NLP Assignment 1
Xinyu Jiang
09/07/2020

2.

a. The paper “Understanding the Origins of Bias in Word Embeddings” reveals the stereotypical biases in word embedding algorithms. To solve the problem, the author traces the origin of words (that leads to bias) back to the original training sets, then identify and remove the subset which cause stereotypical biases. In addition, the author decomposes the identify/removal into two subproblem: examine how the pre-tubing training data affect the learned word embedding, and the how transform those words affect the bias.

b. The essay “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings” uncovers the gender stereotype in nature language processing task. To remove this gender bias, the author provides two properties: gender stereotype is first shown been captured in one direction of word embedding, gender neutral words are been detached from gender definition words. By analyzing those properties, the author gives a solution to reduce the gender bias: remove the “unnecessary” association while maintaining the desired association. And by using this method, the author empirically demonstrates the gender bias is significantly reduced.

c. The essay “Learning Gender-Neutral Word Embeddings” provides a new technique (Gender-Neutral Global Vectors) to solve gender stereotype. The method keeps gender information in certain dimension of word vectors while forcing other dimensions to be independent. In addition to that, Gender-Neutral Global Vectors (GN-GloVe) identifies gender-neutral words and learn word vectors at the same time. The author also claims that GN-GloVe does not remove any gender information from the word and reduce the possibility that the word is misclassified so that it can be applied in any languages.