Lab Practicum:
Detecting Bias in Language Models

Prof. Ameet Soni and Prof. Krista Thomason
Swarthmore College

Introduction

This lab will introduce you to a popular tool in natural language processing known as *word embeddings*. Word embeddings are a mapping of a word (or phrase) to a vector of numbers. These types of methods aims to provide a representation of words that incorporate the context in which words get used. For example, while "motel" and "hotel" are distinct words, we understand them to be fairly similar in meaning; a computer, however, would not know this if given only the words. By placing these two words close to each other (i.e. assigning similar vector values) a word embedding model would provide a computer algorithm clues to the underlying meaning of these words. A typical application that uses word embeddings is a language translator (e.g., Google Translate).

To learn these embeddings, algorithms such as GloVe and word2vec use corpora, or large data sets of examples of human language, to train the model. For example, you will use models trained either on Twitter, Wikipedia, or generic web pages. In this lab, you will demonstrate the usefulness of these word embeddings. But you will also show that, by training on natural data sets, these models also incorporate cultural biases. Before continuing, please read the following paper for more background:

"Semantics derived automatically from language corpora contain human-like biases". Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind. *Science*, Vol. 356, No. 6334, 14.04.2017, p. 183-186.

Lab Instructions

Before beginning the assignment, you will first work through the provided code to understand the purpose of the two main programs. You should refer to the `README.md` file in your code directory for details on each of these steps.

1. Follow the **Setup** instructions in the `README.md` file. This will provide you with pretrained word embedding models; it is suggested that you download the Twitter, Wikipedia, and Common Crawl models. If given a choice for dimensionality of the word vectors, choose one of the smaller ones to make your experiments run faster.

2. Next, follow the instructions to run `findSimilarWords.py`. This program provides a simple proof of concept – do word embedding models do a good job of mapping similar words to a similar location in the embedding (i.e., are similar words close to each other)? Try several search terms on each of the data sets e.g., "dog", "baseball", "swarthmore", etc. and verify that the resulting words are similar to the search term.

3. Finally, run `weatTest.py` according to the provided instructions to perform the evaluation in the Caliskan, Bryson, and Narayan paper. This program performs a Word Embedding Association Test to detect biases in the word embedding models. You will need to provide four word lists - one each for the two concepts (e.g., European names and African names) and one each for the attribute you are testing for a bias (e.g. pleasant and unpleasant).

Once you have completed these steps, you are able to move on to the assignment. This is a good time to pause and to clear up any confusions or to restate what you have learned to this point.

Assignment Instructions

Turn in your written response individually. For Part 1, you should design and conduct your experiment as a pair but write your responses independently. You may share the results (i.e., the numerical findings) with each other. Please indicate who you worked with in your response. Part 2 should be done independently.

Part 1: Bias in word embeddings (1-1.5 pages)

Using the tools provided for testing associations with pairs of words, explore the implicit bias that can be encoded in natural language word embeddings. First, begin by replicating some of the pairings from the readings (e.g., European names vs. African names when paired with pleasant/unpleasant words). Then, respond to the following two prompts.

1. Rerun the pairings by varying the underlying training corpus used to learn the word embeddings. Discuss what impact, if any, this has on introducing biases into the trained word embedding model. Be sure to try each of the three datasets on both a task related to race/ethnicity and a task related to gender. Is there a noticeable change in each case?

2. Propose a new set of WEAT pairings to see if there are additional biases (e.g., religion, nationality, class). Generate a list of words for your target pair and use one of the existing attribute pairs (unpleasant/pleasant, family/career, male/female) or create your own. Describe your hypothesis (including how you created your lists of words) and then your findings. Discuss the implications of your results (e.g., what does this tell us about the algorithm, training corpus, and/or the experimental design itself).

Part 2: Ethical case study (2 pages)

Consider the following proposal you have been asked to review by a regional hospital:

> The emergency room has had difficulties with their triage system for prioritizing patients - it is both labor intensive (nurses spend less time caring for the patient and more time filling out paperwork and assessing the severity of a patients case) and error prone (incorrect prioritization can cause prolonged suffering and worsening health). The hospital would like to use natural language processing to suggest prioritization for a patient. The patient submits a description of their condition which can be supplemented by a medical assessment (e.g., paramedic or nurse) and the system outputs a triage score. The system undergoes significant training using notes/scores from triage nurses on prior patients.

Identify ethical issues you believe need to be addressed by the hospital and/or developers. You can use relevant readings from class to help you craft your response. It's better to choose one or two of the most important concerns and explain them in detail rather than try to cover too many things in this short assignment. For each concern, be sure to a) state your concern, b) justify the concern (i.e., why the issue is a possibility, including referencing our readings) and c) explain the potential ethical implications if the issue is ignored.