Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem

Danielle Saunders and **Bill Byrne**
Department of Engineering, University of Cambridge, UK
{ds636, wjb31}@cam.ac.uk

Abstract

Training data for NLP tasks often exhibits gender bias in that fewer sentences refer to women than to men. In Neural Machine Translation (NMT) gender bias has been shown to reduce translation quality, particularly when the target language has grammatical gender. The recent WinoMT challenge set allows us to measure this effect directly (Stanovsky et al., 2019).

Ideally we would reduce system bias by simply debiasing all data prior to training, but achieving this effectively is itself a challenge. Rather than attempt to create a 'balanced' dataset, we use transfer learning on a small set of trusted, gender-balanced examples. This approach gives strong and consistent improvements in gender debiasing with much less computational cost than training from scratch.

A known pitfall of transfer learning on new domains is 'catastrophic forgetting', which we address both in adaptation and in inference. During adaptation we show that Elastic Weight Consolidation allows a performance trade-off between general translation quality and bias reduction. During inference we propose a lattice-rescoring scheme which outperforms all systems evaluated in Stanovsky et al. (2019) on WinoMT with no degradation of general test set BLEU, and we show this scheme can be applied to remove gender bias in the output of 'black box' online commercial MT systems. We demonstrate our approach translating from English into three languages with varied linguistic properties and data availability.

1 Introduction

As language processing tools become more prevalent concern has grown over their susceptibility to social biases and their potential to propagate bias (Hovy and Spruit, 2016; Sun et al., 2019). Natural language training data inevitably reflects biases present in our society. For example, gender bias manifests itself in training data which features more examples of men than of women. Tools trained on such data will then exhibit or even amplify the biases (Zhao et al., 2017).

Gender bias is a particularly important problem for Neural Machine Translation (NMT) into gender-inflected languages. An over-prevalence of some gendered forms in the training data leads to translations with identifiable errors (Stanovsky et al., 2019). Translations are better for sentences involving men and for sentences containing stereotypical gender roles. For example, mentions of male doctors are more reliably translated than those of male nurses (Sun et al., 2019; Prates et al., 2019).

Recent approaches to the bias problem in NLP have involved training from scratch on artificially gender-balanced versions of the original dataset (Zhao et al., 2018; Zmigrod et al., 2019) or with debiased embeddings (Escudé Font and Costa-jussà, 2019; Bolukbasi et al., 2016). While these approaches may be effective, training from scratch is inefficient and gender-balancing embeddings or large parallel datasets are challenging problems (Gonen and Goldberg, 2019).

Instead we propose treating gender debiasing as a domain adaptation problem, since NMT models can very quickly adapt to a new domain (Freitag and Al-Onaizan, 2016). To the best of our knowledge this work is the first to attempt NMT bias reduction by fine-tuning, rather than retraining. We consider three aspects of this adaptation problem: creating less biased adaptation data, parameter adaptation using this data, and inference with the debiased models produced by adaptation.

Regarding data, we suggest that a small, trusted gender-balanced set could allow more efficient and effective gender debiasing than a larger, noisier set. To explore this we create a tiny, handcrafted profession-based dataset for transfer learning. For contrast, we also consider fine-tuning on a coun-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7724–7736

terfactual subset of the full dataset and propose a straightforward scheme for artificially gender-balancing parallel text for NMT.

We find that during domain adaptation improvement on the gender-debiased domain comes at the expense of translation quality due to catastrophic forgetting (French, 1999). We can balance improvement and forgetting with a regularised training procedure, Elastic Weight Consolidation (EWC), or in inference by a two-step lattice rescoring procedure.

We experiment with three language pairs, assessing the impact of debiasing on general domain BLEU and on the WinoMT challenge set (Stanovsky et al., 2019). We find that continued training on the handcrafted set gives far stronger and more consistent improvements in gender-debiasing with orders of magnitude less training time, although as expected general translation performance as measured by BLEU decreases.

We further show that regularised adaptation with EWC can reduce bias while limiting degradation in general translation quality. We also present a lattice rescoring procedure in which initial hypotheses produced by the biased baseline system are transduced to create gender-inflected search spaces which can be rescored by the adapted model. We believe this approach, rescoring with models targeted to remove bias, is novel in NMT. The rescoring procedure improves WinoMT accuracy by up to 30% with no decrease in BLEU on the general test set.

Recent recommendations for ethics in Artificial Intelligence have suggested that social biases or imbalances in a dataset be addressed prior to model training (HLEG, 2019). This recommendation presupposes that the source of bias in a dataset is both obvious and easily adjusted. We show that debiasing a full NMT dataset is difficult, and suggest alternative efficient and effective approaches for debiasing a model after it is trained. This avoids the need to identify and remove all possible biases prior to training, and has the added benefit of preserving privacy, since no access to the original data or knowledge of its contents is required. As evidence, in section 3.4.5, we show this scheme can be applied to remove gender bias in the output of 'black box' online commercial MT systems.

