Word Embeddings to Quantify Cross-Lingual Semantic Variation and Bias - Experimental Protocol

Sandhini Agarwal
sandhini@stanford.edu

Jacqueline Ennis
jqennis

Olivia Li
oli2

1 Introduction and Hypothesis

For our final project, we hope to leverage monolingual word embeddings to gain insight about the ways biases, stereotypes, and concepts differ across languages. We hypothesize that biases surrounding words denoting gender, ethnicity, and historic events vary across languages and that this variation can be quantified using word embeddings, as demonstrated by Garg et al. (Garg et al., 2018).

Furthermore, we posit that this difference in concepts across monolingual embeddings will result in interesting, nuanced biases across multilingual embeddings. This might be because of differences in quality of data across languages or embedding creation methods. We then hope to probe into the nature of these differences and assess if multilingual embeddings adhere more to the norms of one language over the other.

2 Data and Preprocessing

2.1 Social Datasets

We examine sources of social data to use as validation against the calculated embedding biases. We look at employment data from the most recent U.S. Census (cen, a) and the Census of India (cen, b) to map the percentage of workers in a given occupation in each gender and ethnic group. These percentages will be used in comparison with the calculated bias in the word embeddings. Additionally, we aim to use the Sex Stereotype Index from the Williams et al. study on sex stereotypes in order to validate our findings for adjective stereotypes (1977).(Williams and Best, 1977)

2.2 English-Hindi Translation

Data provided by the Census of India is in English as are the adjectives dataset we found, so we use Google Translate to translate the occupation categories and adjectives to Hindi. A random subset of these translations were then validated by a native speaker of Hindi. We found that the quality of translation was usually high since the terms being translated were simply occupation words or single word adjectives.

3 Metrics

3.1 Bias Measurement

We plan to use a number of evaluation metrics to measure possible relationships between the embeddings. We calculate an embedding bias metric in order to represent the strength of association between neutral words and a given group. Common ways to measure this bias are as follows:

Cosine Distance:

$$\cos(U, V) = u \cdot v$$

Negative Norm Distance:

$$(u, v) = -||u - v||^2$$

In this baseline, we define this metric as the average cosine distance between the group vector, computed as the average of the vectors for each word in the given gender/ethnicity group, and each vector in the neutral word list, which can be occupations or adjectives. We also aim to use ordinary least-squares regressions to measure the associations themselves between the two words, from which we will report r2 and the p-value.

3.2 Shared Event Analysis

In the future, we plan to analyze variance in the relative perception of a shared historic event across two languages. We would measure this using sentiment analysis scores and noun-adjective

distance. The latter metric will take embeddings of the shared historic event and embeddings of a wide range of positive and negatively charged adjectives, and find the distance between them according to the formula outlined above. The former method would use the SentProp algorithm for sentiment analysis of unlabeled documents to assign a sentiment score to the relevant historic terms. This would act as a proxy for quantifying the difference in perspectives surrounding a given event, depending on the history of the language.

4 Models

4.1 Word Embeddings

We will be using FastText and MUSE embeddings for this project.

FastText is a library for text classification and representation (embeddings) created by Facebook, which uses state-of-art machine learning and NLP techniques to classify words and learn their word vector representations. To do learning on an immense dataset, it employs a hierarchical softmax that is able to organize categories of words in a tree, rather than a list. This structure also takes advantage of categorical frequencies - the depth for more frequent categories is smaller than infrequent ones - which also speeds up computation. FastTexts embeddings use n-grams to hold information about context, and overall show the same kind of results as state-of-art deep learning without the compute time. It provides pretrained embeddings in many different languages. These are trained on Common Crawl and Wikipedia using continuous bag of words with position-weights. (Grave et al., 2018)(Mikolov et al., 2017)

For our baseline, weve utilized pretrained FastText embeddings in English and Hindi. We download word vectors from Facebook Open Sources fastText (Mikolov et al., 2017; Grave et al., 2018) and focus on extracting words from two categories: identity words, that represent gender, or neutral words with no direct semantic link to the identity words (e.g. inherently gendered words like policeman are excluded). A list of representative identity words, including gender words like man, woman, is taken from Garg et al.

