Exploring Debiasing Techniques in Neural Machine Translation

Arun Govind M
arungm@iisc.ac.in

Ponsuganth Ilangovan
ponsuganthp@iisc.ac.in

Abstract

Neural machine translation has significantly pushed forward the quality of the field. However, fairness in machine translation is a area that has r. Neural models are trained on large text corpora which contains biases and stereotypes. As a consequence, models inherit these social biases. Recent methods have shown results in reducing gender bias in other natural language processing applications such as word embeddings.We take advantage of the fact that word embeddings are used in neural machine translation to propose the first debiased machine translation system. Specifically, we propose, experiment and analyze the integration of two debiasing techniques over GloVe embeddings in the Transformer translation architecture. We evaluate our proposed system on a generic English-German task, showing gains up to one BLEU point. As for the gender bias evaluation, we generate a test set of occupations and we show that our proposed system learns to equalize existing biases from the baseline system.

1 Introduction

Neural machine translation (NMT) is a recent approach in MT which learns patterns between source and target language corpora to produce text translations using deep neural networks (Sutskever et al., 2014). The downside to training models trained on parallel corpora is that social biases present in training data are also learned by the model. (Bolukbasi et al., 2016) showed that word embeddings learn gender bias from the training corpus. This bias not only propagates to downstream applications (Zhao et al., 2018) but also amplify the bias (Zhao et al., 2017). Thus fairness in AI is more important than ever before as models get integrated into real-world systems in every day life. Touching on the wider implications of the gender translation issue, (Vanmassenhove et al., 2018) said recent research has shown that neural models do not just reflect controversial societal asymmetries but exaggerate them.

The objective of this work is to study gender bias in MT and how NMT perpetuates biases in word embeddings. We also hope to give insights on the effectiveness of debiasing pre-trained embeddings towards mitigating this bias.

2 Related Work

2.1 Embedding layer initialization

While pre-trained word embeddings have been successfully employed to improve the performance of many downstream NLP tasks, their utility in NMT has not been explored extensively. (Qi et al., 2018) show that embeddings can be surprisingly effective in some cases providing gains of up to 20 BLEU points in the most favorable setting. Initializing the embedding later with pre-trained embeddings has been shown to be effective particularly in low-resource settings. Moreover, using monolingual source embeddings on the encoder side shows much more significant impact on the performance than monolingual target embeddings (Qi et al., 2018). This indicates that the of the gain from pre-trained word embeddings is largely due to a better encoding of the source sentence. (Neishi et al., 2017) showed that initializing the embedding layer with pre-trained word embeddings learned from the parallel corpus alone leads to marginal improvements in BLEU score even in high resource settings.

2.2 Debiasing Word Embeddings

The dangers of learning from human generated corpora was first brought to light when (Bolukbasi et al., 2016) explored the presence of gender bias in word embeddings. Word vectors inherit biases from broader society in their geometry.

The extensive use of these biased word vectors as basic features, leads to the perpertuation of bias in downstream tasks. Word embeddings not only reflect such stereotypes but can also amplify them. This poses a significant risk and challenge for machine learning and its applications.In recent years, several debiasing techniques have been proposed to mitigate this bias. We explore two such debiasing methods

Debiaswe is a post processing technique proposed by (Bolukbasi et al., 2016) to remove gender bias from word embeddings. First, the gender subspace is identified and projection of gender neutral words on this subspace is computed. This component is then removed from the said vector rendering the vector orthogonal to the gender subspace. The gender subspace is defined by the first k principal components in PCA decomposition of gendered difference vectors. A set of gender specific pairs like $she - he$, $man - woman$, $he - she$ are used to find the gender subspace.