1.1 Related work

Vanmassenhove et al. (2018) treat gender as a domain for machine translation, training from scratch by augmenting Europarl data with a tag indicating the speaker's gender. This does not inherently remove gender bias from the system but allows control over the translation hypothesis gender. Moryossef et al. (2019) similarly prepend a short phrase at inference time which acts as a gender domain label for the entire sentence. These approaches are not directly applicable to text which may have more than one gendered entity per sentence, as in coreference resolution tasks.

Escudé Font and Costa-jussà (2019) train NMT models from scratch with debiased word embeddings. They demonstrate improved performance on an English-Spanish occupations task with a single profession and pronoun per sentence. We assess our fine-tuning approaches on the WinoMT coreference set, with two entities to resolve per sentence.

For monolingual NLP tasks a typical approach is gender debiasing using counterfactual data augmentation where for each gendered sentence in the data a gender-swapped equivalent is added. Zhao et al. (2018) show improvement in coreference resolution for English using counterfactual data. Zmigrod et al. (2019) demonstrate a more complicated scheme for gender-inflected languages. However, their system focuses on words in isolation, and is difficult to apply to co-reference and conjunction situations with more than one term to swap, reducing its practicality for large MT datasets.

Recent work recognizes that NMT can be adapted to domains with desired attributes using small datasets (Farajian et al., 2017; Michel and Neubig, 2018). Our choice of a small, trusted dataset for adaptation specifically to a debiased domain connects also to recent work in data selection by Wang et al. (2018), in which fine-tuning on less noisy data reduces translation noise. Similarly we propose fine-tuning on less biased data to reduce gender bias in translations. This is loosely the inverse of the approach described by Park et al. (2018) for monolingual abusive language detection, which pre-trains on a larger, less biased set.

2 Gender bias in machine translation

We focus on translating coreference sentences containing professions as a representative subset of the gender bias problem. This follows much recent work on NLP gender bias (Rudinger et al., 2018; Zhao et al., 2018; Zmigrod et al., 2019) including the release of WinoMT, a relevant challenge set for NMT (Stanovsky et al., 2019).

A sentence that highlights gender bias is:

*The **doctor** told the nurse that **she** had been busy.*

A human translator carrying out coreference resolution would infer that 'she' refers to the doctor, and correctly translate the entity to German as *Die Ärztin*. An NMT model trained on a biased dataset in which most doctors are male might incorrectly default to the masculine form, *Der Arzt*.

Data bias does not just affect translations of the stereotyped roles. Since NMT inference is usually left-to-right, a mistranslation can lead to further, more obvious mistakes later in the translation. For example, our baseline en-de system translates the English sentence

*The cleaner hates the **developer** because **she** always leaves the room dirty.*

to the German

*Der Reiniger haßt **den Entwickler**, weil **er** den Raum immer schmutzig lässt.*

Here not only is 'developer' mistranslated as the masculine *den Entwickler* instead of the feminine *die Entwicklerin*, but an unambiguous pronoun translation later in the sentence is incorrect: *er* ('he') is produced instead of *sie* ('she').

In practice, not all translations with gender-inflected words can be unambiguously resolved. A simple example is:

The doctor had been busy.

This would likely be translated with a masculine entity according to the conventions of a language, unless extra-sentential context was available. As well, some languages have adopted gender-neutral singular pronouns and profession terms, both to include non-binary people and to avoid the social biases of gendered language (Misersky et al., 2019). However, the target languages supported by WinoMT lack widely-accepted non-binary inflection conventions (Ackerman, 2019). This paper addresses gender bias that can be resolved at the sentence level and evaluated with existing test sets, and does not address these broader challenges.

2.1 WinoMT challenge set and metrics

WinoMT (Stanovsky et al., 2019) is a recently proposed challenge set for gender bias in NMT. Moreover it is the only significant challenge set we are aware of to evaluate translation gender bias comparably across several language pairs. It permits automatic bias evaluation for translation from English to eight target languages with grammatical gender. The source side of WinoMT is 3888 concatenated sentences from Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018). These are coreference resolution datasets in which each sentence contains a primary entity which is co-referent with a pronoun – *the doctor* in the first example above and *the developer* in the second – and a secondary entity – *the nurse* and *the cleaner* respectively.

WinoMT evaluation extracts the grammatical gender of the primary entity from each translation hypothesis by automatic word alignment followed by morphological analysis. WinoMT then compares the translated primary entity with the gold gender, with the objective being a correctly gendered translation. The authors emphasise the following metrics over the challenge set:

- **Accuracy** – percentage of hypotheses with the correctly gendered primary entity.
- **ΔG** – difference in F_1 score between the set of sentences with masculine entities and the set with feminine entities.
- **ΔS** – difference in accuracy between the set of sentences with pro-stereotypical ('pro') entities and those with anti-stereotypical ('anti') entities, as determined by Zhao et al. (2018) using US labour statistics. For example, the 'pro' set contains male doctors and female nurses, while 'anti' contains female doctors and male nurses.

Our main objective is increasing accuracy. We also report on ΔG and ΔS for ease of comparison to previous work. Ideally the absolute values of ΔG and ΔS should be close to 0. A high positive ΔG indicates that a model translates male entities better, while a high positive ΔS indicates that a model stereotypes male and female entities. Large negative values for ΔG and ΔS, indicating a bias towards female or anti-stereotypical translation, are as undesirable as large positive values.

