Machine translation performance within non-binary gender contexts

Author - Devanshi Garg, A69036540
d2garg@ucsd.edu

Abstract

Bias and fairness in Machine Translation are critical issues that have gained a lot of attention in the last few years. They perpetuate harmful stereotypes in the society due to the biased training data. While researchers have made efforts to mitigate these biases, the inclusion of non-binary genders remains a relatively unexplored area. Most existing techniques focus on binary gender distinctions, inadvertently reinforcing traditional gender norms. This project aims to address this limitation by developing a more inclusive MT model for Hindi-English translation. By combining data augmentation with adversarial learning, we seek to mitigate gender bias and promote a broader representation of gender identities.

1 Introduction

Large language models, pre-trained without supervision and then fine-tuned for specific applications, have become a dominant paradigm in natural language processing (NLP). However, these models exhibit bias and perpetuate harmful stereotypes. A lot of research has begun to examine how biases in NLP systems could be measured and reduced with a focus on binary gender representations. As society increasingly acknowledges diverse gender identities, it becomes essential to adapt these models to ensure inclusivity across all gender types. In our project, we analyzed the gender bias in the pre-trained Marian Machine Translation model and proposed an approach to mitigate it by integrating adversarial learning. The following steps were undertaken as part of this work:

- Used IIT Bombay English-Hindi Corpus and preprocessed it: DONE
- Enhanced the database to identify the gender inclination of sentences based on specific keywords, to train the adversarial network: DONE
- Identified issues in the baseline Marian MT Model: DONE
- Incorporated an adversarial network into the Marian MT model and fine-tuned it on our augmented dataset: DONE
- Evaluated the model and validated the result: DONE

2 Related work

Machine Translation models are state-of-the-art for machine translation, but these models are prone to various social biases, especially gender bias. As discussed by **Michela Menegatti** and **Monica Rubini** in **2017** in their paper **Gender Bias and Sexism in Language** (Menegatti and Rubini, 2017), language models subtly reproduce the societal asymmetries of status and power in favor of men, which are attached to the corresponding social roles. These asymmetries are particularly more prevalent in languages like Hindi where roles have different forms depending upon the gender being talked about. Although, this work depicts that gender-fair linguistic expressions can effectively prevent negative societal impact, the paper does not talk about ways to make the language inclusive for all gender types.

Mitigating Gender Bias in Machine Translation through Adversarial Learning by **Eve Fleisig** and **Christiane Fellbaum** (Fleisig and Fellbaum, 2022) solved the biasness issue in 2022 by presenting an adversarial learning framework. The paper talks about how data-centric approaches to mitigate bias are not a good idea due to the difficulty in collecting gender based data, specifically for the low level languages. Therefore,

an adversary was trained to predict the gender, while the model learned to prevent the adversary from predicting the gender. Binary projections were used to predict the gender, which lead to the problem of neglect towards non-binary genders.

In **2022**, the quality of bias mitigation was tried and tested by using 1) data augmentation to inform and support the main task, 2) adversarial techniques that actively discourage the model from learning the bias. With information from **Bias Mitigation in Machine Translation Quality Estimation** by **Hanna Behnke** and others (Behnke et al., 2022), we tried to incorporate both the techniques in our project.

Gender Inflected or Bias Inflicted: On Using Grammatical Gender Cues for Bias Evaluation in Machine Translation from **2023** by **Pushpdeep Singh** (Singh, 2023) evaluates various Neural Machine Translation models for their gender bias on low resource language Hindi. The paper talks about huge bias in the models predictions and hence it was concluded that it is important that NMT models can identify the correct gender based on the grammatical gender cues in the source sentence rather than just relying on biased correlations.

None of the above papers addressed the issue of inclusivity in gender predictions by machine translation models until the publication of **Beyond Binary Gender: Evaluating Gender-Inclusive Machine Translation with Ambiguous Attitude Words** by **Yijie Chen, Yijin Liu**, and others (Chen et al., 2024). This paper introduced a benchmark for evaluating gender inclusive translations with ambiguous attitude words. Their work focused on English-to-Chinese translation, where gender bias is not embedded in grammatical structures but is primarily present in nouns and pronouns. Therefore, this approach is not directly applicable to languages like Hindi.

