Analyzing Gender Translation Errors to Identify Information Flows between the Encoder and Decoder of an NMT System

Guillaume Wisniewski and **Lichao Zhu**
LLF, Université Paris Cité & CNRS F-75013 Paris, France
{guillaume.wisniewski,lichao.zhu}@u-paris.fr

Nicolas Ballier
CLILLAC-ARP, Université Paris Cité
F-75013 Paris, France
nicolas.ballier@u-paris.fr

François Yvon
Université Paris-Saclay & CNRS, LISN
91403 Orsay, France
francois.yvon@cnrs.fr

Abstract

Multiple studies have shown that existing NMT systems demonstrate some kind of gender bias. As a result, MT output appears to err more often for feminine forms and to amplify social gender misrepresentations, which is potentially harmful to users and practioners of these technologies.

This paper continues this line of investigations and reports results obtained with a new test set in strictly controlled conditions. This setting allows us to better understand the multiple inner mechanisms that are causing these biases, which include the linguistic expressions of gender, the unbalanced distribution of masculine and feminine forms in the language, the modelling of morphological variation and the training process dynamics. To counterbalance these effects, we formulate several proposals and notably show that modifying the training loss can effectively mitigate such biases.

1 Introduction

State-of-the-art machine translation models (TMs) have been shown to suffer from gender-bias (Prates et al., 2020) and works trying to mitigate this problem constitute a very active line of research (e.g. Costa-jussà and de Jorge (2020); Saunders and Byrne (2020); Savoldi et al. (2021)). We here adopt a different point of view and try to understand why the TM, often incorrectly, chooses a masculine rather than a feminine form. For this, we identify the mechanisms which the neural network uses to extract gender information from the source side and transfer it to the target side. This study of the causes of gender bias illustrates in a more general way the inner working of neural translation systems and notably reveals information flows between the encoder and the decoder involved in a TM.

To study gender transfer, we introduce a new French-English test set specifically designed to highlight the difficulties of translating gender information between these two languages. Using a controlled test set allows us to precisely pinpoint where and how gender is expressed in source and target sentences and to quantify the information flow in an encoder-decoder architecture. Our experiments rely on both well-known methods such as probing or new methods tailored to identify gender bias such as comparing the predictions of a language model and those of a TM to better analyze the possible causes of gender bias. Considering a controlled experimental setting also allows us to assess the impact of training conditions such as subword segmentation or training data distributions.

The rest of the paper is organized as follows. In §2, we describe our test set and use it to highlight gender bias in state-of-the-art systems. In §3 we describe several experiments aiming to study how information flows between the encoder and the decoder can explain these biases. To counterbalance these effects, we formulate several proposals in §4, and notably show that a simple modification of the training loss can effectively mitigate gender bias.

2 Observing Gender biases in MT

2.1 A Controlled Set to Study Gender Bias

We first describe the controlled test set used in our experiments and explain why (and how) we use it to identify the flow of information in an encoder-decoder architecture.

This test set, built on an idea introduced in Wis-

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 153 - 163
December 8, 2022 ©2022 Association for Computational Linguistics

niewski et al. (2021), is made of 3,394 parallel sentences perfectly balanced between genders. All sentences use the following template:

- `[DET] [N]` a terminé son travail.
- The `[N]` has finished `[PRO]` work.

where `[N]` is an occupational noun chosen from the list of Dister and Moreau (2014) that matches feminine and masculine professions and occupations in French. This list was automatically translated in English with `DeepL` and manually corrected by two professional translators.[1] `[DET]` is the French determiner in agreement with the `[N]` (the feminine form la_F, the masculine form le_M or the epicene form *l'* that is used for both grammatical genders when the job noun starts with a vowel); `[PRO]` is the English possessive pronoun `her` $_F$ or `his` $_M$. For English, in the case of indefinite reference or generic NPs like "the writer", since the 1980's different strategies have been used to avoid the use of resumption by "his" and style guides for authors have recommended the use of "his or her", then "her" for generic references as well and the same period has seen the rise of the use of singular "their" (see, among others, (Bodine, 1975; Pauwels, 2000)). For the sake of simplicity and to keep our test set gender-balanced, we have only considered two forms of the possessive pronoun and our test set contains a feminine sentence (with `her`) and a masculine sentence (with `his`) for each occupational noun, even if its gender is ambiguous in the source sentence (see below).