We note that ΔS can be significantly skewed by low-accuracy systems. A model generating male forms for most test sentences, stereotypical roles or not, will have very low ΔS, since its pro- and anti-stereotypical class accuracy will both be about 50%. Consequently in Appendix A we report:

- **M:F** – ratio of hypotheses with male predictions to those with female predictions.

This should be close to 1.0, since WinoMT balances male- and female-labelled sentences. M:F correlates strongly with ΔG, but we consider M:F

easier to interpret, particularly since very high or low M:F reduce the relevance of ΔS.

Finally, we wish to reduce gender bias without reducing translation performance. We report **BLEU** (Papineni et al., 2002) on separate, general test sets for each language pair. WinoMT is designed to work without target language references, and so it is not possible to measure translation performance on this set by measures such as BLEU.

2.2 Gender debiased datasets

2.2.1 Handcrafted profession dataset

Our hypothesis is that the absence of gender bias can be treated as a small domain for the purposes of NMT model adaptation. In this case a well-formed small dataset may give better results than attempts at debiasing the entire original dataset.

We therefore construct a tiny, trivial set of gender-balanced English sentences which we can easily translate into each target language. The sentences follow the template:

The [*PROFESSION*] *finished* [*his*|*her*] *work.*

We refer to this as the *handcrafted* set[1]. Each profession is from the list collected by Prates et al. (2019) from US labour statistics. We simplify this list by removing field-specific adjectives. For example, we have a single profession 'engineer', as opposed to specifying industrial engineer, locomotive engineer, etc. In total we select 194 professions, giving just 388 sentences in a gender-balanced set.

With manually translated masculine and feminine templates, we simply translate the masculine and feminine forms of each listed profession for each target language. In practice this translation is via an MT first-pass for speed, followed by manual checking, but given available lexicons this could be further automated. We note that the handcrafted sets contain no examples of coreference resolution and very little variety in terms of grammatical gender. A set of more complex sentences targeted at the coreference task might further improve WinoMT scores, but would be more difficult to produce for new languages.

We wish to distinguish between a model which improves gender translation, and one which improves its WinoMT scores simply by learning the vocabulary for previously unseen or uncommon professions. We therefore create a *handcrafted no-overlap* set, removing source sentences with professions occurring in WinoMT to leave 216 sentences. We increase this set back to 388 examples with balanced adjective-based sentences in the same pattern, e.g. *The tall* [*man*|*woman*] *finished* [*his*|*her*] *work*.

[1]Handcrafted sets available at https://github.com/DCSaunders/gender-debias

2.2.2 Counterfactual datasets

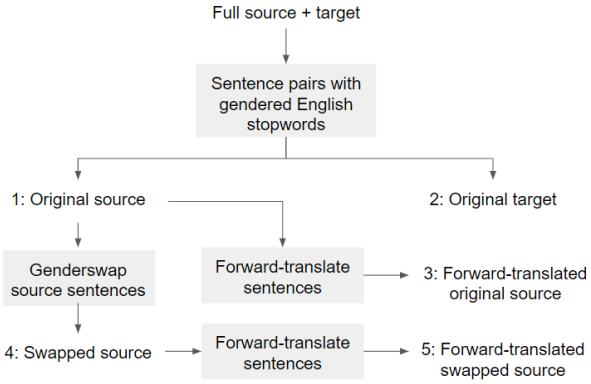


Figure 1: Generating counterfactual datasets for adaptation. The **Original** set is 1||2, a simple subset of the full dataset. **FTrans original** is 1||3, **FTrans swapped** is 4||5, and **Balanced** is 1,4||2,5

For contrast, we fine-tune on an approximated counterfactual dataset. Counterfactual data augmentation is an intuitive solution to bias from data over-representation (Lu et al., 2018). It involves identifying the subset of sentences containing bias – in this case gendered terms – and, for each one, adding an equivalent sentence with the bias reversed – in this case a gender-swapped version.

While counterfactual data augmentation is relatively simple for sentences in English, the process for inflected languages is challenging, involving identifying and updating words that are co-referent with all gendered entities in a sentence. Gender-swapping MT training data additionally requires that the same entities are swapped in the corresponding parallel sentence. A robust scheme for gender-swapping multiple entities in inflected language sentences directly, together with corresponding parallel text, is beyond the scope of this paper. Instead we suggest a rough but straightforward approach for counterfactual data augmentation for NMT which to the best of our knowledge is the first application to parallel sentences.

We first perform simple gender-swapping on the subset of the English source sentences with gendered terms. We use the approach described in Zhao et al. (2018) which swaps a fixed list of gen-

dered stopwords (e.g. *man* / *woman*, *he* / *she*).[2]. We then greedily forward-translate the gender-swapped English sentences with a baseline NMT model trained on the the full source and target text, producing gender-swapped target language sentences.

This lets us compare four related sets for gender debiasing adaptation, as illustrated in Figure 1:

- **Original**: a subset of parallel sentences from the original training data where the source sentence contains gendered stopwords.
- **Forward-translated (FTrans) original**: the source side of the *original* set with forward-translated target sentences.
- **Forward-translated (FTrans) swapped**: the *original* source sentences are gender-swapped, then forward-translated to produce gender-swapped target sentences.
- **Balanced**: the concatenation of the *original* and *FTrans swapped* parallel datasets. This is twice the size of the other counterfactual sets.

