Evaluating the Underlying Gender Bias in Contextualized Word Embeddings

Christine Basta **Marta R. Costa-jussà** **Noe Casas**
Universitat Politècnica de Catalunya
{christine.raouf.saad.basta,marta.ruiz,noe.casas}@upc.edu

Abstract

Gender bias is highly impacting natural language processing applications. Word embeddings have clearly been proven both to keep and amplify gender biases that are present in current data sources. Recently, contextualized word embeddings have enhanced previous word embedding techniques by computing word vector representations dependent on the sentence they appear in.

In this paper, we study the impact of this conceptual change in the word embedding computation in relation with gender bias. Our analysis includes different measures previously applied in the literature to standard word embeddings. Our findings suggest that contextualized word embeddings are less biased than standard ones even when the latter are debiased.

1 Introduction

Social biases in machine learning, in general and in natural language processing (NLP) applications in particular, are raising the alarm of the scientific community. Examples of these biases are evidences such that face recognition systems or speech recognition systems work better for white men than for ethnic minorities (Buolamwini and Gebru, 2018). Examples in the area of NLP are the case of machine translation that systems tend to ignore the coreference information in benefit of a stereotype (Font and Costa-jussà, 2019) or sentiment analysis where higher sentiment intensity prediction is biased for a particular gender (Kiritchenko and Mohammad, 2018).

In this work we focus on the particular NLP area of word embeddings (Mikolov et al., 2010), which represent words in a numerical vector space. Word embeddings representation spaces are known to present geometrical phenomena mimicking relations and analogies between words (e.g. *man* is to *woman* as *king* is to *queen*). Following this property of finding relations or analogies, one popular example of gender bias is the word association between *man* to *computer programmer* as *woman* to *homemaker* (Bolukbasi et al., 2016). Pre-trained word embeddings are used in many NLP downstream tasks, such as natural language inference (NLI), machine translation (MT) or question answering (QA). Recent progress in word embedding techniques has been achieved with contextualized word embeddings (Peters et al., 2018) which provide different vector representations for the same word in different contexts.

While gender bias has been studied, detected and partially addressed for standard word embeddings techniques (Bolukbasi et al., 2016; Zhao et al., 2018a; Gonen and Goldberg, 2019), it is not the case for the latest techniques of contextualized word embeddings. Only just recently, Zhao et al. (2019) present a first analysis on the topic based on the proposed methods in Bolukbasi et al. (2016). In this paper, we further analyse the presence of gender biases in contextualized word embeddings by means of the proposed methods in Gonen and Goldberg (2019). For this, in section 2 we provide an overview of the relevant work on which we build our analysis; in section 3 we state the specific request questions addressed in this work, while in section 4 we describe the experimental framework proposed to address them and in section 5 we present the obtained and discuss the results; finally, in section 6 we draw the conclusions of our work and propose some further research.

2 Background

In this section we describe the relevant NLP techniques used along the paper, including word embeddings, their debiased version and contextualized word representations.

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 33–39
Florence, Italy, August 2, 2019.

2.1 Words Embeddings

Word embeddings are distributed representations in a vector space. These vectors are normally learned from large corpora and are then used in downstream tasks like NLI, MT, etc. Several approaches have been proposed to compute those vector representations, with word2vec (Mikolov et al., 2013) being one of the dominant options. Word2vec proposes two variants: continuous bag of words (CBoW) and skipgram, both consisting of a single hidden layer neural network trained on predicting a target word from its context words for CBoW, and the opposite for the skipgram variant. The outcome of word2vec is an embedding table, where a numeric vector is associated to each of the words included in the vocabulary.

These vector representations, which in the end are computed on co-occurrence statistics, exhibit geometric properties resembling the semantics of the relations between words. This way, subtracting the vector representations of two related words and adding the result to a third word, results in a representation that is close to the application of the semantic relationship between the two first words to the third one. This application of analogical relationships have been used to showcase the bias present in word embeddings, with the prototypical example that when subtracting the vector representation of *man* from that of *computer* and adding it to *woman*, we obtain *homemaker*.

