Debiasing multilingual word embeddings : A case study for three Indian languages

G Vishal

June 5

Introduction

- Word embeddings are ubiquitous across many downstream NLP applications
- Bolukbasi et al. 2016 showed embeddings are 'blatantly sexist' thus introducing a bias in the applications built on top of them.
- For example there are some words in english which must be gender neutral(Computer Programmer, Homemaker).
- There also some gendered words like (man,woman),(he,she) which have gender connotations associated to them
- Bolukbasi et al. 2016 used the well-known analogy puzzle on Google-News embeddings to motivate their work. They showed that mining analogies produces some gender inappropriate analogies like man:woman :: computer-programmer : homemaker.
- To mitigate this problem they introduced different flavours of debaising algorithms that can quite successfully remove the biases in the word embeddings.Dev and Phillips 2019 showed using simple linear projections can be more effective in attenuating bias in word vectors than complex algorithms.

An open question: How well the above debiasing algorithms work for other non-English languages ?

- The semantics of gender words may vary from one language to another. Hassan and Alamgir 2013 point out for the sentence কেউ তার ছাতা ভুলে গেছে ।, 'তার' can refer to both he or she.
- While Bolukbasi et al. 2016 leverages the pronouns (e.g., *she/he*) to construct gendered directions this might not be possible for many languages.
- For many low-resource languages, translation to a high-resource language may distort the gender orientations native to that language.

The present work: We demonstrate through a series of experiments that the debiasing algorithms meant for English do not generalize well for other languages.

- We build sets of gendered words for three different Indian languages – Hindi, Bengali and Telugu. We also identify a set of gender neutral words for these languages which includes both profession words as well as a set of adjectives.
- We propose two different approaches – Language dependent and Language independent debiasing to remove bias from the multilingual MUSE word embeddings.
- We compare our approaches with the LP algorithm (Dev and Phillips 2019) using three different metrics. Our methods are generic and can be easily extended for any other language.

Related Works

- Zhou et al. 2019 and Burns et al. 2018 show that machine learning algorithms and their output show more bias than the data they are generated from. Word vector embeddings are used in machine learning towards applications which significantly affect people's lives. So it is paramount that efforts are made to identify and if possible to remove bias inherent in them.
- Bolukbasi et al. 2016 provides motivation for this problem. They define a gender direction using the difference between male- and female-definition word embeddings and demonstrate that women and men are associated with different professions.
- Dev and Phillips 2019 further advance this work by proposing simpler algorithms for debiasing based on linear projections. They also suggest metrics for evaluating the quality of the generated embeddings.

Dataset Description

Dataset Creation

- We create a set of gender neutral words from vocabulary $\mathcal{V}$ and denote it as $N_{all} \subset \mathcal{V}$. We also create gender defining pairs $D_{all} \subset \mathcal{V} \times \mathcal{V}$ for English and three Indian languages: Hindi, Bengali and Telugu.
- $N_{lang} = \{w \mid w \in N_{all} \ \ \& \ \ w \in lang\}$, $N_{all} = \bigcup_{lang} N_{lang}$ for $lang \in \{en, hi, be, te\}$
- $D_{all} = \{D_{all,1}, D_{all,2}, \ldots, D_{all,m}\}$ where $m = |D_{all}|$. Each $D_{all,i}$ represents a tuple of male-female words (δ_i^+, δ_i^-) where $\delta_i^+, \delta_i^- \in \mathcal{V}$ and let $d = dim(W)$.
- $D_{lang} = \{(\delta_i^+, \delta_i^-) \mid (\delta_i^+, \delta_i^-) \in D_{all} \ \ \& \ \ \delta_i^+, \delta_i^- \in lang\}$, $D_{all} = \bigcup_{lang} D_{lang}$

Dataset Description

Dataset Creation(Continued)

- N_{all} includes profession words and adjectives. For the three Indian languages it also includes the relevant transliterated-English profession
- Since Hindi, Bengali and Telugu have gendered nouns, we cross-check the transliterated profession words for gender neutrality.

Lang	$\|N_{prof}\|$	$\|N_{adj}\|$	$\|N_{tr}\|$	$\|N_{lang}\|$	$\|D_{lang}\|$
en	59	50	-	109	20
hi	28	44	14	86	20
be	29	43	15	87	21
te	18	54	18	90	15
All	134	191	47	**372**	**76**

Table: Datasets statistics: $|N_{prof}|$ - neutral profession words, $|N_{adj}|$ - neutral adjectives, $|N_{tr}|$ - neutral English transliterated words, $|N_{lang}|$ - total number of neutral words for language *lang*, $|D_{lang}|$ - total number of gender pairs for language *lang*.