Comparing performance in adaptation of *FTrans swapped* and *FTrans original* lets us distinguish between the effects of gender-swapping and of obtaining target sentences from forward-translation.

2.3 Debiasing while maintaining general translation performance

Fine-tuning a converged neural network on data from a distinct domain typically leads to catastrophic forgetting of the original domain (French, 1999). We wish to adapt to the gender-balanced domain without losing general translation performance. This is a particular problem when fine-tuning on the very small and distinct handcrafted adaptation sets.

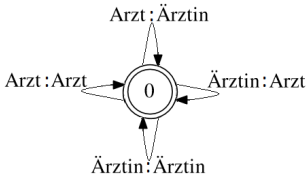
2.3.1 Regularized training

Regularized training is a well-established approach for minimizing catastrophic forgetting during domain adaptation of machine translation (Barone et al., 2017). One effective form is Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) which in NMT has been shown to maintain or even improve original domain performance (Thompson et al., 2019; Saunders et al., 2019). In EWC a regularization term is added to the original log likelihood loss function L when training the debiased model (DB):

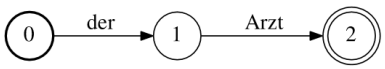
$$L'(\theta^{DB}) = L(\theta^{DB}) + \lambda \sum_j F_j(\theta_j^{DB} - \theta_j^B)^2 \quad (1)$$

θ_j^B are the converged parameters of the original biased model, and θ_j^{DB} are the current debiased model parameters. $F_j = \mathbb{E}\left[\nabla^2 L(\theta_j^B)\right]$, a Fisher information estimate over samples from the biased data under the biased model. We apply EWC when performance on the original validation set drops, selecting hyperparameter λ via validation set BLEU.

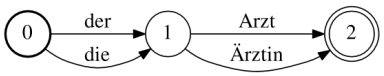
2.3.2 Gender-inflected search spaces for rescoring with debiased models



(a) A subset of flower transducer T. T maps vocabulary to itself as well as to differently-gendered inflections.



(b) Acceptor Y_B representing the biased first-pass translation $\mathbf{y_B}$ for source fragment 'the doctor'. The German hypothesis has the male form.



(c) Gender-inflected search space constructed from the biased hypothesis 'der Arzt'. Projection of the composition $Y_B \circ T$ contains paths with differently-gendered inflections of the original biased hypothesis. This lattice can now be rescored by a debiased model.

Figure 2: Finite State Transducers for lattice rescoring.

An alternative approach for avoiding catastrophic forgetting takes inspiration from lattice rescoring for NMT (Stahlberg et al., 2016) and Grammatical Error Correction (Stahlberg et al., 2019). We assume we have two NMT models. With one we decode fluent translations which contain gender bias (B). For the one-best hypothesis we would translate:

$$\mathbf{y_B} = \text{argmax}_{\mathbf{y}} p_B(\mathbf{y}|\mathbf{x}) \quad (2)$$

The other model has undergone debiasing (DB) at a cost to translation performance, producing:

$$\mathbf{y_{DB}} = \text{argmax}_{\mathbf{y}} p_{DB}(\mathbf{y}|\mathbf{x}) \quad (3)$$

[2]The stopword list and swapping script are provided by the authors of Zhao et al. (2018) at https://github.com/uclanlp/corefBias

We construct a flower transducer T that maps each word in the target language's vocabulary to itself, as well as to other forms of the same word with different gender inflections (Figure 2a). We also construct Y_B, a lattice with one path representing the biased but fluent hypothesis $\mathbf{y_B}$ (Figure 2b).

The acceptor $\mathcal{P}(\mathbf{y_B}) = \text{proj}_{\text{output}}(Y_B \circ T)$ defines a language consisting of all the gender-inflected versions of the biased first-pass translation $\mathbf{y_B}$ that are allowed by T (Figure 2c). We can now decode with lattice rescoring (LR) by constraining inference to $\mathcal{P}(\mathbf{y_B})$:

$$\mathbf{y_{LR}} = \text{argmax}_{\mathbf{y} \in \mathcal{P}(\mathbf{y_B})} p_{DB}(\mathbf{y}|\mathbf{x}) \quad (4)$$

In practice we use beam search to decode the various hypotheses, and construct T using heuristics on large vocabulary lists for each target language.

3 Experiments

3.1 Languages and data

WinoMT provides an evaluation framework for translation from English to eight diverse languages. We select three pairs for experiments: English to German (en-de), English to Spanish (en-es) and English to Hebrew (en-he). Our selection covers three language groups with varying linguistic properties: Germanic, Romance and Semitic. Training data available for each language pair also varies in quantity and quality. We filter training data based on parallel sentence lengths and length ratios.

For **en-de**, we use 17.6M sentence pairs from WMT19 news task datasets (Barrault et al., 2019). We validate on newstest17 and test on newstest18.

For **en-es** we use 10M sentence pairs from the United Nations Parallel Corpus (Ziemski et al., 2016). While still a large set, the UNCorpus exhibits far less diversity than the en-de training data. We validate on newstest12 and test on newstest13.