In 2024, the case study **Post-Editing Machine Translation Beyond the Binary: Insights into Gender Bias and Screen Activity** by **Manuel Lardelli** (Lardelli, 2024) had a focus on on bias in the machine-translated outputs and on the gender-fair-language post-editing process. Findings from the analysis of the machine translations showed that DeepL systematically misgenders and erases non-binary identities. Also, with a high variability among participants in editing the Machine Translations utilising GFL approach, highlighted how there is no one-fits-all solution to non-binary representation. Therefore, this field needs a lot of research to come up with fairer translation systems.

3 Dataset and problem analysis

We have used the **IIT Bombay English-Hindi Corpus** (Kunchukuttan et al., 2018) which contains parallel corpus for English-Hindi as well as monolingual Hindi corpus collected from a variety of existing sources and corpora developed at the Center for Indian Language Technology, IIT Bombay over the years. This corpus was loaded as a dataframe via the HuggingFace Datasets repository. It has a total of 1,659,082 entries. The dataset consists of pair of english hindi sentences such as

The default plugin layout for the top panel: ऊपरी पटल के लिए डिफ़ॉल्ट प्लग-इन खाका

Our task was to analyse how the pretrained Machine translation models interpreted the gender during the translation **(Figure 1)**. For this, various Hindi sentences were inputted to the Marian Machine Translation Model. Some of the examples of input output are shown in the table below:

Input Hindi Sentence	Translated Sentence
वह एक डॉक्टर है	He's a doctor
वह एक नर्स है	She's a nurse
उसे फुटबॉल खेलना पसंद है	He likes playing football
विश्वामित्र और उनके परिजन	Sarah and her family

The above examples shows a clear bias in the data where the gender neutral term **वह** was translated as **He** in case of a doctor and **She** in case of a nurse. This is due to the skewed training data and this can lead to inappropriate or inaccurate translations, and addressing these biases will improve model fairness and usability, that is the aim of this project.

As discussed by Michela Menegatti and Monica Rubini, these biases arise from historical texts, biased human translations, and imbalanced

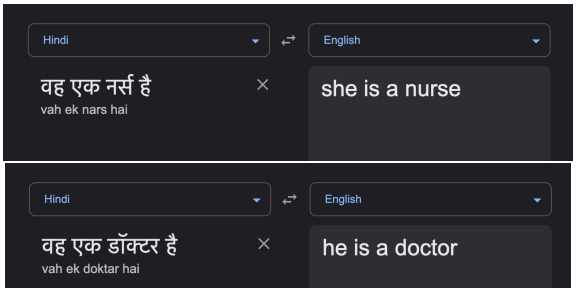

Figure 1: Example of gender bias in Hindi-English translation using Google Translate. The system translates ”nurse” to ”She” a female nurse, and ”doctor” to ”He” a male doctor.

linguistic representations. If we go deep into our training data, we see that the translations having terms ***”he”, ”his”, ”him”, ”man”, ”boy”, ”father”, ”brother”*** are ***132,433*** as compared to the translations having terms ***”she”, ”her”, ”woman”, ”girl”, ”mother”, ”sister”*** are ***15,394***, which is almost 1/10th of the former. Many languages like do not inherently encode gender in neutral terms like "वह". However, when translating to a language like English, which requires a gendered pronoun (he/she), the model is forced to choose. In the absence of additional context, it defaults to the statistically dominant gender learned from the data. Hence, it is also crucial to incorporate terms like ***”they”, ”them”, ”their”, ”non-binary”, ”queer”, ”gay”, ”lesbian”, ”transgender”, ”bisexual”, ”pansexual”, ”genderqueer”, ”agender”*** into the dataset. This allows the model to learn fairness not only between binary genders (male and female) but also among non-binary and LGBTQ+ identities. so that the model not only learns fairness amongst male/female genders but also makes it more inclusive. It equips the model to better handle gender-neutral and inclusive language, which is increasingly important in modern contexts.

3.1 Data Pre-processing

By adding adversarial layer for gender detection, we tried to mitigate gender bias in machine translation models. The pre-trained model Marian MT was fine-tuned by adding this adversarial network, the gradient reversal layer, and adjusting the training process accordingly. To do this, it was necessary to have gender labels for each translation to train the adversarial layer. Dataset was a JSON in a dictionary format such as

”en”: ”Hello”, ”hi”: "नमस्ते"

The keys (”en”, ”hi”) were extracted as column names and their corresponding values as rows in a new DataFrame.
Next, Hugging Face tokenizer of the Marian MT model was used for tokenizing Hindi and English sentences. These tokenized sentences were converted into tensors.

