Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

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

Abstract

We introduce a new benchmark, WinoBias, for coreference resolution focused on gender bias. Our corpus contains Winograd-schema style sentences with entities corresponding to people referred by their occupation (e.g. the nurse, the doctor, the carpenter). We demonstrate that a rule-based, a feature-rich, and a neural coreference system all link gendered pronouns to pro-stereotypical entities with higher accuracy than anti-stereotypical entities, by an average difference of 21.1 in F1 score. Finally, we demonstrate a data-augmentation approach that, in combination with existing word-embedding debiasing techniques, removes the bias demonstrated by these systems in WinoBias without significantly affecting their performance on existing coreference benchmark datasets. Our dataset and code are available at `http://winobias.org`.

1 Introduction

Coreference resolution is a task aimed at identifying phrases (mentions) referring to the same entity. Various approaches, including rule-based (Raghunathan et al., 2010), feature-based (Durrett and Klein, 2013; Peng et al., 2015a), and neural-network based (Clark and Manning, 2016; Lee et al., 2017) have been proposed. While significant advances have been made, systems carry the risk of relying on societal stereotypes present in training data that could significantly impact their performance for some demographic groups.

In this work, we test the hypothesis that coreference systems exhibit gender bias by creating a new challenge corpus, WinoBias.This dataset follows the winograd format (Hirst, 1981; Rahman and Ng, 2012; Peng et al., 2015b), and contains references to people using a vocabulary of 40 occupations. It contains two types of challenge sentences that require linking gendered pronouns to either male or female stereotypical occupations (see the illustrative examples in Figure 1). None of the examples can be disambiguated by the gender of the pronoun but this cue can potentially distract the model. We consider a system to be gender biased if it links pronouns to occupations dominated by the gender of the pronoun (pro-stereotyped condition) more accurately than occupations not dominated by the gender of the pronoun (anti-stereotyped condition). The corpus can be used to certify a system has gender bias.[1]

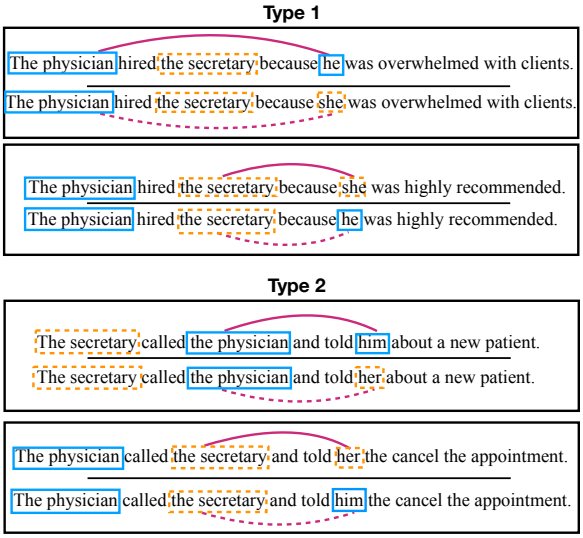

Figure 1: Pairs of gender balanced co-reference tests in the WinoBias dataset. Male and female entities are marked in solid blue and dashed orange, respectively. For each example, the gender of the pronominal reference is irrelevant for the co-reference decision. Systems must be able to make correct linking predictions in pro-stereotypical scenarios (solid purple lines) and anti-stereotypical scenarios (dashed purple lines) equally well to pass the test. Importantly, stereotypical occupations are considered based on US Department of Labor statistics.

We use three different systems as prototypical examples: the Stanford Deterministic Coreference System (Raghunathan et al., 2010), the

[1]Note that the counter argument (i.e., systems are gender bias free) may not hold.

Berkeley Coreference Resolution System (Durrett and Klein, 2013) and the current best published system: the UW End-to-end Neural Coreference Resolution System (Lee et al., 2017). Despite qualitatively different approaches, all systems exhibit gender bias, showing an average difference in performance between pro-stereotypical and anti-stereotyped conditions of 21.1 in F1 score. Finally we show that given sufficiently strong alternative cues, systems can ignore their bias.

