

Second Order WinoBias (SoWinoBias) Test Set for Latent Gender Bias Detection in Coreference Resolution

Hillary Dawkins

University of Guelph, Ontario, Canada
Vector Institute, Toronto, Ontario, Canada
hdawkins@uoguelph.ca

Abstract

We observe an instance of gender-induced bias in a downstream application, despite the absence of explicit gender words in the test cases. We provide a test set, SoWinoBias, for the purpose of measuring such latent gender bias in coreference resolution systems. We evaluate the performance of current debiasing methods on the SoWinoBias test set, especially in reference to the method’s design and altered embedding space properties. See <https://github.com/hillary-dawkins/SoWinoBias>.

1 Introduction

Explicit (or first-order) gender bias was observed in coreference resolution systems by [Zhao et al. \(2018a\)](#), by considering contrasting cases:

1. The doctor hired the secretary because he was overwhelmed. [he → doctor]
2. The doctor hired the secretary because she was overwhelmed. [she → doctor]
3. The doctor hired the secretary because she was highly qualified. [she → secretary]
4. The doctor hired the secretary because he was highly qualified. [he → secretary]

Sentences 1 and 3 are pro-stereotypical examples because gender words align with a socially-held stereotype regarding the occupations. Sentences 2 and 4 are anti-stereotypical because the correct coreference resolution contradicts a stereotype. It was observed that systems performed better on pro cases than anti cases, and the WinoBias test set was developed to quantify this disparity.

Here we make a new observation of gender-induced (or second-order) bias in coreference resolution systems, and provide the corresponding test set SoWinoBias. Consider cases:

1. The doctor liked the nurse because they were beautiful. [they → nurse]
2. The nurse dazzled the doctor because they were beautiful. [they → nurse]
3. The nurse admired the doctor because they were beautiful. [they → doctor]

The examples do not contain any explicit gender cues at all, and yet we can observe that sentences 1 and 2 align with a gender-induced social stereotype, while sentence 3 opposes the stereotype. The induction occurs because “nurse” is a female-coded occupation ([Bolukbasi et al., 2016](#); [Zhao et al., 2018b](#)), and women are also more likely to be described based on physical appearance ([Hoyle et al., 2019](#); [Williams and Bennett, 1975](#)). A coreference resolution system is gender-biased if correct predictions on sentences like 1 and 2 are more likely than on sentence 3.

The difference between first-order and second-order gender bias in a downstream application is especially interesting given current trends in debiasing static word embeddings. Early methods ([Bolukbasi et al., 2016](#); [Zhao et al., 2018b](#)) focused on eliminating direct bias from the embedding space, quantified as associations between gender-neutral words and an explicit gender vocabulary. In response to an influential critique paper by [Gonen and Goldberg \(2019\)](#), the current trend is to focus on eliminating indirect bias from the embedding space, quantified either by gender-induced proximity among embeddings ([Kumar et al., 2020](#)) or by residual gender cues that could be learned by a classifier ([Ravfogel et al., 2020](#); [Davis et al., 2020](#)).

Indirect bias in the embedding space was viewed as an undesirable property a priori, but we do not yet have a good understanding of the effect on downstream applications. Here we test debiasing methods from both camps on SoWinoBias, and

make a series of observations on sufficient and necessary conditions for mitigating the latent gender-biased coreference resolution.

Additionally, we consider the case that our coreference resolution model employs both static and contextual word embeddings, but debiasing methods are applied to the static word embeddings only. Post-processing debiasing techniques applied to static word embeddings are computationally inexpensive, easy to concatenate, and have a longer development history. However contemporary models for downstream applications are likely to use some form of contextual embeddings as well. Therefore we might wonder whether previous work in debiasing static word embeddings remains relevant in this setting. The WinoBias test set for instance was developed and tested using the “end-to-end” coreference resolution model (Lee et al., 2017), a state-of-the-art model at that time using only static word embeddings. Subsequent debiasing schemes reported results on WinoBias using the same model, just plugging in different debiased embeddings, for the sake of fair comparison. However this is becoming increasingly outdated given the progress in coreference resolution systems. A contribution of this work is to report WinoBias results for previous debiasing techniques using a more updated model, one that makes use of unaltered contextual embeddings in addition to the debiased static embeddings.

