Title: Man is to Computer Programmer as Woman is to Homemaker?
Authors: Tolga Bolukbasi et al.
Published in:

Name: Yuxin Huang
Engineering Research Paper
Question–Answer Form

- What is your take-away message from this paper?

The concept of word embeddings : a framework to represent text data as vectors which has been used in many machine learning and NLP.
Gender bias in word embeddings can be divided to direct and indirect bias.
While reducing bias in a gender subspace, we also want to maintain the properties of the word other than gender.

- What is the motivation for this work (both people problem and technical problem), and its distillation into a research question? *Why doesn't the people problem have a trivial solution? What are the previous solutions and why are they inadequate?*

The blind application of machine learning runs the risk of amplifying biases present in data, influencing social bias in the real world. The essay does not present any previous solution, yet without their debiasing algorithms, gender bias maintains in the word embeddings.

- What is the proposed solution (hypothesis, idea, design)? *Why is it believed it will work? How does it represent an improvement? How is the solution achieved?*

They present a debiasing model that can largely reduce the gender bias in word embeddings. They believe it will work because they implement their model into w2vNEWS and find that they largely reduce gender bias in the embeddings.
After identifying gender subspace, they neutralize the gender component of gender-neutral words and equalize the distance between a gender-definition word pair with respect to neutral words. At last, they use soft debiasing to maintain some properties of the words that are not related to gender.

- What is the author's evaluation of the solution? *What logic, argument, evidence, artifacts (e.g., a proof-of-concept system), or experiments are presented in support of the idea?*

They implement their debiasing model into w2cNEWS embeddings and compare to the original w2cNEWS embeddings. They found that the percentage of either direct or indirect bias is largely reduced after debiasing.

- What is your analysis of the identified problem, idea and evaluation? *Is this a good idea? What flaws do you perceive in the work? What are the most interesting or controversial ideas? For work that has practical implications, ask whether this will work, who would want it, what it will take to give it to them, and when might it become a reality?*

The debiasing algorithm is effective, yet their solution of indirect bias is vague. For the hard-debiasing part, I only get the solution of gender neutral words and gender pair words. How do they eventually reduce the indirect bias in word embeddings?

- What are the paper's contributions (author's and your opinion)? *Ideas, methods, software, experimental results, experimental techniques...?*

The author: by reducing gender bias in word embeddings, one can lessen the opportunity that gender stereotype appears in the real world context (e.g. Google News)
My opinion: Although experienmental techniques are not so sophisticated, the fact that they are sending out questionnaires and interact with real population is outstanding.

- What are future directions for this research (author's and yours, perhaps driven by shortcomings or other critiques)?

I think if they want to extent it further, work more on indirect bias' debiasing.

- What questions are you left with? *What questions would you like to raise in an open discussion of the work (review interesting and controversial points, above)? What do you find difficult to understand? List as many as you can.*

What is principal components and what role does it play in the debiasing algorithms?
For hard debiasing formula, w belongs to E, is it separating a word vector to gender-neutral and gender subpace?
The math of soft bias correction.
How do they determine whether it is a biased word?