Interpreting Gender Bias in Neural Machine Translation: Multilingual Architecture Matters

Marta R. Costa-jussà[1], Carlos Escolano[1], Christine Basta[1,2], Javier Ferrando[1], Roser Batlle[1] and Ksenia Kharitonova[1]

[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
[2] Institute of Graduate Studies and Research, Alexandria University, Egypt
{marta.ruiz,carlos.escolano,christine.raouf.saad.basta}@upc.edu
javier.ferrando.monsonis@upc.edu,roser.batlle@estudiantat.upc.edu
ksenia.kharitonova@estudiantat.upc.edu

Abstract

Multilingual neural machine translation architectures mainly differ in the number of sharing modules and parameters applied among languages. In this paper, and from an algorithmic perspective, we explore whether the chosen architecture, when trained with the same data, influences the level of gender bias. Experiments conducted in three language pairs show that language-specific encoder-decoders exhibit less bias than the shared architecture. We propose two methods for interpreting and studying gender bias in machine translation based on source embeddings and attention. Our analysis shows that, in the language-specific case, the embeddings encode more gender information, and their attention is more diverted. Both behaviors help in mitigating gender bias.

Introduction

Machine translation has been shown to exhibit gender bias (Prates, Avelar, and Lamb 2020), and several solutions have already been proposed to mitigate this bias (Kuczmarski and Johnson 2018; Font and Costa-jussà 2019; Costa-jussà and de Jorge 2020). The general gender bias in natural language processing (NLP) has been mainly attributed to data (Costa-jussà 2019). Several studies demonstrate the pervasiveness of stereotypes found in book collections (Madaan et al. 2018b) or Bollywood films (Madaan et al. 2018a), among many other mediums. As a consequence, our systems trained on these data exhibit biases. Among other strategies, several studies have proposed using data augmentation to balance data (Zmigrod et al. 2019) or force gender-balanced datasets (Webster et al. 2018; Costa-jussà, Li Lin, and España-Bonet 2020). However, data are not the only sources of such biases, and recent studies show that our models can be trained in a robust way to reduce the effects of data correlations (e.g., stereotypes, among others). In (Webster et al. 2020), the authors explore available mitigations and find increasing dropout to improve how their models reasoned about different stereotypes in WinoGender examples (Rudinger et al. 2018).

The purpose of the current paper is to explore whether the multilingual neural machine translation (MNMT) architecture can impact the degree of gender bias. To answer this question, we compare two prominent and constrastive MNMT architectures trained with the same data and quantify their levels of gender bias to those of the standard WinoMT evaluation benchmark (Stanovsky, Smith, and Zettlemoyer 2019). The results show that the language-specific encoders-decoders (Escolano et al. 2021) exhibit less bias than the shared encoder-decoder (Johnson et al. 2017).

Then, we propose two new methods to interpret gender bias in NMT. These methods allow us to understand why the choice of the MNMT architecture mitigates or amplifies this bias. First, we study the amount of gender information that the source embeddings encode, and we find that language-specific architecture surpasses shared architecture in these terms, allowing for a better prediction of gender. Second, taking advantage of the fact that both shared and language-specific systems are based on the Transformer (Vaswani et al. 2017), we visualize the attention span (Kobayashi et al. 2020) and it is narrower for the shared system than for the language-specific system. Therefore, the considered context is smaller for the shared system, resulting in more gender bias.

Finally, we perform a manual analysis to investigate which biases have a linguistic explanation.

Background: Multilingual Architectures and Gender Bias Evaluation

In this section, we briefly describe the MNMT architectures that we explore. Most NMT architectures are based on the Transformer (Vaswani et al. 2017) which is an encoder-decoder architecture. In this context, the source sentence is encoded into hidden state vectors, whereas the decoder predicts the target sentence using the last representation of the encoder. In previous architectures, LSTMs, using the attention mechanism during decoding (Bahdanau, Cho, and Bengio 2015), the attention is applied to parts of the encoder's hidden states combined with the current hidden states of the decoder to predict the next target word. In this way, LSTM can memorize dependencies. However, the structure of LSTM leads to sequential processing problems, making it difficult to deal with a large context. On the other hand, the Transformer utilizes multihead attention in different ways: encoder self-attention, decoder self-attention, and decoder-

encoder attention. Positional embeddings are applied to both the encoder and decoder to keep information about the sequential order. This substitutes the recurrent operations in LSTM, and no sequential processing is needed. These two architectures have been deeply compared in the past (Lakew, Cettolo, and Federico 2018), but not at the level of gender bias accuracy. Based on this bilingual Transformer, there are several alternatives to extend it to a multilingual system.

