Screening Gender Transfer in Neural Machine Translation

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, François Yvon

Screening Gender Transfer in Neural Machine Translation

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

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

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

Abstract

This paper aims at identifying the information flow in state-of-the-art machine translation systems, taking as example the transfer of gender when translating from French into English.

Using a controlled set of examples, we experiment several ways to investigate how gender information circulates in a encoder-decoder architecture considering both probing techniques as well as interventions on the internal representations used in the MT system. Our results show that gender information can be found in all token representations built by the encoder and the decoder and lead us to conclude that there are multiple pathways for gender transfer.

1 Introduction

The existence of translation divergences (i.e. cross-linguistic distinctions) raises many challenges for machine translation (MT) (Dorr, 1994): when translating a sentence, some information or constructions are specific to the target language and, consequently, can only be inferred by the decoder from the target context; some are only found in the source language and have to be ignored; finally, some information has to be adapted and transferred from the encoder to the decoder. Contrary to previous generations of MT engines where transfer rules were quite transparent, understanding this *information flow* within state-of-the-art neural MT systems is a challenging task, and a key step for their interpretability.

To illustrate these alternatives and the difficulty they raise, we focus in this work on one specific translation problem: the transfer of gender information from French, where grammatical gender is a property of all nouns, and agreement rules exist within the noun phrase, to English, where gender is only overtly used in rare constructs involving human agents and pronoun coreference.[1] More specifically, we focus on the English translation of French sentences such as "*L'actrice$_F$ a terminé son travail.*" (the actress$_F$ has finished her$_F$ job).[2] Translating this kind of sentences is problematic for state-of-the-art MT systems, notably because i) the coreference has to be correctly identified and ii) it can result in gender-biased translations due to stereotypical associations such as *nurses* are always female.

Using a controlled test set, we are able to screen the different information flows at stake when transferring gender information from French into English using two families of methods. The first one relies on linguistic probes to find in which parts of the NMT system gender information is represented; the second one is based on causal models and consists in intervening on the different parts of the source sentence and of the decoder representations in order to reveal their impact on the predicted translations. While this work focuses on one translation phenomenon and on one translation direction only, we believe that our observations shed a new light on how translation systems work and that the methods we describe can be used to analyze other translation divergences.

The rest of the paper is organized as follows. In section 2, we first introduce our controlled dataset and explain how gender is expressed in the two languages. Then, in Section 3 we describe our MT system and evaluate the system outputs in Section 4 its capacity to translate gender information. To explain these results, we then describe two sets of experiments that rely on different ways to analyze neural networks: in Section 5, we use linguistic probes to find out in which source and target tokens gender information is present and in Section 6 we report experiments modifying the token rep-

[1] We follow (Huddleston et al., 2002) presentation of gender in English, mostly grammaticalized in pronouns, as evidenced by *herself, himself, itself*.

[2] In all examples, we tag words either with F or M to indicate a feminine or masculine grammatical gender.

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 311–321

resentation to determine when this information is used. Finally, In Section 7, we relate our findings to previous research before summarizing our results in Section 8.

2 A Controlled Test Set to Study Gender Transfer between French and English

Corpus Creation Following (Saunders and Byrne, 2020),[3] we consider parallel sentences with the following pattern to study gender transfer between French and English:

(1) `[DET] [N]` a terminé son travail.

(2) The `[N]` has finished `[PRO]` work.

where `N` is a job noun that can be either masculine or feminine (e.g. in English, actor$_M$/actress$_F$; in French, *acteur*$_M$, *actrice*$_F$), `DET` is the French determiner in agreement with the noun (either the feminine form *la*$_F$, the masculine form *le*$_M$ or the epicene form *l'*[4]) and `PRO` is the English possessive pronoun '*her*' or '*his*'.

We use the complete list of professions and occupations for French from (Dister and Moreau, 2014) to fill the French `[N]` slot, and select the associated determiner accordingly. This list contains the feminine and masculine forms of each profession, allowing us to create a list of 3,394 sentences, perfectly balanced between genders.[5] Most of these occupational nouns are rare compound nouns unseen in the training corpus: as reported in Figure 1, only 1,707 of the 2,393 occupational nouns used to create the corpus can be found in the training set. This is also reflected in Figure 2, where we see that most occupational nouns are tokenized into multiple BPE units. These sentences were automatically translated and manually verified to produce the corresponding English list.