In English, gender is unambiguously expressed in the possessive pronoun; it may also be expressed by the occupational noun, when it has different feminine and masculine forms (e.g. *actress-actor*). In most sentences, however, the occupational noun is epicene and its gender can not be inferred from the surface form. In French, gender can be expressed by the determiner, the occupational noun, or both; in rarer cases, both the determiner and the occupational noun are epicene, and the feminine and masculine versions are identical. In the latter case, as explained above, the English translation should use the possessive pronoun `their` to mark that gender is not specified. The choice we made to associate the same sentence once with the masculine pronoun and once with the feminine pronoun is a way to identify the biases of the translation system: an unbiased system would be expected to err one out of two times in the choice of pronoun, a higher error rate indicates that the system prefers one pronoun to the other. Table 1 illustrates the various ways that gender can be expressed in English and French as well as their proportion in our test set.

When translating sentences from our controlled set from French into English, the prediction of the English possessive pronoun can rely on two kinds of evidence: *i)* using cross-attention, the model can encode information about the French subject gender into the representation of the possessive pronoun;[2] *ii)* because of the decoder self-attention, the possessive pronoun representation can also encode information about the target context. This is notably the case of the English subject that encodes gender information either directly, or because its representation depends on the French subject (through cross-attention).

2.2 Direct Evidence of a Gender Bias

Before investigating the roots of gender bias in MT systems, we would first like to describe the experimental setting that will be used throughout this work and highlight the difficulties of predicting gender information. Most observations reported in this section have already been described for other models, language pairs and datasets (see e.g. (Stanovsky et al., 2019; Saunders et al., 2020) or (Stanczak and Augenstein, 2021) for an overview).

NMT System We use `JoeyNMT` (Kreutzer et al., 2019), an implementation of a translation system based on the `Transformer` model of Vaswani et al. (2017). Encoder and decoder are composed of 6 layers, each with 8 attention heads; hidden representations have dimension $d = 512$, while *feed-forward* layers has dimension 2,048. Our model comprises a grand total of about 76M parameters.

We consider the English-French parallel corpus from the WMT'15 'News' task (Bojar et al., 2015) that contains 4.8M sentences and nearly 141M French running words. All raw corpora

[1] As our sentences have a fixed structure, most translation issues were related to occupational nouns, and were resolved by mining reference dictionaries and corpora such as the COCA corpus (Davies, 2009). The resulting resource is more than three times larger than the list used in (Niu et al., 2021).

[2] The French subject can have either a direct impact through cross-attention or an indirect impact as the representation of all source tokens depends on it (via encoder self-attention). We will not try to distinguish these two effects.

Gender-marked		Proportion	Example
Determiner	Noun		
French			
yes	yes	53.0%	• (**la**$_F$ **boulangère**$_F$|**le**$_M$ **boulanger**$_M$) a fini son travail • the baker has finished (his$_M$|her$_F$) job.
yes	no	24.2%	• (**la**$_F$ **cinéaste**|**le**$_M$ **cinéaste**) a fini son travail • the film-maker has finished (his$_M$|her$_F$) job.
no	yes	14.9%	• (**l'adjointe**$_F$|**l'adjoint**$_M$) a fini son travail • the assistant has finished (his$_M$|her$_F$) job.
no	no	7.9%	• **l'artiste** a fini son travail • the artist has finished (his$_M$|her$_F$) job.
English			
no	yes	5.5%	• (**the actress**$_F$|**the actor**$_M$) has finished her$_F$|his$_M$ work • (**l'actrice**$_F$|**l'acteur**$_M$) a terminé son travail.
no	no	94.5%	• **the user** has finished her$_F$|his$_M$ work • (**l'usagère**$_F$**x**|**l'usager**$_M$) a terminé son travail.

Table 1: Examples of the various ways by which gender is expressed in our test corpus.

are segmented into sub-lexical units using the unigram model of `SentencePiece` (Kudo, 2018); the vocabularies contain 32,000 units in each language. Our model is trained by optimizing the cross-entropy with ADAM and achieves a BLEU score of 34.0 on the WMT'14 test set.