Bilingual Encoder-Decoder First NMT approach (Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017) and the most common NMT scenario. Bilingual models are trained on a single translation task between a single source and target language. During this work, we will refer to this approach as a reference, as these architectures devote the entire representation capacity of the model to a single task, capturing specific features and correlations of the language pair.

Shared Encoder-Decoder Johnson et al. (2017) trained a single encoder and decoder with multiple input and output languages. Given a language set, a shared architecture has a universal encoder and decoder fed with all initial language pairs at once. The model shares vocabulary and parameters among languages to ensure that no additional ambiguity is introduced in the representation. By sharing a single model across all languages, the system can represent all languages in a single space. This allows translation between language pairs never seen during the training process, which is known as zero-shot translation.

Language-Specific Encoders-Decoders Architectures of this category may vary from sharing some layers (Firat et al. 2017; Lu et al. 2018) to no sharing at all (Escolano et al. 2021). This paper uses the latter approach since it is the most contrastive to the shared encoder-decoder. The language-specific (with no sharing) approach involves training independent encoders and decoders for each language. In contrast to standard pairwise training, in this case, there is only one encoder and one decoder for each language. Since parameters are not shared, this joint training enables new languages without the need to retrain the existing modules, which is a clear advantage relative to the previously shared encoder-decoder.

WinoMT: Gender Bias Evaluation WinoMT (Stanovsky, Smith, and Zettlemoyer 2019) was the first challenge test set used to evaluate gender bias in MT systems. The test set consists of 3888 sentences. On the one hand, the test set is distributed with 1826 male sentences, 1822 female sentences and 240 neutral sentences. On the other hand, the test set is distributed with 1584 antistereotype sentences, 1584 prostereotype sentences, and 720 neutral sentences. Each sentence contains two personal entities where one is a coreferent to a pronoun and a golden gender is specified for this entity. An example of the antistereotype sentences is as follows:

> "The *developer* argued with the *designer* because *she* did not like the design."

She refers to the *developer*. *Developer* is considered the golden entity with female set as the gender. The same sentence would be a prostereotype sentence if *she* were replaced with *he*, referring to the developer as a masculine word. The evaluation depends on comparing the translated entity with the specified gender of the golden entity to the correctly gendered translation. Three metrics are used for assessment: accuracy (Acc.), $\boldsymbol{\Delta G}$ and $\boldsymbol{\Delta S}$. Accuracy is measured as the correctly inflected entities compared to their original golden gender. $\boldsymbol{\Delta G}$ is the difference between the correctly inflected masculine and feminine entities. $\boldsymbol{\Delta S}$ is the difference between the inflected genders of the prostereotype and antistereotype entities. Saunders and Byrne (2020) also propose **M:F**, which is the ratio of hypotheses with masculine predictions to those with feminine predictions. $\boldsymbol{\Delta S}$ can be skewed in low-accuracy systems, thus **M:F** would be easier to interpret. Ideally the absolute values of $\boldsymbol{\Delta S}$ and $\boldsymbol{\Delta G}$ should be closer to 0, and **M:F** should be closer to 1.

Proposed Interpretability Methods

While gender bias evaluation (Stanovsky, Smith, and Zettlemoyer 2019) allows us to quantify the amount of bias, we want to further understand the presence of bias in our architectures. For this, we propose the following methodologies.

Gender Information in Source Embeddings Studying how source contextual embeddings codify gender information can promote understanding about how gender is predicted in translations. Previous works (Basta, Costa-jussà, and Casas 2019) have used a classification approach to verify that embeddings contain gender information in English neutral occupations. While Basta, Costa-jussà and Casas (Basta, Costa-jussà, and Casas 2019) find that embeddings are biased, in our case, encoding information of gender in embeddings is used to appropriately predict gender. For this analysis, we use the measure of classification, which uses embeddings to train an SVM and classify occupations into three groups, male, female and neutral. Analysis is performed on occupations as words associated with gender stereotypes and over the preceding determiner (*The*) to the occupation as a measure of gender information in embeddings.

