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 perceives the actor [1].

Verb choice can make a person weak or strong.

She desires to get paid.  She demands to get paid.

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

Agency: How active is a person?

X accepts their future.  X chooses their future.

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

X begs their opponent for mercy.  X demands mercy from their opponent.

Research Question
How to generate counterfactuals for gender bias mitigation...

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

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.

Evaluation Results
Automatic Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Agency Accuracy ↑</th>
<th>Power Accuracy ↑</th>
<th>Meaning BERTScore ↑</th>
<th>Fluency Perplexity ↓</th>
<th>Repetition Repeat ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ma et al. (2020) [4]</td>
<td>0.544</td>
<td>0.353</td>
<td>0.908</td>
<td>134.2</td>
<td>0.189</td>
</tr>
<tr>
<td>Approach w/o classification</td>
<td>0.464</td>
<td>0.495</td>
<td>0.931</td>
<td>161.5</td>
<td>0.127</td>
</tr>
<tr>
<td>Approach</td>
<td>0.448</td>
<td>0.484</td>
<td>0.931</td>
<td>158.2</td>
<td>0.132</td>
</tr>
</tbody>
</table>

Manual Evaluation
- Three annotators per instance
- Ranking the model outputs

<table>
<thead>
<tr>
<th>Model</th>
<th>Agency Mean rank ↓</th>
<th>Power Mean rank ↓</th>
<th>Meaning Mean rank ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ma et al. (2020) [4]</td>
<td>1.96</td>
<td>1.99</td>
<td>2.15</td>
</tr>
<tr>
<td>Approach w/o classification</td>
<td>1.74</td>
<td>1.77</td>
<td>1.73</td>
</tr>
<tr>
<td>Approach</td>
<td>1.67</td>
<td>1.69</td>
<td>1.69</td>
</tr>
</tbody>
</table>

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