The motivations for using these fixed syntactic patterns are many. First, they limit the only source of variability between sentences to the `[N]` slots, allowing us to perform controlled experiments. Second, they simplify the analysis and manipulation

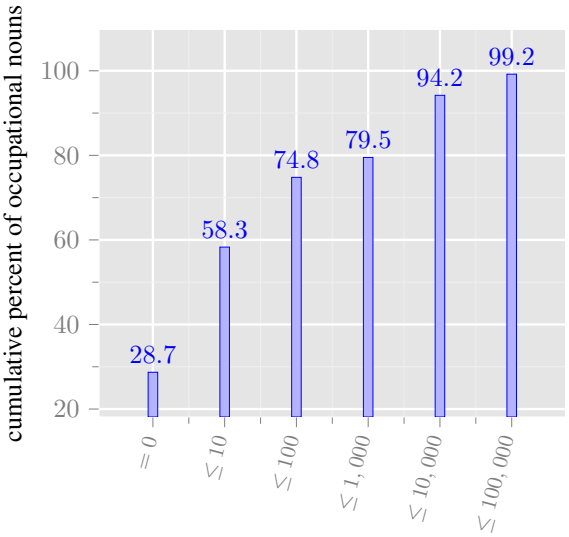

Figure 1: Cumulative frequency of occupational nouns in the training data.

of the trained representations, as the position of each word is almost constant throughout the entire dataset.[6] Despite its simplicity, this dataset enables rich analyses, as the large coverage of the set of nouns enables us to analyze the result with respect to the noun frequency, length, stereotypicality, and also with respect to the amount of gender information available in each language. Furthermore, the asymmetry between the gender carrying words in French and English will be a facilitating factor for generating interesting contrasts. In this paper we focus on French to English translations,[7] using variations in French determiners to generate interesting contrasts in the source. The possessive marker *son* is similarly epicene if the following word begins with a vowel. It should also be noted that there are sociolinguistic implications beyond the remit of this paper, such as a debated preference to refer to women's occupations favouring the masculine form of the occupational noun or exclusive uses of the masculine form to express generic uses (Brauer, 2008).

Expression of Gender When translating from French, four situations can occur:

(i) gender information can be inferred from both determiner `DET` and noun `N`, as in *le*$_M$

[3] Using a simplified list from (Prates et al., 2020), Saunders and Byrne (2020) created a "handcrafted" dataset of 388 parallel sentences of the type *The [PROFESSION] finished [his|her] work.* for three translation directions (English-Spanish, English-German and English-Hebrew). In this paper, we adapted this approach for a new translation direction (French to English) using a much larger list of occupational nouns: our corpus contains 3,394 sentences.

[4] The French determiner is *l'* for both genres if the job noun begins with a vowel.

[5] This dataset can be found at `https://github.com/neuroviz/neuroviz/tree/main/blackbox2021`.

[6] Notwithstanding small variations due to the BPE segmentation of the noun `N`, which can be split into one, two, three, and up to seven subword units.

[7] For a preliminary analysis of the English to French direction, see (Wisniewski et al., 2021).

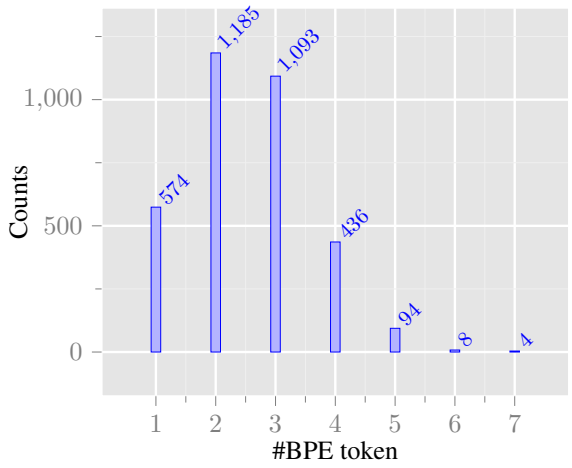

Figure 2: Distribution of the number of tokens in French occupational nouns after BPE tokenization.