Representative examples from our dataset

	Profession	Adjective	Transliterated	Gender defining pairs
en	professor, boss, singer	rich, powerful, bright	-	father-mother, king-queen, he-she
hi	माली[1], पत्रकार[2], न्यायाधीश[3]	कठिन[4], स्वच्छ[5], भारी[6]	कंडक्टर[7], नर्स[8], प्रोफेसर[9]	लड़का-लड़की[10], आदमी-महिला[11], पति-पत्नी[12]
be	মালী[13] সাংবাদিক[14] বিচারক[15]	নোনতা[16],সহজ[17], কম[18]	কন্ডাকটর[19],নার্স[20], অধ্যাপক[21]	পুরুষ-মহিলা[22], ছেলে-মেয়ে[23] , ভাইপো-ভাইঝি [24]
te	న్యాయమూర్తి[25],న్యాయవాది[26] నిర్వాహకుడు[27]	శక్తివంతమైన[28], శాంతిగా[29], నిశితం[30]	కమిషనర్[31],కన్సల్టెంట్ [32], ఎడిటర్[33]	అంకుల్-అత్త[34], సోదరుడు-సోదరి[35], దేవుడు-దేవత[36]

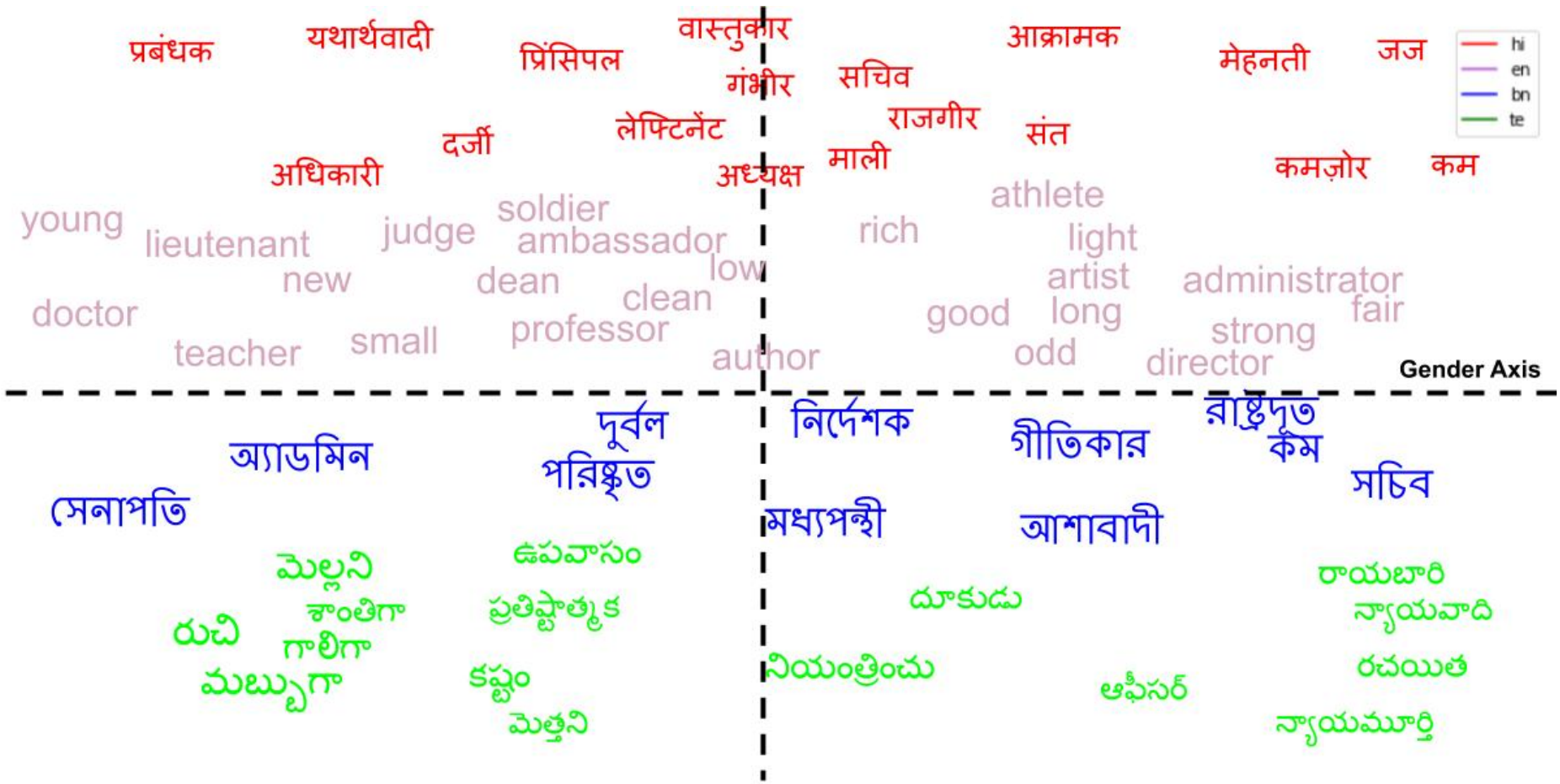
[1]gardner [2]reporter [3]judge [4]tough [5]clean [6]heavy [7]conducter [8]nurse [9]professor [10]boy-girl [11]man-woman [12]husband-wife [13]gardner [14]journalist [15]judge [16]salty [17]easy [18]low [19]conducter [20]nurse [21]professor [22]male-female [23]son-daughter [24]nephew-niece [25]judge [26]lawyer [27]administrator [28]powerful [29]peaceful [30]careful [31]commissioner [32]consultant [33]editor [34]uncle-aunt [35]brother-sister [36]god-goddess

