

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@bracu.ac.bd*

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—Gender bias in neural machine translation is not a new topic. However, works on reducing such bias from Natural Language Processing tasks are fairly new. In this paper, we have tried to analyze the gender bias that is observed during Machine Translation, one of the most notable applications of Natural Language Processing, of three of the South Asian languages and have tried to reduce that by one of the pre-existing methods, specifically by using more data for the biased segment of the population. 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 existence 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 data points for the minor genders.

One important observation is that, among all the existing works, we have not found any that works with gender bias in South Asian languages. Therefore, ours is possibly the start of the analysis an important topic that needs more attention.

III. EXPERIMENTS

To begin with our approach, we divided our works into three experiments in the first two of which we first confirm our claim that gender bias is indeed present in the current neural machine translation systems and later we try to implement our own model to counter the problem. The next subsections explain the experimental setup from data collection to model building in detail.

A. Analyzing the bias of other systems for single sentences

For our first experiment, we follow [4]’s method by adding having more data on each of the occupations and then analyzing the existing bias. However, knowing the countries where our focused languages are spoken may have a different occupational distribution, we first made our own dataset for the occupations by using the authorized government and private sites. The steps that we have followed for collecting the occupations are given in Fig. 1. The sites used for this step are Bangladesh Bureau

of Statistics¹, Pakistan Bureau of Statistics² and ACBAR, the most popular job site in Afghanistan³.

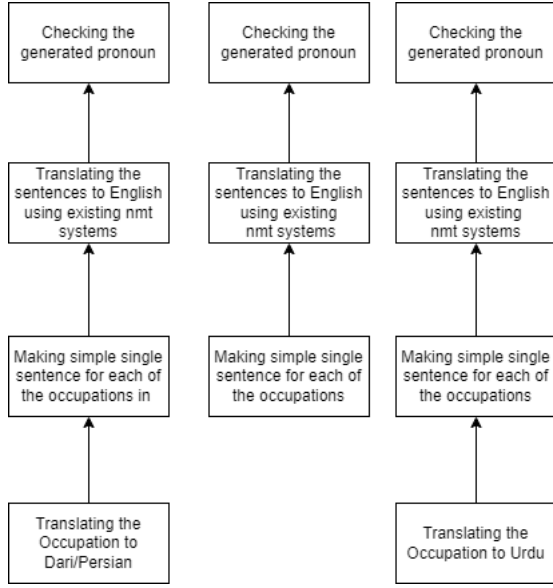


Fig. 1. Steps followed in the first experiment

For the occupations in Bangladesh, we skipped step 1 as the used document for collecting the data already provided the translation. We use two translation systems for each of the target languages. The translation systems used for Bengali and Urdu were Google Translate and Systran, while the ones used for Dari/Afghan Persian were Google Translate and Microsoft Translator. One factor to be noted here is that, as of May, 2023 Google translate does not provide translation for Dari. However, because of its similarity with Persian, Dari is often called "Afghan Persian" in the literature, and is still commonly called Farsi [20], and that is why for the translations using Google translate, we chose the Persian translations. Additionally, as confirmed by one of our authors, who is an Afghan native, the little difference observed between Iranian Persian and Afghan Persian due to the cultural differences does not affect any of the translations for our task.

Also, we observed that each of the languages provide alternate translations, which we also analyze those in both this and the next task.

B. Analyzing the bias other systems for multiple sentences

For our second experiment, we checked the gender bias present in the selected system on a document level. Here, we used texts where the number of sentences are two. To get the accurate picture, we clearly mention the gender of the subject in the first sentence, followed by the same single sentence used in analyzing the single sentence translations. The steps followed in this experiment are given in Fig. 2. All the sentences used in this experiment were basic ones as well. Here, the first

sentence contained the sentence "I met that woman yesterday." Followed by the sentence used in the first experiment. Our expectation in this experiment was to observe an increased count of feminine pronoun in the sentences translated by the mentioned neural machine translation systems.

