Are We Paying Attention to *Her*? Investigating Gender Disambiguation and Attention in Machine Translation

Chiara Manna **Afra Alishahi** **Frédéric Blain** **Eva Vanmassenhove**

{c.manna, a.alishahi, f.l.g.blain, e.o.j.vanmassenhove}@tilburguniversity.edu

CSAI, Tilburg University

Netherlands

Abstract

While gender bias in modern Neural Machine Translation (NMT) systems has received much attention, the traditional evaluation metrics for these systems do not fully capture the extent to which models integrate contextual gender cues. We propose a novel evaluation metric called Minimal Pair Accuracy (MPA) which measures the reliance of models on gender cues for gender disambiguation. We evaluate a number of NMT models using this metric, we show that they ignore available gender cues in most cases in favour of (statistical) stereotypical gender interpretation. We further show that in anti-stereotypical cases, these models tend to more consistently take male gender cues into account while ignoring the female cues. Finally, we analyze the attention head weights in the encoder component of these models and show that while all models to some extent encode gender information, the male gender cues elicit a more diffused response compared to the more concentrated and specialized responses to female gender cues.[1]

1 Introduction

The field of Machine Translation (MT) has undergone significant technological shifts over the past decades, moving from transparent rule-based systems to increasingly opaque probability-based ones such as statistical and neural MT. Furthermore, the complexity and scale of current Transformer-based (Vaswani et al., 2017) architectures, which underpin both neural MT (NMT) and Large Language Models (LLMs), are making it more challenging to trace back model decisions and understand the underlying processes. This growing opacity raises concerns for AI governance where transparency, fairness and risk mitigation are becoming increasingly important for a responsible deployment of MT technology.

At the same time, research on (gender) bias in MT has been on the rise, reflecting more general tendencies in the field of Natural Language Processing (NLP) (Sun et al., 2019; Costa-jussà, 2019; Blodgett et al., 2020; Stanczak and Augenstein, 2021). The increasing awareness has led to concerns related to the flaws, inconsistencies and biases that models inherit, propagate and potentially exacerbate – especially with the increasing integration of NLP tools in people's everyday lives (Bansal, 2022). In response, AI governance policies are emerging worldwide, such as the European Union's AI Act (2024), aiming to regulate the development and deployment of AI systems to ensure ethical standards and mitigate potential risks. For MT specifically, the nature of the translation task itself further complicates matters due to cross-linguistic differences in gender representation and expression across languages, where social gender, linguistic gender and diverse cultural contexts intersect.

EN: The cook prepared a soup for the **housekeeper** because **he** helped clean the room.

IT: Il cuoco ha preparato una zuppa per **la governante** perché ha aiutato a pulire la stanza.

Figure 1: Example from the WinoMT dataset (Stanovsky et al., 2019) illustrating gender bias in an English-Italian translation. While the English sentence establishes the referent as male (using the pronoun *he*), the translation[2] uses a feminine form *la governante*, thereby disregarding the contextual gender cue.

[1]The code used in this work is made publicly available at github.com/chiaramanna/gender-cue-integration-MT.

[2]Generated by ChatGPT on March 6th, 2025.

Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025), pages 1–16, June 23, 2025

Languages encode gender in different ways and to varying degrees (Ackerman, 2019; Cao and Daumé III, 2020). While some, such as English or Danish, rely predominantly on pronouns, others, such as Italian, require morphological agreement across multiple parts of speech (Stahlberg et al., 2007). This implies that – in certain translation contexts – implicit source information must be made explicit in the target (Vanmassenhove et al., 2018). Figure 1 illustrates this through an example from the WinoMT dataset (Stanovsky et al., 2019). The English word *housekeeper* is translated into the Italian feminine form *la governante*, despite the broader sentence context indicating that the referent identifies as male (*he*). When this happens on a large scale and in a systematic way, it can result in representational and allocational harms, disproportionately affecting more marginalized groups (Blodgett et al., 2020) while simultaneously eroding linguistic diversity (Vanmassenhove et al., 2019, 2021b).

Despite the increasing awareness and research efforts over the past decade (Savoldi et al., 2024), gender bias in MT remains a complex, largely unsolved challenge (Vanmassenhove, 2024; Zhao et al., 2024). While current evaluation metrics offer a broad bias assessment, they do not capture whether models actively integrate contextual cues or default to learned statistical associations when disambiguating gendered nouns. This limitation makes it challenging to determine whether observed errors stem from a failure to process contextual information, the reinforcement of pre-existing biases, or internal shortcomings in how gender information is encoded and utilized. This hinders the development of targeted interventions and effective mitigation strategies.

To address this gap, we provide a nuanced evaluation framework that moves beyond a surface-level assessment of gender realization in an English-Italian translation context. Our main contribution is two-fold:

- We introduce **Minimal Pair Accuracy (MPA)**, a novel metric that measures whether models consistently rely on gender cues for gender disambiguation, rather than defaulting to learned priors. By leveraging the WinoMT dataset (Stanovsky et al., 2019), we construct minimal pairs, *i.e.*, sentence pairs that only differ in the gendered pronoun, and compute the proportion of cases where the model correctly adjusts the target gender.

- We conduct an exploratory **Attention-Based Analysis** to better understand how gender information is encoded within Transformer models. Specifically, we examine the extent to which profession nouns attend to gender cues at different layers and attention heads, and whether this behavior varies based on gender (masculine *vs.* feminine) or alignment with gender-role stereotypes (pro-stereotypical *vs.* anti-stereotypical contexts).

Our evaluation reveals that the assessed NMT models do not consistently leverage the contextual gender cues provided. Instead, they often seem to revert back to statistical (and thus stereotypical) patterns rather than context. We furthermore observe a discrepancy between the integration of masculine versus feminine cues. The presence of a masculine pronoun with pro-stereotypically female professions often enables the model to correctly infer the gender of a lexically gender-ambiguous target word while the reverse does not hold. Additionally, our analysis of attention head weights in the encoder component indicates that, although all models encode gender information to some extent, masculine cues elicit a more diffused response, whereas feminine ones generate more concentrated and specialized attention patterns.

