GFST: Gender-Filtered Self-Training for More Accurate Gender in Translation

Prafulla Kumar Choubey∗†  Anna Currey∗  Prashant Mathur, Georgiana Dinu
Texas A&M University  Amazon AI Translate
prafulla.choubey@tamu.edu  {ancurrey,pramathu,gddinu}@amazon.com

Abstract

Targeted evaluations have found that machine translation systems often output incorrect gender in translations, even when the gender is clear from context. Furthermore, these incorrectly gendered translations have the potential to reflect or amplify social biases. We propose gender-filtered self-training (GFST) to improve gender translation accuracy on unambiguously gendered inputs. Our GFST approach uses a source monolingual corpus and an initial model to generate gender-specific pseudo-parallel corpora which are then filtered and added to the training data. We evaluate GFST on translation from English into five languages, finding that it improves gender accuracy without damaging generic quality. We also show the viability of GFST on several experimental settings, including re-training from scratch, fine-tuning, controlling the gender balance of the data, forward translation, and back-translation.

1 Introduction

Recent work has drawn attention to the harms that machine learning algorithms can cause by reflecting or even amplifying data biases against protected groups (Barocas et al., 2019; Kearns and Roth, 2019). For the most part, machine translation (MT) studies on bias have focused on gender bias in neural machine translation (NMT) and have identified a series of representational harms and stereotyping. For example, on input sentences that are underspecified in terms of gender, MT models often default to masculine or gender-stereotypical outputs (Cho et al., 2019; Prates et al., 2018), which can exclude female and non-binary people (e.g., the sentence I am a doctor spoken by a woman may be translated incorrectly as I am a (male) doctor). Even on unambiguously gendered inputs, NMT models can exhibit poorer performance, in terms of overall quality or gender translation accuracy, on content with non-masculine referents (Bentivogli et al., 2020; Stanovsky et al., 2019).

In this paper, we take on the task of improving gender translation accuracy, focusing on unambiguous inputs where there is only one correct translation with respect to gender. This task is especially difficult when translating from languages with very limited grammatical gender (such as English) into languages with extensive gender markings (such as German).

Known sources of gender bias in MT include sample bias (a.k.a. selection bias), which occurs when the input (source) distribution differs from that of the target application; label bias, which in MT occurs when gender-neutral sentences are translated predominantly into a specific gender or when the gender is translated incorrectly in the training data; and over-amplification, which is a property of the machine learning model (Shah et al., 2020). In this paper we focus on sample bias, starting from the observation that MT training data is often underspecified in terms of gender, MT models often default to masculine or gender-stereotypical outputs (Cho et al., 2019; Prates et al., 2018), which can exclude female and non-binary people (e.g., the sentence I am a doctor spoken by a woman may be translated incorrectly as I am a (male) doctor). Even on unambiguously gendered inputs, NMT models can exhibit poorer performance, in terms of overall quality or gender translation accuracy, on content with non-masculine referents (Bentivogli et al., 2020; Stanovsky et al., 2019).

In this paper, we take on the task of improving gender translation accuracy, focusing on unambiguous inputs where there is only one correct translation with respect to gender. This task is especially difficult when translating from languages with very limited grammatical gender (such as English) into languages with extensive gender markings (such as German).

Known sources of gender bias in MT include sample bias (a.k.a. selection bias), which occurs when the input (source) distribution differs from that of the target application; label bias, which in MT occurs when gender-neutral sentences are translated predominantly into a specific gender or when the gender is translated incorrectly in the training data; and over-amplification, which is a property of the machine learning model (Shah et al., 2020). In this paper we focus on sample bias, starting from the observation that MT training data is often underspecified in terms of gender, MT models often default to masculine or gender-stereotypical outputs (Cho et al., 2019; Prates et al., 2018), which can exclude female and non-binary people (e.g., the sentence I am a doctor spoken by a woman may be translated incorrectly as I am a (male) doctor). Even on unambiguously gendered inputs, NMT models can exhibit poorer performance, in terms of overall quality or gender translation accuracy, on content with non-masculine referents (Bentivogli et al., 2020; Stanovsky et al., 2019).

In this paper, we take on the task of improving gender translation accuracy, focusing on unambiguous inputs where there is only one correct translation with respect to gender. This task is especially difficult when translating from languages with very limited grammatical gender (such as English) into languages with extensive gender markings (such as German).

Known sources of gender bias in MT include sample bias (a.k.a. selection bias), which occurs when the input (source) distribution differs from that of the target application; label bias, which in MT occurs when gender-neutral sentences are translated predominantly into a specific gender or when the gender is translated incorrectly in the training data; and over-amplification, which is a property of the machine learning model (Shah et al., 2020). In this paper we focus on sample bias, starting from the observation that MT training data is often underspecified in terms of gender, MT models often default to masculine or gender-stereotypical outputs (Cho et al., 2019; Prates et al., 2018), which can exclude female and non-binary people (e.g., the sentence I am a doctor spoken by a woman may be translated incorrectly as I am a (male) doctor). Even on unambiguously gendered inputs, NMT models can exhibit poorer performance, in terms of overall quality or gender translation accuracy, on content with non-masculine referents (Bentivogli et al., 2020; Stanovsky et al., 2019).

In this paper, we take on the task of improving gender translation accuracy, focusing on unambiguous inputs where there is only one correct translation with respect to gender. This task is especially difficult when translating from languages with very limited grammatical gender (such as English) into languages with extensive gender markings (such as German).

