Mitigating Gender Bias in Machine Translation through Adversarial Learning

Anonymous ACL submission

Abstract

Machine translation and other NLP systems often contain significant biases regarding sensitive attributes, such as gender or race, that worsen system performance and perpetuate harmful stereotypes. Recent preliminary research suggests that adversarial learning can be used as part of a model-agnostic bias mitigation method that requires no data modifications. However, adapting this strategy for machine translation and other modern NLP domains requires (1) restructuring training objectives in the context of fine-tuning pretrained large language models and (2) developing measures for gender or other protected variables for tasks in which these attributes must be deduced from the data itself.

We present an adversarial learning framework that addresses these challenges to mitigate gender bias in seq2seq machine translation. Our framework improves the disparity in translation quality for sentences with male vs. female entities by 86% for English-German translation and 91% for English-French translation, with minimal effect on translation quality. The results suggest that adversarial learning is a promising technique for mitigating gender bias in machine translation.

1 Introduction

To avoid perpetuating harm, recent research has begun to examine how biases in NLP systems could be measured and reduced. Efforts to mitigate biases that rely on extensive dataset curation may be infeasible in some applications, such as translation of low-resource or morphologically complex languages. However, recent work suggests that adversarial learning can help to mitigate biases during training without the need to provide additional unbiased data or restructure the original model (Zhang et al., 2018).[1] The method has shown promise in simple proof-of-concept applications, such as mitigating bias in word embeddings for use in analogies.

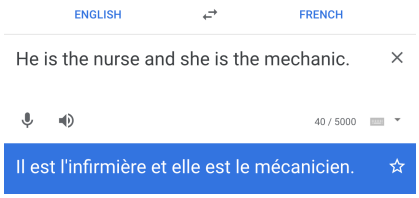

Figure 1: Example of gender bias in English-French translation using Google Translate. The system translates "nurse" to "l'infirmière," a female nurse, and "mechanic" to "le mécanicien," a male mechanic.

Large language models, pre-trained without supervision and then fine-tuned for specific applications, have become a dominant paradigm in NLP. However, using adversarial learning for bias mitigation in such frameworks raises several questions. How can we define a protected variable in the context of these models? How can we apply an adversarial framework for bias mitigation to a pre-training/fine-tuning setup? Finally, how can we quantitatively evaluate the extent to which this method mitigates gender bias?

We present a framework for leveraging adversarial learning to mitigate gender bias in machine translation that advances solutions to several issues faced when using this framework in modern NLP domains: we propose two measures for gender as a protected variable in the context of large language models, discuss how an adversarial framework can be applied during model fine-tuning, and present quantitative results on the effectiveness of this method at mitigating gender bias in machine translation. Our model reduces translation gender bias in the model T5 with little to no adverse effect on translation quality.

[1] This approach is sometimes referred to as "adversarial debiasing," but following the authors themselves, we use "adversarial bias mitigation" to avoid the implication that all forms of bias are completely removed.

2 Background and Related Work

Recent work in the NLP community has stressed the need for studies of bias in NLP systems to discuss the normative reasoning behind why, how, and to whom an NLP system is harmful and ground this research in the literature outside NLP that examines how social processes lead to inequity (Blodgett et al., 2020). Beukeboom and Burgers (2019) define linguistic bias as a "systematic asymmetry in language choice" that reflects stereotypical beliefs about social categories, as applied to either the category as a whole or its members. Under their Social Categories and Stereotypes Communication (SCSC) framework, these stereotypes skew perception of others by (1) preventing members of a social category from being viewed as distinct individuals (perceived entitativity), (2) reinforcing expectations about the social category, and (3) implying that characteristics are immutable and inherent to the group (perceived essentialism). Overtly or implicitly, stereotypes threaten or demean their targets. As a result, cognitive biases harm stereotyped individuals by causing people to fulfill stereotypical expectations, lowering their self-esteem, barring access to opportunities, and harming their mental and physical health (Beukeboom and Burgers, 2019).

