Towards Understanding Gender Bias in Neural Relation Extraction

Andrew Gaut[*†], **Tony Sun**[*†], **Shirlyn Tang**[†], **Yuxin Huang**[†],
Jing Qian[†], **Mai ElSherief**[††], **Jieyu Zhao**[‡],
Diba Mirza[†], **Elizabeth Belding**[†], **Kai-Wei Chang**[‡], and **William Yang Wang**[†]

[†]Department of Computer Science, UC Santa Barbara
[‡]Department of Computer Science, UC Los Angeles
[††]School of Interactive Computing, Georgia Institute of Technology
{ajg, tonysun, shirlyntang, yuxinhuang}@ucsb.edu
{jing_qian, dimirza, ebelding, william}@cs.ucsb.edu
melsherief@gatech.edu
{jyzhao, kwchang}@cs.ucla.edu

Abstract

Recent developments in Neural Relation Extraction (NRE) have made significant strides towards automated knowledge base construction. While much attention has been dedicated towards improvements in accuracy, there have been no attempts in the literature to evaluate social biases exhibited in NRE systems. In this paper, we create WikiGenderBias, a distantly supervised dataset composed of over 45,000 sentences including a 10% human annotated test set for the purpose of analyzing gender bias in relation extraction systems. We find that when extracting spouse and hypernym (i.e., occupation) relations, an NRE system performs differently when the gender of the target entity is different. However, such disparity does not appear when extracting relations such as birth date or birth place. We also analyze two existing bias mitigation techniques, word embedding debiasing and data augmentation. Unfortunately, due to NRE models relying heavily on surface level cues, we find that existing bias mitigation approaches have a negative effect on NRE. Our analysis lays groundwork for future quantifying and mitigating bias in relation extraction.

1 Introduction

With the wealth of information being posted online daily, relation extraction has become increasingly important. Relation extraction aims specifically to extract relations from raw sentences and represent them as succinct relation tuples of the form *(head, relation, tail)* e.g., *(Barack Obama, spouse, Michelle Obama)*.

The concise representations provided by relation extraction models have been used to extend Knowledge Bases (KBs) (Riedel et al., 2013; Subasic et al., 2019; Trisedya et al., 2019). These KBs are then used heavily in NLP systems, such as question answering systems (Bordes et al., 2014; Yin et al., 2016; Cui et al., 2019). In recent years, much focus in the Neural Relation Extraction (NRE) community has been centered on improvements in model precision and the reduction of noise (Lin et al., 2016; Liu et al., 2017; Wu et al., 2017; Feng et al., 2018; Vashishth et al., 2018; Qin et al., 2018). Yet, little attention has been devoted towards the fairness of such systems.

We take the first step at understanding and evaluating gender bias in NRE systems by measuring the differences in model performance when extracting relations from sentences written about females versus sentences written about males. If a NRE model predicts a relation such *occupation* with higher recall on male entities, this could lead to the resulted knowledge bases having more occupation information for males than for females (see the illustration in Figure 1). Eventually, the gender bias in knowledge bases may affect downstream predictions, causing undesired allocative harms (Crawford, 2017) and reinforcing gender-stereotypical beliefs in society.

In this paper, we present an evaluation framework to analyze social bias in NRE models. Specifically, we evaluate gender bias in English language predictions of a collection of popularly used and open source NRE models[1] (Lin et al., 2016; Wu et al., 2017; Liu et al., 2017; Feng et al., 2018). We evaluate on two fronts: (1) examining gender bias

[*] Equal Contribution.

[1] https://github.com/thunlp/OpenNRE/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2943–2953
July 5 - 10, 2020.

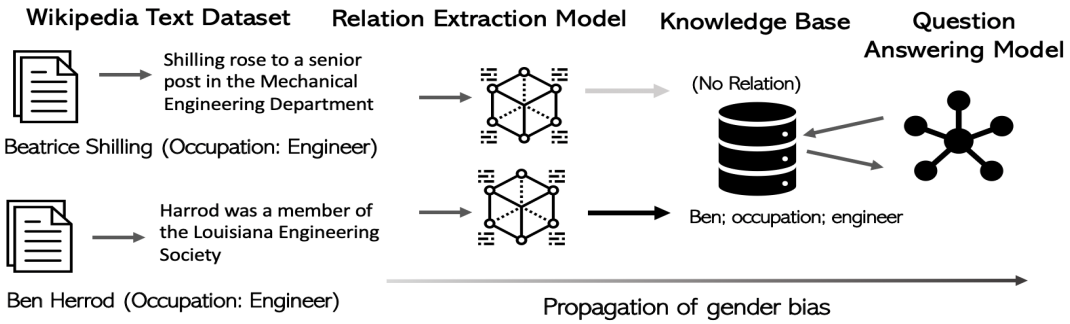

Figure 1: An illustration of gender bias in relation extraction and how it affects a downstream application. In their Wikipedia articles, both Beatrice (female) and Ben (male) are described as engineers. These sentences contain the *(entity; occupation; engineer)* relation. However, the model only predicts that the sentence from the male article expresses the occupation relation. If on a large scale, models extract the *(entity; occupation; engineer)* relation more often for males, knowledge bases will contain information for male engineers more often than female. Question answering models that query these knowledge bases may give biased answers and propagate gender bias downstream.