Known sources of gender bias in MT include sample bias (a.k.a. selection bias), which occurs when the input (source) distribution differs from that of the target application; label bias, which in MT occurs when gender-neutral sentences are translated predominantly into a specific gender or when the gender is translated incorrectly in the training data; and over-amplification, which is a property of the machine learning model (Shah et al., 2020). In this paper we focus on sample bias, starting from the observation that MT training data is often underspecified in terms of gender, MT models often default to masculine or gender-stereotypical outputs (Cho et al., 2019; Prates et al., 2018), which can exclude female and non-binary people (e.g., the sentence I am a doctor spoken by a woman may be translated incorrectly as I am a (male) doctor). Even on unambiguously gendered inputs, NMT models can exhibit poorer performance, in terms of overall quality or gender translation accuracy, on content with non-masculine referents (Bentivogli et al., 2020; Stanovsky et al., 2019).
Table 1: Distribution of feminine (Fem), masculine (Msc), and mixed data in our parallel training data. Data is from WMT/IWSLT (described in Appendix A).

gender imbalanced. Indeed, Table 1 shows the relative proportion of masculine-referring vs. femininereferring\(^3\) sentences in our training data (extracted using the FILTERSRC algorithm described in Section 2). Though over 90% of the data is not specific to one gender (Mix in Table 1), there are at least 2.6 times more masculine-specific than feminine-specific sentences in all of our training sets. Similarly, Vanmassenhove et al. (2018) showed that, across 10 languages, only 30% of Europarl (Koehn, 2005) has female speaker gender.

This paper proposes a data augmentation-based method to address sample bias using only source-language monolingual data. Our approach, dubbed gender-filtered self-training (GFST), consists of self-training the NMT model using gender-balanced monolingual data that is filtered to reduce error propagation. Our framework is simple, generic, and easily scalable to any target language for which a morphological tagger is available. Our main contributions are:

1. We propose GFST, a broadly applicable self-training technique that leverages natural monolingual corpora exhibiting diverse gender phenomena.
2. We show that GFST yields significant improvements in gender translation accuracy on both feminine and masculine gendered input without harming overall translation quality.
3. We perform a wide set of experiments that show that these results hold on several language pairs and settings, including adapting to fine-tuning and to back-translation.

2 Gender-Filtered Self-Training (GFST)

In this paper, we propose gender-filtered self-training (GFST) for improving gender translation accuracy on unambiguously gendered input sentences. We use filtering and self-training to augment the data used to train the MT model. Our GFST approach is illustrated in Figure 1.

GFST assumes access to a parallel corpus \(D_{\text{par}}\) and a monolingual source corpus \(D_{\text{src}}\). We first train an initial model \(\Theta_{\text{ini}}\) on \(D_{\text{par}}\). Due to the skewed gender representation of the training data (see Table 1), \(\Theta_{\text{ini}}\) may fail to use relevant gender cues from context, incorrectly translating gender-unmarked feminine words (such as friend in the sentence She is my friend) as masculine or vice versa. The extent of such errors can vary with the amount and quality of the training data, the domain of the data, or linguistic features of the languages. Nonetheless, we assume that our baseline models can render the correct gender for at least some inputs (Escudé Font and Costa-jussà, 2019).

Therefore, we use \(\Theta_{\text{ini}}\) to generate translations for gender-specific sentences extracted from \(D_{\text{src}}\). This forward-translated data is then filtered to ensure that the translations accurately reflect the gender of the source, balanced by gender, and used as additional training data. Note that filtering is only done on the additional pseudo-parallel data; the original parallel data is used in its entirety. The full process is illustrated in Algorithm 1, and we describe each step in detail below.

### Algorithm 1 GFST for NMT.

**Require:** Parallel and src mono data \(D_{\text{par}}, D_{\text{src}}\)

1. Train \(\Theta_{\text{ini}}\) on \(D_{\text{par}}\)
2. For gen in \{fem, msc\}
3. \(D_{\text{gen}}^{\text{src}} \leftarrow \text{FILTERSRC}(D_{\text{src}}, \text{gen})\)
4. \(D_{\text{gen}}^{\text{trg}} \leftarrow \text{Translate} D_{\text{gen}}^{\text{src}} \text{using } \Theta_{\text{ini}}\)
5. \(D_{\text{gen}}^{\text{par}} \leftarrow \text{FILTERTRG}(D_{\text{gen}}^{\text{src}}, D_{\text{gen}}^{\text{trg}}, \text{gen})\)
6. Train \(\Theta_{\text{fin}}\) on \(D_{\text{par}} + D_{\text{gen}}^{\text{par}} + D_{\text{par}}\)
7. return \(\Theta_{\text{fin}}\)

**FILTERSRC:** We extract a feminine and a masculine subset of sentence candidates (\(D_{\text{gen}}^{\text{src}}\) for \(\text{gen} \in \{\text{fem}, \text{msc}\}\)) from the source-language (in our case, English) monolingual corpus \(D_{\text{src}}\). Specifically, given lists of feminine and masculine words, we consider a source sentence masculine if it meets all of the following criteria:
1. Has at least one masculine pronoun
2. Does not have any feminine pronouns
3. Does not contain any feminine words

We use an equivalent set of criteria to extract feminine sentence candidates from the data. To define

\[\begin{array}{cccccc}
\text{Fem} & \text{en-de} & \text{en-fr} & \text{en-he} & \text{en-it} & \text{en-ru} \\
0.7\% & 0.9\% & 1.9\% & 1.9\% & 0.5\% & \\
\text{Msc} & 4.8\% & 3.1\% & 4.9\% & 4.9\% & 2.4\% & \\
\text{Mix} & 94.5\% & 96.0\% & 93.2\% & 93.2\% & 97.1\%
\end{array}\]
Table 2: Sentence pairs removed in the FILTERTRG step. Both pairs are removed because the target sentences contain words with masculine grammatical gender (underlined along with their aligned source words). The source sentences were selected by the FILTERSRC step due to the feminine words in bold.

| Source | Target |
|--------|--------|
| My daughter is hurt at being rejected by the girl she called her best friend | Meine Tochter ist verletzt, weil sie von dem Mädchen, das sie als ihren besten Freund [...] |
| Another passenger was held for three days for using her phone on a flight | Ein weiterer Passagier wurde drei Tage lang festgehalten, weil er sein Telefon [...] |

gender-specific words, we use a list from Zhao et al. (2018)\(^4\) that contains a total of 104 pairs of words (such as brother/sister or boy/girl).

