The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations

João Sedoc and Lyle Ungar
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
joao@cis.upenn.edu, ungar@cis.upenn.edu

Abstract

Systemic bias in word embeddings has been widely reported and studied, and efforts made to debias them; however, new contextualized embeddings such as ELMo and BERT are only now being similarly studied. Standard debiasing methods require large, heterogeneous lists of target words to identify the "bias subspace". We show that using new contextualized word embeddings in conceptor debiasing allows us to more accurately debias word embeddings by breaking target word lists into more homogeneous subsets and then combining ("Or'ing") the debiasing conceptors of the different subsets.

1 Introduction

Contextualized word representations are replacing word vectors in many natural language processing (NLP) tasks such as sentiment analysis, coreference resolution, question answering, textual entailment, and named entity recognition (Peters et al., 2018; Devlin et al., 2018). However, ELMo and BERT have bias similar (Wang et al., 2019; May et al., 2019; Kurita et al., 2019) to the well documented bias in traditional word embedding methods (Bolukbasi et al., 2016; Bhatia, 2017; Caliskan et al., 2017; Nikhil Garg and Zou, 2018; Kiritchenko and Mohammad, 2018; Rudinger et al., 2018; Zhang et al., 2018), and this could cause bias in NLP pipelines used for high stakes downstream tasks such as resume selection or bail setting algorithms (Hansen et al., 2015; Bolukbasi et al., 2016; Ayres, 2002). Traditional word embeddings, such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and Fasttext (Bojanowski et al., 2017) require large sets of target words, since debiasing is generally done in the space of the PCA of the word embeddings. (If one only uses a two words, like ”man” and ”woman”, the PCA space is just a single vector pointing in the difference between those two vectors.) Context-sensitive embedding such as ELMo and BERT give an embedding for every token (based on its context), giving large numbers of embedding for each word (such as ”man”), so that principal components can be calculated even for word lists of size two as shown in Figure 1.

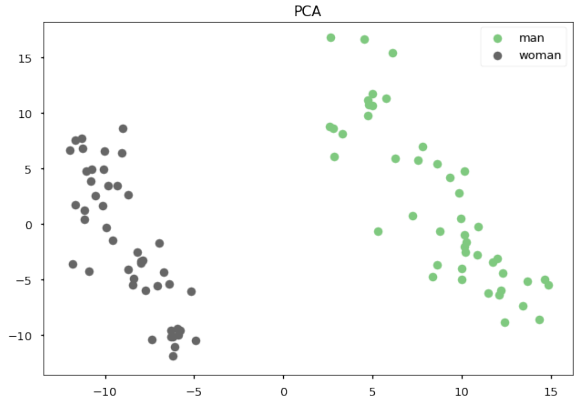


Figure 1: ELMo word representations of tokens of *man* and *woman* projected onto their first and second principal components.

Use of contextualized word embedding allows better debiasing by allowing one (as will be described below) to break up target word lists into smaller homogeneous subsets; it also gives better insight into where the bias may be coming from.

Word embeddings capture distributional similarities; just as humans come to associate certain professions (homemaker or computer programmer) with certain genders (woman or man), word embeddings capture very similar associations (Bolukbasi et al., 2016). Such embedding biases tend to track statistical regularities such as percentage of people with a given occupation (Nikhil Garg and Zou, 2018) but sometimes deviate from them (Bhatia, 2017).

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 55–61
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

A number of debiasing methods have been proposed. Most of them use hard debiasing – zeroing out one or more directions in the embedding space, generally selected using principal components (Bolukbasi et al., 2016; Wang et al., 2019). In this paper, we use a soft debiasing method, *conceptor debiasing*, which also works in the principal component space, but does a softer shrinkage of the bias and close-by directions (Liu et al., 2018).

Many debiasing algorithms rely entirely on so called "target lists" of protected classes in order to identify and mitigate the "bias subspace"; however, to our knowledge no work examines the role of these target lists in defining this space. This in part due to the fact that in standard word embeddings there is only one embedding for a token. In contrast, new contextualized word representations such as BERT and ELMo have a different embedding for each word token in a context. This allows us an opportunity to more closely examine what information target word lists are capturing.

This paper:

- Examines bias in ELMo and BERT, taking advantage of their context-sensitivity to give better visualizations.
- Shows how heterogeneity in content and size of the "target list" of gendered or racially marked terms interferes with debiasing, and how conceptors on contextual embeddings can be used to address such target list heterogeneity.