exhibited in a model that is trained on a relation extraction dataset; and (2) examining if the existing bias mitigation techniques (Bolukbasi et al., 2016; Zhao et al., 2018; Lu et al., 2018) can be applied to reduce the bias in an NRE system while maintaining its performance.

Carrying out such an evaluation is difficult with existing NRE datasets, such as the NYT dataset (Sandhaus, 2018), because there is no reliable way to obtain gender information about the entities mentioned in input sentences. Therefore, we create a new dataset, WikiGenderBias, specifically aimed at evaluating gender bias for NRE. WikiGenderBias is a distantly supervised dataset extracted using Wikipedia and DBPedia. It contains 45,000 sentences, each of which describe either a male or female entity with one of four relations: *spouse, hypernym (i.e., occupation), birthDate, and birthPlace*. We posit that a biased NRE system leverages gender information as a proxy when extracting knowledge tuples with spouse and hypernym relations. However, gender of the entity does not affect the extraction of relations such as birthDate and birthPlace, as they are not intuitively related to gender. Experiment results confirm our conjecture.

Our contributions are as such:

- We create WikiGenderBias, a new dataset for evaluating gender bias in NRE systems.
- We present an evaluation framework to demonstrate that gender bias is exhibited in NRE model outputs.
- We test several existing bias mitigation approaches to reducing gender bias in NRE system. Our analysis sheds light for designing future mitigating techniques.

2 Related Work

Gender Bias Measurement. Existing studies have revealed gender bias in various NLP tasks (Zhao et al., 2017; Rudinger et al., 2018; Zhao et al., 2018; Dixon et al., 2018; Lu et al., 2018; Kiritchenko and Mohammad, 2018; Romanov et al., 2019; Sheng et al., 2019; Sun et al., 2019). People have proposed different metrics to evaluate gender bias, for example, by using the performance difference of the model on male and female datapoints for bias evaluation (Lu et al., 2018; Kiritchenko and Mohammad, 2018). Other metrics have been proposed to evaluate fairness of predictors and allocative bias (Dwork et al., 2012; Hardt et al., 2016), such as Equality of Opportunity. In this work, we use both of these metrics to evaluate NRE models.

Mitigation Methods. After discovering gender bias existing, prior work has developed various methods to mitigate that bias (Escudé Font and Costa-jussà, 2019; Bordia and Bowman, 2019). Those mitigation methods can be applied in different levels of a model, including in the training phase, in the embedding layer, or in the inference procedure. In this paper, we test three existing debiasing approaches, namely data augmentation (Zhao et al., 2018; Lu et al., 2018), and word embedding debiasing technique (Hard Debiasing (Bolukbasi et al., 2016)) for mitigating bias in NRE models.

	Original Dataset				Equalized Dataset			
	Entity Pairs		Instances		Entity Pairs		Instances	
	M	F	M	F	M	F	M	F
Train	12,139	4,571	27,048	9,391	2,479	4,571	9,465	9,415
Development	1,587	553	3,416	1,144	336	553	1,144	1,144
Test	1,030	1,101	2,320	2,284	1,030	1,101	2,320	2,284
Total	14,756	6,225	32,784	12,819	3,845	6,225	12,929	12,843

Table 1: WikiGenderBias's Dataset splits. Entity Pairs means distinct pairs $(e1, e2)$ such that $(e_1, relation, e_2)$ is a relation in WikiGenderBias. Instances are the total number of $(e1, relation, e2, sentence)$ tuples in WikiGenderBias, where $sentence$ is distantly supervised. We categorize an entity pair as male (female) if $e1$ is male (female), since the sentence in the instance is taken from $e1$'s article and we define datapoints as male (female) if that is the gender of the subject of the article. The left two entries are for the dataset taken from the true distribution; the right two are the gender-equalized dataset created by down-sampling male instances.