One way in which language encodes gender stereotypes is through the use of gendered terms. For example, studies examining job advertisements for male-dominated occupations found that female applicants were not only judged a poorer fit, but were also less likely to apply when a position was advertised in a masculine form (e.g., "chairman") versus a gender-neutral form (e.g., "chairperson") (Menegatti and Rubini, 2017). These effects could be particularly strong in languages with gender inflection, where most terms for professions have different forms depending on the person's gender (e.g., *infirmier/infirmière* for "nurse"). Thus, biases in NLP systems are destructive because they reproduce and reinforce pernicious societal power structures. Interventions in NLP that combat these biases present an opportunity to create more ethical and equitable systems that benefit all members of society.

2.1 What Constitutes a Biased Translation?

The harms of gender-stereotypical translations take the form of representation bias (misrepresenting a social category) and allocation bias (decreased performance for that social category) (Crawford, 2017). Allocation bias with respect to gender in machine translation can occur when the accuracy of translation decreases according to a linguistic bias. This includes: (1) mistranslating sentences when they contain a female entity, but not when they contain a male entity and (2) mistranslating sentences when they contain a counter-stereotypical association (such as a female doctor or male nurse, as in Figure 1), but not when they contain a stereotypical association. Mistranslations of sentences that contain a counter-stereotypical association (e.g., a female mechanic) simultaneously display allocation bias, because they fail to provide equal performance to different genders, and representational bias, because they reinforce gendered stereotypes.

This research aims to minimize allocative and representational bias perpetuated *within* a machine translation system as measured by the failure to meet a statistical fairness criterion. Statistical fairness criteria that have been proposed include *demographic parity*, *equality of odds*, and *equality of opportunity* (Hardt et al., 2016; Beutel et al., 2017); we use demographic parity, which defines a fair classifier as one in which predictions $\hat{Y}$ and the protected variable Z are independent. That is,

$$P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$$

The adversarial method for bias mitigation used in this paper can be quickly extended to work with equality of odds and equality of opportunity (see Section 3).

2.2 Documenting Bias

Caliskan et al. (2017) found that word embeddings exhibited gender and racial bias similar to those exhibited by humans and that machine translation systems exhibited gender bias in its translation of pronouns; subsequent studies found similar biases across other NLP tasks (May et al., 2019; Zhao et al., 2017; Rudinger et al., 2018). The translation biases found by Caliskan et al. (2017) raised awareness of bias in machine translation, leading some translation systems to introduce limited gender-specific translations as recently as 2020. However, more recent studies by Kocmi et al. (2020a) and Stanovsky et al. (2019) found that evidence of gender bias persisted across 10 languages over a total of 23 translation systems, including Google Translate, Microsoft Translator, Amazon Translate, and Systran.

Biases can be incorporated into machine learning systems during different stages of model devel-

opment. Stereotyped associations and unbalanced representation of different demographics in training corpora ("dataset bias"), along with bias amplification effects during model training, result in models that exhibit biases far beyond real-world disparities (Rudinger et al., 2018; Lu et al., 2018).

2.3 Mitigating Bias

Escudé Font and Costa-jussà (2019) propose reducing gender bias in English-Spanish machine translation by adjusting word embeddings and suggest that this method improves translation gender bias on some examples. Meanwhile, several efforts at mitigating translation bias have intervened through dataset curation, either by refining or annotating existing training sets or creating new datasets for fine-tuning. Vanmassenhove et al. (2018) tagged sentences with information on the speaker's gender, which affects the grammatical gender of words in some languages and may inform word choice more generally, to improve the translation quality of sentences spoken by women. Saunders et al. (2020) and Stafanovičs et al. (2020) similarly use training data annotated with gender tags. Saunders and Byrne (2020) addressed gender bias using corrective fine-tuning with a smaller, handcrafted dataset of gender-balanced sentences and suggested methods for swapping the genders of entities in languages with gender inflection. The authors note that there is usually a tradeoff between bias mitigation on the WinoMT dataset (see Section 3) and translation quality.