2 Related Work

NLP has begun tackling the problems that are limiting the achievement of fair and ethical AI (Hovy and Spruit, 2016; Friedler et al., 2016), including techniques for mitigating demographic biases in models. In brief, a demographic bias is taken to mean a difference in model output based on gender (either of the data author or within the content itself) or selected demographic dimension ("protected class") such as race. Demographic biases manifest in many ways, from disparities in tagging and classification accuracy depending on author age and gender (Hovy, 2015; Dixon et al., 2018), to over-amplification of demographic differences in language generation (Yatskar et al., 2016; Zhao et al., 2017), to diverging implicit associations between words or concepts within embeddings or language models (Bolukbasi et al., 2016; Rudinger et al., 2018).

Recent work of Wang et al. (2019) shows bias in ELMo and presents several examples of successful debiasing. However, May et al. (2019) found that bias in BERT may be more difficult to identify, but Kurita et al. (2019) did indeed find bias in BERT. However, prior work has not focused on identifying word lists as a potential area of research.

3 Target Word Lists

To debias word embeddings, an appropriate word list representing the bias in question needs to be used to define the subspace. [1] For example, a gender word list might be a set of pronouns which are specific to a particular gender such as *he / she* or *himself / herself* and gender specific words representing relationships like *brother / sister* or *uncle / aunt*. We test conceptor debiasing both using the list of such pronouns[2] used by Caliskan et al. (2017) and using a more comprehensive list of gender-specific words that also includes gender-specific terms related to occupations, relationships and other commonly used words such as *prince / princess* and *host / hostess*[3]. We further tested conceptor (Jaeger, 2014; Liu et al., 2018) (soft) debiasing using male and female names such as *Aaron / Alice* or *Chris / Clary*.[4]

Previous researchers used a variety of different word lists, but did not study the effect of word list selection; we show below that the word list matters. However, we leave systematic study for future work.

3.1 Word Lists and Principal Components

Recall most debiasing methods rely on principal components of the matrix of embeddings of the target words. Hard debiasing methods remove the first or first several principal components (Bolukbasi et al., 2016; Mu and Viswanath, 2018). Conceptors, as explained below, do soft debiasing in the same principal component space.

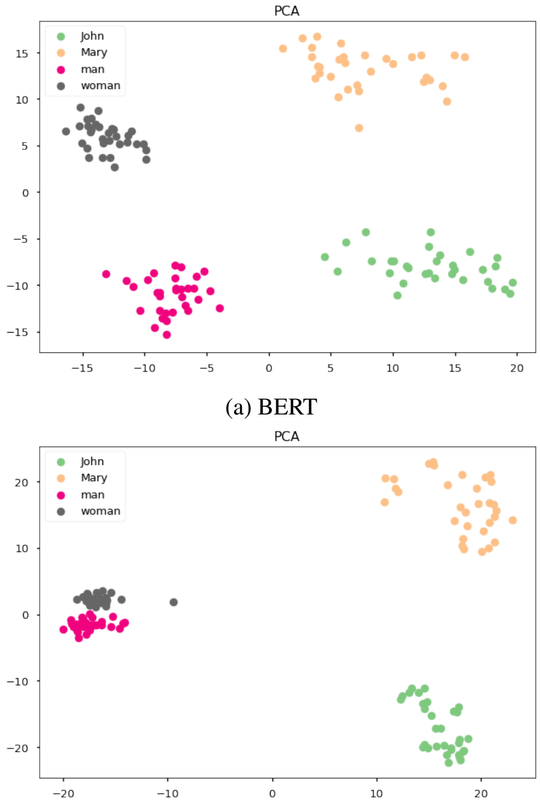
Paired nouns and pronouns should provide better support for debiasing than names if we assume that the linguistic markers are unambiguous

[1] Some methods also require a list of unbiased words as well, but we will not address those since conceptor debiasing does not require them.

