Discussion: Man is to Computer Programmer as Woman is to Homemaker? Debiassing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

Computational Social Media
Alejandro Ramírez Atrio
22/05/2020
Structure

0. In perspective: more recent tests on debiasing (Gonen & Goldberg 2019)

1. How to test for bias?

2. Identifying gender-related information

3. Problems with biased AI
0. Gonen & Goldberg (2019) Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them
Are the embeddings actually debiased?

They compare this paper’s method and Zhao et al *Learning gender-neutral word embeddings* (2018) (they debias at training time).

**Experiment 1**

Take 1000 gender-neutral most biased words from the original embeddings of our paper (500 male-associated and 500 female-associated)

Cluster them with $k$-means, then again with the debiased embeddings
Original: 99.9% accuracy

Debiased: 92.5% accuracy

Gonen & Goldberg (2019)
Experiment 2

Take 5,000 most biased words (2,500 each gender)

Train SVM classifier to predict the gender for 1,000 (500 each) and test on 4,000

Original: 98.25% accuracy

Debiased: 88.88% accuracy
The other method does not get much better results.

Their conclusion: gender-related information is present in many different dimensions of a word embedding.

It cannot be easily deleted from a word’s representation when it has already been learned by the model.
(Reminder) Bolukbasi et al (2016) 150 analogies generation task:

Original: 19% show some gender stereotype
Debiased: 6% show some gender stereotype

Aren’t the debiased embeddings much better?
1. How to test for bias?
Bolukbasi et al (2016) test by asking people if generated analogies reflect any gender stereotypes

QUESTION(s):

Should we test for bias by asking people or by performing statistical analysis on the words’ representations?

Do you think a machine may be able to “see” some gender bias that humans do not? Is it possible that the people that they use for testing might have some (unconscious?) biases?

Or maybe the way to come up with analogies is not sophisticated enough?
2. Identifying gender-related information
Defining the gender subspace

“[...] we asked crowdworkers to generate two lists of words: one list corresponding to words that they think are gendered by definition (waitress, menswear) and a separate list corresponding to words that they believe captures gender stereotypes (e.g., sewing, football).

From this we generated the most frequently suggested 50 male and 50 female words for each list to be used for a classification task. For each candidate pair, for example (she, he), we say that it accurately classifies a crowd suggested female definition (or stereotype) word if that word vector is closer to she than to he.

[...] The accuracies are high, indicating that these pairs capture the intuitive notion of gender.”
Defining gender-specific list of words

List of 6,499 words

Which are the result of applying an SVM which was trained with 218 words to the total 3M word list (Appendix C)

Which were semi-automatically gathered by searching wordnet definitions that contain one word of the 7 following pairs:
<table>
<thead>
<tr>
<th>female</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>woman</td>
<td>man</td>
</tr>
<tr>
<td>girl</td>
<td>boy</td>
</tr>
<tr>
<td>sister</td>
<td>brother</td>
</tr>
<tr>
<td>daughter</td>
<td>son</td>
</tr>
<tr>
<td>grandmother</td>
<td>grandfather</td>
</tr>
<tr>
<td>wife</td>
<td>husband</td>
</tr>
</tbody>
</table>

“Note that the choice of words is subjective and ideally should be customized to the application at hand.”

The gender subspace is (simplifying) defined by what people perceive as gender stereotypes (in particular, by the words that they can come up with)

Gender-specific words are ultimately (semi-automatically) learned from very simple gendered words
QUESTION(s):

Are these good ways to capture gender-related information? (Both gender stereotypes and/or gender-specific/gender-neutral words)

Can you imagine a problem with this either of the two methods?

Can you think of some alternative method(s) that might be better for one of the two tasks?

Is there a clear “gender-relevant subspace” in a words’ vectorial representation, or maybe gender-related information is much more spread out and difficult to isolate?
3. Problems with biased AI
Likely that in the not so distant future (and already now) AI will be used for informing decision making in:

- College admissions
- Hiring process
- Credit and loans, etc.

QUESTION(s)

Can you foresee a situation in which it will affect you in your future (bias may not be just gender, also ethnicity, religion, economic class, language, etc.)?

Do you think it may have already affected you?

Should the law regulate whether AI systems must not show any biases (anti-discrimination laws)? Or maybe only some systems? If so, how should it be enforced to test for so many possible biases?
Thanks!