2.2 Debiased Word Embeddings

Human-generated corpora suffer from social biases. Those biases are reflected in the co-occurrence statistics, and therefore learned into word embeddings trained in those corpora, amplifying them (Bolukbasi et al., 2016; Caliskan et al., 2017).

Bolukbasi et al. (2016) studied from a geometrical point of view the presence of gender bias in word embeddings. For this, they compute the subspace where the gender information concentrates by computing the principal components of the difference of vector representations of male and female gender-defining word pairs. With the gender subspace, the authors identify direct and indirect biases in profession words. Finally, they mitigate the bias by nullifying the information in the gender subspace for words that should not be associated to gender, and also equalize their distance to both elements of gender-defining word pairs.

Zhao et al. (2018b) proposed an extension to GloVe embeddings (Pennington et al., 2014) where the loss function used to train the embeddings is enriched with terms that confine the gender information to a specific portion of the embedded vector. The authors refer to these pieces of information as *protected attributes*. Once the embeddings are trained, the gender protected attribute can be simply removed from the vector representation, therefore eliminating any gender bias present in it.

The transformations proposed by both Bolukbasi et al. (2016) and Zhao et al. (2018b) are downstream task-agnostic. This fact is used in the work of Gonen and Goldberg (2019) to showcase that, while apparently the embedding information is removed, there is still gender information remaining in the vector representations.

2.3 Contextualized Word Embeddings

Pretrained Language Models (LM) like ULMfit (Howard and Ruder, 2018), ELMo (Peters et al., 2018), OpenAI GPT (Radford, 2018; Radford et al., 2019) and BERT (Devlin et al., 2018), proposed different neural language model architectures and made their pre-trained weights available to ease the application of transfer learning to downstream tasks, where they have pushed the state-of-the-art for several benchmarks including question answering on SQuAD, NLI, cross-lingual NLI and named identity recognition (NER).

While some of these pre-trained LMs, like BERT, use subword level tokens, ELMo provides word-level representations. Peters et al. (2019) and Liu et al. (2019) confirmed the viability of using ELMo representations directly as features for downstream tasks without re-training the full model on the target task.

Unlike word2vec vector representations, which are constant regardless of their context, ELMo representations depend on the sentence where the word appears, and therefore the full model has to be fed with each whole sentence to get the word representations.

The neural architecture proposed in ELMo (Peters et al., 2018) consists of a character-level convolutional layer processing the characters of each word and creating a word representation that is then fed to a 2-layer bidirectional LSTM (Hochreiter and Schmidhuber, 1997), trained on language modeling task on a large corpus.

3 Research questions

Given the high impact of contextualized word embeddings in the area of NLP and the social consequences of having biases in such embeddings, in this work we analyse the presence of bias in these contextualized word embeddings. In particular, we focus on gender biases, and specifically on the following questions:

- Do contextualized word embeddings exhibit gender bias and how does this bias compare to standard and debiased word embeddings?
- Do different evaluation techniques identify bias similarly and what would be the best measure to use for gender bias detection in contextualized embeddings?

To address these questions, we adapt and contrast with the evaluation measures proposed by Bolukbasi et al. (2016) and Gonen and Goldberg (2019).

4 Experimental Framework

As follows, we define the data and resources that we use for performing our experiments. The approach motivation is applying the experiments on contextualized word embeddings.

We worked with the English-German news corpus from the WMT18[1]. We used the English side with 464,947 lines and 1,004,6125 tokens.

To perform our analysis, we used a set of lists from previous work (Bolukbasi et al., 2016; Gonen and Goldberg, 2019). We refer to the list of definitional pairs [2] as 'Definitonal List' (e.g. *she-he*, *girl-boy*). We refer to the list of female and male professions [3] as 'Professional List' (e.g. *accountant*, *surgeon*). The 'Biased List' is the list used in the clustering experiment and it consists of biased male and female words (500 female biased tokens and 500 male biased token). This list is generated by taking the most biased words, where the bias of a word is computed by taking its projection on the gender direction ($\overrightarrow{he}$-$\overrightarrow{she}$) (e.g. *breast-feeding*, *bridal* and *diet* for female and *hero*, *cigar* and *teammates* for male). The 'Extended Biased List' is the list used in classification experiment, which contains 5000 male and female biased tokens, 2500 for each gender, generated in the same way of the Biased List[4]. A note to be considered, is that the lists we used in our experiments (and obtained from Bolukbasi et al. (2016) and Gonen and Goldberg (2019)) may contain words that are missing in our corpus and so we cannot obtain contextualized embeddings for them.