determiner	job gender	case	#sentences
l'	fem.	(iii)	251
	epicene	(iv)	272
	masc.	(iii)	251
la	fem.	(i)	895
	epicene	(ii)	415
le	epicene	(ii)	417
	masc.	(i)	893

Table 1: Number of sentences for each way of expressing gender in French sentences.

couturier$_M$/*la*$_F$ *couturière*$_F$ — both translated by 'the stylist' or 'the seamstress' ;

(ii) gender information can be inferred from the determiner `DET` but not from the noun `N`, as in *le*$_M$ *cinéaste*/*la*$_F$ *cinéaste* — 'the film-maker' in both cases;

(iii) gender information can be inferred from the noun `N` but not from the determiner `DET`, as in *l'assistant*$_M$/*l'assistante*$_F$ — 'the assistant' in both cases;

(iv) gender information can not be inferred at all, as in *l'illusioniste* — 'the illusionist'.

Contrary to case (iv), in situations [i-iii], the translation system has the information to predict the right pronoun. Table 1 reports the number of sentences for each of these cases.

Conversely, in English sentences, gender information is always overtly expressed in the English pronoun, and in rare cases, also in the English noun `[N]`, as in the *actor/actress* pair of words.[8]

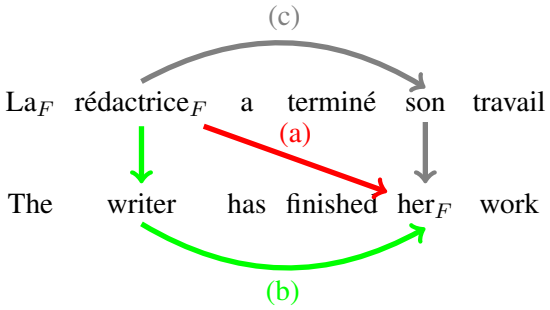

Figure 3: Gender transfer, from French to English: three possible influences on the choice of the gender of the possessive pronoun in English.

Pathways to Transfer Gender Information Looking at the example in Figure 3, we see that to correctly translate the gender of the French profession into English, three main hypotheses can be entertained:

- (a) a *direct influence* through the cross-lingual attention(s) computed when generating the English pronoun that should attend to the French noun;
- (b) an *indirect influence* through the (monolingual) encoding of gender in the representation of the English noun, the contextualized embedding of which should encode (through cross-lingual projection) the gender of the corresponding French NP;
- (c) an *indirect influence* through the (cross-lingual) attention to the French possessive adjective, the contextualized representation of which should then encode the French Noun gender.

Note that these three possibilities are not mutually exclusive, and gender may well be transferred through a combination of the three influences, and also through the representations computed for the other words in the sentence.

Our main objective in this paper is to explore various ways to assess these hypotheses and try to reach a conclusion regarding the way gender is actually transferred.

[8]Note that two English sentences (one with '*her*' the other with '*him*') were created when the gender of the profession can not be inferred from the French sentence. For instance, the French sentence '*l'artiste a terminé son travail*' appears twice in the parallel corpora: the first time as the translation of 'the artist has finished *her*$_F$ work', the second time as the translation of 'the artist has finished *his*$_M$ work'.

3 Experimental Setting

In all our experiments, we use JoeyNMT[9] (Kreutzer et al., 2019), an educational implementation of a translation system based on the Transformer model of Vaswani et al. (2017). The simplicity of the codebase, which nonetheless allowed us to achieve near SOTA performance on our data, made it a perfect choice for our endeavor. In our system, encoder and decoder are composed of 6 layers, each with 8 attention heads; the *feed-forward* layers have 2,048 parameters and the dimension of lexical embeddings is 512. Our model comprises a grand total of 76,596,736 parameters. The system was trained with data from the 'News' task of the WMT'2015 evaluation campaign.[10] It includes the Europarl, NewsCommentary and CommonCrawl corpora, and altogether contains 4,813,682 sentences and nearly 141 million French running words. All the corpora were tokenized and segmented into sub-lexical units using the unigram model of SentencePiece (Kudo, 2018); the resulting vocabularies contain 32,000 units in each language. The model is trained by optimizing the cross-entropy using the ADAM strategy. This system achieves a BLEU score of 34.0 for the French-English direction.

