How sensitive are translation systems to extra contexts? Mitigating gender bias in Neural Machine Translation models through relevant contexts

Shanya Sharma
Walmart Labs, India
shanya.sharma@walmart.com

Manan Dey
SAP Labs, India
manan.dey@sap.com

Koustuv Sinha
McGill University, Montreal, Canada,
Mila - Quebec AI Institute,
koustuv.sinha@mail.mcgill.ca

Abstract

Neural Machine Translation systems built on top of Transformer-based architectures are routinely improving the state-of-the-art in translation quality according to word-overlap metrics. However, a growing number of studies also highlight the inherent gender bias that these models incorporate during training, which reflects poorly in their translations. In this work, we investigate whether these models can be instructed to fix their bias during inference using targeted, guided instructions as contexts. By translating relevant contextual sentences *during inference* along with the input, we observe large improvements in reducing the gender bias in translations, across three popular test suites (WinoMT, BUG, SimpleGen). We further propose a novel metric to assess several large pre-trained models (OPUS-MT, M2M-100) on their sensitivity towards using contexts during translation to correct their biases. Our approach requires no fine-tuning and thus can be used easily in production systems to de-bias translations from stereotypical gender-occupation bias [1]. We hope our method, along with our metric, can be used to build better, bias-free translation systems.

1 Introduction

Despite the ongoing success of large pre-trained Transformer (Vaswani et al., 2017) based models in Neural Machine Translation (NMT), these systems are immensely prone to various forms of gender biases in their learned representations. Recent work (Stanovsky et al., 2019; Prates et al., 2020) has found out that a specific kind of gender bias exists

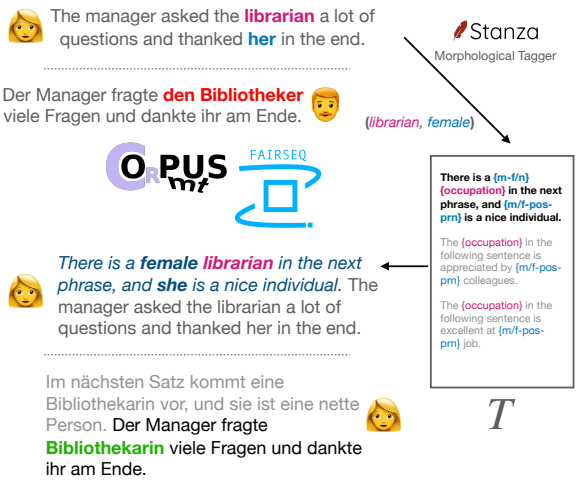


Figure 1: An example of our de-biasing pipeline, using OPUS-MT (Tiedemann and Thottingal, 2020) model on English to German translation on a sample drawn from WinoMT (Stanovsky et al., 2019) dataset. We correct the bias during inference by providing additional relevant *context* (built using our template bank T) to the input, which we remove after the translation.

in the translation of NMT models. Specifically, sentences containing stereotypical occupations[2] which are typically gender-unbalanced in the training data are translated per their respective stereotypes (e.g., *nurse* tends to be associated to *female* pronouns) intact in the output (Figure 1). It is, therefore, imperative to study effective de-biasing techniques to instruct a translation model to output unbiased translations.

De-biasing biased gender associations have been thoroughly investigated in the light of static word embeddings (Bolukbasi et al.; Zhao et al., 2018a; Elazar and Goldberg, 2018). Relatively fewer

[1] Our evaluation data and code are publicly available at https://github.com/manandey/bias_machine_translation

[2] Occupations obtained from the US Bureau of Labor Statistics, http://bls.gov/cps/cpsaat11.htm

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1968 - 1984
December 7-11, 2022

works exist in de-biasing on contextual word embeddings (such as Transformer-based models). The prevalent approach is to fine-tune a pre-trained contextualized embedding while balancing gender-specific associations (Zhao et al., 2019), or by fine-tuning the word embeddings of the pre-trained model itself (Kaneko and Bollegala, 2021). However, in the case of NMT, fine-tuning is expensive as it requires massive parallel corpora, and to this date, we do not have a massive, gender-balanced parallel corpora to begin with.

Thus, in this work, we investigate whether context can be leveraged as a way to improve the bias of a translation system. NMT systems have been reported to be highly sensitive to input sentences (Fadaee and Monz, 2020; Dankers et al., 2022). In this work, we aim to use it to our advantage to de-bias a model. Instead of fine-tuning a translation model, we embark on improving the generation by allowing the model to focus on contextual information during inference. Concretely, we expose the model to unambiguous context alongside to the input to translate, containing the unbiased gender association of the entity in the input[3]. Specifically, we do not modify the input: instead we either *prepend* or *append* the context to the input as a separate sentence, separated by delimiters (Figure 1). We observe that using this context, state-of-the-art translation models are able to reduce their gender association biases for occupations considerably in the output translation, purely during inference. This improvement varies according to model, language-pair and occupations.

Concretely, in this work we systematically study the effect of contexts in the source language (English) affecting the occupation gender-bias in the target translation (German, French, and Spanish) using two popular state-of-the-art NMT models: M2M-100 (Fan et al., 2021) and OPUS-MT (Tiedemann and Thottingal, 2020), in three publicly available occupation-gender bias datasets: WinoMT (Stanovsky et al., 2019), BUG (Levy et al., 2021) and SimpleGen (Renduchintala and Williams, 2022). We find that both models can be de-biased significantly by adding unambiguous, relevant contexts, with the largest improvement being that of M2M-100, which exhibits higher sensitivity towards additional contexts. Our proposed method thus introduces a simple and effective way to de-bias translations during inference.

[3]Unlike Fadaee and Monz (2020), we do not modify the source sentence in itself, instead we append or prepend extra contextual information, which can also be compared to tuning-free prompt mining setup in the literature (Liu et al., 2021).

2 Approach

To mitigate gender-stereotyped bias for occupations in NMT, we consider the approach of adding a context to the input, as a separate sentence either prepended or appended with the original input sentence. This context is generated using hand-crafted templates, which contain unambiguous signals about the gender of the profession in the source sentence. In particular, we investigate the capability of NMT models to extract signals from these contexts and mitigate gender bias in the translation during inference. Our approach can also be thought of as an in-context learning scheme, as popularized by Brown et al. (2020).

2.1 Template construction and usage

We start with a parallel corpora D consisting of sentence pairs (X, Y), where each source sentence X contains a target entity of gender *male* or *female*[4] and associated with an occupation. X also contain gender-specific pronoun(s) to indicate the gender of the target entity. Y denotes the gold translation of X. We use a pre-trained translation model to translate X to $\hat{Y}$. This translation permeates the stereotypical bias of the target entity in the source, which we aim to fix in this work, by providing a context during translation.