2 Bias Statement

We define gender bias in MT as the tendency of models to default to learned statistical associations rather than systematically relying on contextual information for gender disambiguation. We focus on cases where gender is unambiguously expressed in the source sentence – typically through pronouns referring to human entities – capturing one subtype of gender bias. Ambiguous cases – lacking explicit gender cues – fall outside the scope of this paper.

While our framework targets the English-Italian (EN–IT) language pair, it is broadly applicable to any setting where gender must be explicitly marked in the target language. We particularly highlight stereotypical bias, for which models successfully generate feminine translations when the target word (*i.e.*, the profession noun) is already associated with women (*e.g. librarian* → *bibliotecaria*), but struggle to override male defaults in anti-stereotypical contexts. This asymmetry suggests that gender disambiguation might be driven by learned priors rather than syntactic dependencies, reinforcing a male-as-norm bias (Danesi, 2014). Such bias can lead to both representational harm, by perpetuating traditional gender roles, and allocational harm, by systematically underrepresenting women in male-dominated professions (Blodgett et al., 2020).

Our analysis only considers binary gender due to the constraints of the WinoMT dataset, which relies on U.S. Labor Statistics and morphological analysis tools that categorize gender along a binary axis. While we acknowledge that this is a major limitation and gender is not a binary construct, there is no standardized approach to systematically evaluate non-binary gender bias in MT. Broader inclusivity challenges persist and underscore the need for future work to develop more inclusive methodologies that better reflect gender as a spectrum.

3 Related Work

Research on gender bias in MT has largely focused on: analyzing MT output (e.g. Rescigno et al. (2020); Ramesh et al. (2021)...); rewriting into gendered (e.g. Vanmassenhove et al. (2018); Moryossef et al. (2019); Habash et al. (2019) or neutral outputs (e.g. Vanmassenhove et al. (2021a); Sun et al. (2021)...); word-embedding debiasing techniques (e.g. Hirasawa and Komachi (2019); Font and Costa-jussà (2019)....), domain adaptation (e.g. Saunders and Byrne (2020)), counterfactual data augmentation (e.g. Zmigrod et al. (2019)) and/or the development novel benchmarks and evaluation sets (e.g. Stanovsky et al. (2019); Luisa et al. (2020)....). Given that several studies (Blodgett et al., 2020; Stanczak and Augenstein, 2021; Savoldi et al., 2021) already offer a more comprehensive overview of broader discussions and research on (gender) bias in language technology, we specifically dedicate this related work section to the limited body of work focusing on the internal mechanisms underlying gender bias in (MT) models and interpretability techniques.

MT-specific research on interpretability techniques has largely focused on linguistic competence through probing (Belinkov et al., 2017a,b; Conneau et al., 2018), or by analyzing contrastive translation (Sennrich, 2017; Burlot and Yvon, 2017; Rios Gonzales et al., 2017; Vamvas and Sennrich, 2021, 2022). More recent work investigated how MT systems process intra- and intersentential context and whether their context usage aligns with human expectations (Goindani and Shrivastava, 2021; Voita et al., 2021; Sarti et al., 2024; Mohammed and Niculae, 2024). Despite high overall performance, these studies highlight how models often struggle to effectively leverage contextual information, either failing to integrate necessary information or attending to irrelevant tokens when resolving ambiguities (Kim et al., 2019; Yin et al., 2021), an interesting finding raising concerns about gender disambiguation which indeed could be driven by biased statistical patterns rather than reliance on relevant contextual cues.

The problem of context integration is not only relevant to model decision-making but also affects how gender bias is evaluated. Template-based evaluation frameworks, such as WinoMT (Stanovsky et al., 2019), provide controlled settings to measure surface-level accuracy metrics, and have been widely used to quantify gender bias across different language pairs and MT systems (Kocmi et al., 2020; Costa-jussà et al., 2020; Choubey et al., 2021). However, as these primarily rely on the alignment and morphosyntactic analysis of lexically gender-ambiguous words, they do not reveal whether models actively integrate contextual cues when making gender-related decisions. These limitations underscore the need for more nuanced evaluation methods.

A promising avenue for investigating how gender cues influence model decisions is through the study of context mixing, *i.e.*, the ability of Transformer-based models to dynamically incorporate information from the broader context into token representations. This process is largely governed by the attention mechanism, which plays a central role in these models. While attention-based analyses have been criticized for their reliability (Jain and Wallace, 2019; Bibal et al., 2022), and more advanced interpretability methods have been introduced (Kobayashi et al., 2020, 2021; Modarressi et al., 2022; Ferrando et al., 2022; Mohebbi et al., 2023b), attention weights remain a popular choice for analyzing model behavior due to their ability to provide direct insights into token interactions across layers and heads. As a matter of fact, they have been extensively leveraged to track token dependencies, revealing that specific attention heads may specialize in distinct linguistic functions (Xu et al., 2015; Rocktäschel et al., 2016; Wang et al., 2016; Lee et al., 2017; Vaswani et al., 2017; Kovaleva et al., 2019; Reif et al., 2019; Lin et al., 2019; Voita et al., 2019; Jo and Myaeng, 2020).

To the best of our knowledge, only the study by Bau et al. (2018) attempted to control gender through internal mechanisms in an MT setting. They explored this by probing and deactivating specific neurons associated with gender in an Long Short-Term Memory (LSTM) architecture. Their findings showed that gender-related properties are widely distributed across the network, making effectively controlling the output very difficult.

4 Experimental Setup

In order to examine the extent to which contextual gender cues contribute to the representation of profession nouns for different models, we analyzed how multiple state-of-the-art models (Section 4.1) integrate contextual gender cues provided in the WinoMT challenge set in the gender disambiguation process (Section 4.2).

4.1 Models

We investigate three pre-trained encoder-decoder models for English-to-Italian translation, selecting them based on their widespread use and high ranking among open source translation models on Hugging Face[3], allowing for greater transparency in analyzing their internal mechanisms. While we focus on encoder-decoder models, the framework can be extended to encoder-only or decoder-only architectures, adopted by LLMs.