Embeddings

- We use MUSE[1] (Multilingual Unsupervised and Supervised Embeddings) model for multilingual word embeddings (dimension $d = 300$ per word). We use pre-trained fastText embeddings (Bojanowski et al. 2016) aligned in a common vector space using pre-trained transformation matrix (Smith et al. n.d.).
- Motivation for choosing embeddings aligned in a common space comes from that fact that, if we have L languages some of which are high-resource (e.g., English) and some low-resource (e.g., Telugu) then debiasing Telugu wrt English gender space may cause performance penalties as the concept of gender in different languages, maybe semantically very different (see Table 2).

[1] https://github.com/facebookresearch/MUSE

Pictorial Representation of Bias in MUSE



Projection of selected neutral words along mean-gender direction.

Definition of Bias in Embeddings

Bias Definition

Overall "proximity" of neutral words with respect to the notion of gender in that vector space. Intuitively, we expect the vector difference of a gender defining pair $D_i = \{man, woman\}$, i.e., $\overrightarrow{man} - \overrightarrow{woman}$ to capture the gender direction in the embedding space.

Construction of Bias Spaces

Let D be any set such that $D \subseteq D_{all}$. We can represent D as $D = \{D_1, D_2, \ldots, D_n\}$. For each $D_i = \{\delta_i^+, \delta_i^-\}$, *difference vector* can be defined as $\vec{\delta_i} = \vec{\delta_i^+} - \vec{\delta_i^-}$. We can stack these difference vectors to form a matrix $Q = [\vec{\delta_1}\vec{\delta_2} \ldots \vec{\delta_n}]^T$. Now, gender subspace **B** can be obtained from Q by finding the *basis vectors* in two different ways as follows.

Construction of Bias Spaces

- **PCA**: We compute the top-k principal components of the vector differences which account for the maximum amount of variation.
- **PPA**: Projection pursuit (Friedman and Tukey 1974), attempts to find interesting projections in the data according to maximizing or minimizing a projection index (variance in the PCA framework).We use kurtosis-based projection pursuit (the fourth statistical moment) proposed by Hou and Wentzell 2011 as the projection index and minimize the kurtosis using a quasi power learning algorithm. (Orthonormalize k most significant projections using Gram-Schmidt process)

Defining **B** from k orthonormal unit projections

$\mathbf{B} = span\{\vec{b_1}, \vec{b_2}, \ldots, \vec{b_k}\}$ where $\vec{b_i} \in \mathbb{R}^d$ for integer parameter $k > 1$.