4 Evaluation of Gender Translation

4.1 Experimental Results

We evaluate the ability of our system to predict the gender of occupational nouns using the corpus described in Section 2 and consider, as a point of comparison, the translations generated by e-translation, a translation system developed by the European Commission that is freely accessible for academic research.[11] When translating into English, this evaluation is straightforward and simply amounts to checking the pronoun gender: does the translation hypothesis of a feminine (resp. masculine) occupational noun contains *her* (resp. *his*)? We therefore evaluate the two considered systems by the percentage of sentences for which the possessive pronoun is correct.

It should be noted that the gender information of a translation hypothesis can not always be determined: in some cases, the system produces a correct translation that does not contain *her* nor *his* (e.g. *the programmer has finished working*); in other cases, the translation is completely wrong or the determiner is translated as *its* (901 sentences mostly corresponding to situations in which the job noun was not translated correctly) or as *their* (52 sentences). For the sake of clarity, we do not distinguish these cases in our analyses.

It appears that our system is able to correctly predict the possessive pronoun in only 52.4% of the English sentences (the gender information could not be extracted in 1.4% of the sentences); on the contrary, e-translation achieves near perfect results: in 90.9% of the translation hypotheses, the gender of the pronoun is correct, which strongly suggests that this system integrates a specific process to transfer gender information.

Table 2 details these scores for the various situations identified in Section 2. These results show that our system (trained on 'standard' MT corpora) has a clear tendency to favor the translation of *son* by a masculine pronoun even in situations in which there is no ambiguity on the gender of the nominal group (e.g. when both the determiner and the noun both have a form specific to the feminine). Overall, our system achieves a precision of only 26.3% for the feminine pronouns, but correctly predicts the pronoun for 78.5% of masculine sentences. These observations are in line with the conclusions drawn by Saunders and Byrne (2020) on English-German, English-Spanish and English-Hebrew. Similar observations are also reported in (Renduchintala and Williams, 2021) when translating out of English for a larger set of target languages. On the contrary, e-translation is able to correctly infer the gender information in almost all cases and most of the errors are due to the French sentences in which the gender is not expressed (case (iv) in the description of Section 2).

4.2 Predicting Failure

We conducted two experiments to better understand the reasons why the gender of the possessive pronoun is not correctly predicted in the translations of JoeyNMT.

First, we looked at the number of times the possessive pronoun *son* was translated by *his* or *her* in the training data. For this purpose, we used eflomal (Östling and Tiedemann, 2016) to align French and English tokens of the training set[12] and

[9] https://github.com/joeynmt/joeynmt

[10] This is the most recent evaluation campaign for English-French organized in the context of the WMT conference (see http://statmt.org/wmt15).

[11] ec.europa.eu/cefdigital/eTranslation

[12] Alignment was performed after BPE tokenization and

Determiner	Job gender	predicted pronoun	JoeyNMT % sentences	JoeyNMT accuracy	e-translation % sentences	e-translation accuracy
l'	epicene	her	0.7%		4.4%	
		his	80.1%	40.4%	94.1%	49.3%
		other	19.2%		1.5%	
	fem.	**her**	7.2%		91.6%	
		his	59.4%	7.2%	4.0%	91.6%
		other	33.4%		4.4%	
	masc.	her	0.4%		0%	
		his	73.7%	73.7%	96%	96.0%
		other	25.9%		4.0%	
la	epicene	**her**	31.6%		93%	
		his	43.9%	31.6%	0.3%	93.0%
		other	24.5%		6.7%	
	fem.	**her**	33.3%		94%	
		his	18.5%	33.3%	0%	94.0%
		other	48.2%		6%	
le	epicene	her	0.7%		0%	
		his	84.4%	84.4%	95.4%	95.4%
		other	14.9%		4.6%	
	masc.	her	0.2%		0%	
		his	76.8%	76.8%	95.4%	95.6%
		other	21.2%		4.4%	