[2] https://github.com/jsedoc/ConceptorDebias/tree/master/lists

[3] https://github.com/uclanlp/corefBias, https://github.com/uclanlp/gn_glove

[4] https://www.cs.cmu.edu/Groups/AI/areas/nlp/corpora/names/



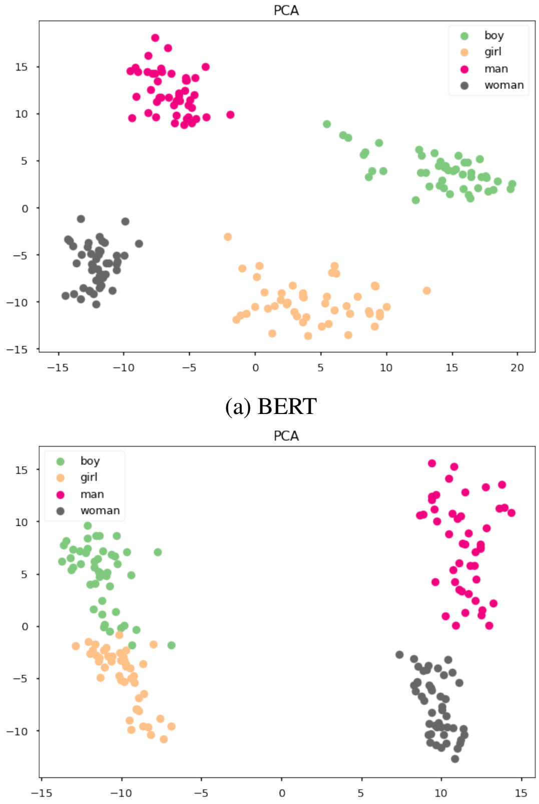
(a) BERT

(b) ELMo

Figure 2: BERT and ELMo word representations of the union of the set of contextualized word representations of the pairs *man / woman* and *Mary / John* projected onto their first and second principal components.

(a counterexample is ”guys”), and there is no polysemy. Names of people can also be ambiguous (e.g. “Pat”). A possible solution for this which we leave for future work is to regress the first few principal components of a word pair with the binary attribute to verify that the pair is properly captures the attribute of interest on out-of-sample lists. In fact, for racial names Gaddis (2017)’s method (using linear regression) can be used to both filter and pair names. While this is difficult to achieve using word embeddings which are at the type level (i.e. one vector per word as in Fasttext and word2vec), for contextualized word representations, which are token level (i.e. one vector per word and context), this is completely feasible.

Figure 1 shows how the pair *man / woman* cleanly separates across the first principal component of the space of their contextualized representations. However, even though one word pair give good results, combining it with a second word pair can have unfortunate effects; debiasing becomes more complicated if we add another pair of words, say *Mary / John* to the pair *man / woman*, as shown in Figure 2. The first principal component is now capturing pronoun vs proper noun difference, which we do not desire to remove after debiasing.



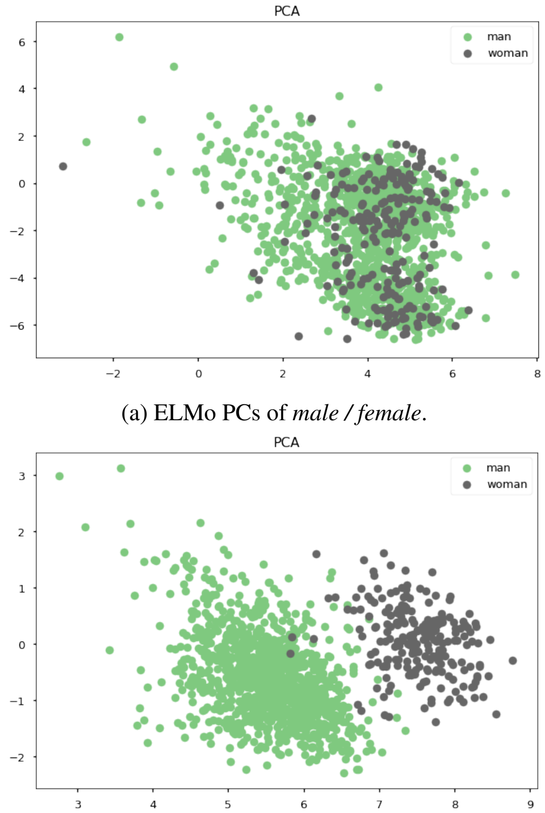
(a) BERT

(b) ELMo

Figure 3: BERT and ELMo word representations of the union of the set of contextualized word representations of the pairs *man / woman* and *boy / girl* projected onto their first and second principal components.

