Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] **Tianlu Wang**[†] **Mark Yatskar**[‡]
Ryan Cotterell[ℵ] **Vicente Ordonez**[†] **Kai-Wei Chang**[§]
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org
[ℵ]University of Cambridge rdc42@cam.ac.uk

Abstract

In this paper, we quantify, analyze and mitigate gender bias exhibited in ELMo's contextualized word vectors. First, we conduct several intrinsic analyses and find that (1) training data for ELMo contains significantly more male than female entities, (2) the trained ELMo embeddings systematically encode gender information and (3) ELMo unequally encodes gender information about male and female entities. Then, we show that a state-of-the-art coreference system that depends on ELMo inherits its bias and demonstrates significant bias on the WinoBias probing corpus. Finally, we explore two methods to mitigate such gender bias and show that the bias demonstrated on WinoBias can be eliminated.

1 Introduction

Distributed representations of words in the form of word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and contextualized word embeddings (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018; McCann et al., 2017; Radford et al., 2019) have led to huge performance improvement on many NLP tasks. However, several recent studies show that training word embeddings in large corpora could lead to encoding societal biases present in these human-produced data (Bolukbasi et al., 2016; Caliskan et al., 2017). In this work, we extend these analyses to the ELMo contextualized word embeddings.

Our work provides a new intrinsic analysis of how ELMo represents gender in biased ways. First, the corpus used for training ELMo has a significant gender skew: male entities are nearly three times more common than female entities, which leads to gender bias in the downloadable pre-trained contextualized embeddings. Then, we apply principal component analysis (PCA) to show that after training on such biased corpora, there exists a low-dimensional subspace that captures much of the gender information in the contextualized embeddings. Finally, we evaluate how faithfully ELMo preserves gender information in sentences by measuring how predictable gender is from ELMo representations of occupation words that co-occur with gender revealing pronouns. Our results show that ELMo embeddings perform unequally on male and female pronouns: male entities can be predicted from occupation words 14% more accurately than female entities.

In addition, we examine how gender bias in ELMo propagates to the downstream applications. Specifically, we evaluate a state-of-the-art coreference resolution system (Lee et al., 2018) that makes use of ELMo's contextual embeddings on WinoBias (Zhao et al., 2018a), a coreference diagnostic dataset that evaluates whether systems behave differently on decisions involving male and female entities of stereotyped or anti-stereotyped occupations. We find that in the most challenging setting, the ELMo-based system has a disparity in accuracy between pro- and anti-stereotypical predictions, which is nearly 30% higher than a similar system based on GloVe (Lee et al., 2017).

Finally, we investigate approaches for mitigating the bias which propagates from the contextualized word embeddings to a coreference resolution system. We explore two different strategies: (1) a training-time data augmentation technique (Zhao et al., 2018a), where we augment the corpus for training the coreference system with its gender-swapped variant (female entities are swapped to male entities and vice versa) and, afterwards, retrain the coreference system; and (2) a test-time embedding neutralization technique, where input contextualized word representations are averaged with word representations of a sentence with entities of the opposite gender. Results show that test-time embedding neutralization is only partially effective, while data augmentation largely mitigates bias demonstrated on WinoBias by the coreference

system.

2 Related Work

Gender bias has been shown to affect several real-world applications relying on automatic language analysis, including online news (Ross and Carter, 2011), advertisements (Sweeney, 2013), abusive language detection (Park et al., 2018), machine translation (Font and Costa-jussà, 2019; Vanmassenhove et al., 2018), and web search (Kay et al., 2015). In many cases, a model not only replicates bias in the training data but also amplifies it (Zhao et al., 2017).

For word representations, Bolukbasi et al. (2016) and Caliskan et al. (2017) show that word embeddings encode societal biases about gender roles and occupations, e.g. engineers are stereotypically men, and nurses are stereotypically women. As a consequence, downstream applications that use these pretrained word embeddings also reflect this bias. For example, Zhao et al. (2018a) and Rudinger et al. (2018) show that coreference resolution systems relying on word embeddings encode such occupational stereotypes. In concurrent work, May et al. (2019) measure gender bias in sentence embeddings, but their evaluation is on the aggregation of word representations. In contrast, we analyze bias in contextualized word representations and its effect on a downstream task.

To mitigate bias from word embeddings, Bolukbasi et al. (2016) propose a post-processing method to project out the bias subspace from the pre-trained embeddings. Their method is shown to reduce the gender information from the embeddings of gender-neutral words, and, remarkably, maintains the same level of performance on different downstream NLP tasks. Zhao et al. (2018b) further propose a training mechanism to separate gender information from other factors. However, Gonen and Goldberg (2019) argue that entirely removing bias is difficult, if not impossible, and the gender bias information can be often recovered. This paper investigates a natural follow-up question: What are effective bias mitigation techniques for contextualized embeddings?

3 Gender Bias in ELMo

In this section we describe three intrinsic analyses highlighting gender bias in trained ELMo contextual word embeddings (Peters et al., 2018). We show that (1) training data for ELMo contains significantly more male entities compared to female entities leading to gender bias in the pre-trained contextual word embeddings (2) the geometry of trained ELMo embeddings systematically encodes gender information and (3) ELMo propagates gender information about male and female entities unequally.

	#occurrence	#M-biased occs.	#F-biased occs.
M	5,300,000	170,000	81,000
F	1,600,000	33,000	36,000

Table 1: Training corpus for ELMo. We show total counts for male (M) and female (F) pronouns in the corpus, and counts corresponding to their co-occurrence with occupation words where the occupations are stereotypically male (M-biased) or female (F-biased).

3.1 Training Data Bias

Table 1 lists the data analysis on the One Billion Word Benchmark (Chelba et al., 2013) corpus, the training corpus for ELMo. We show counts for the number of occurrences of male pronouns (*he*, *his* and *him*) and female pronouns (*she* and *her*) in the corpus as well as the co-occurrence of occupation words with those pronouns. We use the set of occupation words defined in the WinoBias corpus and their assignments as prototypically male or female (Zhao et al., 2018a). The analysis shows that the Billion Word corpus contains a significant skew with respect to gender: (1) male pronouns occur three times more than female pronouns and (2) male pronouns co-occur more frequently with occupation words, irrespective of whether they are prototypically male or female.

3.2 Geometry of Gender

Next, we analyze the gender subspace in ELMo. We first sample 400 sentences with at least one gendered word (e.g., *he* or *she* from the OntoNotes 5.0 dataset (Weischedel et al., 2012) and generate the corresponding gender-swapped variants (changing *he* to *she* and vice-versa). We then calculate the difference of ELMo embeddings between occupation words in corresponding sentences and conduct principal component analysis for all pairs of sentences. Figure 1 shows there are two principal components for gender in ELMo, in contrast to GloVe which only has one (Bolukbasi et al., 2016). The two principal components in ELMo seem to represent the gender from the contextual information (Contextual Gender) as well as the gender embedded in the word itself (Occupational Gender).