PCA and PPA top components

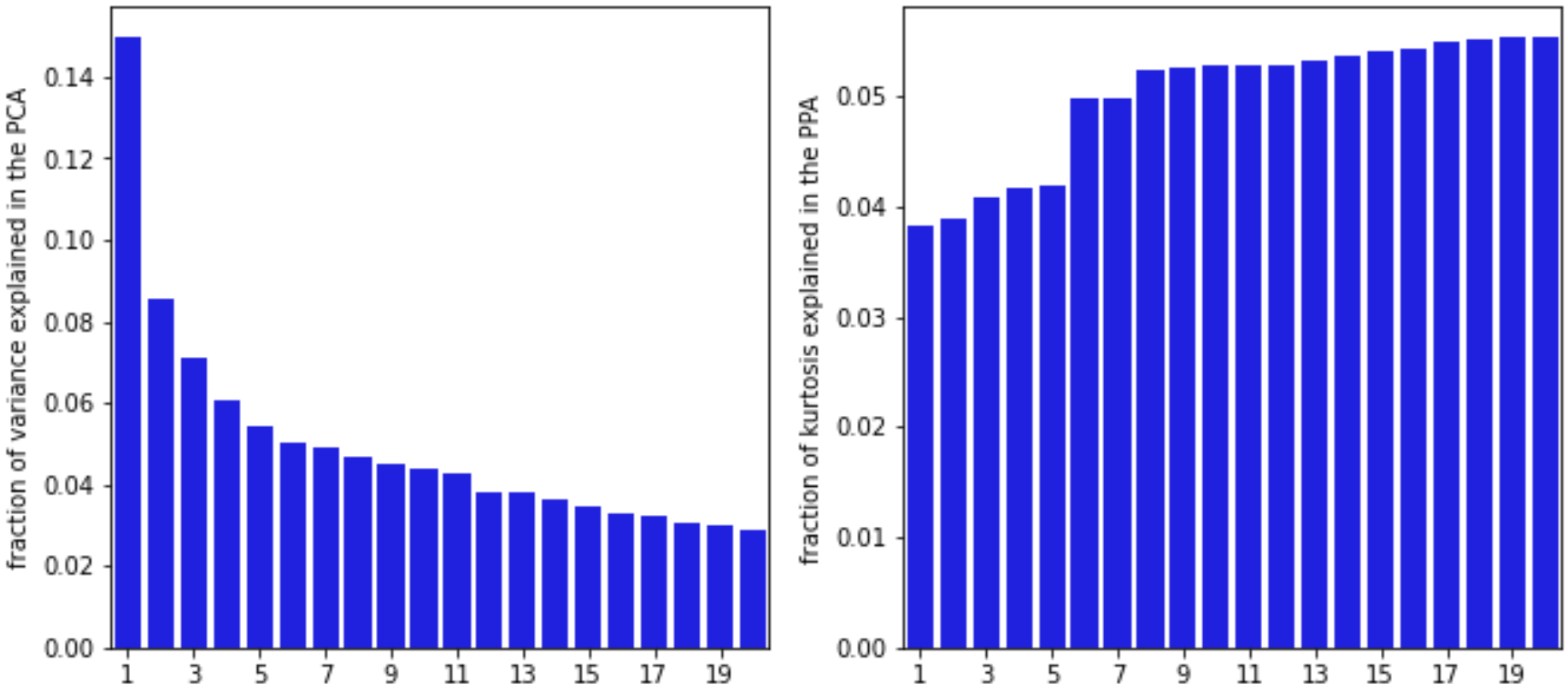


Figure: Fraction of projection index (Left : variance explained for PCA and Right : kurtosis explained for PPA)

Debiasing

Decomposing a word

For word vector: $\vec{w} \in W$, we denote its projection onto the gender subspace **B** as $\vec{w_B} = \sum_{i=1}^{k} \langle \vec{w}, \vec{b_j} \rangle \vec{b_j}$ Thus, a word vector $\vec{w}$ can be decomposed as $\vec{w} = \vec{w_B} + \vec{w_\perp}$, where $\vec{w_\perp}$ is the projection onto the orthogonal space

Debiasing Algorithm

We use a variant of the linear projection algorithm (Dev and Phillips 2019) for debiasing embeddings along the gender subspace. For a given word vector $\vec{w} \in \mathbb{R}^d$, the debiased embedding such that $\vec{w'} \in \mathbb{R}^d$ is $\vec{w'} = \vec{w} - \vec{w_B}$ such that $dim(W') = d - dim(\mathbf{B}) = d - k$.
This ensures that the updated embedding W' has no component along bias subspace.

Proposed Algorithms : LDD

Language-dependent debiasing (LDD)

LDD algorithm focuses on debiasing a specific language $\{en, hi, be, te\}$ based on the language-specific gender subspace $\mathbf{B}_{lang}$ obtained using language specific gender defining set D_{lang}. This algorithm is effective when only a specific language embedding is used for building downstream applications.

- Let $\mathbf{B}_{lang} = span\{\vec{b}_{lang,1}, \vec{b}_{lang,2}, \ldots, \vec{b}_{lang,k}\}$
- We hypothesize that for two languages $l_1, l_2 \in \{en, hi, be, te\}$ s.t. $l_1 \neq l_2$, their corresponding gender subspace $\mathbf{B}_{l_1}$ and $\mathbf{B}_{l_2}$ are inherently different from each other.(maybe due to different gender semantics)
- To verify our hypothesis, we use linear projection algorithm for debiasing neutral words $N_{l_2} \subset N_{all}$ using gender direction $\vec{b}_{l_1,1}$ obtained from top-most PCA component of matrix Q_{l_1}

Cross Language Performance

Scoring

- $S_{l_1,l_2} = \frac{1}{|N_{l_2}|} \sum_{\substack{i=1 \\ \forall w_i \in N_{l_2}}}^{|N_{l_2}|} \frac{||\langle \vec{w_i'}, \vec{b}_{l_2,1} \rangle| - |\langle \vec{w_i}, \vec{b}_{l_2,1} \rangle||}{|\langle \vec{w_i}, \vec{b}_{l_2,1} \rangle|}$ where $\vec{w'}$ corresponds to the debiased embedding of word w wrt $\vec{b}_{l_1,1}$ having original word embedding $\vec{w} \in \mathbb{R}^d$. Note that $\langle \vec{w_i'}, \vec{b}_{l_1,1} \rangle = 0$ for $w_i \in N_{l_1} \implies S_{l_1,l_1} = 1$
- The results clearly supports our hypothesis that gender subspace $\mathbf{B}_{en}$, $\mathbf{B}_{hi}$, $\mathbf{B}_{te}$ and $\mathbf{B}_{be}$ are significantly different. Thus multilingual debiasing cannot be accomplished using a single gender subspace.