It is also critical to note that contextualized word embeddings are very rich, so while one might think the union of the contextualized word representations of *man / woman* and *boy / girl* would yield a good gender direction, in fact we find that the first principal component of these four words is along the direction of adults vs. children (see Figure 3). There is some separation between “husband” and “wife” in this dimension, but none between “boy” and “girl”. Similarly when names such as *Mary / John* are projected onto this subspace, little separation occurs. Since most debiasing methods remove or shrink these principal component directions, this combined word list does poorly for debiasing.

Furthermore, some apparently sensible target word lists are not useful for debiasing contextu-



(a) ELMo PCs of *male / female*.

(b) ELMo PC of *John / Mary*

Figure 4: ELMo word representations of *man / woman* projected onto the first and second principal components defined by the pair (a) *male / female* and (b) *John / Mary*.

alized representations. Figure 4 shows that the ELMO vectors of *man* and /*woman* separate nicely when projected on to the first two principal components of the *Mary / John* but fail to separate when projected on to the first two principal components of *male / female*. The word *male / female* word pair form a poor target list since they are, in fact, rarely used to refer to people; They are instead applied to animals (*the male parrot*) or to distinguish a break of the social bias (*the male model*).

Note that none of the above figures could have been generated using traditional word embeddings; one cannot get two PCA dimensions for a target word list of only two words.

3.2 Conceptors

Conceptors provide and effective, computationally cheap and mathematically elegant a way of doing soft debiasing of word embeddings. As with many debiasing methods, the input is matrix Z of word embeddings corresponding to a set of target

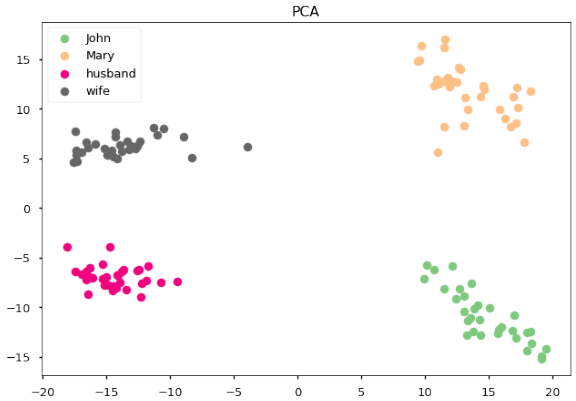


Figure 5: BERT word representations of the union of the set of contextualized word representations of the pairs *husband / wife* and *Mary / John* projected onto the first and second principal components.

words, $\mathcal{Z}$. (These can either be one embedding per word type, for conventional embeddings, or one vector per word token, as we use here for context-sensitive embeddings; for best results, $\mathcal{Z}$ should be mean-centered.) A conceptor matrix, C, is a regularized identity map (in our case, from the original word embeddings to their biased versions) that minimizes

$$\|Z - CZ\|_F^2 + \alpha^{-2}\|C\|_F^2. \tag{1}$$

where α^{-2} is a scalar parameter. As described in the orignal work on matrix conceptors (Jaeger, 2014; He and Jaeger, 2018; Liu et al., 2019b,a) C has a closed form solution:

$$C = \frac{1}{k}ZZ^\top(\frac{1}{k}ZZ^\top + \alpha^{-2}I)^{-1}. \tag{2}$$

Intuitively, C is a soft projection matrix on the linear subspace that gives the largest shrinkage where the word embeddings Z have the highest variance. Once C has been learned, it can be 'negated' by subtracting it from the identity matrix and then applied to any word embeddings to shrink their bias directions.