**FILTERTRG**: Filtering on the target side of the data is done to exclude sentence pairs for which the model failed to preserve the gender of the source sentence. We run morphological analysis on the translations \(D_{\text{trg}}^{\text{masc}}\) of \(D_{\text{src}}^{\text{masc}}\) and keep only those sentences that have:

1. No grammatically feminine words, and
2. At least one grammatically masculine word and similarly for the translations of \(D_{\text{trg}}^{\text{fem}}\) of \(D_{\text{src}}^{\text{fem}}\). This results in parallel datasets \(D_{\text{par}}^{\text{masc}}\) and \(D_{\text{par}}^{\text{fem}}\). Table 2 shows examples of sentences that passed FILTERSRC but were removed during FILTERTRG.

Note that FILTERTRG suffices to generate gender-specific sentence pairs. However, FILTERSRC reduces computational cost by limiting the search space for the candidate sentences and reduces the risk of introducing wrongly translated sentence pairs that may pass FILTERTRG.

**Self-Training NMT**: After obtaining gender-specific pseudo-parallel corpora, the larger of the two is sub-sampled to balance the pseudo-parallel data. Finally, the original parallel corpus \(D_{\text{par}}\) is concatenated with the two pseudo-parallel corpora \(D_{\text{par}}^{\text{masc}}\) and \(D_{\text{par}}^{\text{fem}}\) and used to train a final MT model \(\Theta_{\text{fin}}\).

### 3 Evaluation

#### 3.1 Gender Accuracy on WinoMT

We evaluate on the WinoMT (Stanovsky et al., 2019) gender-annotated test sets. WinoMT contains 3888 English sentences taken from the Wino-gender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) datasets. Each sentence contains a target occupation that lacks gender marking at the lexical level, such as *salesperson*. The gender of the referent is implicitly defined by a coreferential pronoun in the sentential context, leading to sentences such as *The salesperson sold some books to the librarian because it was her job, where salesperson is implicitly but unambiguously feminine. The dataset distinguishes between anti- and pro-stereotypical occupations, and contains 3648 sentences equally balanced between masculine and feminine as well as pro-stereotypical and anti-stereotypical occupations. Target occupations in the remaining 240 sentences are identified with neutral gender (e.g. *The technician told someone that they could pay with cash*) and are excluded from the stereotype annotation.

**WinoMT Metrics**: On the WinoMT data, the automated evaluation strategy first uses *fast_align* (Dyer et al., 2013) to find the alignment for the target occupation in the translation. Then, using heuristic rules over language-specific morphological analysis, it identifies the gender of the translated occupation and uses three metrics to estimate the overall bias. **Accuracy** is the percentage of translations that correctly reflect the gender of target occupation, while \(\Delta G\) and \(\Delta S\) are defined as the difference in \(F_1\) scores between masculine and feminine and between pro-stereotypical and anti-stereotypical target occupations respectively.

**\(\Delta R\)**: \(\Delta G\) may not give a complete picture of gender bias when the test set includes samples with unambiguously neutral gender (e.g. WinoMT sentences with *they*). To understand how this can happen, consider two hypothetical MT models that both have equal accuracy on feminine and masculine inputs but differ in how they treat neutral
inputs. Model A translates all neutral inputs as masculine, whereas model B translates half of the neutral inputs as masculine and half as feminine. In this scenario, model A will have a lower $\Delta G$ because it has lower precision on masculine inputs but the same recall for masculine and feminine inputs. However, we argue that model A may still be biased towards the masculine gender, since it defaults to masculine when the inputs are neutral.

Therefore, we propose a new metric for the WinoMT test suite: $\Delta R$, which we define as the difference in recall between masculine and feminine samples. This metric complements the existing metrics and gives a more complete picture of model biases. $\Delta R$ decouples precision from the $\Delta G$ metric by excluding neutral inputs from consideration and only evaluating on unambiguously gendered input sentences. Thus, it is an indicator of the model’s bias towards outputting masculine vs. feminine gender. We use $\Delta R$ because GFST does not specifically address translation of neutral inputs, and we do not take a stance on how such inputs should be translated.

**Human Evaluations:** The WinoMT automatic gender accuracy metric was originally validated using human annotators. While the metric was shown to be relatively accurate, with an agreement between annotators and the metric of over 85% across all languages and systems, in this paper we complement the automatic metric with a small-scale human evaluation. Fluent speakers of German, Italian, and Russian were asked to annotate the gender translation accuracy of a random subset of 100 unambiguously gendered sentences from WinoMT (balanced for masculine/feminine and pro-/anti-stereotype). Annotators were instructed to classify a translation with one of five discrete labels: besides *masculine* or *feminine* (as in automatic evaluations), we added *inconsistent* (if some words in the translation indicate one gender and some indicate another for the same referent), *ambiguous* (if the translation is valid for both masculine and feminine referents), and *N/A* (if the referent of interest is completely omitted from the translation). We classify translations as *incorrect* if they are inconsistent, N/A, or a different gender from the unambiguous source (e.g. masculine if the source sentence is feminine), and *correct* if they are ambiguous or the same gender as the source.

### 3.2 Gender Accuracy on MuST-SHE

In addition to WinoMT, we also use the MuST-SHE gender-specific translation test set (Bentivogli et al., 2020) to evaluate gender translation accuracy. MuST-SHE consists of roughly 1000 triples of audio, transcript, and reference translations taken from MuST-C (Di Gangi et al., 2019) for en-fr and en-it. Each triple is identified with either masculine or feminine gender based on speaker gender (category 1) or explicit gender markers such as pronouns (category 2). Furthermore, for each correct reference translation, the dataset includes a wrong alternative translation that changes the gender-marked words (e.g. feminine words are changed to masculine). MuST-SHE is balanced between masculine and feminine and between categories 1 and 2.

**Automatic Metrics for MuST-SHE:** We use the category 2 samples (which contain explicitly marked gender words on the source side) from MuST-SHE to evaluate our en-fr and en-it models. Following Bentivogli et al. (2020), we evaluate the gender accuracy for translations and also look at $\Delta Acc$, which is the difference between the gender accuracy of translation with respect to correct and counterfactual references. Higher $\Delta Acc$ is better, as this indicates that the model is closer to the correct reference than to the counterfactual one.

