Debiasing Word Embeddings

First and last name
School

UB University at Buffalo

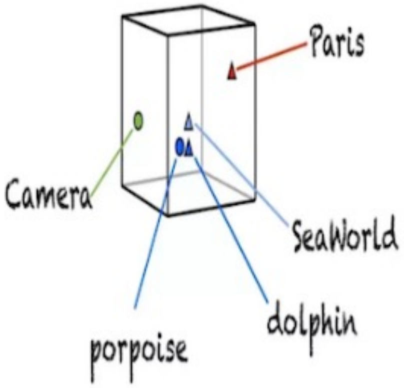
CLEMSON

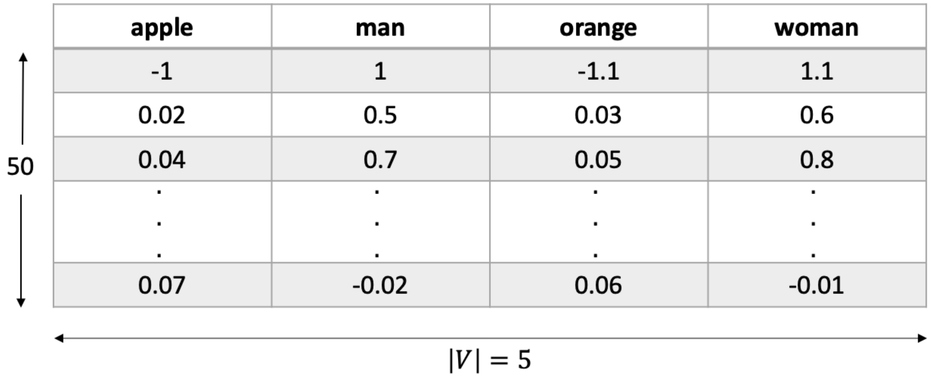
Outline

- What is word embeddings
- Pre-trained word embeddings
- Using word embeddings
- Properties of word embeddings
- Issue of bias in word embeddings
- Addressing bias in word embeddings

What is Word Embeddings?

- Word embedding represents words as vectors
- Words are represented as a d-dimensional vector
- They are vectors that carry meaning
- The word “apple” would be represented as a vector: [-1, 0.02, 0.04, …, 0.07]





apple	man	orange	woman
-1	1	-1.1	1.1
0.02	0.5	0.03	0.6
0.04	0.7	0.05	0.8
. . .	. . .	. . .	. . .
0.07	-0.02	0.06	-0.01

50

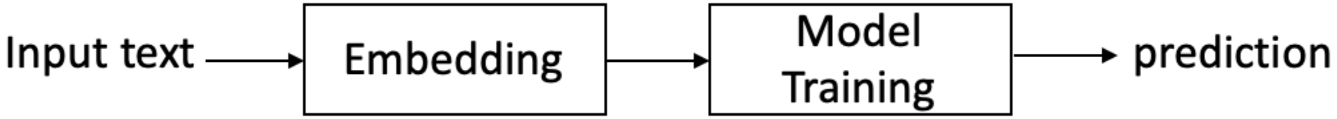
$|V| = 5$

Pre-trained Word Embeddings

- Word embedding is trained on word co-occurrence in a text corpora using neural networks
- The embedding matrix in the previous slide is learned after training
- Word embedding trained on billions of text is known as pre-trained word embedding
- Computationally intensive to train so it is better to use pre-trained embeddings.

Using word embeddings

- We used pre-trained word embeddings in lab 1 to train our cyber-bully detection model
- Pre-trained word embeddings are commonly used when we don’t have enough training data for a new task



Properties of Word Embeddings

- Words with similar semantic meaning will be close to each other in high dimensional space
- Can represent relationships between words
- For example, given the analogy, "*man is to king as woman is to x*" simple arithmetic of the embedding vectors of *man*, *king*, *woman*, and all words in our vocabulary finds that "*x = queen*". This is because:

$$\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{king} - \overrightarrow{queen}$$

- Similarly, *x = Japan* for "*Paris is to France as Tokyo is to x*"
- *X* is found by finding the most similar word to $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ using a similarity measure such as cosine similarity.

Issue of Bias in Word Embeddings

- Pre-trained word embeddings can propagate the bias contained in the dataset used in training the model that learned the embeddings
- That can have negative impact when such embedding is used in real world applications such as cyber-bully detection
- To understand this, the embedding system offensively answer *x* = homemaker for the analogy *“man is to computer programmer as woman is to x*”.
- It also outputs *x* = nurse for the analogy *“father is to a doctor as mother is to a x*”
- Word embeddings reflect gender stereotypes present in the society

Addressing Bias in Word Embeddings

- Bias in word embeddings is addressed in three steps
- Identify a gender direction in geometric space
  - Identify the embedding that captures bias. Gender direction can be found by taking a simple vector difference of gender pairs such as ($\overrightarrow{she}$ - $\overrightarrow{he}$) or ($\overrightarrow{woman}$ - $\overrightarrow{man}$)
- Neutralize
  - Removes values from the components of gender-neutral word vector. This ensures gender neutral words are zeros in the gender direction and projects the word to the non-bias direction.
  - Gender neutral words are words not specific to any gender such as shoe or flight attendant
- Equalize
  - Equalizes gender specific words to be equidistant (equal distance) to each other.
  - Ensures that a neutral word like “babysit” is equidistant to {*grandmother*, *grandfather*} and {*guy*, *gal*}
  - Gender specific words are words that are definitionally associated with gender such as *brother*, *sister*, etc.

Lab

See the lab 4 manual to access the notebook