To construct unambiguous contexts, we carefully create templates t, which can be used to generate a contextual sentence, c. This context provides enough signal unambiguously to convey the correct gender of the target entity in X (Figure 1). For example, given a context template t = *"The* {occupation} *in the next sentence identifies* {male or female self-reference pronoun} *using the pronouns* {male or female subject pronoun}/{male or female object pronoun}", and given an occupation gender pair `(nurse, male)`, we construct the following context c: *"The nurse in the next sentence identifies himself using the pronouns he/him."*

Thus, given the input sentence X, we prepend or append the context c to construct a new input for translation, $X_c = [c \| X]$ or $X_c = [X \| c]$, where $\|$

[4]We acknowledge that a limitation of this work is that we do not consider the non-binary genders, and we leave it for potential future work to explore the context sensitivity of non-binary genders in language. We also hope our work will lead to new ideas and better methods for mitigating biases about non-binary and transgender people in the future.

is the delimiter which separates the two sentences. We translate this sentence X_c to a target language $\mathcal{L}$ using the pre-trained translation model to output the translation $\hat{Y}_c$. To extract the intended translation, we drop the translated context by splitting the output using the delimiter $\|$, to get $\hat{Y}_{c'}$, which is the translation of X_c after removing c.

2.2 Choosing a template

Using our template construction strategy, we create T unique templates which could be applied to a given input sentence. We use a greedy strategy to choose a template to apply for a given sentence X. Following the formulation of Stanovsky et al. (2019), we first use a heuristic morphological tagger to extract the gender of the target entity from the source (g_X) and from the translation ($g_{\hat{Y}}$). We use Stanza (Qi et al., 2020) as the morphological tagger and `AWeSOME` aligner (Dou and Neubig, 2021) to align the source and target entities. $g_X \neq g_{\hat{Y}}$ indicates the presence of stereotypical bias [5]. In those sentences, we iteratively search for a relevant context $c, \forall t \in T$ such that $g_X = g_{\hat{Y}_{c'}}$. We stop this search once we exhaust our set of templates in T.

2.3 Experiment Details

Models: For our experiments, we consider using the two most commonly used open-source multilingual translation models, M2M-100 and OPUS-MT, for evaluation. M2M-100 (Fan et al., 2021) which is a many-to-many multilingual encoder-decoder translation model that can translate directly between any pair of 100 languages, based on the Transformer(Vaswani et al., 2017) architecture. We use the 418 Million parameters version of the model. OPUS-MT (Tiedemann and Thottingal, 2020) is a collection of bilingual and multilingual models based on the standard 6-layer 8-head Transformer architecture. We use HuggingFace (Wolf et al., 2020) model hub to load and run inference for both of the above models.

Languages & Datasets: We perform our experiments using translations from English to three target languages: German, French and Spanish. We chose these three languages as they are well-supported by Stanza for performing morphological analysis, while also being supported by the NMT models we consider above. For each of these languages, we carry the evaluation on three datasets, WinoMT (Stanovsky et al., 2019), BUG (Levy et al., 2021) and SimpleGen (Renduchintala and Williams, 2022).

Generating templates: We construct $|T| = 87$ unique templates[6] with varying linguistic properties. For each model and language-pair, we prune a subset of templates that evoke stereotypical gender biases in their translations. We then apply these templates to the input sentences in the dataset and translate the combined sentence. We observe that the choice of delimiter used to combine the input sentence with the context has a significant effect on the actual translation quality of the input sentence. In our primary experiments, we choose hash(#) as our delimiter since it provides a substantial improvement in the bias while also ensuring minimal change in translation quality. We discuss more about the choice and impact of delimiters in §3.3.

Evaluation. To compute the gender translation *accuracy*, we extract the predicted gender from the translation, $g_{\hat{Y}}$, using the morphological tagger. Then, we measure if this predicted gender $g_{\hat{Y}}$ is the same as that of an annotated, gold truth gender g_Y for the same entity. We use BLEU scores to evaluate the translation of the combined sentence X_c using the same setup, which contains the source sentence X and a context c. In this case, post translation, we drop the translated output of c and evaluate $Y_{c'}$ using the same method as described above.

3 Results & Analysis

In this section, we conduct a series of experiments and analysis to understand the viability of correcting gender bias in translation by using contexts.

3.1 Does addition of templates allow the model to correct its bias?

Setup. Concretely, we want the model to generate the appropriate gender-specific pronouns for the occupations in the translation Y according to the gender g_X in the input sentence X. We first apply our greedy template selection strategy to the WinoMT, BUG and SimpleGen datasets, using OPUS-MT and M2M-100 models. We compute the greedy search accuracy $\mathcal{A}_C$.

Results. We find that by the application of greedy strategy (§2.2), the accuracy $\mathcal{A}_C$ is significantly high (Table 1), with the highest performance improvement in BUG dataset (**87.36%** for M2M-100

[5] Note, we do not use the ground truth annotation g_Y to decide this, as it is unavailable during testing. Our method relies on the accuracy of the heuristic morphological tagger.

[6] Table 12 in the Appendix contains a full list of templates.

Dataset	Model	Target Language	Without Context	With Context				
			$\mathcal{A}(\%)$	$\mathcal{A}_C(\%)$	$\mathcal{A}_{all}$ (%)	$\mathcal{CSS}$	$\mathcal{C}_U$	$\mathcal{C}_L$
WinoMT	OPUS-MT	German (de)	60.57 (8.64)	**82.13**	**62.09 (9.81)**	0.27 (0.38)	54.60	10.64
		French (fr)	57.70 (12.20)	**73.64**	**59.88 (13.67)**	0.34 (0.43)	37.60	3.80
		Spanish (es)	60.10 (10.18)	**76.79**	58.97 (8.64)	0.31 (0.41)	41.80	9.96
	M2M-100	German (de)	58.71 (10.82)	**80.65**	58.25 (10.86)	0.25 (0.35)	53.13	6.30
		French (fr)	49.43 (19.09)	**76.83**	**54.03 (7.42)**	0.37 (0.44)	54.18	17.40
		Spanish (es)	56.56 (10.23)	**85.35**	**59.25 (11.69)**	0.35 (0.41)	66.27	18.45
BUG	OPUS-MT	German (de)	70.72 (17.08)	**85.60**	66.91 (16.09)	0.20 (0.30)	51.43	8.40
		French (fr)	55.96 (19.27)	**75.90**	**61.34 (16.91)**	0.22 (0.35)	45.70	11.00
		Spanish (es)	74.85 (21.64)	**86.74**	**75.50 (18.31)**	0.16 (0.29)	47.72	7.20
	M2M-100	German (de)	58.13 (10.97)	**87.36**	**67.09 (15.20)**	0.39 (0.42)	70.39	25.95
		French (fr)	48.08 (16.89)	**78.84**	**57.92 (20.22)**	0.40 (0.43)	59.76	22.27
		Spanish (es)	63.19(17.10)	**82.25**	**72.34(19.75)**	0.32(0.40)	61.11	18.18
SimpleGen	OPUS-MT	German (de)	58.03 (15.52)	**83.37**	**59.95 (15.57)**	0.33 (0.40)	60.37	6.70
		French (fr)	57.28 (12.04)	**83.29**	**62.70 (12.37)**	0.29 (0.39)	60.89	16.90
		Spanish (es)	67.34 (9.70)	**86.48**	**70.65 (11.41)**	0.21 (0.34)	58.62	8.30
	M2M-100	German (de)	54.05 (5.72)	**79.84**	**56.10 (8.13)**	0.29 (0.39)	56.12	10.00
		French (fr)	53.34 (9.80)	**81.45**	**60.00 (8.45)**	0.33 (0.41)	60.25	13.75
		Spanish (es)	59.75 (7.76)	**87.42**	**64.22 (11.49)**	0.23 (0.32)	68.75	5.20

Table 1: Full results containing per dataset, per model and per language pairs. In "Without Context", $\mathcal{A}$ reflects the accuracy of correct gender associations in the translation. In "With Context", $\mathcal{A}_C$ is the accuracy of the overall dataset using the greedy approach. $\mathcal{A}_{all}$ reflects the average accuracy of correct gender associations for all templates applied on all sentences. The values in 'green' represent the highest accuracy score obtained by a language-model pair and the values in bold represent the values where average accuracy improved after applying all templates to the sentences. $\mathcal{CSS}$ represents the `CSS Score` (§3.4). The values in bracket represent the standard deviation for the corresponding metric. $\mathcal{C}_U$ and $\mathcal{C}_L$ represents respectively the percentage of the biased sentences where at least one template / all templates yields the correct prediction of the gender association.

