Gender Bias Amplification During Speed-Quality Optimization in Neural Machine Translation

Adithya Renduchintala, Denise Diaz[*1], Kenneth Heafield, Xian Li, Mona Diab
Facebook AI, [1]Independent Researcher
{adirendu,kheafield,xianl,mdiab}@fb.com
denisedeediaz@gmail.com

Abstract

Is bias amplified when neural machine translation (NMT) models are optimized for speed and evaluated on generic test sets using BLEU? We investigate architectures and techniques commonly used to speed up decoding in Transformer-based models, such as greedy search, quantization, average attention networks (AANs) and shallow decoder models and show their effect on gendered noun translation. We construct a new gender bias test set, SimpleGEN, based on gendered noun phrases in which there is a single, unambiguous, correct answer. While we find minimal overall BLEU degradation as we apply speed optimizations, we observe that gendered noun translation performance degrades at a much faster rate.

1 Introduction

Optimizing machine translation models for production, where it has the most impact on society at large, will invariably include speed-accuracy trade-offs, where accuracy is typically approximated by BLEU scores (Papineni et al., 2002) on generic test sets. However, BLEU is notably not sensitive to specific biases such as gender. Even when speed optimizations are evaluated in shared tasks, they typically use BLEU (Papineni et al., 2002; Heafield et al., 2020) to approximate quality, thereby missing gender bias. Furthermore, these biases probably evade detection in shared tasks that focus on quality without a speed incentive (Guillou et al., 2016) because participants would not typically optimize their systems for speed. Hence, it is not clear if Neural Machine Translation (NMT) speed-accuracy optimizations amplify biases. This work attempts to shed light on the *algorithmic choices* made during speed-accuracy optimizations and their impact on gender biases in an NMT system, complementing existing work on data bias.

source	That physician is a funny lady!
reference	¡Esa médica/doctora es una mujer graciosa!
system A	¡Ese médico es una dama graciosa!
system B	¡Ese médico es una dama divertida!
system C	¡Ese médico es una mujer divertida!
system D	¡Ese médico es una dama divertida!

Table 1: Translation of a simple source sentence by 4 different commercial English to Spanish MT systems. All of these systems fail to consider the token "lady" when translating the occupation-noun, rendering it in with the masculine gender "doctor/médico".

We explore optimizations choices such as (i) search (changing the beam size in beam search); (ii) architecture configurations (changing the number of encoder and decoder layers); (iii) model based speedups (using Averaged attention networks (Zhang et al., 2018)); and (iv) 8-bit quantization of a trained model..

Prominent prior work on gender bias evaluation forces the system to "guess" the gender (Stanovsky et al., 2019a) of certain occupation nouns in the source sentence. Consider, the English source sentence "That physician is funny.", containing no information regarding the physician's gender. When translating this sentence into Spanish (where the occupation nouns are explicitly specified for gender), an NMT model is forced to guess the gender of the physician and choose between masculine forms, doctor/médico or feminine forms doctora/médica. While investigating bias in these settings is valuable, in this paper, we hope to highlight that the problem is much worse — despite an explicit gender reference in the sentence, NMT systems still generate the wrong gender in translation (see Table 1), resulting in egregious errors where not only is the gender specification incorrect but the generated sentence also fails in morphological gender

[*] This work conducted while author was working at Facebook AI.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 99–109
August 1–6, 2021. ©2021 Association for Computational Linguistics

Templates			That `f/m-occ-sg` is a funny `f/m-n-sg`! My `f/m-rel` is a `f/m-occ-sg`.

Keywords			`f-occ-sg` = {nurse, nanny...}
			`m-occ-sg` = {physician, mechanic...}
			`f-rel` = {sister, mother..}
			`m-rel` = {brother, father...}
			`f-n-sg` = {woman, gal, lady...}
			`m-n-sg` = {man, guy...}

Generated	pro.	MoMc	That engineer is a funny guy! My father is a mechanic.
		FoFc	That nanny is a funny lady! My mother is a nurse.
	anti.	MoFc	That mechanic is my funny woman! My sister is a physician.
		FoMc	That nurse is funny man! My brother is a nanny.