The remainder of the paper is organized as follows: In section 2, we further define the type of bias being measured by the SoWinoBias test set and discuss some limitations. In section 3, we review the 4 word embedding debiasing methods that we will analyze, in the context of how each method aims to alter the word embedding space. In section 4, we provide details of the experimental setup and report results on both coreference resolution test sets, the original WinoBias and the newly constructed SoWinoBias. In section 5, we discuss the results with respect to the geometric properties of the altered embedding spaces. In particular, we review whether mitigation of intrinsic measures of bias on the embedding space, quantified as direct bias and indirect bias by various definitions, are related to mitigation of the latent bias in a downstream application.

2 Bias Statement

Within the scope of this paper, bias is defined and quantified as the difference in performance of a

coreference resolution system on test cases aligning with a socially-held stereotype vs. test cases opposing a socially-held stereotype. We observe that gender-biased systems perform significantly better in pro-stereotypical situations. Such difference in performance creates representational harm by implying (for example) that occupations typically associated with one gender cannot have attributes typically associated with another.

Throughout this paper, the term “second-order” is used interchangeably with “latent”. Characterizing the observed bias as “second-order” follows from the observation of a gender-induced bias in the absence of gender-definitional vocabulary, resting on the definition of “they” as a gender-neutral pronoun.

Therefore, a limitation in the test set construction is the possible semantic overloading of “they”. As discussed, the intention throughout this paper is to use the singular “they” as a pronoun that does not carry any gender information (and could refer to someone of any gender). However, different contexts may choose to treat “they” exclusively as a non-binary gender pronoun.

The gender stereotypes used throughout this paper are sourced from peer-reviewed academic journals written in English, which draw from the US Labor Force Statistics, as well as US-based crowd workers. Therefore a limitation may be that stereotypes used here are not common to all languages or cultures.

3 Debiasing methods

3.1 Neutralization of static word embeddings

3.1.1 Methods addressing direct bias

The first attempts to debias word embeddings focused on the mitigation of direct bias (Bolukbasi et al., 2016). The definition of direct bias assumes the presence of a “gender direction” \vec{g} ; a subspace that mostly encodes the difference between the binary genders. A non-zero projection of word \vec{w} onto \vec{g} implies that \vec{w} is more similar to one gender over another. In the case of ideally gender-neutral words, this is an undesirable property. Direct bias quantifies the extent of this uneven similarity¹:

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

¹The original definition included a strictness exponent c , here set to 1 as has commonly been done in subsequent works.

The Hard Debias method (Bolukbasi et al., 2016) is a post-processing technique that projects all gender-neutral words into the nullspace of \vec{g} . Therefore, the direct bias is made to be zero by definition. We measure the performance of **Hard-GloVe**² on the coreference resolution tasks.

A related retraining method used a modified version of GloVe’s original objective function with additional incentives to reduce the direct bias for gender-neutral words, resulting in the GN-GloVe embeddings (Zhao et al., 2018b). Rather than allowing for gender information to be distributed across the entire embedding space, the method explicitly sequesters the protected gender attribute to the final component. Therefore the first $d - 1$ components are taken as the gender-neutral embeddings, denoted **GN-GloVe**(w_a)³.

3.1.2 Methods addressing indirect bias

The indirect bias is less well defined, and loosely refers to the gender-induced similarity measure between gender-neutral words. For instance, semantically unrelated words such as “sweetheart” and “nurse” may appear quantitatively similar due to a shared gender association.

