Does Context Help Mitigate Gender Bias in Neural Machine Translation?

Harritxu Gete[1,2] **Thierry Etchegoyhen**[1]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]University of the Basque Country UPV/EHU
{hgete,tetchegoyhen}@vicomtech.org

Abstract

Neural Machine Translation models tend to perpetuate gender bias present in their training data distribution. Context-aware models have been previously suggested as a means to mitigate this type of bias. In this work, we examine this claim by analysing in detail the translation of stereotypical professions in English to German, and translation with non-informative context in Basque to Spanish. Our results show that, although context-aware models can significantly enhance translation accuracy for feminine terms, they can still maintain or even amplify gender bias. These results highlight the need for more fine-grained approaches to bias mitigation in Neural Machine Translation.

1 Introduction

Neural machine translation (NMT) models tend to exhibit gender bias, originating from their training data (Stanovsky et al., 2019; Saunders and Byrne, 2020). A typical example is the translation of gender-neutral professions in a language like English, into languages with differentiated feminine and masculine forms. In this case, NMT systems often produce translations that reflect gender-stereotypical biases (Troles and Schmid, 2021). Beyond translation errors, bias perpetuation has a clear negative impact overall.

Several studies have addressed gender bias in NMT, highlighting various sources and manifestations of gender bias in NMT models. For example, Prates et al. (2019) and Rescigno and Monti (2023) examined gender bias in commercial machine translation systems, revealing a systematic bias towards masculine translation. A variety of approaches have been explored to mitigate these effects, such as data augmentation by swapping gender-specific words (Zmigrod et al., 2019; Wang et al., 2022), or the incorporation of gender tags in the input to guide the translation process (Vanmassenhove et al., 2018; Corral and Saralegi, 2022). The use of context has also been studied as a potential solution, as context-aware models have been shown to significantly enhance translation quality for specific linguistic phenomena, including gender agreement (Bawden et al., 2018). Sharma et al. (2022) thus demonstrated that using artificial context specifying the expected gender could mitigate gender bias, while Basta et al. (2020) and Currey et al. (2022) suggested that real context could as well. However, a more detailed study is still warranted.

In this work, we further explore the role of context in mitigating gender bias in NMT, by addressing the following question: does context always help mitigate bias in NMT or can it have bias perpetuation effects? To tackle this question, we studied two specific phenomena related to gender bias.

First, we evaluated the performance of context-aware models in the translation of stereotypical professions from English into German and French, measuring translation accuracy on gender-based subsets of the data. Our results in this case indicate that, although context-aware models lead to significantly more feminine forms, this was achieved mainly for professions that are stereotypically viewed as feminine, thus with limited bias mitigation or an actual increase in gender bias.

We then studied the impact on gender bias of non-informative context in Basque to Spanish, i.e., where context lacks gender disambiguating information but nonetheless provides information that may impact the translation. In this case, our results showed significant increases in accuracy for masculine translation options, but notable losses for feminine ones. The use of context was thus detrimental in this case, exacerbating gender biases present in sentence-level models.

Although context can contribute positively to more accurate gender translation, our results highlight the complexity of gender bias in context-aware NMT systems and the need for more fine-grained approaches to mitigate this type of bias.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14788–14794

2 Experimental Setup

2.1 Data

As training data for our sentence-level baselines, for English-German, we selected the data from the WMT 2017 news translation task; for English-French, we used a mix of publicly available sentence-level parallel data to train baseline models, namely Europarl v7, NewsCommentary v10, CommonCrawl, UN, Giga from WMT 2017 and the IWSLT17 TED Talks (Cettolo et al., 2012). We then selected the document-level IWSLT17 dataset to train our context-aware models. For evaluation, we selected the contextual subset of MT-GenEval (Currey et al., 2022) for both language pairs.

For Basque to Spanish, we selected the TANDO^{+} (Gete et al., 2024) dataset to train our sentence- and context-level models, and the COH-TGT:GENDER challenge test set for evaluation, which features gender-related context phenomena where the disambiguating information only occurs on the target side. When using models that only have access to the source target context, this test will allow us to measure the impact of non-informative context.

Table 1 describes corpora statistics for all language-pairs.

		Train	Dev
EN-DE	Doc. Level	5,852,458	2,999
EN-FR	Sent. Level	11,221,790	4,992
	Doc. Level	234,738	5,818
EU-ES	Doc. Level	1,753,726	3,051

Table 1: Training corpora statistics (# parallel sentences)

2.2 Models

We trained sentence-level baselines and concatenation-based context-aware models. Since the MT-GenEval test set contains only one context sentence in the source language, our analysis is focused solely on models with one source context sentence and no target context (2to1 model), following the concatenation approach of Tiedemann and Scherrer (2017).