Table 2: Example Templates, Keywords and a sample of the resulting generated source sentences.

agreement. To focus on these egregious errors, we construct a new data set, SimpleGEN. In SimpleGEN, all source sentences include an occupation noun (such as “mechanic”, “nurse” etc.) *and* an unambiguous “signal” specifying the gender of the person being referred to by the occupation noun. For example, we modify the previous example to “That physician is a funny *lady*”. We call our dataset “Simple” because it contains all the information needed by a model to produce correctly gendered occupation nouns. Furthermore, our sentences are short (up to 12 tokens) and do not contain complicated syntactic structures. Ideally, SimpleGEN should obviate the need for an NMT model to incorrectly guess the gender of occupation nouns, but using this dataset we show that gender translation accuracy, particularly in female context sentences (see Section 2), is negatively impacted by various speed optimizations at a *greater rate* than a drop in BLEU scores. A small drop in BLEU can hide a large increase in biased behavior in an NMT system. Further illustrating how insensitive BLEU is as a metric to such biases.

2 SimpleGEN: A gender bias test set

Similar to Stanovsky et al. (2019b), our goal is to provide English input to an NMT model and evaluate if it correctly genders occupation-nouns. We focus on English to Spanish (En-Es) and English to German (En-De) translation directions as occupation-nouns are explicitly specified for gender in these target languages while English is underspecified for such a morphological phenomenon which forces the model to attend to contextual clues. Furthermore, these language directions are considered “high-resource” and often cited as exemplars for advancement in NMT.

A key differentiating characterization of our test set is that there is no ambiguity about the gender of the occupation-noun. We achieve this by using carefully constructed templates such that there is enough contextual evidence to *unambiguously specify* the gender of the occupation-noun. Our templates specify a “scaffolding” for sentences with *keywords* acting as placeholders for *values* (see Table 2). For the occupation keywords such as `f-occ-sg` and `m-occ-sg`, we select the occupations for our test set using the U.S Department of Labor statistics of high-demand occupations.[1] A full list of templates, keywords and values is in table A6. Using our templates, we generate English source sentences which fall into two categories: (i) *pro-stereotypical* (pro) sentences contain either stereotypical male occupations situated in male contexts (MOMC) or female occupations in female contexts (FOFC), and (ii) *anti-stereotypical* (anti) sentences in which the context gender and occupation gender are mismatched, i.e. male occupations in female context (MOFC) and female occupations in male contexts (FOMC). Note that we use the terms “male context” or “female context” to categorize sentences in which there is an unambiguous signal that the occupation noun refers to a male or female person, respectively. We generated 1332 pro-stereotypical and anti-stereotypical sentences, 814 in the MOMC and MOFC subgroups and 518 in the FOMC and FOFC subgroups (we collect more male stereotypical occupations compared to female, which causes this disparity).

To evaluate the translations of NMT models on SimpleGEN, we also create an occupation-noun bilingual dictionary, that considers the number and gender as well as synonyms for the occupations. For example for the En-Es direction, the English occupation term ‘physician”, has corresponding entries for its feminine forms in Spanish as “doctora” and “médica” and for its masculine forms “doctor” and “médico” (See table A8 for our full dictionary). By design, non-occupation keywords such as `f-rel` and `f-n-sg` specify the expected gender of the occupation-noun on the target side, enabling dictionary based correctness verification.

3 Speeding up NMT

There are several “knobs” that can be tweaked to speed up inference for NMT models. Setting the beam-size (bs) to 1 during beam search is likely the

[1] https://www.dol.gov/agencies/wb/data/high-demand-occupations

Source	That physician is a funny lady!	Label
Translations	¡Esa doctora es una mujer graciosa!	Correct
	¡Esa médica es una mujer feliz!	Correct
	¡Ese médico es una mujer graciosa!	Incorrect
	¡Ese medicación es una mujer graciosa!	NA

Table 3: Our evaluation protocol with an example source sentence and four example translations.

simplest approach to obtain quick speedups. Low-bit quantization (INT8) is another recent approach which improves decoding speed and reduces the memory footprint of models (Zafrir et al., 2019; Quinn and Ballesteros, 2018).

For model and architecture based speedups, we focus our attention on Transformer based NMT models which are now the work-horses in NLP and MT (Vaswani et al., 2017). While transformers are faster to train compared to their predecessors, Recurrent Neural Network (RNN) encoder-decoders (Bahdanau et al., 2014; Luong et al., 2015), transformers suffer from slower decoding speed. Subsequently, there has been interest in improving the decoding speed of transformers.

