Boosting the Performance of Gender Subspace in Domain-Specific Gender Bias Analysis

Yanqin Tan
Workhuman
yanqin.tan@workhuman.com

Cassandra L. Jacobs
University at Buffalo
Workhuman
mimi.zhang@workhuman.com

Mimi Zhang
Workhuman
cxjacobs@buffalo.edu

Marvin Thielk
Workhuman
marvin.thielk@workhuman.com

Yi Chu
Workhuman
yi.chu@workhuman.com

Abstract

Gender subspace analysis is a useful means of quantifying gender bias using word embeddings. However, this technique is challenging to apply when the domain is shifted from off-the-shelf word embeddings and/or when the training dataset size is smaller. This study reports our qualitative and quantitative analyses of the learned gender subspace in a workplace communication dataset. We highlight how word sparsity and homonyms (i.e. words that have multiple semantic meanings) affect word embeddings, leading to degraded performance on subspace learning. To improve the robustness of a learned gender subspace for domain-specific data, we then propose word replacement and substitution to exclude non-gendered information from word embeddings for empirical gender bias analysis.

1 Introduction

Gender bias is a serious ethical concern in artificial intelligence that takes historical language data as input. Static and contextualized word embeddings are known to encode gender stereotypes and bias (Bolukbasi et al., 2016; Bhardwaj et al., 2021). Bolukbasi et al. (2016) proposed to use the concept of a gender subspace to measure and remove gender bias from embeddings. This concept has been extended to measure gender bias in corpora (Babaeianjelodar et al., 2020). In principle, the words of interest are words that are used to describe different gender, which are supposed to be ideally gender-neutral. However, this measurement is reliable only if the learned gender subspace is consistent and robust to noise.

Inspired by these works, we extend Bolukbasi et al. (2016) to analyze gender bias in workplace communication, where gender bias has deep implications for women’s career performance and development (Heilman, 2012; Correll et al., 2020). The dataset we use comes from Workhuman’s recognition platform, which contains messages users written to express praise or gratitude to their coworkers. This newer, lower-resource domain is critical to estimate the robustness of gender subspace representations. We show that empirical use of gender subspace requires prior study of gender-specific pairs in order to perform well (Ethayarajh et al., 2019) and word substitution based on domain knowledge is an effective way to boost the performance of gender subspace analysis in new domains.

2 Methods and Experiments

In Bolukbasi et al. (2016), a gender space is defined by the vector differences of 10 gender-specific pairs (e.g. “he-she”, “his-her”, “himself-herself”, “male-female”, “man-woman”, “father-mother”, “son-daughter”, “boy-girl”, “guy-gal”, “Mary-John”). A gender subspace is the first principal component extracted from these 10 pair-wise (e.g., “he-she”) vector differences, which encodes binary gender continuum in one dimension. The cosine similarity between a target word’s vector representation and this gender subspace results in a “bias score” for words of interest. The more similar a word’s vector is to the gender subspace, the more biased it is.

Bolukbasi et al. (2016) train the gender subspace on the Google News dataset, which contains 100 billion tokens and 3 million unique words from diverse topics. By comparison, our workplace context-specific dataset contains only 10 million tokens and 180k unique words. This poses two challenges to the application of gender subspace: (1) because of the domain-specific nature and the smaller size, the occurrence of some gendered pairs might be sparse; (2) the (homonymous) use of these terms in non-gendered contexts has an amplified impact on the vector representations of these words.

To verify these assumptions and their impacts, we first conduct a statistical analysis on the
Workhuman data. As shown in Table 1, one pair (“Mary-John”) is not present and several other words are rare. To measure how relevant these sparse gender-specific words are to a gender subspace, we compare the cosine similarity of a new gender subspace created by adding a new pair to two benchmarks: (1) the gender subspace created by one gender-specific pair “he-she” and (2) the gender subspace created by five gender-specific pairs - “he-she”, “his-her”, “himself-herself”, “man-woman” and “male-female”. For comparison, we also applied the same procedure to the Google News embeddings. As shown in Figure 1, the pretrained model has a more stable gender space, changing only slightly as additional pairs are added. However, the gender subspace built with these same gender pairs in our model changes dramatically, with the cosine similarity dropping to 0.24 when the “father-mother” pair is added and to 0.08 when the “guy-gal” pair is added.

We hypothesize that lexical ambiguity jeopardizes the reliability of gender subspace and acts as a source of noise in our model, as one irrelevant meaning of the homonym may obstruct the other. Homonymous word embeddings are subject to catastrophic forgetting, losing what was learned earlier in a dataset while reinforcing what is learned more recently (Mannering and Jones, 2021). In our case, some words in the gender-neutral word lists are used in non-gendered context (Ethayarajh et al., 2019). For example, “mother” is used in phrases like, “Mother Goose” and “boy” is used as colloquial expressions of interjection in phrases like, “oh boy!” Similarly, plural “guys” (in contrast to “gals”) may refer to a group of people (“you guys”) without necessarily encoding gender. The interference between word senses might make the vector representation of low-resource pairs much less accurate in our domain-specific model.

To overcome this, our method is to use two rounds of manual replacement to transform gender-irrelevant usages into a non-gendered term, i.e., replacing “father” or “mother” with “parent”, “boy” and “girl” with “kid” and “gal” and “guy(s)” with “pal”, to eliminate the non-gendered meaning. Specific to our goal of measuring bias in workplace language as seen in recognition messages to coworkers of different genders. In the first round, gender-specific words used in absolute non-gendered context like those in the examples above are replaced. Then, we exclude gendered language that does not pertain to the recipient of the recognition (e.g., when ‘mother’ is mentioned but not to identify the addressee as a mother). “Girl” was replaced with “gal” when “girl” is used in a similar context as “guy” to better represent the “guy-gal” pair.

After replacement, these words generate a more reliable gender subspace. Figure 2 shows that the two rounds of replacement improves the gender subspace’s cosine similarity to Benchmark 2 (gender subspace from the five pairs) substantially.

### Conclusion

Gender subspace can inform empirical gender bias analysis in different social domains. However, the use of gender-specific pairs requires prior research for the sake of reliability and robustness. In this work, we show that word replacement and substitution based on domain knowledge is a reliable method to counter word sparsity and noise introduced by homonyms in shifted domains.
References


