Impact of Debiasing Word Embeddings on Information Retrieval

Emma Gerritse

Institute for Computing and Information Sciences
Radboud University, Nijmegen, the Netherlands
emma.gerritse@ru.nl

Abstract. Word embeddings are a core technology in neural methods for information retrieval. However, previous work has suggested undesirable biases in word embeddings, in particular against gender. In this paper, we look at the extent of the bias in different cases. Presumably, not all biased analogies are 'robust' and can sometimes give unexpected results. We discuss some ways in which bias in word embeddings could affect systems in information retrieval, which is the topic of our future research.

Keywords: Word embeddings · Bias · Information Retrieval.

1 Introduction

The vector representations of words generated by neural network methods are now commonly used for various information retrieval applications. Word2Vec [6] and Glove [8] are amongst the best-known word embeddings and are used in various downstream tasks, including document retrieval [2, 3, 5, 7, 9]. However, Bolukbasi et al. [1] raises the concern that pre-trained word embeddings are biased and exhibit female/male gender stereotypes to a disturbing extent. They show that some analogies give sexist results; e.g., the answer to the analogy *man : computer programmer as woman : x* solves for $x = homemaker$. They further proposed two methods of removing gender stereotypes from word embeddings and transforming embeddings such that gender-neutral words are not closer to one gender than to another. In this paper, we provide an overview of our research project on the effect of word embedding bias for information retrieval tasks. Our main two research questions are:

- RQ1: To what extent are word embeddings biased?
- RQ2: Which information retrieval tasks are affected by this bias effect?

 FDIA 2019, 17-18 July 2019, Milan, Italy.

First, we discuss the related work in Section 2. In Section 3, we look at word embedding bias based on word analogies and explore geometric properties of word vectors. For the analogies, we observe that some analogies are biased, but we also observe that they are not 'robust'. Some analogies present unexplained behavior, which needs further investigations. Considering the geometric properties of vectors, we find that the distribution of male and female words is not equal, while the PCA visualization of vectors does not show gender biased clusters. Next, in Section 4, we discuss the effect of bias in word embedding for information retrieval tasks and highlight future directions.

2 Related work

2.1 Word embeddings

Word embeddings are representations of words in a lower dimensional space, which capture relations between words. In the embedding space, similar words are mapped close together. On top of that, the difference between vectors has meaning. For example, the result of $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$ should hold. Well-known methods of word embeddings are Word2Vec [6] and Glove [8]. Word2Vec is trained by predicting a word depending on a surrounding window of context words (Continuous Bag of Words) or by predicting the surrounding window of words using the current word (skip-gram). Pre-trained embeddings of Word2Vec are shared for researchers to use, most famously the embeddings trained by Mikolov et al. [6] , on the Google News dataset in 2013. Since these embeddings are easily available, they are often used in research in information retrieval.

2.2 Bias in word embeddings

Though word embeddings are very useful, Bolukbasi et al. [1] found that word embeddings can exhibit biases. Their paper focuses on the pre-trained embeddings on the Google News Dataset. Considering analogies found in the embedding space, they sometimes can give sexist results like $\overrightarrow{computer\_programmer} - \overrightarrow{he} + \overrightarrow{she} = \overrightarrow{homemaker}$. To find biased words, Bolukbasi et al. [1] calculated the projection of words in the dictionary on the he-she plane; by calculating $\overrightarrow{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$. The larger this value, the larger the bias associated with that word is. They identified many analogies which could be biased, and Mechanical Turkers rated these analogies for their level of bias. They constructed two ways of debiasing, referred to as soft and hard debiasing. With these methods, words which should be gender neutral are mapped to have the same distance between clearly male and clearly female terms.
However, Gonen and Goldberg [4] have shown that debiasing is harder than previously thought. They show that after debiasing with the methods proposed by Bolukbasi et al. [1], the original bias can still be recovered. They show this in multiple ways. One of them is by applying k-means clustering on the 1000 most

biased words, both before and after debiasing. They find that with an accuracy of 92.5%, words are assigned to the same gender cluster as before debiasing. They show that when training a classifier on the debiased words, they can still predict with high accuracy whether they belong to the male or female group.

2.3 Use of word embeddings in IR

Word Embeddings are extensively used in Information Retrieval [2, 3, 5, 7, 9]. One example of this is Dehghani et al. [2], where a neural ranking model is trained with weak supervision. As an input representation, word vectors are used. This method seems to gain a big improvement on simply using BM25. Another example is Diaz et al. [3], who have shown that using locally trained word embeddings can be useful for query expansion.

