Towards Reducing Gender Bias in South Asian Language Translations

Nazmul Hasan Wanjan
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
nazmul.hasan.wanjan@g.bracu.ac.bd

Basit Hussain
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
basit.hussain@g.bracu.ac.bd

Malika Muradi
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
malika.muradi@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@gmail.com

MD. Humaion Kabir Mehedi
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

MD. Mustakin Alam
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

***Abstract*—While we have come very far in reducing gender bias in the society by taking steps towards that for a long time, works on reducing such bias from Natural Language Processing tasks are fairly new. In this paper, we have tried to add to the existing works on that topic by looking into the gender bias that is observed during Machine Translation, one of the most notable applications of Natural Language Processing, and have tried to reduce that with a few pre existing methods. The purpose of this paper therefore is to help the current machine translation systems reach a better state with more accurate prediction of the subject's gender with lesser dependency on other factors (mostly occupation) that are often unrelated.**

***Index Terms*—Gender Bias, Machine Translation, Natural Language Processing, Bias**

I. Introduction

Machine Learning fairness is a relatively new area which studies how to reduce favoring any specific segment of the population in machine learning models that may be the consequence of data or model inaccuracies. The end goal is to ensure equality for every segment of the population when they interact with the systems developed using machine learning. Older systems suffer from the problems of wrongly identifying segments of the population either due to the algorithmic or the dataset biases that are presented to it during the time of training. These biases, either coming willingly or unwillingly, hamper the modeling process and make it less generalized and usable for the whole population it is designed for. Additionally, it runs the risk of creating conflict among individuals in the real world as these models are often designed with the goal of being used in the real world systems.

Of the topics that concern machine learning fairness, bias reduction is a notable one. It is bias that leads us to stereotyping demographic groups, suppressing some and even leading the individuals within those demographic groups to believe certain aspects about themselves which are often misleading and wrong. Bias is an umbrella term to define the pre existing beliefs or shortcuts that may give a certain segment of the population some sort of edge over others. This has many types, one being gender bias (which is closely related to the topic of discussion of this paper), that comes from the beliefs about certain genders that exist in the society. For instance, historically, the majority of the existing cultures have been patriarchal. This means that the chief earning member is the father and he is also the head decision maker. Due to that, women are viewed as the segment who are weak and not suited to do physically and mentally challenging jobs. These pre-existing and dated thoughts, although useful in some situations in the past, are getting more and more inaccurate in the modern world where women, as well as other segments of the population, are joining the workforce for these demanding jobs.

In this work, we have tried to peer into the gender bias that may arise from the faulty and incomplete data and algorithms used for training the machine translation models, which is one of the most popular areas of Natural Language Processing. We specifically look into three of the South Asian languages: Bengali, Dari, and Urdu, and measure the gender bias that may occur when translating to and from English. Then we try to reduce the bias using different existing methodologies and compare the results to check which one performs the best for

this particular task.

It has to be noted, however, that although we have used only three languages for this work, it is nowhere near the number of languages used in the Southern part of the Asian subcontinent. Therefore, while this generalization may work for some related languages in the same family, it may not be suitable for all.

II. Literature Review

Previous works done by the researchers have found significant bias in the existing machine translation systems. Machine translation, as we know now, is not a single faceted task. One of the most crucial sub-tasks of it is coreference resolution where prevalence of gender bias has been observed in three of the existing systems each of which use one of the 3 machine learning paradigms by [1] using their Winogender schemas. Another approach by [2] introduces WinoBias benchmark that can be used for the task of coreference resolution, specifically focused on gender bias. Here, the authors used sentences with Winograd schema style and demonstrate better result in coreference resolution. One work by [3] mentions how the faulty word embeddings used for machine translation causes bias in the system and tries to use debiasing approaches to have a better model.

Reviews on the bias reduction method are useful to get the whole picture of the current landscape, scope and state-of-the-art methods of the process. For instance, [4] analyzes the gender bias in NLP systems across multiple languages by using Wikipedia Corpora across languages and extends Professional and Corpora Level Gender Bias Metrics. [5] summarizes the current conceptualizations of bias, the previous works aimed at assessing the gender bias in Machine translation and mitigation strategies so far by looking at different existing benchmarks.

With the view to shed further light on the phenomenon of Machine Bias and raising awareness about updating the current automatic translation tools using debiasing techniques, [6] collects a comprehensive list of job positions from the BLS dataset provided by the U.S. Bureau of Labor Statistics. They then build sentences using a general pattern to backtranslate using Google Translate API. The authors found a strong tendency of the system towards male default which leads to inaccurate representation of the real world statistics, with a more pronounced inaccuracy in case of STEM jobs. [4] works towards developing a new method to measure gender bias in 9 gendered languages [7] including English. They perform this by extending [4]'s method that utilizes word embedding systems like Word2Vec [8] to measure gender bias.

