Diachronic Evaluation of Gender Bias in English Using Word Embeddings

Ben Hinthorne bhinthorne@hmc.edu

Abstract

Natural Language Processing allows for quantitative analysis of gender bias in language. Specifically, *word embeddings*, a technique to represent words from a corpora as vectors, allow for in depth analysis of word relationships. Through computations with word embeddings, we can understand which words in the English language are more closely related to others based on the frequency and context in which they appear. In this paper, we build on work done by Bolukbasi et. al. (2016) to understand the similarity of occupational words, like nurse and architect, to gendered pronouns. By quantifying the similarity of occupations and gendered pronouns through word embeddings, we can understand the extent to which occupational gender bias appears in word embeddings trained on English corpora. In this paper, we are specifically focused on the way in which the relationships of occupational and gendered pronouns change over time. We introduce a metric called *bias score* to quantify the gender bias of an occupation word and investigate how the bias score of occupation words change over time. Through this investigation, while we do not find a significant pattern across all occupational words, we highlight trends of individual occupation-pronoun pairs and discuss their significance.

1 Introduction

Word embeddings have become a popular technique in Natural Langauge Processing to understand word relationships. Scholars like Bolukbasi et al. (2016), Greyson et. al. (2017), and Gonen and Goldberg (2019) have specifically used word embeddings to understand gender bias in the English language. Furthermore, studies like those conducted by Hamilton et. al., highlight the ways in which word embeddings can be used to understand how English semantics have changed over time. While word embeddings have proven useful in our understanding of gender bias and semantic change in the English language, little work has been done to understand the overlap between these two areas of study. In this paper, we set out to use word embeddings to explore diachronic change in gender bias in the english language. We focus specifically on gender bias in occupation words, like architecture or nurse, and the gendered pronouns 'he' and 'she'. Occupational gender bias can be understood as when the an occupational word, which should not be gendered, is more closely related to one gendered pronoun than the other. We quantify this similarity by calculating the cosine similarity of the word embedding representing the occupation and those representing each gendered pronoun. For example, Bolukbasi et. al. (2016) found that in embeddings trained on Google News articles, 'homemaker' had a much higher cosine similarity to 'she' than it did to 'he', therefore quantifying the gender bias of the use of 'homemaker' in Google News articles. While such bias has been called out in embeddings trained from corpora of a single time period, we investigate how gender bias has changed over time in english corpora. By training word embeddings on english corpora which has been partitioned diachronically by decade, we can better understand the diachronic change in occupational gender bias in English corpora. Specifically, we set out to answer the question: how does gender bias in occupational words in English corpora change over time?

2 Related Work

2.1 Word Embeddings and Semantic Change

Systematic semantic change of language is not a new phenomenon, having been studied by scholars such as Breal (1897) and Stuartevant (1917) in the late 20th and early 19th century. Breal (1897)

and Stuartevant (1917) early work suggested two theories of semantic change: the Law of Differentiation, which states that near-synonyms tend to differentiate in meaning over time, and the Law of Parallel change, which states that that related words tend to undergo parallel changes in meaning. More recent advances in the field of Natural Language Processing, the use of word embeddings in particular, and growing availability of historical corpora have been a recent driving force for the study of semantic change of language. A recent study by Xu and Kemp (2015) used the Google Million Corpus containing English words from 1899 to 1999 to create meaning vectors of words grouped by decades to evaluate the early theories of semantic change proposed by Breal. By analyzing semantic change on large scale corpora and using large scale statistical analysis, Xu and Kemp (2015) concluded that the Law of Parallel change applies more broadly than the Law of Differentiation.

Additionally, in their paper "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change", Hamilton et. al. (2016) were able to use diachronic word embeddings to hypothesize their own laws of semantic change. By measuring the pairwise cosine similarity over a time series of word embeddings, Hamilton et. al. (2016) formulated the Law of Conformity and the Law of Innovation. The Law of Conformity states that the rate of semantic change scales with an inverse power-law of word frequency, and the Law of Innovation states that independent of frequency, more polysemous words have higher rates of semantic change. Thus, recent advancements in NLP and word embeddings have allowed researchers to create a further understanding of not just semantic change itself but also the rate of semantic changes, which, as noted by Kulkrani et. al. (2014), is increasing with the growth of micro-blogging and large scale sharing of language on sites like Twitter. In this study, we focus the use of word embeddings to understand semantic change on one specific subsections of English semantics: occupation and gender.