One definition (first given in (Bolukbasi et al., 2016)) measures the relative change in similarity after removing direct gender associations as

$$\beta(\vec{w}, \vec{v}) = \frac{1}{\vec{w} \cdot \vec{v}} \left(\vec{w} \cdot \vec{v} - \frac{\vec{w}_\perp \cdot \vec{v}_\perp}{\|\vec{w}_\perp\| \|\vec{v}_\perp\|} \right), \quad (2)$$

where $\vec{w}_\perp = \vec{w} - (\vec{w} \cdot \vec{g})\vec{g}$, however this relies on a limited definition of the original gender association.

The Repulse-Attract-Neutralize (RAN) debiasing method attempts to repel undue gender proximities among gender-neutral words, while keeping word embeddings close to their original learned representations (Kumar et al., 2020). This method quantifies indirect bias by incorporating β into a graph-weighted holistic view of the embedding space (more on this later). In this paper, we will measure the performance of **RAN-GloVe**⁴ on the coreference resolution tasks.

A related notion of indirect bias is to measure whether gender associations can be predicted from the word representation. The Iterative Nullspace

²Hard debias: <https://github.com/tolga-b/debiaswe>. All base (unbiased) embeddings are GloVe trained on the 2017 January Wikipedia dump (vocab contains 322,636 tokens). Available at <https://github.com/uclanlp/gnglove>, based on the work of Pennington et al. (2014).

³<https://github.com/uclanlp/gnglove>

⁴<https://github.com/TimeTraveller-San/RAN-Debias>

Linear Projection method (INLP) achieves linear guarding of the gender attribute by iteratively learning the most informative gender subspace for a classification task, and projecting all words to the orthogonal nullspace (Ravfogel et al., 2020). After sufficient iteration, gender information cannot be recovered by a linear classifier. We will measure the performance of **INLP-GloVe**⁵.

3.2 Data augmentation

In addition to debiasing methods applied to word embeddings, we measure the effect of simple data augmentation applied to the training data for our coreference resolution system. The goal is to determine whether data augmentation can complement the debiased word embeddings on this particular test set. The training data is augmented using a simple gender-swapping protocol, such that binary gender words are replaced by their equivalent form of the opposite gender (e.g. “he” \leftrightarrow “she”, etc.).

4 Detection of gender bias in coreference resolution: Experimental setup

All systems were built using the “Higher-order coreference resolution with coarse-to-fine inference” model (Lee et al., 2018)⁶. It is important to keep in mind that this model uses both static word embeddings and contextual word embeddings (specifically ELMo embeddings (Peters et al., 2018)). Our experimental debiasing methods were applied to static word embeddings only, and contextual embeddings are left unaltered in all cases.

All systems were trained using the OntoNotes 5.0⁷ train and development sets, using the default hyperparameters⁸, for approximately 350,000 steps until convergence. Baseline performance was tested using the OntoNotes 5.0 test set (results shown in Table 1). Baseline performance is largely consistent across all models, indicating that neither debiased word embeddings nor gender-swapped training data significantly degrades the performance of the system overall.

4.1 WinoBias

The WinoBias test set was created by Zhao et al. (2018a), and measures the performance of coreference systems on test cases containing explicit bi-

⁵<https://github.com/shauli-ravfogel/nullspaceprojection>

⁶<https://github.com/kentonl/e2e-coref>

⁷<https://catalog.ldc.upenn.edu/LDC2013T19>

⁸“best” configuration at <https://github.com/kentonl/e2e-coref/blob/master/experiments.conf>

Table 1: Results on coreference resolution test sets. OntoNotes (F_1) performance provides a baseline for “vanilla” coreference resolution ($n = 348$). WinoBias (F_1) measures explicit gender bias, observable as the diff. between pro ($n = 396$) and anti ($n = 396$) test sets. SoWinoBias (% accuracy) measures second-order gender bias, likewise observable as the diff. between pro ($n = 4096$) and anti ($n = 4096$) test sets. Note: accuracy is the relevant metric to report on the SoWinoBias test set, rather than F_1 , due to our assertion that “they” is not a new entity mention.