In order to study the source of this bias, we analyze the training corpus used by these systems, Ontonotes 5.0 (Weischedel et al., 2012).[2] Our analysis shows that female entities are significantly underrepresented in this corpus. To reduce the impact of such dataset bias, we propose to generate an auxiliary dataset where all male entities are replaced by female entities, and vice versa, using a rule-based approach. Methods can then be trained on the union of the original and auxiliary dataset. In combination with methods that remove bias from fixed resources such as word embeddings (Bolukbasi et al., 2016), our data augmentation approach completely eliminates bias when evaluating on WinoBias , without significantly affecting overall coreference accuracy.

2 WinoBias

To better identify gender bias in coreference resolution systems, we build a new dataset centered on people entities referred by their occupations from a vocabulary of 40 occupations gathered from the US Department of Labor, shown in Table 1.[3] We use the associated occupation statistics to determine what constitutes gender stereotypical roles (e.g. 90% of nurses are women in this survey). Entities referred by different occupations are paired and used to construct test case scenarios. Sentences are duplicated using male and female pronouns, and contain equal numbers of correct coreference decisions for all occupations. In total, the dataset contains 3,160 sentences, split equally for development and test, created by researchers familiar with the project. Sentences were created to follow two prototypical templates but annotators were encouraged to come up with scenarios where entities could be interacting in plausible ways. Templates were selected to be challenging and designed to cover cases requiring semantics and syntax separately.[4]

Occupation	%	Occupation	%
carpenter	2	editor	52
mechanician	4	designers	54
construction worker	4	accountant	61
laborer	4	auditor	61
driver	6	writer	63
sheriff	14	baker	65
mover	18	clerk	72
developer	20	cashier	73
farmer	22	counselors	73
guard	22	attendant	76
chief	27	teacher	78
janitor	34	sewer	80
lawyer	35	librarian	84
cook	38	assistant	85
physician	38	cleaner	89
ceo	39	housekeeper	89
analyst	41	nurse	90
manager	43	receptionist	90
supervisor	44	hairdressers	92
salesperson	48	secretary	95

Table 1: Occupations statistics used in WinoBias dataset, organized by the percent of people in the occupation who are reported as female. When woman dominate profession, we call linking the noun phrase referring to the job with female and male pronoun as 'pro-stereotypical', and 'anti-stereotypical', respectively. Similarly, if the occupation is male dominated, linking the noun phrase with the male and female pronoun is called, 'pro-stereotypical' and 'anti-steretypical', respectively.

Type 1: [`entity1`] [interacts with] [`entity2`] [conjunction] [pronoun] [circumstances]. Prototypical WinoCoRef style sentences, where co-reference decisions must be made using world knowledge about given circumstances (Figure 1; Type 1). Such examples are challenging because they contain no syntactic cues.

Type 2: [`entity1`] [interacts with] [`entity2`] and then [interacts with] [pronoun] for [circumstances]. These tests can be resolved using syntactic information and understanding of the pronoun (Figure 1; Type 2). We expect systems to do well on such cases because both semantic and syntactic cues help disambiguation.

Evaluation To evaluate models, we split the data in two sections: one where correct coreference decisions require linking a gendered pronoun to an occupation stereotypically associated with the gender of the pronoun and one that requires linking to the anti-stereotypical occupation. We say that a model passes the WinoBias

[2]The corpus is used in CoNLL-2011 and CoNLL-2012 shared tasks, http://www.conll.org/previous-tasks

[3]Labor Force Statistics from the Current Population Survey, 2017. https://www.bls.gov/cps/cpsaat11.htm

[4]We do not claim this set of templates is complete, but that they provide representative examples that, pratically, show bias in existing systems.

test if for both Type 1 and Type 2 examples, pro-stereotyped and anti-stereotyped co-reference decisions are made with the same accuracy.

3 Gender Bias in Co-reference

In this section, we highlight two sources of gender bias in co-reference systems that can cause them to fail WinoBias: training data and auxiliary resources and propose strategies to mitigate them.

