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). To this end, the authors focus on gender bias with respect to occupations, evaluating the accuracy of Rule-based, Statisical, 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 redunancy, 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 or method for selecting them (e.g., qualification tests). With the Winogender schemas in hand, an unbiased model is expected to exhibit no sensitivity to pronoun gender in its accuracy, and to resolve a male or female pronoun to an occupation or participant with equal likelihood.

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 referent (a specific participant, e.g., "the passenger," and a generic paricipant, "someone."). Thus, the resulting evaluation set contains 720 sentences: 60 occupations $\times$ 2 sentence templates per occupation $\times$ 2 participants $\times$ 3 pronoun genders.

A key observation is that when each CRS's predictions diverge based on pronoun gender, they do so in ways that reinforce and magnify real-world occupational gender disparities. 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), which these systems access directly. A notable case is "gotcha" sentences in which pronoun gender does not match the occupation's majority gender (BLS) if the occupation is the correct answer; all models perfomed worse in this scenario. The paper discussed a potential case of bias amplification involving the occupation manager: 38.5% female in the U.S. 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. The mechanism in which dataset bias could be amplified at the system level depends on the system; a Rule-based system is susceptible to the biases of the human designer and the resulting hand-crafted rules, a Statistical system is vulnerable to the biases of a feature function associating an occupation with a gender, and a Neural system's pre-trained embeddings is prone to capturing latent biases between occupations and genders due to the biases in its pre-training data. System-level bias can lead to further amplification in society as the market for consumer chat-based LLM's continues to grow. For example, the integration of Gemini in Google search is prone to gender bias in queries such as "most impactful computer scientists", where the contributions of Ada Lovelace are likely to be overlooked in comparison with Alan Turing.

For many occupations, the tested systems strongly prefer to resolve pronouns of one gender over another. The paper demonstrates that this preferential behavior correlates both with real world employment statistics and the text statistics that these systems use. They posited that these systems overgeneralize the attribute of gender, leading them to make errors that humans do not make. The authors developed a simple and effective way to measure the degree of gender bias in the tested models. However, they note the limitations of Winogender schemas, viewing them as having high positive predictive value and low negative predictive value. In other words, the schemas may demonstrate the presence of gender bias in a system, but not prove its absence. This follows from their dataset's focus on occupation bias, as the models may be good at coreference resolution in this setting, but exhibit gender bias in different topics.

In conclusion, the paper presents precise schemas for measuring the presence of gender bias in a CRS. Their dataset has gone through rigorous validation through crowdsourcing, and they use appropriate data (BLS and B&L) to compare these systems' biases with. The authors do no explore or inquire the generalizability of these results across more models, and there is no discussion on the importance of using an evaluation dataset whose national origins are the same as that of the models being evaluated. This is critical due to the varying degrees of gender bias across nations. Furthermore, there is no discussion of how the training method and national origin of training corpora for the CRS models impacts their gender bias in coreference resolution. In the end, the authors measured the presence occupational gender bias in Rule-based, Statistical, and Neural Coreference Resolution Systems successfully, but Winogender 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. In *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.

- https://rshiny.ilo.org/dataexplorer39/?lang=en&id=SDG_T552_NOC_RT_A