However, data-centric approaches to bias mitigation suffer from the difficulty of collecting data for specific applications, particularly for low-resource and morphologically complex languages. In addition, approaches that intervene before training may address dataset bias but not bias amplification. In the case of gender bias, data-centric approaches may also be more difficult to apply to languages with grammatical gender, in which gender may be represented not only in pronouns or occupations, but also through the inflection of nouns, adjectives, and other parts of speech (Zmigrod et al., 2019).

2.4 Adversarial Learning Approaches

Zhang et al. (2018) proposed an adversarial technique for general bias mitigation during training. An adversary is trained to predict a *protected variable* (i.e., gender), while the model learns to prevent the adversary from predicting the protected variable (see Section 3). They applied adversarial bias mitigation to two proof-of-concept tasks: income prediction on the UCI Adult dataset and analogy completion (e.g., "man : woman :: king : __"). For the analogy completion task, the study defined gender bias according to the notion of a "gender direction" proposed by Bolukbasi et al. (2016). This method measures gender bias as the magnitude of the projection $proj_g y$ of a sentence y onto the "gender direction" g of a word embedding space.

Zhang et al. found that the method substantially reduced bias in the income prediction task. They also gave examples of bias reduction in the analogy completion task, such as a decreased likelihood of choosing "nurse" as the female equivalent of "doctor"; however, they do not provide evidence of systematic bias mitigation in the model overall, possibly due to the scarcity of datasets for testing gender bias at the time of the study's publication. Thus, the exact degree to which this method can mitigate bias remained an open question.

Kumar et al. (2019) independently introduced an adversarial framework for text classification to prevent confounding variables, such as the mention of a particular country, from overly affecting classification, such as language identification. Xia et al. (2020) drew on this vein of research to mitigate racial bias in a small LSTM-based hate speech detection model, using tweets that were pre-annotated with the probable race of the author. However, previous work on adversarial bias mitigation has yet to examine issues pertaining to measuring gender as a protected variable when not prelabeled or apply the adversarial technique to large language models.

3 Approach

The adversarial framework for bias mitigation has several advantages that make it suitable for machine translation and LLM-based tasks more broadly. Adversarial bias mitigation is a model-agnostic strategy: so long as the model trains using gradient descent, the complexity of the model being trained does not affect the overall framework. This advantage makes it suitable for mitigating biases under the common framework of pre-training a large language model on a general language understanding task, then fine-tuning on machine translation or other specific applications. By modifying the training process itself, it also works to mitigate the effects of bias amplification as well as dataset bias. In addition, unlike bias mitigation techniques that intervene before or after the training process,

adversarial bias mitigation does not require extensive modifications to the training data or additional data collection, which makes it easier to extend to new tasks or low-resource domains.

3.1 Framework

We use an adversarial approach that mitigates gender bias by defining a training objective that encourages a model to minimize the gendered information encoded in output sentence embeddings beyond what is strictly necessary to translate the sentence correctly. In this approach, a prediction model M with weights W learns to predict an output Y from input X while remaining neutral with respect to the protected variable Z. The adversary A attempts to predict Z from the model's output predictions $\hat{Y}$. Then, W is updated according to:

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

where α is a tuneable hyperparameter. This training objective penalizes the prediction model for helping the adversary to determine the value of the protected variable (Zhang et al., 2018).

In this work, we define fairness through demographic parity (see Section 2.4). Replacing the objective of demographic parity used in this paper with either equality of odds (conditional independence between $\hat{Y}$ and Z, given Y) or equality of opportunity for a group y (independence between $\hat{Y}$ and Z, conditioned on $Y = y$) requires minimal changes: for equality of odds, the adversary can be given access to the target translation Y as well as the prediction $\hat{Y}$; for equality of opportunity on a class y, the adversary should only train on examples for which $Y = y$.

