On the Unintended Social Bias of Training Language Generation Models with Latin American Newspapers

Anonymous Author(s)

Abstract

Gender bias is a significant problem when generating text, and its unintended memorization could impact the user experience of many applications (e.g., the auto-complete feature in Gmail). In this abstract, we introduce a novel architecture that decouples the representation learning of a neural model from its memory management role. This architecture allows us to update a memory module with an equal ratio across gender types breaking biased correlations directly in the latent space. We show that our approach can mitigate gender bias amplification in the automatic generation of articles from Latin American newspapers.

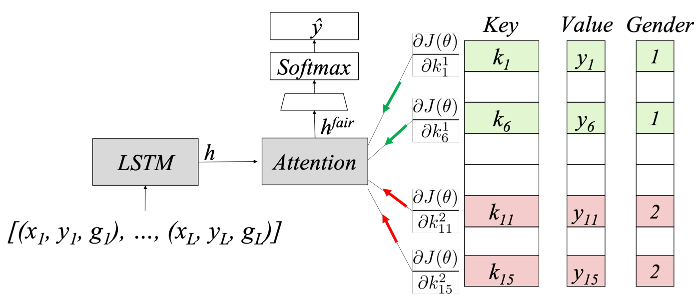

Figure 1: A Fair Region in memory M consists of the most similar keys to h given a uniform distribution over genders (e.g., 1: *male*, and 2: *female*). The input consists of a sequence tokens annotated with gender information, e.g., (*The*, 0), (*president*, 0), (*gave*, 0), (*her*, 2), (*speech*, 0).

1 Memory Networks and Fair Region

As illustrated in Figure 1, the memory M consists of arrays K and V that store *keys* (embeddings) and *values* (class labels), respectively as in Kaiser et al. (2017). We extend this model with an array G to store the *gender* associated to each word, (e.g., *actor* is *male*, *actress* is *female*, and *scientist* is *no-gender*). A neural encoder $f(x, \theta)$ with trainable parameters θ receives an observation x and generates the activations h in a hidden layer. We want to incorporate a normalized h to M. Hence, let i_{max} be the index of the most similar key ($i_{max} = argmax_i\{h \cdot K[i]\}$), then writing the triplet (x, y, g) to M consist of: $K[i_{max}] = \|h + K[i_{max}]\|$, $V[i_{max}] = y$, and $G[i_{max}] = g$. However, word embeddings are severely biased in natural language. For example, it has been shown that *man* is closer to *programmer* than *woman*, Bolukbasi et al. (2016). Similar problems have been recently observed in embedding algorithms such as Word2Vec, Glove, and BERT, Kurita (2019). For this purpose, we propose the use of a subset of keys to define a Fair Region in which we can control an equal ratio of gender types in such region. These keys will compute error signals and generate gradients that will flow through the entire architecture with backpropagation, as depicted in Figure 1. We define this region as follows.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

	Perplexity			Bias Amplification		
Model	All	Peru	Mexico	All	Peru-Mexico	Mexico-Peru
Seq2Seq	13.27	15.31	15.61	+0.18	+0.25	+0.21
Seq2Seq+Attention	**10.73**	13.25	14.08	+0.25	+0.32	+0.29
SeqSeq+FairRegion	10.79	**13.04**	**13.91**	**+0.09**	**+0.17**	**+0.15**

Table 1: Perplexity and Bias Amplification results on the datasets of crawled newspapers.

Definition 1.1. (Fair Region) Let h be an latent representation of the input and M be an external memory. The *male*-neighborhood of h is represented by the indices of the n-nearest keys to h in decreasing order and that share the same gender type *male* as $\{i_1^m, ..., i_k^m\} = KNN(h, n, male)$. Repeating the same process for each gender type estimates the indices i^f and i^{ng} for the *female* and *non-gender* neighborhoods. Then, the *FairRegion* of M given h consists of $K[i^m; i^f; i^{ng}]$.

2 Bias Amplification

Inspired by Zhao et al. (2017), we compute the bias score of a word x considering its word embedding $h^{fair}(x)$[1] and two gender indicators (words *man* and *woman*). For example, the bias score of *scientist* is: $b(scientist, man) = \frac{\left\|h^{fair}(scientist)\right\| \cdot \left\|h^{fair}(man)\right\|}{\|h^{fair}(scientist) \cdot h^{fair}(man) + h^{fair}(scientist) \cdot h^{fair}(woman)\|}$. If the bias score during testing is greater than the one during training, $b^{test}(scientist, man) - b^{train}(scientist, man) > 0$, then the bias of *man* towards *scientist* has been amplified by the model while learning such representation, given training and testing datasets similarly distributed.

3 Experiments

Figure 1 illustrates our proposed model (**Seq2Seq+FairRegion**), which we use as the LSTM decoder of a Seq2Seq architecture Sutskever, Vinyals, and Le (2014). This decoder attends to a Fair Region of 9 entries (3 for each gender type) given Definition 1.1. We crawl the websites of three newspapers from Chile, Peru, and Mexico to collect a working dataset. To enable a fair comparison, we limit the number of articles to 20,000 for each domain and a vocabulary of 18,000 most common words. Datasets are split into 60%, 20%, and 20% for training, validation, and testing. We want to see if there are correlations showing stereotypes across different nations. *Does the biased correlations learned by an encoder transfer to the decoder considering word sequences from different countries?* We compare our approach with these baseline models: 1) **Seq2Seq** Sutskever, Vinyals, and Le (2014): An encoder-decoder architecture that maps between sequences. 2) **Seq2Seq+Attention** Bahdanau, Cho, and Bengio (2015): A Seq2Seq that attends to parts of the input to predict the target word.

3.1 Fair Region Results in Similar Perplexity

We evaluate all the models with test *perplexity*, which is the exponential of the loss. We report in Table 1 the average perplexity of the aggregated dataset from Peru, Mexico, and Chile, and also from specific countries. Our main finding is that our approach (Seq2Seq+FairRegion) shows similar perplexity values (10.79) than the Seq2Seq+Attention baseline model (10.73) when generating word sequences despite using the Fair Region strategy. These results encourage the use of a controlled region as an automatic technique that maintains the efficacy of generating text. We observe a larger perplexity for country-based datasets, likely because of their smaller training datasets.

3.2 Fair Region Controls Bias Amplification

We compute the *bias amplification* metric for all models, as defined in Section 2, to study the effect of amplifying potential bias in text for different language generation models. Table 1 shows that using Fair Regions is the most effective method to mitigate bias amplification when combining all the datasets (+0.09). Instead, both Seq2Seq (+0.18) and Seq2Seq+Attention (+0.25) amplify gender bias

[1]For Seq2Seq neural models, this word embedding is the output of the decoder component $h^{deco}(x)$

for the same corpus. Interestingly, feeding the encoders with news articles from different countries decreases the advantage of using a Fair Region and also amplifies more bias across all the models. In fact, training the encoder with news from Peru has, in general, a larger bias amplification than training it with news from Mexico. This could have many implications and be a product of the writing style or transferred social bias across different countries. We take its world-wide study as future work.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. NIPS'16, 4356–4364.

Kaiser, L.; Nachum, O.; Roy, A.; and Bengio, S. 2017. Learning to remember rare events.

Kurita, K. 2019. Acl web. 166–172.

Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1400–1409. Austin, Texas: Association for Computational Linguistics.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 3104–3112.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.