For **en-he** we use 185K sentence pairs from the multilingual TED talks corpus (Cettolo et al., 2014). This is both a specialized domain and a much smaller training set. We validate on the IWSLT 2012 test set and test on IWSLT 2014.

Table 1 summarises the sizes of datasets used, including their proportion of gendered sentences and ratio of sentences in the English source data containing male and female stopwords. A gendered sentence contains at least one English gendered stopword as used by Zhao et al. (2018).

Interestingly all three datasets have about the same proportion of gendered sentences: 11-12% of the overall set. While en-es appears to have a much more balanced gender ratio than the other pairs, examining the data shows this stems largely from sections of the UNCorpus containing phrases like 'empower women' and 'violence against women', rather than gender-balanced professional entities.

	Training	Gendered training	M:F	Test
en-de	17.5M	2.1M	2.4	3K
en-es	10M	1.1M	1.1	3K
en-he	185K	21.4K	1.8	1K

Table 1: Parallel sentence counts. A gendered sentence pair has minimum one gendered stopword on the English side. M:F is ratio of male vs female gendered training sentences.

For en-de and en-es we learn joint 32K BPE vocabularies on the training data (Sennrich et al., 2016). For en-he we use separate source and target vocabularies. The Hebrew vocabulary is a 2k-merge BPE vocabulary, following the recommendations of Ding et al. (2019) for smaller vocabularies when translating into lower-resource languages. For the en-he source vocabulary we experimented both with learning a new 32K vocabulary and with reusing the joint BPE vocabulary trained on the largest set – en-de – which lets us initialize the en-he system with the pre-trained en-de model. The latter resulted in higher BLEU and faster training.

3.2 Training and inference

For all models we use a Transformer model (Vaswani et al., 2017) with the 'base' parameter settings given in Tensor2Tensor (Vaswani et al., 2018). We train baselines to validation set BLEU convergence on one GPU, delaying gradient updates by factor 4 to simulate 4 GPUs (Saunders et al., 2018). During fine-tuning training is continued without learning rate resetting. Normal and lattice-constrained decoding is via SGNMT[3] with beam size 4. BLEU scores are calculated for cased, detokenized output using SacreBLEU (Post, 2018)

3.3 Lattice rescoring with debiased models

For lattice rescoring we require a transducer T containing gender-inflected forms of words in the target vocabulary. To obtain the vocabulary for German we use all unique words in the full target training dataset. For Spanish and Hebrew, which have smaller and less diverse training sets, we use 2018

[3] https://github.com/ucam-smt/sgnmt

OpenSubtitles word lists[4]. We then use DEMorphy (Altinok, 2018) for German, spaCy (Honnibal and Montani, 2017) for Spanish and the small set of gendered suffixes for Hebrew (Schwarzwald, 1982) to approximately lemmatize each vocabulary word and generate its alternately-gendered forms. While there are almost certainly paths in T containing non-words, we expect these to have low likelihood under the debiasing models. For lattice compositions we use the efficient OpenFST implementations (Allauzen et al., 2007).

3.4 Results

3.4.1 Baseline analysis

In Table 2 we compare our three baselines to commercial systems on WinoMT, using results quoted directly from Stanovsky et al. (2019). Our baselines achieve comparable accuracy, masculine/feminine bias score ΔG and pro/anti stereotypical bias score ΔS to four commercial translation systems, outscoring at least one system for each metric on each language pair.

The ΔS for our en-es baseline is surprisingly small. Investigation shows this model predicts male and female entities in a ratio of over 6:1. Since almost all entities are translated as male, pro- and anti-stereotypical class accuracy are both about 50%, making ΔS very small. This highlights the importance of considering ΔS in the context of ΔG and M:F prediction ratio.

3.4.2 Counterfactual adaptation

Table 3 compares our baseline model with the results of unregularised fine-tuning on the counterfactual sets described in Section 2.2.2.

Fine-tuning for one epoch on *original*, a subset of the original data with gendered English stopwords, gives slight improvement in WinoMT accuracy and ΔG for all language pairs, while ΔS worsens. We suggest this set consolidates examples present in the full dataset, improving performance on gendered entities generally but emphasizing stereotypical roles.

On the *FTrans original* set ΔG increases sharply relative to the *original* set, while ΔS decreases. We suspect this set suffers from bias amplification (Zhao et al., 2017) introduced by the baseline system during forward-translation. The model therefore over-predicts male entities even more heavily than we would expect given the gender makeup of the adaptation data's source side. Over-predicting male entities lowers ΔS artificially.

Adapting to *FTrans swapped* increases accuracy and decreases both ΔG and ΔS relative to the baseline for en-de and en-es. This is the desired result, but not a particularly strong one, and it is not replicated for en-he. The *balanced* set has a very similar effect to the *FTrans swapped* set, with a smaller test BLEU difference from the baseline.

We do find that the largest improvement in WinoMT accuracy consistently corresponds to the model predicting male and female entities in the closest ratio (see Appendix A). However, the best ratios for models adapted to these datasets are 2:1 or higher, and the accuracy improvement is small.

