Gender Bias in Neural Natural Language Processing

Kaiji Lu
Carnegie Mellon University
Moffiet Field, CA 94035
kaijil@andrew.cmu.edu

Piotr Mardziel
Carnegie Mellon University
Moffiet Field, CA 94035
piotrm@cmu.edu

Fangjing Wu
Carnegie Mellon University
Moffiet Field, CA 94035
fangjinw@andrew.cmu.edu

Preetam Amancharla
Carnegie Mellon University
Moffiet Field, CA 94035
pamancha@andrew.cmu.edu

Anupam Datta
Carnegie Mellon University
Moffiet Field, CA 94035
danupam@cmu.edu

Abstract

We examine whether neural natural language processing (NLP) systems reflect historical biases in training data. We define a general benchmark to quantify gender bias in a variety of neural NLP tasks. Our empirical evaluation with state-of-the-art neural coreference resolution and textbook RNN-based language models trained on benchmark data sets finds significant gender bias in how models view occupations. We then mitigate bias with *counterfactual data augmentation (CDA)*: a generic methodology for corpus augmentation via causal interventions that breaks associations between gendered and gender-neutral words. We empirically show that CDA effectively decreases gender bias while preserving accuracy. We also explore the space of mitigation strategies with CDA, a prior approach to word embedding debiasing (WED), and their compositions. We show that CDA outperforms WED, drastically so when word embeddings are trained. For pre-trained embeddings, the two methods can be effectively composed. We also find that as training proceeds on the original data set with gradient descent the gender bias grows as the loss reduces, indicating that the optimization encourages bias; CDA mitigates this behavior.

1 Introduction

Natural language processing (NLP) with neural networks has grown in importance over the last few years. They provide state-of-the-art models for tasks like coreference resolution, language modeling, and machine translation [Clark and Manning, 2016a,b, Lee et al., 2017, Jozefowicz et al., 2016, Johnson et al., 2017]. However, since these models are trained on human language texts, a natural question is whether they exhibit bias based on gender or other characteristics, and, if so, how should this bias be mitigated. This is the question that we address in this paper.

Prior work provides evidence of bias in autocomplete suggestions [Lapowsky, 2018] and differences in accuracy of speech recognition based on gender and dialect [Tatman, 2017] on popular online platforms. Word embeddings, initial pre-processors in many NLP tasks, embed words of a natural language into a vector space of limited dimension to use as their semantic representation. Bolukbasi et al. [2016] and Caliskan et al. [2017] observed that popular word embeddings including *word2vec*

Preprint. Work in progress.

$1_{\square}$: The **doctor** $\overbrace{\text{ran because}}^{5.08}$ **he** is late.

$1_{\bigcirc}$: The **doctor** $\overbrace{\text{ran because}}^{1.99}$ **she** is late.

$2_{\square}$: The **nurse** $\overbrace{\text{ran because}}^{-0.44}$ **he** is late.

$2_{\bigcirc}$: The **nurse** $\overbrace{\text{ran because}}^{5.34}$ **she** is late.

(a) Coreference resolution

	A	B	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	**doctor**.	-9.77
$2_{\square}$:	**He** is a	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	**nurse**.	-8.97

(b) Language modeling

Figure 1: Examples of gender bias in coreference resolution and language modeling as measured by coreference scores (left) and conditional log-likelihood (right).

[Mikolov et al., 2013] exhibit gender bias mirroring stereotypical gender associations such as the eponymous [Bolukbasi et al., 2016] "Man is to computer programmer as Woman is to homemaker".

Yet the question of how to measure bias in a general way for neural NLP tasks has not been studied. Our first contribution is a general benchmark to quantify gender bias in a variety of neural NLP tasks. Our definition of bias loosely follows the idea of causal testing: matched pairs of individuals (instances) that differ in only a targeted concept (like gender) are evaluated by a model and the difference in outcomes (or scores) is interpreted as the causal influence of the concept in the scrutinized model. The definition is parametric in the scoring function and the target concept. Natural scoring functions exist for a number of neural natural language processing tasks.

We instantiate the definition for two important tasks—coreference resolution and language modeling. Coreference resolution is the task of finding words and expressions referring to the same entity in a natural language text. The goal of language modeling is to model the distribution of word sequences. For neural coreference resolution models, we measure the gender coreference score disparity between gender-neutral words and gendered words like the disparity between "doctor" and "he" relative to "doctor" and "she" pictured as edge weights in Figure 1a. For language models, we measure the disparities of emission log-likelihood of gender-neutral words conditioned on gendered sentence prefixes as is shown in Figure 1b . Our empirical evaluation with state-of-the-art neural coreference resolution and textbook RNN-based language models Lee et al. [2017], Clark and Manning [2016b], Zaremba et al. [2014] trained on benchmark datasets finds gender bias in these models [1].

