MAN IS TO COMPUTER PROGRAMMER AS WOMAN IS TO HOMEMAKER? DEBIASING WORD EMBEDDINGS

introduction

Bias especially gender stereotypes in word embeddings:

e.g. Man - woman = programmer - homemaker

Pretrained embeddings: word2vec / 300dimensions / Google News

Quantify bias:

Compare a word vector to the vectors of a pair of gender-specific words. for example, nurse close to woman is not bias itself, because nurse close to humans, but closer than man suggest bias.

consider the distinction between gender specific words that are associated with a gender by definition (e.g. brother / sister), which close to a specfic gender is not bias, and the remaining gender neutral words (e.g. programmer / nurse).

We will use the gender specific words to learn a gender subspace (Surprisingly, there exists a low dimensional subspace in the embedding that captures much of the gender bias.) in the embedding. Removes the bias only from the gender neutral words while respecting gender specific words.

Gender biases in English

Implicit Association Tests have uncovered gender-word biases that people do not self-report and may not even be aware of. Biases are shown in morphology as well as while there are more words referring to males, there are many more words that sexualize females than males.

Biases in algorithms

A number of online systems have been shown to exhibit various biases.Schmidt identified the bias present in word embeddings and proposed debiasing by entirely removing multiple gender dimensions. His approach is entirely remove gender from embeddings. At the same time, the difficulty of evaluating embedding quality (as compared to supervised learning) parallels the difficulty of defining bias in an embedding.

word embeddings

Embeddings form: $w \epsilon R^d$,||w||=1. Assume F-M pair $P \epsilon R^d * R^d$, gender neutral word $N \epsilon W$, similiarity is cosine similarity:

$$cos(u,v) = \frac{u*v}{|u|*|v|}(1),$$

so similarity between embeddings is

$$cos(w_1, w_2) = w_1 * w_2(2)\,.$$

Geometry of Gender and Bias in Word Embeddings

understand biases present in embeddings(i.e which words more close to he/she etc.) and to which extent biases agree with human notion of stereotypes.

Occupational stereotypes

Ask the crowdworkers to evaluate whether an occupation is considered female-stereotypic, male-stereotypic, or neutral. Spearman r=.51(strongly correlated):

the geometric biases of embedding vectors is aligned with crowd judgment.

Alsoly, word2vec on Google News performs highly consistent with GloVe on web-crawl corpus when projecting embeddings on he/she axis.

Analogies exhibiting stereotypes

Analogy task:

- Standard: give 3 words (he, she, king) to predict the forth word
- In our task: give 2 words, a and b (he,she). a,b determine a seed direction ($\vec{a} - \vec{b}$). thus we have:

  $$S_{(a,b)}(x,y) = \begin{cases} \cos(\vec{a} - \vec{b}, \vec{x} - \vec{y}) & \text{if } ||\vec{x} - \vec{y}|| \le \delta \\ 0 \quad \text{else} \end{cases}$$

  in which δ is a threshold which restricts the distance between $\vec{x}$ and $\vec{y}$ in order to keep their semantically relavant. Outputing pairs with highest scores while refusing multiple analogies sharing the same $\vec{x}$ to reduce redundancy.

Evaluating analogies by crowd on:

1. Whether the pairing makes sense as ab analogy
2. Whether it reflects a gender stereotype

29 out of 150 analogies were picked as gender stereotype

Identifying the gender subspace

Find the gender-related word-pairs, in which both sets of definitional and stereotypical words are included. compute the difference ($\vec{x} - \vec{y}$). They are not exactly the same because of different biases, polysemy and randomness. Put 10 gender pairs together, conduct a PCA on them and then we can see one component is bigger than others significantly; to prove it is not a result of randomness, we conduct the same PCA on a set of random vectors and find the decrease far more gradual and uniform than the former one. then we denote the Principle component as vector $\vec{g}$. In general, the gender subspace could be higher dimensional and all of our analysis and algorithms (described below) work with general subspaces.

Biases

N is a set of gender neutral words. Define the direct bias: $\frac{1}{|N|}\sum_{w \epsilon N}|\cos(\vec{w},g)|^c$, where c is a hyper parameter that determines how strict we want to measure biases.