2.2 Word Embeddings and Gender Bias

In addition to their role in driving our understanding of semantic change, word embeddings have helped us grow our understanding of gender bias in language. In their paper "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", Bolukbasi et. al. (2016) call out significant gender bias in word embeddings trained on Google News articles. Bolukbasi et. al. (2016) explicitly call out occupational stereotypes represented in word embeddings by listing the occupations that are closest to the gendered pronouns 'he' and 'she' using cosine similarity, and demonstrate additional bias by using word embeddings to generate analogies which they then show exhibit gender stereotypes.

In the results of their occupational stereotype study, Bolukbasi et. al. (2016) found that 'homemaker' was the occupational word most similar to the pronoun 'she', while 'maestro' was the most similar occupation to the pronoun 'he', based on the cosine similarity values comparing each occupation to each pronoun. In addition, Bolukbasi et. al. (2016) highlighted further gender bias in the embeddings by using simple vector addition to create predictions of analogies using the word embeddings. For example, given the analogy Man:Woman::ComputerProgrammer::X, vector addition of the word embeddings gives 'homemaker' as the best fitting value for X, clearly indicating gender bias contained in these word embeddings. Having called out the blatant gender bias in these embeddings, Bolukbasi et al. (2016) and others such as Gonen and Goldberg (2019) have even researched and proposed techniques to potentially "debias" word embeddings to try and eliminate gender bias in the embeddings.

While these papers examine gender bias in word embeddings trained on more recent corpora, work by Greyson et al. (2017) has demonstrated that word embeddings can also be used to understand gender bias and roles in English corpora dating back to the 19th century. While these studies have clearly called out gender bias in word embeddings on single sets of english corpora, they have not studied the change in bias on corpora as it changes over time. In this study, we use techniques like those presented by Bolukbasi et al. (2016) on embeddings trained on corpora from separate decades, to understand how occupational gender bias in english corpora changes over time.

3 Methods

In order to carry out this study, we utilize diachronic word embeddings previously trained by Hamilton et. al. (2016) on the Corpus of Historical American English (COHA). As described by

Davies (2012), COHA contains hundreds of millions of words from thousands of texts dated from the early 19th century to the early 21st century. The corpus contains a wide variety of text, drawing from fiction and non-fiction books, magazines, and newspapers. The word embeddings were created using the Skip Gram with Negative Sampling (SGNS) method presented by Mikolov et. al. and described by Goldberg et. al. (2014). SGNS represents each word with two vectors, a word vector and a context vector to form the encompassing word embedding. Hamilton et. al. (2016) provide separate word embeddings for each decade from 1800-2000 trained from COHA. In our study, we use embeddings ranging from 1900-2000, and do not use the earlier embeddings due to an inconsistent representation of the occupation words which we focus on.

To analyze gender bias using these constructed diachronic word embeddings, we use a combination of evaluation techniques from diachronic semantic change and gender bias studies done by Bolukbasi et. al. (2016) and Hamilton et. al. (2016). First, we selected 20 occupation words to investigate in our experiment. We chose the 10 words most closely related to the word 'he', in addition to the 10 words most closely related to the word 'she', as determined by Bolukbasi. et. al. (2016) in their study on Google News articles. These words were chosen because they have been shown to exhibit the gender bias which we wish to study.

In order to understand the change in gender bias over time, we investigate the cosine similarity of the selected occupation words and gendered pronouns in embeddings trained on corpora from different decades. Specifically, we use two separate techniques to investigate the occupation pronoun pairs. The first calculates the cosine similarity of the occupation word and each individual gendered pronoun over time. For example, we calculate and plot the cosine similarity of the word pair ('he','homemaker') and the word pair ('she','homemaker') at each decade from 1900 to 2000. We can then observe whether each individual pair of words becomes more or less similar over time(e.g. did 'homemaker' become more or less similar to 'she' over time). In order to ensure that the change in similarity is not simply a reflection of an overall trend on the entire corpus, we also calculate the average cosine similarity of 50,000 randomly chosen word pairs at each time to be used as a baseline. By comparing the change in similarity of each pair, we can observe the change in an occupation word's similarity with each gender pronoun individually.

