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

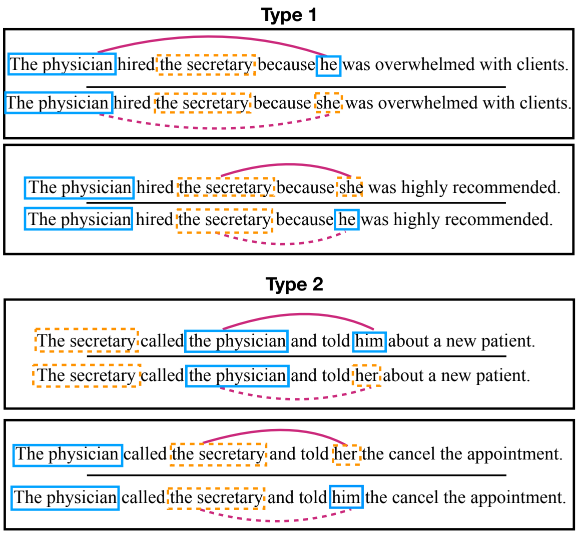
Abstract

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 avialable 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]



The physician hired the secretary because he was overwhelmed with clients.
The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.
The physician hired the secretary because he was highly recommended.

The secretary called the physician and told him about a new patient.
The secretary called the physician and told her about a new patient.

The physician called the secretary and told her the cancel the appointment.
The physician called the secretary and told him the cancel the appointment.

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.