### 3.3 Generic Quality

Our main goal is to improve gender translation accuracy. Additionally, we measure generic quality using BLEU and human evaluations to investigate any potential overall quality loss. Generic human quality evaluations on WinoMT also allow us to investigate whether changes in gender accuracy lead to noticeable quality improvements.

### 4 Experiments

With source language as English (EN), we experiment on five target languages from four families, all of which have grammatical gender: French (FR), Italian (IT), Russian (RU), Hebrew (HE), and German (DE). Our experiments include low-, medium- and high-resource settings. Table 3 shows the number of parallel training sentences after preprocess-
Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{msc}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

| Dataset | $en-de$ | $en-fr$ | $en-he$ | $en-it$ | $en-ru$ |
|---------|---------|---------|---------|---------|---------|
| $D_{\text{par}}$ | 5.2M | 35.7M | 180k | 161k | 1.6M |
| $D_{\text{fem}}^{\text{src}}$ | 1.8M | 4.2M | 1.8M | 1.8M | 1.8M |
| $D_{\text{fem}}^{\text{par}}$ | 428k | 150k | 29k | 81k | 184k |

Table 4: Comparison of automatic models.

| Dataset | $en-de$ | $en-fr$ | $en-he$ | $en-it$ | $en-ru$ |
|---------|---------|---------|---------|---------|---------|
| $D_{\text{par}}$ | 5.2M | 35.7M | 180k | 161k | 1.6M |
| $D_{\text{fem}}^{\text{src}}$ | 1.8M | 4.2M | 1.8M | 1.8M | 1.8M |
| $D_{\text{fem}}^{\text{par}}$ | 428k | 150k | 29k | 81k | 184k |

Table 5: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 5: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 5: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.

Table 3: Number of sentences in each training set. $D_{\text{par}}$ is the original parallel training data, $D_{\text{fem}}^{\text{src}}$ the source-filtered feminine monolingual data, and $D_{\text{fem}}^{\text{par}}$ the feminine data after target filtering. We downsample the larger masculine data $D_{\text{msc}}^{\text{par}}$ to match the size of $D_{\text{par}}$.
of the guard, although it is marked as feminine and thus incorrect by the automatic WinoMT evaluations (because the source unambiguously indicates that the guard is masculine).

**Automatic MuST-SHE Accuracy:** In Table 6, we report accuracy, as well as $\Delta \text{Acc}$ between correct and gender-swapped references, using category 2 data from MuST-SHE for en-it and en-fr. For en-fr, GFST increases both accuracy and $\Delta \text{Acc}$ for feminine and masculine data. For the high-resource pair en-fr, there is an increase in accuracy and $\Delta \text{Acc}$ for feminine data, but a (smaller) decrease in both metrics for masculine data.

### 5.2 Generic Quality

**Automatic Translation Quality:** Table 8 reports case-sensitive de-tokenized BLEU\(^{14}\) for all language pairs on the generic (WMT or IWSLT) test sets. The results confirm that our proposed GFST method does not come at a trade-off in generic translation quality, compared to a baseline that does not use the gender-filtered data. We also observe a general trend of small improvements from self-training, irrespective of data selection method.

**Human Quality Evaluations:** To better understand how GFST affects overall translation quality, we perform human quality evaluations on a balanced, 300-sentence subset of WinoMT. For each language pair, baseline vs. GFST quality is evaluated on a six-point Likert scale by two professional translators. The scores, averaged between annotators, are shown in Table 9. For en-de, en-he, and en-it, GFST significantly improves overall quality. For en-fr and en-ru, there is no significant difference in overall quality between the two models.

### 6 Analysis

#### 6.1 Retraining vs. Fine-Tuning

The main experiments (section 5) used the data generated by the GFST method to train the final models from scratch. In this section, we further explore the utility of GFST by fine-tuning the existing models that were used for forward translation, instead of retraining them. We fine-tune these models using the feminine and masculine samples, and additionally mix in an equal number of sentences from the original training corpus to avoid catastrophic forgetting (following the mixed fine-tuning approach of Chu et al., 2017; Freitag and Al-Onaizan, 2016).

BLEU scores for baselines, retrained GFST models, and fine-tuned GFST models are shown in Table 10. For four of the five language pairs, there is no significant drop in quality between the baseline and the fine-tuned models; en-ru loses 1.1 BLEU. Table 11 shows WinoMT accuracy and $\Delta R$ results for the three models. On average, retraining on GFST data outperforms fine-tuning. However, fine-tuning on GFST data consistently improves gender accuracy over the baseline, making fine-tuning a viable low-cost alternative to retraining.

#### 6.2 Single-Gender Data Augmentation

Although our original motivation (see Table 1) was to address gender imbalance in the training data, the proposed GFST models use gender-balanced augmented data, i.e. the same amount of feminine-specific and masculine-specific sentences in the pseudo-parallel data. In this section, we investigate the relative contribution of each corpus using:

---

14SacreBLEU (Post, 2018) signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.10.
Table 6: MuST-SHE performance measured in Accuracy (Acc) and ∆Accuracy (ΔAcc).

| Model     | en-it | en-fr | Msc | en-it | Msc |
|-----------|-------|-------|-----|-------|-----|
| Baseline  | Acc   | ΔAcc  | Acc | ΔAcc  | Acc |
| GFST      |       |       |     |       |     |
| Baseline  | 32.5  | 2.5   | 38.8| 48.8  | 57.5|
| GFST      | 41.9  | 21.9  | 61.6| 54.3  | 66.8|

Table 7: Human quality scores (average of two annotators) on a balanced subset of WinoMT. Differences outside the 95% confidence interval are shown in bold.

| Model     | en-de | en-fr | en-he | en-it | en-ru |
|-----------|-------|-------|-------|-------|-------|
| Baseline  | Acc   | ΔAcc  | Acc   | ΔAcc  | Acc   |
| GFST      |       |       |       |       |       |
| Baseline  | 4.52  | 4.55  | 2.84  | 3.50  | 3.86  |
| GFST      | 4.70  | 4.47  | 3.05  | 3.59  | 3.96  |

Table 8: BLEU on generic test sets for the baselines and GFST models. The test sets are from WMT (en-de, en-fr, en-ru) and IWSLT (en-it, en-he), corresponding to the training data for each language pair.

| Model | en-de | en-fr | en-he | en-it | en-ru |
|-------|-------|-------|-------|-------|-------|
| Baseline | Acc   | ΔAcc  | Acc   | ΔAcc  | Acc   |
| GFST   |       |       |       |       |       |
| Baseline | 41.7  | 40.5  | 23.4  | 34.5  | 25.7  |
| +HD    | 41.8  | 40.7  | 23.5  | 34.4  | 25.5  |
| GFST   | 42.4  | 40.8  | 23.9  | 34.4  | 26.9  |
| +HD    | 41.8  | 40.2  | 22.8  | 34.6  | 26.6  |

Table 9: Sentence and its ambiguous output that was marked incorrect by the WinoMT automatic metric.