3.1 Training Data Bias

Bias in OntoNotes 5.0 Resources supporting the training of co-reference systems have severe gender imbalance. In general, entities that have a mention headed by gendered pronouns (e.g."he", "she") are over 80% male.[5] Furthermore, the way in which such entities are referred to, varies significantly. Male gendered mentions are more than twice as likely to contain a job title as female mentions.[6] Moreover, these trends hold across genres.

Gender Swapping To remove such bias, we construct an additional training corpus where all male entities are swapped for female entities and vice-versa. Methods can then be trained on both original and swapped corpora. This approach maintains non-gender-revealing correlations while eliminating correlations between gender and co-reference cues.

We adopt a simple rule based approach for gender swapping. First, we anonymize named entities using an automatic named entity finder (Lample et al., 2016). Named entities are replaced consistently within document (i.e. "Barak Obama ... Obama was re-elected." would be annoymized to "E1 E2 ... E2 was re-elected."). Then we build a dictionary of gendered terms and their realization as the opposite gender by asking workers on Amazon Mechnical Turk to annotate all unique spans in the OntoNotes development set.[7] Rules were then mined by computing the word difference between initial and edited spans. Common rules included "she $\rightarrow$ he", "Mr." $\rightarrow$ "Mrs.", "mother" $\rightarrow$ "father." Sometimes the same initial word was edited to multiple different phrases: these were resolved by taking the most frequent phrase, with the exception of "her $\rightarrow$ him" and "her $\rightarrow$ his" which were resolved using part-of-speech. Rules were applied to all matching tokens in the OntoNotes. We maintain anonymization so that cases like "John went to his house" can be accurately swapped to "E1 went to her house."

3.2 Resource Bias

Word Embeddings Word embeddings are widely used in NLP applications however recent work has shown that they are severely biased: "man" tends to be closer to "programmer" than "woman" (Bolukbasi et al., 2016; Caliskan et al., 2017). Current state-of-art co-reference systems build on word embeddings and risk inheriting their bias. To reduce bias from this resource, we replace GloVe embeddings with debiased vectors (Bolukbasi et al., 2016).

Gender Lists While current neural approaches rely heavily on pre-trained word embeddings, previous feature rich and rule-based approaches rely on corpus based gender statistics mined from external resources (Bergsma and Lin, 2006). Such lists were generated from large unlabeled corpora using heuristic data mining methods. These resources provide counts for how often a noun phrase is observed in a male, female, neutral, and plural context. To reduce this bias, we balance male and female counts for all noun phrases.

4 Results

In this section we evaluate of three representative systems: rule based, Rule, (Raghunathan et al., 2010), feature-rich, Feature, (Durrett and Klein, 2013), and end-to-end neural (the current state-of-the-art), E2E, (Lee et al., 2017). The following sections show that performance on WinoBias reveals gender bias in all systems, that our methods remove such bias, and that systems are less biased on OntoNotes data.

WinoBias Reveals Gender Bias Table 2 summarizes development set evaluations using all three systems. Systems were evaluated on both types of sentences in WinoBias (T1 and T2), separately in pro-stereotyped and anti-stereotyped conditions (T1-p vs. T1-a, T2-p vs T2-a). We evaluate the effect of named-entity anonymization (Anon.), debiasing supporting resources[8] (Re-

[5] To exclude mentions such as "his mother", we use Collins head finder (Collins, 2003) to identify the head word of each mention, and only consider the mentions whose head word is gender pronoun.

[6] We pick more than 900 job titles from a gazetteer.

[7] Five turkers were presented with anonymized spans and asked to mark if it indicated male, female, or neither, and if male or female, rewrite it so it refers to the other gender.