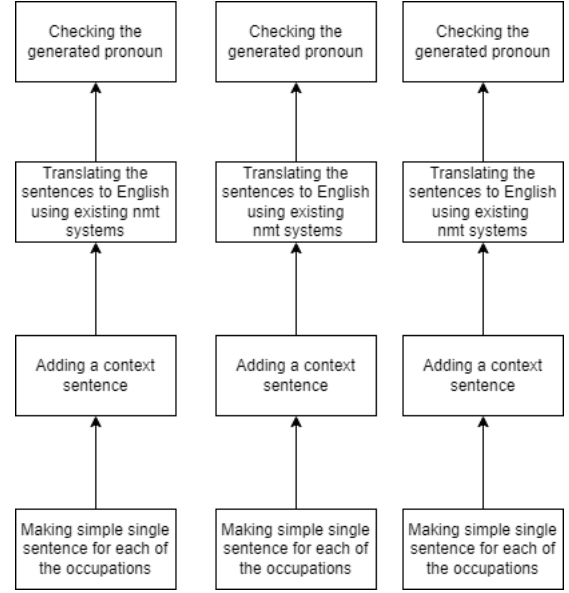


Fig. 2. Steps followed in the second experiment

As mentioned in the previous subsection, we observe that google translate provides alternate translation for each of the translations and for each of the sentences. Ignoring the translation errors that may be present in the system, we only analyzed the alternate translation for the second sentence in addition to the default ones.

C. Analyzing the bias in our model

For this experiment, we first built our own model and then checked how well it works in reducing the bias that was observed in the previous experiments. The detailed steps for this whole experiment has been depicted in Fig. 3. Here, we use the same sentences as before without any additional processing.

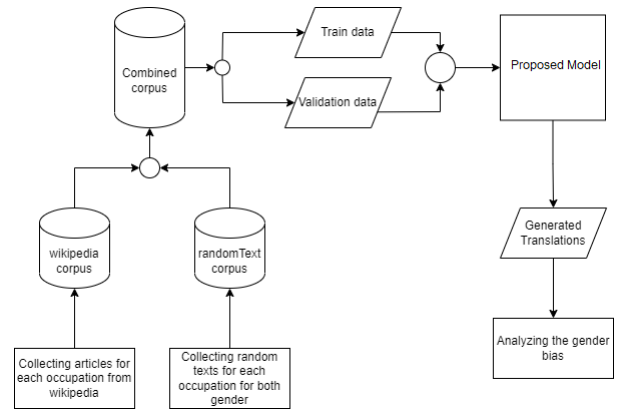


Fig. 3. Steps followed in the third experiment

¹ <http://www.bbs.gov.bd/>

² <https://www.pbs.gov.pk/>

³ <https://www.acbar.org/>

Our workflow in this step can be divided into three subtasks,

- 1 Building the multilingual parallel corpus
- 2 Splitting the corpus into train and validation sets
- 3 Building the model
- 4 Analyzing the gender bias

each of which we briefly explain next.

1) *Building the multilingual parallel corpus*: As the focus of this work is not to get accurate translation, we detained from using any large dataset or parallel corpus that are publicly available for each of our target languages. Rather, we chose to use the available English⁴, Persian⁵, Urdu⁶ and Bengali⁷ articles on each of the occupations from Wikipedia. We then translate each of the articles in one language to the other languages, giving priority to the English articles. Here, we collected about 450 articles in 4 languages.

Then, with the goal of balancing the training dataset, we collected random texts using artificial text generation tool⁸, where for each occupation, we generated the texts by providing the occupation title, the gender mention and the pronoun. To obtain an equal distribution in each occupation for both our analyzed genders, we generated the texts twice, changing the pronoun and the gender mention. We only chose the relevant portions with additional context. So, each of the corresponding texts was 5-6 sentences long even though the actual texts were over 3 paragraphs long in most cases.

Finally, we combined all the Wikipedia articles and the random texts which, after careful segmentation were converted into the multilingual corpus which contains 152, 122 lines of text. We used this corpus to train and validate our model.

2) *Splitting the corpus into train and validation sets*: This step is self explanatory. In this step, we split the corpus into test and validation sets, keeping 0.85 portion for training.

3) *Building the model*: For our own model of gender bias free translation system, to avoid repetition, we focused on the translation to only one of the target languages, which is Bengali. Here we used the encoder-decoder architecture, both of which uses Recurrent Neural Network in the input and output layer. Further details have been mentioned in Fig. 4.