OPUS-MT EN–IT[4] (Tiedemann et al., 2023) is a bilingual model specifically trained for English-to-Italian translation using supervised learning on parallel corpora from the OPUS dataset (Tiedemann, 2012). It consists of 6 encoder layers, 6 decoder layers, and 8 attention heads per layer.

NLLB-200[5] (Costa-jussà et al., 2022) is a multilingual model trained to support 200 languages. We make use of the distilled version, which contains 12 encoder layers, 12 decoder layers, and 16 attention heads per layer, with 600M parameters. The model is trained with a combination of supervised and self-supervised learning on multilingual corpora.

mBART[6] (Liu et al., 2020) is a denoising autoencoder-based multilingual model designed for cross-lingual tasks, including translation. Unlike the previous models, which rely primarily on parallel corpora, mBART is first pretrained on monolingual text using a denoising autoencoding objective, where the model learns to reconstruct text from a corrupted version. As NLLB-200, it features 12 encoder layers, 12 decoder layers, and 16 attention heads per layer.

[3]https://huggingface.co/
[4]huggingface.co/Helsinki-NLP/opus-mt-en-it
[5]huggingface.co/facebook/nllb-200-distilled-600M
[6]huggingface.co/facebook/mbart-large-50-many-to-many-mmt

4.2 Data

We use the **WinoMT** challenge set (Stanovsky et al., 2019), which was designed to evaluate gender bias in MT systems from English to 8 target languages with grammatical gender. It includes a regular set of 3888 synthetic sentences derived from Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) – two benchmark datasets for coreference resolution – in which a primary entity (*e.g.*, a profession noun) and a pronoun are in a coreference relation. Instances are balanced for masculine and feminine genders, as well as pro- and anti-stereotypical gender-role assignment based on U.S. Labor Statistics (Zhao et al., 2018).

PRO-S: The chief gave the **housekeeper** a tip because **she** was helpful.

ANTI-S: The chief gave the **housekeeper** a tip because **he** was helpful.

Figure 2: Example of a pro-stereotypical (PRO-S) and anti-stereotypical (ANTI-S) gender role assignment from the WinoMT challenge set.

Additionally, two sets of 1584 instances each are provided – en_pro and en_anti – where the same profession nouns are paired with pronouns based on pro- and anti-stereotypical gender-roles, respectively. To illustrate this, we present the same sentence from both sets in Figure 2. In the pro-stereotypical sentence (PRO-S), the gender of *housekeeper* aligns with the gender that most often carries out this particular profession according to the U.S. Bureau of Labour Statistics.[7] Conversely, in the anti-stereotypical (ANTI-S) setting the gender role assigned can be considered more challenging as statistically[8] men are less likely to carry out the job of *housekeeper*.

[7]In 2024, 87.7% of housekeepers are women – see: bls.gov/cps/cpsaat11.htm.
[8]Again, based on US statistics.

Set	Model	Overall	Male	Female
REG	OPUS_MT	42.6%	70.1%	20.6%
	NLLB-200	57.0%	79.6%	41.8%
	mBART	**60.9%**	**83.2%**	**46.5%**
PRO-S	OPUS_MT	55.7%	77.3%	34.1%
	NLLB-200	74.9%	87.4%	**62.5%**
	mBART	**76.6%**	**92.2%**	61.0%
ANTI-S	OPUS_MT	34.2%	59.1%	9.2%
	NLLB-200	47.3%	70.4%	24.2%
	mBART	**54.0%**	**71.9%**	**35.9%**

Table 1: Overall, male and female accuracy on WinoMT for the OPUS_MT, NLLB-200, and mBART models on the regular (REG), pro-stereotypical (PRO-S), and anti-stereotypical (ANTI-S) sets.

5 Evaluating Context Integration in Gender Disambiguation

In this section, we will first delve into the evaluation of contextual cue integration through our novel metric. Next, in Section 6, we continue with the analysis of the encoder attention head weights to investigate how gender cues are integrated into the target representations.

5.1 WinoMT Evaluation

WinoMT provides an integrated evaluation pipeline that relies on automatic word alignment and morphological analysis to extract the grammatical gender of the primary entity from each translated sentence. Comparing the extracted gender information with the gold label enables us to compute three accuracy measures:

Overall Accuracy: Percentage of correctly gendered entities.

Male Accuracy: Accuracy for entities with a masculine gold label.

Female Accuracy: Accuracy for entities with a feminine gold label.

Table 1 presents the gender accuracy for all models across the regular, pro- and anti-stereotypical sets. First of all, we observe that all models consistently perform better for: (i) masculine referents and (ii) in stereotypical settings where the gender aligns with the societal expectations. When comparing the models, mBART outperforms NLLB-200 and OPUS-MT on all three sets (regular, stereotypical and anti-stereotypical) in terms of overall and male accuracy. Only on the stereotypical set, NLLB-200 (62.5%) slightly outperforms BART (61.0%) in terms of accuracy for female referents.

Model	MPA
OPUS_MT	6.12%
NLLB-200	30.24%
mBART	**38.45%**

Table 2: MPA presents the percentage of cases where the model correctly disambiguated the source word based on the contextual gender cue.

Model	Pro-F	Pro-M
OPUS_MT	**82.29%**	17.71%
NLLB-200	69.10%	30.90%
mBART	61.90%	**38.10%**

Table 3: A breakdown of the MPA. **Pro-F** refers to the percentage of correctly disambiguated minimal pairs where the profession would stereotypically be associated with women. **Pro-M** refers to the ones where the profession would stereotypically be associated with men.

5.2 Minimal Pair Accuracy

While the aforementioned standard metrics provide an overall performance measure, they do not assess whether models effectively leverage contextual gender cues to resolve gender ambiguity during translation. In an attempt to move beyond these rather surface-level accuracy scores, we introduce and analyze the **Minimal Pair Accuracy (MPA)**, a novel metric designed to assess the degree to which models are able to effectively leverage contextual gender cues for disambiguation during translation.

