1 Description of the reading material

The paper **“Gender Bias in Coreference Resolution”** [4] investigates **gender bias in coreference resolution systems**. By using **Winogender schemas** [3], a dataset of 720 sentences modeled after Winograd schemas, the study evaluates the gender bias in three types of systems: **rule-based**, **statistical**, and **neural**. The findings reveal that these systems show **bias in resolving pronouns**, correlating with real-world occupational gender statistics and textual biases. Male pronouns are disproportionately linked to occupations compared to female or neutral pronouns. This work highlights **bias amplification** in AI models, urging for the development of less biased systems.

► ***Strength.*** One of the main strengths of the paper is the creation of a **specialized evaluation set** in the form of Winogender schemas. This **novel dataset** is well-designed, providing a **reliable framework** to assess how pronouns are resolved in relation to occupational gender stereotypes. Additionally, the **comprehensive evaluation** across three different system paradigms (rule-based, statistical, and neural) highlights the **widespread and systemic nature** of **gender bias in AI models**. By correlating system outputs with **real-world labor statistics** (i.e., BLS [2]) and **textual data** (i.e., B&L [1]), the paper establishes **clear evidence** of how biases are **reflected and amplified** in machine learning systems, making its findings highly relevant to real-world applications.

► ***Weakness.*** The scope of the analysis is limited to occupational gender bias, **neglecting other potential manifestations of gender bias**. Another limitation is its **diagnostic nature**—while the schemas effectively demonstrate the **presence of bias**, they do not comprehensively assess **bias absence**. While this paper focuses on the **validation and analysis** of observed system bias, it would benefit from including a discussion on **potential methods for debiasing** based on their findings. Despite a parallel paper focusing on debiasing, incorporating such discussions would provide greater technical depth. Last, this paper **does not explicitly include the models used in three paradigms**.

► ***Questions.*** The study identifies correlations between system bias and real-world/textual gender statistics (in Fig.4). Since neural paradigm with millions/billions of parameters seems to be less biased, **is biases related to model architectures and number of parameters**? If we manually remove the gender-based rules/features, **will there be (perhaps obvious) gender biases? Will systems indirectly learn biases from other rules/features?**

2 Discussion

► ***Measure gender bias and unbiased system.*** This paper measures gender bias using Winogender schemas, a dataset focused on **OCCUPATION**, **PARTICIPANT**, **PRONOUN**, and **templates**. It incorporates **real-world occupational statistics** and **textual gender distributions** for further comparisons. An unbiased system should resolve pronouns based solely on **sentence semantics**, treating male, female, and neutral pronouns **equally**. Regarding the experiment results, the correlation between gender and occupations should be **zero** in Tab.1, with no direct associations between them. Ideally, there should be **no/weak relationship** between textual resources, occupational statistics, and any system paradigm. Similarly, an unbiased system should display **zero differences** between male and female predictions in Fig.4, regardless of the x-axis metric (**B&L or BLS**).

► ***Mechanism that amplifies dataset biases at the system level.*** The bias amplification mechanism consists of two steps: **dataset bias leading to system bias amplification** and **system bias leading to societal bias amplification**. Dataset bias originates from **skewed training data**, such as BLS reporting 38.5% female managers, B&L mentions only 5.18% female manager. Systems trained on biased data inherit these patterns: **rule-based systems rigidly apply predefined rules**, **statistical systems overlearn correlations**, and **neural systems depend on biased embeddings**. Consequently, systems disproportionately resolve **male pronouns** to certain occupations, amplifying bias. **System-level bias propagates downstream** in real-world applications like **hiring tools**, **reinforcing stereotypes**, **discouraging women**, and **skewing societal perceptions and decisions**.

► ***Explanations for low negative predictive value.*** The paper has **low negative predictive value** because the proposed Winogender schemas are designed to **demonstrate the presence of gender bias** in coreference resolution systems but cannot conclusively prove its **absence**. In other words, while the schemas are effective at **identifying bias when it exists** (high positive predictive value), the absence of bias in the test results does not guarantee that the system is unbiased. This limitation arises because the test focuses on **specific scenarios** (i.e., occupational gender bias) and cannot comprehensively evaluate all potential forms or sources of gender bias in the system.

References

[1] Shane Bergsma and Dekang Lin. 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, 2006.

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

[3] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

[4] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018.