Lang	N_{en}	N_{hi}	N_{be}	N_{te}
b_{en}	1.0	0.143	0.137	0.038
b_{hi}	0.105	1.0	0.083	0.023
b_{be}	0.345	0.126	1.0	0.075
b_{te}	0.049	0.054	0.157	1.0

Table: S_{l_1,l_2} for debiasing N_{l_2} neutral words using gender direction $\vec{b}_{l_1,1}$

Proposed Algorithms : LID

Language-independent debiasing (LID)

LID algorithm concentrates on debiasing multilingual embedding for all languages ensuring the common space constraint. We construct a gender subspace by combining the gender defining set of all languages D_{all}. The gender subspace $\mathbf{B}_{all}$ thus obtained will have contributions from all language's gender pairs.

- We next verify our intuition that $\mathbf{B}_{all}$ is indeed representative of all the languages under consideration by labelling the directions $\vec{b}_{all,i} \in \mathbf{B}_{all}$ with their language orientation l_i $(1 \leq i \leq |D_{all}|)$ based on their similarity with language specific mean bias direction $\bar{b}_{lang}$

Language Orientation

$$\mathbf{B}_{lang} = span\{\vec{b}_{lang,1}, \ldots, \vec{b}_{lang,k}\} \quad k \leq |D_{lang}|$$

$$\bar{b}_{lang} = \frac{1}{k}\sum_{i=1}^{k} \vec{b}_{lang,i} \quad \forall\ lang$$

$$\mathbf{B}_{all} = span\{\vec{b}_{all,1}, \vec{b}_{all,2}, \ldots, \vec{b}_{all,k'}\} \quad k' = |D_{all}|$$

$$l_i = argmax_{lang}\langle \vec{b}_{all,i}, \bar{b}_{lang} \rangle \quad \forall\ \vec{b}_{all,i} \in \mathbf{B}_{all}$$

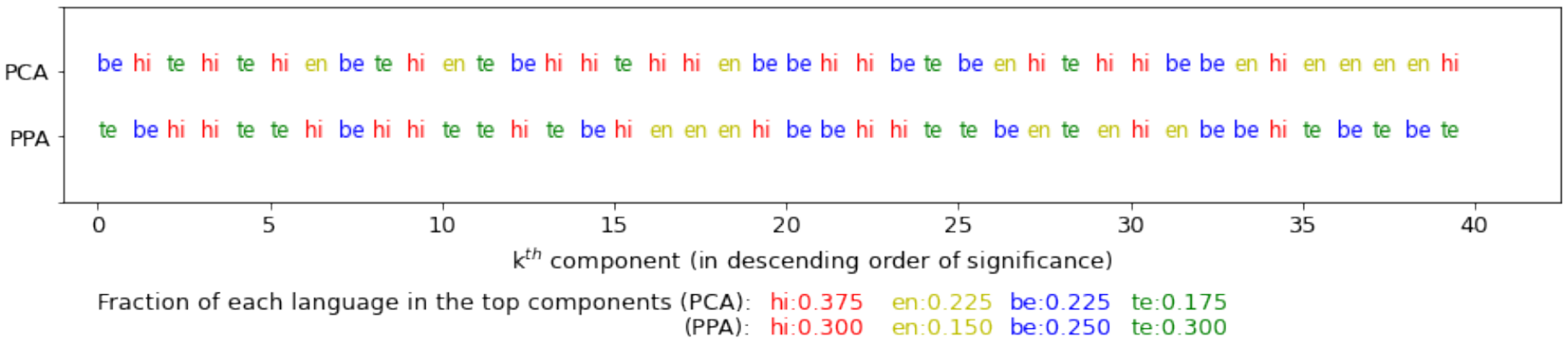


Figure: Language orientation l_i of top PCA and PPA components.

Equal Representation

- Different languages contribute in different proportions to the bias subspace which might lead to preference of certain language over others in terms of capturing the gender semantics

LID : Equal Representation (EQR)

EQR algorithm constructs bias subspace $\mathbf{B}_{equal\_rep} \subset \mathbf{B}_{all}$ $(k' = |D_{all}|)$ ensuring equal representation of each language, treating gender from a multilingual perspective.

- For L languages under consideration and gender subspace having k'' basis vectors s.t. $k'' \leq k' = |\mathbf{B}_{all}|$, we choose $\frac{k''}{L}$ top basis vectors of each language.
- This ensures that the multilingual perspective of gender is not over represented by any particular language. Thus, this approach restricts k'' to integral multiples of L.

Evaluation Metrics

- Neutral words $N \subseteq N_{all}$ $(N \in \{N_{en}, N_{hi}, N_{be}, N_{te}, N_{all}\})$
- Test gender pairs $D \subseteq D^{test}_{all}$ $(D \in \{D^{test}_{en}, D^{test}_{hi}, D^{test}_{be}, D^{test}_{te}, D^{test}_{all}\})$
- mean *male* vector : $\vec{o}_{male} = \frac{1}{|D|} \sum_{i=1}^{|D|} \vec{\delta_i^+}$
- mean *female* vector : $\vec{o}_{female} = \frac{1}{|D|} \sum_{i=1}^{|D|} \vec{\delta_i^-}$
- $N = N_{lang} \implies D = D^{test}_{lang}$ or $N = N_{all} \implies D = D^{test}_{all}$

Embedding coherence test (ECT) proposed by Dev and Phillips 2019

The metric quantifies the aggregate similarity of the neutral words wrt *male* and *female* gender orientations.