We construct minimal pairs by extracting `PRO-S` and `ANTI-S` sentences, with the only distinguishing factor being the gendered pronoun. This allows us to analyze the effect of the gender cue on translation. For instance, as shown in Figure 3, the profession noun *librarian* is translated into Italian as either *la bibliotecaria* (feminine) or *il bibliotecario* (masculine), depending on the pronoun (e.g., *she* or *he*) in the English source sentence. We then focus on pairs where the model successfully disambiguates the profession noun in both cases by utilizing the contextual gender information provided by the pronoun. This way, **MPA** can give us an indication of how consistently the models integrate contextual gender cues.

The overall low MPA results presented in Table 2 indicate that models struggle to consistently leverage contextual gender cues for disambiguation. However, NLLB-200 and mBART perform notably better, with an accuracy of 30.24% and 38.45% – respectively – as compared to OPUS-MT's significantly lower 6.12%.

PRO-S: The analyst consulted with the **librarian** because **she** knows a lot about books.

ANTI-S: The analyst consulted with the **librarian** because **he** knows a lot about books.

OPUS-MT: L'analista si è consultato con **la bibliotecaria** perché sa molto sui libri.

OPUS-MT: L'analista si è consultato con **il bibliotecario** perché sa molto sui libri.

NLLB-200: L'analista ha consultato **la bibliotecaria** perché sa molto di libri.

NLLB-200: L'analista ha consultato **il bibliotecario** perché sa molto di libri.

mBART: L'analista ha consultato **la bibliotecaria** perché sa molto sui libri.

mBART: Il analista ha consultato **il bibliotecario** perché sa molto sui libri.

Figure 3: Example of accurate minimal pair translations constructed from the WinoMT challenge set. The left side (pro-stereotypical) assigns the feminine pronoun *she* to the profession *librarian*, while the right side (anti-stereotypical) replaces it with the masculine pronoun *he*. The Italian translations correctly adapt the grammatical gender (*la bibliotecaria vs. il bibliotecario*) across all models. Therefore, this pair contributes positively to the Minimal Pairs Accuracy (MPA) for each model.

A closer examination of those accurate minimal pairs reveals yet another layer of asymmetry. Table 3 presents the percentage of accurately translated minimal pairs where the profession is stereotypically associated with women (Pro-F) versus those where the profession is stereotypically associated with men (Pro-M). The results show that correctly disambiguating a profession noun based on a gender cue is much easier when the profession is stereotypically associated with women. In other words, stereotypical female professions are relatively easy to override with a masculine cue.

An example can be found in Figure 3, where all models correctly disambiguate a stereotypically female profession *librarian*[9] in both a stereotypical (PRO-S) and anti-stereotypical (ANTI-S) setting. Even in the ANTI-S condition, where the context provides a masculine cue (*he*), the correct anti-stereotypical masculine form *il bibliotecario* is generated by all three models. Overriding a stereotypically male-dominated profession is more difficult for all three models. When *mechanic* – a profession predominantly held by men[10] – is paired with *she*, none of the models succeed in generating the expected feminine form, *la meccanica*.

Specifically, OPUS_MT shows that only 17.71% of accurate minimal pairs successfully utilize an anti-stereotypical context to disambiguate a feminine referent. While this percentage increases slightly with the other models, it remains below 40%, indicating a general difficulty in overriding male defaults.

These findings indicate that feminine cues only trigger gender disambiguation when the profession noun they refer to is stereotypically associated with the feminine gender. Otherwise, the investigated models often default to masculine terms, reinforcing an inherent male-as-norm bias (Danesi, 2014). Previous work supports this pattern, showing that language models, in fact, tend to follow a default-to-masculine reasoning process when assigning gender (Jumelet et al., 2019).

6 Investigating Context Integration Through Attention

To gain further insight into how contextual gender information is encoded within Transformer models, we further investigate the extent to which gender cues are integrated into the representation of target words. For example, if a model correctly translates both **PRO-S** and **ANTI-S** examples in Figure 3, we expect the representation of the target word *librarian* to be heavily influenced by the gender cue *she/he* in the original sentence. More specifically, we are interested in analyzing whether the attention mechanism contributing to the input representation of the target word attends to the gender cue and, if so, whether there are specific attention layers and heads that specialize in encoding gender cues.

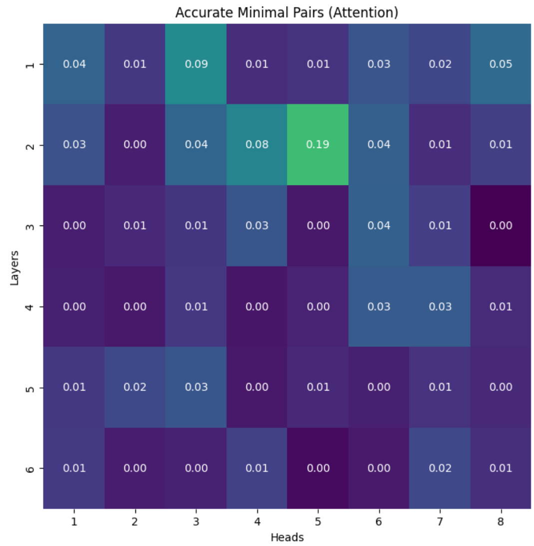
[9]In 2024, based on the US Labor Force Statistics, 89.2% of librarians are women – see: bls.gov/cps/cpsaat11.htm.

[10]In 2024, based on the US Labor Force Statistics, only 3.2% of mechanics are women – see: bls.gov/cps/cpsaat11.htm.