Table 2: Percentage of translation hypotheses that contain each possessive pronoun according to the way gender is expressed in the French subject. In each case the correct English pronoun is in bold.

translation	frequency
_its	27.94%
_his	**18.28%**
_the	7.24%
_her	**6.42%**
_a	3.34%
_their	2.92%
_it	2.45%
_sound	1.37%
s	1.33%
_he	0.76%
__OTHER__	22.13%

Table 3: Most frequent translation of the French token *son* according to the word alignment links. *son* is aligned with 3,658 different types. Those which do not appear in the table are grouped in the special token __OTHER__.

use the alignment link to find all possible translation of the French *son* token.[13] Results reported in Table 3 show that translations of *son* by *his* are three times more frequent than translations by *her*.

Second, we considered a simple logistic regression model that, given a sentence, predicts whether the pronoun gender will be correct or not. We used a small set of surface features to describe a French sentence: the gender of the determiner, the gender of the occupational noun (both can be either masculine, feminine or epicene), a binary feature that is true when both the determiner and the noun have an explicit gender marker, a feature describing the number of BPE units into which the occupational noun has been encoded and three Boolean features to describe the number of occurrences of the occupational noun in the train set. These features are respectively true when the occupational noun does not appear in the training set, when it occurs 10 times or less in the training set and when it occurs 100 times or more in the training set.

This model is trained on 75% of the examples and we evaluate the accuracy of its predictions on the remaining 25% of examples. To assess the stability of the model we consider 100 train-test splits and report the 95% confidence interval.

We report in Table 4 the accuracy achieved using all of these features as well as each of this feature individually.[14] Results show that, overall, the quality of the prediction is pretty high even when considering a single feature. This observation suggests the choice of the possessive pronoun in English is mostly based on surface information and does not result from a 'linguistic' analysis of the input sentence. In particular, the high precision achieved when considering solely the number of

symmetrized using the `grow-diag-final-and` heuristic.

[13]Note that, in French, *son* can either be a possessive pronoun or a noun meaning 'sound'.

[14]Regression models have been computed using `sklearn` (Pedregosa et al., 2011).

feature	accuracy
single features	
occupational noun gender	73.9% ± 0.2
determiner gender	74.5% ± 0.2
number of BPE tokens	74.5% ± 0.2
explicit gender	74.5% ± 0.2
occurrences in train set	74.5% ± 0.2
combination of features	
all gender features	81.5% ± 0.2
all features	82.2% ± 0.2

Table 4: Accuracy achieved when predicting whether the possessive pronoun will be correctly translated. Features are described in Section 4.2; 'all gender features' denotes the combination of three features: occupational noun gender, determiner gender and explicit gender.

BPE tokens, corroborated by the observation reported in Figure 4, shows that the system is not able to correctly predict gender information for occupational nouns that it did not see during training. As expected, the best feature to predict whether the gender of the English pronoun will be correct is the combination between the gender of the determiner and the gender of the job name, a feature that is closely related to the different ways gender is expressed in the French sentence as described in Section 2.

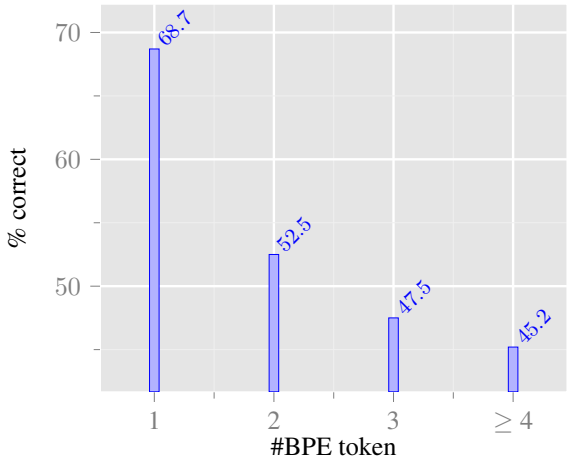

Figure 4: Sentences in which the English possessive pronoun is correctly translated according the number of tokens in the French occupational nouns after BPE tokenization.

	decoder	
layer	the	all 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 5: Precision of a probe predicting the gender of the French occupational noun given the decoder representation.