(German) compared with baseline 58.13%). This is a promising result, as even accounting for morphological and heuristic gender detection approximation, it is possible to effectively de-bias a gender stereotype of a profession in translation by adding extra context.

Takeaway. *For most sentences, there exists at least one template which is able to correct the bias.*

3.2 Is the non-greedy strategy also an effective method to reduce translation bias?

Setup. Since the greedy strategy stops the search once it finds a working template, we also investigate a non-greedy strategy. Specifically, we naively apply all T templates to all data points in D, and compute an average accuracy over $D \times T$. We denote this as the *average accuracy*, $\mathcal{A}_{\text{all}}$.

Results. We also observe a marked improvement over averaged accuracy ($\mathcal{A}_{\text{all}}$) across most language-model pairs (Table 1) when we apply all templates from T, *without* using our greedy template selection strategy. While we see slight performance dips in WinoMT, the least improvement is for BUG dataset using OPUS-MT model for German, where we see a significant decrease in performance after adding contexts. However, with M2M model, we observe the highest improvement in the same dataset across all languages. This result is possibly due to OPUS-MT German model being significantly worse in raw translation quality, as we observe in Table 2.

Unsurprisingly, we observe consistent improvement in SimpleGen dataset across all language-model pairs. Since the SimpleGen dataset is constructed from artificially generated data, it is effectively the least ambiguous among the three datasets (Renduchintala and Williams, 2022), hence enabling the models to fully exploit the additional context.

Takeaway. *On average, contexts correct the bias of translations across different datasets, language pairs and models.*

3.3 Does the context impact the translation quality?

Setup. An important consideration for any debiasing measure is to ensure the overall translation quality does not get impacted as an unwanted side effect. In our approach, we observed the choice of delimiter used during the addition of contexts has

Target Language	OPUS-MT		M2M-100	
German(de)	original	45	original	25
	hash (#)	**45**	hash (#)	**25**
	period (.)	44	period (.)	23
	colon (:)	42	colon (:)	24
	semicolon (;)	44	semicolon (;)	24
French(fr)	original	56	original	37
	hash (#)	**54**	hash (#)	**38**
	period (.)	52	period (.)	32
	colon (:)	53	colon (:)	35
	semicolon (;)	53	semicolon (;)	35
Spanish(es)	original	62	original	42
	hash (#)	**62**	hash (#)	**42**
	period (.)	53	period (.)	34
	colon (:)	61	colon (:)	40
	semicolon (;)	61	semicolon (;)	39

Table 2: BLEU Scores of translations for each delimiter

Target Language	Delimiter	OPUS-MT	M2M-100
		$\mathcal{A}_{all}$ (%)	$\mathcal{A}_{all}$ (%)
German(de)	colon (:)	**65.37**	**62.51**
	semi-colon (;)	64.00	62.04
	hash (#)	62.09	58.25
	period (.)	61.39	59.73
French(fr)	colon (:)	58.97	58.46
	semi-colon (;)	58.26	**58.53**
	hash (#)	**59.88**	54.03
	period (.)	56.49	57.25
Spanish(es)	colon (:)	63.38	**67.92**
	semi-colon (;)	**63.98**	66.73
	hash (#)	58.97	59.25
	period (.)	60.45	59.40

Table 3: Results for different delimiters per model and per language pairs for the WinoMT dataset. $\mathcal{A}_{all}$ (%) reflects the average accuracy of correct gender associations for all templates applied to all sentences.

an impact on the translation quality. This was a critical deciding factor in choosing the right delimiter for our de-biasing approach, to ensure negligible impact on the translation quality while balancing for the precision in de-biasing.

For each of the language pairs, we draw a parallel corpus from Tatoeba[7] and draw 300 sentences from this dataset such that they contain the occupations used in the WinoMT dataset. We then translate these sentences after applying the top 50 contexts per sentence for each occupation and calculate the BLEU score after removing the contexts. We conduct our experiments testing the following delimiters: hash (#), period (.), colon (:) and semicolon (;).

Results. We observe that when hash (#) is used as a delimiter, the translation quality of the input sentence does not change compared to its translation without context (Table 2). However, the translation quality degrades when period, colon and semicolon are used as delimiters. This is due to the fact that the delimiters period, and semicolon also naturally occur in the training corpora, leading to their presence within the output translation, making it hard for our post-processing pipeline to ascertain the exact boundary between the context and the sentence.

Thus, while we observe the best de-biasing performance by the use of the delimiter colon (:) (Table 3), we recommend using hash (#) as it provides competitive de-biasing performance while maintaining the best translation quality.

Takeaway. *The choice of delimiter governs the translation quality of the input sentence. Using hash(#) as the delimiter provides the best trade-off between translation quality and removal of bias.*

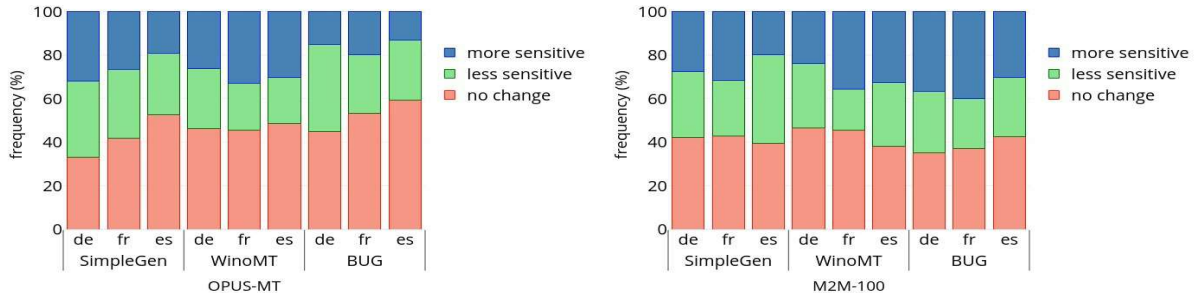
[7] https://tatoeba.org/en/downloads.

