Research Paper Summary

Summary of research paper #1 "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings"

This research study uses algorithms to greatly eliminate gender bias in embeddings using crowd-worker evaluation as well as standard benchmarks. To eliminate gender inequality, the embedding of gender-neutral words is modified in this analysis by eliminating their gender connection. This is accomplished by first capturing gender inequality in the direction of word embedding and then linearly separating the gender-neutral words from the gender definition words in the word embedding. The research shows, for example, that the nurse has been moved towards gender equality. It also provides an example of a grandmother and grandfather who are closer to wisdom than a gal and a guy. It asserts that grandmothers are closer to babysit than grandfathers, which means they are more likely to babysit, reflecting a gender bias. When gender inequality is removed from the term "babysit," a neutral word, both grandmother and grandfather are equally matched closer to the word "babysit," thus eliminating gender inequality.

This study uses the hard-debiasing algorithm and soft-debiased algorithm to replicate the analogy generation task (for example, pairs of words that are similar to she-he). The first stage in a debiasing algorithm is to define the gender subspace, which is accompanied by defining a direction of the embedding that detects bias. Second stage involves two options: Neutralize and Equalize. Neutralize means that gender neutral words have a value of zero in the gender subspace. Equalize absolutely equalizes sets of words outside the subspace, imposing the rule that every neutral word is equidistant to all words in each equality set. Furthermore, to determine if the pairing reflected gender stereotypes, yes/no questions were asked of crowd workers in the United States. The findings show that before the debiasing approach was used, 19% of the top 150 analogies displayed gender stereotypes, while when hard-debiasing was used, only 6% showed gender stereotypes. When compared to the soft-denoising process, hard-debiasing produced better performance by significantly reducing both explicit and subtle gender disparities while maintaining embedding utility. The debiasing algorithm used excludes sexism only from gender neutral phrases while retaining the definitions of gender specific words. The study has also shed light on how society is currently skewed, which has a significant impact on word embedding. As a result, rather than word embedding, one should attempt to debias society.

Summary of research paper #2 "Word embeddings quantify 100 years of gender and ethnic stereotypes"

This research creates a context to explain how the temporal complexities of embedding serve to measure shifts in stereotypes and attitudes against women and ethnic minorities in the United States between the twentieth and twenty-first centuries. To ensure that the bias in the embedding accurately reflects sociological trends, this analysis reveals patterns in the embeddings with quantifiable population trends in occupations involvement, as well as historical surveys of stereotypes. The study uses 100 years of text data to train word embeddings to quantify the biases for occupations and adjectives. This research makes use of the embedding bias to investigate historical shifts that would otherwise be difficult to measure. The study shows both gender and racial occupation biases in the embeddings which associate substantially with real occupation participation rates. It shed light on how particular biases diminish over time when other forms of

stereotyping increase. It also illustrates adjective correlations in embeddings that offer details about how different groups of individuals are interpreted over time.

The method in the research study employs two kinds of word lists: group words and neutral words. Group words represent different types of individuals, such as gender and race. Words that are not inherently gendered or racial are considered neutral. The approach entails averaging the group words and then computing the average euclidean distance between each representative group vector and each vector in the neutral word list of interest. If the effect is negative, the embedding is most closely correlated with men's professions. The researcher mainly used a linear regression method to fit the relationship between embedding bias and various external metrics. In this study various occupations and adjectives were used as neutral words. However, there are several drawbacks to this research. For example, true correlations between embedding racism and different external metrics may be nonlinear, which is a problem when researching racial stereotypes. Additionally, particular word lists used in the methodology can yield specific results. Overall, this analysis offers a strong foundation for future studies into the temporal dynamics of assumptions through the prism of word embedding. It also demonstrated certain themes in US culture, such as how some movements, such as the women's movement, are correlated with a dramatic change in the prejudices associated with profession and gender biased adjectives (for example hysterical vs emotional).

Research study comparison: By word embedding complexities, both research papers present the prejudices that exist in society. One paper illustrates the steps that should be taken to demonstrate how attitudes toward women and people of color have shifted from decade to decade. The second research paper, on the other hand, proposes a novel algorithm that greatly reduces the issue of bias in word embedding. We discovered that Google News word2vec vectors were used as a common data source in both research papers. Both studies used cosine similarity to determine the degree of similarity between the groups and the neutral words, and they also used crowd worker assessment in their studies. Overall, both studies demonstrated how society's bias evolved over time due to various movements in history, as well as how society's bias made machine learning more biased than ever.

Questions: 1) I'm interested in learning how words can be classified as gender neutral.
2) The second research paper states that the embeddings used are completely "black box," which I'm not sure what that means.