5 Probing Representations

In this section, we conduct an analysis of the representations computed by the encoder when translating from French into English. Our goal is to evaluate how well the gender information spreads through the transformer network from the initial French occupational noun (either `DET`, `N` or both of them) to the other French words, as well as to their English counterparts. Following a standard practice, we use probing (Belinkov and Glass, 2019) to analyze which words in the source and target sentence convey gender information: a *probe* (Alain and Bengio, 2017) is trained to predict linguistic properties (here the gender of the French subject) from the representations of language; achieving high accuracy at this task implies these properties were encoded in the representation.

Experimental setup We extract and collect the 512 dimensional hidden representations at the output of each layer of the encoder for all French lexical tokens following the job noun (i.e. *a*, *terminé*, *son*, *travail*, . and `<eos>`), as well as the first token (*the*)[15] of the English sentence in all decoder layers. All these words are frequent enough to correspond to one single BPE unit. We also consider a probe that is trained on all tokens of the target sentence (i.e. we collect the token representations of all translation hypotheses and associate each of them to a label indicating whether occupational name in the French sentence refers to a woman or a man), as the diversity of the translation structures makes it impossible to carry out a position-by-position analysis.

For each word, we randomly split our 3,394 sentences between train (75%) and test (25%), and use

[15] This is the only word that is always predicted correctly.

`scikit-learn` (Pedregosa et al., 2011) to learn a logistic regression model that predicts the gender of the occupational noun using the hidden representation of one single word. We use ℓ_1 penalty to regularize this model. The same data is also used to predict a random binary labeling: this is to control the capacity of our probing model (Hewitt and Liang, 2019). This experiment is repeated on 100 random train/test splits and 95% confidence intervals are computed.

Results Table 6 reports the accuracy achieved by our probes considering the representation of the source tokens as features. It appears that the representation of *son* (the translation of the possessive pronoun in French) is not the same when the occupational noun is masculine as when it is feminine: the representation of the French possessive pronoun encodes gender information even if the form of the word does not. It also appears that this information is more present in the deepest layers of the encoder: the probe achieves an accuracy of 80% when representations from the first layer are considered and of more than 90% when representations are extracted from any of the last three layers of the encoder. This observation confirms that there is actually an information flow between the possessive pronoun *son* and its antecedent the French occupational noun, corresponding to the path denoted (c) in Figure 3.

More surprisingly, accuracies achieved by the probe when the representations of other source tokens are considered are also very high: these accuracies are comparable or only slightly less than the ones achieved with *son*, showing that the gender information has an impact on the representations of all source tokens, even when these tokens have no direct syntactic relations with the subject phrase.

The results of the probe considering the decoder representations of '*the*' (Table 5) show a similar trend: the gender information is encoded in the representation even if the token generated by the decoder does not change with this information. It also appears that the probe is still able to predict the gender of the French occupational noun with a high accuracy when the representation of any token predicted by the decoder is considered as features, showing that, as for the encoder representations, gender information is encoded in all target tokens, even those for which this information is useless.

Results for predicting random labels (column 'random labels' in Table 6) finally show that the information is actually present in the representations and that the probe is not capturing spurious correlations in our data (Hewitt and Liang, 2019).

6 Manipulating Representations

The probing experiments described in the previous section show that gender information is encoded in all tokens representations built by the encoder and the decoder. However, it is not possible to identify from these observations if and when this information is used. To answer this question, the second method we propose to analyze gender transfer in our MT system relies on an *intervention*. It consists in replacing the embedding of the French possessive pronoun (i.e. the *son* token that, intuitively triggers the generation of '*her*' or '*his*') at the output of the encoder by either a *neutral* version of this embedding, obtained by averaging the representations of *son* on the whole test set (it should be borne in mind that the design of our corpus ensures that genders are balanced) or a *prototypical* version of a masculine *son* embedding or a feminine *son* embedding. These embeddings are extracted from the encoder representations of these two sentences:

(3) le facteur a terminé son travail.
the postman has finished his work.

(4) la pharmacienne a terminé son travail.
the pharmacist has finished her work.