3.4 Are certain NMT models more sensitive towards contexts?

Setup. In our preliminary experiments we observe that certain sentences are more sensitive to the addition of contexts than others. NMT systems reportedly are highly sensitive to modifications and perturbations in the input (Fadaee and Monz, 2020; Dankers et al., 2022). Thus, in this study, we aim to quantify the sensitivity of the translation model towards contexts. Concretely, we evaluate the sensitivity of a source sentence X towards *both* relevant and counterfactual contexts - i.e. applying T unique templates with *both* male and female gender signals, irrespective of the gender of the entity in the source sentence (g_X). Our objective is to observe if providing *any* context triggers the model to *change* the gender association of the entity in the translation, i.e., how *sensitive* the model ($g_{\hat{Y}} \neq g_{\hat{Y}_t}, t \in T$) is. Thus, we define a context-sensitivity score (CSS) for sentences X in dataset D as the *percentage of instances where the model changed the target entity gender association on the application of context*.

$$\texttt{CSS}_X = \frac{1}{|D|} \sum_{X \in D} \frac{\sum_{t=1}^{T} I(g_{\hat{Y}} \neq g_{\hat{Y}_t})_{X, X_c \xrightarrow{\mathbb{M}} \hat{Y}, \hat{Y}_t}}{2|T|}$$

where, $|T|$ is the total number of templates, used twice for male and female genders. We compute the CSS score for each language-model pair, and based on empirical evidence we categorize sentences into three distinct bins: *no-change* for CSS



(a) OPUS-MT (b) M2M-100

Figure 2: Distribution of sentences on the basis of sensitivity to the contexts for OPUS-MT and M2M-100 model. We define three bins: *no change* (CSS score is 0), *less sensitive* (CSS Score <= 0.5) and *more sensitive* (CSS Score > 0.5) on addition of contexts.

score = 0, *less sensitive*, for CSS $<= 0.5$; and *more sensitive*, for CSS > 0.5.

Results. Figure 2 shows the frequency distribution of the sentences according to sensitivity range for each model and dataset. We observe M2M-100 model exhibiting higher sensitivity towards the addition of contexts. However, M2M-100 was also the most biased model before the application of contexts (Table 1). This result highlights that the many-to-many pre-training method (as used by M2M-100) is sensitive towards the change in input, which can be leveraged to develop a better, unbiased translation system using our method.

Takeaway. *M2M-100 model proves to be more sensitive to contexts than OPUS-MT.*

3.5 Does the NMT models understand the semantics of the context?

Setup. While we observe marked improvements after adding contexts to the input, it is important to understand whether the semantics of the contexts matter for the model. We thus construct *irrelevant* contexts containing no information about the gender of the target occupation, but having similar syntactic markup as T [8]. We then compute the average accuracy after the application of these contexts using the evaluation approach mentioned in §3.2.

Results. We observe that adding gender-irrelevant contexts results do result in a significant decrease in accuracy (Table 4). These results indicate that the contexts containing gender-relevant information are indeed useful for the model to de-bias the stereotypical occupation bias in translation, and irrelevant contexts hurt the performance by making the sentence more ambiguous for the NMT model.

[8]We provide the full list of irrelevant contexts in Appendix, Table 11

Target Language	Model	$\mathcal{A}(\%)$	$\mathcal{A}_{all}$ (%)	$\mathcal{A}_{all\_gi}$ (%)
German(de)	OPUS-MT	58.03	59.95	53.27
French(fr)		57.28	62.70	57.44
Spanish(es)		67.34	70.65	62.74
German(de)	M2M-100	54.05	56.10	53.09
French(fr)		53.34	60.00	55.23
Spanish(es)		59.75	64.22	58.05

Table 4: Results from adding gender irrelevant contexts to SimpleGen dataset. $\mathcal{A}$ represents the original accuracy of the dataset for the language model pairs. $\mathcal{A}_{all}$ is the average accuracy when correct contexts (with gender signals) are added, whereas $\mathcal{A}_{all\_gi}$ represents the accuracy when gender-irrelevant contexts are added to the dataset.

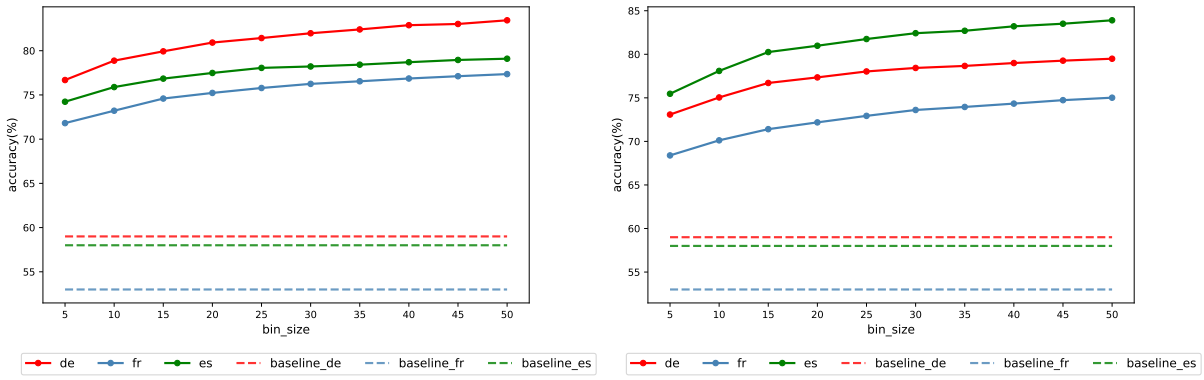
Takeaway. *Both OPUS-MT and M2M-100 understand the meaning of the contexts.*

3.6 Does the de-biasing accuracy vary with gender?

Dimension	Gender	$\mathcal{A}(\%)$	$\mathcal{A}_{all}$ (%)	$\delta(\%)$
Strong Stereotypes	Female	62.66	63.35	0.69
	Male	56.76	59.98	3.22
Weak Stereotypes	Female	43.26	47.76	4.50
	Male	77.81	76.45	-1.36

Table 5: Comparision between male and female on gender on the basis of degree of stereotypes. $\mathcal{A}$ represents the aggregated accuracy of the original dataset without adding context. $\mathcal{A}_{all}$ is the aggregated accuracy with context added and δ represents the improvement in the accuracy after adding relevant contexts.

Setup. In this section, we critically analyze the difference in the accuracy among different gender-stereotypical occupations before and after adding contexts. We hypothesize that the degree of stereotype of the occupation is an important factor to consider while applying a de-biasing technique, as certain occupations can be more challenging for a



(a) OPUS-MT (b) M2M-100

Figure 3: Variation in performance of the accuracy when using greedy strategy with increasing template space evaluated over WinoMT dataset. Baseline values represent the accuracy on the base dataset (with no context).

model to reduce its bias. For example, the sentence *"That nurse is a funny man."* is much harder to debias compared to the sentence *"That teacher is a funny man."*, as the societal stereotype of (`nurse`, `female`) is more prevalent in the data distribution. Thus, we perform the comparison among the degree of stereotypes in the occupations. We classify certain occupations which conform strongly to societal stereotypes using the WinoMT dataset (Stanovsky et al., 2019) and the US Current Population Survey (CPS). The remaining occupations are deemed as "weakly stereotyped". We compare the performance of our approach on these two bins.

Results. From our analysis (Table 5), we first observe that before the addition of the context, the gender-association accuracy is better for strong female stereotypes than male stereotypes. However, post application of contexts, we observe larger improvement in male stereotypical occupations ($\delta = 3.22$) compared to their female counterparts ($\delta = 0.69$). The trend is opposite in case of weakly-stereotypical occupations. In fact, the gender association of male entities suffer post-addition of contexts, while female weakly-stereotypical entities are more correctly annotated ($\delta = 4.50$). Our results indicate that in the case of strong stereotypes, female occupations are much harder to de-bias, owing to their higher prevalence in the data used to train NMT models.