Debiasing algorithms

Notations: subspace B:$\{b_1, b_2, \ldots b_k\}$,projection of v on B: $v_B = \sum_{j=1}^{k}(v * b_j)b_j$

Step1: **Identifiying gender space**

to specify:

In practice, we have more than one set of word-pairs in order to de-biasing other kinds of stereotypes.

Input: we n set of pairs $D_1, D_2, \ldots D_n \in W$, and an integer parameter $k \geq 1$

Compute the co-variance matrix: each set's average: $\mu_i = \sum_{w \in D_i} \vec{w}/|D_i|$

Co-variance matrix of each set: $C_i = \sum_{w \in D_i}(\vec{w} - \mu_i)^T(\vec{w} - \mu_i)/|D_i|$,thus:

Co-variance matrix: $C = \sum_{i=1}^{n} C_i$,

so the subspace B is the first k rows of SVD(C).

Step2:

a) **Neutralize and Equalize** :

Neutralize ensures that gender neutral words are zero in the gender subspace.

Input: word to neutralize: $N \subset W$. for each word, re-embed $\vec{w}$ to :

$\vec{w} = (\vec{w} - \vec{w}_B)/||\vec{w} - \vec{w}_B||$, ensure zero in subspace and unit it.

Equalize enforces neutral word is equidistant to all words in each equality set(e.g. grandpa & grandma).

Input: family of equality sets $\varepsilon =\{E_1, E_2, \ldots E_m\}$. Separate the adjust embeddings into 2 components: $\vec{w}_B, \vec{w}_{\perp B}$. For $\vec{w}_{\perp B}$, we simply regard it as the projection of average vector on orthogonal subspace: for each $E_i \subset \varepsilon$, let $\mu = \sum_{w \in E} w/|E|$, then let it be zero in subspace B, $\vec{w}_{\perp B} = \nu = \mu - \mu_B$; for $\vec{w}_B$, we center and unit the vector: $\vec{w}_B = \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||}$, as for why we center the value,consider that: if we got the gender dimension and a pair of words:{male.female}. Happenly, both male and female is positive(towards female) in the dimension for some reasons, and if we do not center the embeddings, we are going to the same embedding on $\vec{w}_B$. To make the vector to be a unit one, we multiply the $\vec{w}_B$ with a parameter $\sqrt{1-||\nu||^2}$, thus we have:

$\vec{w} = \nu + \sqrt{1-||\nu||^2}\dfrac{\vec{w}_B-\mu_B}{||\vec{w}_B-\mu_B||}$, then output the subspace B and the new embeddings.

b) **Soften** : reduces the differences between these sets while maintaining as much similarity to the original embedding as possible, with a parameter that controls this trade-off.

Notations: let $W \in R^{d*|vocab|}$ denote matrix of all embeddings, N to be matrix of embeddings corresponding to gender neutral words. $T \in R^{d*d}$ is the de-biasing transformation matrix. To balance former embedings and reduce the stereotype, we have loss function:

$$\arg\min_{T} ||(TW)^2(TW) - W^TW||_F^2 + \lambda||(TN)^T(TB)||_F^2$$

the first component aims to minimize the difference and the latter one aims to minimize the biases inside the neutral words. λ is the parameter tuning these 2 parts.(In this paper, we set $\lambda = 0.2$); and we normalize the output as unit length, $\hat{w} = Tw/||Tw||_2$.

Determining gender neutral words

Since there are many fewer gender specific words, it is more efficient to enumerate the set of gender specific words S(S_0=218), whose selection is subjective. Starting with the 50,000 most frequent words, we selected only lower-case words and phrases consisting of fewer than 20 lower-case characters. After this filtering, 26,377 words remained. then training a SVM (10-fold cross validation) to classify the remaining corpus (size=3M) to select other 6449 words.

De-biasing results

for both hard-de-biased and soft-de-biased results, conduct the analogy generation task: he-she as the seed; ask crowds to evaluate whether these pairs show stereotypes. for top 150 pairs, 19% show biases for soft-de-biasing and 6% for the hard one. Soft is weaker than hard. Both of them are also tested in other benchmarks.