These two sentences were chosen because, in both cases, gender information is carried by both the determiner and the noun and the translation of these sentences by our system is correct. After plugging the chosen representation in the last layer of the encoder and keeping the representation of the other tokens of the source sentence unchanged, the rest of the translation proceeds without any further modification.

Results of this manipulation are in Table 7: like in Table 2, we have reported the proportion of translation hypotheses in which the possessive pronoun is feminine, masculine or is neither *her* nor *his*.[16] Contrary to what was expected, changing the representation of the French possessive *son* has little impact (if any) on the choice of the English pronoun. These observations suggest that the representations of *son* built by the MT system are not the only evidence used during the generation of the translation hypothesis, even if the results reported in the

[16] As in Section 4, this last category includes both sentences that cannot be analyzed and those in which the possessive pronoun is *its* or *their*.

layer	encoder						random labels
	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% ±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% ±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% ±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% ±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% ±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% ±0.8

Table 6: Precision of a probe predicting the gender of the French subject given the encoder representations.

intervention	English pronoun	% sentences
none	her	13.4%
	his	57.1%
	other	29.5%
feminine	her	17.3%
	his	56.8%
	other	25.9%
gender-neutral	her	13.2%
	other	29.4%
	his	57.4%
masculine	her	13.8%
	other	29.2%
	his	57.0%

Table 7: Intervention on *son* representations: proportion of translation hypotheses in which the English possessive pronoun is *her*, *his* or neither of these two values, depending on the intervention on *son*.

previous section show that these representations are particularly relevant for making the correct prediction. This counter-intuitive result is consistent with several observations made in the literature: the fact that a 'linguistic' information is encoded in the neural representations does not imply that it will be used by the neural network (see, for instance, (Belinkov and Glass, 2019)). This suggests that the information flow along the path denoted (c) in Figure 3 should be small and the choice of the English possessive pronoun is based on other information than the representation of *son*.

7 Related Work

Our work is part of a very active line of research aiming to analyze, interpret, and evaluate neural networks used in NLP. Belinkov and Glass (2019) present a detailed overview of these papers and of the different tools and methods that can be used to uncover the linguistic information represented in the hidden layers of neural networks. Experiments reported in Section 5 are based on the probing approach of Alain and Bengio (2017) and have been used in many works (see (Belinkov and Glass, 2019) for an overview). This approach has also been used in several works to study the information flow within an encoder-decoder architecture: for instance, Belinkov et al. (2020) rely on probes to find which components of a NMT system encode linguistic information when translating morphologically rich languages. However, to the best of our knowledge, this work is the first to use the differences between gender expression in French and English to get insights into the inner representations used in NMT systems based on the Transformer architecture. Experiments reported in Section 6 are inspired by causal analysis, a type of analysis that has been used by Vig et al. (2020) to analyze gender bias in neural monolingual NLP models.

Several studies have investigated gender bias using dedicated datasets, some of them presented at the ACL Workshop on Gender Bias in Natural Language Processing (Costa-jussà et al., 2019, 2020b; Costa-jussa et al., 2021). Savoldi et al. (2021) synthesizes the studies and datasets on gender bias for translation. In particular, the controlled test set considered in our work builds on the works of Stanovsky et al. (2019) and Saunders and Byrne (2020), who both propose challenge test sets to evaluate gender bias in MT systems. The corresponding datasets consider the translation of occupational nouns with an anaphoric reference that makes gender explicit: the former contains instances of difficult translation patterns inspired by the WinoGender dataset of Rudinger et al. (2018); similar to our work, the latter contains a smaller set of simple sentences following a fixed template. Working with a slightly more varied set of sentence templates chosen to unambiguously express the gender of the occupational noun, (Renduchintala and Williams, 2021) also found that a generic multilingual system translating out of English made more errors for feminine than for masculine nouns, a trend that is observed in 20 languages.

Noting the limitations of artificial datasets, (Gonen and Webster, 2020) develop a methodology to

mine actual instances of likely biased translations in large corpora: these are found by automatically generating minimal contrasts in English source (e.g. replacing one noun by another) yielding a gender change in the target sentence. For instance, replacing 'doctor' by 'nurse' in the English might trigger a gender change in the corresponding translation in Russian.

