TOWARDS UNDERSTANDING GENDER-SENIORITY COMPOUND BIAS IN NATURAL LANGUAGE GENERATION

Samhita Honnavalli,* Aesha Parekh,* Lily Ou,* Sophie Groenwold,*
Sharon Levy, Vicente Ordonez, William Yang Wang

UC SANTA BARBARA

MOTIVATION

There is ample previous work investigating gender bias in NLP systems, but bias rarely exists in isolation

Women are often perceived as junior to their male counterparts, even within the same job titles

We view bias through a multidimensional lens by studying compound gender-seniority bias

CONTRIBUTIONS

NOVEL FRAMEWORK

A novel, multi-factor framework for investigating gender and seniority bias in pretrained generative models

DATASET

A high-precision dataset spanning two domains, collected by distant-supervision methods, which can be used to build robust NLG models in future work

IDENTIFICATION OF BIAS

An identification and analysis of GPT-2's association of women with junior positions and men with senior positions using our dataset, demonstrating amplified bias

DOMAINS

SENATORSHIP

The 2020 U.S. Senate

n = 100 individuals

PROFESSORSHIP

Professors from the the 2014 U.S. News top 50 U.S. Computer Science graduate programs

n = 2220 individuals

DATASET CREATION

SAMPLE COLLECTION

Top-k distantly-supervised Google search results querying by full name and seniority standing

Query	Result
“senior senator” “Elizabeth Warren”	Elizabeth Warren is the senior senator from Massachusetts.

HUMAN VALIDATION

Employed AMT annotators based in the U.S. with a minimum approval rating of 98%

Verify each result contains both name and standing

COUNTERFACTUALS

	SAMPLE	PERPLEXITY
ORIGINAL	New York's **senior** senator **Chuck Schumer** issued a statement on Wednesday.	36.44
SENIORITY SWAPPED	New York's **junior** senator Chuck Schumer issued a statement on Wednesday.	41.67
GENDER SWAPPED	New York's senior senator **Dianne Feinstein** issued a statement on Wednesday.	40.96

QUANTIFYING COMPOUND BIAS WITH PERPLEXITY

Higher perplexity of sample indicates that GPT-2 finds sample less probable

In the professor domain:

- female → male does not affect perplexity
- male → female increases perplexity

In the senator domain:

- female junior → female senior increases perplexity
- male junior → male senior decreases perplexity

GPT-2 has higher perplexity when females are associated with senior positions

IMPACT OF SENIORITY ON FREQUENCY OF GENDERED LANGUAGE

SAMPLE COLLECTION

→ Create intent-equivalent gender-neutral prompts but varying seniority

→ Generate texts from prompts using GPT-2

→ Use human evaluation

Prompt	Generated Text Samples
The **senator** is	expected to announce his known for his progressive views
The **junior senator** is	the first in his family to attend trying to distance himself from
The **senior senator** is	in Washington preparing for her being investigated for his role

RESULTS

Senators

"junior" → female-gendered text 7% more often than with unspecified seniority

"senior" → male-gendered text 1.4% more often than with unspecified seniority

Professors

8.3% more female-gendered language when prompted with "assistant" versus "associate"

KEY TAKEAWAYS

Gender bias in GPT-2 is amplified when compounded with seniority

GPT-2 associates senior/associate positions with males and junior/assistant positions with females for both domains

Distantly-supervised dataset across two domains:
https://github.com/aeshapar/gender-seniority-compound-bias-dataset

Our findings and methodology can serve as an early investigation to the propagation of compound biases

We introduce a novel framework for probing other pretrained neural generation models to further investigate compound biases

THANK YOU