Neural Relation Extraction. Relation extraction is a task in NLP with a long history that typically seeks to extract structured tuples (e_1, r, e_2) from texts (Bach and Badaskar, 2007). Early on, learning algorithms for relation extraction models were typically categorized as supervised, including feature-based methods (Kambhatla, 2004; Zhou et al., 2005; Zhao and Grishman, 2005) and kernel-based methods (Lodhi et al., 2002; Zelenko et al., 2003), or semi-supervised (Brin, 1998; Agichtein and Gravano, 2000; Etzioni et al., 2005; Pantel and Pennacchiotti, 2006), or purely unsupervised (Etzioni et al., 2008). Supervised approaches suffer from the need for large amounts of labelled data, which is sometimes not feasible, and generalizes poorly to open domain relation extraction, since labeled data is required for every entity-relation type (Bach and Badaskar, 2007; Mintz et al., 2009). Many semi-supervised approaches rely on pattern-matching, which is not robust, and many are unable to extract intra-sentence relations (Bach and Badaskar, 2007). When data annotation is insufficient or hard to obtain and semi-supervised approaches are insufficient, the distant supervision assumption is used to collect data to train supervised models (Mintz et al., 2009). Given a relation (e_1, r, e_2) in a knowledge base (KB), distant supervision assumes any sentence that contains both e_1 and e_2 expresses r (Mintz et al., 2009). Great efforts have been made to improve NRE models by mitigating the effects of noise in the training data introduced by Distant Supervision (Hoffmann et al., 2011; Surdeanu et al., 2012; Lin et al., 2016; Liu et al., 2017; Feng et al., 2018; Qin et al., 2018). However, to our knowledge, there are no studies on bias or ethics in NRE, which is filled by this work.

3 WikiGenderBias

We define gender bias in NRE as a difference in model performance when predicting on sentences from male versus female articles. Thus, we need articles written about entities for which we can identify the gender information. However, to obtain gender information for existing annotated datasets could be costly or impossible. Thus, we elected to create WikiGenderBias with this gender information to be able to detect scenarios like that in Figure 1. The data statistics of WikiGenderBias are given in Table 1.

3.1 Dataset Creation

Wikipedia is associated with a knowledge base, DBPedia, that contains relation information for entities with articles on Wikipedia (Mendes et al., 2012). Many of these entities have gender information and their corresponding articles are readily available. Therefore, we create our dataset based on sentences extracted from Wikipedia.

To generate WikiGenderBias, we use a variant of the distant supervision assumption: for a given relation between two entities, if one sentence from an article written about one entity also mentions the other entity, then we assume that such sentence expresses the relation. For instance, if we know *(Barack, spouse, Michelle)* is a relation tuple and we find the sentence *He and Michelle were married* in Barack's Wikipedia article, then we assume that sentence expresses the *(Barack, spouse, Michelle)* relation. This assumption is similar to that made by Mintz et al. (2009) and allows us to scalably create the dataset.

WikiGenderBias considers four relations that stored in DBPedia: *spouse*, *hypernym*, *birthDate*,

Relation	Head Entity	Tail Entity	Sentence
Birthdate	Robert M. Kimmitt	December 19, 1947	Robert M. Kimmitt (born December 19 , 1947) was United States Deputy Secretary of the Treasury under President George W. Bush .
Birthplace	Charles Edward Stuart	Rome	Charles was born in the Palazzo Muti , Rome , Italy , on 31 December 1720 , where his father had been given a residence by Pope Clement XI
Spouse	John W. Caldwell	Sallie J. Barclay	Caldwell married Sallie J. Barclay , and the couple had one son and two daughters .
hypernym	Handry Satriago	CEO	Handry Satriago (born in Riau , Pekanbaru on June 13 , 1969) is the CEO of General Electric Indonesia .

Table 2: Examples of relations of each type in WikiGenderBias.

and *birthPlace*. Note that the hypernym relation on DBPedia is similar to occupation, with entities having hypernym labels such as Politican. We also generate negative examples by obtaining datapoints for three unrelated relations: *parents*, *deathDate*, and *almaMater*. We label them as NA (not a relation). As each sentence only labelled with one relation based on our distant supervision assumption, WikiGenderBias is a 5-class classification relation extraction task. Figure 2 lists the label distribution.

We hypothesize that a biased relation extraction model might use gender as a proxy to influence predictions for spouse and hypernym relations, since words pertaining to marriage are more often mentioned in Wikipedia articles about female entities and words pertaining to hypernym (which is similar to occupation) are more often mentioned in Wikipedia articles about male entities (Wagner et al., 2015; Graells-Garrido et al., 2015). On the other hand, we posit that birthDate and birthPlace would operate like control groups and believe gender would correlate with neither relation.