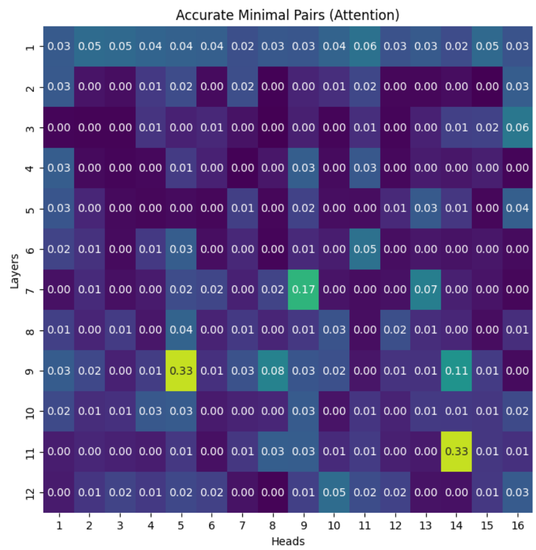
6.1 Setup

Contextual information is leveraged through a multi-head attention mechanism in Transformer models. This operates at three levels in encoder-decoder architectures: self-attention in the encoder, self-attention in the decoder and cross-attention between the decoder and encoder representations (Vaswani et al., 2017). Previous work on context mixing in Transformer models has shown that encoder-only models effectively integrate contextual cues in their representations, while encoder-decoder models seem to relegate this task to the decoder (Mohebbi et al., 2023a,b). However – in our setup – the gender cue (*i.e.*, the pronoun) is preceded by the target word (*i.e.*, the profession noun) (see Figure 1). As a result, decoder self-attention cannot account for it, as it only captures dependencies within already-generated tokens. Therefore, we focus on the self-attention patterns observed within the encoder in our analysis.[11]

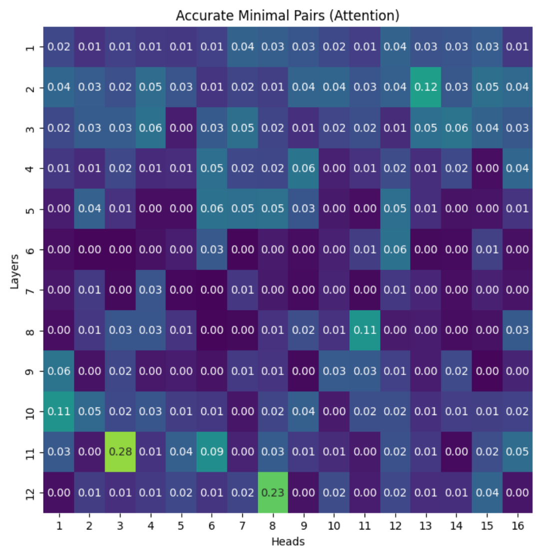
Given that the gender cue serves as the only explicit indicator of the primary entity's gender in the source language (EN), it is expected to play a key role in the gender disambiguation of the target word in the target language (IT). To examine this, we focus on accurately gendered minimal pairs. We begin by identifying the target word's source-side index by leveraging the annotations in WinoMT. We then align source and target sentences using `fast_align`[12] to retrieve the target word's corresponding index in the generated translation. The gender cue is identified by detecting a predefined set of pronouns (*he, she, him, her, his*) in the source sentence, from which we extract their corresponding index. Since both target word and gender cue may be tokenized into multiple subwords, we map them accordingly by iterating through the tokenized sequence, incrementally matching subword segments. Once the relevant (subword) indices are obtained, we extract the corresponding attention weights from the model's attention matrices. To account subword tokenization, we compute the average attention weights across subword tokens before aggregating the values across all instances. Since attention weights sum to 1 across all context tokens, no further normalization is required.



(a) OPUS-MT



(b) NLLB-200

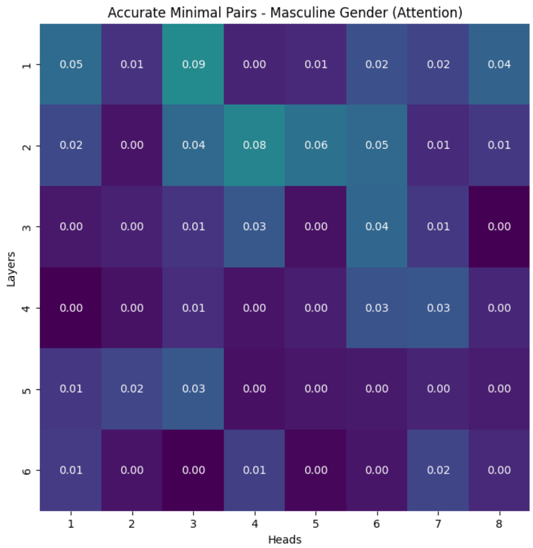


(c) mBART

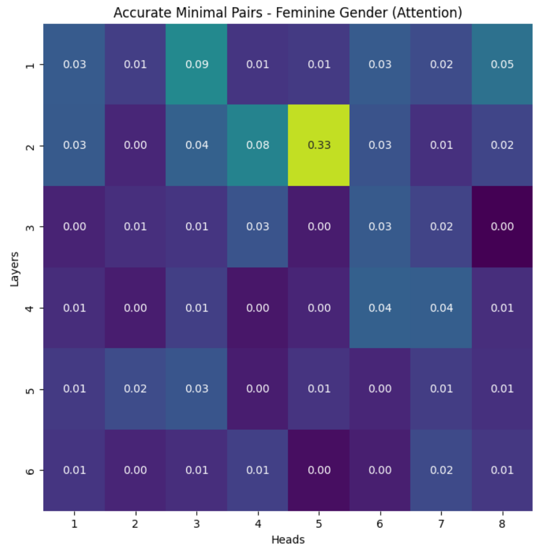
Figure 4: Heatmaps of average encoder self-attention weights between gender cue (*i.e.*, pronoun) and profession noun across accurate minimal pairs. A standardized colormap is applied.

[11]The analysis of cross-attention heads did not reveal notable patterns, but for the sake of completeness, the full set of cross-attention results are reported in the Appendix.

