word embedding bias

论文：Joint Multiclass Debiasing of Word Embeddings

The Word Embedding Association Test (WEAT)

It considers two sets of attribute words (A, B), e.g., family and career related words, and two target sets (X, y), e.g., black and white names.

NULL hypothesis: the relative association of target sets' words to both attribute sets' words are equally strong. Thus, rejecting the null hypothesis asserts bias.

Test statistic s:

$$s(X, Y, A, B) = \Sigma_{\vec{x} \in X} h(\vec{x}, A, B) - \Sigma_{\vec{y} \in Y} h(\vec{y}, A, B)$$
$$where\ h(\vec{w}, A, B) = mean_{\vec{a} \in A} cos(\vec{w}, \vec{a}) - mean_{\vec{b} \in B} cos(\vec{w}, \vec{b})$$

h describes the relative association between a single target word $x \in X$ compared to the two attribute sets in a range [-2, 2].

$$d(X, Y, A, B) = \frac{mean_{\vec{x} \in X} h(\vec{x}, A, B) - mean_{\vec{y} \in Y} h(\vec{y}, A, B)}{std_{w \in X \cup Y} h(\vec{w}, A, B)}$$

Existing debiasing techniques

Method 1 (by Bolukbasi)

Post-processing step.

- Hard debiasing
  - employs a neutralize operation that removes, e.g., all non-gender related words N from a gender subspace by deducting from vectors their bias subspace projection.
  - Subsequently, an equalize operation positions opposing gender pair words (e.g., king, queen) to share the same angle with neutral words.
- Soft debiasing
  - enables more gradual bias removal by utilizing a tuning parameter λ: an embedding $W \in R^{d \times |vocab|}$ is being transformed by optimizing a transformation matrix $T \in R^{d \times d}$:

$$min_T \| (TW)^T (TW) - W^T W \|_F^2 + \lambda \| (TN)^T (TB) \|_F^2$$

Bias subspace B, computed via PCA on word pairs differences, such as he-she, man-woman.

讨论：后面的penalty的意思大概是要让与bias无关的词经过transform以后尽量垂直于经过transform以后的bias subspace。

Method 2 (by Manzini)

It expands and evaluates results not only on gender, but separately also on race and religion.

A bias subspace definition for non-binary class environment, which is formulated via PCA of mean shifted k-tuples (k is the number of subclasses) of the definitional words.

Others

- pre-processing techniques:
  - Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning gender-neutral word embeddings. In: Conference on Empirical Methods in Natural Language Process- ing. pp. 4847{4853 (2018)
- post-processing:
  - Dev, S., Phillips, J.: Attenuating bias in word vectors. In: International Conference on Artificial Intelligence and Statistics. pp. 879{887 (2019)
  - Font, J.E., Costa-jussa, M.R.: Equalizing gender bias in neural machine transla- tion with word embeddings techniques. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 147{154 (2019)

论文：Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Dataset

w2vNEWS: word2vec trained on Google News texts consisting of 3 million English words and terms into 300 dimension.

Unconver gender stereotypes

Issue: not just a trivial matter of couting pairs of words that occur together, since the counts are often misleading. E.g.: male nurse is several times more frequent than female nurse.

Use gender specific words to learn a gender subspace in the embedding, and the debiasing algorithm removes the bias only from the gender neutral words while respecting the definitions of these gender specific words.

- Two types of bias:
  - direct bias: an association between a gender neutral word and a clear gender pair;
  - indirect bias: association between gender neutral words that are clearly arising from gender;

Debiasing

- Reduce gender biases in the word embedding while preserving the useful properties of the embedding.
- Leverage the fact that there exists a low dimensional subspace in the embedding that empirically captures much of the gender bias.

Notes: 他们做了crowd experiments.

Understand the bias

and the extent to which these geometric biases aggree with human notion of gender stereotypes.

- evaluate whether the embedding has stereotypes on occupation words;
- evaluate whether the embedding produces analogies that are judged to reflect stereotypes by humans;

Project the occupation word onto the $she - he$ direction.