The purpose of EWC regularization is to avoid catastrophic forgetting of general translation ability. This does not occur in the counterfactual experiments, so we do not apply EWC. Moreover, WinoMT accuracy gains are small with standard fine-tuning, which allows maximum adaptation: we suspect EWC would prevent any improvements.

Overall, improvements from fine-tuning on counterfactual datasets (*FTrans swapped* and *balanced*) are present. However, they are not very different from the improvements when fine-tuning on equivalent non-counterfactual sets (*original* and *FTrans original*). Improvements are also inconsistent across language pairs.

3.4.3 Handcrafted profession set adaptation

Results for fine-tuning on the handcrafted set are given in lines 3-6 of Table 4. These experiments take place in minutes on a single GPU, compared to several hours when fine-tuning on the counterfactual sets and far longer if training from scratch.

Fine-tuning on the handcrafted sets gives a much faster BLEU drop than fine-tuning on counterfactual sets. This is unsurprising since the handcrafted sets are domains of new sentences with consistent sentence length and structure. By contrast the counterfactual sets are less repetitive and close to subsets of the original training data, slowing forgetting. We believe the degradation here is limited only by the ease of fitting the small handcrafted sets.

Line 4 of Table 4 adapts to the handcrafted set, stopping when validation BLEU degrades by 5% on each language pair. This gives a WinoMT accuracy up to 19 points above the baseline, far more improvement than the best counterfactual result. Difference in gender score ΔG improves by at least

[4]Accessed Oct 2019 from https://github.com/hermitdave/FrequencyWords/

	en-de			en-es			en-he		
	Acc	ΔG	ΔS	Acc	ΔG	ΔS	Acc	ΔG	ΔS
Microsoft	**74.1**	**0.0**	30.2	47.3	36.8	23.2	48.1	14.9	32.9
Google	59.4	12.5	12.5	53.1	23.4	21.3	**53.7**	**7.9**	37.8
Amazon	62.4	12.9	16.7	**59.4**	**15.4**	22.3	50.5	10.3	47.3
SYSTRAN	48.6	34.5	**10.3**	45.6	46.3	15.0	46.6	20.5	**24.5**
Baseline	60.1	18.6	13.4	49.6	36.7	**2.0**	51.3	15.1	26.4

Table 2: WinoMT accuracy, masculine/feminine bias score ΔG and pro/anti stereotypical bias score ΔS for our baselines compared to commercial systems, whose scores are quoted directly from Stanovsky et al. (2019).

	en-de				en-es				en-he			
	BLEU	Acc	ΔG	ΔS	BLEU	Acc	ΔG	ΔS	BLEU	Acc	ΔG	ΔS
Baseline	42.7	60.1	18.6	13.4	27.8	49.6	36.7	2.0	**23.8**	51.3	15.1	26.4
Original	41.8	60.7	15.9	15.6	**28.3**	53.0	**24.3**	10.8	23.5	**53.6**	**12.2**	31.7
FTrans original	43.3	60.0	20.0	13.9	27.4	51.6	31.6	-4.8	23.4	48.7	23.0	**20.9**
FTrans swapped	**43.4**	63.0	15.4	12.7	27.4	**53.7**	24.5	-3.8	23.7	48.1	20.7	22.7
Balanced	42.5	**64.0**	**12.6**	**12.4**	27.7	52.8	26.2	**1.9**	**23.8**	48.3	20.8	24.0

Table 3: General test set BLEU and WinoMT scores after unregularised fine-tuning the baseline on four gender-based adaptation datasets. Improvements are inconsistent across language pairs.

		en-de				en-es				en-he			
		BLEU	Acc	ΔG	ΔS	BLEU	Acc	ΔG	ΔS	BLEU	Acc	ΔG	ΔS
1	Baseline	**42.7**	60.1	18.6	13.4	**27.8**	49.6	36.7	2.0	23.8	51.3	15.1	26.4
2	Balanced	42.5	64.0	12.6	12.4	27.7	52.8	26.2	**1.9**	23.8	48.3	20.8	24.0
3	Handcrafted (no overlap)	40.6	71.2	3.9	10.6	26.5	64.1	9.5	-10.3	23.1	56.5	-6.2	28.9
4	Handcrafted	40.8	78.3	**-0.7**	6.5	26.7	68.6	5.2	-8.7	22.9	65.7	-3.3	20.2
5	Handcrafted (converged)	36.5	**85.3**	-3.2	6.3	25.3	**72.4**	**0.8**	-3.9	22.5	**72.6**	-4.2	21.0
6	Handcrafted EWC	42.2	74.2	2.2	8.4	27.2	67.8	5.8	-8.2	23.3	65.2	**-0.4**	25.3
7	Rescore 1 with 3	**42.7**	68.3	7.6	11.8	**27.8**	62.4	11.1	-9.7	**23.9**	56.2	2.8	23.0
8	Rescore 1 with 4	**42.7**	74.5	2.1	6.5	**27.8**	64.2	9.7	-10.8	**23.9**	58.4	2.7	18.6
9	Rescore 1 with 5	42.5	81.7	-2.4	**1.5**	27.7	68.4	5.6	-8.0	23.6	63.8	0.7	**12.9**

Table 4: General test set BLEU and WinoMT scores after fine-tuning on the handcrafted profession set, compared to fine-tuning on the most consistent counterfactual set. Lines 1-2 duplicated from Table 3. Lines 3-4 vary adaptation data. Lines 5-6 vary adaptation training procedure. Lines 7-9 apply lattice rescoring to baseline hypotheses.

a factor of 4. Stereotyping score ΔS also improves far more than for counterfactual fine-tuning. Unlike the Table 3 results, the improvement is consistent across all WinoMT metrics and all language pairs.