Results We report in Table 2 the accuracy with which our NMT systems are able to predict the English possessive pronoun gender and therefore to correctly capture and transfer gender information between French and English. These numbers are broken down by context, where we distinguish between the various cases of Table 1. Note that when both the determiner and the noun are epicene, both `his` and `her` (as well as `their`) are equally correct: we nonetheless report the number of mispredictions for this category - as our test set is perfectly balanced, an unbiased system predicting only `his` and `her` should have an accuracy of 50%. We also report the performance achieved by `mBART`,[3] the multilingual model of Tang et al. (2020). Note that `mBART` is a strong baseline with 610M parameters trained on, at least, a hundred times more English sentences than our system.

It appears that our system, `JoeyNMT`, has many difficulties in predicting the correct form of the possessive pronoun, with a striking difference between the accuracy of the prediction of `his` and `her`. For `mBART`, which is doing much better overall, the difference in accuracy between the two genders is about 20 points. For both systems, the accuracy is slightly better when both the determiner and noun are unambiguous. Also note that mispredictions are not only due to error on gender: the system sometimes generates sentences that do not contain any possessive pronoun or in which the possessive pronoun was either `their` or `its`. This may be because in French, occupational nouns may also refer to technical devices or machines that are used to perform the occupation, especially for the feminine form: one such example is *cafetière* that can either mean *coffeemaker*, the machine that makes coffee or (female) barkeeper, the person that makes coffee (in a bar).

Gender in source		Accuracy	
Determiner	Epicene Noun	`JoeyNMT`	`mBART`
Masc.	no	76.5	72.8
	yes	84.7	84.1
Fem.	no	33.1	60.6
	yes	31.9	65.4
Epicene	no	40.4	66.1
	yes	40.4	45.9

Table 2: Accuracy (in %) of possessive pronoun prediction by `JoeyNMT` and a strong baseline (`mBART`).

3 Uncovering the Flows of Gender Information

In this Section, we report experiments aimed to explain the results reported in Table 2 using either well-known probing methods (§3.1) or, as suggested by Fernandes et al. (2021), by comparing the prediction of a translation and a language models (§3.2). In Section 3.3 we describe the impact of our findings on training.

[3] We used the model through the HuggingFace API (Wolf et al., 2020).

	layer	encoder					
		a	terminé	son	travail	.	eos
Gender weakening							
chaque surveillant a terminé son travail.	1	73.1	73.6	65.7	63.5	53.9	56.7
	6	71.0	71.4	70.4	68.2	71.2	69.7
Gender strengthening							
le surveillant *français* a terminé son travail.	1	99.9	98.5	95.0	80.6	62.0	80.4
	6	100.0	99.7	99.7	98.9	98.8	96.9
Gender change for the direct object							
le surveillant a terminé son *travail.*	1	79.4	74.6	79.0	75.0	58.8	72.0
	6	90.3	88.8	89.2	85.3	86.2	83.3
le surveillant a terminé son *activité.*	1	80.5	75.5	78.6	62.6	57.6	67.2
	6	89.7	88.3	89.6	84.3	86.1	84.1
Syntactical distancing							
le surveillant *qui a chanté formidablement hier* a terminé son travail.	1	71.1	66.3	68.8	81.1	56.8	65.4
	6	91.5	91.0	90.5	86.8	81.2	82.1
Distractor							
.without gender weakening							
le surveillant *que cette femme critiquait* a terminé son travail.	1	65.7	66.6	69.3	79.50	62.8	68.5
	6	90.6	89.6	89.1	85.91	81.9	80.2
le surveillant *que cet homme critiquait* a terminé son travail.	1	65.4	67.0	68.7	80.0	63.4	68.2
	6	90.3	89.3	89.7	86.6	81.0	79.9
.with gender weakening							
chaque surveillant *que cet homme critiquait* a terminé son travail.	1	63.1	63.5	64.3	62.4	56.2	55.8
	6	72.1	71.4	69.7	69.9	71.8	69.2
chaque surveillant *que cette femme critiquait* a terminé son travail.	1	63.3	64.6	65.9	63.4	55.4	55.2
	6	71.8	71.8	70.0	69.2	70.2	69.5

Table 3: Accuracy (in %) of probes when manipulating the test sentences corpus

3.1 Probing

We use *probing* (Belinkov, 2022) to analyze which words in the source sentences convey gender information: a *probe* (Alain and Bengio, 2017) is trained to predict linguistic properties from the representations of language (e.g. token embeddings); achieving high accuracy at this task implies these properties are encoded in the representations.