GN-Glove Instead of trying to mitigate the bias in the trained word embeddings, GN- Glove introduced by (Zhao et al., 2018) attempts to debias the embeddings while training itself. The gender neutral glove method attempts to change the loss function in vanilla glove (Pennington et al., 2014) and assumes the biases in gender are lost in training itself. In this embedding model, a word vector **w** consists of two parts $w = [w^{(a)}; w^{(b)}]$ where $w^{(a)} \in R^{d-k}, w^{(b)} \in R^k$ stand for neutralized and gendered components respectively, where k is the number of dimensions reserved for gender information. This gender neutralizing scheme is to reserve the gender feature, known as protected attribute into $w^{(g)}$. Therefore, the information encoded in $w^{(a)}$ is independent of gender influence. Also $v^g \in R^{d-k}$ denote the direction of gender in the embedding space. This is done by defining a set of gender specific words and gender neutral words from wordnet. Two terms are added for minimization in addition to the vanilla glove loss function. We minimize the negative euclidean distance between the gender component of word vectors for words in the male centric gender set and female centric gender set. The second term minimises the dot product of gender neutral component of the word vectors in the gender neutral set and the gender subspace $v^{(g)}$.

3 Method

3.1 Machine Translation

We train the neural machine translation model on the english-german parallel corpora with diferent types of initializations of word embeddings to study the bias in the model and data. We train on four kinds of settings.

Random initialization: First, for the baseline model , we train a transformer model end to end on 2 million sentence pairs with (Glorot and Bengio, 2010) uniform initialization for the embedding layer.

Glove embeddings: The source embedding layer was initialized with glove word vectors. These glove embeddings were trained from the parallel corpora itself and fed in to the encoder.

Deabiaswe: The word embeddings were trained with the method adopted from (Bolukbasi et al., 2016) on the parallel corpora itself and these embeddings were used to initialize the encoder side word embeddings.

GN-Glove: The word embeddings were trained with the method adopted from (Zhao et al., 2018) on the parallel corpora itself and these embeddings were used to initialize the encoder side word embeddings.

3.2 Effect of Training data size

(Qi et al., 2018) studied the impact of initialization with pre-trained embeddings for low-resource and high-resource scenarios. We repeat the aforementioned experiments on one low-resource and one high-resource setting each. For the low-resource setting, we down-sample the 2 million parallel corpus to 1/20 it's size i.e, 100k sentences. This is to validate the interesting implications regarding available data size and the efficacy of initialization with pre-trained word embeddings.

4 Datasets

The dataset used was the common crawl English German parallel corpora consisting of 2 million sentences and the vocabulary size was set to 50000. The vocabulary was formed based on the

frequency count and less frequent words were replaced with unk tokens. The validation and test sets used are newstest2012 and newstest2013 from the Workshop on Machine Translation.

To measure the bias in occupation among genders, we formulate a method of measuring the bias.

5 Metrics

Firstly we use the quality of translation using BLEU score (Papineni et al., 2002). To quantify the bias, we use count the gender of the translated possessive pronoun in sentences of the format "I called my <occupation>".
German is an inflectional language. Inflection is a process of word formation, where the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender and mood. For the phrase *"my <occupation>"*, *<occupation>* is in *"my"* is a possessive pronoun which needs to be inflected to be in accordance with the gender of the word *<occupation >*.

Ideally, *my* would get translated to *mein* the neutral accusativ case for possessive pronoun. Any deviation from *mein* is a case of bias. The metric, therefore,is the percentage occurrence of *meinen*(male) and *meine*(female).

Case	Masculine	Feminine	Neutral
Nominative	mien	meine	mein
Accusative	**meinen**	**meine**	**mein**
Dative	meinem	meiner	meinem

Table 1: German Grammar

6 Baselines

In our baseline model, we train a standard transformer model with random initialization of the embedding layer. We quantify the bias in translation as mentioned above. Additionally, to evaluate the effectiveness of debiasing, we train the same transformer model with Glove initialization. Bias reflected in the translation of these two models forms the baseline. The baselines, remain the same for both high and low resource scenarios.

7 Experiments

For our experiments, we use a standard transformer model with multiplicative attention (Luong et al., 2015) with a beam size of 5 implemented in OpenNMT with pytorch (Klein et al., 2017). Training uses a batch size of 1024 with a dropout probability of 0.1 applied to LSTM stacks and Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 2, with noam decay method as is used for transformers (Vaswani et al., 2017). We evaluate the models performance using BLEU metric (Papineni et al., 2002).