While this method allows us to observe an occupation word's relationship with each individual gendered pronoun, in order to understand bias we want to investigate the *difference* in cosine similarity of an occupation word and each gendered pronoun. We thus introduce a new metric called *bias score* to quantify the gender bias of an occupation word. We calculate the bias score of a word w at time i as the difference in cosine similarity of the word w with each gendered pronoun, calculated as:

$$b(w_i) = |csim(w_i, he_i) - csim(w_i, she_i)|$$

A higher magnitude of this difference indicates a larger difference in similarity of the word w with each gendered pronoun, thus indicating a higher level of gender bias.

In order to understand how gender bias changes over time, we calculate the bias score of each selected occupation word at each time from 1900-2000. In a slight deviation of 'pairwise similarity time series' method discussed by Hamilton et. al. (2016), we investigate a time series of bias scores. Specifically, we plot the bias score of each occupation word from 1900 to 2000 and attempt to fit the series of points with a linear function. We can then observe the slope of such a linear function, and conclude that a positive slope means that the gender bias of the occupation word increased in general from 1900 to 2000, while a negative slope indicates it decreased in general. By observing the average rate of change of the bias score of each occupational word, we can quantify the change in gender bias of occupational words in English corpora.

4 Results

Through observing the change in similarity of each occupation word with each individual gendered pronoun from 1900-2000, we cannot make a definitive overarching conclusion about the change in cosine similarity of occupation words and gendered pronouns over time. In most cases, we observed that the change in similarities did not demonstrate a consistent and clear trend of either increasing or decreasing similarity, but rather increased or decreased inconsistently over time. Figure 1 displays an individual example of analysis on the occupation

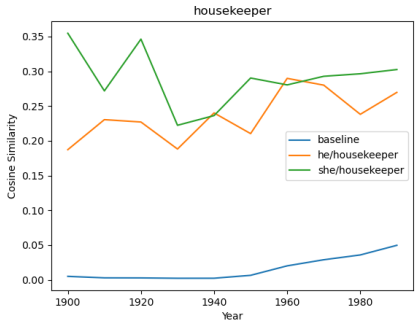

Figure 1: Cosine similarity of 'housekeeper' and the pronouns 'he' and 'she' from 1900-2000.

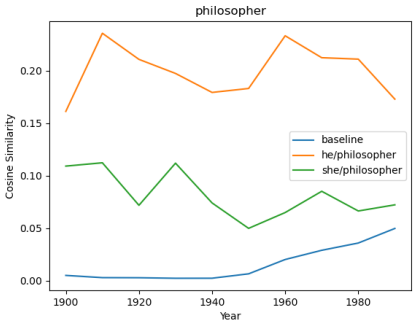

Figure 3: Cosine similarity of 'philosopher' and the pronouns 'he' and 'she' from 1900-2000.

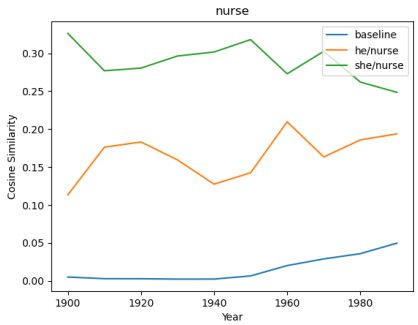

Figure 2: Cosine similarity of 'nurse' and the pronouns 'he' and 'she' from 1900-2000.

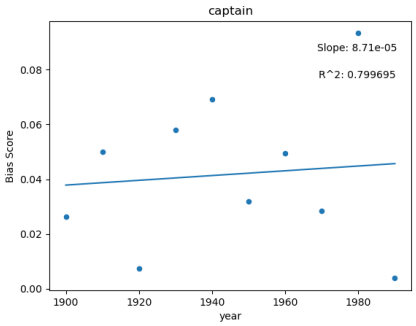

Figure 4: Bias Score of 'captain' from 1900-2000 and a linear fit.

word 'housekeeper' in which the graph indicates sporadic changes in cosine similarity with both 'he' and 'she' across decades. We observe sharp jumps in cosine similarity in either direction for both pronouns from decade to decade, with little significant trend. While for this specific word it may look as though the similarity of 'housekeeper' and 'he' is increasing in general, the given 10 data points are not enough definitively conclude that such a trend will continue as it is not consistent throughout. Additionally, the smoothness of the increase in similarity does not reflect the smoothness of the increase in the baseline similarity, thus indicating that the observed trend may have been more subject to randomness. Plots of additional words 'philosopher' and 'nurse' can be viewed in Figures 2 and 3 respectively. Much like the results of the word 'captain', these figures indicate inconsistent changes in cosine similarity over time, representative of what we observed throughout the entire set of 20 chosen occupation words.