Experimental Setup We collected the 512 dimensional representations at the output of the first and last layer of the encoder and the decoder for all tokens except the French subject and associate each of them to a label indicating the occupational noun gender in the French sentence.

For each of these examples, we randomly split all sentences between a train (75%) and a test (25%) set. We use `scikit-learn` (Pedregosa et al., 2011) to learn a logistic regression to predict the occupational noun gender from a single token representation. This experiment is repeated on 10 random train/test splits and 95% confidence intervals are computed. As advocated by Hewitt and Liang (2019), we use a linear classifier to be sure that gender information is actually encoded in token representations and not learned by the probe and report, as a control score, the performance achieved after a random permutation of labels.

layer	decoder	
	the	other tokens
1	89.5 ±0.2	71.6 ±0.6
2	92.0 ±0.1	76.3 ±0.7
3	91.8 ±0.1	78.1 ±0.6
4	90.9 ±0.2	79.1 ±0.6
5	89.3 ±0.2	82.4 ±0.5
6	87.7 ±0.2	84.7 ±0.3

Table 4: Accuracy (in %) of probing the gender of the French occupational noun in decoder representations.

Results The probe achieves an average accuracy of 74.1% (resp. 87.9%) for the first (resp. last) layer of the encoder and of 80.5% and 86.2% for the decoder (detailed results are in Table 4 for the decoder and Table 5 for the encoder), showing that gender information is encoded in the representations of all source and target tokens. Note that, for the decoder, the diversity of the automatically generated structures makes it impossible to carry out a position-by-position analysis.

In the spirit of the analyses of Marvin and Linzen (2018) for monolingual representations, we have also manipulated source sentences to evaluate the robustness of our observations. We consider transformations that consist in:

1. weakening the gender expression in the subject by replacing `[DET]` (which can vary in gender) by *chaque* (each), which is epicene;

	encoder						random labels
layer	a	terminé	son	travail	.	eos	son
1	80.4 ± 1.1	75.1 ± 0.3	80.6 ± 0.3	76.4 ± 0.6	59.5 ± 1.0	73.3 ± 1.0	$45,3 \pm 0.9$
2	85.8 ± 1.0	80.8 ± 0.2	81.6 ± 0.3	78.3 ± 0.7	87.6 ± 0.6	88.3 ± 0.7	$50,7 \pm 0.8$
3	89.5 ± 0.6	88.2 ± 0.2	89.2 ± 0.2	82.0 ± 1.1	86.5 ± 1.0	87.6 ± 0.6	$48,8 \pm 0.9$
4	90.8 ± 0.4	89.3 ± 0.2	90.6 ± 0.2	85.9 ± 0.9	85.7 ± 1.0	85.6 ± 0.7	$48,6 \pm 0.8$
5	90.4 ± 1.0	89.3 ± 0.2	90.4 ± 0.2	85.5 ± 0.8	86.4 ± 0.8	85.2 ± 1.2	$49,6 \pm 0.8$
6	91.0 ± 0.6	89.3 ± 0.2	90.0 ± 0.2	86.0 ± 1.0	86.4 ± 1.1	85.1 ± 0.8	$49,2 \pm 0.8$

Table 5: Accuracy (in %) of a probe predicting the gender of the French subject given the encoder representations

2. strengthening the gender expression in the subject group by introducing an adjective that is always marked in gender (*français$_M$*/*française$_F$*, French);
3. replacing the direct object *travail$_M$* (work) by *activité$_F$* (activity) to evaluate the impact of the object noun phrase on the gender information encoded in the sentence;
4. increasing the distance between the subject group in which the gender is expressed and the possessive pronoun by inserting a relative clause containing no word marked in gender;
5. inserting a distractor (e.g. a word whose gender is different from the subject gender) between the subject, likely to introduce noise in the propagation of gender information.

Results in Table 3 show that the encoder is able to capture gender information even in convoluted contexts (e.g. presence of a distractor, insertion of an extraneous relative clause, etc.). In all these cases, accuracy remains much higher than chance for each input token, especially for the last layer of the encoder. Note, however, the strong effect obtained with weakening, where we observe a drop in accuracy of about 20 points. This corroborates the analysis of Table 2, where we see a drop in accuracy with an epicene determiner. A cumulative effect seems to be at play, whereby the encoding of gender is stronger with a double marking (on `[DET]` and `[N]`), and even more so when an additional unambiguous adjective also comes into play (the strengthening condition).