Embedding	Data Aug.	OntoNotes	WinoBias				SoWinoBias			
			pro	anti	avg.	diff.	pro	anti	avg.	diff.
GloVe		72.3	77.8	48.8	63.8	29.0	64.2	46.8	55.5	17.4
GloVe	✓	72.0	67.0	59.0	63.0	8.0	62.8	56.5	59.7	6.4
Hard-GloVe		72.2	66.5	59.1	62.8	7.4	63.6	49.2	56.4	14.3
Hard-GloVe	✓	71.8	64.0	61.9	63.0	2.1	77.1	50.1	63.6	27.0
GN-GloVe(w_a)		72.2	63.4	61.1	62.3	2.3	68.0	49.7	58.9	18.3
GN-GloVe(w_a)	✓	71.4	59.0	66.0	62.5	7.0	72.1	69.7	70.9	2.4
RAN-GloVe		72.4	72.8	53.2	63.0	19.6	70.2	60.0	65.1	10.2
RAN-GloVe	✓	71.1	60.1	63.8	62.0	3.7	69.5	59.4	64.5	10.0
INLP-GloVe		71.6	67.5	57.5	62.5	10.0	68.4	46.1	57.3	22.4
INLP-GloVe	✓	72.1	66.2	59.1	62.7	7.1	73.4	65.1	69.3	8.3

nary gender words. In particular, pro-stereotypical sentences contain coreferents where an explicit gender word (e.g. he, she) is paired with an occupation matching a socially held gender stereotype. Anti-stereotypical sentences use the same formulation but gender swap the explicit gender words such that coreferents now oppose a socially held gender stereotype. Gender bias is measured as the difference in performance on the pro. versus anti. test sets, each containing $n = 396$ sentences.

Recall that here we are reporting WinoBias results using a system incorporating unaltered contextual embeddings, in addition to the debiased static embeddings. Previously reported results on the “end-to-end” coreference model (Lee et al., 2017), using only debiased static word embeddings, are compiled in the Appendix for reference.

In this setting, we observe that debiasing methods addressing direct bias are more successful than those addressing indirect bias. In particular, without the additional resource of data augmentation, RAN-GloVe struggles to reduce the difference between pro and anti test sets (in contrast to RAN-GloVe’s great success in the end-to-end model setting, as reported by Kumar et al. (2020)). Data augmentation is found to be a complementary resource, providing further gains in most cases. Overall, Hard-GloVe with simple data augmentation successfully reduces the difference in F_1 from 29% to 2.1%, while not significantly degrading the average performance on WinoBias or baseline performance on OntoNotes. This suggests that debiasing the con-

textual word embeddings is not needed to mitigate the explicit gender bias in coreference resolution, as measured by this particular test set.

4.2 SoWinoBias

The SoWinoBias test set measures second-order, or latent, gender associations in the absence of explicit gender words. At present, we measure associations between male and female stereotyped occupations with female stereotyped adjectives, although this could easily be extended in the future. Adjectives with positive and negative polarities are represented evenly in the test set. We will denote the vocabularies of interest as

$$M_{occ} = \{\text{doctor, boss, developer, ...}\} \quad (3)$$

$$F_{occ} = \{\text{nurse, nanny, maid, ...}\}$$

$$F_{adj}^+ = \{\text{lovely, beautiful, virtuous, ...}\}$$

$$F_{adj}^- = \{\text{hysterical, unmarried, prudish, ...}\},$$

where $|M_{occ}| = |F_{occ}| = |F_{adj}^+| = |F_{adj}^-| = 16$, and the full sets can be found in the appendix. Stereotypical occupations were sourced from the original WinoBias vocabulary (drawing from the US labor occupational statistics), as well as the SemBias (Zhao et al., 2018b) and Hard Debias analogy test sets (drawing from human-annotated judgements). Stereotypical adjectives with polarity were sourced from the latent gendered-language model of Hoyle et al. (2019), which was found to be consistent with the human-annotated corpus of Williams and Bennett (1975).

