Gender bias in dictionary-derived word embeddings

Edward Lee
edlee1@stanford.edu

Abstract

Existing methods for generating word embeddings offer great results and coverage of non-standard or colloquial terms but can often suffer from bias. Such bias may arise from the corpora used like Wikipedia and Google News. We suspect that dictionaries may be less likely to embed biases, and thus methods that are able to generate embeddings from dictionaries may side-step these biases. To do this, we evaluate the bias found in 2 dictionary-based methods (CPAE and dict2vec) and 2 common methods (word2vec and GLoVe) for generating word embeddings. We find that dictionary-based methods do seem promising in containing less bias, but are unable to completely avoid the issue. We then make a few attempts at identifying why bias may appear in these embeddings, and find evidence for two hypotheses that may contribute.

1 Introduction

Most word embeddings are derived from large text corpora of language like Google News [1] and Wikipedia [2]. While these methods offer great results and coverage of non-standard or colloquial terms, they can often suffer from bias. For example, it has been found for word2vec that man is analogous to computer programmer in the way woman is related to homemaker. These relations likely do exist in colloquial use, as computer programmers are currently male-associated and homemakers currently female-associated. However, it could be harmful for such relations to exist in the word embeddings we train, as that may affect the results of important downstream tasks that use such embeddings (e.g. resume filtering, bail judgements, etc.). Much work has been done to both understand and remove these biases from existing embeddings [3, 4], to varying degrees of success.

We examine one possible method that might side-step these issues by taking a completely different approach to learning word embeddings. Rather than use a corpora of colloquial language (which likely contains many instances of biases in action), we look into the possibility of generating word embeddings based solely on a dictionary, which may be less likely to directly embed biases. Existing methods may not work as well on a dictionary due to the stricter and more structured language used, and as such, we primarily look into CPAE [5] as a method for generating embeddings from dictionaries and thus possibly reducing bias.

2 Related Work

This paper combines concepts from two primary fields: methods for constructing word embeddings (more specifically, those leveraging structured data like dictionaries), and methods of addressing bias in such word embeddings.

CS230: Deep Learning, Fall 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2.1 Dictionary-based Word Embeddings

Only two methods for constructing word embeddings from dictionaries were discovered during our research: Consistency Penalized AutoEncoder (CPAE) [5] and dict2vec [6]. CPAE attempts to derive the embedding of a word to allow for re-construction of the definition of that word after being passed through an LSTM, with an additional "Consistency Penalized" aspect aimed at dealing with the recursive nature of words being used in both the input and output of the auto-encoder. On the other hand, dict2vec attempts to augment existing embeddings with knowledge derived from dictionaries by identifying additional relations between words and the words present in their definition.

2.2 Addressing bias in Word Embeddings

There have been quite a few attempts at identifying and addressing bias in word embeddings. [4], for example, attempts to address gender bias by modifying the embeddings themselves. They attempt to identify a gender subspace within the embedding space, and move embeddings along this subspace based on whether they appear to be gender-related or not. [3] attempts to identify the documents in a corpora that most affect gender bias, and evaluates the effect of removing these documents on both performance and bias metrics.

A commonly used metric for evaluating bias is known as WEAT [7], which is a hypothesis test attempting to check whether two groups of words (e.g. math- and arts-related) are equally similar to another two groups of words (e.g. male- and female-related).

3 Methods [1]

3.1 Models Evaluated

As this is not a conventional machine learning project, we don't directly use any datasets. Instead, we evaluate bias on the two discovered dictionary-based approaches and two widely-used corpus-based approaches to training embeddings: CPAE trained with $\lambda = 4$ [5], dict2vec over Wikipedia [6], word2vec trained on GoogleNews [1], and GloVE trained on Wikipedia [2]. We train CPAE ourselves, and use their provided configurations for reproducible results.

3.2 Evaluating Bias

As stated earlier, WEAT (Word Embedding Association Test) is the most common method for evaluating bias of word embeddings [7]. A single WEAT takes 2 sets of target words (e.g. ["programmer", "engineer", ...] and ["nurse", "teacher", ...]), and 2 sets of attribute words (e.g. ["man", "male", ...] and ["woman", "female", ...]). It then runs a permutation test, with the null hypothesis being there is no difference in the relative similarity between the sets of target words and sets of attribute words. The resulting effect size and p-value can be used to measure how biased word embeddings are for that given test and how likely this result was, respectively.