3.2 Translation Model as Conditional Language Model

Our probing experiments have confirmed that information about the gender of the nominal subject in the source was actually encoded in the source representations, and also available in the hidden states of target tokens. We now investigate whether this information is actually used: a well-known weakness of probes is that they can detect the presence of linguistic information in representations, but cannot measure how much it is used in the model predictions (Ravichander et al., 2021).

For this, we compare the predictions of a target language model (LM), that only knows about previous target tokens $t_{<i} = t_0, ..., t_{i-1}$; with the predictions of a translation model (TM), which additionally conditioned the output probabilities on the entire source sequence $\mathbf{s}$. The TM can be viewed as a *conditional language model* which computes $p(t_i|t_{<i}, \mathbf{s})$ where the LM computes $p(t_i|t_{<i})$. By comparing the predictions of these two models, we can evaluate the impact of information from the source. Other attempts at disentangling the influence of the source *vs.* the target context in NMT, using other methods and tools, are in (Ma et al., 2018; Fernandes et al., 2021; Voita et al., 2021).

Experimental Setting We compare the predictions of our NMT system (described in §2.2) with our in-house implementation of a TRANSFORMER language model with the same dimensions as the MT decoder using the PYTORCH library (Paszke et al., 2019).[4] To mimic the decoder, we use an autoregressive (causal) LM in which the representation of the i-th token is computed based on the $(i-1)$ previous tokens.[5] The model is trained by optimizing the cross-entropy with ADAM on the same corpus as for our TM (considering only the English side of the parallel corpus) and achieves a perplexity of 43.0 on the WMT'14 test set.

Method We investigate the ability of a TM or an LM to predict the correct form of the possessive pronoun in English sentences by comparing $p(\texttt{her}|c)$ and $p(\texttt{his}|c)$, where the context c is either the target prefix *The [occupational noun] has finished* for a LM or the target prefix and the

[4] Code and models are available at `https://github.com/neuroviz/neuroviz/tree/main/blackbox2022`

[5] An alternative to this experiment, more economical in terms of computational cost, would have been to consider only the predictions of a translation system in which all the words of the source sentence would be masked.

source sentence, for a TM. These probabilities can be easily computed with a forced decoding, e.g. by compelling the model to generate the reference English sentence up to the possessive pronoun. Comparing the probabilities rather than the predicted token allows us to conduct a more precise analysis, as we can include all sentences (rather than only the ones in which the sentence contains `his` or `her`) and also evaluate the confidence with which the model prefers one form to the other. This "contrastive" methodology has also been used in (Sennrich, 2017; Müller et al., 2018) to evaluate the quality of pronoun translation in MT.

Results Table 6 reports the average values for these probabilities.The LM that has only access to the prefix (and not to the source sentence) always generates `his` with a much higher probability than `her`. This is because the prefix context rarely contains information about the gender (Table 1). In this situation the model can only rely on associations between the target words and their frequency, and prefers to generate `his`, which is twice as frequent in the train set as `her`.

The TM probabilities of generating `his` or `her` are much higher than those estimated by the LM, showing that the TM actually uses source information, notably the presence of `son` in the French input, to increase these two probabilities. In almost all cases, nevertheless, the probability of generating `his` is reinforced more strongly than that of generating `her`, except when the determiner is la_F. Even in the latter case, where the context unambiguously marks feminine (cf. Table 6, the TM fails to give a clear preference for `her`). This result shows that the initial preference for a masculine pronoun is hard to be overturned, even in the presence of strong evidence that the feminine should be preferred.

Impact of tokenization Several factors may influence these observations and explain why the TM is able (or not) to predict the correct gender. The first factor is the tokenization of the occupational noun into lexical subword units. Recall that the number of subword units that a word is tokenized into is directly related to its frequency in the train set.

To assess the impact of tokenization, we report in Table 7 the probability that a TM generates `her` or `his` as a function of the number of subword units the occupational noun is broken into. Here we only consider the source sentences in which gender is expressed (either by the determiner, the occupational noun or both of them). These results show that the TM is more likely to increase the probability of generating `her` for a feminine source context when the occupational noun is sufficiently frequent to be kept as one single token. In all other cases, the increase in probability is not large enough to surpass that of the masculine form. This effect of subword splitting on gender prediction confirms the hypothesis of Savoldi et al. (2021) regarding the effect of morphological variation on gender bias.