Sr  The guard saved the editor from the criminals because he was on duty.
Trg La guardia ha salvato l’editore dai criminali perché era in dovere.

Table 10: BLEU scores on the generic test data for the baseline model and the GFST models: retrained (GFST-RT) and fine-tuned (GFST-FT).

| Model     | en-de | en-fr | en-he | en-it | en-ru |
|-----------|-------|-------|-------|-------|-------|
| Baseline  | Acc   | ΔAcc  | Acc   | ΔAcc  | Acc   |
| GFST      |       |       |       |       |       |
| Baseline  | 41.7  | 40.5  | 23.4  | 34.5  | 25.7  |
| GFST-RT   | 41.8  | 40.2  | 23.3  | 34.6  | 26.6  |
| GFST-FT   | 41.8  | 41.1  | 23.6  | 34.3  | 24.6  |

Table 8: BLEU on generic test sets for the baselines and GFST models. The test sets are from WMT (en-de, en-fr, en-ru) and IWSLT (en-it, en-he), corresponding to the training data for each language pair.

- **GFST*Fem**: models trained on the original bitext Dpar + feminine sentence pairs Dpar* fem.
- **GFST*Msc**: models trained on Dpar + down-sampled masculine sentence pairs Dpar* msc.

In overall translation quality, all models perform similarly (see Appendix C). In Table 12, we compare feminine-only, masculine-only, and joint self-training models to the baseline on the WinoMT benchmark using accuracy and ΔR. As expected, GFST*Fem reduces the gap between recall for feminine and masculine inputs, lowering ΔR by up to 19.8 points with respect to the baseline. At the same time, GFST*Msc increases ΔR overall, suggesting that GFST works as hypothesized and can be used to balance the training data distribution between masculine and feminine genders.

On gender accuracy, GFST*Fem outperforms the baseline for all five language pairs and yields similar accuracy to the original GFST model. On the other hand, GFST*Msc performs very closely to the baseline. This result highlights the under-representation of feminine samples in the existing training corpora. The original GFST model, which is trained on both masculine and feminine additional data, outperforms GFST*Fem in accuracy but underperforms it in ΔR. This is not surprising since the GFST*Fem training data is more gender-balanced than the original GFST training data (which contains additional masculine data).

6.3 Forward Translation vs. Back-Translation

So far, our experiments have used forward translation (FT) to generate gender-balanced data through self-training. Here, we extend the approach to back-translation (BT) on a monolingual target-language corpus (Sennrich et al., 2016a). Back-translation is potentially preferable because the automatically translated data is on the source side rather than the target. Thus, BT is less likely to damage generic translation quality (although our evaluations in Section 5.2 indicate that FT does not damage generic quality either).

The BT model is created by running FILTERTRG on target monolingual data, using a target→source system for translation, and applying FILTERSRC on the resulting source-language output15. We use German News Crawl 2015, 2016, and 2017 (Bojar et al., 2018) as the monolingual data for back-translation.

In Table 13, we compare BT and FT for en-de. We use the same amount of pseudo-parallel data for both (although the data itself is not the same, as it comes from different languages).

The results highlight the flexibility of GFST, in that it can be applied to both source and target...
Table 11: Accuracy and $\Delta R$ scores on the WinoMT test data for the baseline model and the GFST models: retrained (GFST-RT) and fine-tuned (GFST-FT).

| Model      | en-de | en-fr | en-he | en-it | en-ru | Avg Acc |
|------------|-------|-------|-------|-------|-------|---------|
| Baseline   | 75.5  | 18.8  | 66.2  | 13.3  | 50.0  | 47.8    |
| GFST-RT    | 85.4  | -0.3  | 71.0  | 10.2  | 50.0  | 41.3    |
| GFST-FT    | 83.2  | 3.5   | 72.2  | 4.8   | 50.0  | 39.0    |

Table 12: Accuracy and $\Delta R$ scores on WinoMT for the baseline model, the original GFST model, and models that use only feminine-specific additional data (GFST$_{Fem}$) and masculine-specific additional data (GFST$_{Msc}$).

| Model       | en-de | en-fr | en-he | en-it | en-ru | Avg Acc |
|-------------|-------|-------|-------|-------|-------|---------|
| Baseline    | 75.5  | 0.4   | 66.2  | 18.8  | 50.0  | 47.8    |
| GFST        | 85.4  | -4.4  | 71.0  | 10.2  | 50.0  | 41.3    |
| GFST$_{Fem}$| 84.0  | -1.0  | 69.3  | 5.6   | 48.2  | 25.1    |
| GFST$_{Msc}$| 75.5  | 22.0  | 64.8  | 25.0  | 48.8  | 34.7    |

Table 13: WinoMT performance for en-de for the baseline and forward- and back-translated GFST models.

| Model         | Acc  | $\Delta G$ | $\Delta R$ |
|---------------|------|------------|------------|
| Baseline      | 75.5 | 0.4        | 18.8       |
| Forward Trans | 85.4 | -4.4       | -0.3       |
| Back-Trans    | 87.7 | -4.6       | 2.3        |

Related Work

Gender Translation Accuracy in MT: A large body of work has addressed bias in natural language processing (NLP) and MT, surveyed in Blodgett et al. (2020); Costa-Jussà et al. (2019, 2020); Savoldi et al. (2021); Sun et al. (2019); and others. In MT, several papers address the topic of gender in the context of ambiguous input (Cho et al., 2019) and propose methods to control for gender or augment data with gender (Elaraby et al., 2018; Morossef et al., 2019; Prates et al., 2018; Saunders et al., 2020; Stananović et al., 2020; Vanmassenhove et al., 2018). By contrast, in this paper we instead address the problem of gender accuracy for unambiguous inputs through gender-balancing techniques.