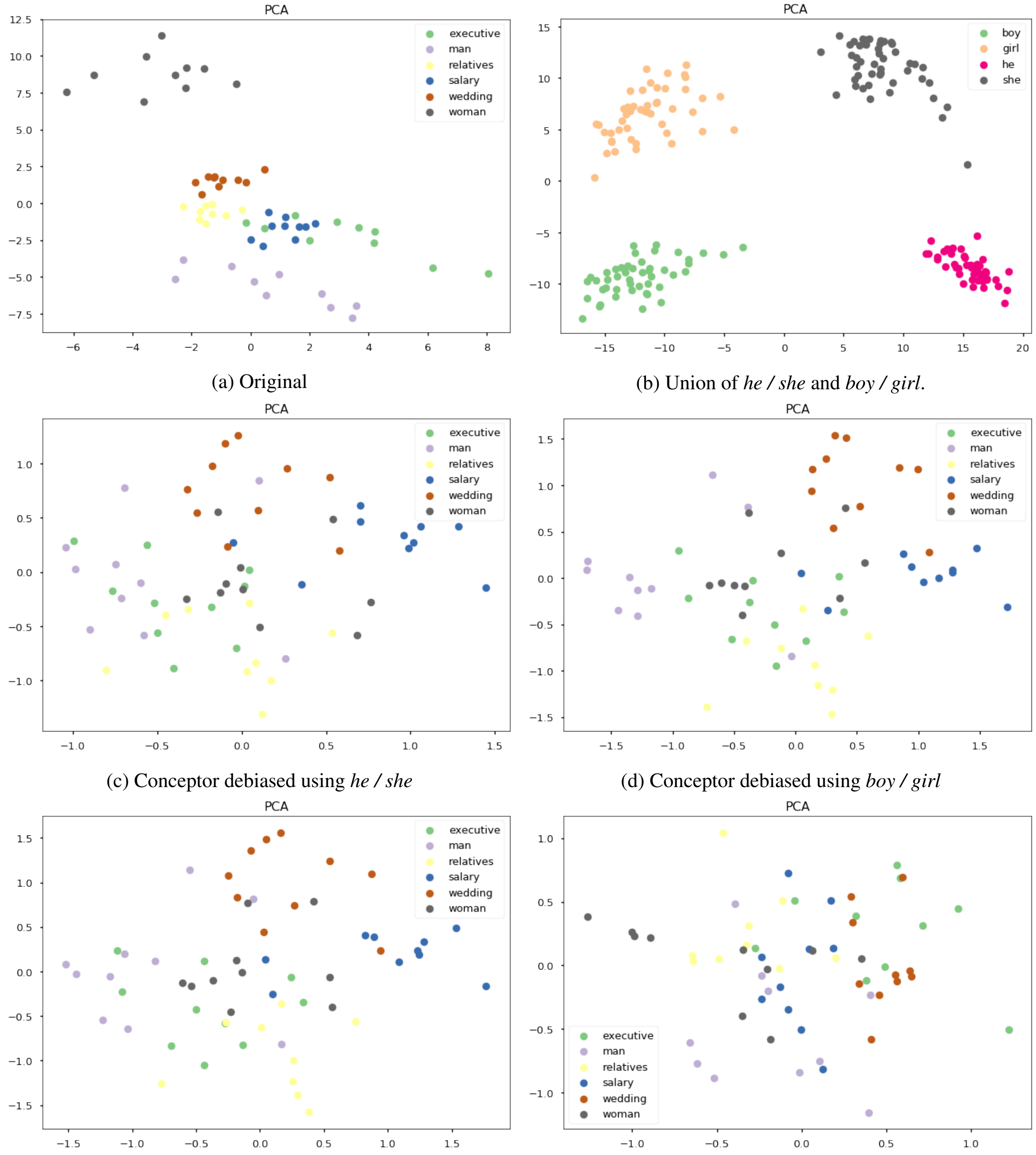
Conceptors can represent laws of Boolean logic, such as NOT $\neg$, AND $\wedge$, and OR $\vee$. For two conceptors C and B, we define the following operations:

$$\neg C := \mathbf{I} - C, \tag{3}$$

$$C \wedge B := (C^{-1} + B^{-1} - \mathbf{I})^{-1} \tag{4}$$

$$C \vee B := \neg(\neg C \wedge \neg B) \tag{5}$$

Thus, to minimize bias, we apply the negated conceptor, NOT C (see Equation 3) to an embedding space and reduce its bias. We call NOT C the



(a) Original

(b) Union of *he / she* and *boy / girl*.

(c) Conceptor debiased using *he / she*

(d) Conceptor debiased using *boy / girl*

(e) Conceptor debiased using the **union** of *he / she, boy / girl*

(f) Conceptor debiased using **OR** *he / she, boy / girl*

Figure 6: Effect of target word lists on debiasing BERT word representations. The union of the set of contextualized word representations of *career, business, family, children, man, woman* projected on to the first two principal components of *he / she*.

debiasing conceptor. More generally, if we have K conceptors, C_i derived from K different word lists, we call NOT $(C_1 \vee ... \vee C_K)$ a debiasing conceptor.