Going one step further, we observe (Table 8) the 15 most frequent suffixes in feminine occupational nouns (e.g. the last token in their segmentation) in our test set. These suffixes appear to often correspond to feminine endings (*-ienne*, *-ière*, *-atrice*), with several of them only appearing in feminine nouns. Yet, except for two cases, `his` remains more likely than `her` showing that TM is not able to take advantage of the suffixes uncovered by the subword segmentation. We see here that using statistical subword segmentation yields morphologically inconsistent segmentations,[6] which has the effect of weakening gender information that could be learned from good predictors of the feminine form.

Replacing `SentencePiece` with a morphological segmentation would likely result in more consistent analyses and may have positive effects on gender transfer, at least for rare words. However, for the frequent words, which are not split into subwords, the effect could be somehow reversed. It is safe to say, still, that architectural decisions related to word segmentations have an impact on the transfer of gender across languages, at least for languages where gender is morphologically expressed like French. The impact of using a morphological *vs.* non-morphological segmentation in NMT is also documented in e.g. (Huck et al., 2017; Ataman et al., 2017; Banerjee and Bhattacharyya, 2018; Weller-Di Marco and Fraser, 2020).

3.3 Impact on training

The results reported in the previous Sections show that, for a TM, the target side priors for predicting `his` are much larger than for predicting `her`,

[6] Note, for instance, the competition between the two "suffixes" *-use* and *-euse*, the former corresponding to a "wrong" morphological segmentation of the latter.

Gender source sentence		TM		LM	
Determiner	Epicene noun	$p\,(\texttt{his}\|c)$	$p\,(\texttt{her}\|c)$	$p\,(\texttt{his}\|c)$	$p\,(\texttt{her}\|c)$
Feminine	no	0.278	0.216	0.158	0.022
	yes	0.213	0.248	0.098	0.022
Masculine	no	0.589	0.037	0.158	0.022
	yes	0.548	0.036	0.104	0.016
Epicene	no	0.588	0.055	0.182	0.021
	yes	0.485	0.074	0.109	0.016

Table 6: Average probabilities of his and her when conditioning on the source sentence (TM) or not (LM).

gender	# tokens	# occ.	$p\,(\texttt{his}\|c)$	$p\,(\texttt{her}\|c)$
Feminine	1	155	0.232	0.315
	2	601	0.251	0.238
	3	526	0.295	0.185
	≥ 4	279	0.289	0.176
Masculine	1	386	0.569	0.040
	2	529	0.547	0.035
	3	460	0.546	0.038
	≥ 4	200	0.544	0.039

Table 7: Probability of his or her estimated by a TM broken down by the number of tokens of the French occupational noun. Only French sentences in which the gender is marked are considered.

and that they get an additional boost when taking the source context into account. In comparison, the LM probabilities of her are always very small, and on average very similar to his for a feminine source context.

This unbalance is likely to have a negative impact during training. To see this, recall that the gradient of the training loss, the cross-entropy, is small when the system makes correct predictions with a high confidence. This is the case for the predictions of his, which also happens to be much more frequent in the data than her. As a result, the cumulated gradient flow that propagates to the encoder layers through the cross-attention module is not sufficiently strong for a system to correctly learn the dependency between the gendered words in the source and the target pronoun prediction.

To measure the strength of this effect, we perform the following experiment: after training the MT system, we consider all sentences with a feminine source subject in our test set and perform a single learning step (forward pass, cross-entropy computation and backward pass) and compute for each layer of the encoder and of the decoder the gradients accumulated during the backward pass. We carry out the same computation with sentences with a masculine subject. Note that we have ignored the 136 sentences having an epicene subject. Overall, there are as many words in sentences with a feminine subject as with a masculine subject.

We report in Table 9 the ratio between the norms of gradients computed on feminine and masculine examples. As predicted, gradients computed on feminine examples are on average larger than those computed on masculine examples, showing that the system errs more for the former cases. More interestingly, the difference is more significant for the encoder's parameters than for the decoder's: to correct the mispredictions for feminine examples, the training process attempts, as it were, to update the encoder parameters so as to better extract gender information from the source. When processing masculine sentences, the errors are less common and the parameter updates on the encoder side are comparatively smaller.

