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

Adithya Renduchintala, Denise Diaz\textsuperscript{1}, Kenneth Heafield, Xian Li, Mona Diab
Facebook AI, \textsuperscript{1}Independent Researcher
\{adirendu,kheafield,xian1,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.

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-noun 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

| source                        | reference                                      |
|-------------------------------|-----------------------------------------------|
| That physician is a funny lady! | ¡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”.

\textsuperscript{*}This work conducted while author was working at Facebook AI.
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

1https://www.dol.gov/agencies/wb/data/high-demand-occupations
Table 3: Our evaluation protocol with an example source sentence and four example translations.

| 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 shows an example of our evaluation protocol for an example source sentence 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
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 

\[
\frac{(3662.8 - 3662.8)}{3662.8} = 0\% 
\]

(a) Each speed-up optimization individually.

(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 ≈ 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
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), Stafanovićs 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.
8 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.

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 Battle, 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, Hiroyuki 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, Jindrich 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 Computer 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.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

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 Zafir, 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. arXiv preprint arXiv:1805.00631.

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.
Appendices

A.1 Full Template and Terms

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-.

| 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.

A.2 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.3 Evaluation Dictionary

Table A8 shows the dictionary we use for evaluation.

| 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, groundkeeper |
| f-occ-pl            | clerks, designers, hairdressers, housekeepers, nannies, nurses, secretaries |
| m-occ-pl            | directors, engineers, truck drivers, farmers, laborers, mechanics, physicists, presidents, plumbers, carpenters, groundkeepers |
| 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, groundkeeper |
| 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, groundkeepers |
| f-occ-sg-V          | engineer, |
| m-occ-sg-V          | engineer, |
| f-occ-pl-V          | engineers, |
| 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. |