Among many reasons, popular dynamic pre-trained word embeddings are shown to be responsible for augmenting bias in natural language tasks [3] [14]. This bias present in the word embeddings come from the corpus collected from the sources which promote the androcentric view in the society, and hence in the data [15]. The level of bias in the pre-trained word embeddings vary from one to the next [9]. [10] tries to look into the behavior of word embedding on the modification of the training corpus. While the existance of these biases have been helpful in analyzing cross-cultural disparities [11], with the increasing applications of these models in our daily life decisions, the technologies using them may, unknownst to the users and developers, spread social unfairness [12].

One of the mitigation methods therefore is to use a debiased word embedding which removes the gender stereotypes in the word association and still keeps the desired associations [16]. The authors of [9] mention one way to reduce the bias from the word embedding is to trace the original document for the corpus, identify the subsets that cause the bias to take place in the first place, and remove them. Another similar approach by [13] finds and omits the words that are responsible for linking the attribute words to the target words. Training with the newly formed corpus gives a significantly better result in Word2Vec and GloVe models without any effect on the capturing of the semantic information. Another method by [17] suggests counterfactual data augmentation by swapping the gender in the sentences to increase the datapoints for the minor groups.

While all the previous methods focus on removing the bias from the high resource language pairs, works on low resource ones are significantly less. This is especially challenging because getting the gender right in these translations are generally not the first priority. One early work by [18] successfully obtained a maximum of 1.96 BLEU improvement in case of low resource language pair in Neural Machine Translation using monolingual corpora. Also, different counterfactual methods mentioned by [19] can help in augmenting the data for the low resource language pairs before following the augmentation of datapoints for the minor genders.

III. Workflow

- Gathering the job position data from different job sites in 3 countries where the selected languages are spoken (Bn, Da, and Ur)
- If the dataset is small, augment with the BLS dataset using backtranslation
- Adding texts with the occupation to make full sentences
  - without context,
  - with context
- Check the no. of instances the predictions match
- Compare with the real world stats. Apply different bias reduction techniques on the generated sentences Check the results for each using different benchmarks

References

[1] Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. (2018). Gender bias in coreference resolution. arXiv preprint arXiv:1804.09301.

[2] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

[3] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

[4] Matthews, A., Grasso, I., Mahoney, C., Chen, Y., Wali, E., Middleton, T. , Njie, M., and Matthews, J. 2021. Gender Bias in Natural Language Processing Across Human Languages. In Proceedings of the First Workshop on Trustworthy Natural Language Processing, pages 45–54, Online. Association for Computational Linguistics.

[5] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, Marco Turchi; Gender Bias in Machine Translation. Transactions of the Association for Computational Linguistics 2021; 9 845–874.

[6] Prates, M.O.R., Avelar, P.H. Lamb, L.C. Assessing gender bias in machine translation: a case study with Google Translate. Neural Comput Applic 32, 6363–6381 (2020).

[7] Eckert, P., McConnell-Ginet, S. (2013). Language and gender. Cambridge University Press.

[8] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[9] Emeralda Sesari, Max Hort, and Federica Sarro. 2022. An Empirical Study on the Fairness of Pre-trained Word Embeddings. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 129–144, Seattle, Washington. Association for Computational Linguistics.

[10] Brunet, M., Alkalay-Houlihan, C., Anderson, A. amp; Zemel, R.. (2019). Understanding the Origins of Bias in Word Embeddings. ¡i¿Proceedings of the 36th International Conference on Machine Learning¡/i¿, in ¡i¿Proceedings of Machine Learning Research¡/i¿ 97:803-811

[11] Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating Word Embedding Gender Biases to Gender Gaps: A Cross-Cultural Analysis. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 18–24, Florence, Italy. Association for Computational Linguistics.

[12] Aylin Caliskan et al. ,Semantics derived automatically from language corpora contain human-like biases.Science356, 183-186(2017).DOI:10.1126/science.aal4230

[13] Singla, S. Machine Unlearning Human Biases: Inclusive Word Embeddings by Excluding Biased Texts.

[14] Li, A., Bamler, R. (2020). Quantifying Gender Bias Over Time Using Dynamic Word Embeddings.

[15] Petreski, D., Hashim, I.C. Word embeddings are biased. But whose bias are they reflecting?. AI Soc (2022). https://doi.org/10.1007/s00146-022-01443-w

[16] Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

[17] Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A. (2020). Gender bias in neural natural language processing. Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday, 189-202.

[18] Gülçehre, Ç., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., Bengio, Y. (2015). On Using Monolingual Corpora in Neural Machine Translation. ArXiv, abs/1503.03535.

[19] Maimaiti, M, Liu, Y, Luan, H, Sun, M. Data augmentation for low-resource languages NMT guided by constrained sampling. Int J Intell Syst. 2022; 37: 30- 51. https://doi.org/10.1002/int.22616