Among different approaches to contextualized word embeddings (mentioned in section 2), we choose ELMo (Peters et al., 2018) as contextualized word embedding approach. The motivation for using ELMo instead of other approaches like BERT (Devlin et al., 2018) is that ELMo provides word-level representations, as opposed to BERT's subwords. This makes it possible to study the word-level semantic traits directly, without resorting to extra steps to compose word-level information from the subwords that could interfere with our analyses.

5 Evaluation measures and results

There is no standard measure for gender bias, and even less for such the recently proposed contextualized word embeddings. In this section, we adapt gender bias measures for word embedding methods from previous work (Bolukbasi et al., 2016) and (Gonen and Goldberg, 2019) to be applicable to contextualized word embeddings.

We start by computing the gender subspace from the ELMo vector representations of gender-defining words, then identify the presence of direct bias in the contextualized representations. We then proceed to identify gender information by means of clustering and classifications techniques. We compare our results to previous results from debiased and non-debiased word embeddings (Bolukbasi et al., 2016) .

Detecting the Gender Space Bolukbasi et al. (2016) propose to identify gender bias in word representations by computing the direction between representations of male and female word pairs from the Definitional List ($\overrightarrow{he}$-$\overrightarrow{she}$, $\overrightarrow{man}$-$\overrightarrow{woman}$) and computing their principal components.

In the case of contextualized embeddings, there is not just a single representation for each word, but its representation depends on the sentence it

[1] http://data.statmt.org/wmt18/translation-task/training-parallel-nc-v13.tgz

[2] https://github.com/tolga-b/debiaswe/blob/master/data/definitional_pairs.json

[3] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json

[4] Both 'Biased List' and 'Extended Biased List' were kindly provided by Hila Gonen to reproduce experiments from her study (Gonen and Goldberg, 2019)

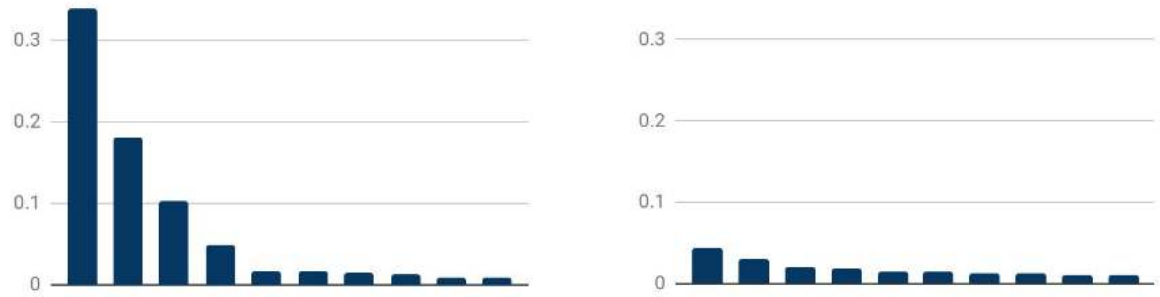

Figure 1: (Left) the percentage of variance explained in the PC of definitional vector differences. (Right) The corresponding percentages for random vectors.

appears in. Hence, in order to compute the gender subspace we take the representation of words by randomly sampling sentences that contain words from the Definitional List and, for each of them, we swap the definitional word with its pair-wise equivalent from the opposite gender. We then obtain the ELMo representation of the definintional word in each sentence pair, computing their difference. On the set of difference vectors, we compute their principal components to verify the presence of bias. In order to have a reference, we computed the principal components of representation of random words.