[8] Word embeddings for E2E and gender lists for Feature

Method	Anon.	Resour.	Aug.	OntoNotes	T1-p	T1-a	Avg	| Diff |	T2-p	T2-a	Avg	| Diff |
E2E				**67.7**	**76.0**	49.4	62.7	26.6*	**88.7**	75.2	82.0	13.5*
E2E	✓			66.4	73.5	51.2	62.6	21.3*	86.3	70.3	78.3	16.1*
E2E	✓	✓		66.5	67.2	59.3	63.2	7.9*	81.4	82.3	81.9	0.9
E2E	✓		✓	66.2	65.1	59.2	62.2	5.9*	86.5	**83.7**	**85.1**	2.8*
E2E	✓	✓	✓	66.3	63.9	**62.8**	**63.4**	**1.1**	81.3	83.4	82.4	**2.1**
Feature				**61.7**	**66.7**	56.0	61.4	10.6*	**73.0**	57.4	65.2	15.7*
Feature	✓			61.3	65.9	56.8	61.3	9.1*	72.0	58.5	65.3	13.5*
Feature	✓	✓		61.2	61.8	**62.0**	**61.9**	**0.2**	67.1	63.5	65.3	3.6
Feature	✓		✓	61.0	65.0	57.3	61.2	7.7*	72.8	63.2	68.0	9.6*
Feature	✓	✓	✓	61.0	62.3	60.4	61.4	1.9*	71.1	**68.6**	**69.9**	**2.5**
Rule				57.0	76.7	37.5	57.1	39.2*	50.5	29.2	39.9	21.3*

Table 2: F1 on OntoNotes and WinoBias development set. WinoBias results are split between Type-1 and Type-2 and in pro/anti-stereotypical conditions. * indicates the difference between pro/anti stereotypical conditions is significant ($p < .05$) under an approximate randomized test (Graham et al., 2014). Our methods eliminate the difference between pro-stereotypical and anti-stereotypical conditions (Diff), with little loss in performance (OntoNotes and Avg).

Method	Anon.	Resour.	Aug.	OntoNotes	T1-p	T1-a	Avg	| Diff |	T2-p	T2-a	Avg	| Diff |
E2E				**67.2**	**74.9**	47.7	61.3	27.2*	**88.6**	77.3	**82.9**	11.3*
E2E	✓	✓	✓	66.5	62.4	**60.3**	**61.3**	**2.1**	78.4	**78.0**	78.2	**0.4**
Feature				**64.0**	**62.9**	58.3	60.6	4.6*	**68.5**	57.8	63.1	10.7*
Feature	✓	✓	✓	63.6	62.2	**60.6**	**61.4**	**1.7**	**70.0**	**69.5**	**69.7**	**0.6**
Rule				58.7	72.0	37.5	54.8	34.5*	47.8	26.6	37.2	21.2*

Table 3: F1 on OntoNotes and Winobias test sets. Methods were run once, supporting development set conclusions.

Model	Original	Gender-reversed
E2E	66.4	65.9
Feature	61.3	60.3

Table 4: Performance on the original and the gender-reversed developments dataset (anonymized).

sour.) and using data-augmentation through gender swapping (Aug.). E2E and Feature were retrained in each condition using default hyperparameters while Rule was not debiased because it is untrainable. We evaluate using the coreference scorer v8.01 (Pradhan et al., 2014) and compute the average (Avg) and absolute difference (Diff) between pro-stereotyped and anti-stereotyped conditions in WinoBias.

All initial systems demonstrate severe disparity between pro-stereotyped and anti-stereotyped conditions. Overall, the rule based system is most biased, followed by the neural approach and feature rich approach. Across all conditions, anonymization impacts E2E the most, while all other debiasing methods result in insignificant loss in performance on the OntoNotes dataset. Removing biased resources and data-augmentation reduce bias independently and more so in combination, allowing both E2E and Feature to pass WinoBias without significantly impacting performance on either OntoNotes or WinoBias . Qualitatively, the neural system is easiest to de-bias and our approaches could be applied to future end-to-end systems. Systems were evaluated once on test sets, Table 3, supporting our conclusions.