Shallow Decoders (SD): Shallow decoder models simply reduce the decoder depth and increase the encoder depth in response to the observation that decoding latency is proportional to the number of decoder layers (Kim et al., 2019; Miceli Barone et al., 2017; Wang et al., 2019; Kasai et al., 2020). Alternatively, one can employ SD models without increasing the encoder layers resulting in smaller (and faster) models.

Average Attention Networks (AAN): Average Attention Networks reduce the quadratic complexity of the decoder attention mechanism to linear time by replacing the decoder-side self-attention with an average-attention operation using a fixed weight for all time-steps (Zhang et al., 2018). This results in a $\approx$ 3-4x decoding speedup over the standard transformer.

4 Experimental Setup

Our objective is not to compare the various optimization methods against each other, but rather surface the impact of these algorithmic choices on gender biases. We treat all the optimization choices described in section 3 as data points available to conduct our analysis. To this end, we train models with all combinations of optimizations described in section 3 using the Fairseq toolkit (Ott et al., 2019). Our baseline is a standard large transformer with a $(6, 6)$ encoder-decoder layer configuration. For our SD models we use the following encoder-decoder layer configurations $\{(8, 4), (10, 2), (11, 1)\}$. We also train smaller shallow decoder (SSD) models without increasing the encoder depth $\{(6, 4), (6, 2), (6, 1)\}$. For each of these 7 configurations, we train AAN versions. Next, we save quantized and non-quantized versions for the 14 models, and decode with beam sizes of 1 and 5. We repeat our analysis for English to Spanish and English to German directions, using WMT13 En-Es and WMT14 En-De data sets, respectively. For the En-Es we limited the training data to $4M$ sentence pairs (picked at random without replacement) to ensure that the training for the two language directions have comparable data sizes. We apply Byte-Pair Encoding (BPE) with $32k$ merge operations to the data (Sennrich et al., 2016).

We measure decoding times and BLEU scores for the model's translations using the WMT test sets. Next, we evaluate each model's performance on SimpleGEN, specifically calculating the percent of correctly gendered nouns, incorrectly gendered nouns as well as inconclusive results. Table 3 shows an example of our evaluation protocol for an example source sentences and four possible translations. We deem the first two as correct even though the second translation incorrectly translates "funny" as "feliz" since we focus on the translation of "physician" only. The third translation is deemed incorrect because the masculine form "médico" is used and the last translation is deemed inconclusive since it is in the plural form. We average these metrics over 3 trials, each initialized with different random seeds. We obtained 56 data points for each language direction.

5 Analysis

Table 4a shows the performance of 6 selected models including a baseline transformer model with 6 encoder and decoder layers. The first two columns (time and BLEU) were computed using the WMT test sets. The remaining columns report metrics using SimpleGEN. The algorithmic choices resulting in the highest speed-up, result in a 1.5% and 4% relative drop in BLEU for En-Es and En-De, respectively (compared to the baseline model). The pro-stereotypical (pro) column shows the percentage correct gendered translation for sentences where the occupation gender matches the context gender. As expected the accuracies are relatively high (80.9 to 77.7) for all the models. The

direction	model	time(s)	BLEU	pro	anti	Δ	FOFC	MOFC	ΔFC	MOMC	FOMC	ΔMC
En-Es	baseline (bl)	3,662.8	33.2	80.9	44.2	36.7	69.4	41.7	27.7	88.2	48.1	40.0
	bl w/ bs=1	2,653.1	32.7	79.5	44.9	34.6	68.4	42.8	25.6	86.6	48.2	38.4
	bl w/ AAN	3,009.4	32.9	78.6	37.8	40.8	67.4	33.6	33.8	85.6	44.3	41.3
	bl w/ SD(10, 2)	2,241.7	32.9	77.9	38.1	39.8	67.3	35.9	31.4	84.6	41.7	42.9
	bl w/ SSD(6, 2)	1,993.5	32.7	77.7	38.7	39.0	66.0	33.8	32.2	85.1	46.3	38.8
	bl w/ quantization	2,116.1	32.7	79.8	41.4	38.4	67.0	37.2	29.8	88.0	48.1	39.8
	max rel. % drop	45.6	1.5	3.9	15.1		4.9	21.4		4.0	13.5	
En-De	baseline (bl)	3,653.0	27.2	67.7	39.7	28.0	57.5	31.6	25.9	74.2	52.3	21.8
	bl w/ bs=1	2,504.5	26.7	65.0	39.2	25.8	51.5	29.7	21.8	73.5	54.0	19.5
	bl w/ AAN	2,600.0	27.1	68.5	33.0	35.5	58.0	23.9	34.1	75.3	47.4	27.8
	bl w/ SD(10, 2)	1,960.8	27.1	67.5	32.6	35.0	57.7	26.5	31.2	73.8	46.7	27.1
	bl w/ SSD(6, 2)	2,091.0	27.0	66.9	35.9	31.0	56.6	30.3	26.2	73.5	44.6	28.9
	bl w/ quantization	2,205.1	26.1	63.2	33.2	30.0	50.5	24.6	25.9	71.3	46.8	24.6
	max rel. % drop	46.3	4.0	6.5	17.9		13.0	22.1		5.3	9.5	