3.2 Defining the Protected Variable Z

3.2.1 Method 1: Gender Direction from Sentence Encodings

To define the protected variable Z, we extend the notion of a "gender direction" g from Bolukbasi et al. (2016) and Zhang et al. (2018). Zhang et al. defined 10 male/female word pairs (e.g., he/she, him/her), and from these defined a *bias subspace*, the space spanned by the top principal component of the differences. The unit vector g representing the bias subspace thus approximates the "she-he direction" of the word embedding space. They then defined the protected variable for the task of completing analogies based on word embeddings as $proj_g y$, the projection along the gender direction of the word y that completes the analogy.

Extending this formulation to sentence embeddings, we calculate the bias subspace from the top principal component of the model M's sentence encodings for the 10 male/female word pairs to find g for the sentence embedding space of the output of the model.[2] We then define Z as $proj_g y$, the projection of the sentence encoding along the gender direction. After masking all pronouns in the model's predicted translation $\hat{y}$ of a sentence, the adversary attempts to predict $proj_g \hat{y}$, while the model is trained to avoid providing information that allows the adversary to do so.[3]

3.2.2 Method 2: Pronoun Usage Heuristic

An open question is whether there are ways of defining the protected variable Z that are more effective at mitigating bias or otherwise useful for NLP tasks. Thus, we also implemented a *pronoun usage heuristic* for defining the protected variable Z. Under this alternative metric, Z is defined as 1 if a female-gendered pronoun occurred, -1 if a male-gendered pronoun occurred, and 0 if both occurred. This simpler metric has some advantage in terms of computational cost, since the principal component analysis and matrix multiplications needed to calculate the projection of each encoded sentence on g require some expensive calculations before training. On the other hand, evidence that the gender direction metric is more effective than the pronoun usage metric would indicate that calculating Z from how the model encodes the sentence provides additional information that is useful for mitigating bias in that model.

4 Implementation

We fine-tuned the model T5 (Raffel et al., 2020) on English-French and English-German translation under our adversarial bias mitigation paradigm.[4] The adversarial intervention occurred during fine-tuning alone, without intervening during the pre-training stage. Since T5 is an encoder-decoder

[2] g is reduced to 30,000 entries to prevent the principal component analysis from becoming prohibitively expensive.

[3] First names were uncommon enough in the data that we found masking them was not needed for the method to work.

[4] In an effort to examine performance on low-resource languages, we also performed initial experiments on English-Czech translation with an order of magnitude smaller dataset; however, issues with translation quality suggest more extensive modifications are necessary to extend this technique to low-resource settings.

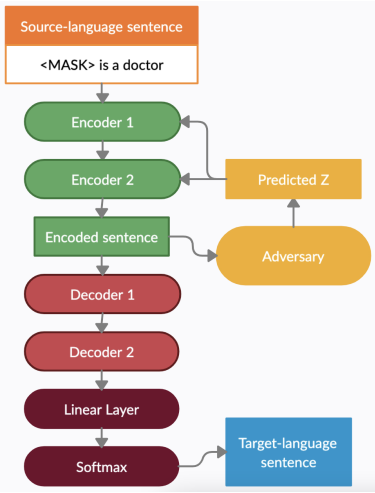

Figure 2: Our framework for adversarial bias mitigation in machine translation with T5.

model, $\hat{Y}$ (the representation of the encoded sentence) is the output of the second encoder of T5. For the gender direction method, the protected variable Z for a sentence S was found through principal component analysis on $\hat{Y}$ on the pretrained model before fine-tuning. During training, $\hat{Y}$ is then sent to the adversary A, which attempts to predict Z (Figure 2).

We used the WMT-2014 corpus (Bojar et al., 2014) to train the model and evaluate for translation quality (see Section 5). For each translation pair, the model was fine-tuned on a subset of 100,000 examples that contained at least one gendered pronoun. This was done to ensure that the training set included a higher proportion of sentences with gendered entities, since the majority of sentences in the original dataset contained no gendered entities at all. We masked all source sentences' gendered pronouns in the training data.

The development and test sets each consisted of 50,000 random unseen sentence pairs from the corpus, including sentences without gendered pronouns, to ensure that the intervention during fine-tuning did not hinder the model's ability to translate in general. (See Appendix B for hyperparameter details.)