8 Results

	random	GLove	**debiaswe**	gn-glove
Neutral	9.95	14.93	**30.32**	31.22
Female	26.7	31.22	**29.86**	71.04
Male	63.35	54.3	**32.58**	46.15

Table 2: Low Resource setting

	random	Glove	debiaswe	gn-glove
Neutral	11.29	8.38	11.63	10.52
Female	28.06	28.38	28.66	39.38
Male	60.65	63.25	59.71	50.11

Table 3: High Resource setting

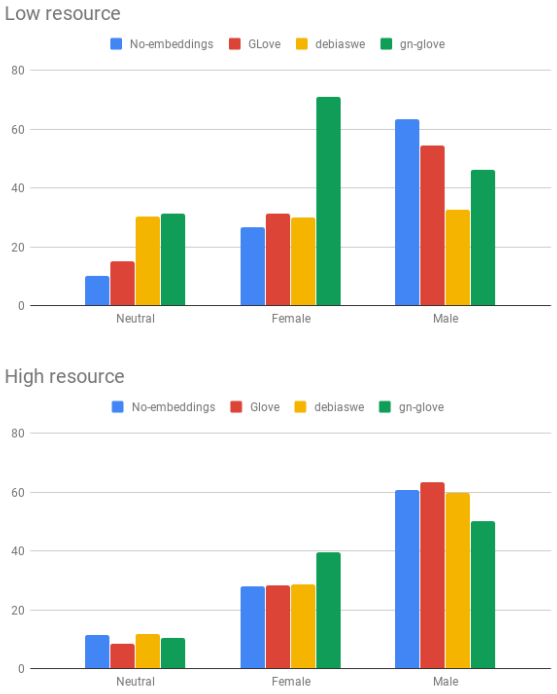

Figure 1: Gender Bias Results from Translation

	Male	Female
Count from BLS	140	59
Count matched	108	20
Percentage	77.14	33.89

Table 4: Male and Female dominated occupation counts from Bereau of Labor Statistics

9 Discussion

Firstly, we see that initializing the embedding layer with pre-trained embeddings has little to no effect on translation in the high-resource setting. While we were not able to see any improvement in the BLEU score in the low-resource setting due to initialization, we can see that BLEU score doesn't degrade significantly.

The baseline model with random initialization can be interpreted as the bias the model has learnt from the dataset. To validate this, we check the actual dominant gender of employees in those occupations that show gender bias in our translation. The results are shown in Table 4 This data is obtained from Bureau of labor statistics (US Dept. of Labor)[1]

We see that initializing with GLOVE alone leads to a slight improvement in the bias statistics. Initializing with Debiaswe embeddings show a dramatic improvement in the bias statistics and is our best model. GN-Glove embeddings, however, tend to drastically increase the ratio of female pronouns. From Figure 1 we see that while the the bias in fact decreases.

10 Future Work

One direction, we find worth exploring, is cross-lingual training of word embeddings. These models train their embeddings on a parallel corpus and optimize a cross-lingual constraint between embeddings in different languages that encourages embeddings of similar words to be close to each other in a shared vector space.

Additionally, effect of initialization with contextualized word vectors (Peters et al., 2018) could be studied in low and high resource setting.

Further, analysing groups of occupations that show gender bias would be interesting. ELMO? use bilingual word embeddings. why GN glove ?

[1] https://www.bls.gov/

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv e-prints*, page arXiv:1607.06520.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *arXiv e-prints*, page arXiv:1508.04025.

Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 99–109.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (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. *arXiv e-prints*, page arXiv:1802.05365.

Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. 2018. When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation? *arXiv e-prints*, page arXiv:1804.06323.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv e-prints*, page arXiv:1409.3215.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine

translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762.

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. *arXiv e-prints*, page arXiv:1707.09457.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. *arXiv e-prints*, page arXiv:1809.01496.