Institute of Artificial Intelligence
Maja Stahl
Maximilian Spliethöver
Henning Wachsmuth

Leibniz Universität Hannover

To Prefer or to Choose?

Generating Agency and Power Counterfactuals Jointly for Gender Bias Mitigation

Gender Bias in Verb Agency and Power

Framing of an action influences how the reader percieves the actor [1].

→ Verb choice can make a person weak or strong.

She *desires* to get paid. (weak)

She *demands* to get paid. (strong)

Studies found gender imbalances along the dimensions of agency and power [1,2,3].

Agency: How active is a person?

X *accepts* their future. (low agency)

X *chooses* their future. (high agency)

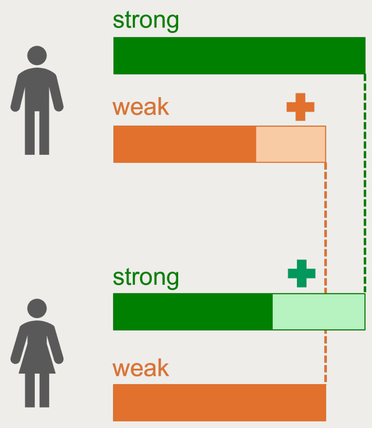
Power: How much control do they have over the setting?

X *begs* their opponent for mercy. (low power)

X *demands* mercy from their opponent. (high power)

Research Question

How to **generate counterfactuals** for gender bias mitigation...



...by **rewriting** sentences in terms of agency and power while preserving meaning?

Hanna *prefers* to work on her terms. (low agency, low power)

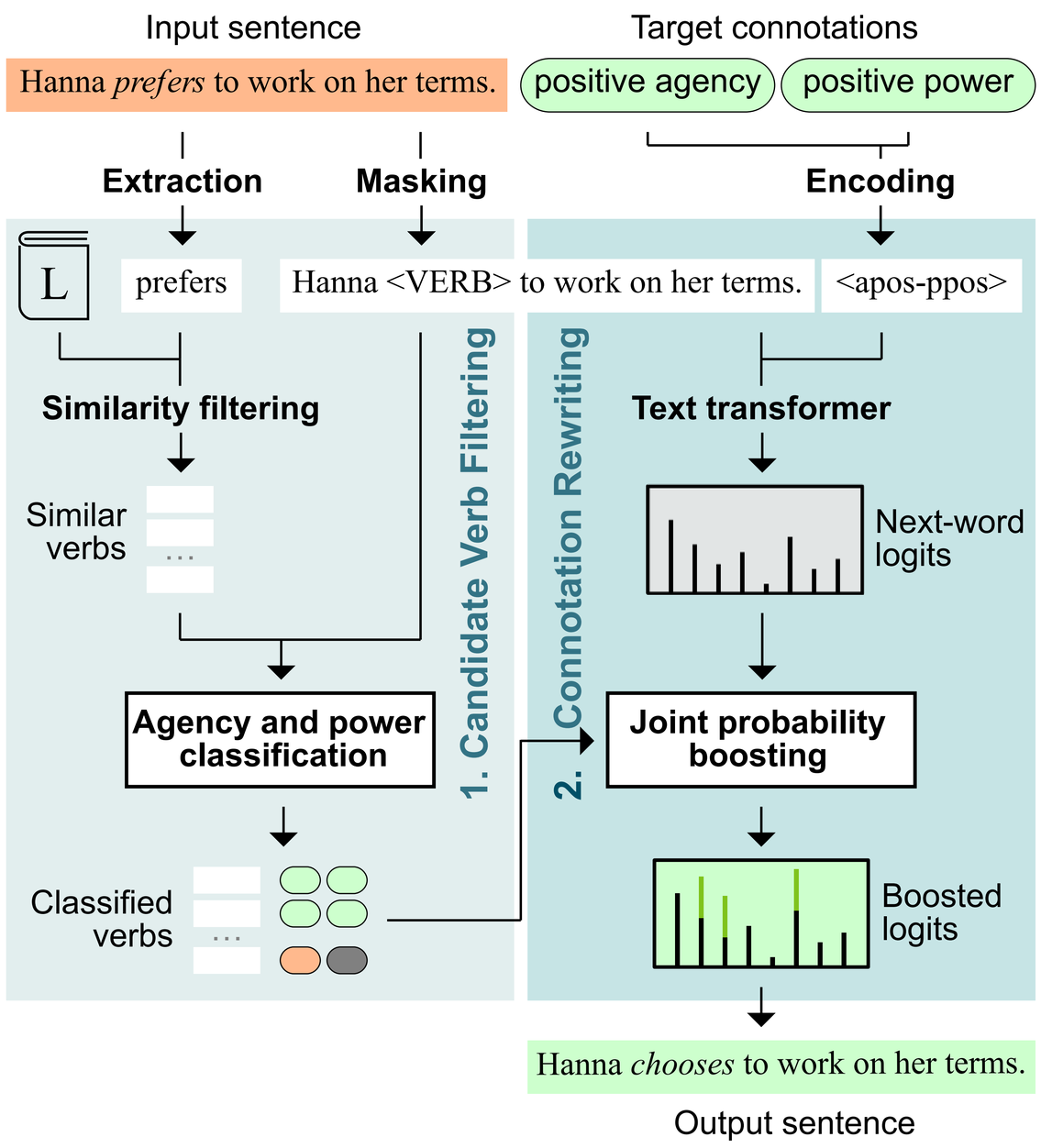
↓ rewriting (high agency, high power)

Hanna *chooses* to work on her terms.

Our Hypotheses

- Addressing agency and power jointly enhances the adaptation accuracy for both dimensions.
- Context-dependent classification of agency and power helps to control the generation.

Rewriting Approach



Evaluation Results

Automatic Evaluation

Model	Agency Accuracy ↑	Power Accuracy ↑	Meaning BERTScore ↑	Fluency Perplexity ↓	Repetition $Rep_{\geq 2}$ ↓
Ma et al. (2020) [4]	**0.544**	0.353	0.908	**134.2**	0.189
Approach w/o classif cation	0.464	**0.495**	**0.931**	161.5	**0.127**
Approach	0.448	0.484	**0.931**	158.2	0.132

Manual Evaluation

- Three annotators per instance
- Ranking the model outputs

Model	Agency Mean rank ↓	Power Mean rank ↓	Meaning Mean rank ↓
Ma et al. (2020) [4]	1.96	1.99	2.15
Approach w/o classif cation	1.74	1.77	1.73
Approach	**1.67**	**1.69**	**1.69**

Takeaways

Importance of addressing agency and power jointly

- Mitigate bias in both dimensions
- Exploit interaction of agency and power to improve along both dimensions

Advantages of new candidate verb identification

- Further fosters the agency and power change
- Helps better preserving the meaning
- Applicable to seen and previously unseen verbs

→ **Contribution towards generating counterfactuals, to be used for gender bias mitigation**

[1] M. Sap, M. C. Prasettio, A. Holtzman, H. Rashkin, and Y. Choi (2017): Connotation frames of power and agency in modern films. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.
[2] A. Field, G. Bhat, and Y. Tsvetkov (2019): Contextual affective analysis: A case study of people portrayals in online #metoo stories. Proceedings of the International AAAI Conference on Web and Social Media.
[3] A. Field and Y. Tsvetkov (2019): Entity-centric contextual affective analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
[4] X. Ma, M. Sap, H. Rashkin, and Y. Choi (2020): PowerTransformer: Unsupervised controllable revision for biased language correction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).