Work addressing the gender data imbalance issue in NLP (Zhao et al., 2018) is closely related to our proposal, as the GFST method for creating gender-specific data is motivated by data imbalance. In MT, Saunders and Byrne (2020) show that gender translation accuracy for unambiguous inputs can be improved through fine-tuning on small gender-balanced counterfactual data. Specifically, they extract a subset of source sentences containing gender-specific words (e.g. woman, she) and swap the gender of these words (e.g. man, he). The subsequent translations are used to create a dataset for fine-tuning the original model. Tomalin et al. (2021) take a similar approach of fine-tuning a trained model on a small, constructed, counterfactual dataset, while Costa-Jussà and de Jorge (2020) fine-tune a model on a small parallel Wikipedia corpus. Unlike counterfactual data augmentation, GFST does not alter the source data or generate artificial source data according to specific patterns. It instead uses naturally occurring, diverse data that is filtered for gender phenomena. Additionally, GFST requires only monolingual data, which increases its flexibility. In particular, we can generate relatively large pseudo-parallel corpora, which can be used for fine-tuning (as in prior work) as well as for train-time data augmentation.

Another popular approach to reducing gender bias in NLP is to use embedding debiasing techniques (Bolukbasi et al., 2016). In NMT, Escudé Font and Costa-Jussà (2019) use pre-trained debiased word embeddings and show that hard-debiased embeddings improve gender accuracy. This approach is orthogonal to GFST; in Section 5, we showed experiments combining both methods.

Self-Training for MT: Monolingual data has been exploited via self-training to enhance statistical (Schwenk, 2008; Ueffing, 2006) and neural MT (Wu et al., 2019) through forward translation of
source data (Imamura and Sumita, 2018; Zhang and Zong, 2016) or back-translation of target data (Sen-
nrich et al., 2016a). Additionally, unfiltered for-
ward translation has been effective in NMT for
model compression (Kim and Rush, 2016), non-
autoregressive translation (Zhou et al., 2020), and
domain adaptation (Currey et al., 2020). Here, we
experiment with forward and back-translation, and
add filtering to reduce error propagation.

8 Conclusion

This paper addresses gender translation accuracy
for unambiguously gendered inputs. The proposed
gender-filtered self-training approach creates ad-
tional gender-specific training data by filtering
source monolingual data by gender, translating the
data, and filtering the translations to remove gen-
der errors. Using this additional data, the mod-
els achieve large gains in gender accuracy without
damaging overall translation quality.

In the future, we plan to extend GFST to other
genders and language pairs. This will not be trivial:
the self-training aspect of GFST assumes that the
initial model is good enough at gender translation,
which may not be the case for other genders and
languages. In particular, the use of morphological
analysis for FILTER might limit GFST’s ap-
pliability to other genders or very low-resource
target languages. Thus, we will explore alternative
approaches to self-training (e.g. synthetic data gen-
eration) and filtering (e.g. using round-trip transla-
tion (Moon et al., 2020)).

Acknowledgments

We would like to thank Margo Lynch, Tanya
Badeka, Sony Trenous, and Felix Hieber for their
help in evaluations. We would also like to thank
the anonymous reviewers for their feedback.

9 Broader Impact

This paper has presented an approach for reducing
the gap in accuracy between masculine-referring
and feminine-referring inputs. This work addresses
potential representational harms that can come
from bias affecting feminine gender. We use only
gender-marked words, with gender being marked
either lexically (English) or morphologically (Ger-
man, French, Hebrew, Italian, and Russian), as
the basis for our definitions of feminine- and
masculine-referring inputs. Thus, we do not use hu-
man subjects, ascribe gender to any specific person,
or use gender as a variable in our work.

This work has shown improvements in gender
translation accuracy for translation from English
into several relatively diverse languages. In addi-
tion, improvements on translation accuracy for
feminine inputs do not harm overall translation
quality or gender translation accuracy for mascu-
line inputs. The approach can be generalized to
other source languages with only lexical gender
(e.g. Chinese) and to other target languages with
grammatical gender (e.g. Hindi), using a gendered
wordlist in the former case and a morphological an-
alyzer in the latter case. While our technique does
not completely close the gap in accuracy between
masculine and feminine inputs, it does significantly
improve over the baselines and as such it is a step
in the right direction.

Relying exclusively on the WinoMT benchmark
may give practitioners and users false confidence
about the level of gender bias in their machine
translation systems. While the method proposed
uses a generic monolingual corpus as the basis for
the gender-specific data, our evaluation is limited
to the available benchmarks: WinoMT and MuST-
SHE. In order to mitigate the risk of overfitting to a
specific benchmark, we have included human eval-
uations of accuracy and quality in addition to the
standard automatic evaluations. However, given the
availability of evaluation data for this task, we are
not able to thoroughly test if the method proposed
introduces other biases with respect to gender or
other protected groups. For future work, we plan to
expand existing evaluation benchmarks and use any
additional benchmarks that may become available
to the community.

This paper has only considered two genders
(masculine and feminine). The proposed self-
training approach relies on the baseline model be-
ing able to correctly translate the under-represented
gender (in this case, feminine) for at least some in-
puts. This assumption is unlikely to hold for other
under-represented genders, at least for the com-
monly used machine translation training corpora.
Additionally, the filtering step relies on a morpho-
logical analyzer to detect grammatical gender of the
target words, which may not be straightforward for
non-binary genders. Finally, although the WinoMT
dataset used for evaluation covers neutral gender,
it does not cover non-binary gender, making this
difficult to evaluate. In the future, we plan to ex-
pand our work towards covering other genders by creating additional evaluation benchmarks.

References

Duygu Altinok. 2018. DEMorphy, German language morphological analyzer. arXiv preprint arXiv:1803.00902.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. fairmlbook.org. http://www.fairmlbook.org.