Next we turn our attention to mitigating the bias. Bolukbasi et al. [2016] introduced a technique for *debiasing* word embeddings which has been shown to mitigate unwanted associations in analogy tasks while preserving the embedding's semantic properties. Given their widespread use, a natural question is whether this technique is sufficient to eliminate bias from downstream tasks like coreference resolution and language modeling. As our second contribution, we explore this question empirically. We find that while the technique does reduce bias, the residual bias is considerable. We further discover that debiasing models that make use of embeddings that are co-trained with their other parameters [Clark and Manning, 2016b, Zaremba et al., 2014] exhibit a significant drop in accuracy.

Our third contribution is *counterfactual data augmentation (CDA)*: a generic methodology to mitigate bias in neural NLP tasks. For each training instance, the method adds a copy with an *intervention* on its targeted words, replacing each with its partner, while maintaining the same, non-intervened, ground truth. The method results in a dataset of *matched pairs* with ground truth independent of the target distinction (see Figure 1a and Figure 1b for examples). This encourages learning algorithms to not pick up on the distinction.

Our empirical evaluation shows that CDA effectively decreases gender bias while preserving accuracy. We also explore the space of mitigation strategies with CDA, a prior approach to word embedding debiasing (WED), and their compositions. We show that CDA outperforms WED, drastically so when word embeddings are co-trained. For pre-trained embeddings, the two methods can be effectively

[1] Note that these results have practical significance. Both coreference resolution and language modeling are core natural language processing tasks in that they form the basis of many practical systems for information extraction[Zheng et al., 2011], text generation[Graves, 2013], speech recognition[Graves et al., 2013] and machine translation[Bahdanau et al., 2014].

Task / Dataset	Model	Loss via	Trainable embedding	Pre-trained embedding
coreference resolution / CoNLL-2012 [Pradhan et al., 2012]	Lee et al. [2017]	coref. score		✓
	Clark and Manning [2016b]	coref. clusters	✓	✓
language modeling / Wikitext-2 [Merity et al., 2016]	Zaremba et al. [2014]	likelihood	✓	

Table 1: Models, their properties, and datasets evaluated.

composed. We also find that as training proceeds on the original data set with gradient descent the gender bias grows as the loss reduces, indicating that the optimization encourages bias; CDA mitigates this behavior.

In the body of this paper we present necessary background (Section 2), our methods (Sections 3 and 4), their evaluation (Section 5), and speculate on future research (Section 6).

2 Background

In this section we briefly summarize requisite elements of neural coreference resolution and language modeling systems: scoring layers and loss evaluation, performance measures, and the use of word embeddings and their debiasing. The tasks and models we experiment with later in this paper and their properties are summarized in Table 1.

Coreference Resolution The goal of a coreference resolution [Clark and Manning, 2016a] is to group *mentions*, base text elements composed of one or more consecutive words in an input instance (usually a document), according to their semantic identity. The words in the first sentence of Figure 1a, for example, include "the doctor"and "he". A coreference resolution system would be expected to output a grouping that places both of these mentions in the same cluster as they correspond to the same semantic identity.

Neural coreference resolution systems typically employ a *mention-ranking* model [Clark and Manning, 2016a] in which a feed-forward neural network produces a coreference score assigning to every pair of mentions an indicator of their coreference likelihood. These scores are then processed by a subsequent stage that produces clusters.

The ground truth in a corpus is a set of mention clusters for each constituent document. Learning is done at the level of mention scores in the case of Lee et al. [2017] and at the level of clusters in the case of [Clark and Manning, 2016b] . The performance of a coreference system is evaluated in terms of the clusters it produces as compared to the ground truth clusters. As a collection of sets is a partition of the mentions in a document, partition scoring functions are employed, typically MUC, B^3 and $CEAF_{\phi 4}$ [Pradhan et al., 2012], which quantify both precision and recall. Then, standard evaluation practice is to report the average F1 score over the clustering accuracy metrics.

Language Modeling A language model's task is to generalize the distribution of sentences in a given corpus. Given a sentence prefix, the model computes the likelihood for every word indicating how (un)likely it is to follow the prefix in its text distribution. This score can then be used for a variety of purposes such as auto completion. A language model is trained to minimize *cross-entropy loss*, which encourages the model to predict the right words in unseen text.

Word Embedding Word embedding is a representation learning task for finding latent features for a vocabulary based on their contexts in a training corpus. An embedding model transforms syntactic elements (words) into real vectors capturing syntactic and semantic relationships among words.

Bolukbasi et al. [2016] show that embeddings demonstrate bias. Objectionable analogies such as "man is to woman as programmer is to homemaker" indicate that word embeddings pick up on historical biases encoded in their training corpus. Their solution modifies the embedding's parameters so that gender-neutral words no longer carry a gender component. We omit here the details of how the