Takeaway. *Sentences having strongly stereotyped male-centric gender bias can be corrected more effectively by using contexts than their female counterparts.*

3.7 What factors determine an effective template?

Setup. We perform an empirical analysis to understand which factors determine the effectiveness of a template for correcting the gender bias in a sentence containing stereotypical occupation bias. To have a diverse set of contexts, we vary the templates with respect to the length of tokens, the number of gender signals (whether the gender is referred to by one or more nouns/pronouns) and the minimum distance of a gender signal from the target profession (i.e, how close the occupation word and the gender signal word(s) are in terms of token distance) [9].

We compare the following properties of the templates: length of tokens (l), number of tokens with gender connotations (s) and the minimum distance from gender connotations from the token containing the target profession (d). For example, the context template *"The* {occupation} *in the next sentence identifies* {m/f-ref-prn} *using the pronouns* {m/f-sbj-prn}/{m/f-obj-prn}" has the features $s = 2$, $d = 3$ and $l = 11$.

Results. As can be observed in Table 6, the number of gender signals (s) and token length (l) is positively correlated to the accuracy of the templates, i.e., longer templates with more number of gender signals lead to a larger improvement in the accuracy. Interestingly, the relative distance(d) of the gender signals from the target profession token is negatively correlated with the accuracy, highlighting issues in co-reference resolution.

Takeaway. *The token length and the number of gender signals used in the template is directly re-*

[9] The full list of gender signal keywords and values used in the templates are described in Appendix Table 8.

Model	Target Language	d	s	l
OPUS-MT	German(de)	-0.18	0.27	0.13
	French (fr)	-0.16	0.26	0.15
	Spanish (es)	-0.20	0.23	0.06
M2M-100	German(de)	-0.17	0.30	0.14
	French (fr)	-0.16	0.32	0.19
	Spanish (es)	-0.24	0.30	0.14

Table 6: Pearson correlation between the sentence accuracy of the templates for each language model pairs and various factors such as token length(l) , number of gender signals(s) and distance of the gender signal (d) from the profession across each language-model pair.

sponsible for its effectiveness.

3.8 What is the time complexity of the de-biasing pipeline?

Setup. While our approach uses a simple methodology to de-bias translations, we understand that iterating through 50 contexts, as we do in our experiments, might seem relatively costlier in production systems. To find the minimum size of the template set that can still reduce the bias in translations, we construct multiple subsets of randomly sampled templates, each having an increasing number of elements ranging from 5 to 50. We then perform our evaluation using the greedy strategy to choose a template from the given sample and evaluate it over the WinoMT dataset. We bootstrap the experiments 100 times to ensure statistical significance.

Results. While the performance does improve with an increase in sample size, even the smallest bin of 5 samples performs significantly well on the given dataset (Figure 3). Thus, we highlight that while the performance of our approach is directly proportional to the number of templates considered, even smaller samples can also lead to considerable improvements in reducing the bias in translations.

Takeaway. *Inference time complexity can be reduced by using less number of templates while maintaining competitive de-biasing accuracy.*

4 Related Work

Several approaches to mitigate gender bias in Neural Machine Translation models have been proposed in the literature. Escudé Font and Costa-jussà (2019) de-bias pre-trained embeddings using hard de-biasing methods proposed by Bolukbasi et al. which removes the gender associations from the representation of English gender-neutral words. However, the effectiveness of this approach has been debatable (Gonen and Goldberg, 2019; Nissim et al., 2020; Goldfarb-Tarrant et al., 2021). Costa-jussà and de Jorge (2020) propose a fine-tuning method using gender-balanced datasets containing an equal amount of masculine and feminine references and observe an improvement in the feminine forms. Closely related to our approach, Vanmassenhove et al. (2018) make use of gender tags (M or F) and prepend them to the source sentence during training and inference. Stanovsky et al. (2019) propose bias reduction with addition of pro-stereotypical adjectives. Saunders et al. (2020) explore the addition of gender tags at the word level. However, all of these approaches require the knowledge of gender metadata, which might not always be feasible to acquire. In our work, we bypass this limitation by using morphological taggers to extract the gender of the target entity. Basta et al. (2020) make use of the preceding sentence as context and concatenate it to the previous sentence. This context doesn't ensure gender-specific information and can be irrelevant with respect to the gender of the target entity. In our experiments, we find that such irrelevant contexts do not help with the de-biasing and in fact hurt the performance in some cases (§3.5). Contrary to this, we use relevant contexts which contain information about the gender of the target entity in the input. Our work is, to the best of our knowledge, the first attempt in exploring the effect of adding relevant sentences as contexts to de-bias translations during inference.

5 Conclusion

In this work, we propose a simple and effective approach to correct stereotypical gender associations for occupations in translations during inference. Specifically, we add an unambiguous context to the input to state-of-the-art NMT models and observe that it enables the model to fix its own gender-bias towards gender-stereotyped occupations. Popular NMT models, such as M2M-100 and OPUS-MT, can effectively learn "in-context" how to correct their own biased translations provided the relevant context along the input. Future work could consist of automatically choosing the correct template to add to the model during inference, or even generating "prompts" dynamically (Shin et al., 2020).

Limitations

- *Cost of iterating through template collection.* For larger models, iterating through a large set

of templates to find the one that fixes the bias can be a costly process. Although we do analyze the complexity of our de-biasing pipeline in §3.8 and show that even a small set of templates can lead to significant improvement, iterating through a larger set for improved performance can be time-consuming.

- *Accuracy of morph tagger.* An important limitation of our approach is that we rely on the accuracy of heuristic morphological taggers based on Stanza (Qi et al., 2020) to determine the gender of the source and target sentences before applying the corresponding context. Thus, a promising future work is to develop a robust gender extraction mechanism for effective de-biasing in complex sentence constructions.

- *Proxy for Gender Bias.* Our approach evaluates gender bias using occupation, which are a commonly used proxy for gender in NLP (Renduchintala et al. (2021), Stanovsky et al. (2019), Bolukbasi et al., Sharma et al. (2021), Zhao et al. (2018b)). We acknowledge that this proxy might provide only a narrow view of gender bias in the NLP domain. Future work should investigate the removal of non-occupation gender-biases using our methodology.

- *Applicability to only binary genders.* We use the existing gender-bias evaluation datasets that take only binary genders (man/woman) into consideration. Building upon this, our approach also takes the narrow view of binary gender. Furthermore, our reliance on US-Census-based occupations during evaluation might be covering only a limited set of occupations and stereotypes. We acknowledge these limitations and advocate that future work should consider non-binary gender as well as intersectional identities.

Ethics Statement

Neural Machine Translation models have been shown to exhibit gender-bias, which also impacts their translation. Our work intends to investigate the usage of contextual information to fix this bias during inference. Our approach can thus contribute to the development of translation systems that are fairer and potentially less harmful. However, the removal of gender bias from translation is an active research problem, and our method can only improve bias to a certain extent. Special care should be taken while deploying NMT systems in production such that specific, harmful biases are pruned before they are served to the end user.

References

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker ? debiasing Word Embeddings. pages 1–25.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

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.

Marzieh Fadaee and Christof Monz. 2020. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Adi Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2021. Evaluating gender bias in natural language inference.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018b. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

A Appendix

A.1 Evaluation dataset details

Our evaluation was carried out on the following datasets.