3.2 Data annotation

Gender labels for the dataset was defined to indicate whether a sentence is male-biased or female-biased. This required annotated data where the gender of sentences is known. Simple rules were used detect gender-specific terms (e.g., ”he,” ”she”, ”they”) in the target sentences. The gender labels were annotated as:

Label	Example Terms
1: Male Bias	”he”, ”his”, ”him”, ”man”, ”boy”, ”father”, ”brother”, etc.
2: Female Bias	”she”, ”her”, ”woman”, ”girl”, ”mother”, ”sister”, etc.
3: Non-binary	”they”, ”them”, ”their”, ”non-binary”, ”queer”, ”gay”, ”lesbian”, ”transgender”, ”bisexual”, ”pansexual”, ”genderqueer”, ”agender”
0: Others	No specific gender terms.

The biggest challenge encountered during data annotation was determining how to project the gender of a sentence. Due to limited resources, we had to rely on simpler methods, annotating based on terms within the sentence. However, this approach is unreliable for gender labeling. Terms within a sentence may hint at a particular gender, but other terms can introduce ambiguity. For example, if both ”mother” and ”father” appear in a sentence, our model may automatically label it with a female bias. Additionally, pronouns like ”they,” ”them,” and ”their” can refer to groups of people, not just non-binary individuals.

A more effective annotation approach would involve greater contextual understanding of the text. This could be achieved by having a team of human annotators manually verify machine-generated labels, ensuring that gender labels are accurate and promote fairness in language. This human oversight would allow for nuanced

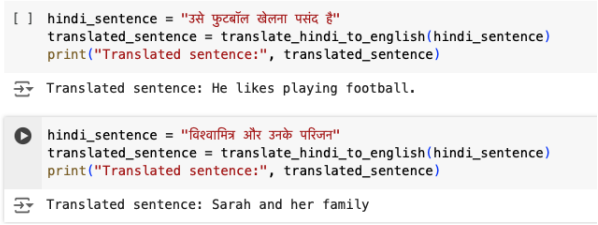

Figure 2: Biases in the Marian MT model reveal that football is often associated with masculinity, while family-related contexts are deemed more feminine, perpetuating gender norms and limiting inclusivity.

decisions that consider context, reducing bias and improving label consistency.

4 Baselines

A neural machine translation (NMT) model Marian which is developed by the Hugging Face team based on the Marian architecture (Junczys-Dowmunt et al., 2018) was used for this task. This model is trained to handle a wide variety of languages and language pairs efficiently, making them suitable for tasks that require translating between multiple languages. It is available for use via the Hugging Face transformers library for Hindi to English language pairs.

As an initial attempt, the pre-trained Marian MT model was tested with various Hindi sentences to investigate potential inbuilt gender bias. The experiment revealed that the model exhibits gender stereotypes based on certain occupations. For instance, the Hindi sentence "वह एक डॉक्टर है" (ideally, "He/She is a doctor") was translated as "He's a doctor", while the sentence "वह एक नर्स है" (ideally, "He/She is a nurse") was translated as "She's a nurse" **(Figure 2)**. This indicates a gender bias in the translation model, where traditionally male-dominated professions (like doctors) are associated with men, and traditionally female-dominated professions (like nurses) are associated with women. Such biases highlight the influence of societal stereotypes on language models and the need for addressing these biases to make machine translation systems more equitable and inclusive.

For mitigating bias in machine translation, adversarial network was implemented. Adversarial training was used to confuse the encoder of the pre-trained model forcing it to create neutral or

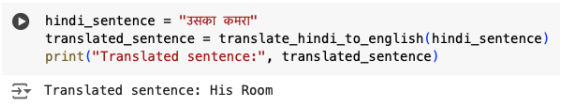

Figure 3: Room is always predicted as "His" because it is a masculine noun, despite the fact that who's room it is being referred to. It neglects the feminine and other non-binary genders.

non-biased representations. Translation loss was combined with the adversarial loss, regularizing the model to reduce bias while maintaining translation quality.