Systems Demonstrate Less Bias on OntoNotes While we have demonstrated co-reference systems have severe bias as measured in WinoBias , this is an out-of-domain test for systems trained on OntoNotes. Evaluating directly within OntoNotes is challenging because sub-sampling documents with more female entities would leave very few evaluation data points. Instead, we apply our gender swapping system (Section 3), to the OntoNotes development set and compare system performance between swapped and unswapped data.[9] If a system shows significant difference between original and gender-reversed conditions, then we would consider it gender biased on OntoNotes data.

Table 4 summarizes our results. The E2E system does not demonstrate significant degradation in performance, while Feature loses roughly 1.0-F1.[10] This demonstrates that given sufficient alternative signal, systems often do ignore gender biased cues. On the other hand, WinoBias provides an analysis of system bias in an adversarial setup, showing, when examples are challenging, systems are likely to make gender biased predictions.

[9]This test provides a lower bound on OntoNotes bias because some mistakes can result from errors introduce by the gender swapping system.

[10]We do not evaluate the Rule system as it cannot be train for anonymized input.

5 Related Work

Machine learning methods are designed to generalize from observation but if algorithms inadvertently learn to make predictions based on stereotyped associations they risk amplifying existing social problems. Several problematic instances have been demonstrated, for example, word embeddings can encode sexist stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017). Similar observations have been made in vision and language models (Zhao et al., 2017), online news (Ross and Carter, 2011), web search (Kay et al., 2015) and advertisements (Sweeney, 2013). In our work, we add a unique focus on co-reference, and propose simple general purpose methods for reducing bias.

Implicit human bias can come from imbalanced datasets. When making decisions on such datasets, it is usual that under-represented samples in the data are neglected since they do not influence the overall accuracy as much. For binary classification Kamishima et al. (2012, 2011) add a regularization term to their objective that penalizes biased predictions. Various other approaches have been proposed to produce "fair" classifiers (Calders et al., 2009; Feldman et al., 2015; Misra et al., 2016). For structured prediction, the work of Zhao et al. (2017) reduces bias by using corpus level constraints, but is only practical for models with specialized structure. Kusner et al. (2017) propose the method based on causal inference to achieve the model fairness where they do the data augmentation under specific cases, however, to the best of our knowledge, we are the first to propose data augmentation based on gender swapping in order to reduce gender bias.

Concurrent work (Rudinger et al., 2018) also studied gender bias in coreference resolution systems, and created a similar job title based, winograd-style, co-reference dataset to demonstrate bias [11]. Their work corroborates our findings of bias and expands the set of systems shown to be biased while we add a focus on debiasing methods. Future work can evaluate on both datasets.

6 Conclusion

Bias in NLP systems has the potential to not only mimic but also amplify stereotypes in society. For a prototypical problem, coreference, we provide a method for detecting such bias and show that three systems are significantly gender biased. We also provide evidence that systems, given sufficient cues, can ignore their bias. Finally, we present general purpose methods for making coreference models more robust to spurious, gender-biased cues while not incurring significant penalties on their performance on benchmark datasets.

[11] Their dataset also includes gender neutral pronouns and examples containing one job title instead of two.

Acknowledgement

This work was supported in part by National Science Foundation Grant IIS-1760523, two NVIDIA GPU Grants, and a Google Faculty Research Award. We would like to thank Luke Zettlemoyer, Eunsol Choi, and Mohit Iyyer for helpful discussion and feedback.

References

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *ACL*.

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

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*. IEEE, pages 13–18.

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

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational linguistics* 29(4):589–637.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. pages 259–268.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *WMT@ ACL*. pages 266–274.

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

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases* pages 35–50.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, pages 643–650.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Human Factors in Computing Systems*. ACM, pages 3819–3828.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. pages 4069–4079.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* .

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

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*. pages 2930–2939.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015a. A joint framework for coreference resolution and mention head detection. In *CoNLL*. page 10.

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

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *ACL*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *EMNLP*. pages 492–501.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *EMNLP*. pages 777–789.

Karen Ross and Cynthia Carter. 2011. Women and news: A long and winding road. *Media, Culture & Society* 33(8):1148–1165.

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

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

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

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