SoWinoBias test sentences are constructed as “The [occ1] (dis)liked the [occ2] because they were [adj]”, where “(dis)liked” is matched appropriately to the adjective polarity, such that “they” always refers to “occ2”. Each sentence selects one occupation from M_{occ} , and the other from F_{occ} . In pro-stereotypical sentences, $occ2 \in F_{occ}$, such that the adjective describing the (they, occ2) entity matches a social stereotype. In anti-stereotypical sentences, $occ2 \in M_{occ}$, such that the adjective describing the (they, occ2) entity contradicts a social stereotype. Example sentences in the test set include:

1. The doctor liked the nurse because they were beautiful. (pro)
2. The nurse liked the doctor because they were beautiful. (anti)
3. The ceo disliked the maid because they were unmarried. (pro)
4. The maid disliked the lawyer because they were unmarried. (anti)

In total, there are $n = 4096$ sentences in each of the pro and anti test sets. Due to the simplicity of our constructed sentences, plus our desire to measure gendered associations, we further assert that “they” should refer to one of the two potential occupations (i.e. “they” cannot be predicted as a new entity mention). As with WinoBias, gender bias is observed as the difference in performance between the anti and pro test sets.

Firstly, we observe that the second-order gender bias is more difficult to the correct than the explicit bias, given access to the debiased embeddings alone. Methods that made good progress in reducing the WinoBias diff. make little to no progress on the SoWinoBias diff. However, even simple data augmentation was found to be a valuable resource. When combined with GN-GloVe(w_a), the difference is reduced to 2.4% while increasing average performance significantly. Again, we observe that good bias reduction can be achieved, even before incorporating methods to debias the contextual word embeddings. It is interesting that debiasing methods explicitly designed to address indirect bias in the embedding space do not do better at mitigating second-order bias in a downstream task. Further discussion in relation to the embedding space properties is provided in the following section.

5 Relationship to embedding space properties

5.1 Single-attribute WEAT

The Word Embedding Association Test (WEAT) measures the association strength between two concepts of interest (e.g. arts vs. science) relative to two defined attribute groups (e.g. female vs. male) (Caliskan et al., 2017). It was popularized as a means for detecting gender bias in word embeddings by showing that (arts, science), (arts, math), and (family, careers) produced significantly different association strengths relative to gender.

Here we adapt the original WEAT to measure relative association across genders given a single concept of interest. This provides a means to measure whether the set of female-stereotyped adjectives F_{adj} are quantitatively gender-marked in the embedding space.

The relative association of a single word t across attribute sets A_1, A_2 is given by

$$s(t, A_1, A_2) = \frac{1}{|A_1|} \sum_{a_1 \in A_1} \cos(t, a_1) - \frac{1}{|A_2|} \sum_{a_2 \in A_2} \cos(t, a_2) \quad (4)$$

where $s(t, A_1, A_2) > 0$ indicates that t is more closely related to attribute A_1 than A_2 . The average relative association of concept T is then

$$S(T, A_1, A_2) = \frac{1}{|T|} \sum_{t \in T} s(t, A_1, A_2). \quad (5)$$

The significance of a non-zero association strength can be assessed by a partition test. We randomly sample alternate attribute sets of equal size A_1^* and A_2^* from the union of the original attribute sets. The significance p is defined as the proportion of samples to produce $S(T, A_1^*, A_2^*) > S(T, A_1, A_2)$. Small p values indicate that the defined grouping of the attributes sets (here defined by gender) are meaningful compared to random groupings.

Table 2 shows the results of the single-attribute WEAT. We measure association strength of the female adjectives relative to gender in two ways: i) gender is defined using a “definitional” vocabulary ($A_1 = F_{def} = \{she, her, woman, \dots\}$, $A_2 = M_{def} = \{he, him, man, \dots\}$), and ii) gender is defined using a latent vocabulary – the stereotypical occupations ($A_1 = F_{occs}$, $A_2 = M_{occs}$).

As shown, the F_{adj} embeddings are strongly associated with the explicit gender vocabulary in

Table 2: Single-Attribute WEAT association strength between gender and female-stereotyped adjectives with significance values. Lower association strength (S) values are better. Smaller significance values indicate that the observed association strength is meaningful with respect to gender.

