CIS399: The Science of Data Ethics **Lecture 13: Feb 28, 2019**
Instructor: Michael Kearns & Ani Nenkova **Scribes: Kennedy Ellison & Sophia Trump**

1 Review of Word Embeddings

A word embedding is a representation of the meaning of a word through a vector of real numbers which are derived from a task. An example of such a task is classification. Embeddings can be learned on any corpus that is large enough and online, such as news corpora or hollywood scripts.

There are two principle ways to learn word embeddings:

1. **Counting Method:** learns by looking at words that co-occur and then using PCA (principle component analysis) to project the high dimensional space into a lower dimension. However, dimensionality reduction on large matrices is computationally expensive.

2. **Predict the word:** learns by deleting a word from a text and seeing if it can predict the missing word based on the context of the surrounding words. For example, Word2Vec is a trained neural network trained on a task (predict a deleted word). The features in the task are the representations of surrounding words.

As discussed in Tuesday's lecture, the cosine similarity, which is the primary similarity function used in NLP, is a formula through which the similarity between word embeddings is quantified. More specifically,

$$cos(x,y) = \frac{x_1y_1 + x_2y_2 + ... + x_ny_n}{\sqrt{\sum_{i=1}^{m} x_i^2}\sqrt{\sum_{i=1}^{m} y_i^2}} \tag{1}$$

The concepts above are not proved by any theorem but are the results of repeated studies. Studies have found that directly learning the word representation by predicting a word performs better than the counting method. However, saying one "performs better" than another is not saying much since there is is only a 1.2% degradation in accuracy. In terms of geometric explanations of word embeddings, there are no guarantees, but they have been shown to work empirically. Word embeddings can be understood without geometry, and can instead be thought of as a dictionary containing 300 real numbers.

1.1 Science Paper

In the reading assigned for 26 February[1], word embeddings were applied in order to measure human biases. The study cited previous studies done using the Implicit Association Test (IAT) to measure ethically neutral biases (such as flower vs insect) and problematic biases (such as mental vs physical disease) in human participants. The researchers from the reading used Word2Vec embeddings trained on a news corpus. Using the Word2Vec embeddings, the researchers approximated the IAT by using the cosine similarity (1) to measure bias between word categories. In this context, a larger cosine similarity between two groups is interpreted as a shorter delay (and a stronger association) in human participants taking the IAT for those groups.

For example, the study measured bias using cosine similarity for the following groups (where each quadrant of the table represents a list of words):

[1] http://science.sciencemag.org/content/356/6334/183/tab-pdf

African American Names	European Names
Pleasant	Unpleasant

Using the cosine similarity for each group, the researchers found that African American names were more closely associated with unpleasant words. This demonstrates the Word2Vec system has a negative bias toward African American names. The ethical implications of this bias are discussed in the following section.

1.2 Additional Information about Word Embeddings

The representation of an entire text can be obtained by looking up all the words in the text and taking component by component averages. However, there is no guarantee that every word will have a vector representation. In the Science paper, if the word is not in Word2Vec, the researchers skipped the word. Another option is to inject a random vector for the OOV (out of vocabulary) word.

There is also another linguistic application of word embeddings. With parsing, word embeddings can be used to obtain word similarities, which can then be used to understand the meaning of a sentence. Once the semantics of a sentence is understood, the syntax can be determined. For example, in the sentence “I ate the sushi with chopsticks”, it is unclear what “with chopsticks” refers to without an understanding of semantics. If the sentence were “I ate the icecream with sprinkles”, then “with sprinkles” is describing the icecream, while in the earlier sentence “with chopsticks” describes the manner of eating.

2 Discussion

Keeping in mind the ethical issues raised by the Science paper, the following questions were discussed

1. **What did the researchers of the science paper mean by “our methods may yield an efficient way to explore previously unknown implicit associations”? In other words, can we use computerized bias to demonstrate bias in humans? Can we fix the bias in humans instead of fixing the computers?**

   As an answer to this question, consider a study that revealed implicit bias in letters of recommendation. In this study, it was found that women’s recommendation letters were seven times more likely to mention their personal lives, and more likely to discuss their work ethic than their accomplishments, compared to men’s recommendation letters.

2. **In the discussion of debiasing, is there bias we want to keep in our system because it’s good? Another way to think about this is, how do we determine what bias to keep and to change, especially in terms of stereotypes.**

   There was not a succinct answer, but the class discussed stereotypes such as African Americans being good at sports and women not being good in STEM. Studies of stereotype bias have shown that women perform worse on a computer science exam if a professor reminds the class prior to the exam that man typically outperform women. How might stereotypes in AI impact performance of different groups? It probably isn’t possible to go into the embeddings and fix biases against all groups, especially since people have different opinions of various biases. Further, no groups have an absolute advantage or disadvantage, and it will depend on the task to determine which groups need to be protected.

3. **Is the data biased? Or is the learning biased?**

   As an answer, consider googling homemaker vs programmer. In the case of homemaker, the results are all women. In the case of programmer, search results are more mixed. However, these results are not representative of the real world, in which 20% of people at Google are women. Thus, the data is biased.

3 After Spring Break: System Performance

After break we will talk about systems that use word embeddings to track all mentions of an item/individual in a text. For example, “Ani entered the room. She closed the door”. Such systems track that “Ani” is the antecedent of “She”. Even the best systems for these tasks note a disparity in performance between male and female subjects (69% accuracy for men and 52% accuracy for women). This means that if you run a program to disambiguate a pronoun, it is 64% accurate. The program chooses the person a male pronoun refers to correctly 69% of the time and only finds what female pronouns refer to 52% of the time. A closer analysis revealed a gender disparity in the training data (news corpora) for these systems (men were mentioned four times as often and named four times as often as women). These systems works much better for men over women. Thus, Google has developed a new corpus that has roughly the same male and female individuals in order to lessen the disparity between genders.

The above begs the following question: is it better to have a system with lower accuracy and less bias, or vice versa? And what if it is impossible to equalize performance across groups due to implicit differences in their feature sets? These will be discussed further after the break.