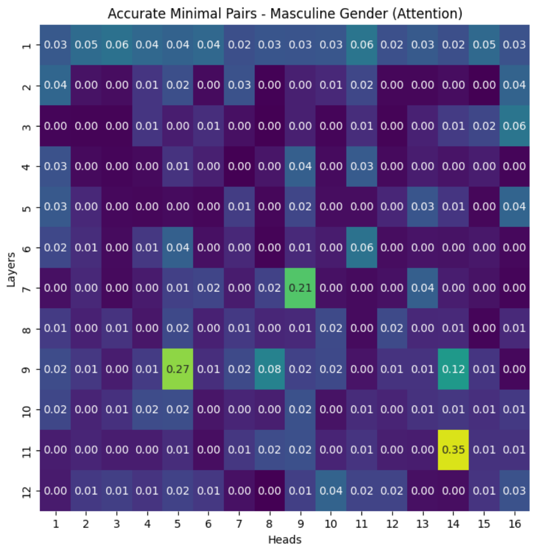
[12]github.com/clab/fast_align



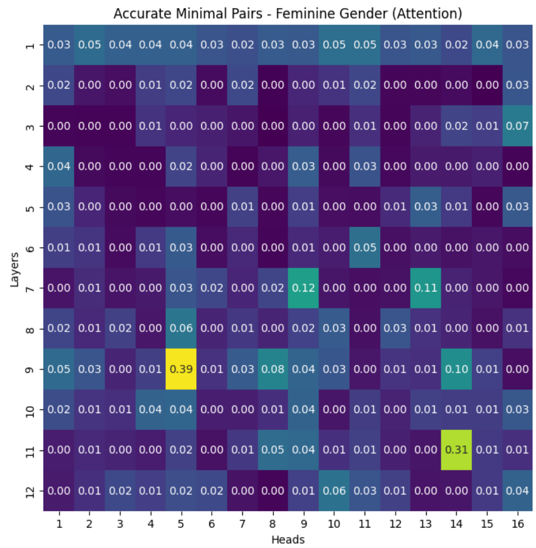
(a) OPUS-MT



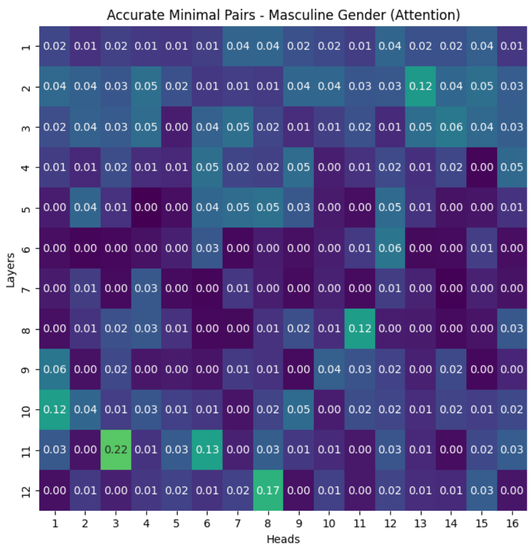
(b) OPUS-MT



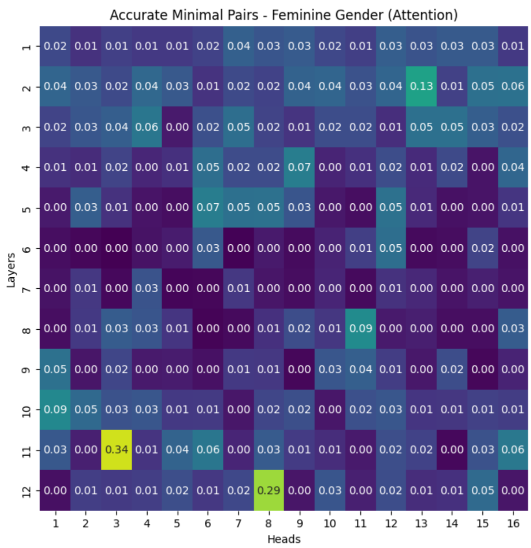
(c) NLLB-200



(d) NLLB-200



(e) mBART



(f) mBART

Figure 5: Heatmaps of average encoder self-attention weights between gender cue (*i.e.*, pronoun) and profession noun across accurate minimal pairs. Each row contrasts masculine (left) *vs.* feminine (right) referents. A standardized colormap is applied.

6.2 Results and Analysis

The heatmaps in Figure 4 illustrate the self-attention weights between the gender cue and the target word, averaged across all sentences. These scores indicate how much the target word attends to the gender cue, *i.e.*, the contribution of the cue to the target word's contextualized representation.

To identify attention heads that may play a more specialized role in gender disambiguation, we establish a threshold of relevance. Given that minimal pair sentences have an average length of ≈ 13 words, a uniform attention distribution would allocate a weight of approximately $1/13$ (≈ 0.08) to each word. Therefore, we consider attention heads that exceed this baseline by a notable margin as potentially relevant for gender cue integration.

Comparing the models, we observe distinct attention patterns. While for OPUS_MT a single attention head stands out at an early stage of encoding (layer 2), the other two models display a more distributed pattern, with at least two potentially influential attention heads emerging in deeper layers. This seems to indicate a more diffuse integration and a multi-layered processing of the gender cues.

To further investigate whether models encode masculine and feminine gender cues differently, we separately report the attention scores for masculine and feminine pronouns in Figure 5. This reveals that feminine pronouns elicit more localized activations, while masculine ones tend to receive weaker, more dispersed attention, especially for OPUS-MT and mBART. Finally, NLLB-200 exhibits a different type of asymmetry, in which distinct attention heads appear to specialize in encoding gender-specific patterns – some being more responsive to feminine pronouns, others playing a stronger role in encoding masculine ones.

While informative, these results must be interpreted with caution. As most feminine examples are found in pro-stereotypical settings (Table 3), the observed attention patterns may reflect a form of training or dataset bias, where models have learned to associate certain professions with feminine pronouns due to their statistical distribution in the training data, rather than consistently relying on syntactic dependencies. Furthermore, combining this with the way minimal pairs are constructed, an inherent gender composition imbalance emerges. Since feminine entities are predominantly featured in pro-stereotypical examples, masculine ones are mostly found in anti-stereotypical settings. As a result, there are relatively fewer observations for pro-stereotypical masculine and anti-stereotypical feminine cases, making it difficult to draw definitive conclusions about gender cue integration in these underrepresented scenarios.

7 Discussion

In this section, we reflect on the key findings from our two-fold analysis, their implications, as well as potential avenues for future research.