(a) Each speed-up optimization individually.

direction	model	time(s)	BLEU	pro	anti	Δ	FOFC	MOFC	ΔFC	MOMC	FOMC	ΔMC
En-Es	baseline	3,662.8	33.2	80.9	44.2	36.7	69.4	41.7	27.7	88.2	48.1	40.0
	+bs=1	2,653.1	32.7	79.5	44.9	34.6	68.4	42.8	25.6	86.6	48.2	38.4
	+AAN	1,971.8	32.5	77.4	38.5	38.9	67.4	34.9	32.5	83.7	44.0	39.7
	+SD(10, 2)	1,164.2	32.1	75.3	36.2	39.1	57.1	31.7	25.3	86.8	43.2	43.6
	+SSD(6, 2)	1,165.7	31.9	78.6	40.4	38.2	66.9	36.3	30.5	86.0	46.8	39.2
	+quantization	679.6	31.1	73.1	34.9	38.2	58.7	29.5	29.2	82.3	43.4	38.8
	max rel. % drop	81.4	6.3	9.6	22.3		17.7	31.0		6.7	10.4	
En-De	baseline	3,653.0	27.2	67.7	39.7	28.0	57.5	31.6	25.9	74.2	52.3	21.8
	+bs=1	2,504.5	26.7	65.0	39.2	25.8	51.5	29.7	21.8	73.5	54.0	19.5
	+AAN	2,176.6	26.3	66.7	32.2	34.5	54.6	22.1	32.5	74.4	48.1	26.3
	+SD(10, 2)	1,332.3	25.8	64.2	29.1	35.1	50.3	22.2	28.1	73.0	44.7	28.3
	+SSD(6, 2)	1,153.2	25.7	64.7	28.9	35.9	53.9	19.9	34.1	71.6	43.0	28.6
	+quantization	732.6	24.7	61.0	23.3	37.6	46.3	14.8	31.5	70.3	36.7	33.6
	max rel. % drop	79.9	9.2	9.9	41.3		19.5	53.2		5.5	29.8	

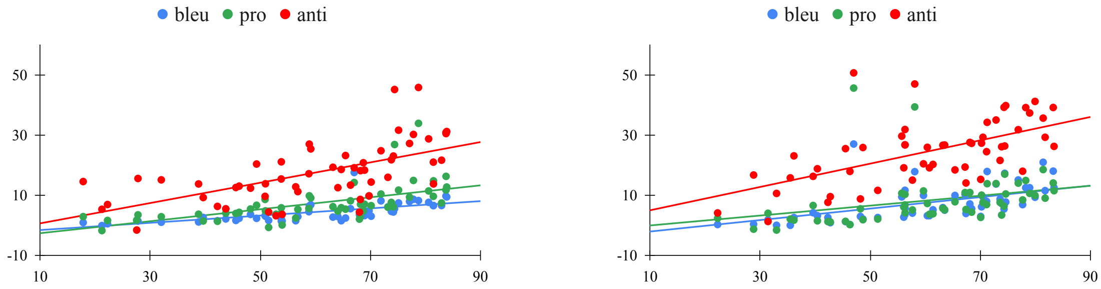
(b) "Stacked" speed-up optimizations.