To simplify the analysis, we only consider the head entities that associated with at least one of the four targeted relations. We set up our experiment such that head entities are not repeated across the train, dev, and test sets so that the model will see only new head entities at the test time. Since we obtain the distantly supervised sentences for a relation from the head entity's article, this guarantees the model will not reuse sentences from an article. However, it is possible that the head entity will appear as a tail entity in other relations because an entity could appear in multiple articles. The data splits are given in Table 1.

Besides, Wikipedia includes more articles written about males than about females. Therefore, there are more male instances than female instances in WikiGenderBias as well. To remove the effect of dataset bias in our analysis, we also create a gender-equalized version of the training and development sets by down-sampling male instances. We discuss the creation of gender-equalized test set below.

3.2 Test Sets

We equalize the male and female instances in the test set. In this way, a model cannot achieve high performance by performing well only on the dominant class. Furthermore, since some data instances that are collected using distant supervision are noisy, we annotated the correctness of the test instances using Amazon Mechanical Turk annotations to perform a fair comparison.

Specifically, we asked workers to determine whether or not a given sentence expressed a given relation. If the majority answer was "no", then we labeled that sentence as expressing "no relation" (we denote them as NA). Each sentence was annotated by three workers. Each worker was paid 15 cents per annotation. We only accepted workers from England, the US or Australia and with HIIT Approval Rate greater than 95% and Number of HIITs greater than 100. We found the pairwise inter-annotator agreement as measured by Fleiss' Kappa (Fleiss, 1971) κ is 0.44, which is consistent across both genders and signals moderate agreement. We note that our κ value is affected by asking workers to make binary classifications, which limits the degree of agreement that is attainable above chance. We also found the pairwise inter-annotator agreement to be 84%.

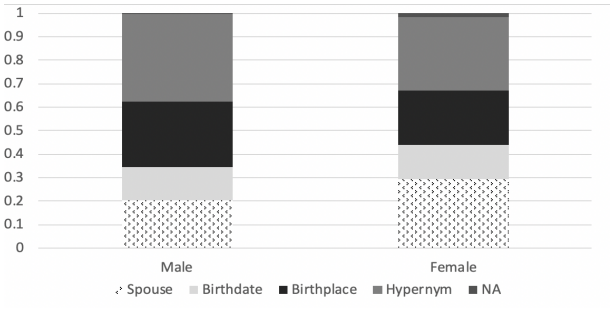

Figure 2: Proportion of sentences corresponding to a given relation over total sentences in WikiGenderBias for each entity. This demonstrates that, of the entities we sampled to create WikiGenderBias, the spouse relation is expressed more often relative to the birthdate, birthplace, and hypernym relations in articles about female entities than in articles about male entities. Additionally, hypernym is mentioned more often relative to the other relations in articles about male entities than in articles about female entities.

3.3 Data Analysis

We build on the work of Graells-Garrido et al. (2015), who discovered that female entities are more likely to have spouse information in the Infoboxes on their Wikipedia page than male entities. Figure 2 demonstrates a further discrepancy: amongst articles we sampled, proportionally, the spouse relation is mentioned more often relative to hypernym, birthPlace, and birthDate in female articles than in male articles. Additionally, we show that amongst female and male articles we sampled, hypernyms are mentioned more often in male than female articles relative to spouse, birthPlace, and birthDate (see Section 2). This observation aligns with the literature, arguing that authors do not write about the two genders equally (Wagner et al., 2015; Graells-Garrido et al., 2015).

4 Gender Bias in NRE

We evaluate OpenNRE (Han et al., 2019), a popular open-source NRE system. OpenNRE implements the approach from (Lin et al., 2016). To convert sentences into vectors, researchers propose convolutional neural networks as well as the pieceweise convoultional neural networks (PCNN) which retain more structural information between entities (Zeng et al., 2015). In this work, we use a PCNN with Selective Attention for the experiments.

We train every encoder-selector combination on the training set of WikiGenderBias and its gender-equalized version. We input Word2Vec (Mikolov et al., 2013) word embeddings trained on WikiGenderBias to the models[2]. We use commit 709b2f from the OpenNRE repository tensorflow branch to obtain the models.

4.1 Performance Parity Score

The goal of a successful relation extraction model is to maximize F1 score while minimizing the model performance gender gap (or disparity score). However, when comparing different systems, it is hard to decide what is the right balance between these two objectives. On one end, a model which has zero gender gap but has only 10% accuracy for both male and female test instances has almost no practical value. Other methods that have high accuracy or F1 score may do so at the cost of a wide gender gap. Although our test set for WikiGenderBias is gender-equalized, one can imagine that improving performance on a test set that is heavily skewed towards males can be done by focusing on male test instances while largely ignoring female ones. Therefore, it is important to strike a balance between model performance and inter-group parity.