5 Evaluation Results

The model was tested on both translation accuracy (BLEU score on the WMT-2014 data) and minimization of bias. For bias mitigation, the model was tested on the WinoMT dataset (Stanovsky et al., 2019), a challenge set for gender bias in machine translation consisting of sentences balanced between male and female genders and between male and female gender role assignments (e.g. male doctor/female doctor, female nurse/male nurse).

The primary metrics used to measure performance on the WinoMT dataset are ΔG, ΔS, and the overall accuracy of preserving the gender of an entity upon translation ("acc."). ΔG denotes the difference in F1 scores between sentences involving masculine entities and sentences involving feminine entities. ΔS denotes the difference in accuracy on correctly translating the antecedent between sentences involving pro-stereotypical (e.g., male doctor/female nurse) and anti-stereotypical (e..g, female doctor/male nurse) role assignments.

5.1 Results on Bias Mitigation

Both methods of bias mitigation significantly reduced gender bias in the translated output, with minimal change in translation accuracy (Table 1). The disparity in F1 scores between sentences involving male and female entities (ΔG) decreased from 18.8 (EN-FR) and 27.5 (EN-DE) in the baseline to 4.8 and 10.0 under the pronoun usage heuristic and to 2.0 and 2.4 using the gender direction (86% and 91% improvement, respectively).

The accuracy of preserving the genders of entities in translation increased from 53.5% and 53.7% in the baseline to 62.9% and 78.9% under the pronoun usage heuristic and to 64.9% and 72.7% using the gender direction (Figure 3). The disparity in accurate translation of antecedents in sentences with stereotypical and reverse-stereotypical role assignments (ΔS) decreased in most cases, from 15.3 and 1.5 in the baseline to 14.9 and 0.8 under the pronoun usage heuristic and to 9.0 and 7.3 using the gender direction. [5] (The greater effect on ΔG compared to the other metrics is consistent with other studies on the WinoMT dataset, in which ΔG usually displays the most visible change after bias mitigation (Bergmanis et al., 2020; Kocmi et al., 2020b).)

[5] The WinoMT dataset for gender bias evaluation is fairly new, which places a limitation on comparing the results on these evaluation metrics to previous studies of bias mitigation in machine translation. One study whose evaluation bears some similarity is by Bergmanis et al. (2020), who added gender annotations to WMT-2015 to mitigate bias in the Sockeye MT model (with 40 times as many training examples as this work). Their model's performance went from a baseline of $\Delta G = 29.8, \Delta S = 11.8$ (EN-FR) and $\Delta G = 10.2, \Delta S = 14.4$ (EN-DE) to $\Delta G = 1.6$ and $\Delta S = 10.1$ (EN-FR) and $\Delta G = -4.7, \Delta S = 1.7$ (EN-DE).