1. WinoMT - This dataset by ((Stanovsky et al., 2019)) consists of 3888 sentences consisting of male, female and neutral entities. For our experiments, we filter out the neutral ones, leading to a total of 3648 sentences with 1826 male and 1822 female entities.

2. BUG - (Levy et al., 2021) (Balanced-Bug) is a large-scale corpus of 108K diverse real-world English sentences, collected via semi-automatic grammatical pattern matching to evaluate gender bias in various coreference resolution and machine translation models. Of these, we consider the sentences that have a complete overlap with the professions used in WinoMT dataset, leading to 3290 female and 3057 male entities.

3. SimpleGEN - (Renduchintala and Williams, 2022) is a gender translation evaluation set based on gendered noun phrases in which there is a single, unambiguous, correct answer. There are a total of 2664 sentences with 1260 female and 1404 male entities.

A.2 Evaluating the performance of the heuristic morphological tagger

As described in §2, our greedy algorithm ($\mathcal{A}_c$) heavily depends on the heuristic morphological tagger to extract the gender association(g_X) from the source and from the translation ($g_{\hat{Y}}$). Our tagger makes use of the morphological tagger provided by Stanza (Qi et al., 2020) while using heuristics such as the presence of gender-specific words (e.g., he/him in English) to predict the gender. To ensure that the gender predicted by this tagger is accurate, we measure the accuracy of the tagger in Table 7. We observe the accuracy of the morphological tagger is specific to the dataset, and the performance of the same varies according to the complexity of sentence constructions and ambiguity within a dataset. Thus, our morphological tagger performs flawlessly on SimpleGen dataset, which is not surprising given the same dataset is constructed from artificially generated templates and thus exhibits the least ambiguity.

Dataset	Accuracy (%)
WinoMT	99.21
BUG	98.70
SimpleGen	100

Table 7: Accuracy of the custom tagger on various datasets for English (en)

Keywords	Values
f-n	female
m-n	male
f-n-pl	women
m-n-pl	men
f-sbj-prn	she
m-sbj-prn	he
f-n-sg	gal, woman
m-n-sg	guy, man
f-pos-prn	her
m-pos-prn	his
f-obj-prn	her
m-obj-prn	him
f-ref-prn	herself
m-ref-prn	himself

Table 8: Key-Value pairs used to fill the placeholders in the templates while creating the contexts.

In some cases, we observed that the heuristic tagger is unable to predict any gender, leading to the label *unknown* (Table 7). Only a few professions (such as nurse) across particular languages appear to be affected by this issue. On average, we observe this issue in 5% sentences in the entire datasets, with the highest in BUG dataset (7%). Out of the two models, M2M-100 translations appear to display a higher propensity for this issue. In our experiments, these unknown labels contribute to the error of the model, and thus our method could be improved by the use of a better morphological tagger in the future.

B Computational Budget

Our experiments are fairly lightweight in terms of the compute required, as we only run inference, and we avoid training any models. However, inference in NMT is a slow process due to beam search (we used beam size 5), and thus translating the three datasets along with all possible contexts requires approximately 1 day of GPU usage using two NVIDIA P100 GPUs in parallel. We use the HuggingFace (Wolf et al., 2020) repository to

Delimiter	Model	Dataset	Without Context	With Context				
			$\mathcal{A}(\%)$	$\mathcal{A}_C(\%)$	$\mathcal{A}_all$ (%)	CSS	$\mathcal{C}_U$	$\mathcal{C}_L$
Period (.)	OPUS-MT	WinoMT	59.45 (10.34)	**77.81**	59.44 (12.87)	0.20 (0.29)	59.00	1.70
		BUG	67.17 (19.33)	**82.22**	65.16 (17.58)	0.21 (0.20)	55.39	15.17
		SimpleGen	60.88 (12.42)	**90.29**	66.18 (14.67)	0.24 (0.28)	75.94	1.06
	M2M-100	WinoMT	54.9 (13.38)	**80.15**	58.67 (13.10)	0.21 (0.31)	57.84	0.71
		BUG	56.46 (14.98)	**82.54**	59.65 (16.92)	0.38 (0.41)	66.27	23.70
		SimpleGen	55.71 (7.76)	**94.82**	62.08 (11.00)	0.31 (0.69)	88.54	1.51
Hash (#)	OPUS-MT	WinoMT	59.45 (10.34)	**77.52**	60.31 (10.70)	0.31 (0.41)	44.67	8.13
		BUG	67.17 (19.33)	**82.75**	67.92 (17.10)	0.19 (0.31)	48.28	8.87
		SimpleGen	60.88 (12.42)	**84.38**	64.43 (13.12)	0.28 (0.38)	59.96	10.63
	M2M-100	WinoMT	54.9 (13.38)	**80.94**	57.18 (9.99)	0.32 (0.40)	57.86	14.05
		BUG	56.46 (14.98)	**82.82**	65.78 (18.39)	0.37 (0.42)	63.75	22.13
		SimpleGen	55.71 (7.76)	**82.90**	60.11 (9.36)	0.28 (0.37)	61.71	9.65

Table 9: Results aggregated over the language pairs (EN-DE, EN-FR,EN-ES) for each model/dataset. In "Without Context", $\mathcal{A}$ reflects the accuracy of correct gender associations in the translation. In "With Context", $\mathcal{A}_C$ is the accuracy after addition of extra-sentential context using the greedy strategy. $\mathcal{A}_{all}$ reflects the average accuracy of correct gender associations for all templates applied to all sentences, and CSS represents the Context-Sensitivity score. The values in bracket are the standard deviation for the corresponding metric. $\mathcal{C}_U$ and $\mathcal{C}_L$ represent respectively the percentage of the biased sentences where at least one/all templates yields the correct prediction of the gender association.

run inference on the NMT models (OPUS-MT and M2M-100).

Dataset	Model	Language	Without Context		With Context	
			F1_Male	F1_Female	F1_Male	F1_Female
WinoMT	M2M-100	German (de)	69.7	39.2	67.2	44.8
		French (fr)	65.1	34.8	66.7	28.3
		Spanish (es)	67.6	43.2	67.3	47.8
	OPUS-MT	German (de)	66.7	53.0	69.5	51.4
		French (fr)	69.1	34.9	70.6	38.7
		Spanish (es)	66.5	51.1	66.1	48.8
BUG	M2M-100	German (de)	62.7	58.5	72.5	64.5
		French (fr)	59.9	49.0	65.5	54.2
		Spanish (es)	65.6	61.0	77.0	69.7
	OPUS-MT	German (de)	74.9	69.7	70.6	68.5
		French (fr)	62.8	58.1	67.4	59.5
		Spanish (es)	78.8	73.9	77.1	76.2
SimpleGen	M2M-100	German (de)	66.8	31.4	63.8	46.1
		French (fr)	67.4	33.4	67.0	49.8
		Spanish (es)	71.2	39.8	71.9	53.4
	OPUS-MT	German (de)	67.3	50.0	66.7	58.0
		French (fr)	66.3	48.0	66.5	58.5
		Spanish (es)	72.3	61.5	72.9	68.5

Table 10: F1 scores aggregated over the language pairs (EN-DE, EN-FR,EN-ES) for each model/dataset. F1_Male and F1_Female represents the F1 scores of entities with gold label as male and female respectively. Scores before and after adding context are labelled as "Without Context" and "With Context" respectively