One conclusion of this experiment is that feminine and masculine examples do not have the same impact on the parameter estimation and the learning procedure fails to faithfully capture the dependency between source and target: only parameter updates for feminine occurrences that are often mispredicted, go a long way towards correcting the prior preference for the masculine pronoun.

4 Towards Mitigating Gender Bias

4.1 Increasing the cost of gender errors

In support of our analyses, we propose to slightly modify the loss function and to replace the cross-entropy, which penalizes mispredicted words according to the system confidence, by the softmax-margin loss (Gimpel and Smith, 2010):

$$-\,\text{logit}[y_{\text{gold}}]+ \log\left(\sum_{y\in\mathcal{V}} \exp\left(\text{logit}[y] + \text{cost}\left(y, y_{\text{gold}}\right)\right)\right) \qquad (1)$$

where $\text{logit}[y]$ is the score computed by the translation model (the logit) for token y, $\mathcal{V}$ is the system vocabulary and y_{gold} is the gold token that

	-iste	**-use**	**-euse**	-ologue	**-ière**	-e	**-atrice**	-graphe	**-ienne**	-ologiste	**-trice**	-liste	-niste	**-rice**
# occurrences in occupational noun														
masculine	168	0	0	76	0	18	0	32	0	27	0	18	17	0
feminine	167	208	199	76	129	110	76	30	54	25	46	18	17	30
$p(\texttt{his}\|c)$	**0.51**	**0.25**	**0.27**	**0.61**	**0.26**	**0.34**	0.28	**0.40**	**0.31**	**0.55**	0.26	**0.47**	**0.56**	0.30
$p(\texttt{her}\|c)$	0.12	0.19	0.21	0.08	0.15	0.18	**0.34**	0.08	0.18	0.11	**0.30**	0.07	0.07	0.27

Table 8: Most frequent sub-lexical suffixes in feminine occupational nouns and the probability, estimated by the TM, that the translation hypothesis contains `her` or `his`. Suffixes in bold only appear in feminine nouns.

	layer	$\frac{\nabla \text{param}_{\text{masc}}}{\nabla \text{param}_{\text{fem}}}$
decoder	0	0.719
	1	0.756
	2	0.758
	3	0.720
	4	0.780
	5	0.950
encoder	0	0.652
	1	0.649
	2	0.713
	3	0.661
	4	0.729
	5	0.770

Table 9: Ratio of the gradients norm for masculine and feminine sentences, in each of the encoder and decoder layer. Feminine sentences yield larger gradients, especially on the encoder side.

Gender in source		Accuracy	
Determiner	Epicene Noun		
Masculine	no	76.6	+0.1
	yes	84.5	-0.2
Feminine	no	35.3	+2.2
	yes	35.1	+3.2
Epicene	no	39.2	-1.2
	yes	41.1	+0.7

Table 10: Accuracy (in %) of possessive pronoun prediction when the TM is trained with the softmax-margin loss and difference with the accuracy achieved by system trained with cross-entropy loss.

should have been predicted. In comparison to the standard cross-entropy (in black), the loss function of Equation (1) includes an additional term (in cyan) that can been implemented using a $\mathcal{V} \times \mathcal{V}$ cost matrix cost $\left(y^{(i)}, y\right)$ where we specify how much the system should be penalized (in addition to the usual penalty) when predicting y instead of $y^{(i)}$. This loss function allows us to associate an extra penalty for each pair of (predicted word, gold word). In practice, in our experiments, we use a cost matrix where all elements except four are zero: the system receives an additional penalty when it predicts `its` instead of `her`, `its` instead of `his`, `his` instead of `her` and `her` instead of `his`. This means that for all other tokens, the system is trained as usual, meaning that the overall impact on translation quality remains circumscribed to these words.This penalty is fixed at 10% of the average value of the logits on the last layer of the decoder. Preliminary experiments with other values of the penalty or other pairs of tokens that are penalized did not improve these results.

By increasing the penalty incurred by mistakes in predicting the possessive pronoun, we are actually instructing the TM to be more careful when choosing them. This is mathematically expressed through reinforced gradients for these examples, which should ultimately result in parameters that are better at extracting and transferring gender information between the source and the target.