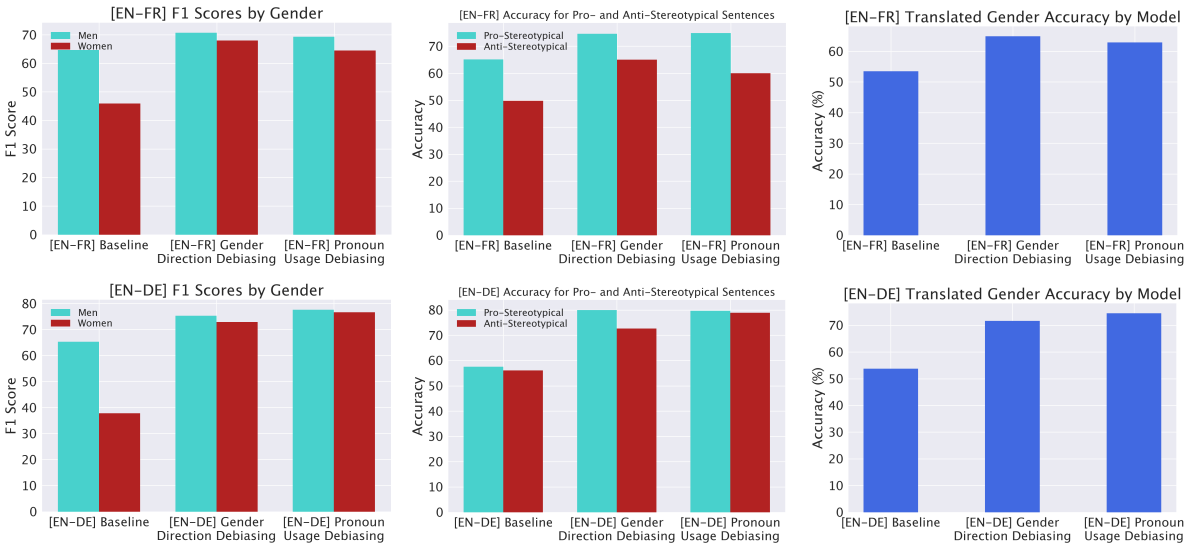

Figure 3: **Left**: Difference in F1 scores on WinoMT for sentences involving male vs. female entities. **Middle**: Accuracy scores on WinoMT dataset for sentences involving pro- vs. anti-stereotypical role assignments. **Right**: Overall accuracy of preserving the gender of an entity upon translation in WinoMT.

		BLEU	ΔG	Acc.	ΔS
EN-FR	**Baseline**	30.7	18.8	53.5	15.3
	Gender direction method	29.2	**2.7**	**64.9**	**9.0**
	Pronoun usage method	**33.0**	4.8	62.9	14.9
EN-DE	**Baseline**	28.4	27.5	53.7	1.5
	Gender direction method	**31.5**	**2.4**	72.7	7.3
	Pronoun usage method	29.9	10	**78.9**	**0.8**

Table 1: Results for bias mitigation and translation quality on all three models. BLEU scores were evaluated on the WMT-2014 data. ΔG denotes the difference in F1 scores on the WinoMT dataset between sentences with masculine entities and sentences with feminine ones. ΔS denotes the difference in accuracy scores on the WinoMT dataset between sentences with pro-stereotypical and anti-stereotypical role assignments.

5.2 Results on Translation Quality

The translation quality, as measured by BLEU score, displayed only small variations under both bias mitigation methods. In English-French translation, BLEU score decreased slightly from 30.7 to 29.2 when using the gender direction, and in fact increased from 30.7 to 33.0 when using the pronoun usage heuristic. In English-German translation, the BLEU score increased in both cases, from 28.4 in the baseline to 31.5 when using the gender direction method and 29.9 with pronoun usage method. Since a slight decrease in translation quality might be expected when training with more complex objectives for bias mitigation, the increase in translation quality under most circumstances might suggest that the adversarial bias mitigation framework provides some additional benefit in terms of avoiding overfitting: training the model to avoid gendered associations could permit it to generalize more easily to unseen data.

The gender direction method provided the best results in terms of bias mitigation on English-French translation; in English-German translation, the pronoun usage method was more suited to removing stereotypical associations, whereas the gender direction method was more suited to improving performance on sentences with female entities. Both methods performed comparably in terms of translation accuracy, though the pronoun usage method provided a significantly greater boost in English-French translation. Linguistic differences between French and German, such as the existence of a neuter gender in German, could account for these differences; future work could examine more languages across of a variety of language families to further understand this behavior. In addition, these results suggest that when choosing methods

Source sentence	Translation (baseline)	Translation (after bias mitigation)
The developer argued with the designer because **she** did not like the design.	**Le promoteur** a eu des échanges avec l'auteur parce qu'**il** n'a pas aimé le dessin.	**La développeure** a contesté la conception parce qu'**elle** n'aimait pas le design. *(gender direction)*
The physician hired **the secretary** because **he** was highly recommended.	Le médecin embaucha **la secrétaire** parce qu'**elle** était recommandée à la haute recommandation.	Le médecin a embauché **le secrétaire** parce qu'**il** était hautement recommandé. *(pronoun usage)*

Table 2: Sample translations by the baseline model and by the models trained with each of the bias mitigation methods. Colors indicate entities' grammatical genders: **blue** for male entities and pronouns, **red** for female ones, and **orange** for neutral ones.

for bias mitigation, there may be a tradeoff, albeit small, between the best possible fairness and the best possible accuracy. For example, in English-French translation, the pronoun usage method provides the best improvement in overall translation quality, while the gender direction method is best at improving gender bias across all metrics.