A {occupation} appears in the following statement	In the next sentence, there is a revered {occupation}
The {occupation} in the next sentence is alive.	A well-known {occupation} is mentioned in the preceding statement.
The following statement refers to a respected {occupation}	A humble {occupation} is talked about in the following statement.
In the preceding sentence there is a {occupation} who is kind.	In the preceding sentence, there is a helpful {occupation}
The {occupation} in the next sentence is active.	The {occupation} in the next sentence works on the weekend.
The {occupation} in the preceding sentence is talented.	The {occupation} in the previous sentence is often praised by others.
The {occupation} in the following statement likes to travel.	An excellent {occupation} is mentioned in the following phrase.
There is a {occupation} in the next sentence who is trustworthy.	The {occupation} in the next sentence is short-tempered.
There is a loyal {occupation} in the next sentence.	In the previous sentence, there is a {occupation} who paints well.
The next sentence mentions a {occupation}	A {occupation} who is careful is mentioned in the previous phrase.
The {occupation} in the next sentence is energetic.	There is an accomplished {occupation} in the next sentence.
The person in the next sentence is a good {occupation}	The {occupation} in the next sentence will be on holiday next week.
The person in the previous sentence is a famous {occupation}	The following line is about an individual who works as a {occupation}
The next sentence mentions a {occupation} who is a learned professional.	The {occupation} in the following line is a nice person.
In the preceding sentence there is a funny {occupation}	A German {occupation} is mentioned in the previous line.
A renowned {occupation} is mentioned in the following statement.	The {occupation} in the previous sentence is fond of writing.
The {occupation} in the next sentence works hard.	A {occupation} is mentioned in the previous line who has two kids.
Everyone is fond of the {occupation} in the previous sentence.	The {occupation} in the next sentence likes gardening.
The {occupation} in the next sentence is afraid of dogs.	The {occupation} in the following statement is a great individual.
The next sentence talks about a {occupation} who is very polite.	The {occupation} in the next sentence knows Japanese.
There is a competent {occupation} in the next sentence.	There is a {occupation} in the previous statement who owns a car.
The {occupation} in the next statement is smart.	The {occupation} in the next sentence sings well.
The following sentence mentions a strict {occupation}	The {occupation} in the preceding sentence has diabetes.
In the next sentence, there is an experienced {occupation}	People talk highly about the {occupation} in the previous sentence.
The {occupation} in the next sentence is from India.	The {occupation} in the previous sentence was awarded.

Table 11: Templates used for creating gender irrelevant contexts

Templates used for creating contexts	$\mathcal{A}_{mean}$	l	d	s
The {occupation} in the following sentence is appreciated by {m/f-pos-prn} colleagues.	68.82 (0.47)	11	7	1
While referring to {m/f-ref-prn}, the mentioned {occupation} uses the pronouns {m/f-sbj-prn}/{m/f-obj-prn}.	67.77 (0.47)	11	3	1
The {occupation} in the next sentence identifies {m/f-ref-prn} using the pronouns {m/f-sbj-prn}/{m/f-obj-prn}.	67.71 (0.43)	12	5	2
The {occupation} in the following sentence is excellent at {m/f-pos-prn} job.	65.95 (0.48)	11	7	1
The {m/f-n} {occupation} in the next sentence is a responsible {m/f-n-sg}.	65.85 (0.48)	11	0	2
There is a {m/f-n} {occupation} in the following sentence and {m/f-sbj-prn} is a polite {m/f-n-sg}.	65.8 (0.48)	15	0	3
The {m/f-n} {occupation} in the next sentence is valued at {m/f-pos-prn} workplace.	65.64 (0.48)	12	0	2
The {m/f-n} {occupation} in the next sentence is a {m/f-n-sg}.	65.3 (0.48)	10	0	2
A talented {occupation} appears in the next sentence and {m/f-sbj-prn} identifies {m/f-ref-prn} as a {m/f-n-sg}.	65.21 (0.48)	15	6	3
The {m/f-n} {occupation} mentioned in the previous sentence is excellent at {m/f-pos-prn} job.	64.96 (0.48)	13	0	2
A {m/f-n} {occupation} is being talked about in the next sentence.	64.88 (0.48)	11	0	1
The {occupation} in the next sentence is liked by {m/f-pos-prn} coworkers.	64.78 (0.48)	11	7	1
In the following sentence there is a {m/f-n} {occupation} and {m/f-sbj-prn} is a humble {m/f-n-sg}.	64.75 (0.48)	15	0	3
In the following sentence is a {m/f-n} {occupation} and {m/f-sbj-prn} is a polite {m/f-n-sg}.	64.54 (0.48)	14	0	3
The next sentence speaks of a {m/f-n} {occupation}.	64.49 (0.48)	8	0	1
The next sentence talks about a {m/f-n} {occupation}.	64.45 (0.48)	8	0	1
The {occupation} in the following sentence is the best among {m/f-pos-prn} peers.	64.4 (0.48)	12	8	1
The {m/f-n} {occupation} in the previous sentence is well-known for {m/f-pos-prn} expertise.	64.32 (0.48)	12	0	2
The preceding sentence's {m/f-n} {occupation} is well-liked by {m/f-pos-prn} coworkers.	64.14 (0.48)	10	0	2
The person in the following sentence is a {m/f-n} and is the only {m/f-n} {occupation} among {m/f-pos-prn} peers.	64.13 (0.48)	18	0	2
A {m/f-n} {occupation} has been mentioned in the next sentence.	64.09 (0.48)	10	0	1
The {occupation} in the following line enjoys {m/f-pos-prn} work.	63.97 (0.48)	9	5	1
The {occupation} in the next sentence is fond of {m/f-pos-prn} job.	63.45 (0.48)	11	7	1
The preceding sentence's {m/f-n} {occupation} is respected by {m/f-pos-prn} colleagues.	63.39 (0.48)	10	0	2
The {occupation} in the next sentence loves {m/f-pos-prn} job.	63.38 (0.48)	9	5	1
The following statement refers to a {m/f-n} {occupation} who is valued at {m/f-pos-prn} workplace.	63.36 (0.49)	14	0	2
There is a {occupation} in the following sentence and {m/f-pos-prn} gender is {m/f-n}.	63.36 (0.48)	13	5	2
A confident {m/f-n} {occupation} is being spoken about in the next sentence.	63.36 (0.48)	12	0	1
The {m/f-n} {occupation} who was mentioned in the preceeding phrase is well-known for {m/f-pos-prn} knowledge and experience.	63.31 (0.48)	17	0	2
There is a {m/f-n} {occupation} in the next phrase, and {m/f-sbj-prn} is a nice individual.	63.26 (0.49)	15	0	2

