# 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).

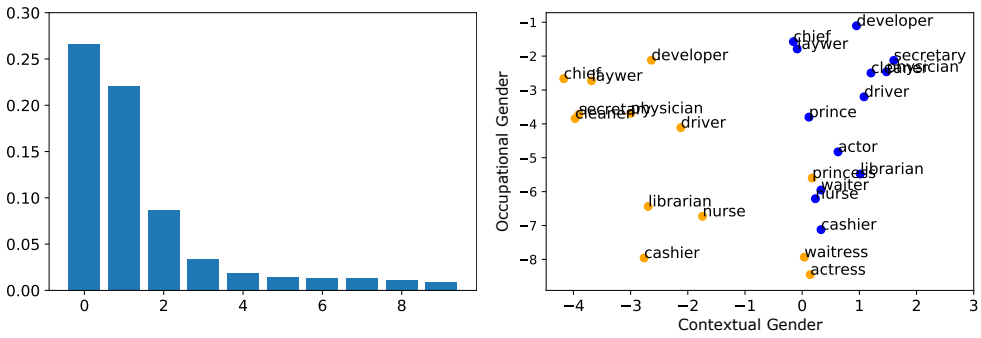


Figure 1: Left: Percentage of explained variance in PCA in the embedding differences. Right: Selected words projecting to the first two principle components where the blue dots are the sentences with male context and the orange dots are from the sentences with female context.

To visualize the gender subspace, we pick a few sentence pairs from WinoBias (Zhao et al., 2018a). Each sentence in the corpus contains one gendered pronoun and two occupation words, such as "The developer corrected the secretary because she made a mistake" and also the same sentence with the opposite pronoun (he). In Figure 1 on the right, we project the ELMo embeddings of occupation words that are co-referent with the pronoun (e.g. *secretary* in the above example) for when the pronoun is male (blue dots) and female (orange dots) on the two principal components from the PCA analysis. Qualitatively, we can see the first component separates male and female contexts while the second component groups male related words such as *lawyer* and *developer* and female related words such as *cashier* and *nurse*.

### 3.3 Unequal Treatment of Gender

To test how ELMo embeds gender information in contextualized word embeddings, we train a classifier to predict the gender of entities from occupation words in the same sentence. We collect sentences containing gendered words (e.g., *he-she*, *father-mother*) and occupation words (e.g., *doctor*)[1] from the OntoNotes 5.0 corpus (Weischedel et al., 2012), where we treat occupation words as a mention to an entity, and the gender of that entity is taken to the gender of a co-referring gendered word, if one exists. For example, in the sentence "the engineer went back to her home," we take *engineer* to be a female mention. Then we split all such instances into training and test, with 539 and 62 instances, respectively and augment these sentences by swapping all the gendered words with words of the opposite gender such that the numbers of male and female entities are balanced.

We first test if ELMo embedding vectors carry gender information. We train an SVM classifier with an RBF kernel[2] to predict the gender of a mention (i.e., an occupation word) based on its ELMo embedding. On development data, this classifier achieves 95.1% and 80.6% accuracy on sentences where the true gender was male and female respectively. For both male and female contexts, the accuracy is much larger than 50%, demonstrating that ELMo does propagate gender information to other words. However, male information is more than 14% more accurately represented in ELMo than female information, showing that ELMo propagates the information unequally for male and female entities.

## 4 Bias in Coreference Resolution

In this section, we establish that coreference systems that depend on ELMo embeddings exhibit significant gender bias. Then we evaluate two simple methods for removing the bias from the systems and show that the bias can largely be reduced.

### 4.1 Setup

We evaluate bias with respect to the WinoBias dataset (Zhao et al., 2018a), a benchmark of paired male and female coreference resolution examples following the Winograd format (Hirst, 1981; Rahman and Ng, 2012; Peng et al., 2015). It contains two different subsets, pro-stereotype, where pronouns are associated with occupations predominately associated with the gender of the pronoun, or anti-stereotype, when the opposite relation is true.

[1] We use the list collected in (Zhao et al., 2018a)

[2] We use the $\nu$-SVC formulation and tune the hyperparameter $\nu$ (Chang and Lin, 2011) in the range of $[0.1, 1]$ with a step 0.1.

| Embeddings | Data Augmentation | Neutralization | | OntoNotes | Semantics Only | | | | w/ Syntactic Cues | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GloVe | ELMo | | Pro. | Anti. | Avg. | \|Diff\| | Pro. | Anti. | Avg. | \|Diff\| |
| GloVe | | | | 67.7 | 76.0 | 49.4 | 62.7 | 26.6* | 88.7 | 75.2 | 82.0 | 13.5* |
| GloVe | ✓ | | | 65.8 | 63.9 | 62.8 | 63.4 | 1.1 | 81.3 | 83.4 | 82.4 | 2.1 |
| GloVe+ELMo | | | | 72.7 | 79.1 | 49.5 | 64.3 | 29.6* | 93.0 | 85.9 | 89.5 | 7.1* |
| GloVe+ELMo | ✓ | | | 71.0 | 65.9 | 64.9 | 65.4 | 1.0 | 87.8 | 88.9 | 88.4 | 1.2 |
| GloVe+ELMo | | ✓ | | 71.0 | 72.6 | 57.8 | 64.9 | 14.3* | 90.2 | 88.6 | 89.4 | 1.6 |
| GloVe+ELMo | | ✓ | ✓ | 71.1 | 71.7 | 60.6 | 66.2 | 11.1* | 90.3 | 89.2 | 89.8 | 1.1 |