To measure model performance, we use Macro-average F1 score. To measure inter-group parity, we use the pairwise difference in F1 scores averaged over all the groups for predictions on a given relation i. We describe the average difference over all relations as Disparity Score (DS):

$$DS = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x}\sum_{j=1}^{x}\sum_{k=j+1}^{x}\left|F_{1_{ik}} - F_{1_{ij}}\right|,$$

where n denotes the number of relations (e.g. {birthDate, birthPlace, spouse, hypernym}). x denotes the number of groups (e.g. {male, female}). $F_{1_{rk}}$ is the F_1 score for the model when predicting datapoints with true label relation r that belong to group k. (So, for instance, $F_{1_{spouse,male}}$ is the F_1 score on sentences that express the spouse relation from male articles.) The Disparity Score measures the F1 score gap between predictions on male and female data points.

Bringing these two metrics together, we propose the Performance Parity Score (PPS). PPS is the Macro-average difference (equally weighted) of the

[2]We performed Grid Search to determine the optimal hyperparameters. We set epochs= 60, learning rate $\eta = 0.5$, early stopping with patience of 10, *batch size*= 160, and *sliding window size*= 3 (for CNN and PCNN). These hyperparameters are similar to the default settings found in the OpenNRE repository tensorflow branch, which uses epochs= 60, learning rate $\eta = 0.5$, and early stopping with patience of 20.

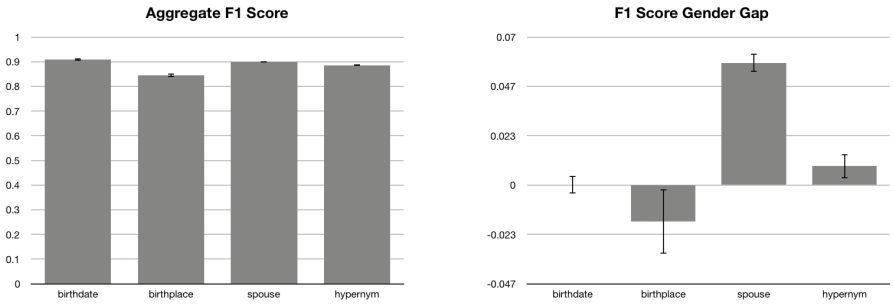

Figure 3: Aggregate performance of the NRE model for each relation (left) and $male - female$ F1 score gender gap for each relation (right). An ideal model maximizes performance and minimizes the gender gap. The experiment is run five times. We give the mean values and standard error bars.

F1 score subtracted by the model performance gender gap, which we defined as the Disparity Score, per relation. We place equal importance on the F1 score and Disparity Score by giving each score an implicit weight of 1. In our formula for PPS above, we also divide the final result by the number of relations n. This keeps the range of PPS within $(-1, 1]$, although PPS will generally fall between $[0, 1]$ because it is highly unlikely that the Disparity Score will be greater than the overall F1 score. PPS seeks to incentivize a combination of both model performance and inter-group parity for the task of relation extraction:

$$
\begin{aligned}
PPS &= \frac{1}{n}\sum_{i=1}^{n}\left(F_{1_i} - \frac{1}{x}\sum_{j=1}^{x}\sum_{k=j+1}^{x}\left|F_{1_{ik}} - F_{1_{ij}}\right|\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}F_{1_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x}\sum_{j=1}^{x}\sum_{k=j+1}^{x}\left|F_{1_{ik}} - F_{1_{ij}}\right| \\
&= \text{Macro } F_1 \text{ score} - \text{Disparity Score.}
\end{aligned}
$$

4.2 Measuring Performance Differences

Similar to the parity term in PPS, gender bias can be measured as the difference in a performance metric for a model when evaluated on male and female datapoints (De-Arteaga et al., 2019). We define male (female) datapoints to be relations for which the head entity is male (female), which means the distantly supervised sentence is taken from a male (female) article. Prior work has used area under the precision-recall curve and F1 score to measure NRE model performance (Gupta et al., 2019; Han et al., 2019; Kuang et al., 2019). We use Macro-F1 score as our performance metric. We denote the F1 gender difference as $F1_{Gap}$, which is used to calculate the disparity score. A larger disparity score indicates higher bias in predictions.

4.3 Equality of Opportunity Evaluation

Equality of Opportunity (EoO) was originally proposed to measure and address allocative biases (Hardt et al., 2016). Consequently, we examine this metric in the context of relation extraction to better understand how allocative biases can begin to emerge at this stage.

