Jieyu Zhao  https://jyzhao.net

University of Maryland, College Park

**Motivation**

NLP plays an important role in many applications, such as resume filtering, text analysis and information retrieval. Despite remarkable accuracy enabled by current machine learning models, it may discover and generalize the societal biases implicit in the data. For example, an automatic resume filtering system may unconsciously select candidates based on their gender and race due to implicit associations between applicant names and job titles, causing the societal disparity. Inspired by such observations, my research goal is to analyze potential stereotypes exhibited in various machine learning models and to develop computational approaches to enhance the fairness in a wide range of NLP applications.

**Contribution**

1. Propose metrics as well as build new datasets to evaluate bias in different models
2. Qualify the bias in different applications
3. Propose different ways to mitigate the bias from different layers of the model

**Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints**

**Summary**

1. Dataset contains high gender bias
2. Model will amplify the existing gender bias
3. We can reduce bias amplification around 50% with insignificant performance loss

**Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints**

**1. Dataset Bias**

<table>
<thead>
<tr>
<th>Gender</th>
<th>WW</th>
<th>Dataset</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>64%</td>
<td>66%</td>
<td>66.4%</td>
</tr>
<tr>
<td>Female</td>
<td>36%</td>
<td>34%</td>
<td>33.6%</td>
</tr>
</tbody>
</table>

**2. Bias Amplification**

- Model will amplify the existing gender bias
- We can reduce bias amplification around 50% with insignificant performance loss

**Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods**

**Summary**

1. Build a gender balanced test set, WinoBias.
2. Demonstrate that different coreference systems show gender bias in this WinoBias dataset.
3. A data-augmentation, in combination with word-embedding debiasing technique, removes the bias without significantly affecting their performance

**Bias mitigation**

- **Gender Swapping**
  I. Build a dictionary of gendered terms
  II. An additional training corpus where all male entities are swapped for female entities and vice-versa (Aug.)
  III. Anonymize the dataset (Anon.)
- **Modify the resource**
  I. Debiased word embeddings or balanced gender list

**Others**

- **Word Representations:** Detect and reduce biases from representations. EMNLP 2018, NAACL 2019.
- **Multilingual:** Understanding the bias in other languages besides English. EMNLP 2019, ACL 2020.
- **Evaluation Metric:** Corpus-level performance gap for group bias detection could be incomplete → LOGAN.
- **Model Intervention:** Adding ethical instructions to intervene in a model → LEI. ACL 2021

Next ?