Reading Assignment 4
Hamza Rashid, 260971031
COMP 550, Fall 2024

The paper, *"Gender Bias in Coreference Resolution"* (Rudinger et al., 2018), proposes dataset schemas for measuring gender bias in Coreference Resolution Systems (CRS). The authors focus on gender bias with respect to occupations, evaluating the accuracy of rule-based, statistical, and neural coreference systems in resolving a pronoun (male, female, or neutral) to a coreferent antecedent that is either an occupation or a participant. They constructed a challenge dataset, *Winogender schemas*, in the style of *Winograd schemas*, wherein a pronoun must be resolved to one of two previously mentioned entities in a sentence. The authors followed good practice by validating their hand-crafted dataset on Amazon's Mechanical Turk (MTurk) with 10-way redundancy, with 94.9% of responses agreeing with their intended answers. This shows that the authors designed test sentences where correct pronoun resolution is not a function of gender. However, they do not report on their MTurk workers' approval ratings, nor do they use the Winograd schemas to filter annotators. They measure gender bias in coreference resolution systems by varying only the pronoun's gender and examining the impact of this change on resolution (revealing cases where coreference systems may be more or less likely to recognize a pronoun as coreferent with a particular occupation based on pronoun gender). An unbiased model is expected to not exhibit sensitivity to pronoun gender in its resolution accuracy, resolving a male or female pronoun to an occupation or participant with equal likelihood. They correlate this bias with real-world and textual gender statistics. The models tested were: the Stanford multi-pass sieve system (Lee et al., 2011; rule-based), Durrett and Klein's (2013; statistical) system, and the Clark and Manning (2016a; neural) deep reinforcement system.

To construct the dataset, the authors used a list of 60 one-word occupations obtained from Caliskan et al. (2017), with corresponding gender percentages available from the U.S. Bureau of Labor Statistics (BLS). For each occupation, there are two similar sentence templates: one in which the pronoun is coreferent with the occupation, and one in which it is coreferent with the participant. For each sentence template, there are two instantiations for the participant (a specific participant, e.g., "the passenger," and a generic participant, "someone"). Thus, the resulting evaluation set contains 720 sentences: 60 occupations $\times$ 2 sentence templates per occupation $\times$ 2 participants $\times$ 3 pronoun genders.

The Winogender schemas revealed varying degrees of gender bias in all three systems. In particular, 68% of male-female minimal pair test sentences are resolved differently by the rule-based system; 28% for statistical; and 13% for neural. Overall, male pronouns were more likely to be resolved to the occupation antecedent than female or neutral pronouns across all systems. As shown in Figure 4 of the paper, the systems' gender preferences for occupations correlate with BLS and the gender statistics from text (Bergsma and Lin, 2006; B&L), which these systems access directly. All models performed worse in "gotcha" sentences, in which the pronoun gender does not match the majority gender (BLS) of the occupation (correct resolution). The paper discussed potential bias amplification involving the occupation *manager*: 38.5% female according to BLS, and mentions of *manager* in the B&L resource are only 5.18%, yet no managers were predicted to be female by any of the coreference systems (percentage-wise differences in real-world statistics may translate into absolute differences in system predictions). During evaluation, a rule-based system may amplify the biases of its hand-crafted rules (which may amplify the biases in the task dataset(s) and external resources). A statistical system is vulnerable to the bias of a feature function associating an occupation with a pronoun (which can be informative, yet biased, for occupations occurring less frequently in the data), and a neural system's pre-trained embeddings are prone to encoding latent biases from its pre-training data. Gender bias is often introduced into the system as an unintended consequence of task-specific model construction or training. System-level biases can lead to further amplification in society through human-AI interaction, causing a cycle of bias.

The authors note that the Winogender schemas have high positive predictive value but low negative predictive value. That is, they may demonstrate the presence of gender bias in a system, but not prove its absence. This follows from the dataset's focus on gender bias in occupations; if a model does not exhibit sensitivity to pronoun gender in this setting, it may still exhibit gender bias in different topics (e.g., crime data across genders).

In conclusion, the paper presents precise schemas for measuring the presence of gender bias in a CRS. Their dataset underwent rigorous validation through crowdsourcing, and they used appropriate data (BLS and B&L) to compare these systems' biases; they are all of North American origin. The dataset is small, and the authors do not explore or inquire about the generalizability of these results across more models. Further, there is no discussion of the importance of using an evaluation dataset whose national origins are the same as those of the models being evaluated. This is critical due to the varying degrees of gender bias across nations. Similarly, there is no discussion of how the national origins of the models' training corpora impact their gender bias in coreference resolution. The Winogender schemas successfully revealed varying degrees of gender bias in all three systems, but the schemas may be extended broadly to probe for other manifestations of gender bias.

References

- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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

- Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.

- Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Conference on Natural Language Learning (CoNLL) Shared Task*.

- 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*, Seattle, Washington. Association for Computational Linguistics.

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

- Crime data. `https://www.ussc.gov/research/quick-facts/individuals-federal-bureau-prisons#:~:text=Individual%20and%20Offense%20Characteristics,93.0%25%20are%20men.`