All models follow the Transformer-base architecture (Vaswani et al., 2017), with the embeddings for source, target and output layers tied. The training was performed with MarianNMT (Junczys-Dowmunt et al., 2018), using the Adam optimiser (Kingma and Ba, 2015) with $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate increased linearly for the first 16,000 training steps and decreased thereafter proportionally to the inverse square root of the corresponding step. The validation data was evaluated every 5,000 checkpoints, and the training process ended if there was no improvement in the perplexity of 10 consecutive checkpoints.

We allocated 8,000MB of memory for training, and automatically selected the largest mini-batch that could fit within this memory on two GPUs for the English to German and Basque to Spanish. For English to French, due to varying data sizes, we used four GPUs for the sentence-level baseline and a single GPU for the context-aware model.

Context-aware models were initialised with sentence-level weights, reinitialising training for English to German and Basque to Spanish, and fine-tuning for English to French.

2.3 Evaluation

Since both selected challenge test sets provide contrastive translations, we calculated models accuracy based on their preference for one translation over the other. As proposed by Post and Junczys-Dowmunt (2024), we also translated the source sentences and classified the translations as correct or incorrect. We will refer to this type of evaluation as *Generative*. To do this, we first identified correct and incorrect tokens by comparing correct translations with their contrastive counterparts. We then measured accuracy by categorising a translation as successful if it generated a correct token without producing any incorrect ones[1]. We also measured incorrect instances, where tokens determined as incorrect are present, and categorised as neutral those cases with neither correct nor incorrect tokens.

3 Results and Analysis

3.1 Stereotypical Professions

We compare sentence-level models with context-aware models in their ability to translate professions from English into German and French, where, contrary to English, gender-specific forms for professions exist. The test selected for evaluation, MT-GenEval, is balanced in terms of feminine and masculine instances, and contains professions stereotypically viewed as masculine or feminine.

[1]This differs from Currey et al. (2022), where any case with no incorrect tokens is categorised as successful.

	Contrast.	Gen. Correct (↑)	Gen. Incorrect (↓)
Sent-level	54.18%	38.00%	36.18%
2to1	**69.27%**	**45.09%**	**27.27%**

Table 2: Overall accuracy in MT-GenEval (EN-DE)

	Contrast.	Gen. Correct (↑)	Gen. Incorrect (↓)
Sent-level	53.41%	42.95%	39.40%
2to1	**69.24%**	**53.41%**	**27.93%**

Table 3: Overall accuracy in the MT-GenEval (EN-FR)

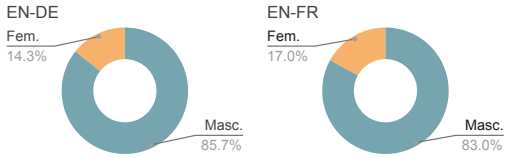

Figure 1: Distribution of sentence-level predictions in stereotipically feminine professions.

The results for English to German and English to French are shown in Tables 2 and 3, respectively. Overall, context-aware models significantly improved accuracy, in both contrastive and generative evaluations for both language pairs. To assess whether these improvements correlate with a reduction in gender bias, we analysed these results in more details.

We manually divided the test into two subsets based on the expected gender and calculated accuracy for each category, with the results shown in Tables 4 to 7. The high scores in the masculine category, and the low scores in the feminine category, in both language pairs with the sentence-level baselines, suggest a significant bias in the data, causing the models to predominantly translate professions into the masculine form. Even when translating stereotypically feminine professions, the models generally tend to favour the masculine form as shown in Figure 1.

When comparing the results with and without context, improvements are most notable in the feminine category, which increases by about 30 percentage points in both language pairs. In contrast, in the masculine category the increase is less than 4 points for English to French, and there is no significant difference for English to German. It thus seems that context might help mitigate the bias towards masculine forms when translating professions.

We further analysed these results by focusing on

	Contrast.	Gen. Correct (↑)	Gen. Incorrect (↓)
Sent-level	**92.00%**	**71.09%**	**5.64%**
2to1	**91.45%**	67.27%	**6.73%**

Table 4: Accuracy over masculine forms (EN-DE)

	Contrast.	Gen. Correct (↑)	Gen. Incorrect (↓)
Sent-level	92.55%	**76.91%**	**6.00%**
2to1	**96.36%**	**77.09%**	**4.73%**

Table 5: Accuracy over masculine forms (EN-FR)

	Contrast.	Gen. Correct (↑)	Gen. Incorrect (↓)
Sent-level	16.36%	4.91%	66.73%
2to1	**47.09%**	**22.91%**	**47.82%**

Table 6: Accuracy over feminine forms (EN-DE)

	Contrast.	Gen. Correct (↑)	Gen. Incorrect (↓)
Sent-level	14.21%	8.93%	72.86%
2to1	**42.08%**	**29.69%**	**51.18%**

Table 7: Accuracy over feminine forms (EN-FR)

professions categorised as stereotypically feminine or masculine, dividing the results into four subcategories based on the type of profession and the expected gender, and assessing accuracy using contrastive evaluation. Results in Tables 8 and 9 show that, for both language pairs, the largest gains from context originate from the feminine category as expected, but to a much larger degree for professions stereotypically seen as feminine.