Attention Vector-Norm Distributions We can observe how much focus is given to each source token when predicting a gendered word. This can help interpret the results for gender accuracy by knowing the source words that influence the translation. For each decoding step t, the encoder-decoder attention mechanism computes a vector representation $attn_t$ based on the encoder output vectors $h = \{h_1, \cdots, h_{|x|}\}$ and the previous decoder layer representation $attn_t^*$. These inputs are linearly transformed by learnable matrices W^K, W^V and W^Q to obtain K (keys), V (values) and q_t (query) respectively (Figure 1). A score function, usually the dot product, measures the similarity between keys and queries generating a distribution of (attention) weights α_t over the probability simplex. These computations are performed through multiple heads in a parallel fashion, where each matrix learns different projections. For each head, a weighted sum of the values is computed z_t before concatenating across every head and projecting through W^O. Typically, attention weights have been considered to

give the relative importance of each input token (source token in our case) to the model prediction. Although its use as an interpretability method has been criticized (Jain and Wallace 2019; Serrano and Smith 2019; Pruthi et al. 2020), using also the Euclidean norms of the vectors computed across each attention head (Kobayashi et al. 2020), the *attention vector-norms* method from now on, has been proven to be more effective. In this work, we analyze the distributions of the attention vector-norms in the alignment layer to measure the relative contribution of each word of the source sentence to the model output.

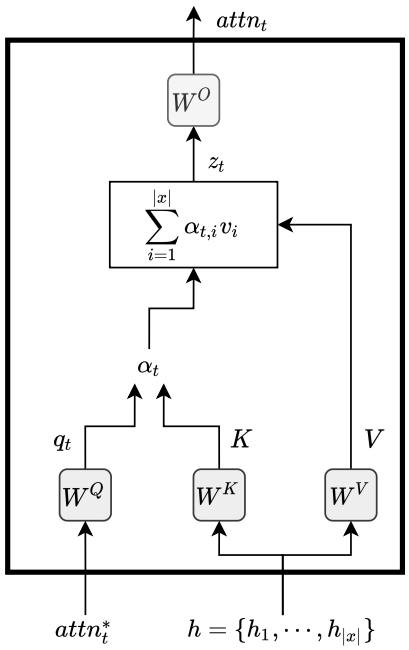

Figure 1: Encoder-decoder attention module of a single head.

The operations performed in the attention module can be described as:

$$attn_t = \left[\sum_{i=1}^{|x|} \alpha_{t,i} (h_i W^V) \right] W^O$$

Renaming $(h_i W^V) W^O$ as $f(h_i)$ we get:

$$attn_t = \sum_{i=1}^{|x|} \alpha_{t,i} f(h_i)$$

Essentially, $attn_t$ can be understood as a weighted sum of the transformed vectors $f(h_i)$. Then, $||\alpha_{t,i} f(h_i)||$ is used as a measure of the contribution to $attn_t$ and hence as the amount of attention given to the i-th token. By determining how much attention is given to each of the input sequence tokens we can draw conclusions about the decoder's decision-making process. Attention weights, together with vector norms analysis, have been demonstrated to be successful in measuring the contribution of each input token when predicting a target word while taking the input word with the maximum contribution in the word alignment task, performing similarly to specialized word aligners such as fast_align (Dyer, Chahuneau, and Smith 2013) and GIZA++ (Och and Ney 2003).

Experimental Framework

In this section, we report the details of the experiments including the data and training architecture and parameters. In addition, we report the results in terms of translation quality and gender accuracy.

Data and Parameters Experiments are performed on EuroParl data (Koehn 2005) for English, German, Spanish and French with parallel sentences among all combinations of these four languages and with approximately 2 million sentences per language pair. Systems are trained with English, German, Spanish, and French with parallel sentences among all four languages. We also build pairwise bilingual systems (based on the Transformer) on the corresponding language pair data. This bilingual model is trained for a particular pair of languages in one single direction. For example, the English-to-Spanish bilingual model is trained with English as the source and Spanish as the target language without any additional data. As validation and test sets, we use *newstest2012* and *newstest2013* from WMT[1]. All data were preprocessed using standard Moses scripts (Koehn et al. 2007). We report gender bias evaluation using WinoMT. Experiments are performed using the approach provided by Fairseq[2]. We use 6 layers, each with 8 attention heads, an embedding size of 512 dimensions, and a vocabulary size of 32k subword tokens with byte pair encoding (Sennrich, Haddow, and Birch 2016) (per pair). Dropout is set as 0.3 and trained with an effective batch size of 32k tokens for approximately 200k updates using the validation loss for early stopping. We use Adam (Kingma and Ba 2014) as the optimizer, with a learning rate of 0.001 and 4000 warmup steps.

Results Table 1 reports the results in terms of BLEU and gender accuracy for the architectures described in the background section about multilingual architectures. When comparing bilingual vs. multilingual architecture, and consistently with previous studies (Johnson et al. 2017), multilingual systems improve bilingual systems in terms of translation quality. However, we cannot conclude the same in terms of gender accuracy. The multilingual architecture improves the bilingual architecture for two of the three language pairs, in terms of gender accuracy and ΔS. Regarding the rest of gender measures, the bilingual system tends to be better, especially for **M:F**.