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

| Template | Description |
|----------|-------------|
| **My {f/m}-rel is a {f/m}-occ-sg-C.** | My {f/m}-rel is an {f/m}-occ-sg-C. |
| Alex's {f/m}-rel became a {f/m}-occ-sg-C. | Alex's {f/m}-rel became an {f/m}-occ-sg-V. |
| His {f/m}-rel is an {f/m}-occ-sg-V. | My {f/m}-rel enjoys {f/m}-pos-prn work a {f/m}-occ-sg-C. |
| His {f/m}-rel is an {f/m}-occ-sg-V. | 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. | Those {f/m}-n-pl work hard {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 a {f/m}-occ-sg-C. | {f/m}-sbj-prn works an {f/m}-occ-sg-V. |
| {f/m}-sbj-prn works an {f/m}-occ-sg-V. | People laughed at the {f/m}-occ-sg behind {f/m}-obj-pos-prn back. |
| People laughed at the {f/m}-occ-sg behind {f/m}-obj-pos-prn back. | The {f/m}-occ-sg who arrived early for {f/m}-pos-prn shift. |
| Those {f/m}-occ-pl are talented {f/m}-n-pl! | Those {f/m}-occ-pl are talented {f/m}-n-pl! |
| Those {f/m}-occ-pl are talented {f/m}-n-pl! | Their {f/m}-occ-sg is a smart {f/m}-n-sg! |
| 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 very dynamic {f/m}-n-sg! | The {f/m}-occ-sg is a {f/m}-n-sg with a lot of experience. |
| 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. |
| 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 works hard and {f/m}-sbj-prn excels! | The {f/m}-occ-sg drove fast and {f/m}-sbj-prn was on time. |
| 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 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. |
| 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. |
| When the {f/m}-occ-sg arrived, everyone applauded {f/m}-obj-prn. | The {f/m}-occ-sg is also my {f/m}-rel. |
| The {f/m}-occ-sg is also my {f/m}-rel. | The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. |
| The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. | The {f/m}-occ-sg retired early, good for {f/m}-obj-prn. |
| The {f/m}-occ-sg retired early, good for {f/m}-obj-prn. | The {f/m}-occ-sg is a very dynamic {f/m}-n-sg! |
| The {f/m}-occ-sg is a very dynamic {f/m}-n-sg! | The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. |
| The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. | As an {f/m}-occ-sg-V, {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. |
| When the {f/m}-occ-sg arrived, everyone applauded {f/m}-obj-prn. | The {f/m}-occ-sg is also my {f/m}-rel. |
| The {f/m}-occ-sg is also my {f/m}-rel. | The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. |
| The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. | As an {f/m}-occ-sg-V, {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. |
| When the {f/m}-occ-sg arrived, everyone applauded {f/m}-obj-prn. | The {f/m}-occ-sg is also my {f/m}-rel. |
| The {f/m}-occ-sg is also my {f/m}-rel. | The {f/m}-occ-sg is from Ethiopia where {f/m}-sbj-prn is from. |
| English       | Spanish-Masc                                      | Spanish-Fem                                      | German-Masc           | German-Fem           |
|--------------|--------------------------------------------------|-------------------------------------------------|-----------------------|-----------------------|
| clerk        | funcionario)(empleado administrativo             | funcionaria)(empleada administrativa             | Mitarbeiter           | Mitarbeiterin         |
| designer     | diseñador                                         | diseñadora                                      | Designer              | Designerrin           |
| hairdresser  | peluquero                                        | peluquera                                       | Friseur               | Friseurin             |
| housekeeper  | amo de llaves)(amo de casa                       | ama de llaves)(ama de casa                      | Haushalter            | Haushalterin          |
| nanny        | niño                                             | niña                                           | Kinderbetreuer        | Kindermädchen         |
| nurse        | enfermero                                        | enfermera                                       | Krankenpfleger        | Krankenschwestern     |
| 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                  | Arztin                |
| president    | presidente                                      | presidenta                                      | Präsident             | Präsidentin           |
| plumber      | plomero)(fontanero                              | plomera)(fontanera                              | Klemmpner             | Klemmpferin           |
| carpenter    | carpintero                                       | carpintera                                      | Tischler              | Tischlerin            |
| groundskeeper| jardinero)(guardián                             | jardinera)(guardiana                           | Gärtnere              | Gärtnerin             |
| clerks       | funcionarios                                      | funcionarias                                    | Mitarbeiter           | Mitarbeiterinnen      |
| designers    | diseñadores                                      | diseñadoras                                     | Designerrin           | Designerrinnen        |
| hairdressers | peluqueros                                       | peluqueras                                      | Friseur               | Friseurin             |
| housekeepers | amos de llaves)(amos de casa                     | amas de llaves)(amas de casa                    | Haushalter            | Haushalterin          |
| nannies      | niñeros                                          | niñeras                                         | Kinderbetreuer        | Kindermädchen         |
| nurses       | enfermeros                                       | enfermeras                                      | Krankenpfleger        | Krankenschwestern     |
| secretaries  | secretarios                                      | secretarias                                     | Sekretäre             | Sekretärinnen         |
| directors    | directores                                       | directoras                                      | Direktoren            | Direktorinnen         |
| engineers    | ingenieros                                       | ingenieras                                      | Ingenieuren           | Ingenieurnin          |
| truck drivers| camioneros                                       | camioneras                                      | Lastwagenfahrer        | Lastwagenfahrerin     |
| farmers      | granjeros                                        | granjeras                                       | Bauern                | Bäuerinnen            |
| laborers     | obreros                                          | obreras                                         | Arbeiter              | Arbeiterinnen         |
| mechanics    | mecánicas                                        | mecánicas                                       | Mechaniker            | Mechanikerinnen       |
| physicians   | médico                                           | médicas                                        | Ärzte                 | Arztinnen             |
| presidents   | presidentes                                      | presidentas                                     | Präsidenten           | Präsidentinnen        |
| plumbers     | plomeros                                         | plomeras                                        | Klemmpner             | Klemmpferin           |
| carpenters   | carpinteros                                      | carpinteras                                     | Tischler              | Tischlerinnen         |
| groundskeepers| jardineros)(guardianes                          | jardineras)(guardianas                         | Gärtnere              | Gärtnerin             |

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