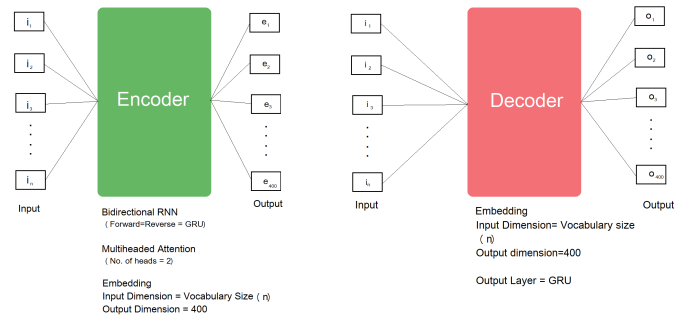


Fig. 4. Model Architecture

The training and validation accuracy and loss of our model with 40 epochs has been shown in Fig. 5. In fact, with 61.006% validation accuracy, this is the best result we could get. Table I shows two of the other results during the experiment. In all the cases, we kept early stopping as the callback function.

In choosing the final model, rather than focusing on building a perfect translation system, we focused on getting a workable model so that the gender and occupational information are properly translated. Therefore, we could not utilize some of the other models we came up with by changing the hyperparameters which gave better performance in terms of accuracy and loss in the validation set.

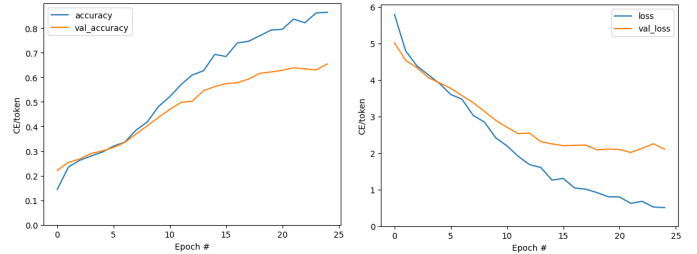


Fig. 5. Accuracy and Loss Progression with Epoch=40

Epochs	Steps/epoch	val steps	Patience	Accuracy	Loss
20	50	50	2	0.4837	0.3812
50	100	40	3	0.5891	0.3308
40	130	20	3	0.61006	0.2253

TABLE I
GENDER DISTRIBUTION RESULTS FOR SINGLE SENTENCE TRANSLATIONS TO URDU

4) *Analyzing the gender bias*: After building our model, we move on to analyzing the gender bias using that. Following the previous experiments, we use the same single and multiple sentences translations that we built earlier to evaluate and later compare the results with that of the state-of-the-art neural machine translation systems.

IV. RESULTS AND DISCUSSION

A. Analyzing the bias other systems for single sentences

Looking at the performance of other systems in this step by checking the gender distribution, we find that all our selected

⁴<https://en.wikipedia.org/>

⁵<https://fa.wikipedia.org/>

⁶<https://ur.wikipedia.org/>

⁷<https://bn.wikipedia.org/>

⁸Anyword: <https://anyword.com/>

systems perform poorly in both single sentence and multi sentences translation. The results of the analysis are given in table II, III and IV. Here, we added the alternate Google translation for only Persian because it is the only language we could find that provides multiple gender-specific translations. The alternate translations of the other languages give the rephrasals of the same sentence, which is not the topic of our discussion.

	Female	Male	Others
GTranslate	347	177	46
SYSTRAN	437	71	62

TABLE II

GENDER DISTRIBUTION RESULTS FOR SINGLE SENTENCE TRANSLATIONS TO URDU

	Female	Male	Others
GTranslate (default)	254	33	1
GTranslate (alternate)	23	265	0
Microsoft Translator	265	23	0

TABLE III

GENDER DISTRIBUTION RESULTS FOR SINGLE SENTENCE TRANSLATIONS TO PERSIAN/DARI

	Female	Male	Others
GTranslate (default)	599	23	0
SYSTRAN	563	58	1

TABLE IV

GENDER DISTRIBUTION RESULTS FOR SINGLE SENTENCE TRANSLATIONS TO BENGALI

One important factor to note here is that, although we label all translations that contain neither male nor female pronouns, we found that none of those translations contain other pronouns (eg. they [21]) that are newly emerging to refer to the non-binary gender groups. Therefore, the "Others" labels may be taken as mistranslations where the pronouns of the subjects are not properly recorded.