3 Bias in word embeddings

Word embeddings are well known to solve word analogies of the form "a is to b as c is to d" and to exhibit meaningful distances between similar words. We compute the distance between two word vectors using cosine similarity, so the distance between words a and b would be $1-cos(\overrightarrow{a}, \overrightarrow{b})$. We compute the answer of an analogy by solving for the vector which has the greatest cosine similarity in the following equation:

$$\max_{d \in D \setminus \{a,b,c\}} cos(\overrightarrow{a} - \overrightarrow{b} + \overrightarrow{c}, \overrightarrow{d})$$

We note this as $\overrightarrow{a} - \overrightarrow{b} + \overrightarrow{c} \approx \overrightarrow{d}$. If the reverse of this equation also holds, so $\overrightarrow{d} - \overrightarrow{c} + \overrightarrow{b} \approx \overrightarrow{a}$, we will call the analogy *robust*.

Bolukbasi et al. [1] examined gender biases in the embeddings, both in the analogies and distances between words. We are going to look into these two aspects as well. Gender-biased results can be observed in the analogies generated from Word2Vec embeddings; e.g., $\overrightarrow{computer\_programmer} - \overrightarrow{he} + \overrightarrow{she} \approx \overrightarrow{homemaker}$. We tested some of the analogies given in [1], and found that while the answer to *computer programmer - he + she* is indeed homemaker, this analogy is not robust. Computing the reverse, we find that $\overrightarrow{homemaker} - \overrightarrow{she} + \overrightarrow{he} \approx \overrightarrow{carpenter}$. We further repeat this process and alternate in both directions until we get the same result in both directions. Figure 1 shows several examples. We observe that $\overrightarrow{carpenter} - \overrightarrow{he} + \overrightarrow{she} \approx \overrightarrow{seamstress}$, and for this analogy is robust. While this analogy is still biased, it seems less severe than the *computer programmer* and *homemaker* combination.

Analogies depend on the choice of words examined. When taking the word *programmer* instead of *computer_programmer*, the analogy solves to *programmers* (plural form), which does not seem to be biased. However, this analogy converges to $\overrightarrow{Henrik\_Sørensen} - \overrightarrow{she} + \overrightarrow{he} \approx \overrightarrow{Martin\_Thorborg\_together}$, which are the two founders of a Danish website. This suggests that the behavior of

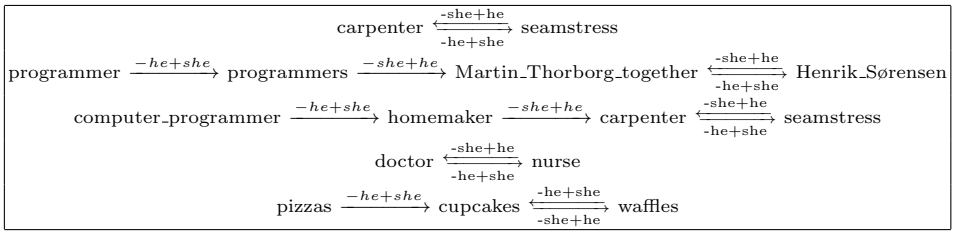

Fig. 1. Analogy convergence over several iterations.

these embedding vectors is not fully explainable; while some of these analogies converge immediately, others converge to seemingly unrelated results. To further investigate the strength of these biases, we would like to construct evaluation measures, taking this convergence into account. We also hope to gain more insights on how embedding analogies behave, and why these seemingly unexpected results are obtained.

Another question is whether biases can be observed from the geometric properties of embedding vectors or not. To this end, we use the list of professions given by Bolukbasi et al. [1] and divide them into female and male words based on their cosine similarity between words *he* and *she*. We then compute $|\cos(\overrightarrow{w}, \overrightarrow{he}) - \cos(\overrightarrow{w}, \overrightarrow{she})|$ and observe how this measure varies for the female and male professions, (see also [1]). The results show more male words for unbiased terms (values closer to 0) and more female words for biased terms (values larger than 0.2). This suggests that there is a bigger bias towards female words than to male words. On the other hand, when applying Principal Component Analysis (PCA) to the same list of word vectors, we see that there is a less severe separation between male and female words; see Figure 2. [1] Therefore, when gender is the most important component of the embeddings, we should be able to see this in the PCA plot. This PCA projection suggests that the first and second components of these vectors are not related to gender, but reflect other properties of the words.