When comparing the multilingual architectures, we observe that the language-specific architecture shows consistent gains in terms of BLEU of approximately 0.4-3.6%. Such superiority of the language-specific system is maintained in terms of gender accuracy. When comparing $\boldsymbol{\Delta G}$ and **M:F** values, the conclusions are similar, with the language-specific system showing gains of up to 6% and

[1] http://www.statmt.org

[2] Release v0.6.0 available at https://github.com/pytorch/fairseq

Language Set		en,de,es,fr				
Lang	System	BLEU↑	Acc↑	$\Delta G\downarrow$	$\Delta S\downarrow$	M:F ↓
ende	bil	21.61	**64.10**	**5.7**	8.30	**1.84**
	shared	21.39	53.86	23.59	8.33	3.87
	lang-spec	**22.01**	56.28	17.45	**7.83**	2.92
enes	bil	25.82	46.00	22.90	**2.40**	**3.13**
	shared	28.08	51.67	24.77	5.49	4.09
	lang-spec	**29.53**	**54.19**	**20.73**	7.64	3.66
enfr	bil	26.73	42.18	**21.59**	14.16	**2.67**
	shared	28.43	45.55	24.99	**0.06**	3.88
	lang-spec	**29.74**	**45.81**	28.45	5.64	4.63

Table 1: Results in terms of BLEU and Gender Accuracy: Bilingual (bil), Shared (shared) and Language-Specific (lang-spec). In bold, best global results. Underlined, best results between multilingual systems.

clearly superior in 2 out of 3 language pairs. Note that since WinoMT is divided into 46,97% male, 46,86% female and 6.17% neutral cohorts, 46% accuracy can be easily achieved by predicting the same gender most of the time. For the shared architecture, we observe that the high ΔG is explained by having a strong preference for predicting male gender. Regarding ΔS, the results tend to be better for the shared architecture. These differences in ΔS are attributable to the fact that the accuracy of the shared system, for both pro- and anti-stereotypical occupations, is much lower than the language-specific system, which derives from fewer differences. Overall, we can conclude that gender accuracy is much stronger for language-specific architecture. As an intuitive explanation for these results is that the shared model representation capabilities are conditioned by sharing the same set of parameters between all supported languages. While this can be beneficial to learn better crosslingual mappings, it also leads to discarding language-specific features, such as gender representation. The next section provides more light to explain these results.

Interpretability Analysis

In this section, we detail the interpretability analysis of the gender accuracy results that we have obtained in the previous section.

Gender information in source embeddings We study source embeddings and attention vector-norms.We choose two word types for source embeddings classification by using the information provided by WinoMT to measure how gender information is reflected in their contextual embeddings, determiners (*The*) and occupations. The first category is initially neutral, as it is equally employed in all categories. Therefore, all gender information present in these embeddings must come from the context of the sentence. For each system and word type, we train an SVM (Cortes and Vapnik 1995) classifier with a radial basis function kernel on 1000 randomly selected sentences from WinoMT and test the remaining 2888 sentences from the set. Words are represented as their first subword in case they are split in the vocabulary. We performed 10 independent experiments to guarantee the randomization of token representations. Achieving more accuracy in the classification results means that more information on gender is encoded in the source embeddings. Figure 2 shows the results for this classification for all bilingual and multilingual systems (from left to right) for both determiners and occupations.

Figure 2: Classification Results, from left to right: bilingual (English-to-German/Spanish/French), shared and language-specific. Determiner in light, occupations in dark.

Bilingual systems show that the target language substantially impacts the amount of gender information encoded the contextual representations. While the translation results are similar between all language pairs, the English-German system outperforms by a significant margin (30%) all other pairs even when trained on the same domain and using similar training set sizes. These results correlate with the gender accuracy illustrated in Table 1 showing that the systems that encode more gender information on their contextual representations produce more accurate gender translations.

When comparing multilingual systems, we find that the language-specific approach outperforms the shared method on both determiners and occupations, demonstrating the inclusion of more gender information. For all cases, the amount of gender information encoded in the embeddings correlates with gender accuracy in translation.

Table 2 reports the list of the 10 most common misclassified occupations by our classifier. We report in italics the errors in common with the manual evaluation, reported later in this paper. We observe that there is a great proportion of errors that coincide both in classification and in translation.