5 Implementation

5.1 Background of the language

Hindi is a gendered language where parts of speech such as nouns, pronouns and adjectives are different based on the gender they refer to. Not only that, it categorizes the verbs according to the gender context of the subject or object within the sentence. Some important points to note in Hindi Language are:

- For neutral genders, Hindi language uses plural forms, similar to how some expressions are handled in English. However, Marian MT model overlooks the representation of non-binary genders in its translation. The model either produces a binary translation (male or female) or incorrectly interprets the neutral input as plural, thereby failing to provide inclusive and accurate translations.
- In Hindi, even the inanimate nouns such as book and room are gendered. For instance, किताब (book) is feminine and कमरा (room) is masculine **(Figure 3)**. This characteristic poses challenges in accurately labeling data for adversarial training as it becomes difficult to generalize or infer gender rules without explicit labeling.
- The pronoun वह in Hindi can mean "he" "she" or "it" but verbs and adjectives clarify the gender. In Machine translation, this pronoun exhibits bias by relying on traditional stereotypes to determine the gender in the translated sentence.

5.2 Implementation Steps

Code consists of three parts, 1) baseline model which is the pre-trained Marian MT model, 2)

training loop for training the gender classifier model which will be later used as an adversary, 3) training loop for model with an adversary network. Baseline model has been discussed above, here, we will be talking more about the model to mitigate bias and make it more gender inclusive.

As mentioned in the Data annotation part, 4 different annotations were used to train the adversary. MT With Adversary model implements a Marian MT model with an additional adversarial gender classifier to detect and mitigate gender biases in the output embeddings of the MT model. First, a **gender classifier** is implemented which is a simple feed-forward neural network designed to classify embeddings into one of four gender-related categories: male, female, LGBTQ, or neutral. It takes the output from the hidden encoder state of the Marian MT model and outputs a 4-class probability distribution. The adversary is trained first on the embeddings produced by the MT model, without updating the MT model's parameters.

After the adversary is trained, it is used in the MT model during joint training, where both the MT loss and the adversarial loss are combined. During the forward pass in the main model, the adversary's parameters are not updated. Only the MT model's parameters are updated during training. This setup decouples the training of the MT model and the adversary, making it easier to train the adversary on its own data and fine-tune it for better performance on gender bias detection.

MT Model which is the base translation model gives the encoder output on the input sequence producing contextual embeddings. Embeddings are averaged across the sequence to produce a single vector per sample. In order to pass these contextual embeddings to the adversary, a projection layer was used to map the output of encoder to a fixed dimension to input it to the adversary network.

A custom PyTorch layer **Gradient Reversal Layer** is used for adversarial learning. In the forward pass, the GRL acts as an identity function, passing inputs unchanged.

$$x_{out} = x_{in}$$

In the backward pass, it reverses the gradients and scales them by a factor lambda.

$$\frac{\partial L}{\partial x_{in}} = -\lambda \cdot \frac{\partial L}{\partial x_{out}}$$

This reversal causes the model to update its weights in a way that makes it more difficult for the adversary to distinguish between the gender attributes it is trying to eliminate.

The embedding tensor is passed to the GRL layer which reverses the gradients during backpropogation. The adversary predicts gender labels from the reversed embeddings. The loss from this classifier is used to discourage the MT model from encoding gendered information. The embeddings from the encoder are also passed to the decoder for generating translations. Decoder outputs represent the final predicted translations.

The model is optimized to minimize the translation loss while maximizing adversarial loss. The adversarial loss helps the model minimize gender-specific information in embeddings while still performing translation accurately.

Two hyperparameters were tuned,

1. **Lambda:** This hyperparameter influences the impact of adversary on the model. Higher the lambda, more the emphasis on removing gender-specific information. This was set to 0.05.

2. **Hidden Dimension of Classifier:** This is the size of the projected embedding space for adversarial training. It was 512 for our model.

A python file namely, "256 Project.py" is submitted as part of this project. Since, the computations done in the entire setup are heavy due to two separate training loops for the Gender classifier to train the adversary and another training loop for training the integrated MT model with Adversary. I used the T4 GPU on colab for training the adversary on a dataset of 50,000 pairs of labeled data. For fine-tuning the translation model with the adversarial network, I used a dataset of 15,000 pairs. The model took around 30 minutes to train both the models.