MUSE is also a library of word embeddings, partially built on FastText. It provides bilingual embeddings created with two methods: supervised or unsupervised. In the former, the embeddings are created using parallel corpora in two languages; in the latter, monolingual embeddings are aligned in vector space to create these bilingual embeddings. The latter method proves to be very advantageous for languages which dont share alphabets or have low-resource data. Thus, we will be training bilingual hindi-english embeddings using unsupervised muse embeddings methods. (Conneau et al., 2017)

5 General reasoning

We propose a two-step investigation for our hypothesis. First, we examine embedding bias between a Western language (English) and an Eastern language (Hindi) for different gender groups, and validate these biases against measurable trends in occupational data. We also explore bias in relation to a wide range of adjectives that have various positive, negative, or other connotations. We validate these findings using survey data from previous sociological and psychological studies.

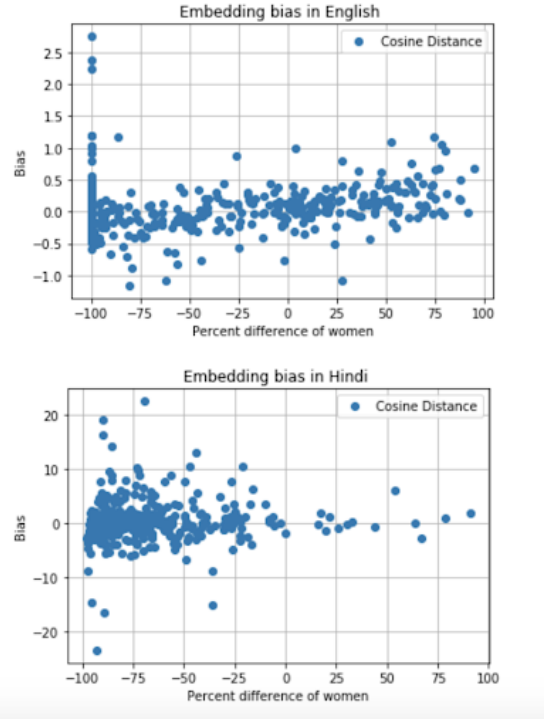
Second, we launch a comparative analysis between word embeddings about a shared historic events using sentiment analysis and noun-adjective distance. Given the two example languages, these will be events involving English speaking nations such as Britain and India, such as colonization, Indian independence, slavery, and so on. We hope to shed light on how these concepts differ across the two languages.

Lastly, we aim to run our experiments above using multilingual MUSE embeddings. We hope to compare how the performance of MUSE differs with respect to the monolingual embeddings. This will help us ascertain whether MUSE has the tendency to lean inherit the biases of one language over another. In order to assess this, we aim to look at the distance MUSE embeddings have from identity group vectors to neutral words and assess if the average distances for the English monolingual embeddings or for Hindi monolingual embeddings are significantly less. Thus, we aim to take the average of (MUSE-ENG) and (MUSE-HIN) and study if there is a substantial difference between these - i.e. if avg(MUSE-ENG) $<$ avg(MUSE-HIN) or vice versa. We posit that this will help us ascertain if these bilingual embeddings lean more towards the norms of one language.

6 Progress to Date

6.1 Gender Occupation

For our baseline, we looked at quantifying gender bias for Fasttext embeddings of occupational roles in the U.S. and in India. We took the cosine-distance of these embeddings and group words related to gender, and compared this bias measure to the gender ratio in the occupation. The X-axis shows the difference between the percentage of women to men, and the Y-axis shows the cosine distance between our neutral words (the occupational data) and group words (related to gender). On both axes, a more positive value corresponds to a stronger association with women.