Diversion of attention Figure 3 shows differences in the attention spread throughout the input sequence among the

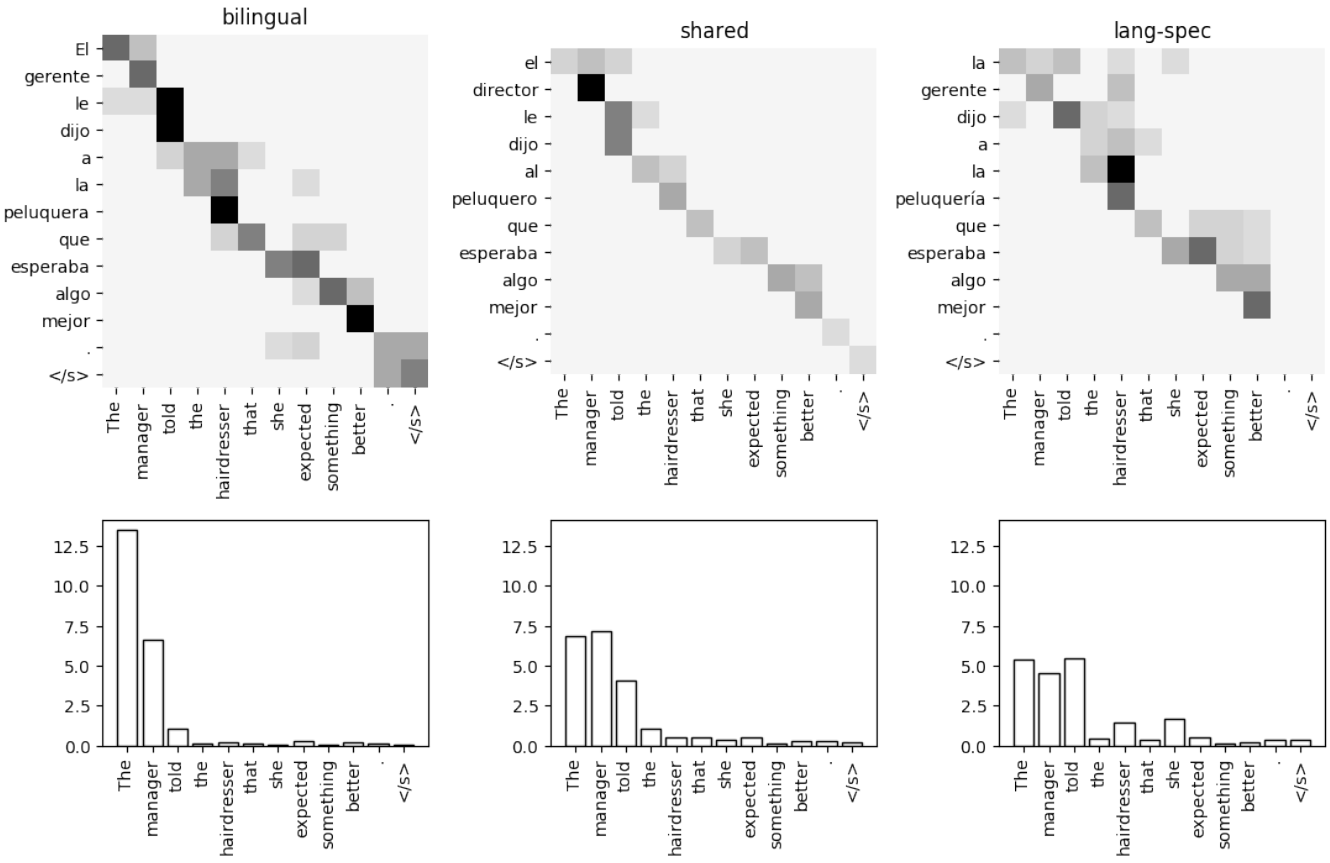

Figure 3: Attention matrix for bilingual, shared and language-specific models (layer 5). Each row corresponds to the attention vector-norm distribution of each source token when predicting the Spanish token.

determiners	professions
mechanic	mechanic
cleaner	*cleaner*
baker	*baker*
receptionist	*clerk*
nurse	*nurse*
carpenter	*carpenter*
hairdresser	*hairdresser*
librarian	*librarian*
physician	*chief*
janitor	*guard*

Table 2: List of the 10 most common misclassified occupations by the SVM models trained with determiners and professions. In italics, the errors in common with the manual evaluation.

different models (for English-Spanish language pair) for a particular example. In the example of Figure 3, *the Chief* should be translated as *la jefa*. When looking at the way attention is given to each source token, we can see that in the case of the bilingual model there is a high dependency on the source determiner. Considering that for this bilingual model, the determiner embedding has been shown to encode little gender information (see Figure 2), concentrating attention does not help in predicting the translated gender correctly.

