Hands-on Tutorial: Quantifying and Reducing Gender Stereotypes in Word Embeddings

Tutorial at Conference on Fairness, Accountability, and Transparency 2018

Kai-Wei Chang, Tolga Bolukbasi, Venkatesh Saligrama
kwchang@cs.ucla.edu, tolgab@bu.edu, srv@bu.edu

Overview

Ensuring fairness in algorithmically-driven decision-making is important to avoid inadvertent cases of bias and perpetuation of harmful stereotypes. However, modern natural language processing techniques, which learn model parameters based on data, might rely on implicit biases presented in the data to make undesirable stereotypical associations. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. Recent results [1,2] show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because of their widespread use, as we describe, often tends to amplify these biases.

In this tutorial, we will provide attendees hands on experience writing small programs to display and quantify the gender stereotypes in word embedding. We will also show how to reduce such a gender stereotype from the word embedding. The first half of the tutorial will be mostly "lecturing" to provide necessary background, in which we will cover the basics of how word embeddings are learned and how they are used in the application domain. The second half of the tutorial will focus on hands-on exploration of biases present in the word-embedding and solutions to reduce the biases. An ipython notebook explaining the interface can be viewed at https://github.com/tolga-b/debiaswe/blob/master/tutorial_example1.ipynb; an elaborated version of this notebook will serve as the backbone for the "hands on" part of the tutorial, paired with exercises. The codebase used in the tutorial can be found at: https://github.com/tolga-b/debiaswe. We will lead the participants to explore the gender and the social biases in the word embedding.

Prerequisites

This tutorial assumes familiarity with *very* basic mathematical background such as linear algebra. Programming experience in python is also expected.