Similarly to Bolukbasi et al. (2016), figure 1 shows that the first eigenvalue is significantly larger than the rest and that there is also a single direction describing the majority of variance in these vectors, still the difference between the percentage of variances is less in case of contextualized embeddings, which may refer that there is less bias in such embeddings. In the right graph of the figure, we can easily note the difference in the case of random, where the data is not concentrated in a specific direction, as the weight is spread among all components.

A similar conclusion was stated in the recent work (Zhao et al., 2019) where the authors applied the same approach, but for gender swapped variants of sentences with professions. They computed the difference between the vectors of occupation words in corresponding sentences and got a skewed graph where the first component represent the gender information while the second component groups the male and female related words.

Direct Bias Direct Bias is a measure of how close a certain set of words are to the gender vector. To compute it, we extracted from the training data the sentences that contain words in the Professional List. We excluded the sentences that have both a professional token and definitional gender word to avoid the influence of the latter over the presence of bias in the former. We applied the definition of direct bias from Bolukbasi et al. (2016) on the ELMo representations of the professional words in these sentences.

$$\frac{1}{|N|}\sum_{w \epsilon N} |cos(\vec{w}, g)| \quad (1)$$

where N is the amount of gender neutral words, g the gender direction, and $\vec{w}$ the word vector of each profession. We got direct bias of 0.03, compared to 0.08 from standard word2vec embeddings described in Bolukbasi et al. (2016). This reduction on the direct bias confirms that the substantial component along the gender direction that is present in standard word embeddings is less for the contextualized word embeddings. Probably, this reduction comes from the fact that we are using different word embeddings for the same profession depending on the sentence which is a direct consequence and advantage of using contextualized embeddings.

Male and female-biased words clustering. In order to study if biased male and female words cluster together when applying contextualized embeddings, we used k-means to generate 2 clusters of the embeddings of tokens from the Biased list. Note that we cannot use several representations for each word, since it would not make any sense to cluster one word as male and female at the same time. Therefore, in order to make use of the advantages of the contextualized embeddings, we repeated 10 independent experiments, each with a different random sentence of each word from the list of biased male and female words.

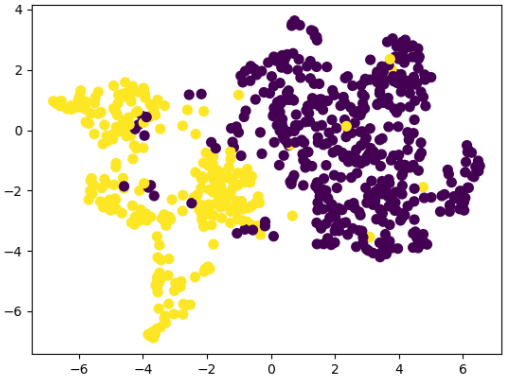

Figure 2: K-means clustering, the yellow color represents the female and the violet represents the male

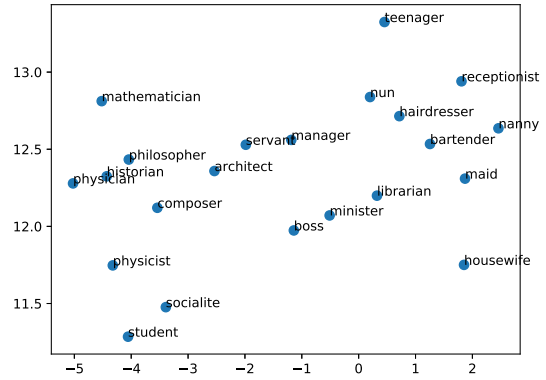

Figure 3: Visualization of contextualized embeddings of professions.

Among these 10 experiments, we got a minimum accuracy of 69.1% and a maximum of 71.3%, with average accuracy of 70.1%, much lower than in the case of biased and debiased word embeddings which were 99.9 and 92.5, respectively, as stated in Gonen and Goldberg (2019). Based on this criterion, even if there is still bias information to be removed from contextualized embeddings, it is much less than in case of standard word embeddings, even if debiased.

The clusters (for one particular experiment out of the 10 of them) are shown in Figure 2 after applying UMAP (McInnes et al., 2018; McInnes et al., 2018) to the contextualized embeddings.