Experimental Results Table 10 reports the accuracy of pronoun prediction achieved by system trained with a softmax-margin loss. This system appears to be, on average, better at predicting the correct form of the possessive pronouns, especially for feminine source subjects. This observation confirms our conclusions: reinforcing gradients does result in better possessive pronoun predictions, illustrating again the fact that the cross-entropy loss fails to fully transfer gender information from the source, mainly because, as explained above, masculine pronoun can be correctly predicted without taking any source side information into account.

5 Conclusions

We have presented a new series of experimental evidence highlighting the causes of gender biases in NMT. Our analyses are based on observing pronoun translation errors in a controlled setting, using a new French-English test set of more than 3,000 occupational nouns. They mostly confirm the findings of previous studies that have amply documented the fact that errors were more likely to occur for feminine than for masculine pronouns, the latter being used as the default option in most context. Additional analyses based (a) on systematic probes, (b) on the comparison of LM *vs.* TM probabilities, (c) on subword splitting, (d) on the gradient flow in the encoder and decoder layers show that the cause of these biases are multifactorial. We have finally proposed a new way to mitigate these errors via a margin augmented training loss, specifically aimed at improving the information flow between source and target.

In our future work, we intend to continue exploring the potential of margin augmented losses, with the aim to also train the cost matrix, and to perform more systematic experiments with other systems and language pairs. Another line of investigation will be considering other linguistic phenomena posing difficult challenges for MT systems, such as the prediction of tense (Vanmassenhove et al., 2017) or mood information (Burchardt et al., 2017).

6 Ethics Statement

This supplemental description tries to follow (Larson, 2017) recommendations for gender as a variable in NLP. We cannot avoid gender as it is necessary to achieve our objectives, that is the study of gender biases. To make our theory of gender explicit, we follow Corbett's analysis of gender as a grammatical category (Corbett, 1991), defined for English as a pronominal gender system. For human referents in our corpus and in our experiments, we operationalize grammatical gender as a binary feature in French and English, which, on the one hand, can be felt as excluding, and, on the other hand, does not take into consideration more recent uses in these languages. As explained in lines 116–125, we resort to this simplistic representation of gender to highlight the gender bias in current NMT systems. For a more complete approach to the variety of gender-inclusive linguistic strategies currently in use in English, see for instance (Cao and Daumé III, 2020).

Acknowledgements

This work was partially funded by the NeuroViz project (Explorations and visualizations of a neural translation system), supported by the Région Ile-de-France within the DIM RFSI 2020 funding framework. This publication has emanated from research supported in part by the 2020 émergence research project SPECTRANS, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université Paris Cité). We gratefully acknowledge support from the CNRS/TGIR HUMA-NUM and IN2P3 Computing Center (Lyon - France) for providing computing and data-processing resources needed for this work.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331 – 342.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Ann Bodine. 1975. Androcentrism in prescriptive grammar: singular they, sex-indefinite he, and he or she. *Language in society*, 4(2):129–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Peter Jan-Thorsten, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Greville Corbett. 1991. *Gender*. Cambridge University Press.

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.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Anne Dister and Marie-Louise Moreau. 2014. *Mettre au féminin : guide de féminisation des noms de métier, fonction, grade ou titre*, 3e édition edition. Fédération Wallonie-Bruxelles.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Kevin Gimpel and Noah A. Smith. 2010. Softmax-margin CRFs: Training log-linear models with cost functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 733–736, Los Angeles, California. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. 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): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Xutai Ma, Ke Li, and Philipp Koehn. 2018. An analysis of source context dependency in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 189–197.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Xing Niu, Georgiana Dinu, Prashant Mathur, and Anna Currey. 2021. Faithful target attribute prediction in neural machine translation. *CoRR*, abs/2109.12105.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Anne Pauwels. 2000. Inclusive language is good business: gender, language and equality in the workplace. In Janet Holmes, editor, *Gendered Speech in Social Context: Perspectives from Gown and Town*, pages 134–151. Victoria University Press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830.

Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. 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. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Eva Vanmassenhove, Jinhua Du, and Andy Way. 2017. Investigating 'aspect' in NMT and SMT: translating the English simple past and present perfect. *Computational Linguistics in the Netherlands Journal (CLIN)*, 7:109–128.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Marion Weller-Di Marco and Alexander Fraser. 2020. Modeling word formation in English–German neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4227–4232. Association for Computational Linguistics.

Guillaume Wisniewski, Lichao Zhu, Nicolas Bailler, and François Yvon. 2021. Screening gender transfer in neural machine translation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 311–321, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.