5.3 Results

An adversarial gender classifier, trained independently, achieved a cross-entropy loss of **0.20**

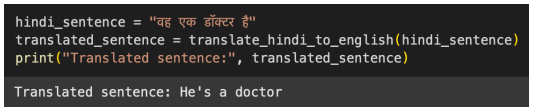

Figure 4: Baseline Output

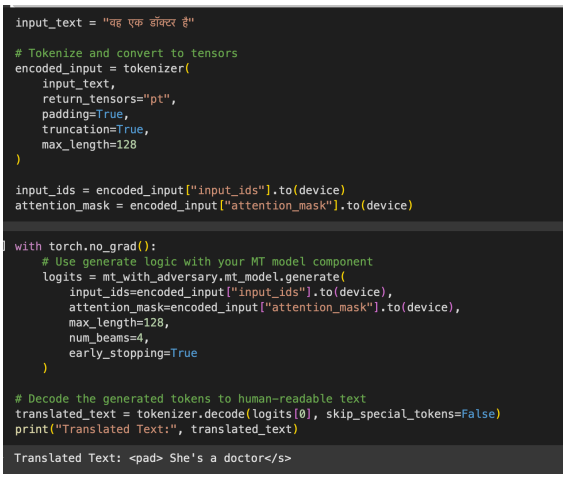

Figure 5: Adversarial Model Output

during training. This classifier was then integrated into a machine translation (MT) model as an adversary to mitigate gender bias.

During fine-tuning, the MT model's translation loss decreased significantly from **11.1995 to 1.11 within 2 epochs**. As expected, the gender classification loss increased throughout the training, indicating that the MT model was learning to obscure gender-specific information. By the end of training, the **gender loss** rose to **38.4**, while the **total loss** stabilized at **3.4829**.

Results on Test Data:

- Average Total Loss: 0.2725
- Average Translation Loss: 0.1404
- Average Adversary Loss: 2.6408

The experiment demonstrated that incorporating the adversarial gender classifier effectively reduced gender bias in the MT model. However, the adversarial setup slightly impacted the baseline translation quality, achieving the goal of promoting gender neutrality in translations.

Input Output Samples:

With the motivation to make language translations gender-inclusive and fair, we analyzed the

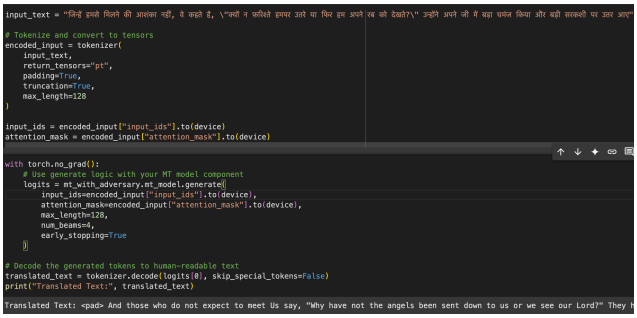

Figure 6: Adversarial Model Output for Longer Input sentences

scenarios where the baseline model—a pre-trained machine translation (MT) model for Hindi-to-English translation—failed. The baseline model often produced biased translations, perpetuating traditional gender stereotypes. For instance: **(Figure 4)**

Input (Hindi): "वह डॉक्टर है।"
Baseline Output (English): "He is a doctor."

To mitigate such biases, we trained the model with an adversarial gender classifier. The adversarial training caused a reduction in translation performance due to the inhibiting factor introduced by the adversarial layer. However, it succeeded in producing some translations away from the existing stereotypes :**(Figure 5)**

Input (Hindi): "वह डॉक्टर है।"
Integrated Model Output (English): "She is a doctor."

Input (Hindi) : "विश्वामित्र और उनके परिजन"
Integrated Model Output (English): "Anandres and Their Family."

Even for longer sentences, the translation quality was maintained. **(Figure 6)**

6 Error analysis

Despite achieving some improvements, the model struggled to maintain translation quality. This was primarily due to limited training data and the suboptimal performance of the adversarial classifier, which itself suffered from the lack of a well-annotated training dataset with appropriate gender labels.

Failed Example 1 - Loss in translation qual-

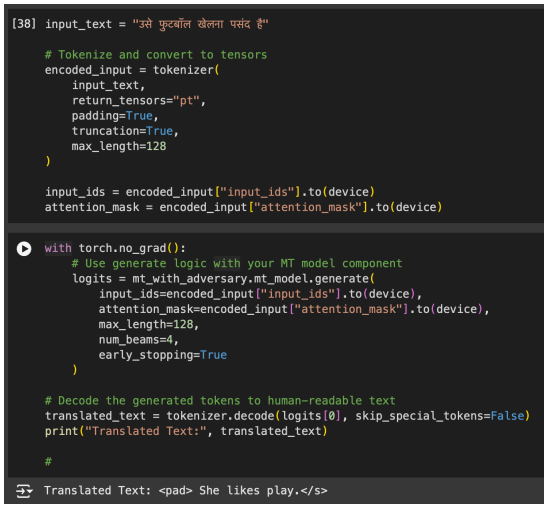

Figure 7: Adversarial Model Gender Inclusive Output at a cost of Loss in Translation Quality

ity **(Figure 7)**
Hindi: "उसे फुटबॉल खेलना पसंद है"
English: "She likes play."
Desired Translation: "He/She/They like to play football"