- $u_{male} = \{1 - \langle \vec{w}, \vec{o}_{male} \rangle \mid \forall w \in N\}$ & $u_{female} = \{1 - \langle \vec{w}, \vec{o}_{female} \rangle \mid \forall w \in N\}$.
- Spearman's rank correlation (in $[-1, 1]$, larger is better) between the two lists u_{male} and u_{female} is the defined as the ECT score.

Evaluation metrics

Overlap test (Ov)*

We propose a new metric based on the overlap of the 2ϵ length window centred at *male* and *female* similarity scores with neutral words N. The extent of bias removed is defined as the overlap between these windows, assuming overlap always occurs.

- *male*-similarity score : $\alpha_w = \langle \vec{w}, \vec{o}_{male} \rangle \ \forall \ w \in N$
- *female*-similarity score : $\beta_w = \langle \vec{w}, \vec{o}_{female} \rangle \ \forall \ w \in N$
- Construct two intervals $[\alpha_w - \epsilon, \alpha_w + \epsilon]$ and $[\beta_w - \epsilon, \beta_w + \epsilon]$; $\epsilon > 0$
- Given ϵ is chosen in such a way that overlap always occurs, overlap is $|\beta_w - \alpha_w - 2\epsilon|$. Hence it follows $\epsilon \geq max(|\beta_w - \alpha_w|)/2, \forall w \in N$
- The overlap score is defined as the average of overlaps for all the words, i.e., $Ov = \frac{1}{|N|} \sum_{w \in N} |\beta_w - \alpha_w - 2\epsilon|$.

Evaluation metrics

Histogram distance (JSD)*

For two random variables X_{male}, X_{female} denoting the similarity with mean *male* and *female* vector of a randomly chosen neutral word $w \in N$, JSD captures how close the similarity distributions $P(X_{male}), P(X_{female})$ are.

- $u_{male} = \{1 - \langle \vec{w}, \vec{o}_{male} \rangle \mid \forall w \in N\}$ & $u_{female} = \{1 - \langle \vec{w}, \vec{o}_{female} \rangle \mid \forall w \in N\}$.
- Generate two histograms corresponding to u_{male}, u_{female}.
- Jensen-Shannon distance (JSD) between these two histograms is a measure of the extent of bias (smaller is better).

Experiment

Baseline

Dev and Phillips 2019 proposed Linear Projection (LP) algorithm for debiasing which removes the bias component w_B for all $w \in \mathcal{V}$. This is equivalent to LDD for $k = 1$. Their algorithm is simple and more powerful than Bolukbasi et al. 2016. It will make the vectors close-by but not exactly equal. For simpler word pairs with fewer word senses, like (*he*/*she*) and (*him*/*her*), we can expect them to be almost at identical positions in the vector space after debiasing while for word pair like (*man*/*woman*) vectors should be close-by as the word *man* is used in many extra senses compared to the word woman; e.g., humankind, and in expressions like "oh man!".

- Our is work is a natural extension of this as, we construct vector spaces for $k > 1$ which is more suitable in the multilingual settings.

Experiment Objective

- Compare the PCA and PPA methods for obtaining top-k components of the gender subspace
- Analyse how these models scale with increasing the dimension of the gender subspace (e.g., $k = 1$, 4 or 8)
- Compare performances of variants of linear projection algorithm (LDD and LID) arising due to the choice of difference of the gender subspace $\{\mathbf{B}_{lang}, \mathbf{B}_{all}, \mathbf{B}_{equal\_rep}\}$.

Train test split of gender pairs:

- *Train gender pairs* used for construction of the gender subspace and *Test gender pairs* used for metric evaluation. $D^{train}_{lang} \subset D_{lang}$ such that $|D^{train}_{lang}| = 10$ & $D^{test}_{lang} \subset D_{lang}$ such that $D^{train}_{lang} \cap D^{test}_{lang} = \phi$
- Evaluating the embeddings on the same directions as used for constructing vector spaces would typically lead to better values of metrics (train-test leakage).

Experimental Setup

LDD Setup

Language-dependent debiasing (LDD) approach constructs language gender subspace $\mathbf{B}_{lang}$ using train gender pairs D_{lang}^{train} for a particular language $lang \in \{en, hi, be, te\}$. k basis-vectors of $\mathbf{B}_{lang}$ obtained using PCA/PPA are used to debias neutral set $N \subset N_{all}$ and its performance is evaluated using test gender pairs.

