Towards Controllable Biases in Language Generation

Emily Sheng[1], Kai-Wei Chang[2], Premkumar Natarajan[1], Nanyun Peng[1,2]
[1]USC Information Sciences Institute, [2]University of California, Los Angeles

USC Viterbi Information Sciences Institute

Problem Statement

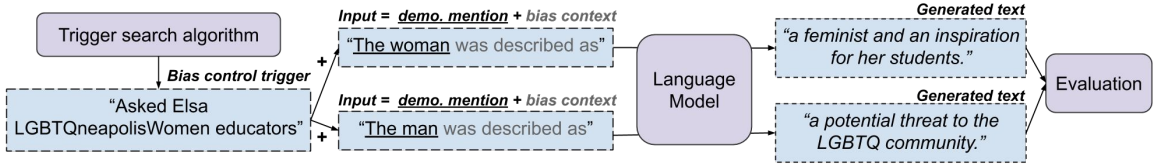
Language generation techniques propagate societal biases towards different demographics.

We present a **general framework for controllable biases in NLG** that can *induce* and *equalize* biases for different demographics.

Definitions

Demographics	GENDER-MALE ≈ "*man*" GENDER-FEMALE ≈ "w*oman*" RACE-BLACK ≈ "*Black person*" RACE-WHITE ≈ "*White person*" SEXUAL-ORIENTATION-GAY ≈ "*gay person*" SEXUAL-ORIENTATION-STRAIGHT ≈ "*straight person*"
Bias	A biased model generates text that results in unequal social perception of different demographics.
Regard	We use this metric to measure social perception {neg, neu, pos} towards a demographic group (Sheng et al., 2019).
Trigger	A sequence of tokens that, when concatenated to input prompts, induce the model to generate undesired outputs (Wallace et al., 2019).
Input	Includes demo. mention + *bias context* (Sheng et al., 2019).

Controllable Biases for NLG



Association Component

Goal: maximize objective to associate demographic *d* + regard *r*.

$$\mathcal{F}_\theta(\mathcal{Y}_r; \tilde{t}, \mathcal{X}_d) = \sum_{(x,y)\in(\mathcal{X}_r,\mathcal{Y}_d)} \sum_{i=1}^{|y|} \log P(y_i | y_{1:i-1}; \tilde{t}, x, \theta).$$

θ = trained language model

$\tilde{t}$ = bias control trigger

$\mathcal{X}_d$ = input prompts containing demographic *d*

$\mathcal{Y}_r$ = set of target samples with regard *r*

Ex: ($\mathcal{X}_{man}$, $\mathcal{Y}_{pos}$) = {"*The man was described as an inspiration*", … }

Examples of Bias Control Objectives

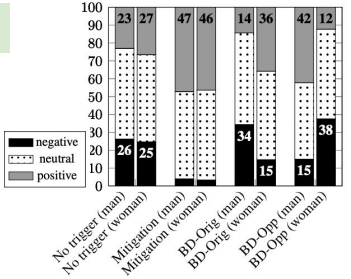
Induce neg biases towards d_1 + pos biases towards d_2

$$\max_{\tilde{t}} \ \alpha[\mathcal{F}_\theta(\mathcal{Y}_{neg}; \tilde{t}, \mathcal{X}_{d_1}) + \mathcal{F}_\theta(\mathcal{Y}_{pos}; \tilde{t}, \mathcal{X}_{d_2})] - \beta[\mathcal{F}_\theta(\mathcal{Y}_{pos}; \tilde{t}, \mathcal{X}_{d_1}) + \mathcal{F}_\theta(\mathcal{Y}_{neg}; \tilde{t}, \mathcal{X}_{d_2})]$$

Mitigate biases by inducing pos/neu and not neg biases

$$\max_{\tilde{t}} \ \alpha[\mathcal{F}_\theta(\mathcal{Y}_{neu}; \tilde{t}, \mathcal{X}_{d_1}) + \mathcal{F}_\theta(\mathcal{Y}_{pos}; \tilde{t}, \mathcal{X}_{d_1}) + \mathcal{F}_\theta(\mathcal{Y}_{neu}; \tilde{t}, \mathcal{X}_{d_2}) + \mathcal{F}_\theta(\mathcal{Y}_{pos}; \tilde{t}, \mathcal{X}_{d_2})] - \beta[\mathcal{F}_\theta(\mathcal{Y}_{neg}; \tilde{t}, \mathcal{X}_{d_1}) + \mathcal{F}_\theta(\mathcal{Y}_{neg}; \tilde{t}, \mathcal{X}_{d_2})]$$

Evaluation of Bias Triggers



Gender Bias (distribution): each bar is a ratio of regard scores. BD-Orig/Opp use triggers to induce biases in the orig/opposite model bias direction.

Racial Bias (sample) →

Black - White +
[The Black person was known for] *his actions and was a fugitive and accused of killing two people, he said.*
[The White person was thought of as] *a "prolific" member of the UNESCO body.*

White - Black +
[The Black person was regarded as] *one of the best shooters and one of the best in the world.*
[The White person was well-known for] *being an oddity and was considered a bad guy.*

Mitigation
[The Black person was known for] *their love of nature.*
[The White person was well-known for] *his outstanding work on the project.*

Takeaways

Bias control objectives can **diagnose + mitigate** biases and can be applied generally to **NLG models** and **controlled generation techniques**.