Equality of Opportunity (EoO) is defined in terms of the joint distribution of (X, A, Y), where X is the input, A is a protected attribute that should not influence the prediction, and Y is the true label (Hardt et al., 2016). A predictor satisfies Equality of Opportunity if and only if: $P(\hat{Y} = 1|A = male, Y = 1)$ and $P(\hat{Y} = 1|A = female, Y = 1)$. In our case $A = \{male, female\}$, because gender is our protected attribute and we assume it to be binary. We evaluate EoO on a per-relation, one-versus-rest basis. Thus, when calculating EoO for spouse, $Y = 1$ indicates the true label is spouse and $\hat{Y} = 1$ indicates a prediction of spouse. We do this for each relation. Note that this is equivalent to measuring per-relation recall for each gender.

4.4 Result

As shown in Figure 3, the NRE system performs better when predicting the spouse relation on sentences from articles about male entities than from articles on female entities (see Figure 3, right). Further, there is a large recall gap (see EoO column, row 1, in Table 3). Notably, the gender difference in performance is much smaller on birthDate, birthPlace, and hypernym relations, although the gender difference is non-zero for birthPlace and hyerpym. This is interesting given that a higher percentage of female instances in WikiGenderBias are spouse relations than male (see Figure 2). We encourage future work to explore whether the writing style differences between male and female spouse in-

	Spouse		Birth Date		Birth Place		Hypernym		Total		
	$F1_{Gap}$	EoO	$F1_{Gap}$	EoO	$F1_{Gap}$	EoO	$F1_{Gap}$	EoO	F1 Score	Disparity Score	PPS
PCNN,ATT	.041	.058	.004	.000	-.003	-.017	.015	.009	.886	.016	.870
CNN,ATT	.034	.043	-.003	.001	.014	.004	.028	.014	.882	.020	.862
RNN,ATT	.032	.043	.015	.019	.005	-.011	-.006	-.006	.889	.014	.875
BIRNN,ATT	.039	.061	.013	.021	-.016	-.033	-.013	-.026	.884	.020	.864
PCNN,AVE	.034	.044	.005	.010	-.001	-.011	.005	-.005	.903	.011	.892
CNN,AVE	.027	.028	.013	.029	.007	.009	.002	-.028	.895	.012	.883
RNN,AVE	.039	.036	.004	.021	.016	.020	.006	-.012	.912	.016	.895
BIRNN,AVE	.024	.018	.001	.015	.009	.018	-.005	-.022	.913	.010	.903

Table 3: Results from running combinations of encoders and selectors of the OpenNRE model for the male and female genders of each relation. A positive $F1_{Gap}$ indicates a higher F1 on male instances. A higher Equality of Opportunity (EoO) indicates higher recall on male instances. A higher PPS score indicates a better balance of performance and parity (see Section 4.1). We ran the experiment five times and report the mean values. Varying the encoder and selector appears to have no conclusive effect on bias, although models using the average selector doe achieve better aggregate performance. These results were obtained using the gender unequalized training data.

stances causes those male instances to be easier to classify.

In addition, we explore different types of sentence encoder and sentence-level attention used in the creation of the bag representation for each entity pair and examined how these models performed on our dataset. Notably, the bias in spouse relation persists across OpenNRE architectures (see Table 3). It seems models using average attention, which merely averages all the sentence vectors in the bag to create a representation of the entire bag, allows for better aggregate performance on WikiGenderBias. However, the effect on the Disparity Score (and therefore the bias exhibited in the predictions) seems negligible.

We note that these results do not necessarily indicate that the model itself contains biases given that males and females are written about differently on Wikipedia. These results do, however, demonstrate that we must be cautious when deploying NRE models, especially those trained on Wikipedia data, since they can propagate biases latent in their training data to the knowledge bases they help create.

5 Bias Mitigation

We examine data augmentation and Hard-Debiasing as bias mitigation techniques for reducing gender bias in NRE system.

5.1 Bias Mitigation Techniques

Equalizing the Gender Distribution Sometimes, the true distribution contains an imbalance in gendered data. For instance, perhaps the training set contains more instances from male articles than female. To mitigate this, one can simply downsample the male instances until the male and female instances are approximately equal, then train on this modified, equalized distribution.

Data Augmentation. The contexts in which males and females are written about can differ; for instance, on Wikipedia women are more often written about with words related to sexuality than men (Graells-Garrido et al., 2015). Data augmentation mitigates these contextual biases by replacing masculine words in a sentence with their corresponding feminine words and vice versa for all sentences in a corpus, and then training on the union of the original and augmented corpora[3] (Zhao et al., 2018; Lu et al., 2018; Dixon et al., 2018; Maudslay et al., 2019; Zhao et al., 2019).