The model adapted to no-overlap handcrafted data (line 3) gives a similar drop in BLEU to the model in line 4. This model also gives stronger and more consistent WinoMT improvements over the baseline compared to the balanced counterfactual set, despite the implausibly strict scenario of no English profession vocabulary in common with the challenge set. This demonstrates that the adapted model does not simply memorise vocabulary.

The drop in BLEU and improvement on WinoMT can be explored by varying the training procedure. The model of line 5 simply adapts to handcrafted data for more iterations with no regularisation, to approximate loss convergence on the handcrafted set. This leads to a severe drop in BLEU, but even higher WinoMT scores.

In line 6 we regularise adaptation with EWC. There is a trade-off between general translation performance and WinoMT accuracy. With EWC regularization tuned to balance validation BLEU and WinoMT accuracy, the decrease is limited to about 0.5 BLEU on each language pair. Adapting to convergence, as in line 5, would lead to further WinoMT gains at the expense of BLEU.

3.4.4 Lattice rescoring with debiased models

In lines 7-9 of Table 4 we consider lattice-rescoring the baseline output, using three models debiased on the handcrafted data.

Line 7 rescores the general test set hypotheses (line 1) with a model adapted to handcrafted data that has no source language profession vocabulary overlap with the test set (line 3). This scheme shows no BLEU degradation from the baseline on any language and in fact a slight improvement on en-he. Accuracy improvements on WinoMT

	en-de			en-es			en-he		
	Acc	ΔG	ΔS	Acc	ΔG	ΔS	Acc	ΔG	ΔS
1	**82.0** (74.1)	-3.0 (0.0)	4.0 (30.2)	65.8 (47.3)	3.8 (36.8)	**1.9** (23.2)	63.9 (48.1)	-2.6 (14.9)	23.8 (32.9)
2	80.0 (59.4)	-3.0 (12.5)	**2.7** (12.5)	68.9 (53.1)	**0.6** (23.4)	4.6 (21.3)	**64.6** (53.7)	-1.8 (7.9)	21.5 (37.8)
3	81.8 (62.4)	**-2.6** (12.9)	4.3 (16.7)	**71.1** (59.4)	0.7 (15.4)	6.7 (22.3)	62.8 (50.5)	**-1.1** (10.3)	26.9 (47.3)
4	78.4 (48.6)	-4.0 (34.5)	5.3 (10.3)	66.0 (45.6)	4.2 (46.3)	-2.1 (15.0)	62.5 (46.6)	-2.0 (20.5)	**10.2** (24.5)

Table 5: We generate gender-inflected lattices from commercial system translations, collected by Stanovsky et al. (2019) (1: Microsoft, 2: Google, 3: Amazon, 4: SYSTRAN). We then rescore with the debiased model from line 5 of Table 4. Scores are for the rescored hypotheses, with bracketed baseline scores duplicated from Table 2.

are only slightly lower than for decoding with the rescoring model directly, as in line 3.

In line 8, lattice rescoring with the non-converged model adapted to handcrafted data (line 4) likewise leaves general BLEU unchanged or slightly improved. When lattice rescoring the WinoMT challenge set, 79%, 76% and 49% of the accuracy improvement is maintained on en-de, en-es and en-he respectively. This corresponds to accuracy gains of up to 30% relative to the baselines with no general translation performance loss.

In line 9, lattice-rescoring with the converged model of line 5 limits BLEU degradation to 0.2 BLEU on all languages, while maintaining 85%, 82% and 58% of the WinoMT accuracy improvement from the converged model for the three language pairs. Lattice rescoring with this model gives accuracy improvements over the baseline of 36%, 38% and 24% for en-de, en-es and en-he.

Rescoring en-he maintains a much smaller proportion of WinoMT accuracy improvement than en-de and en-es. We believe this is because the en-he baseline is particularly weak, due to a small and non-diverse training set. The baseline must produce some inflection of the correct entity before lattice rescoring can have an effect on gender bias.

3.4.5 Reducing gender bias in 'black box' commercial systems

Finally, in Table 5, we apply the gender inflection transducer to the commercial system translations[5] listed in Table 2. We find rescoring these lattices with our strongest debiasing model (line 5 of Table 4) substantially improves WinoMT accuracy for all systems and language pairs.

One interesting observation is that WinoMT accuracy after rescoring tends to fall in a fairly narrow range for each language relative to the performance range of the baseline systems. For example, a 25.5% range in baseline en-de accuracy becomes a 3.6% range after rescoring. This suggests that our rescoring approach is not limited as much by the bias level of the baseline system as by the gender-inflection transducer and the models used in rescoring. Indeed, we emphasise that the large improvements reported in Table 5 do not require any knowledge of the commercial systems or the data they were trained on; we use only the translation hypotheses they produce and our own rescoring model and transducer.