这里主要是用一个网上找的文集，用glove训练embedding出来，然后把这些词project到she-he的方向上，两个embedding的相关性如果高的话，就可以说明这种bias普遍存在。

第二种展示stereotype的方法是analogy： 一个seed pair，比如(she, he)，使用metric:

$$\mathrm{S}_{(a,b)}(x,y) = \begin{cases} \cos\left(\vec{a}-\vec{b}, \vec{x}-\vec{y}\right) & \text{if } \|\vec{x}-\vec{y}\| \leq \delta \\ 0 & \text{otherwise} \end{cases}$$

δ是一个阈值，embedding都是normalized，取 $\delta = 1$ 就代表 $\pi/3$ 。 输出得分最高的那一组x和y，并使用crowd experiment来决定是否存在bias， stereotype.

Gender subspace

By combining serveral directions, such as $she - he$, and $women - man$, we identify a gender direction $g \in R^d$ that largely captures gender in the embedding. This direction helps us to quantify direct and indirect biases in words and associations.

Gender Subspace: Take ten gender pair difference vectors and computed its principal components.

Debiasing algorithms

- debiasing algorithms are defined in terms of sets of words

Step 1: Identify gender subspace

to identify a direction(or a subspace) of the embedding that captures the bias.

Step 2:

Two options:

- Neutralize and Equalize:
  - ensures that gender neutral words are zero in the gender subspace;
  - equalizes sets of words outside the subspace and thereby enforces the property that any neutral word is equidistant to all words in each equality set.
- Soften: reduces the differences between these sets while maintaining as much similarity to the original embedding as possible;

To define the algorithms, it will be convenient to introduce some further notation. A subspace B is defined by k orthogonal unit vectors $B = \{b_1, \ldots, b_k\} \subset \mathbb{R}^d$. In the case $k = 1$, the subspace is simply a direction. We denote the projection of a vector v onto B by,

$$v_B = \sum_{j=1}^{k} (v \cdot b_j) b_j.$$

This also means that $v - v_B$ is the projection onto the orthogonal subspace.

Step 1: Identify gender subspace. Inputs: word sets W, defining sets $D_1, D_2, \ldots, D_n \subset W$ as well as embedding $\{\vec{w} \in \mathbb{R}^d\}_{w \in W}$ and integer parameter $k \geq 1$. Let

$$\mu_i := \sum_{w \in D_i} \vec{w}/|D_i|$$

be the means of the defining sets. Let the bias subspace B be the first k rows of SVD($\mathbf{C}$) where

$$\mathbf{C} := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i)/|D_i|.$$

Step 2a: Hard de-biasing (neutralize and equalize). Additional inputs: words to neutralize $N \subseteq W$, family of equality sets $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$. For each word $w \in N$, let $\vec{w}$ be re-embedded to

$$\vec{w} := (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\|.$$

For each set $E \in \mathcal{E}$, let

$$\begin{aligned} \mu &:= \sum_{w \in E} w/|E| \\ \nu &:= \mu - \mu_B \\ \text{For each } w \in E, \quad \vec{w} &:= \nu + \sqrt{1 - \|\nu\|^2} \frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|} \end{aligned}$$

Finally, output the subspace B and the new embedding $\{\vec{w} \in \mathbb{R}^d\}_{w \in W}$.

具体就是找bias subspace，然后找一堆中性词，debias。

Step 2b: Soft bias correction. Overloading the notation, we let $W \in \mathbb{R}^{d \times |vocab|}$ denote the matrix of all embedding vectors and N denote the matrix of the embedding vectors corresponding to gender neutral words. W and N are learned from some corpus and are inputs to the algorithm. The desired debiasing transformation $T \in \mathbb{R}^{d \times d}$ is a linear transformation that seeks to preserve pairwise inner products between all the word vectors while minimizing the projection of the gender neutral words onto the gender subspace. This can be formalized as the following optimization problem

$$\min_T \|(TW)^T(TW) - W^T W\|_F^2 + \lambda \|(TN)^T(TB)\|_F^2$$

where B is the gender subspace learned in Step 1 and λ is a tuning parameter that balances the objective of preserving the original embedding inner products with the goal of reducing gender bias. For λ large, T would remove the projection onto B from all the vectors in N, which corresponds exactly to Step 2a. In the experiment, we use $\lambda = 0.2$. The optimization problem is a semi-definite program and can be solved efficiently. The output embedding is normalized to have unit length, $\hat{W} = \{Tw/\|Tw\|_2, w \in W\}$.

Results evaluation

还是有crowd experiments。 使用analogy，比如针对she-he，然后生成这个seed pair的analogy，让别人判断analogy是否存在gender stereotype。

第二种indirect bias，他们是直接重复试验，直接说结果变好。

Discussion

其他种类的bias