4 Effect of bias in word embeddings for IR

One of the reasons why bias in word embeddings could be harmful is because the pre-trained word embeddings are often used by other researchers. Bolukbasi et al. [1] give as an example that this bias might apply when looking for people having a certain profession. If the profession is labeled as male, search engines using word embeddings might favor men instead of women. This is especially unwanted for conversational search because users here are often given only one result. However, they did not test this claim in their paper. That is why we want to test this, using some to be determined later task in IR. For example, we could use a neural IR system like Dehghani et al. [2] and see if debiasing gives different

[1] We chose PCA over T-SNE, as PCA reserves the relations in the embedding vectors.

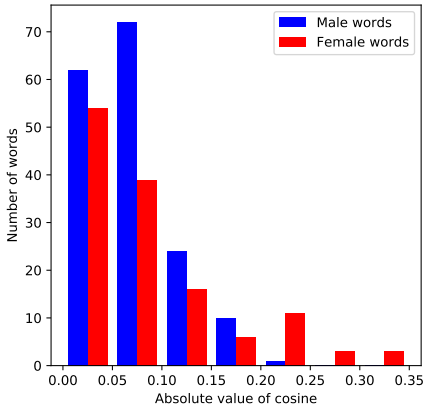

Fig. 2. Frequency of professions listed in [1], based on $|(\cos(\overrightarrow{w}, \overrightarrow{he}) - \cos(\overrightarrow{w}, \overrightarrow{she})|$. Professions with positive value of $\cos(\overrightarrow{w}, \overrightarrow{he}) - \cos(\overrightarrow{w}, \overrightarrow{she})$ are considered as male words, and the other way around.

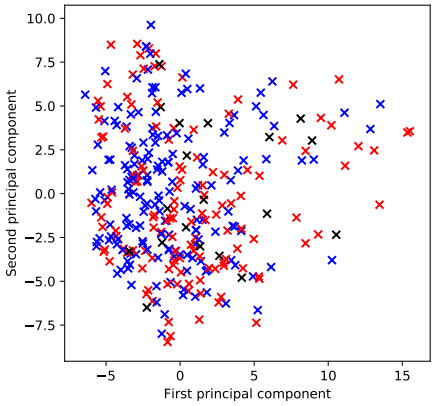

Fig. 3. PCA projection of all professions listed in [1]. , with blue: $\cos(\overrightarrow{w}, \overrightarrow{he}) - \cos(\overrightarrow{w}, \overrightarrow{she}) > 0$, red: $\cos(\overrightarrow{w}, \overrightarrow{he}) - \cos(\overrightarrow{w}, \overrightarrow{she}) < 0$, and black the gender neutral words.

results. We also wonder if there is a difference in the severity of the bias based on the task. If debiasing gives a difference in search results, we want to see what the effects of this debiasing are.
Another question we have is what happens if the user intent is biased. Say someone is looking for example for all female singers in the Eurovision song contest. Since $\overrightarrow{singer} \cdot (\overrightarrow{he} - \overrightarrow{she}) < 0$, this is labeled as a female word. However, you do not want the bias correction to suddenly show male singers as well.

This brings us to another topic: we still do not know why the bias enters our embeddings. Bolukbasi et al. [1] show that there is a high correlation in bias in Word2Vec trained on Google News and Glove trained on the common crawl, so we still cannot infer whether the method or the dataset is more important for creating the bias. This makes us wonder if these kinds of biases also enter in other neural IR systems, like the systems developed by Zamani et al. [10] or Guo et al. [5]. These systems seem to score well. However, to our knowledge, nobody looked into the bias these systems might infer.

5 Conclusion

At this moment, we can conclude that there is still a lot of work needed to understand how word embeddings behave in a real-life setting. As seen in previous work, bias exists in word embeddings, both in the analogies and in the distance between words. However, there are still some open questions of how these biases are formed. We still do not know why bias is more visible when looking at the

cosine distance of professions to gendered words than when looking at the PCA of these same professions. Word embeddings can give weird results, like creating analogies to seemingly unrelated words. We should be careful to include them in our IR systems if we do not fully understand how they work. On top of that, we still do not know how bias and debiasing behave in an actual IR setting. Not only do we not know this for word embeddings, but we also do not know this for other neural methods in IR.

Bibliography

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems 29*, pages 4349–4357, 2016.

[2] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. Neural ranking models with weak supervision. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74, 2017.

[3] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. Query expansion with locally-trained word embeddings. *arXiv preprint arXiv:1605.07891*, 2016.

[4] Hila Gonen 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] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64, 2016.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[7] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. *Proceedings of the 26th International Conference on World Wide Web*, pages 1291–1299, 2017.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[9] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64, 2017.

[10] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. From neural re-ranking to neural ranking. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM 18*. ACM Press, 2018.