Negated conceptors do soft debiasing, shrinking each principal component of the covariance matrix of the target word embeddings $ZZ^\top$ based on the conceptor hyper-parameter α and the eigenvalues σ_i of $ZZ^\top$: $\frac{\alpha^{-2}}{\sigma_i + \alpha^{-2}}$. (Liu et al., 2018).

4 Conceptor Debiasing

Above we showed that visualizations of gender and racial subspaces gives insight for how word lists for embedding can fail to produce good results. We now show how conceptor negation, applied across homogeneous subsets of the word list can improve performance.

Figure 6a shows that there is a gender bias using career versus family words projected onto the gender space. Figure 6 shows that after debiasing using conceptor negation (Liu et al., 2018) (as defined above) there is substantially less bias.

Nonetheless, one should note that gender bias need not be in the first two dimensions. In fact recent work by Gonen and Goldberg (2019) has pointed out that most "debiasing" methods are simply mitigating bias and thus end task methods will potentially be able to undo this mitigation. As a result, we recommend that a method like Gaddis (2017) be used to identify proper word lists.

5 Conclusion

We showed that one should take care when debiasing word embeddings; well-chosen word lists generally yield better subspaces than poorly-chosen ones. Combining heterogeneous words into a single word list presents a host of problems; a couple of 'bad' words like "male/female" can significantly shift the dominant principal components of the bias space. Conversely, since PCA effectively weights words by their frequency of occurrence, combining small word lists (pronouns) with large word lists (names) means that the longer word lists carry more weight in the principal components (unless the rare words 'stick out' a long way in a different direction).

Conceptor debiasing provides a simple way of addressing the problem of combining word lists of different types and sizes, improving performance over state-of-the art 'hard' debiasing methods. Conceptor debiasing has the further benefit that conceptor negation methods allow one to learn separate conceptors for each word subset and then to OR them. The best results are obtained when lists are broken up into subsets of 'similar' words (pronouns, professions, names, etc). Similarly, conceptors for different protected subclasses such as gender and race can be OR'd to simultaneously debias for both classes. OR'ing has the advantage that word lists of different size are still treated as equally important–a key factor when lists such as pronouns, male and female names and black and white names may be of vastly different sizes.

Contextual embeddings such as ELMo and BERT, which give a different vector for each word token, work particularly well with specialized word lists, since they produce a large number of embeddings, allow principal components to be computed and used for debiasing even for lists of two words.

Finally, the main takeaway from this paper is that word lists matter, especially for debiasing contextualized word embeddings. Remember Figures 3 and 4b where intuition fails entirely!

References

Ian Ayres. 2002. Outcome tests of racial disparities in police practices. *Justice research and Policy*, 4(1-2):131–142.

Sudeep Bhatia. 2017. The semantic representation of prejudice and stereotypes. *Cognition*, 164:46–60.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

S Michael Gaddis. 2017. How black are lakisha and jamal? racial perceptions from names used in correspondence audit studies. *Sociological Science*, 4:469–489.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

C Hansen, M Tosik, G Goossen, C Li, L Bayeva, F Berbain, and M Rotaru. 2015. How to get the best word vectors for resume parsing. In *SNN Adaptive Intelligence/Symposium: Machine Learning*.

X. He and H. Jaeger. 2018. Overcoming catastrophic interference using conceptor-aided backpropagation. In *International Conference on Learning Representations*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

H. Jaeger. 2014. Controlling recurrent neural networks by conceptors. Technical report, Jacobs University Bremen.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations.

T. Liu, J. Sedoc, and L. Ungar. 2018. Correcting the common discourse bias in linear representation of sentences using conceptors. In *Proceedings of ACM-BCB- 2018 Workshop on BioCreative/OHNLP Challenge, Washington, D.C., 2018.*

T. Liu, L. Ungar, and J. Sedoc. 2019a. Continual learning for sentence representations using conceptors. In *Proceedings of the NAACL HLT 2019.*

T. Liu, L. Ungar, and J. Sedoc. 2019b. Unsupervised post-processing of word vectors via conceptor negation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-2019), Honolulu.*

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *North American Chapter of the Association for Computational Linguistics (NAACL).*

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

J. Mu and P. Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Dan Jurafsky Nikhil Garg, Londa Schiebinger and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*.

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *North American Chapter of the Association for Computational Linguistics (NAACL).*

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 335–340, New York, NY, USA. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.