Similar to our analysis of changes in individual cosine similarities overtime, our results from investigating the rate of change of the bias score of individual words make it difficult to form succinct and overarching conclusions. Most notably, we found that the majority of the bias score trends could not be fit with a linear line, as each trend had an R^2 value of less than 0.85, with the majority having one between 0.6 and 0.75. Even furthermore, the higher R^2 values most often resulted because each similarity score was 0 because the word did not appear in the corpus enough to give an accurate similarity score, rather than because there was a clear linear trend.

In Figure 4, we display the scatter chart of the bias score for the word 'captain' and the linear fit to the data points. The first takeaway from this chart is that there is no clear linear trend in the bias score, with an R^2 value of only 0.79. If we consider the linear trend for a moment, however, we notice that the slope of the fit line is $8.71e-05$. This positive slope indicates that in general, the bias score of the word captain is increasing over time, indicating that there becomes a larger difference in the cosine similarity of captain and he, and captain and she over time. However, the small magnitude of this slope indicates that this is not an extremely fast change according to this linear trend.

While the slope of the linear fit to the bias score

Word	Slope	R^2
nanny	.000552	0.83
captain	.000087	0.79
architect	-.000103	0.79
financier	-.000895	0.78
hairdresser	.000226	0.77

Table 1: The 5 occupations with the highest R^2 values of the linear fit to their bias score over time, and the slopes of each linear fit.

for captain is positive, for the set of 20 occupational words in general we observed both positive and negative slopes. This mix in sign of slopes of bias score over time indicates that our linear fits suggest that there is not one consistent trend across each of our chosen occupation words. Rather, for some words the linear fit suggests that their bias score increases over time, while for others it decreases. Table 1 reports the slope of the linear fit to the bias scores for those words with the five highest R^2 values. The table demonstrates a variety of positive and negative slopes in the linear fits. Overall, we observed inconsistent trends in the change of the bias score of occupation words over time. Such inconsistent trends may indicate that for the given words and corpus we focused on, there is no clear systematic way in which the gender bias of occupation words changes over time.

5 Discussion

Given the inconsistencies and lack of clear trends in our results, moving forward we search for ways to improve our methods in order to be able to make definitive conclusions. We hypothesize that one way in which we could increase our likelihood of being able to make significant conclusions is by introducing more data into our experiment. Our initial tests only included 20 words, and each word did not have consistent representation across each decade of the corpus. Moving forward, it may be beneficial to choose words which we would expect to be represented across all decades so that we may observe more consistent results. Furthermore, in the analysis of our results it may be beneficial to run more than just a simple linear regression. We cannot expect the change in bias score to necessarily be linear, and thus may benefit from exploring additional functions to determine if there is a consistent fit in the change in gender bias of occupation words. Revising this experiment to include a more consistent representation of occupation words and additional trend analysis may lead to more conclusive results.

6 Conclusion

In this paper we investigated the diachronic change in occupational gender bias in English corpora. In order to quantify change in gender bias, we introduced a new metric called *bias score*, and calculated the bias score of occupation word embeddings from 1900 to 2000. When trying to fit a linear trend to our bias score results, we did not observe a significant or consistent trend in the change in bias score of occupation words. Further investigation into this topic calls for the exploration of additional occupation words with more consistent representation across the historical corpora, and the exploration of non-linear trends.

References

[1] Breal, M. (1897). ´ Essai de semantique ´ . Paris: Hachette.

[2] Davies, Mark. "Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English." Corpora 7.2 (2012): 121-157.

[3] Goldberg, Yoav, and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." arXiv preprint arXiv:1402.3722 (2014).

[4] Gonen, Hila, and Yoav Goldberg. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them." arXiv preprint arXiv:1903.03862 (2019).

[5] Grayson, Siobhán, et al. "Exploring the role of gender in 19th century fiction through the lens of word embeddings." International Conference on Language, Data and Knowledge. Springer, Cham, 2017.

[6] Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." arXiv preprint arXiv:1605.09096 (2016).

[7] Sturtevant, E. H. (1917). Linguistic change: An introduction to the historical study of language. Chicago: The University of Chicago Press.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems.

[9] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically significant detection of linguistic change. In Proc. 24th WWW Conf., pages 625–635. International World Wide Web Conferences Steering Committee.

[10] Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In Proc. 37th Annu. Conf. Cogn. Sci. Soc.