LID Setup

Language-independent debiasing (LID) approach constructs gender subspace $\mathbf{B}_{all}$ and $\mathbf{B}_{equal\_rep}$ using training gender pairs drawn from all languages D_{all}^{train}. k basis-vectors of $\mathbf{B}_{all}$ obtained using PCA/PPA are used to debias neutral set N_{all} and its performance is evaluated using test gender pairs D_{all}^{test}. We test for three values of $k = 1, 4, 8$[a].

[a] Note that $k = 1$ is not possible for the equal representation scheme.

Results : Debiasing language specific neutral words

	N_{en}			N_{hi}			N_{be}			N_{te}		
Algorithm	ECT	Ov	JSD	ECT	Ov	JSD	ECT	Ov	JSD	ECT	Ov	JSD
Orig	0.933	0.937	0.0146	0.824	0.872	0.022	0.725	0.862	0.0241	0.485	0.856	0.0271
LP_{PCA}	0.927	0.921	0.0158	**0.861**	0.891	0.0194	0.829	0.893	0.0207	0.511	0.846	0.0268
LP_{PPA}	0.9	0.927	0.0155	0.815	0.873	0.0221	0.730	0.861	0.0243	0.462	0.856	0.0272
LDD^{4}_{PCA}	0.923	**0.930**	0.0149	0.857	0.900	**0.0184**	0.826	0.908	0.0195	0.561	0.853	0.025
LDD^{8}_{PCA}	0.890	0.919	0.0163	0.835	0.892	0.0194	0.844	0.912	0.0190	0.574	0.863	0.0246
LDD^{4}_{PPA}	0.875	0.914	0.0158	0.845	**0.908**	0.0192	**0.864**	0.911	**0.0188**	0.465	0.861	0.0268
LDD^{8}_{PPA}	**0.928**	0.914	**0.0131**	0.828	0.883	0.0193	0.855	**0.912**	0.0190	**0.621**	**0.863**	**0.0234**
$LID^{1}_{PCA}(all)$	0.930	0.921	0.0143	0.862	0.885	0.0205	0.718	0.856	0.0242	0.470	0.856	0.0273
$LID^{4}_{PCA}(all)$	0.934	0.922	**0.0138**	**0.900**	0.918	0.0170	0.798	0.870	0.0217	0.516	0.849	0.0267
$LID^{8}_{PCA}(all)$	0.900	0.910	0.0148	0.894	**0.921**	**0.0162**	**0.814**	**0.895**	**0.0206**	0.538	0.847	0.0263
$LID^{1}_{PPA}(all)$	0.935	**0.937**	0.0146	0.833	0.868	0.0216	0.726	0.864	0.0241	0.485	0.857	0.0271
$LID^{4}_{PPA}(all)$	0.939	0.936	0.0143	0.831	0.876	0.0210	0.735	0.859	0.0239	0.488	0.857	0.0274
$LID^{8}_{PPA}(all)$	**0.944**	0.935	0.0140	0.865	0.894	0.0183	0.712	0.857	0.0237	0.498	**0.859**	0.0282
$LID^{4}_{PCA}(eqr)$	0.927	0.923	0.0139	0.896	0.921	0.0165	0.770	0.873	0.0221	**0.545**	0.854	**0.0262**
$LID^{8}_{PCA}(eqr)$	0.910	0.916	0.0145	0.899	**0.921**	**0.0162**	0.766	0.883	0.0217	0.535	0.849	0.0264
$LID^{4}_{PPA}(eqr)$	0.933	0.931	0.0142	0.836	0.860	0.0213	0.723	0.870	0.0235	0.489	**0.859**	0.0275
$LID^{8}_{PPA}(eqr)$	0.923	0.933	0.0147	0.823	0.876	0.0209	0.733	0.873	0.0236	0.520	0.854	0.0277

Blue: best LDD results, **black**: best LID results.

Results : Debiasing language specific neutral words

- LDD and LID algorithms are consistently better than the baseline LP and original embedding in debiasing N_{lang} for all the languages.
- The only two exceptions are the ECT (LDD on N_{en}) and Ov (LDD & LID on N_{en}) metrics where the *Orig* is slightly better.
- PPA works better for LDD algorithms while PCA is better choice for LID algorithms. The extra smoothing in PPA is perhaps not necessary when a language-independent setup is required.
- Within any particular language, LDD performs better than LID
- Base ECT values for Telugu are quite low which can be attributed to the fact that it is the least-resource language among others.
- Bengali seems to have the highest difference between $\mathbf{B}_{be}$ and $\mathbf{B}_{equal\_rep}$ performance for its comparatively more gender neutrality.