Embedding	$S(F_{adj}, F_{occ}, M_{occ})$	Significance	$S(F_{adj}, F_{def}, M_{def})$	Significance
GloVe	0.0636	0.0001**	0.0694	0.001**
Hard-GloVe	0.0465	0.0001**	-8.6889e-10	0.512
GN-GloVe(w_a)	0.0664	0.0003**	-0.0015	0.436
RAN-GloVe	0.0402	0.0003**	0.0153	0.177
INLP-GloVe	0.0171	0.0251*	0.0054	0.382

the original GloVe space. However each of the four debiasing methods are successful in removing the explicit gender association, as expected. The Hard Debias method in particular asserts $S(F_{adj}, F_{def}, M_{def}) = 0$ by definition.

In contrast, the F_{adj} embeddings are just as strongly associated with the latent gender vocabulary in the original GloVe space, but this is not undone by any of the debiasing methods. This is somewhat of an unexpected result in the case of the RAN and INLP debiasing methods, as they promised to go beyond direct bias mitigation.

The INLP method makes the most progress in reducing the implicit association strength, however a significant non-zero association remains. Combined with the SoWinoBias test results, we can observe that the WEAT reduction achieved by INLP is not a sufficient condition for mitigating latent gender-biased coreference resolution. Inversely, we observe that reduction of the WEAT measure is not a necessary condition for mitigation when debiased embeddings are combined with data augmentation (demonstrated by GN-GloVe(w_a)).

5.2 Clustering and Recoverability

Clustering and recoverability (C&R) (Gonen and Goldberg, 2019) refer to a specific observation on the embedding space post debiasing; namely, that gender labels of words (assigned according to direct bias in the original embedding space) can be classified with a high degree of accuracy given only the debiased representations. Here we follow the same experimental setup, and report results on an expanded set of embeddings (see Table 3).

In agreement with Gonen and Goldberg (2019), we find that the Hard-GloVe and GN-GloVe embeddings retain nearly perfect recoverability of the original gender labels, indicating high levels of residual bias by this definition.

The INLP method was designed to guard against

linear recoverability, and indeed we find that both C&R by a linear SVM are reduced to near-random performance. Recoverability by an SVM with a non-linear kernel (rbf) achieves 75% accuracy; much reduced compared to other debiasing methods, but still above the baseline of 50%. This result is consistent with Ravfogel et al. (2020).

Of interest are the results obtained for the RAN-GloVe embeddings, which have not previously been reported. RAN was designed to mitigate undue proximity bias, conceptually similar to clustering. Despite this, C&R are still possible with high accuracy given RAN-debiased embeddings. Given RAN’s success on various gender bias assessment tasks (SemBias, and WinoBias using the end-to-end coreference model), this suggests that complete suppression of C&R is unnecessary for many practical applications. Conversely, it may indicate that we have not yet developed any assessment tasks that probe the effect of indirect bias.

In reference to the SoWinoBias results, we can observe that linear attribute guarding (achieved by INLP) is not a sufficient condition for mitigating latent gender-biased coreference resolution. However, even linear guarding is not a necessary condition for mitigating SoWinoBias when retraining with data augmentation is available.

5.3 Gender-based Illicit Proximity Bias

The gender-based illicit proximity bias (GIPE) was proposed by Kumar et al. (2020) as a means to capture indirect bias on the embedding space as a well-defined metric, as opposed to the loosely defined idea of clustering and recoverability. Firstly, the gender-based proximity bias of a single word w , denoted $\eta(w)$, is defined as the proportion of N -nearest neighbours $\{n_i\}$ with indirect bias $\beta(n_i, w)$ above some threshold θ . Intuitively, this is the proportion of words that are close by solely due to a shared gender association. The GIPE extends this

Table 3: Clustering: (reported as accuracy and v -measure (Rosenberg and Hirschberg, 2007)) is performed by taking the $n = 1500$ most biased words in the original embedding space (excluding definitional gender words), and performing k-means clustering ($k = 2$) on the same words in the debiased space. Recoverability: (reported as accuracy) is performed by taking the $n = 5000$ most biased words in the original embedding space, and training a classifier (linear SVM or rbf kernel SVM) on the same words in the debiased space. Smaller values are better (indicating less residual cues that can be used classify gender-neutral words). GIPE: Smaller values are better (indicating less undue proximity bias in the embedding space).