Table 4: Results showing the effect of speed-up optimizations applied individually (in Table 4a) and stacked in Table 4b). We selected 6 models in both sections to highlight their effect on decoding time, BLEU and the % correctness on gender-bias metrics. The last row for each section (and each direction), shows the relative % drops in all the metrics between the fastest optimization method and the baseline. For example, for En-Es the relative % drop of decoding time for Table 4a is calculated as $100 * (3662.8 - 1993.5)/3662.8$.

last row in each section shows the *maximum relative drop* in each metric. We find that for the pro-stereotypical column the maximum relative drop is 1.5 and 6.5 for Spanish and German, respectively, which is similar to the relative change in BLEU scores. However, we find that the models are able to perform better on MOMC compared to FOFC suggesting biases even within the pro-stereotypical setting. In the anti-stereotypical (anti) column, we observe below-chance accuracies of only 44.2% and 39.7% for the two language directions, even from our best model. Columns FOFC and MOFC, show the difference in performance for sentences in the female context (FC) category in the presence of a stereotypical female occupation versus a stereotypical male occupation. We see a large imbalance in performance in these two columns summarized in ΔFC. Similarly, ΔMC summarizes the drop in performance when the model is confronted with stereotypical female occupations in a male context when compared to a male occupation in a male context. This suggests that the transformer's handling of grammatical agreement especially in cases where an occupation and contextual gender mismatch could be improved. The speedups disproportionately affect female context (FC) sentences across all categories.

In terms of model choices, we find that AANs deliver moderate speed-ups and minimal BLEU reduction compared to the baseline. However, AANs suffer the most degradation in terms of gender-bias. Δ, ΔFC and ΔMC are the highest for the ANN model in both language directions. On the other hand, greedy decoding with the baseline model has the smallest degradation in terms of gender-bias.

While Table 4a reveals the effect of select individual model choices, NMT practitioners, typically "stack" the optimization techniques together for large-scale deployment of NMT systems. Table 4b shows that stacking can provide $\approx 80 - 81\%$ relative drop in decoding time. However, we again see a disturbing trend where large speedups and small BLEU drops are accompanied with large drops in gender test performance. Again, FC sentences disproportionately suffer large drops in accuracy, particularly in MOFC in the En-De direction, where



(a) English-Spanish

(b) English-German

Figure 1: Plots showing *relative percentage drop* of BLEU and gender-test metrics on the y-axis and *relative percentage drop* in decoding time in the x-axis FOr the two language directions analyzed. A breakdown of pro and anti into their constituent groups MOMC, FOFC, MOFc and FOMC is shown in Appendix A.3.

we see a 53.2% relative drop between the baseline and the fastest optimization stack.

While tables 4a and 4b show select models, we illustrate and further confirm our findings using all the data points (56 models trained) using scatter plots shown in fig. 1. We see that relative % drop in BLEU aligns closely with the relative % drop in gendered translation in the pro-stereotypical setting. In the case of German, the two trendlines are virtually overlapping. However, we see a steep drop for the anti-stereotypical settings, suggesting that BLEU scores computed using a typical test set only captures the stereotypical cases and even small reduction in BLEU could result in more instances of biased translations, especially in female context sentences.

6 Related Work

Previous research investigating gender bias in NMT has focused on data bias, ranging from assessment to mitigation. For example, Stanovsky et al. (2019b) adapted an evaluation data set for co-reference resolution to measure gender biases in machine translation. The sentences in this test set were created with ambiguous syntax, thus forcing the NMT model to "guess" the gender of the occupations. In contrast, there is always an unambiguous signal specifying the occupation-noun's gender in SimpleGEN. Similar work in speech-translation also studies contextual hints, but their work uses real-world sentences with complicated syntactic structures and sometimes the contextual hints are across sentence boundaries resulting in gender-ambiguous sentences (Bentivogli et al., 2020).

Zmigrod et al. (2019) create a counterfactual data-augmentation scheme by converting between masculine and feminine inflected sentences. Thus, with the additional modified sentences, the augmented data set equally represents both genders. Vanmassenhove et al. (2018), Stafanovičs et al. (2020) and Saunders et al. (2020) propose a data-annotation scheme in which the NMT model is trained to obey gender-specific tags provided with the source sentence. While Escudé Font and Costa-jussà (2019) employ pre-trained word-embeddings which have undergone a "debiasing" process (Bolukbasi et al., 2016; Zhao et al., 2018). Saunders and Byrne (2020) and Costa-jussà and de Jorge (2020) propose domain-adaptation on a carefully curated data set that "corrects" the model's misgendering problems. Costa-jussà et al. (2020) consider variations involving the amount of parameter-sharing between different language directions in multilingual NMT models.