For the English embeddings, our results show a slightly positive correlation between gender bias and the percentage of women employed in that occupation, with a y-intercept approximately near the origin. This seems to match the analysis made in the Garg paper. For Hindi, the relationship and bias and gender ratio is less apparent. The data shows most of these occupations as having much less women than men, while in the American dataset there was a much more even spread. This makes any sort of correlation harder to see because there is just not as much evidence for occupations with higher percentages of women in the Indian Census data as there are in the US Census data.

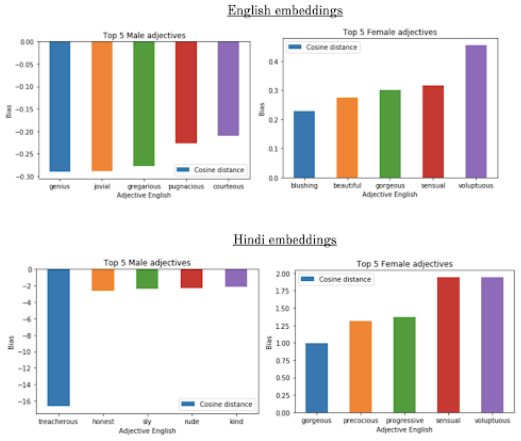
What is apparent, however, is the difference in absolute values between biases in these two languages. Notice that the scales of the y-axis in each plot arent at all the same. The biases in English embeddings are mostly less than 0.5 in either direction, while the Hindi biases are much larger, with many absolute values between 0 and 10. What this could indicate is a generally larger tendency to be biased in toward either gender for occupation words in Hindi.

Additionally, we aim to further preprocess the data to find common occupations across India and America and then further see how the embedding bias and survey difference measure across these comparable datasets. For example, the bias in Hindi for electrical engineer is -0.7227 and 0.07689 percent of electrical engineers in India are female. On the other hand, in English, the bias for electrical engineer is -0.3681 and women represent 9.4 percent of electrical engineers in America. While the bias in Hindi is almost twice that of English, the ground truth number of electrical engineers in India is significantly lower than that of engineers in America.

6.2 Gender Adjectives

In addition to looking at the relationship between gender ratio with embedding bias, we also looked at how embeddings in English and Hindi might reveal gender stereotypes. Similar to the gender occupation method, we find the cosine distance between the embeddings for adjectives and gender group words. This gives us a way to quantify and compare the association between adjectives and gender.

The below plots show the adjectives with the 5 smallest and 5 largest cosine distances. The smallest bias mean that the adjectives are least associated with female (so most associated with male), and a larger bias means more strongly associated with female. We will be validating these findings with survey data across how people associate adjectives with men and women.



Unsurprisingly, the spread of most- and least-associated stereotypes with women seem to follow a common theme. In both sets of embeddings, adjectives that are more strongly associated with women are (for the most part) appearance-related - like voluptuous and gorgeous, while those strongly associated with men are more related to personality or intellect - like genius or honest.

7 Next Steps

For the remainder of the project, we will expand our analysis to look at shared historic events. This will involve looking at word embeddings representing these events in their respective languages. For example, we will compare the word embeddings and associated word lists of an event like colonization to quantify the semantic difference between the two languages perception of the event using the bias metrics mentioned above. We will also carry out sentiment analysis using the SentProp algorithm to assign a sentiment score to the relevant historic terms. We will then carry out an analysis of the studies above with Muse embeddings and see how the compare to the monolingual embeddings.

Our raw project code and data can be found at the following link. https://github.com/jackieennis/cs224u-word-embeddings-semantic-bias.

References

a. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity.

b. Occupational classification of main workers other than cultivators and agricultural labourers by sex 2011(india/state/uts-district level)(total, sc/st).

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

John E Williams and Deborah L Best. 1977. Sex stereotypes and trait favorability on the adjective check list. *Educational and Psychological Measurement*, 37(1):101–110.