More formally, let X, Y are the sets of target words where $|X| = |Y|$, and A, B are the sets of attribute words. Let $\cos(a, b)$ be the cosine similarity between vectors a and b. Define $s(w, A, B)$ as the association of w with some attribute, and $s(X, Y, A, B)$ as the difference in association between the two target words with the attribute. They can be calculated as:

$$s(w, A, B) = \text{mean}_{a\in A}\cos(w, a) - \text{mean}_{b\in B}\cos(w, b)$$

$$s(X, Y, A, B) = \sum_{x\in X} s(x, A, B) - \sum_{y\in Y} s(y, A, B)$$

The effect size of this difference can be expressed as

$$\frac{\text{mean}_{x\in X}s(x, A, B) - \text{mean}_{y\in Y}s(y, A, B)}{\text{std-dev}_{w\in X\cup Y}s(w, A, B)}$$

And the p-value of the test can be expressed as

$$\text{Pr}_i\left[s(X_i, Y_i, A, B) > s(X, Y, A, B)\right]$$

for any i where $X_i \cup Y_i$ is some arbitrary partition of $X \cup Y$ into two sets of equal size.

[1] All code can be found at `https://github.com/ed-w-lee/cpae-bias`

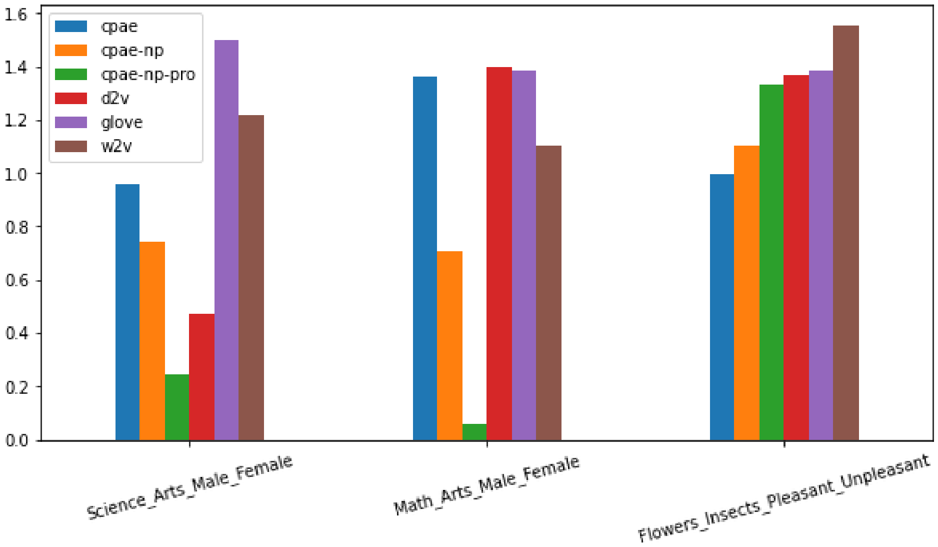

Figure 1: Bias effect size evaluated on the three WEAT tests: WEATs 8, 7, and 1 in order from left to right. Smaller is better. `cpae-np` represents CPAE trained without proper nouns and `cpae-np-pro` represents CPAE trained without proper nouns and balanced pronouns.

4 Initial Results

4.1 Evaluation

We evaluate the 4 different models on 3 different WEATs, a subset of the set of tests used in [7]. The first two are used to measure gender bias: WEAT 7 and 8, which compare *math* and *arts* to *male* and *female*, and *science* and *arts* to *male* and *female*, respectively. The third, WEAT 1, is to check bias in a separate topic, comparing *flowers* and *insects* with *pleasant* and *unpleasant*. Note that some words were missing across the various embeddings, so I slightly modified these tests to account for the missing words and reduce noise when evaluating on each embedding.

As seen in figure 1, there seems to be bias in all four models (ignore `cpae-np` and `cpae-np-pro` for now) for all three of the tests, but the dictionary-based models (CPAE and dict2vec) seem to do slightly better with CPAE being quite good in two out of the three tests. The main outlier is WEAT 7 (`Math_Arts_Male_Female`), where word2vec does better than the rest.

4.2 Discussion

Based on these test results, it seems that bias still exists in the dictionary-based models, albeit typically on-par or better than the other models. Our next goal is to understand why bias might exist in the dictionary-based models, especially CPAE as it's the only model that trains solely off of a dictionary.

Two hypotheses we explore in this paper relate to the dictionary CPAE was trained on [8] (specifically regarding gender bias): (1) definitions of proper nouns include male scientists and mathematicians like "Galileo", which may more strongly associate math and science with male-ness than female-ness, and (2) the pronouns used may be imbalanced – for example, when "he" is used to refer to arbitrary individuals rather than more gender-neutral pronouns like "he or she" and "they".

The first hypothesis seemed promising since 176 different proper nouns (in this case, words where the first letter is capitalized) contain the word "he" and only 37 contain the word "she". An example of one such a proper noun that could impact gender bias is "Kennelly", whose definition (excerpted) is "United States electrical engineer noted for his work on the theory of alternating currents". These words could thus more closely tie science-related words like "currents" and "engineer" to male-related roles and concepts.