7 Conclusion

With the current mainstreaming of machine translation, and its impact on people's everyday lives, bias mitigation in NMT should extend beyond data modifications and counter bias amplification due to algorithmic choices as well. We focus on algorithmic choices typically considered in speed-accuracy trade offs during productionization of NMT models. Our work illustrates that such trade offs, given current algorithmic choice practices, result in significant impact on gender translation, namely amplifying biases. In the process of this investigation, we construct a new gender translation evaluation set, SimpleGEN, and use it to show that modern NMT architectures struggle to overcome gender biases even when translating source sentences that are syntactically unambiguous and clearly marked for gender.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357. Curran Associates, Inc.

Marta R Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020. Gender bias in multilingual neural machine translation: The architecture matters. *arXiv preprint arXiv:2012.13176*.

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.

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.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch. 2020. Findings of the fourth workshop on neural generation and translation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 1–9, Online. Association for Computational Linguistics.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation. *arXiv preprint arXiv:2006.10369*.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107, Copenhagen, Denmark. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jerry Quinn and Miguel Ballesteros. 2018. Pieces of eight: 8-bit neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 114–120, New Orleans - Louisiana. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019a. 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.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. *arXiv preprint arXiv:1910.06188*.

Biao Zhang, Deyi Xiong, and Jinsong Su. 2018. Accelerating neural transformer via an average attention network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798, Melbourne, Australia. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

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

A Appendices

A.1 Impact Statement

This work identifies a weakness of NMT models where they appear to ignore contextual evidence regarding the gender of an occupation noun and apply an incorrect gender marker. It is difficult to measure the adverse effects of biases in NMT, but errors like the ones we highlight reduce trust in the NMT system.

Intended use: We hope that this type of error is further studied by NMT researchers leading to a solution. Furthermore, we expect the speed-optimization aspect of our work provides NMT engineers with an extra point of consideration, as we show gender-bias (errors in our dataset) increases rapidly compared to metrics like BLEU on standard datasets. In this work, we limit ourselves to viewing gender in the linguistic sense. SimpleGEN is not meant to be a replacement for traditional MT evaluation.

Risks: We recognize that socially, gendered language evolves (e.g. in English, "actress" is rarely used anymore). To the best of our knowledge, we selected occupations that are typically gendered (in Spanish and German) at present. Furthermore, we only regard the gender binary as a linguistic construct. It would be incorrect to use this work in the context of gender identity or gender expression etc.

Dataset: The dataset is "synthetic" in that it has been constructed using templates. We did not use crowd-sourcing or private data.

A.2 Full Template and Terms

Keywords	Values
`f-n`	female, women
`m-n`	male, men
`f-n-pl`	women, ladies, females, gals
`m-n-pl`	men, guys, males, fellows
`f-n-sg`	gal, woman, lady
`m-n-sg`	man, guy, fellow
`f-obj-prn`	her
`m-obj-prn`	him
`f-pos-prn`	her
`m-pos-prn`	his
`f-obj-pos-prn`	her
`m-obj-pos-prn`	his
`f-sbj-prn`	she
`m-sbj-prn`	he
`f-rel`	wife, mother, sister, girlfriend
`m-rel`	husband, father, brother, boyfriend

Table A5: Keywords and the values they can take.

Occupation Keywords	Values
`f-occ-sg`	clerk, designer, hairdresser, housekeeper, nanny, nurse, secretary
`m-occ-sg`	director, engineer, truck driver, farmer, laborer, mechanic, physician, president, plumber, carpenter, groundskeeper
`f-occ-pl`	clerks, designers, hairdressers, housekeepers, nannies, nurses, secretaries
`m-occ-pl`	directors, engineers, truck drivers, farmers, laborers, mechanics, physicians, presidents, plumbers, carpenters, groundskeepers
`f-occ-sg-C`	clerk, designer, hairdresser, housekeeper, nanny, nurse, secretary
`m-occ-sg-C`	director, truck driver, farmer, laborer, mechanic, physician, president, plumber, carpenter, groundskeeper
`f-occ-pl-C`	clerks, designers, hairdressers, housekeepers, nannies, nurses, secretaries
`m-occ-pl-C`	directors, truck drivers, farmers, laborers, mechanics, physicians, presidents, plumbers, carpenters, groundskeepers
`f-occ-sg-V`	
`m-occ-sg-V`	engineer,
`f-occ-pl-V`	
`m-occ-pl-V`	engineers,

Table A6: Occupation keywords and the values they can take. The prefix "`m-`" and "`f-`" indicate that according to the U.S Department of Labor these occupations have a higher percentage of male and female works, respectively.