The differences in accuracy for the feminine category, between professions classified as feminine and those classified as masculine, thus increased with the use of context. This was the case both in English-German, where the baselines reflected almost no initial differences between the two groups, and English-French where the initial baseline differences were amplified.

These results indicate that context-aware models can help mitigate the bias in favour of masculine translations, significantly increasing the use of feminine forms, but at the same time maintaining or increasing the differences in accuracy between instances that belong to an existing stereotype and those that do not.

Examples where context either helps or fails to improve gender translation are shown in Table 10.

Sentence-level	Actual gender	
	Feminine	Masculine
Stereot. Fem.	17.33%	88.67%
Stereot. Masc.	18.00%	97.33%

Context-aware	Actual gender	
	Feminine	Masculine
Stereot. Fem.	56.00% (↑ 38.67%)	89.33% (↑ 0.66%)
Stereot. Masc.	40.00% (↑ 22.00%)	93.33% (↓ 4.00%)

Table 8: Accuracy over gender-specific subsets (EN-DE)

Sentence-level	Actual gender	
	Feminine	Masculine
Stereot. Fem.	21.33%	87.33%
Stereot. Masc.	8.72%	97.33%

Context-aware	Actual gender	
	Feminine	Masculine
Stereot. Fem.	49.33% (↑ 28.00%)	96.67% (↑ 9.34%)
Stereot. Masc.	30.87% (↑ 22.15%)	97.33% (=)

Table 9: Accuracy over gender-specific subsets (EN-FR)

Example 1: Context improves gender prediction (EN-DE)			
Context	Source Sentence	Sentence-Level Translation	Context-Aware Translation
She is a divorced single mother with two children who lives (...)	Levavasseur is a trained nurse's aide.	Levavasseur ist ein ausgebildeter Pflegeassistent. ✗ (MASC.)	Levavasseur ist eine ausgebildete Krankenschwester. ✓ (FEM.)
Example 2: Context fails to improve gender prediction (EN-FR)			
Context	Source Sentence	Sentence-Level Translation	Context-Aware Translation
Flinn's case, due in part to her high visibility in Air Force recruitment advertisements, (...)	This is an issue about an officer, entrusted to fly nuclear weapons, who lied.	Il s'agit d'un officier, chargé de piloter des armes nucléaires, qui a menti. ✗ (MASC.)	C'est un problème à propos d'un officier, chargé de faire voler des armes nucléaires, qui a menti. ✗ (MASC.)

Table 10: Examples from MT-GenEval where context either helps or fails to improve gender prediction in translations. Gender markers are underlined in the context.

3.2 Non-informative Context

In this section, we focus on the effect of introducing context that lacks relevant disambiguating information for the translation into the correct gender. Although this type of analysis is unusual, because standard tests aim to evaluate if a model is able to use contextual information to handle extrasentential phenomena, context does not always provide relevant information to solve a specific phenomenon, but is still nonetheless present and impacts the actual translation.

For this analysis, we used the COH-TGT:GENDER test of TANDO^{+} for Basque to Spanish translation, where the disambiguating information is on the target side. Specifically, we analysed the parliamentary domain subset of the test, as the results of Gete et al. (2024) indicate a clear tendency towards masculine translation in this domain. Since we compare a sentence-level model with a 2to1 model that only has access to the source context, we ensure that neither model has access to the disambiguating information on the target side.

The contrastive evaluation results shown in Table 11 indicate that, as expected, the overall accuracy does not vary with or without context with uninformative context. However, when dividing the results by gender category, the accuracy for the masculine category increased to 98%, while the accuracy for the feminine category decreased from 10% to 2%. Using uninformative context actually increased gender bias in this case.

Additionally, we performed a generative evaluation over feminine forms only (Table 11). In line with the contrastive results, the sentence-level results indicate a clear tendency towards masculine translation in this domain, with 70% of incorrect instances in the feminine subset. The results also confirm that the use of context exacerbates this tendency, resulting in even fewer correct feminine forms translation.