Embedding	Acc.	v -measure	linSVM	rbfSVM	GIPE(V_d)	GIPE(V_{So})	Avg. $\eta(w_{So})$
GloVe	99.8	98.4	100	100	0.1153	0.1844	0.1373
Hard-GloVe	79.0	30.2	92.5	94.6	0.0701	0.1020	0.0894
GN-GloVe(w_a)	85.3	49.7	99.1	99.4	0.1173	0.1650	0.1167
RAN-GloVe	80.4	41.9	95.3	96.0	0.0399	0.0827	0.0617
INLP-GloVe	57.1	1.52	52.9	74.8	0.0798	0.1265	0.0967

word-level measure to a vocabulary-level measure using a weighted average over $\eta(w)$.

Table 3 shows the GIPE measure on the entire gender-neutral vocabulary V_d , the gender-neutral vocabulary used to construct SoWinoBias $V_{So} = F_{occ} \cup M_{occ} \cup F_{adj}$, and the simple (unweighted) average $\eta(w_{So})$ on the SoWinoBias vocabulary.

The RAN method mitigates indirect bias as measured by GIPE by design, and therefore achieves the lowest GIPE values as expected (followed by Hard-GloVe, somewhat unexpectedly). However, non-zero proximity bias persists, more so on the stereotyped sub-vocabulary than the total vocabulary. Without extra help from data augmentation, RAN-GloVe achieves the best performance on the SoWinoBias (followed by Hard-GloVe). Therefore further reduction of GIPE may enable further mitigation of the latent gender-biased coreference resolution (cannot be ruled out as a sufficient condition at this time). However, RAN-GloVe does not benefit from the addition of data augmentation, unlike the majority of debiasing methods. Further investigation is needed to determine what conditions of the embedding properties allow for complementary data augmentation.

6 Conclusion

In this paper, we demonstrate the existence of observable latent gender bias in a downstream application, coreference resolution. We provide the first gender bias assessment test set not containing any explicit gender-definitional vocabulary. Although the present study is limited to binary gender, this construction should allow us to assess gender bias (or other demographic biases) in cases where explicit defining vocabulary is limited or unavailable. However, the construction does depend on knowl-

edge of expected relationships or stereotypes (here occupations and adjectives). Therefore interdisciplinary work drawing from social sciences is encouraged as a future direction.

Our observations indicate that mitigation of indirect bias in the embedding space, according to our current understanding of such a notion, does not reduce the latent associations in the embedding space (as measured by WEAT), nor does it mitigate the downstream latent bias (as measured by SoWinoBias). Future work could seek bias assessment tasks in downstream applications that do depend on the reduction of gender-based proximity bias or non-linear recoverability. Currently the motivation for such reduction is unknown, despite being an active direction of debiasing research.

Finally, we do observe that an early debiasing method, GN-GloVe, combined with simple data augmentation, can mitigate the latent gender biased coreference resolution, even when contextual embeddings in the system remain unaltered. Future work could extend the idea of the SoWinoBias test set to more complicated sentences representative of real “in the wild” cases, in order to determine if this result holds.

The SoWinoBias test set, all trained models presented in this paper, and code for reproducing the results are available at <https://github.com/hillary-dawkins/SoWinoBias>.

Acknowledgements

We thank Daniel Gillis, Judi McCuaig, Stefan Kremer, Graham Taylor, and anonymous reviewers for their time and thoughtful discussion. This work is financially supported by the Government of Canada through the NSERC CGS-D program (CGSD3-518897-2018).