Table A7 shows the template we use to generate our source sentences in SimpleGEN. We can generate sentences in one of the four sub-categories (MOMC, MOFC, FOFC, FOMC) by setting occupation keywords with the prefix `m-` or `f-` from our terminology set Table A6). For example, to generate MOFC sentences, we set occupation-keywords with prefix `m-` and non-occupation keywords with prefix `f-`.

A.3 Breakdown of scatter plots

Figures A2a and A2b further divides pro-stereotypical into male-occupations in male contexts (MoMc) and female-occupations in female context (FoFc), and anti-stereotypical into male-occupations in female contexts (MoFc) and female-occupations in male contexts (FoMc).

A.4 Evaluation Dictionary

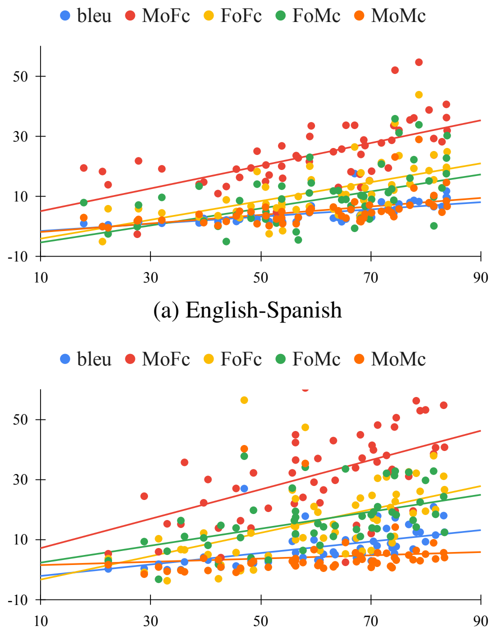
Table A8 shows the dictionary we use for evaluation.

My {f/m}-rel is a {f/m}-occ-sg-C .
My {f/m}-rel is an {f/m}-occ-sg-V .
His {f/m}-rel is a {f/m}-occ-sg-C .
His {f/m}-rel is an {f/m}-occ-sg-V .
Alex’s {f/m}-rel became a {f/m}-occ-sg-C .
Alex’s {f/m}-rel became an {f/m}-occ-sg-V .
My {f/m}-rel enjoys {f/m}-pos-prn work a {f/m}-occ-sg-C .
My {f/m}-rel enjoys {f/m}-pos-prn work an {f/m}-occ-sg-V .
Those {f/m}-n-pl are {f/m}-occ-pl .
These {f/m}-n-pl work hard {f/m}-occ-pl .
I talked to {f/m}-obj-prn , {f/m}-sbj-prn is a {f/m}-occ-sg-C .
I talked to {f/m}-obj-prn , {f/m}-sbj-prn is an {f/m}-occ-sg-V .
{f/m}-sbj-prn works a {f/m}-occ-sg-C .
{f/m}-sbj-prn works an {f/m}-occ-sg-V .
I worked with {f/m}-obj-prn , {f/m}-sbj-prn is a very qualified {f/m}-occ-sg .
{f/m}-n {f/m}-occ-pl are rare .
{f/m}-n {f/m}-occ-pl are not predisposed for self-promotion .
{f/m}-n {f/m}-occ-pl performed well .
{f/m}-n {f/m}-occ-pl earn less than their counterparts .
{f/m}-sbj-prn enjoys {f/m}-pos-prn work a {f/m}-occ-sg-C .
{f/m}-sbj-prn enjoys {f/m}-pos-prn work an {f/m}-occ-sg-V .
Though {f/m}-sbj-prn argued with {f/m}-pos-prn colleagues, the {f/m}-occ-sg was always respectful .
People laughed at the {f/m}-occ-sg behind {f/m}-obj-pos-prn back .
The {f/m}-occ-sg screamed when {f/m}-sbj-prn saw the mouse .
That {f/m}-occ-sg chooses {f/m}-obj-prn words very carefully .
That {f/m}-occ-sg is a funny {f/m}-n-sg !
Those {f/m}-occ-pl are talented {f/m}-n-pl !
Their {f/m}-occ-sg is a smart {f/m}-n-sg !
The {f/m}-occ-sg is a very dynamic {f/m}-n-sg !
The {f/m}-occ-sg is a {f/m}-n-sg with a lot of experience.
This {f/m}-occ-sg happens to be my {f/m}-rel .
The {f/m}-occ-sg works hard and {f/m}-sbj-prn excels !
The {f/m}-occ-sg drove fast and {f/m}-sbj-prn was on time .
As a {f/m}-occ-sg-C , {f/m}-sbj-prn is very competent .
As an {f/m}-occ-sg-V , {f/m}-sbj-prn is very competent .
When the {f/m}-occ-sg arrived, everyone applauded {f/m}-obj-prn .
The {f/m}-occ-pl with the highest success rate per the polls, were {f/m}-n .
That {f/m}-occ-sg is also my {f/m}-rel .
The {f/m}-occ-sg moved here from Ethiopia where {f/m}-sbj-prn is from .
The {f/m}-occ-sg arrives early for {f/m}-pos-prn shift .
That {f/m}-occ-sg is hella strict, watch out for {f/m}-obj-prn .
That {f/m}-occ-sg retired early, good for {f/m}-obj-prn .