The particular goals of the application for which a translation model is deployed could determine which method of measuring gender bias is best for that context: the gender direction method provides more dramatic improvements in some cases, whereas the pronoun usage method provides more consistent bias mitigation across all metrics in both languages. That said, both methods of measuring gender bias in conjunction with adversarial learning resulted in significant decreases in gender biases across nearly all metrics.

5.3 Examples of Bias Mitigation

Table 5.2 gives examples of translations with and without different bias mitigation techniques on the WinoMT dataset. Sentences in the WinoMT dataset are designed such that the model must use context to determine which entity in a sentence (e.g., "the developer" or "the designer" in Table 1) corresponds to the gendered pronoun, since resolving the coreference to either entity would be syntactically correct.

Without using bias mitigation, the baseline model translates both gendered entities and pronouns in stereotypical ways: a female developer becomes *le promoteur*, the male developer, and a male secretary becomes *la secrétaire*, the female secretary. The gendered pronouns associated with these entities are also translated to the stereotypical gender for those professions. Both methods of bias mitigation, however, translate both the gendered entities and the gendered pronouns that correspond to them correctly in these instances. The female developer becomes *la développeure* and the correct pronoun *elle* is used for her; the male secretary becomes *le secrétaire*, using the correct pronoun *il*. These differences illustrate how both the gender direction and pronoun usage methods can successfully mitigate bias through adversarial learning.

6 Conclusion

Linguistic biases serve as a vehicle for harmful stereotypes that demean individuals and communities, harm their mental and physical health, and worsen life outcomes. Recent studies indicate that NLP systems perpetuate these biases, reproducing stereotypes in their output and disproportionately producing demeaning or outright incorrect output for groups that face societal discrimination. For example, translation systems often translate pronouns or gender inflections incorrectly when they correspond to counter-stereotypical professions.

Adversarial bias mitigation intervenes during training by introducing an adversary that attempts to predict a protected variable from the output of the model. It is a model-agnostic strategy, permitting complex models to be substituted into the framework without changing the overall setup, which is convenient for pre-training/fine-tuning setups. It requires no modifications to the training data or additional data collection, permitting it to be extended to new tasks or low-resource domains more easily. By intervening during training itself, this approach also aims to mitigate both dataset bias and bias amplification.

In this work, we addressed several open questions raised by previous research into the adversarial approach to bias mitigation: (1) how to define gender as the protected variable in more complex applications, particularly in language tasks where such information is not prelabeled; (2) how to apply the adversarial framework to a pre-training/fine-

tuning setup with large language models, as has become the norm; and (3) whether the adversarial framework is indeed effective as measured by quantitative evaluations on realistic tasks.

We presented an adversarial framework for mitigating gender bias in machine translation. Our approach proposes two measures for gender as a protected variable in the context of large language models, the *gender direction* method and *pronoun usage* method. We then applied the adversarial framework to English-French and English-German machine translation. For both the gender direction and pronoun usage methods, the difference in F1 scores between sentences in the WinoMT dataset involving male and female entities decreased, and for the pronoun usage method, the difference in accuracy between pro- and anti-stereotypical sentences also decreased. In addition, the accuracy of preserving the gender of an entity upon translation increased and the accuracy of translating pro- and anti-stereotypical sentences increased for both methods. Furthermore, mitigating gender bias did not come at the expense of translation quality. In fact, translation accuracy slightly increased in most cases, suggesting that the method might provide some additional ability to generalize to new examples.

