CIS399: The Science of Data Ethics	Lecture 14: March 11, 2019
Instructor: Michael Kearns & Ani Nenkova Brown	Scribes: Priyansh Sharma, Maxwell

Note: Links to the papers referenced can be found by hovering over the section headers.

1 Review of Word Embeddings

Recent classes have centered around **word embeddings** - numerical vector representations of words and their meaning.

The primary metric used to determine similarity between embeddings in a number of NLP applications is the **cosine similarity**:

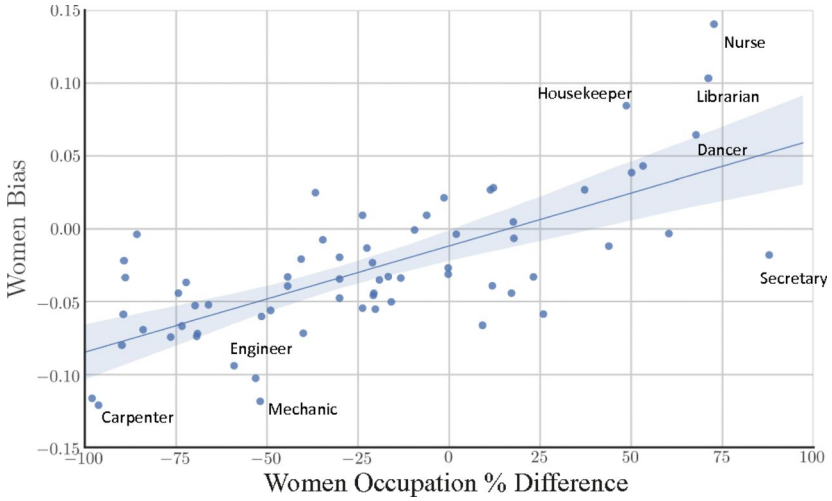
$$sim(\overline{x}, \overline{y}) = \frac{x_1 y_1 + x_2 y_2 ... x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Also discussed previously was how researchers have used a variety of tasks (analogy tasks, searching for words in a static data study, etc.) to demonstrate these embeddings have meaning.

We then went on to discuss that not only do these embeddings have practical meaning, they contain biases (see: Science paper). Some questions that naturally arise are:

- How much of this bias do we want to remove? After all, at some level, we want our systems to recognize existing relationships, but avoid the ones that are offensive or prejudicial.
- How has this bias in embeddings changed over time?

2 Gonen and Goldberg, 2019



As a brief aside, we dug deeper into this second question using a recent publication. The paper in question used Google Books data to train embeddings over 100 years. The researchers then

sought out the association between certain professions and gender over time. The study found that **word representations for a given time period actually fairly accurately represented the proportion of women in that occupation**.

3 Debiasing Word Embeddings

This paper conducted a series of experiments to compare state-of-the-art embeddings to their debiased counterparts. The study looked primarily at the word2vec representation against hard debiasing and the GloVe (Global Vectors for Word Representation) representation against neural network (NN) debiasing.

3.1 Experiment 1

This experiment considered set of 1000 (500 male, 500 female) "most biased" words (computed by finding cosine similarity to "he" and "she", with the most biased having the greatest difference in similarity). These words were separated into two clusters, and the homogoneity of these clusters were measured with the following results:

word2vec - 99.9%, hard-debias - 92.5%
GloVe representation - 100%, NN debiased - 85.6%

While the debiased versions of the embeddings do improve the homogoneity of our clusters to some degree, significant work is necessary for these debiasing methods to create embeddings devoid of gender bias.

3.2 Experiment 2

This experiment examined bias between female names and the arts and male names and the sciences. Once again, our debiasing methods from Experiment 1 decreased these biases, but significant bias still remained.

3.3 Experiment 3

This experiment examined 5,000 biased words, using 1,000 of these to train a binary classifier to classify these embeddings into male and female categories. This classifier was then tested on the remaining 4,000 words. Ideally, we'd like our classify to perform with near 50% accuracy (no better than chance at classifying into male/female), which would imply little to no gender bias in our embeddings. In reality, the classifier has 98-99% accuracy on the original embeddings and 88.88%, 96.53% for hard-debiased, NN debiased embeddings respectively. Again, we see the debiasing displaying some effectiveness, but still lacking.

3.4 Conclusion

These experiments demonstrate the great difficulty in removing all artifacts of gender bias from an embedding. While potentially disheartening, these experiments provide a good methodology for measuring varying degrees of bias in embeddings.

4 Google Autocomplete

We had a brief discussion on problematic autocompletes/search results such as:
"women should" autocompletes to "women should not vote".
"did the holo" autocompletes to "did the holocaust happen", and the first result was a white supremacist, holocaust-denying website. A question that naturally arises and may be beyond the ethical scope of this class is how to effectively make the tradeoff between freedom of expression and discriminatory/offensive behavior.

5 Looking ahead

- differences in classifiers on minority populations
- object recognition of self-driving cars
- medical applications
- differential privacy
- Book Club discussions will be beginning next week or the following