Table 12 continued from previous page

Templates used for creating contexts	$\mathcal{A}_{mean}$	l	d	s
The {occupation} in the next sentence is talented and is excellent at {m/f-pos-prn} job.	63.13 (0.47)	14	10	1
The {occupation} in the next sentence is great at {m/f-pos-prn} work.	63.08 (0.48)	11	7	1
The {occupation} in the next sentence is great at {m/f-pos-prn} job.	62.98 (0.48)	11	7	1
The person in the previous sentence is a {m/f-n-sg} and is the only {m/f-n} {occupation} in {m/f-pos-prn} group.	62.9 (0.49)	18	0	3
The preceding sentence talks about a {m/f-n} {occupation} who loves {m/f-pos-prn} job.	62.73 (0.49)	12	0	2
The {occupation} in the next sentence is a smart {m/f-n-sg}.	62.6 (0.48)	10	7	1
The individual in the preceding sentence is a {m/f-n-sg} and is the only {m/f-n} {occupation} amongst {m/f-pos-prn} peers.	62.47 (0.49)	18	0	3
We are talking about a {m/f-n} {occupation} in the following sentence.	62.42 (0.49)	11	0	1
A {m/f-n} {occupation} who enjoys {m/f-pos-prn} work is described in the previous statement.	62.37 (0.49)	13	0	2
The {occupation} in the previous sentence is respected by {m/f-pos-prn} colleagues.	62.33 (0.49)	11	7	1
A {m/f-n} {occupation} is shown in the previous statement, and {m/f-sbj-prn} is a proud {m/f-n-sg}.	62.23 (0.49)	15	0	2
The {occupation} in the previous sentence is known for {m/f-pos-prn} abilities and is a {m/f-n-sg}.	62.18 (0.49)	15	7	2
In the preceding sentence is a {m/f-n} {occupation} who is valued at {m/f-pos-prn} workplace.	62.11 (0.49)	14	0	2
The {occupation} in the previous sentence is known for {m/f-pos-prn} skills and {m/f-sbj-prn} is a {m/f-n-sg}.	62.02 (0.49)	16	7	3
A {m/f-n} {occupation} appears in the following statement.	61.99 (0.49)	8	0	1
The next sentence talks about a {m/f-n-sg} and {m/f-sbj-prn} is a {occupation}.	61.8 (0.49)	12	2	2
The following sentence is about a {m/f-n-sg} whose occupation is {occupation}.	61.71 (0.49)	11	3	1
According to the previous sentence, the {m/f-n-sg} is a {m/f-n} {occupation}.	61.68 (0.49)	11	0	2
In the previous statement, a {occupation} is mentioned, and {m/f-sbj-prn} is a {m/f-n-sg}.	61.63 (0.49)	13	3	2
In the following phrase is a modest {m/f-n} {occupation}.	61.55 (0.49)	9	0	1
In the previous line, the {occupation} is a {m/f-n} who refers to {m/f-ref-prn} as {m/f-sbj-prn}/{m/f-obj-prn}.	61.15 (0.49)	15	2	3
In the next sentence is a {m/f-n-sg} and {m/f-sbj-prn} has been employed as a {occupation}.	61.01 (0.49)	15	5	2
A {m/f-n} {occupation} was mentioned in the previous statement.	60.95 (0.49)	9	0	1
In the previous sentence, the {occupation} identifies {m/f-ref-prn} as a {m/f-n-sg}.	60.95 (0.49)	11	1	2
The next line is about a {m/f-n-sg} who works as a {occupation}.	60.94 (0.49)	12	4	1
The {occupation} in the previous line is a {m/f-n-sg} of confidence.	60.87 (0.49)	11	6	1
The {occupation} in the next sentence uses the pronouns {m/f-sbj-prn}/{m/f-obj-prn}.	60.83 (0.49)	10	7	1
A {m/f-n} {occupation} appears in the previous statement.	60.77 (0.49)	8	0	1
The previous statement {m/f-n-pl}tions a {occupation} and {m/f-sbj-prn} is a {m/f-n-sg}.	60.72 (0.49)	11	1	2
The {occupation} in the preceding line is a trustworthy {m/f-n-sg}.	60.7 (0.49)	10	7	1

Table 12 continued from previous page

Templates used for creating contexts	$\mathcal{A}_{mean}$	l	d	s
The mentioned {occupation} and {m/f-pos-prn} colleagues are honest {m/f-n-pl}.	60.69 (0.49)	9	1	2
In the previous sentence the {occupation} identifies {m/f-ref-prn} as a {m/f-n-sg}.	60.62 (0.49)	11	1	2
In the previous sentence, there is a {m/f-n} {occupation}.	60.59 (0.49)	9	0	1
The {m/f-n-sg} in the previous sentence is a responsible {occupation}.	60.43 (0.49)	10	7	1
The coworkers of the mentioned {m/f-n} {occupation} are also {m/f-n-pl}.	60.39 (0.49)	10	0	2
The previous sentence is about a {m/f-n} {occupation}.	60.31 (0.49)	8	0	1
As per the previous sentence, there is a {m/f-n} {occupation}.	60.27 (0.49)	10	0	1
The {occupation} in the next sentence is a {m/f-n-sg} and {m/f-sbj-prn} uses the pronouns {m/f-sbj-prn}/{m/f-obj-prn}.	60 (0.49)	15	6	3
The {occupation} mentioned here is a polite {m/f-n-sg}.	59.92 (0.49)	8	5	1
The following line is about a {m/f-n-sg} who works as a {occupation}.	59.77 (0.49)	12	4	1
The following statement refers to a {m/f-n-sg} who works as a {occupation}.	59.74 (0.49)	12	4	1
The {m/f-n-sg} in the previous sentence is a sincere {occupation}.	59.61 (0.49)	10	7	1
The {m/f-n-sg} in the previous sentence is a loyal {occupation}.	59.58 (0.49)	10	7	1
There is a {occupation} in the next sentence and {m/f-pos-prn} gender is {m/f-n}.	59.58 (0.49)	13	5	2
The preceding sentence describes a {m/f-n-sg} {occupation} who likes {m/f-pos-prn} profession.	59.58 (0.49)	11	0	2
The {occupation} in the previous sentence identifies {m/f-ref-prn} as a {m/f-n-sg}.	59.15 (0.49)	11	5	2
In the next sentence there is a {occupation} and {m/f-pos-prn} gender is {m/f-n}.	59.09 (0.49)	13	1	2
The {occupation} in the previous sentence is a {m/f-n} {occupation}.	59.02 (0.49)	10	0	1
The {occupation} in the previous sentence is respected at {m/f-pos-prn} workplace.	58.49 (0.5)	11	7	1
Here, the {occupation} is a {m/f-n-sg}.	57.9 (0.5)	6	2	1
The {occupation} in the previous sentence is a {m/f-n-sg} and {m/f-sbj-prn} likes {m/f-pos-prn} job.	57.8 (0.5)	14	6	3
The {occupation} here is a confident {m/f-n-sg}.	56.93 (0.5)	7	4	1
The {occupation} mentioned here is a humble {m/f-n-sg}.	56.71 (0.49)	8	5	1
The correct gender of the {occupation} in the next sentence is {m/f-n}.	56.33 (0.5)	12	5	1
Here, the {occupation} is a {m/f-n-sg} and uses the pronouns {m/f-sbj-prn}/{m/f-obj-prn}.	53.94 (0.49)	11	2	2
The gender of the {occupation} in the next sentence is {m/f-n}.	51.63 (0.49)	11	5	1
The pronouns {m/f-sbj-prn} and {m/f-obj-prn} are used by the {occupation} in the following phrase.	50.34 (0.5)	14	4	2

Table 12: List of templates used to create the contexts for our evaluation.$\mathcal{A}_{mean}$ represents the accuracy for each template (across all lang-model pairs). The values in brackets represent the corresponding standard deviation. (l), (s), (d) represent token length, number of signals and minimum distance of a gender signal from the target profession respectively.The templates are sorted in decreasing order of the mean sentence accuracy.