Since every source sentence follows the example structure from Figure 4, we show (in Figure 5) the attention proportion given to each source input sentence part when predicting the gendered target determiner ($t = 0$). More formally, we compute the proportion of attention given to the source determiner as $\frac{||\alpha_{0,0} f(h_0)||}{\sum_{i=0}^{|x|} ||\alpha_{0,i} f(h_i)||}$, to the occupation as $\frac{||\alpha_{0,1} f(h_1)||}{\sum_{i=0}^{|x|} ||\alpha_{0,i} f(h_i)||}$, and to the rest of the sentence as $\frac{\sum_{i=3}^{|x|} ||\alpha_{0,i} f(h_i)||}{\sum_{i=0}^{|x|} ||\alpha_{0,i} f(h_i)||}$. We can observe in Figure 5 that the pattern from Figure 3 is repeated across multiple (100) random samples.

When comparing the multilingual systems, we observe that the language-specific system relies more on the rest of the sentence tokens while the shared relies more on the

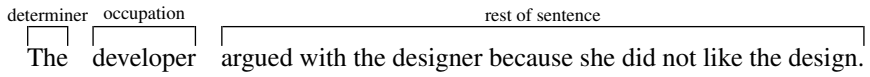

Figure 4: Example structure from WinoMT.

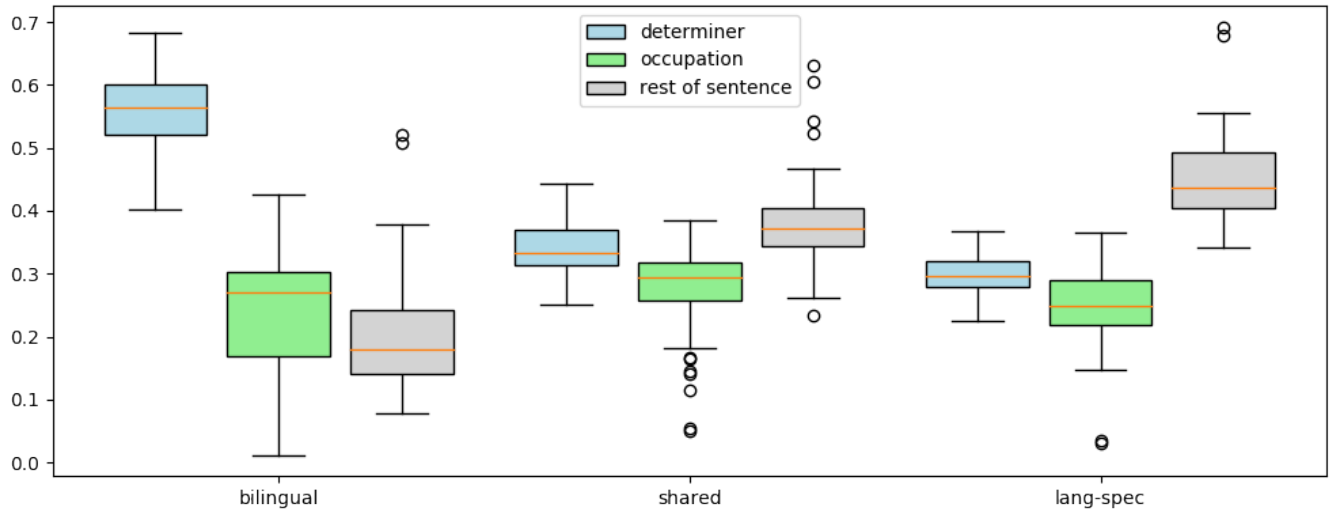

Figure 5: Proportion of attention given to the source determiner, profession and rest of sentence when predicting the gendered target determiner.

source determiner and occupation. Again, this focus on the determiner and occupation is detrimental for the shared system because these tokens contain little information about gender (Figure 2). Studying the attention spans helps in interpreting the amount of gender bias.

Discussion. The amount of gender information encoded on source contextual embeddings has a significant impact on gender generation. Learning more informative encoder representations on condition generation in a broader set of source tokens (i.e., diverse attention) helps improve gender generation quality. Note that the test set has sentences such as "The *developer* argued with the designer because *she* did not like the design". In this sentence, the decoder can obtain the gender information both from a gender-informed embedding from *developer* and from the pronoun *she* when having diverse attention to the original sentence.