B. Analyzing the bias other systems for multiple sentences

In this step, we look at the bias of the selected neural machine translation systems on a document level. The results have been shown in tables V, VI and VII.

Here, we see that in most cases, the result improves with additional context in the first sentence. However, the bias still remains as for all the test sentences we clearly mention that the subject is female. In fact, the English translation for the context sentence that is used to make all the multi sentences is "I met that woman yesterday".

	Female	Male	Others
GTranslate (Default)	520	33	17
GTranslate (Alternate)	347	169	14
SYSTRAN	417	80	73

TABLE V

GENDER DISTRIBUTION RESULTS FOR MULTIPLE SENTENCES TRANSLATIONS TO URDU

	Female	Male	Others
GTranslate (default)	254	34	1
GTranslate (alternate)	23	265	0
Microsoft Translator	265	23	0

TABLE VI

GENDER DISTRIBUTION RESULTS FOR MULTIPLE SENTENCES TRANSLATIONS TO PERSIAN/DARI

	Female	Male	Others
GTranslate (default)	595	27	0
GTranslate (alternate)	542	80	0
SYSTRAN	563	58	1

TABLE VII

GENDER DISTRIBUTION RESULTS FOR MULTIPLE SENTENCES TRANSLATIONS TO BENGALI

C. Analyzing the bias in our model

As mentioned in the previous section, we used the same sentences to analyze our system that was used for others. The pictorial view of our result has been given in Fig. 6 and Table 6. Although we performed the experiment on one language, with a few modifications, the model can be implemented for other languages.

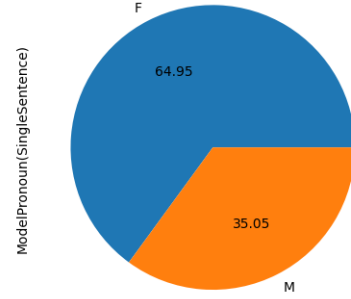


Fig. 6. Model Performance for single sentence translation

Female	Male	Others
404	218	0

TABLE VIII

GENDER DISTRIBUTION RESULTS FOR SINGLE SENTENCES TRANSLATIONS BY OUR PROPOSED MODEL TO BENGALI

It can be observed that although our model is still biased, the male segment of the overall population has become the minority here. One reason for this behavior that we observed is that, in the Wikipedia articles, the female workers have clear mention of gender close to their occupation, which tends to be absent for the male population. It may be because of the acceptance of male defaults on the society that are reflected by the authors of those articles. Even so, it confirms that having more data on the biased segment of a population gives a better distribution in case of single sentences.

V. FUTURE DIRECTIONS

Although in the previous section we saw that our model performs really well in single sentence translations, when we

performed multiple sentence translations, the system failed miserably. That is why we have not included it in the final result, keeping it as the topic to be explored in the future where we try to augment the capability of the current model by providing gender information based on the context sentence.

Also, our model only two genders like most other ones which, in the context of the current society, is obsolete. Therefore, we will try to include the non-binary groups in our future works.

VI. CONCLUSION

In this paper, we have tried to look into the existing gender bias in the state-of-the-art neural machine translation systems and tried to implement the already mentioned methods to mitigate the problem. Although we could successfully reduce the gender bias on a sentence-level translation, we have yet to do the same for document-level ones, which we kept for our later work. Still, with the increasing amount of people gaining interest in the South Asian countries and their culture, we hope that this initial step will encourage future researchers to provide a bias free translation system for them.

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. and Zemel, R.. (2019). Understanding the Origins of Bias in Word Embeddings. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, 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.“ *Science* 356, 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>
- [20] Spooner, B. (2012). 4. Persian, Farsi, Dari, Tajiki: Language Names and Language Policies. In *Language Policy and Language Conflict in Afghanistan and Its Neighbors* (pp. 89–117). Brill.
- [21] LaScotte, D. K. (2016). Singular they: An empirical study of generic pronoun use. *American Speech*, 91(1), 62–80.