A possible explanation for these results is that, even if the context does not contain relevant disambiguating information, it may still include domain-related information. In the political domain, which

	Contrastive accuracy			Generative accuracy	
	Total	Masculine	Feminine	Feminine correct	Feminine incorrect
Sentence-level	50%	90%	10%	4%	70%
2to1	50%	98%	2%	0%	74%
2to1* - ctxTED	50%	88%	12%	4%	70%

Table 11: Accuracy on the parliamentary subsets of COH-TGT:GENDER (EU-ES)

Context	Context Domain	Source Sentence	Translation
-	-	Baina kasu honetan, zure ego-era deitoratuz, berak du arra-zoia.	Pero en este caso, lamentando su situación, **ella** tiene razón. *(But in this case, regretting your situation, **she** is right.)*
Ulertzen ez baduzu ere, bere Gobernukideak zuzendu egin du bere saileko (...) *(Even if you don't understand it, his/her government partner has rectified the department's (...))*	Parliament	Baina kasu honetan, zure ego-era deitoratuz, berak du arra-zoia.	Pero en este caso, lamentando su situación, **él** tiene razón. *(But in this case, regretting your situation, **he** is right.)*
Inoiz ez dit inork hori esan. *(No one has ever told me that.)*	TED	Baina kasu honetan, zure ego-era deitoratuz, berak du arra-zoia.	Pero en este caso, lamentando su situación, **ella** tiene razón. *(But in this case, regretting your situation, **she** is right.)*

Table 12: Examples from COH-TGT:GENDER (EU-ES) where uninformative context impacts gender selection in contrastive evaluation, negatively (Parliament) or neutrally (TED).

strongly favours masculine translations, the introduction of context might reinforce this bias.

To test this hypothesis, we evaluated the same sentences from the political domain, but using context from another domain, namely the TED talks subset of the COH-TGT:GENDER test (ctxTED), which also lacks disambiguating information in the source context side.

In the contrastive evaluation, there was a 10 percentage point difference favoring the feminine category over the original 2to1 model, at the expense of the masculine category. In the generative evaluation, the results were identical to sentence-level results. It thus seems that an uninformative context from the same domain can be a factor in actually increasing gender bias.

Table 12 provides examples of non-informative context impacting gender translation.

4 Conclusions

This study investigated the impact of context-aware models on mitigating gender bias in NMT, focusing on the translation of professions from English into German and French, as well as translation with uninformative context in Basque to Spanish.

Our results show that, although contextual models can significantly improve the overall translation accuracy on gender-specific terms, this was achieved mainly over stereotypically feminine professions. The use of context actually increased the disparity between stereotypical genders in this case. Non-informative context was also shown to increase gender-related bias in a domain with strong latent bias, when using context from a different domain had no such effects.

These results underscore the need for more comprehensive approaches to bias in NMT, including more specific evaluations over balanced datasets. Novel mitigation techniques and a deeper understanding of the impact of context will also be needed to further address translation bias.

5 Limitations

The experiments were conducted exclusively with 2to1 context-aware models. This choice was influenced by the characteristics of one of the test sets, which only provided one context sentence in the source language. As a result, our findings are specific to this type of model, and further research is needed to explore the effects of other context-aware architectures. Expanding the range of test sets and domains would also provide a more comprehensive understanding of biased translation with and without context.

6 Ethical Considerations

Our analysis focused solely on binary gender categories, examining translations in terms of masculine and feminine forms. This binary perspective excludes individuals who do not identify with either of these normative genders. Addressing this issue would require developing and incorporating more inclusive linguistic resources and methodologies that recognise and respect non-binary identities.

Acknowledgments

We wish to thank the anonymous ARR reviewers for their helpful comments. This work was partially supported by the Department of Economic Development and Competitiveness of the Basque Government (Spri Group) via funding for project ADAPT-IA (KK-2023/00035).

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 Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for NMT involving genderless languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 165–176, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Harritxu Gete, Thierry Etchegoyhen, Gorka Labaka, Ander Corral, Xabier Saralegi, Nora Aranberri, David Ponce, Igor Ellakuria Santos, and Maite Martin. 2024. Tandoˆ+: Corpus and baselines for document-level machine translation in basque-spanish and basque-french.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederick P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Matt Post and Marcin Junczys-Dowmunt. 2024. Escaping the sentence-level paradigm in machine translation. *Preprint*, arXiv:2304.12959.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate. *Preprint*, arXiv:1809.02208.

Argentina Rescigno and Johanna Monti. 2023. Gender bias in machine translation: a statistical evaluation of google translate and deepl for english, italian and german. pages 1–11.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984, Abu Dhabi, United Arab Emirates. 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 Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. 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.