7.1 Minimal Pair Accuracy and Default Masculinity

While standard metrics of gender accuracy reveal that the investigated encoder-decoder models perform better for masculine referents and in pro-stereotypical settings, the proposed MPA uncovers another systematic asymmetries in gender disambiguation and exposes a persistent male-as-norm bias (Danesi, 2014).

Although NLLB-200 and mBART showcase a more consistent integration of contextual information as compared to OPUS-MT, all models struggle to correctly disambiguate stereotypically male-dominated professions when provided with a feminine cue word while the reverse does not hold true. Namely, combining a stereotypically male profession with a feminine target cue (e.g., *she*) often fails to trigger the corresponding feminine form, with models defaulting to the masculine variant. This asymmetry suggests a stronger bias towards masculine defaults, particularly in contexts where the feminine form challenges prevailing stereotypes. This asymmetry raises a more fundamental question of whether MT models can indeed consistently process syntactic dependencies for gender disambiguation or whether they are predominantly influenced by entrenched statistical associations. Our results seem to reinforce prior findings that language models often follow a default-to-masculine reasoning process when assigning gender (Jumelet et al., 2019; Danesi, 2014), hence we wonder: *Are we paying attention to **her**?*

7.2 The Role of Attention in Gender Encoding

As the model's primary objective is translation, gender disambiguation is likely treated as an auxiliary task, with responsibility for its resolution distributed across various parts of the network, *i.e.*, specific layers or attention heads within the model (Xu et al., 2015; Wang et al., 2016; Rocktäschel et al., 2016; Lee et al., 2017; Vaswani et al., 2017; Clark et al., 2019; Kovaleva et al., 2019; Reif et al., 2019; Lin et al., 2019; Voita et al., 2019; Jo and Myaeng, 2020).

Having isolated accurate minimal pairs, we can speculate that the identified influential heads may specialize in encoding gender information during translation. Overall, these appear in early layers for OPUS-MT, mid-to-deep layers for NLLB-200, and deeper layers for mBART. Interestingly, gender cue integration is not uniform across all models and presents gender-specific patterns. Specifically, we observe that feminine pronouns elicit more localized activations, while masculine ones tend to receive weaker, more dispersed attention, especially for OPUS-MT and mBART. This aligns with prior research on gender representation in language models, which has shown that masculinity tends to function as the default category, while gender-specific signals – particularly feminine ones – are processed in a more localized manner (Jumelet et al., 2019; Van Der Wal et al., 2022). Notably, NLLB-200 exhibits a different type of asymmetry, where distinct attention heads appear to specialize in encoding gender-specific patterns – some being more responsive to feminine pronouns, others playing a stronger role in encoding masculine ones.

Expanding on these results, we find that models with more distributed and diffuse attention activation – such as mBART and NLLB-200 – perform better in terms of both gender accuracy and MPA compared to OPUS-MT, which attends gender cues in a single early-layer attention head. This suggests that gender disambiguation may benefit from a more adaptable, multi-layered gender encoding mechanism rather than a rigid, localized one.

7.3 Limitations and Future Work

Our findings suggest potential avenues for binary gender bias mitigation strategies. Given that potentially influential attention heads have been identified, targeted interventions could be explored to enhance gender cue integration. Specifically, two promising directions include (i) fine-tuning seemingly specialized attention heads or (ii) enforcing a minimum attention threshold to ensure that gender cues receive sufficient weight when generating target words.

Context mixing scores – such as attention weights – provide useful insights into how models may be processing gender cues and encoding gender-related information, especially when combined with nuanced evaluation metrics such as MPA. However, they should not be taken as definitive explanations of model decision-making as no causal relationship between gender cue integration and translation outputs is established. Although subsetting on accurately gendered minimal pairs partially addresses this limitation, it also introduces additional challenges. As discussed in Section 6.2, the gender composition imbalance within minimal pairs makes it difficult to assess whether observed attention patterns genuinely reflect contextual gender disambiguation or are simply a byproduct of learned statistical associations in the data. To address these challenges, future work should explore mechanistic interpretability methods – such as activation patching (Vig et al., 2020; Meng et al., 2022; Heimersheim and Nanda, 2024) – to directly assess the causal role of gender cues in translation decisions.

8 Conclusion

In this work, we examined how Transformer-based NMT models integrate contextual gender cues and uncovered systematic biases and asymmetries in their processing mechanisms.

Taken together, our findings reinforce previous calls for greater caution when interpreting benchmark scores for gender accuracy in MT (Savoldi et al., 2021). Surface-level improvements, such as higher gender accuracy, can still obscure deeper biases in how and under which conditions these forms indeed appear. More nuanced and comprehensive analyses are needed to to determine whether current systems truly leverage gender-specific cues or merely reinforce statistical stereotypes in subtler ways.

Without a more careful consideration of when, why and how certain patterns emerge, we risk misinterpreting progress and overlooking specific persistent and more structural biases in MT. Ultimately, understanding how gender is encoded in translation models is a crucial component to ensure more fairness, accountability, and transparency in AI systems.

Acknowledgements

We thank the reviewers for their insightful comments and feedback. We further extend our gratitude to our colleague Hosein Mohebbi for his critical suggestions and guidance, which helped shape the direction of this work.

References

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

Rajas Bansal. 2022. A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

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

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

Marcel Danesi. 2014. *Dictionary of Media and Communications*. Routledge.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Akshay Goindani and Manish Shrivastava. 2021. A dynamic head importance computation mechanism for neural machine translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 454–462, Held Online. INCOMA Ltd.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *Preprint*, arXiv:2404.15255.

Tosho Hirasawa and Mamoru Komachi. 2019. Debiasing word embeddings improves multimodal machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 32–42.

Sarthak Jain and Byron C. Wallace. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 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)*, pages 7057–7075, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT‘s linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Bentivogli Luisa, Beatrice Savoldi, Negri Matteo, A Di Gangi Mattia, Cattoni Roldano, Turchi Marco, et al. 2020. Gender in danger? evaluating speech translation technology on the must-she corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2024. On measuring context utilization in document-level MT systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta. Association for Computational Linguistics.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023a. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023b. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8594–8603.