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.

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

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NeurIPS’16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Mauro Cettolo, J Niehues, S Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In IWSLT-International Workshop on Spoken Language Processing, pages 2–17. Marcello Federico, Sebastian Stüker, François Yvon.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation, Heidelberg, Germany.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

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.

Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors. 2019. Proceedings of the First Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, Florence, Italy.

Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors. 2020. Proceedings of the Second Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain (Online).

Anna Currey, Prashant Mathur, and Georgiana Dinu. 2020. Distilling multiple domains for neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4500–4511, Online. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1.
(Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. arXiv preprint arXiv:1612.06897.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of the 2019 Workshop on Widening NLP, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Kenji Imamura and Eiichiro Sumita. 2018. NICT self-training approach to neural machine translation at NMT-2018. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 110–115, Melbourne, Australia. Association for Computational Linguistics.

Michael Kearns and Aaron Roth. 2019. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Inc., USA.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5, pages 79–86. Citeseer.
Artūrs Stafanovics, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations.

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.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

M. Tomalin, B. Byrne, S. Concannon, D. Saunders, and S. Ullman. 2021. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. Ethics and Information Technology. Published online 6 March 2021 (15 pages).

Nicola Ueffing. 2006. Using monolingual source-language data to improve MT performance. In 2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006, pages 174–181. ISCA.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4207–4216, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In 2008 International Workshop on Spoken Language Translation, IWSLT 2008, Honolulu, Hawaii, USA, October 20-21, 2008, pages 182–189. ISCA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5248–5264, Online. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958.

M. Tomalin, B. Byrne, S. Concannon, D. Saunders, and S. Ullman. 2021. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. Ethics and Information Technology. Published online 6 March 2021 (15 pages).

Nicola Ueffing. 2006. Using monolingual source-language data to improve MT performance. In 2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006, pages 174–181. ISCA.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4207–4216, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics.
A Data, Preprocessing, and Hyperparameters

Parallel Data: We train en-fr on the WMT14 news task (Bojar et al., 2014), en-it on the IWSLT13 task (Cettolo et al., 2013), en-ru on WMT16 (Bojar et al., 2016),16 en-he on IWSLT14 (Cettolo et al., 2014), and en-de on WMT18 (Bojar et al., 2018).17 For each language pair, we use the standard validation and test sets from the corresponding shared task.

Monolingual Data: We use English News Crawl 2017 as the monolingual source data for all five language pairs. To balance the larger en-fr parallel corpus, we also obtain feminine samples from English News Crawl 2015 and 2016 for that language pair.

For FilterTrg, we use the spaCy morphological analyzer18 for FR and IT, pymorphy2 (Korobov, 2015) for RU, German-morph-dictionary based on DeMorphy (Altinok, 2018) for DE and character-based rules following Stanovsky et al. (2019) for HE.

Preprocessing: For all language pairs, we follow Edunov et al. (2018) by removing sentences with more than 250 words or with a source/target length ratio higher than 1.5. We tokenize the data using the Moses tokenizer (Koehn et al., 2007). We learn shared BPE vocabularies (Sennrich et al., 2016b) with 32k types for DE and IT and 40k types for FR. For RU and HE, we learn separate BPEs for source and target, source with 32k types for both and target with 2k types for HE and 32k types for RU.

We use all the extracted feminine sentence pairs, and an equal number of masculine sentence pairs, during self-training for all languages except IT, where due to the small parallel data size we pick 30k random pairs. Similarly, due to the large size of the en-fr parallel corpus, we up-sample the gender-specific pseudo-parallel data twenty times for that language pair.

Training: We adopt training hyperparameters from Edunov et al. (2018); Ott et al. (2018), and use the Transformer UMT en_de_big architecture with dropout rate (Srivastava et al., 2014) of 0.3 for en-de/he/it/ru, and dropout rate of 0.1 for en-fr. We use the Adam optimizer (Kingma and Ba, 2014) with β_1=0.9, β_2=0.92 and ϵ=1e-8 (learning rate scheduler proposed by Vaswani et al., 2017), label smoothing (ε=0.1) with uniform prior, and learning rate warm-up for the first 4000 steps when training models. We use learning rate of 1e-3 for training en-de models and for all other language pairs we use learning rate of 5e-4. Baseline en-de and en-fr models are trained for 30K and 180K19 synchronous updates respectively. During self-training, we increase the number of updates in proportion to the number of new samples added. For the other three language pairs, with relatively smaller training data sizes, we stop training when validation perplexity does not improve for 5 consecutive epochs. All models are trained on Nvidia V100 GPUs with 16-bit floating point precision, with parameter update frequency adjusted to simulate training on 64 GPUs for en-de/fr and 8 GPUs for the other three language pairs. Final models are obtained through stochastic averaging of the last 10 checkpoints.

B Full WinoMT Results

Table 14 shows additional metrics on the WinoMT test set that were not shown in Section 5. Specifically, we show the F_1 scores on masculine and feminine inputs, as well as ΔS. We examine the gender-specific F_1 scores to ensure that gains from our proposed GFST model do not harm any specific gender, and indeed we see that our GFST model achieves higher F_1 than both baselines for all language pairs and both genders studied. Our models do not specifically address stereotypicalness, and ΔS scores of our models are comparable to those of the baselines, indicating that our models do not exacerbate stereotype-related bias issues. This is an encouraging initial result, given that GFST’s emphasis on using naturally occurring gendered data

---

16We only use the Common Crawl, News Commentary v11 and Wiki Headlines corpora for training, as we were not able to download the Yandex Corpus.

17Consistent with Edunov et al. (2018), we exclude the ParaCrawl corpus.