Table 2: F1 on OntoNotes and WinoBias development sets. WinoBias dataset is split Semantics Only and w/ Syntactic Cues subsets. ELMo improves the performance on the OntoNotes dataset by 5% but shows stronger bias on the WinoBias dataset. Avg. stands for averaged F1 score on the pro- and anti-stereotype subsets while "Diff." is the absolute difference between these two subsets. * indicates the difference between pro/anti stereotypical conditions is significant ($p < .05$) under an approximate randomized test (Graham et al., 2014). Mitigating bias by data augmentation reduces all the bias from the coreference model to a neglect level. However, the neutralizing ELMo approach only mitigates bias when there are other strong learning signals for the task.

Each subset consists of two types of sentences: one that requires semantic understanding of the sentence to make coreference resolution (Semantics Only) and another that relies on syntactic cues (w/ Syntactic Cues). Gender bias is measured by taking the difference of the performance in pro- and anti-stereotypical subsets. Previous work (Zhao et al., 2018a) evaluated the systems based on GloVe embeddings but here we evaluate a state-of-the-art system that trained on the OntoNotes corpus with ELMo embeddings (Lee et al., 2018).

### 4.2 Bias Mitigation Methods

Next, we describe two methods for mitigating bias in ELMo for the purpose of coreference resolution: (1) a train-time data augmentation approach and (2) a test-time neutralization approach.

**Data Augmentation** Zhao et al. (2018a) propose a method to reduce gender bias in coreference resolution by augmenting the training corpus for this task. Data augmentation is performed by replacing gender revealing entities in the OntoNotes dataset with words indicating the opposite gender and then training on the union of the original data and this swapped data. In addition, they find it useful to also mitigate bias in supporting resources and therefore replace standard GloVe embeddings with bias mitigated word embeddings from Bolukbasi et al. (2016). We evaluate the performance of both aspects of this approach.

**Neutralization** We also investigate an approach to mitigate bias induced by ELMo embeddings without retraining the coreference model. Instead of augmenting training corpus by swapping gender words, we generate a gender-swapped version of the test instances. We then apply ELMo to obtain contextualized word representations of the original and the gender-swapped sentences and use their average as the final representations.

### 4.3 Results

Table 2 summarizes our results on WinoBias.

**ELMo Bias Transfers to Coreference** Row 3 in Table 2 summarizes performance of the ELMo based coreference system on WinoBias. While ELMo helps to boost the coreference resolution F1 score (OntoNotes) it also propagates bias to the task. It exhibits large differences between pro- and anti-stereotyped sets (|Diff|) on both semantic and syntactic examples in WinoBias.

**Bias Mitigation** Rows 4-6 in Table 2 summarize the effectiveness of the two bias mitigation approaches we consider. Data augmentation is largely effective at mitigating bias in the coreference resolution system with ELMo (reducing |Diff | to insignificant levels) but requires retraining the system. Neutralization is less effective than augmentation and cannot fully remove gender bias on the Semantics Only portion of WinoBias, indicating it is effective only for simpler cases. This observation is consistent with Gonen and Goldberg (2019), where they show that entirely removing bias from an embedding is difficult and depends on the manner, by which one measures the bias.

## 5 Conclusion and Future Work

Like word embedding models, contextualized word embeddings inherit implicit gender bias. We analyzed gender bias in ELMo, showing that the corpus it is trained on has significant gender skew and that ELMo is sensitive to gender, but unequally so for male and female entities. We also showed this bias transfers to downstream tasks, such as coreference resolution, and explored two bias mitigation

strategies: 1) data augmentation and 2) neutralizing embeddings, effectively eliminating the bias from ELMo in a state-of-the-art system. With increasing adoption of contextualized embeddings to get better results on core NLP tasks, e.g. BERT (Devlin et al., 2018), we must be careful how such unsupervised methods perpetuate bias to downstream applications and our work forms the basis of evaluating and mitigating such bias.

## Acknowledgement

This work was supported in part by National Science Foundation Grant IIS-1760523. RC was supported by a Facebook Fellowship. We also acknowledge partial support from the Institute of the the Humanities and Global Cultures at the University of Virginia. We thank all reviewers for their comments.

## References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *WMT@ ACL*.

Graeme Hirst. 1981. Anaphora in natural language understanding. *Berlin Springer Verlag*.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Human Factors in Computing Systems*. ACM.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *EMNLP*.

Kenton Lee, Luheng He, and Luke S. Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *NAACL*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *NeurIPS*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *EMNLP*.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *NAACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *EMNLP*.

Karen Ross and Cynthia Carter. 2011. Women and news: A long and winding road. *Media, Culture & Society*, 33(8).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue*, 11(3):10.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *EMNLP*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D Hwang, Claire Bonial, et al. 2012. Ontonotes release 5.0.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *EMNLP*.