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 *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357. Curran Associates, Inc.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. [Semantics derived automatically from language corpora contain human-like biases](#). *Science*, 356(6334):183–186.
- Brent D. Davis, Ethan Jackson, and Daniel J. Lizotte. 2020. [Decision-directed data decomposition](#).
- 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](#). In *Proceedings of NAACL-HLT*.
- Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. [Unsupervised discovery of gendered language through latent-variable modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.
- Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. [Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings](#). *Transactions of the Association for Computational Linguistics*, 8:486–503.
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. [End-to-end neural coreference resolution](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.
- Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. [Higher-order coreference resolution with coarse-to-fine inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Matthew E. 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.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.
- Andrew Rosenberg and Julia Hirschberg. 2007. [V-measure: A conditional entropy-based external cluster evaluation measure](#). In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.
- John E. Williams and Susan M. Bennett. 1975. [The definition of sex stereotypes via the adjective check list](#). *Sex Roles*, 1:327–337.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. [Learning gender-neutral word embeddings](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

A Full test set vocabulary

$F_{occ} = \{ \text{writer, teacher, cleaner, tailor, attendant, librarian, auditor, nurse, nanny, cashier, editor, hairdresser, stylist, maid, baker, counselor} \}$

$M_{occ} = \{ \text{guard, architect, chef, leader, president, developer, lawyer, salesperson, doctor, judge, boss, chief, mover, cook, researcher, physician} \}$

$F_{adj}^+ = \{ \text{sprightly, gentle, affectionate, charming, kindly, beloved, enchanted, virtuous, beautiful, chaste, fair, delightful, lovely, romantic, elegant, fertile} \}$

$F_{adj}^- = \{ \text{fussy, nagging, rattlebrained, haughty, whiny, dependent, sullen, unmarried, prudish, fickle, hysterical, infected, widowed, awful, damned, frivolous} \}$

$M_{def} = \{ \text{man, he, father, brother, his, son, uncle, himself} \}$

$F_{def} = \{ \text{woman, she, mother, sister, her, daughter, aunt, herself} \}$

Table 4: Results on SoWinoBias test set by adjective polarity.

Embedding	Data Aug.	Positive Adj.			Negative Adj.			Total		
		pro	anti	diff.	pro	anti	diff.	pro	anti	diff.
GloVe		69.4	49.2	20.1	58.9	44.3	14.6	64.2	46.8	17.4
GloVe	✓	64.2	60.4	3.9	61.4	52.6	8.8	62.8	56.5	6.4
Hard-GloVe		64.6	49.8	14.7	62.6	48.7	13.9	63.6	49.2	14.3
Hard-GloVe	✓	77.2	51.5	25.8	76.9	48.7	28.2	77.1	50.1	27.0
GN-GloVe(w_a)		71.6	52.9	18.6	64.4	46.5	17.9	68.0	49.7	18.3
GN-GloVe(w_a)	✓	71.5	70.5	1.0	72.7	69.0	3.7	72.1	69.7	2.4
RAN-GloVe		70.9	61.5	9.4	69.4	58.5	11.0	70.2	60.0	10.2
RAN-GloVe	✓	73.6	67.0	6.7	65.3	51.9	13.4	69.5	59.4	10.0
INLP-GloVe		74.2	54.0	20.2	62.7	38.2	24.5	68.4	46.1	22.4
INLP-GloVe	✓	76.4	67.9	8.5	70.4	62.3	8.2	73.4	65.1	8.3

Table 5: Previously reported results on the OntoNotes (baseline) and WinoBias test sets by various debiasing methods when the coreference system was built using the “end-to-end” model (Lee et al., 2017). RAN-GloVe drastically outperforms all methods.

Embedding	OntoNotes	WinoBias			
		pro	anti	avg.	diff.
GloVe	66.5	76.2	46.0	61.1	30.2
Hard-GloVe	66.2	70.6	54.9	62.8	15.7
GN-GloVe	66.2	72.4	51.9	62.2	20.5
GN-GloVe(w_a)	65.9	70.0	53.9	62.0	16.1
RAN-GloVe	66.2	61.4	61.8	61.6	0.4