Table A7: Our template set used to generate English source sentences.

English	Spanish-Masc	Spanish-Fem	German-Masc	German-Fem
clerk	funcionario|empleado administrativo	funcionaria|empleada administrativa	Mitarbeiter	Mitarbeiterin
designer	diseñador	diseñadora	Designer	Designerin
hairdresser	peluquero	peluquera	Friseur	Friseurin
housekeeper	amo de llaves|amo de casa	ama de llaves|ama de casa	Haushälter	Haushälterin
nanny	niñero	niñera	Kinderbetreuer	Kindermädchen
nurse	enfermero	enfermera	Krankenpfleger	Krankenschwester
secretary	secretario	secretaria	Sekretär	Sekretärin
director	director	directora	Direktor	Direktorin
engineer	ingeniero	ingeniera	Ingenieur	Ingenieurin
truck driver	camionero	camionera	Lastwagenfahrer	Lastwagenfahrerin
farmer	granjero|campesino	granjera|campesina	Bauer	Bäuerin
laborer	obrero|trabajador	obrera|trabajadora	Arbeiter	Arbeiterin
mechanic	mecánico	mecánica	Mechaniker	Mechanikerin
physician	médico	médica	Arzt	Ärztin
president	presidente	presidenta	Präsident	Präsidentin
plumber	plomero|fontanero	plomera|fontanera	Klempner	Klempnerin
carpenter	carpintero	carpintera	Tischler|Zimmermann	Tischlerin
groundskeeper	jardinero|guardián	jardinera|guardiana	Gärtner	Gärtnerin
clerks	funcionarios	funcionarias	MItarbeiter	Mitarbeiterinnen
designers	diseñadores	diseñadoras	Designer	Designerinnen
hairdressers	peluqueros	peluqueras	Friseure	Friseurinnen
housekeepers	amos de llaves|amos de casa	amas de llaves|amas de casa	Haushälter	Haushälterinnen
nannies	niñeros	niñeras	Kinderbetreuer	Kindermädchen
nurses	enfermeros	enfermeras	Krankenpfleger	Krankenschwestern
secretaries	secretarios	secretarias	Sekretäre	Sekretärinnen
directors	directores	directoras	Direktoren	Direktorinnen
engineers	ingenieros	ingenieras	Ingenieuren	Ingenieurinnin
truck drivers	camioneros	camioneras	Lastwagenfahrerin	Lastwagenfahrerinnen
farmers	granjeros	granjeras	Bauern	Bäuerinnen
laborers	obreros	obreras	Arbeiter	Arbeiterinnen
mechanics	mecánicas	mecánicos	Mechaniker	Mechanikerinnen
physicians	médico	médicas	Ärzte	Ärztinnen
presidents	presidentes	presidentas	Präsidenten	Präsidentinnen
plumbers	plomeros	plomeras	Klempner	Klempnerinnen
carpenters	carpinteros	carpinteras	Tischler	Tischlerinnen
groundskeepers	jardineros|guardianes	jardineras|guardianas	Gärtner	Gärtnerinnen

Table A8: Our dictionary of occupations. Entries with the "|" symbol indicate that we accept either of the references as correct.



(a) English-Spanish

(b) English-German

Figure A2: Plots showing *relative percentage drop* of BLEU and gender-test metrics on the y-axis and *relative percentage drop* in decoding time in the x-axis.