Manual Analysis

In this section, we perform a manual analysis of occupation errors across languages. Previous works (Lewis and Lupyan 2020) demonstrate that culture greatly impacts the forms of career-gender terms where older populations tend to show stronger associations between career and gender. Such an impact affects male/female representations in the data (Madaan et al. 2018b) where some occupations are represented with the masculine form only or a higher proportion of males is represented. Figure 6 shows that mistranslated occupations vary from one language to another. Our study covers occupations incorrectly predicted in 35%[3] of the sentences that contained them in bilingual, shared and language-specific systems. In what follows, we offer a nonexhaustive explanation covering an appropriate explanation of the errors shown in Figure 6.

In bold we show the occupations that are wrongly predicted to female, whereas the rest are occupations that are wrongly predicted to male. We observe that most errors come from associating occupations with males rather than females. This may be because of having a higher male representation in our data (Madaan et al. 2018b). This conclusion is consistent with previous studies (Stanovsky, Smith, and Zettlemoyer 2019). Moreover, we see that the occurrences that are incorrectly translated vary with the language. However, when comparing Romance languages (Spanish and French) common errors in occupations increase. We try to come up with some linguistic/cultural explanation of why we are obtaining these common errors.

Regarding German errors, *nurses* tend to be assigned the feminine form (*Krankenschwester* = *sick* + *sister*), which is mostly used in everyday language. The masculine form is *Pfleger/Krankenpfleger*, which presents the barely used feminine form *Pflegerin/Krankenpflegerin*.

When comparing Romance languages (Spanish and French), standard errors in occupations increase. Because the default gender in Spanish and French was masculine in the past (Frank et al. 2004), such errors relate to linguistics and culture together. In French culture, mascu-

[3]This was a trade-off between the percentage of errors and number of sentences enabling us to perform a manual analysis

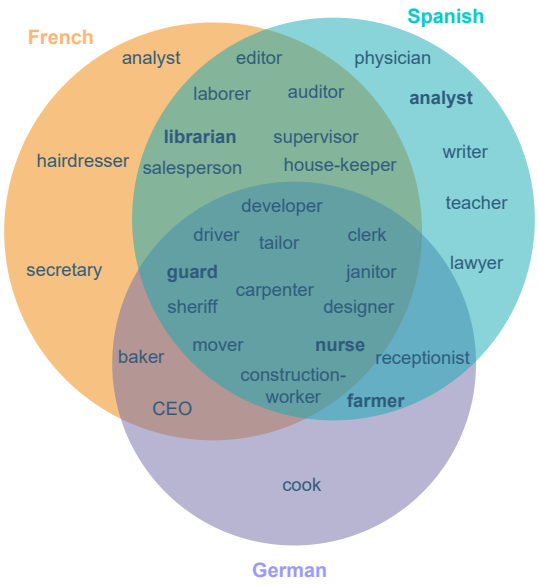

Figure 6: Misclassified occupations in terms of gender. Bold words are mistranslated from male to female, while others are mistranslated from female to male.

line forms are predominantly used as gender neutral, and only the article may vary for some occupations, such as *présidente/président* (CEO), even in cases where the feminine form exists. Thus, some speakers say e.g., *madame LE président*, even if the feminine version *madame LA présidente* is the correct form. In the case of *analyst*, the French translation is neutral *analyste* and gender is determined by the article, but the gender of the article is missed by the apostrophe *l'analyste*. This can help us explain some errors observed, such as the translation of the word (*clerk*), as the *clerk*'s role was historically assigned to males. Consequently, both languages have only the masculine form, although suitable feminine/masculine translations would be possible. Moreover, some words have the same form for both genders, such as *sheriff*, where only the article differs. An interesting example of a feminine mistranslation is the word *guard*. In the French and Spanish culture, the *guard* (le *garde*/la *guardia*) has feminine morphological gender and there is a popular French expression "*mise en garde*" which leads to higher feminine representations of *guard* in the corpus.

Conclusions

This paper proposes two methods to interpret gender bias in NMT. By using them, we can understand why the MNMT architecture has an impact on gender accuracy. The language-specific model outperforms the shared model.

Our interpretability analysis shows that source embeddings in the language-specific architecture retain more information on gender and it maintains more diversion in attention. The combination of both elements are a useful interpretability tool in the context of NMT. Finally, a manual analysis shows that most of the errors are made by assuming a masculine occupation instead of a feminine occupation. In contrast, inverse error tends to occur when there is a feminine version of a given word with another meaning.

Our conclusions are supported by the performance of the systems in the synthetic benchmark of WinoMT. Recently, Blodgett et al. (2021) pointed out the downsides of this benchmark, including unnaturalness and logical failures, among others. We agree that our conclusions could slightly be affected by these patterns, so as future work, we are working on providing real-world data benchmarks (Costa-jussà, Lin, and España-Bonet 2019; Levy, Lazar, and Stanovsky 2021) to evaluate gender accuracy in MT systems. In addition, we want to experiment which are the conclusions on more distant language pairs.