18https://spacy.io/
In this section, we analyze the quality of the target morphological filtering step \textsc{FilterTrg}. In order to reduce error propagation from GFST, this step automatically removes the forward translations that do not correctly reflect the gender of the source. This is done using a morphological tagger and removing all sentences from the feminine-specific corpus that contain a grammatically masculine word (and similarly for the masculine corpus).\footnote{For languages with a neuter gender (DE, RU), we do not filter sentences based on the presence of a neuter gender word.}

Note that this approach conflates grammatical gender and natural gender, which means that sentences with grammatical gender marked on unrelated nouns might be filtered unnecessarily. Table 16 shows two such examples, where the feminine sentence is removed because the translation contains the masculine noun \textit{Anteil} (\textit{share}), and the masculine sentence is removed because of the feminine noun \textit{Arbeit} (\textit{job}). However, with this approach, sentences with \textbf{incorrectly} gendered translations are unlikely to be included in the final pseudo-parallel corpus. Indeed, as shown in Table 3, after \textsc{FilterTrg} we keep only 2-25\% of sentences that were present in the source-filtered data. We consider this to be an acceptable trade-off for the purposes of our work: we prefer to keep high-confidence sentences at the cost of filtering valid sentences so as to minimize error propagation.

We ran a small corpus analysis to estimate the trade-offs of our morphological filtering method. We selected a random 100-sentence sample of the forward-translated \textit{en-de} data and annotated each sentence for whether the gender was preserved in the translation.\footnote{The annotations were done by the authors of the paper, not by language experts.}

We then compared this to the outcome of the filtering in order to estimate the rate of false positives and false negatives coming from this method. These results are shown in Table 17.

As desired, we do not see any false positives coming from morphological target filtering, meaning that errors in gendered translation due to the self-training procedure are unlikely to be propagated. On the other hand, this does come at a trade-off, as most of the sentences in the sample were valid but filtered unnecessarily. It is also important to highlight that this analysis was done on the language pair with the highest baseline gender translation accuracy (\textit{en-de}), meaning that the vast majority of the translations correctly reflected the gender of the source. Despite that, the true negative rate on feminine samples (8\%) is twice the rate on masculine samples (4\%).

To further analyze the importance of the \textsc{FilterTrg} step, we train a new GFST$_{\text{Src}}$ model, which directly uses forward-translated source-filtered samples (without any filtering on the target side). For head-to-head comparison with the standard GFST model (with target filtering), we sam-

Table 14: Additional WinoMT metrics not shown in Table 4 for the baseline, RANDST baseline, and our GFST model. We show $F_1$ score on feminine inputs (Fem) and masculine inputs (Msc), as well as $\Delta S$ score.

| Model   | \textit{en-de} | \textit{en-fr} | \textit{en-he} | \textit{en-it} | \textit{en-ru} |
|---------|---------------|---------------|---------------|---------------|---------------|
|         | Fem | Msc | $\Delta S$ | Fem | Msc | $\Delta S$ | Fem | Msc | $\Delta S$ | Fem | Msc | $\Delta S$ |
| Baseline | 78.0 | 78.4 | 4.1 | 70.9 | 70.7 | 16.9 | 41.0 | 54.8 | 27.0 | 23.3 | 54.8 | 13.9 |
| RANDST  | 82.5 | 80.9 | 4.5 | 68.9 | 69.5 | 15.2 | 42.0 | 54.5 | 26.2 | 21.2 | 55.9 | 11.0 |
| GFST    | 90.8 | 86.4 | 4.5 | 76.7 | 75.4 | 9.3 | 42.1 | 55.9 | 23.6 | 45.8 | 63.8 | 12.1 |
|         |     |     |      | 19.5 | 52.2 | 1.5 |     |     |      | 19.2 | 51.4 | 1.2 |

Table 15: BLEU scores on the generic test sets for the baseline model and GFST models. GFST$_{\text{Fem}}$ uses only the feminine-specific data for augmentation, while GFST$_{\text{Msc}}$ uses only the masculine-specific data.

- **Results on Generic Test Sets for Single-Gender Models**

In this section, we show BLEU scores on the generic test sets for the single-gender models introduced in Section 6.2. Table 15 shows that the single-gender (feminine-only or masculine-only) data augmentation performs similarly to the baseline and to the model augmented with both feminine and masculine data in terms of BLEU score on generic test sets.
Incorrectly target-filtered sentences

| fem | She had her share of sorrows that money could not comfort. | Sie hatte ihren Anteil an den Sorgen, die das Geld nicht trösten konnte. |
|-----|-----------------------------------------------------------|---------------------------------------------------------------|
| msc | He said: 'I would give him a job for life, but this is football. | Er sagte: “Ich würde ihm eine lebenslange Arbeit geben, aber das ist Fußball. |

Table 16: Example sentences incorrectly removed from the *en-de* self-training corpus during FILTER\textsubscript{TRG} (false negatives). The sentences are removed because there is a word in the target with the undesired grammatical gender (which is underlined along with its aligned source word), even though in both cases this word is an inanimate noun. Note that the source sentences passed the FILTER\textsubscript{SRC} step due to the gendered words in **bold**.

| Subset | TP | TN | FP | FN |
|--------|----|----|----|----|
| feminine | 6% | 8% | 0% | 86% |
| masculine | 4% | 4% | 0% | 92% |

Table 17: Percent of true positives and negatives (TP and TN) as well as false positives and negatives (FP and FN) resulting from target morphological filtering on a subset of the *en-de* pseudo-parallel data.

| Model | Acc | $\Delta G$ | $\Delta R$ |
|-------|-----|------------|------------|
| Baseline | 75.5 | 0.4 | 18.8 |
| GFST | 85.4 | -4.4 | -0.3 |
| GFST\textsubscript{Src} | 78.7 | -1.3 | 11.8 |

Table 18: WinoMT scores on *en-de* without target filtering (GFST\textsubscript{Src}), compared to the baseline and the GFST model with target filtering.

pale 428K feminine and masculine samples from the source-filtered EN candidate sentences. As shown in Table 18, GFST\textsubscript{Src} improves the gender translation accuracy when compared to the baseline model, obtaining 3.2% higher accuracy and 7 points lower $\Delta R$. However, the margin of improvement is significantly lower than for the standard GFST model. These results empirically indicate the usefulness of performing target-side filtering with a morphological analysis tool. We hypothesize that even a lower percentage of gender translation errors during self-training can hamper the model. In addition, for our lower-resource language pairs, we believe this aggressive filtering will be even more beneficial than for *en-de*. 

1654