The gender direction method was significantly more successful at mitigating bias in some cases, whereas the pronoun usage method provided more consistent but usually less thorough bias mitigation. Nonetheless, both methods were effective mitigating gender bias in machine translation. The results suggest that the adversarial framework is a promising technique for mitigating biases in common and complex NLP tasks.

6.1 Future Work

A broader avenue of research concerns extending the adversarial framework to other NLP tasks and to protected variables such as race or religion, for which indicators of the protected variable may be more difficult to measure. Possible extensions to this work within machine translation could examine the efficacy of this method on different translation pairs, especially between more dissimilar languages. Translation from languages with more complex systems of gender inflection might require more complex strategies for defining the protected variable tailored to their syntactic and morphological features. Another key direction is to account for nonbinary or transgender users and others who face unique forms of gender discrimination (e.g., misgendering) or for whom gender bias mitigation based on binary notions of gender would result in oversimplified interventions (Cao and Daumé III, 2020).

Complementary research can involve integrating stakeholders in the development of NLP systems, such as by surveying users of various genders or collaborating with language reclamation activists. Allowing users to contest or modify the decisions made by a system, such as by allowing users to correct biased translations or choose between multiple translations, could also improve trust in the fairness of a translation system (Vaccaro et al., 2019). Combining multiple types of interventions can allow the NLP community to address these issues, which require both technical and ethical insights.

References

Toms Bergmanis, Artūrs Stafanovičs, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 629–638.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7:1–37.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. keynote talk at Neural Information Processing Systems (NIPS '17).

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020a. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020b. Gender coreference and bias evaluation at wmt 2020.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. *CoRR*, abs/1909.00453.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv:1807.11714*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem.

Danielle Saunders, Weston Feely, and Bill Byrne. 2020. Inference-only sub-character decomposition improves translation of unseen logographic characters. In *Proceedings of the 7th Workshop on Asian Translation*, pages 170–177, Suzhou, China. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in algorithmic systems. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, page 523–527, New York, NY, USA. Association for Computing Machinery.

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, Brussels, Belgium. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Ethics and Data Statement

The WMT-2014 training data used for the study comes from was chosen because this dataset, as used in the 2014 ACL Workshop in Statistical Machine Translation, was originally used to train T5 for machine translation; therefore, training on this data could be reasonably expected to provide similar results as those on the original model. The data used here comes from parallel English-French and English-German texts. The English data consists primarily of General American English and British English; other dialects of English, such as African-American English, are underrepresented in this corpus. Similarly, the French and German data consists primarily of varieties of these languages as spoken in Europe. The training set therefore cannot not provide a balanced representation of the various dialects and speaker demographics of these languages.

The WinoMT dataset used for evaluation (see Section 5) was chosen because it is the predominant benchmark for evaluating gender bias in machine translation. It consists of English-language templates in General American English that were then translated into other languages. Thus, an important avenue for further research is to examine whether bias mitigation techniques like this one are effective on more translation pairs and language varieties.

Finally, we acknowledge that our approach assumes a binary notion of gender and does not account for other gender identities; we recommend that future work explore avenues for gender-inclusive translation as well.

B Implementation Details

The number of training epochs was manually fine-tuned from 1 to 5 on the development set; all models presented here were trained for 2 epochs, except for the EN-FR gender direction model (1 epoch) and EN-DE baseline (3 epochs). A learning rate of 0.001 was used for both the translation model and the adversary and the Adam optimizer was used for all models. We used T5-base (220 million parameters); training time for each model varied from 24 to 72 hours on one NVIDIA Quadro RTX 8000 GPU.

The development set BLEU scores corresponding to the final models used here were:

		BLEU
EN-FR	**Baseline**	30.6
	Gender direction method	29.3
	Pronoun usage method	33.8
EN-DE	**Baseline**	28.4
	Gender direction method	31.3
	Pronoun usage method	30.3

.

Our datasets are available at `anonymshare.com/396p/inputs.zip`.