Failed Example 2 - Output with no gender inclusion
Hindi: "उसका कमरा"
English: "His room"
Desired Translation: "His/Her/Their room"

Reasons of Failure:

- Hindi assigns grammatical gender not only to people but also to objects and abstract concepts. The heuristic approach we employed for the project—given the dataset's limitations— is unreliable for achieving good results in this scenario.

- The design of adversarial training inherently modifies the machine translation model's parameters to account for the classifier's feedback. Since the classifier's loss is added to the translation loss, the MT model shifts its optimization objectives to reduce bias. But this adjustment also comes at the cost of translation quality.

- Gender annotations in the dataset were insufficient to cover the spectrum of grammatical and semantic gender variations, including non-binary and gender-neutral cases.

- In order to analyze and implement a possible solution for inclusivity of non-binary genders in translation outputs, an unreliable form of gender classifier was used. The integrated model was trained on only 15,000 sentence pairs, which limited its ability to generalize across diverse sentence structures and contexts.

7 Conclusion

This proof-of-concept successfully demonstrated the **feasibility of reducing gender bias** in machine translation, albeit with notable limitations. Despite its challenges, this experiment represents a step toward creating fairer and more inclusive machine translation systems.

The most challenging aspect of this project was understanding the concept of adversarial networks and gradient reversal layers (GRL). Once we grasped these concepts, the next hurdle was designing the neural network architecture and determining the correct flow of data—specifically, deciding which outputs should serve as inputs to subsequent layers.

Another significant challenge was devising a method to evaluate the model's performance effectively. Due to the lack of a comprehensive dataset with inclusive and neutral gender labels, initial testing of the model's performance was limited and not robust. To address this, we trained the adversarial network using basic heuristics by identifying gendered terms in the data to classify gender. While this approach showed some improvement, it remains far from reliable and highlights the need for a more systematically annotated dataset.

Several factors are crucial for improving the model's performance. The gender classifier's reliability was critical for adversarial training success. An inaccurate classifier propagates errors to the translation model. The trade-off between translation quality and gender neutrality requires careful balancing. This can be tuned by choosing an appropriate Lambda hyperparameter.

Future work should focus on increasing dataset diversity, enhancing the adversarial classifier's accuracy, and fine-tuning the balance between

gender fairness and translation quality. If a well-annotated dataset with balanced gender labels—including masculine, feminine, and inclusive gender markers—were available, the adversarial classifier could be trained more effectively. This improvement would likely result in:

Enhanced Translation Quality: A more accurate classifier would reduce unintended perturbations in the MT model, allowing it to focus on translation fidelity while addressing bias.

Improved Gender Neutrality: Incorporating diverse and balanced gender examples would ensure better generalization, especially for underrepresented gender cases.

8 Acknowledgements

Generative AI tools, including ChatGPT, were used in this report for assistance with refining text, improving the clarity of certain sections, as well as to enhance formatting and structure. All outputs from the tool were subjected to major changes to ensure alignment with the project goals, originality, and academic integrity.

References

Behnke, H., Fomicheva, M., and Specia, L. (2022). Bias mitigation in machine translation quality estimation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487, Dublin, Ireland. Association for Computational Linguistics.

Chen, Y., Liu, Y., Meng, F., Xu, J., Chen, Y., and Zhou, J. (2024). Beyond binary gender: Evaluating gender-inclusive machine translation with ambiguous attitude words.

Fleisig, E. and Fellbaum, C. (2022). Mitigating gender bias in machine translation through adversarial learning.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in c++.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lardelli, M. (2024). Post-editing machine translation beyond the binary: Insights into gender bias and screen activity.

Menegatti, M. and Rubini, M. (2017). Gender bias and sexism in language.

Singh, P. (2023). Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, page 17–23. Association for Computational Linguistics.