Classification Approach In order to study if contextualized embeddings learn to generalize bias, we trained a Radial Basis Function-kernel Support Vector Machine classifier on the embeddings of random 1000 biased words from the Extended Biased List. After that, we evaluated the generalization on the other random 4000 biased tokens. Again, we performed 10 independent experiments, to guarantee randomization of word representations. Among these 10 experiments, we got a minimum accuracy of 83.33% and a maximum of 88.43%, with average accuracy of 85.56%. This number shows that the bias is learned in these embeddings with high rate. However, it learns in a lower rate than the normal embeddings, whose classification reached 88.88% and 98.25% for debiased and biased versions, respectively.

K-Nearest Neighbor Approach To understand more about the bias in contextualized embeddings, it is important to analyze the bias in the professions. The question is whether these embeddings stereotype the professions as the normal embeddings. This can be shown by the nearest neighbors of the female and male stereotyped professions, for example 'receptionist' and 'librarian' for female and 'architect' and 'philosopher' for male. We applied the k nearest neighbors on the Professional List, to get the nearest k neighbor to each profession. We used a random representation for each token of the profession list, after applying the k nearest neighbor algorithm on each profession, we computed the percentage of female and male stereotyped professions among the k nearest neighbor of each profession token. Afterwards, we computed the Pearson correlation of this percentage with the original bias of each profession. Once again, to assure randomization of tokens, we performed 10 experiments, each with different random sentences for each profession, therefore with different word representations. The minimum Pearson correlation is 0.801 and the maximum is 0.961, with average of 0.89. All these correlations are significant with p-values smaller than 1×10^{-40}. This experiment showed the highest influence of bias compared to 0.606 for debiased embeddings and 0.774 for biased. Figure 3 demonstrates this influence of bias by showing that female biased words (e.g. *nanny*) has higher percent of female words than male ones and vice-versa for male biased words (e.g. *philosopher*).

6 Conclusions and further work

While our study cannot draw clear conclusions on whether contextualized word embeddings augment or reduce the gender bias, our results show more insights into which aspects of the final contextualized word vectors get affected by such phe-

nomena, with a tendency more towards reducing the gender bias rather than the contrary.

Contextualized word embeddings mitigate gender bias when measuring in the following aspects:

1. Gender space, which is capturing the gender direction from word vectors, is reduced for gender specific contextualized word vectors compared to standard word vectors.

2. Direct bias, which is measuring how close set of words are to the gender vector, is lower for contextualized word embeddings than for standard ones.

3. Male/female clustering, which is produced between words with strong gender bias, is less strong than in debiased and non-debiased standard word embeddings.

However, contextualized word embeddings preserve and even amplify gender bias when taking into account other aspects:

1. The implicit gender of words can be predicted with accuracies higher than 80% based on contextualized word vectors which is only a slightly lower accuracy than when using vectors from debiased and non-debiased standard word embeddings.

2. The stereotyped words group with implicit-gender words of the same gender more than in the case of debiased and non-debiased standard word embeddings.

While all measures that we present exhibit certain gender bias, when evaluating future debiasing methods for contextualized word embeddings it would be worth putting emphasis on the latter two evaluation measures that show higher bias than the first three.

Hopefully, our analysis will provide a grain of sand towards defining standard evaluation methods for gender bias, proposing effective debiasing methods or even directly designing equitable algorithms which automatically learn to ignore biased data.

As further work, we plan to extend our study to multiple domains and multiple languages to analyze and measure the impact of gender bias present in contextualized embeddings in these different scenarios.

Acknowledgements

We want to thank Hila Gonen for her support during our research.

This work is supported in part by the Catalan Agency for Management of University and Research Grants (AGAUR) through the FI PhD Scholarship and the Industrial PhD Grant. This work is also supported in part by the Spanish Ministerio de Economa y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigacin, through the postdoctoral senior grant Ramn y Cajal, contract TEC2015-69266-P (MINECO/FEDER,EU) and contract PCIN-2017-079 (AEI/MINECO).

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 77–91.

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

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. *CoRR*, abs/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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tomas Mikolov, Martin Karafit, Luks Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048. ISCA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Alec Radford. 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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Forthcoming in NAACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.