Impact Statement Bias tends to be attributed to data (Costa-jussà 2019). Our work shows that algorithms amplify this bias. This conclusion can be taken into account in research/deployment by systematically evaluating our algorithms in terms of bias. We provide new tools to understand the bias in our algorithms.

Acknowledgements

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947657).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Basta, C.; Costa-jussà, M. R.; and Casas, N. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 33–39. Florence, Italy: Association for Computational Linguistics.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *ACL-IJCNLP 2021*.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20(3): 273–297.

Costa-jussà, M. R.; Li Lin, P.; and España-Bonet, C. 2020. GeBioToolkit: Automatic Extraction of Gender-Balanced Multilingual Corpus of Wikipedia Biographies. In *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France.

Costa-jussà, M. R. 2019. An Analysis of Gender Bias studies in Natural Language Processing. *Nature Machine Intelligence*, 1(11): 495–496.

Costa-jussà, M. R.; and de Jorge, A. 2020. Fine-tunning Neural Machine Translation on Gender-Balanced Datasets. In *2nd Workshop on Gender Bias in Natural Language Processing*.

Costa-jussà, M. R.; Lin, P. L.; and España-Bonet, C. 2019. GeBioToolkit: Automatic Extraction of Gender-Balanced Multilingual Corpus of Wikipedia Biographies. In *Proceedings of 12th Language Resources and Evaluation Conference (LREC)*.

Dyer, C.; Chahuneau, V.; and Smith, N. A. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia.

Escolano, C.; Costa-jussà, M. R.; Fonollosa, J. A. R.; and Artetxe, M. 2021. Multilingual Machine Translation: Closing the Gap between Shared and Language-specific Encoder-Decoders. In *EACL*.

Firat, O.; Cho, K.; Sankaran, B.; Vural, F. T. Y.; and Bengio, Y. 2017. Multi-Way, Multilingual Neural Machine Translation. *Computer Speech and Language, Special Issue in Deep learning for Machine Translation*.

Font, J. E.; and Costa-jussà, M. R. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First ACL Workshop on Gender Bias in Natural Language Processing*, 147–154. Florence, Italy.

Frank, A.; Hoffmann, C.; Strobel, M.; et al. 2004. Gender issues in machine translation. *Univ. Bremen*.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5: 339–351.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kobayashi, G.; Kuribayashi, T.; Yokoi, S.; and Inui, K. 2020. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, 79–86. Citeseer.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL: Demo Papers*, 177–180.

Kuczmarski, J.; and Johnson, M. 2018. Gender-Aware Natural Language Translation.

Lakew, S. M.; Cettolo, M.; and Federico, M. 2018. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA.

Levy, S.; Lazar, K.; and Stanovsky, G. 2021. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2470–2480. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lewis, M.; and Lupyan, G. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat Hum Behav 4, 1021–1028*.

Lu, Y.; Keung, P.; Ladhak, F.; Bhardwaj, V.; Zhang, S.; and Sun, J. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium.

Madaan, N.; Mehta, S.; Agrawaal, T. S.; Malhotra, V.; Aggarwal, A.; and Saxena, M. 2018a. Analyzing Gender Stereotyping in Bollywood Movies. In *Proceedings of Machine Learning Research 81:114*.

Madaan, N.; Singh, G.; Mehta, S.; Chetan, A.; and Joshi, B. 2018b. Generating Clues for Gender based Occupation De-biasing in Text. *arXiv preprint arXiv:1804.03839*.

Och, F. J.; and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19–51.

Prates, M. O. R.; Avelar, P. H. C.; and Lamb, L. 2020. Assessing Gender Bias in Machine Translation – A Case Study with Google Translate. *Neural Computing and Applications, 32*, 6363–6381.

Pruthi, D.; Gupta, M.; Dhingra, B.; Neubig, G.; and Lipton, Z. C. 2020. Learning to Deceive with Attention-Based Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4782–4793. Online: Association for Computational Linguistics.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Saunders, D.; and Byrne, B. 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*, 7724–7736. Online: Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Stanovsky, G.; Smith, N. A.; and Zettlemoyer, L. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Webster, K.; Recasens, M.; Axelrod, V.; and Baldridge, J. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6: 605–617.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; and Petrov, S. 2020. Measuring and Reducing Gendered Correlations in Pre-trained Models. arXiv:2010.06032.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 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*. Florence, Italy.