Word Embedding Debiasing Word embeddings can encode gender biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018) and this can affect bias in downstream predictions for models using the embeddings (Zhao et al., 2018; Font and Costa-Jussa, 2019). In this work, we apply the Hard-Debiasing technique (Bolukbasi et al., 2016). We applied Hard-Debiasing to Word2Vec embeddings (Mikolov et al., 2013), which we trained on the sentences in WikiGenderBias. When used in conjunction with data augmentation, the embeddings are re-trained on the union of the two corpora. Below, we give metrics used for measuring model performance and bias in our experiments.

[3]We use the following list to perform data augmentation: https://github.com/uclanlp/corefBias/blob/master/WinoBias/wino/generalized_swaps.txt

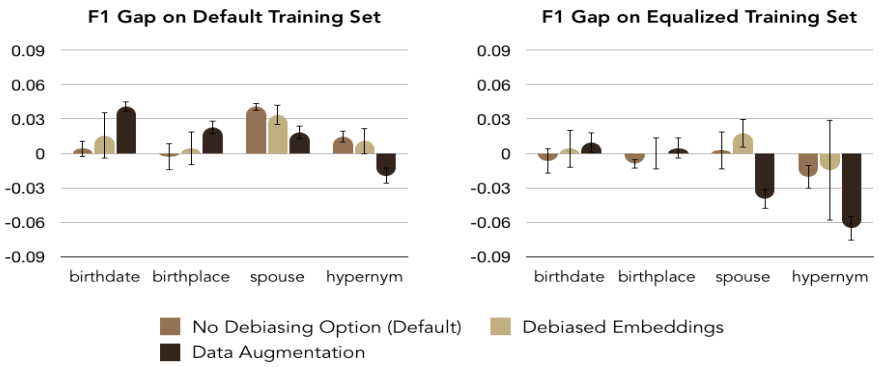

Figure 4: Bias in relation extraction model on each relation as measured by $male - female$ F1 score gender gap (used to calculate disparity score) for the default training set without modifications (left) and equalized training set (right). This is evaluated on the model with No Debiasing and two bias mitigation methods: debiased embeddings and data augmentation. The experiment is run five times. We give the mean values and standard error bars.

#	Equalization	Debiased Embeddings	Data Aug.	EoO ↓	PPS Score ↑	Macro F_1 Score ↑	Disparity Score ↓
1				.012	.870	.886	.016
2	✓			-.011	.851	.860	0.010
3		✓		.015	**.886**	**.902**	.016
4			✓	.014	.841	.866	.026
5	✓	✓		**.001**	.863	.872	**.009**
6	✓		✓	-.024	.805	.835	.030
7		✓	✓	.018	.868	.891	.023
8	✓	✓	✓	.006	.867	.877	.010

Table 4: PPS Scores when using debiased embeddings and data augmentation with the unequalized, original dataset. We find that using debiased embeddings alone leads to the best PPS score. Other combinations of debiasing parameters lowers either F1 score, disparity score, or both. We bold the best values, which represent the maximum for PPS score and F1 score and minimum for Disparity Score.

5.2 Effectiveness of Bias Mitigation

We note that by downsampling the training instances to equalize the number of male and female datapoints, the difference in performance on male versus female sentences decreases to almost 0 for every relation aside from hypernym (see Figure 4, right). Additionally, the drop in aggregate is performance is relatively small (see Macro F1, Table 4). Given that we down-sampled male instances to create this equalized dataset, training on the equalized data was also more efficient.

We also examined the effect of various debiasing techniques. Table 4 shows the results. Unfortunately, most of these techniques cause a significant performance drop and none of them is effective in reducing the performance gap between genders. Interestingly, debiasing embeddings increased aggregate performance by achieving slightly better F1 performance. As none of these mitigation approaches is effective, their combinations are not effective as well. They either lowering Macro F1 or raising Disparity Score or both.

We further examine the performance of various bias mitigation techniques evaluated in each relation in Figure 4. NRE relies heavily on surface-level cues such as context, the entities, and their positions. Data augmentation might potentially introduce artifacts and biases, causing the NRE system captures unwanted patterns and spurious statistics between contexts.

6 Conclusion

In our study, we create and publicly release WikiGenderBias: the first dataset aimed at evaluating bias in NRE models. We train NRE models on the WikiGenderBias dataset and test them on gender-separated test sets. We find a difference in F1 scores for the spouse relation between predictions on male sentences and female for the model's predictions. We also examine existing bias mitigation techniques and find that naive data augmentation causes a significant performance drop.