The second hypothesis also seems somewhat promising since 234 words contain "he" but not "she" while only 36 contain "she" but not "he". And a cursory examination of some of these 280 words shows that "he" is commonly used to refer to arbitrary individuals while "she" is not. One such example is "anesthesiologist", whose definition is "a specialist who administers an anesthetic to a patient before he is treated". Such words may, again, unnecessarily associate concepts or words to one gender over another.

5 Exploring Sources of Gender Bias

Based on these observations, we attempt to re-learn and re-evaluate the embeddings while attempting to address these hypotheses. We do this primarily by modifying the corpus in an attempt to balance the imbalances that we observed.

5.1 Modifying the Training Set

Based on the two hypotheses from section 4.2, we make two changes to the WordNet dictionary before training new embeddings.

The first modification is that of removing proper nouns. A bit more formally, given a dictionary with a set of words W mapped to a set of definitions D, we remove all words w from W where w begins with an upper-case letter. Of note is that we don't modify the set of definitions D, as part of CPAE relies on re-constructing a definition from the embedding and thus modifying D may have unintended side effects on that re-construction process.

The second modification *does* modify D, in an attempt to balance pronoun use. Since the dictionary uses "he" in some definitions and "he or she" in others, we attempt to replace instances of "he" with "he or she" — provided it does not seem like the instance of "he" was a part a more gender-neutral reference like "he or she" and "she or he". We filter out these instances by checking if "she" exists within 3 words of the use of "he". We then repeat this with a second pronoun pair ("his", "her").

We then generate and evaluate two new embeddings using dictionaries with the aforementioned modifications, one with just the 1st modification and another with both. The results for these two embeddings represent `cpae-np` and `cpae-np-pro` in figure 1, respectively.

5.2 Results

As can be seen in figure 1, the two changes do seem to decrease gender bias in the word embeddings. Removing proper nouns improves both WEAT 7 and 8, and balancing pronouns seems to improve them even more, bringing the effect size of bias closer to 0 (especially in WEAT 7's case).

However, these "fixes" seem to make performance on WEAT 1 significantly worse. Since WEAT 1 doesn't test gender bias but rather an unrelated form of bias, it's a bit surprising and worrying that attempting to address bias in one area would drastically affect bias in another. It's hard to say why, and more investigation would need to go into how and why CPAE exhibits this unrelated bias and why changing pronouns may affect this. One possible hypothesis is that this is all due to variation, and re-training may drastically affect the results, but the added time and compute make examining these factors infeasible within the scope of this project.

5.3 Discussion

The two hypotheses seem like promising avenues for understanding the sources of bias in this particular dictionary.

One thing to note is that this section is primarily for exploring sources of bias, and not actually addressing them. The modifications used are likely too coarse as a simple solution to addressing bias. In the first modification, removing every single capitalized word may be too over-reaching since doing so could severely reduce coverage of the generated embeddings. At the same time, the modification may not do enough if some proper nouns aren't capitalized in the source dictionary. And in the second modification, balancing every usage of "he" or "him" is likely too broad. Some words that are male-related (e.g. "gentleman") may lose that association through this process.

An actual solution for addressing gender bias likely requires a more fine-grained approach, either through modifying the existing dataset in a more nuanced manner or finding datasets that have don't include certain classes of issues altogether. For example, one may look into using dictionaries that fully use gender-neutral language when describing arbitrary individuals.

6 Conclusion

Based on our results, we conclude that biases are more minor in dictionary-based embeddings like CPAE than in colloquial-use embeddings like word2vec. However, these biases, gendered or otherwise, still do exist.

Two factors identified in the dataset used that likely contribute to gender bias include proper nouns, which may help perpetuate historical gender roles, and imbalanced pronoun usage. Attempting to address these and other factors may have unintended consequences on other forms of bias if not done in a nuanced manner.

Thus, while dictionary-based embeddings do seem promising as a way to generate high-quality word embeddings while limiting the amount of bias present, more care needs to be placed onto choice of the dataset and how the dataset is cleaned and/or modified to address these biases.

6.1 Future Work

There are many different avenues of investigation for improving upon and examining the bias present in dictionary-based embeddings.

One straightforward avenue is that of identifying other possible sources of gender bias in dictionary-based embeddings. For example, dictionaries often include example sentences using a word which may be inadvertently written in gendered language. One could also attempt to address gender bias using a different, more gender-neutral dictionary, as stated previously.

Another avenue may focus on exploring the unintended consequences of addressing a single form of bias. For example, one could dive deeper into understanding why balancing pronoun usage in CPAE caused `Flowers_Insects_Pleasant_Unpleasant` bias to rise in figure 1.

Last, one may attempt to fix some of the issues with the basic mitigations used in this paper. For example, rather than just remove proper nouns entirely from the dataset, one could attempt to limit the effect proper nouns have on other words by learning embeddings for proper nouns only after the embeddings for other words have been learned and frozen.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[3] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611, 2018.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[5] Tom Bosc and Pascal Vincent. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[6] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2017.

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[8] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.