Result : Debiasing the entire set of neutral words

Algorithm	ECT	Overlap	JSD
Orig	0.744	0.892	0.0209
$\text{LDD}^{4*}_{PCA}(en)$	0.730	0.893	0.0209
$\text{LDD}^{1*}_{PCA}(hi)$	0.680	0.886	0.0222
$\text{LDD}^{8*}_{PCA}(be)$	0.764	0.892	0.0212
$\text{LDD}^{4*}_{PCA}(te)$	0.755	0.894	0.0207
$\text{LDD}^{1*}_{PPA}(en)$	0.724	0.892	0.0207
$\text{LDD}^{4*}_{PPA}(hi)$	0.789	0.903	0.0189
$\text{LDD}^{4*}_{PPA}(be)$	0.765	0.896	0.0206
$\text{LDD}^{8*}_{PPA}(te)$	0.753	0.894	0.0209
$\text{LID}^{1}_{PCA}(all)$	0.819	0.901	0.0177
$\text{LID}^{4}_{PCA}(all)$	**0.845**	0.910	0.0167
$\text{LID}^{8}_{PCA}(all)$	0.840	**0.912**	**0.0166**
$\text{LID}^{1}_{PPA}(all)$	0.733	0.891	0.0212
$\text{LID}^{4}_{PPA}(all)$	0.751	0.892	0.0209
$\text{LID}^{8}_{PPA}(all)$	0.748	0.890	0.0214
$\text{LID}^{4}_{PCA}(eqr)$	0.833	0.906	0.0172
$\text{LID}^{8}_{PCA}(eqr)$	0.830	0.910	0.0171
$\text{LID}^{4}_{PPA}(eqr)$	0.686	0.887	0.0214
$\text{LID}^{8}_{PPA}(eqr)$	0.710	0.886	0.0219

Results : Debiasing the entire set of neutral words

- None of the LDD algorithms (for best k-configuration) can outperform the LID algorithms on N_{all}.
- LID algorithms perform well for both the debiasing scenario. They strike a trade-off for common space with a marginal dip in performance for single-language debiasing whereas completely outperform LDD for multilingual debiasing.
- Equal Representation (EQR) algorithms have comparable performance to LID for all the three metrics. This shows that in a truly multilingual setting LID is a much better choice for the purpose of debiasing.

Discussion

Common-Space alignment Distortion

When we debias each $N_{lang} \in \{N_{en}, N_{hi}, N_{be}, N_{te}\}$ wrt corresponding $\mathbf{B}_{lang}$, the resulting embeddings may distort the common space alignment. We quantify this common space alignment distortion using semantic similarity between English words (high resource language) and their corresponding translated words in Indian languages ($\{hi, be, te\}$).

- We create a set $T \subset N_{en} \times N_{all}$; $T = \{(w_1, w_2) \mid w_1 \in N_{en}, w_2 \in N_{l_2}\}$ where w_1 is translated from English to w_2 in $l_2 \in \{hi, be, te\}$
- N_{lang} are debiased either wrt $\mathbf{B}_{lang}$ (LDD) or wrt $\mathbf{B}_{all}$ or $\mathbf{B}_{equal\_rep}$ (LID).
- We compute the common space distortion $\mathcal{D}$ as

$$\mathcal{D} = \frac{1}{|T|} \sum_{(w_1,w_2)\in T} \frac{\langle \vec{w_1}, \vec{w_2} \rangle - \langle \vec{w_1'}, \vec{w_2'} \rangle}{|\langle \vec{w_1}, \vec{w_2} \rangle|}$$

Discussion : Common-Space Distortion

- We expect $w_1, w_2 \in T$ to be semantically similar since w_2 is translation of w_1). Hence $\langle \vec{w_1}, \vec{w_2} \rangle > \langle \vec{w_1'}, \vec{w_2'} \rangle$ indicates that after debiasing similarity between w_1 and w_2 decreases.
- Thus $\mathcal{D} \leq 0$ suggests common space alignment is preserved while $\mathcal{D} > 0$ suggest common space distortion (smaller is better).
- High distortion value for LDD indicates that resulting multilingual embeddings are not in a common space, thus losing the very purpose of multilingual embeddings.
- $\mathbf{B}_{all}$ & $\mathbf{B}_{equal\_rep}$ attempts to keep inter-language semantics intact (since $\mathcal{D} \leq 0$) due to common gender subspace which uniformly debiases across all the languages with a minimal trade-off on individual language performance.

Algorithm	LDD	LID (*all*)	LID (*eqr*)
PCA^1	0.103	-0.080	-
PCA^4	0.158	-0.114	-0.181
PCA^8	0.377	-0.564	-0.568

Discussion : Semantics distortion

- It is natural to expect that LP which uses gender direction might fail to capture the linguistic concept of gender completely.
- We expect to get better representation of bias subspace as we increase k. But the fact that resulting debiased embedding space has a dimension of $d - k$, higher value of k may degrade the language semantics captured by the embeddings.
- Optimal value of k is required to balance the trade-off. This value needs to be set experimentally and in most cases will depend on the downstream application for which the embeddings are built.

Conclusion

- We proposed different approaches to debias multilingual word embeddings. Our methods not only work well for the Indic languages but also on English.
- We perform both language-dependent as well as language-independent debiasing and show their comparative advantages.
- Our methods consistently outperforms the LP algorithm which is the most competitive baseline known to us.

Scope

We believe that our work will open up many new opportunities for downstream multilingual NLP applications that are dependent on the underlying word embeddings.

The End