in the embedding space (using other words to capture gender, such as *man* and *woman*, gives similar but noisier results due to their multiple meanings). We used a list of 215 common profession names, removing names that are associated with one gender by definition (e.g. waitress, waiter). For each name, v, we computed its projection onto the gender axis: $v \cdot (v_{he} - v_{she})/||v_{he} - v_{she}||_2$. Figure 1 shows the projection of professions on the w2vNEWS embedding (x-axis) and on a different embedding trained by GloVe on a dataset of web-crawled texts (y-axis). Several professions are closer to the *he* or *she* vector and this is consistent