Other studies also investigated the influence of socio-professional parameters such as profession types and the importance and the correlation with qualifying adjectives. Using the European multilingual classification of Skills, Competences and Occupations (ESCO) data, Marzi (2021) suggests insufficient biodiversity of the data in the training sets of neural translation systems. Focusing on 73 hypernyms from the 2,942 ESCO occupational nouns, she evidenced a gender gap by comparing the translations from Google Translate, DeepL and Microsoft Translator in the two directions for the French/Italian language pair. She built a dataset with respectively "competence" (i.e. *intelligent*) and "appearance" (i.e. *beautiful*) adjectives (ADJ) in the following pattern <A very `[ADJ] [N]` entered the room>. The data was manually analyzed. Adjectives seem to have no influence for the translation of masculine nouns, but competence adjectives affect the translation of feminine nouns more severely than appearance adjectives.

Zhao et al. (2018) studies gender bias in ELMo embeddings using probing techniques. In this study, biases in the embeddings also implied biases in a pronoun reference resolution task using the WinoGender dataset. Balancing data, and using averaged representations, to a certain extend, helped remove this bias.

Analyzing misclassified occupations in terms of gender, Costa-jussà et al. (2020a) investigated the architectural bias for the translation of occupational nouns, suggesting that using language-specific encoders and decoder yields less bias than a shared encoder-decoder architecture. Considering the attention patterns in the first two decoder layers, this paper shows that language-specific systems pay more attention to the determiner and occupational nouns, while bilingual models seem to rely more on the determiner. In the language-specific case, the embeddings are reported to encode more gender information.

Other architectural biases are considered by Renduchintala et al. (2021), who observed that gender bias is amplified when the system is optimised for speed. Using a dictionary of occupations for English to Spanish and English to German, they showed that correct translation rates degrade much faster than BLEU scores when limiting the beam-size to 1 during beam search or using low-bit quantization.

Finally, another line of research focuses on *mitigating* gender bias. This can be either achieved by working on the system's internal represention (Escudé Font and Costa-jussà, 2019), or by creating a more balanced training data where occupational roles are equally distributed between genders via counterfactual data augmentation (Hall Maudslay et al., 2019; Zmigrod et al., 2019). As discussed in (Saunders and Byrne, 2020), a cheaper, yet effective alternative to data augmentation, is to resort to domain adaptation techniques.

8 Discussion and Conclusion

Our paper investigated the different pathways for gender transfer. We created a dataset inspired by previous research to test several hypotheses. Our novel contribution is that we simultaneously mobilized several techniques, probing and manipulating. We extended the scope of the investigation of the locus of gender transfer beyond the determiner/noun analysis of Costa-jussà et al. (2020a) and questioned the role of predicates and epicene determiners for French. Our results show that gender information is present on the representation of all tokens built by the encoder and the decoder and suggest that the choice of the English possessive pronoun is distributed and is not based on the sole information contained in the representation of the French possessive pronoun. In our future research, we plan to identify how information is used to choose the form of the English pronoun and to generalize our observations to other languages and to other syntactic divergences.

Acknowledgements

This work was partially funded by the NeuroViz project (Explorations and visualizations of a neural translation system), supported by the Ile-de-France Region within the DIM RFSI 2020 funding framework.

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.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Markus Brauer. 2008. Un ministre peut-il tomber enceinte? L'impact du générique masculin sur les représentations mentales. *L'Année psychologique*, 108(2):243–272.

Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster, editors. 2021. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Online.

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

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

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

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.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

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.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. 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.

Rodney Huddleston, Geoffrey K Pullum, et al. 2002. *The Cambridge Grammar of English*. Cambridge University Press.

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.

Eleonora Marzi. 2021. La traduction automatique neuronale et les biais de genre: le cas des noms de métiers entre l'italien et le français. *Synergies Italie*, 17:19–36.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):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.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization 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 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2021. Investigating failures of automatic translation in the case of unambiguous gender.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 8–14. 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.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2021. Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire (gender bias in neural translation : a preliminary study). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 11–25, Lille, France. ATALA.

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.

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.