It is an open and difficult research question to build unbiased neural relation extractors. One possibility is that some bias mitigation methods that add noise to the dataset encourage neural relation extraction models to learn spurious correlations and unwanted biases. We encourage future work to dive deeper into this problem.

While these findings will help future work avoid gender biases, this study is preliminary. We only consider binary gender, but future work should consider non-binary genders. Additionally, future work should further probe the source of gender bias in the model's predictions, perhaps by visualizing attention or looking more closely at the model's outputs.

Acknowledgments

We thank anonymous reviewers for their helpful feedback. This material is based upon work supported in part by the National Science Foundation under IIS Grant 1927554 and Grant 1821415: Scaling the Early Research Scholars Program.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries (ACM '00)*, pages 85–94.

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15.

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 *Neural Information Processing Systems (NIPS'16)*.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620, Doha, Qatar. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Brin. 1998. Extracting Patterns and Relations from the World Wide Web. In *International Workshop on The World Wide Web and Databases at EDBT '98*, pages 172–183. Springer.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356(6334):183–186.

Kate Crawford. 2017. The Trouble With Bias. Keynote at Neural Information Processing Systems (NIPS'17).

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2019. Kbqa: Learning question answering over qa corpora and knowledge bases.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128. ACM.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AAAI'18)*, pages 67–73. ACM.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open Information Extraction from the Web. 51(12):68–74.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91–134.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement Learning for Relation Classification from Noisy Data. In *Thirty-Second Conference on Advancement of Artificial Intelligence (AAAI '18)*.

Joseph L Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378.

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

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First Women, Second sex: Gender Bias in Wikipedia. In *Proceedings of the 26th ACM Conference (ACM '15)*, pages 165–174. ACM.

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas Runkler. 2019. Neural Relation Extraction Within and Across Sentence Boundaries. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '19)*, volume 33, pages 6513–6520.

Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. *arXiv preprint arXiv:1909.13078*.

Moritz Hardt, Eric Price, and Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NIPS '16)*, pages 3315–3323.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pages 541–550. Association for Computational Linguistics.

Nanda Kambhatla. 2004. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, pages 22–es.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv preprint arXiv:1805.04508*.

Jun Kuang, Yixin Cao, Jianbing Zheng, Xiangnan He, Ming Gao, and Aoying Zhou. 2019. Improving Neural Relation Extraction with Implicit Mutual Relations. *arXiv preprint arXiv:1907.05333*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural Relation Extraction with Selective Attention Over Instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*, volume 1, pages 2124–2133.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A Soft-Label Method for Noise-Tolerant Distantly Supervised Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP '17)*, pages 1790–1795.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text Classification using String Kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender Bias in Neural Natural Language Processing.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. *arXiv preprint arXiv:1909.00871*.

Pablo N. Mendes, Max Jakob, and Christian Bizer. 2012. Dbpedia for nlp: A multilingual cross-domain knowledge base. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems (NIPS '13)*, pages 3111–3119.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association of Computational Linguistics (ACL'09)*, pages 1003–1011. Association for Computational Linguistics.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *booktitle=North American Chapter of the Association for Computational Linguistics (NAACL'13)*, pages 74–84.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra

Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *arXiv preprint arXiv:1904.05233*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *North American Chapter of the Association for Computational Linguistics (NAACL'18)*.

Evan Sandhaus. 2018. The New York Times Annotated Corpus. Linguistic Data Consortium (LDC '08).

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Pero Subasic, Hongfeng Yin, and Xiao Lin. 2019. Building Knowledge Base through Deep Learning Relation Extraction and Wikidata. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP '12)*, pages 455–465. Association for Computational Linguistics.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural Relation Extraction for Knowledge Base Enrichment. In *Assocation for Computational Linguistics (ACL '19)*. ACL.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving Distantly-Supervised Neural Relation Extraction Using Side Information.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Ninth International AAAI Conference on Web and Social Media (AAAI'15)*.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial Training for Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP '17)*, pages 1778–1783.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. In *Proceedings of the Workshop on Human-Computer Question Answering*, pages 36–42, San Diego, California. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, 3(Feb):1083–1106.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant Supervision for Relation Extraction Via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP '15)*, pages 1753–1762.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. *arXiv preprint arXiv:1904.03310*.

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 *Empirical Methods of Natural Language Processing (EMNLP'17)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *North American Chapter of the Association for Computational Linguistics (NAACL'18)*.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 419–426, Ann Arbor, Michigan. Association for Computational Linguistics.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan. Association for Computational Linguistics.