4 Conclusions

We treat the presence of gender bias in NMT systems as a domain adaptation problem. We demonstrate strong improvements under the WinoMT challenge set by adapting to tiny, handcrafted gender-balanced datasets for three language pairs.

While naive domain adaptation leads to catastrophic forgetting, we further demonstrate two approaches to limit this: EWC and a lattice rescoring approach. Both allow debiasing while maintaining general translation performance. Lattice rescoring, although a two-step procedure, allows far more debiasing and potentially no degradation, without requiring access to the original model.

We suggest small-domain adaptation as a more effective and efficient approach to debiasing machine translation than counterfactual data augmentation. We do not claim to fix the bias problem in NMT, but demonstrate that bias can be reduced without degradation in overall translation quality.

Acknowledgments

This work was supported by EPSRC grants EP/M508007/1 and EP/N509620/1 and has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service[6] funded by EPSRC Tier-2 capital grant EP/P020259/1.

[5] The raw commercial system translations are provided by the authors of Stanovsky et al. (2019) at https://github.com/gabrielStanovsky/mt_gender

[6] http://www.hpc.cam.ac.uk

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

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

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. 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*, pages 4349–4357.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, page 57.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

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.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for Neural Machine Translation. *CoRR*, abs/1612.06897.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

AI HLEG. 2019. *Ethics guidelines for trustworthy AI*. High-Level Expert Group on Artificial Intelligence.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Julia Misersky, Asifa Majid, and Tineke M Snijders. 2019. Grammatical gender in German influences how role-nouns are interpreted: Evidence from erps. *Discourse Processes*, 56(8):643–654.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

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

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2018. Multi-representation ensembles and delayed SGD updates improve syntax-based NMT. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–325, Melbourne, Australia. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2019. Domain adaptive inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 222–228, Florence, Italy. Association for Computational Linguistics.

Ora Schwarzwald. 1982. Feminine formation in modern Hebrew. *Hebrew Annual Review*, 6:153–178.

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.

Felix Stahlberg, Christopher Bryant, and Bill Byrne. 2019. Neural grammatical error correction with finite state transducers. 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 4033–4039.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016. Syntactically guided neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 299–305, Berlin, Germany. Association for Computational Linguistics.

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

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.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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 6000–6010.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Belgium, Brussels. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

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

A WinoMT male:female prediction ratio

We report ΔG on WinoMT for easy comparison to previous work, but also find that M:F prediction ratio on WinoMT is an intuitive and interesting metric. Tables 6 and 7 expand on the results of Tables 3 and 4 respectively.

	en-de			en-es			en-he		
	BLEU	Acc	M:F	BLEU	Acc	M:F	BLEU	Acc	M:F
Baseline	42.7	60.1	3.4	27.8	49.6	6.3	**23.8**	51.3	2.2
Original	41.8	60.7	3.1	**28.3**	53.0	**4.0**	23.5	**53.6**	**2.0**
FTrans original	43.3	60.0	3.9	27.4	51.6	5.4	23.4	48.7	3.0
FTrans swapped	**43.4**	63.0	3.1	27.4	**53.7**	**4.0**	23.7	48.1	2.6
Balanced	42.5	**64.0**	**2.7**	27.7	52.8	4.3	**23.8**	48.3	2.7

Table 6: General test set BLEU and WinoMT scores after unregularised fine-tuning the baseline on four gender-based adaptation datasets.

		en-de			en-es			en-he		
		BLEU	Acc	M:F	BLEU	Acc	M:F	BLEU	Acc	M:F
1	Baseline	**42.7**	60.1	3.4	**27.8**	49.6	6.3	23.8	51.3	2.2
2	Balanced	42.5	64.0	2.7	27.7	52.8	4.3	23.8	48.3	2.7
3	Handcrafted (no overlap)	40.6	71.2	1.7	26.5	64.1	2.4	23.1	56.5	0.8
4	Handcrafted	40.8	78.3	1.3	26.7	68.6	1.9	22.9	65.7	0.9
5	Handcrafted (converged)	36.5	**85.3**	**0.9**	25.3	**72.4**	**1.5**	22.5	**72.6**	**1.0**
6	Handcrafted EWC	42.2	74.2	1.6	27.2	67.8	2.0	23.3	65.2	1.2
7	Rescore 1 with 3	**42.7**	68.3	2.2	**27.8**	62.4	2.3	**23.9**	56.2	1.3
8	Rescore 1 with 4	**42.7**	74.5	1.6	**27.8**	64.2	2.1	**23.9**	58.4	1.3
9	Rescore 1 with 5	42.5	81.7	**1.1**	27.7	68.4	1.8	23.6	63.8	1.3

Table 7: General test set BLEU and WinoMT scores after fine-tuning on the handcrafted profession set, compared to fine-tuning on the most consistent counterfactual set. Lines 1-2 duplicated from Table 6. Lines 3-4 vary adaptation data. Lines 5-6 vary adaptation training procedure. Lines 7-9 apply lattice rescoring to baseline hypotheses.