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. *Computational Linguistics CLiC-it 2020*, page 359.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

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.

Beatrice Savoldi, Jasmijn Bastings, Lucia Bentivogli, and Eva Vanmassenhove. 2024. A decade of gender bias in machine translation. *[under review]*.

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.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. *Representation of the sexes in language*, page 163–187.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

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

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

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2023. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

European Union. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75. Association for Computational Linguistics.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *Gendered Technology in Translation and Interpreting: Centering Rights in the Development of Language Technology*, page 225.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021a. Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. *arXiv preprint arXiv:2109.06105*.

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

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021b. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232.

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*, volume 30. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *Preprint*, arXiv:2004.12265.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 606–615.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? 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 788–801, Online. Association for Computational Linguistics.

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

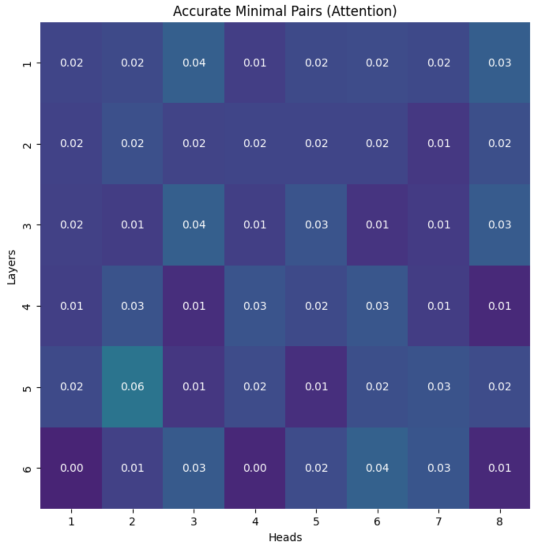
Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv*.

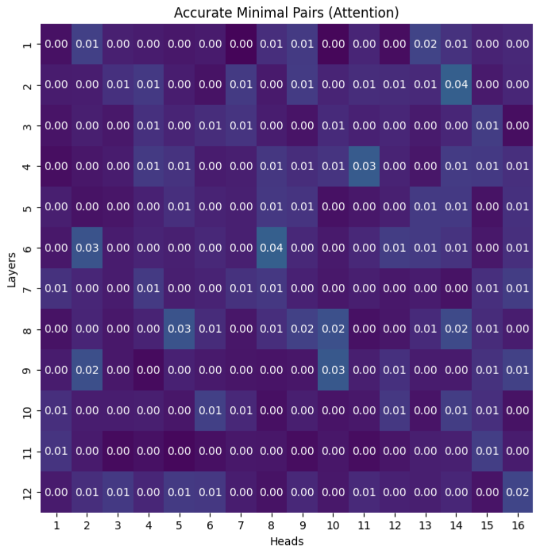
Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Cross-Attention Analysis

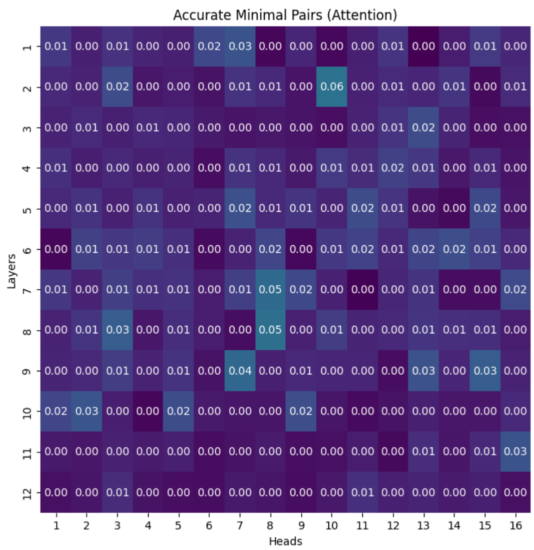
In this section, we present the average cross-attention weights, illustrating how the decoder attends to the gender cue (*i.e.*, the pronoun) in the encoder representations when generating the target word (*i.e.*, the gendered profession).



(a) OPUS-MT

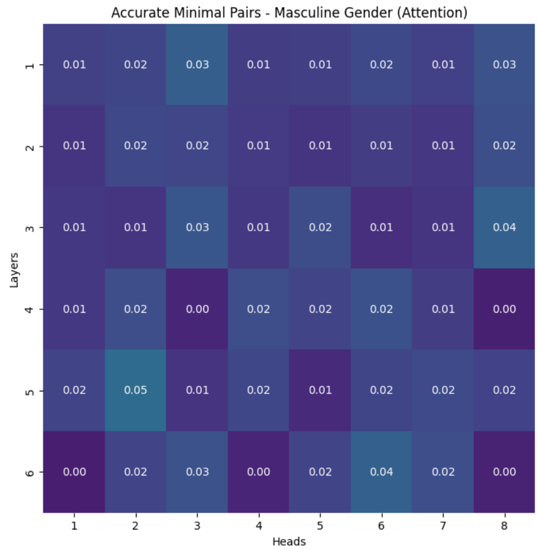


(b) NLLB-200

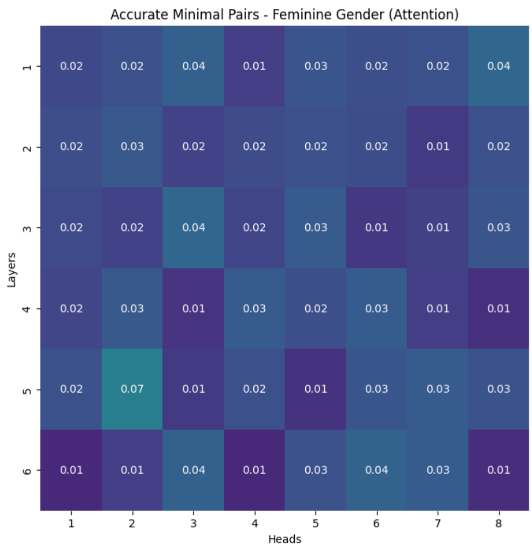


(c) mBART

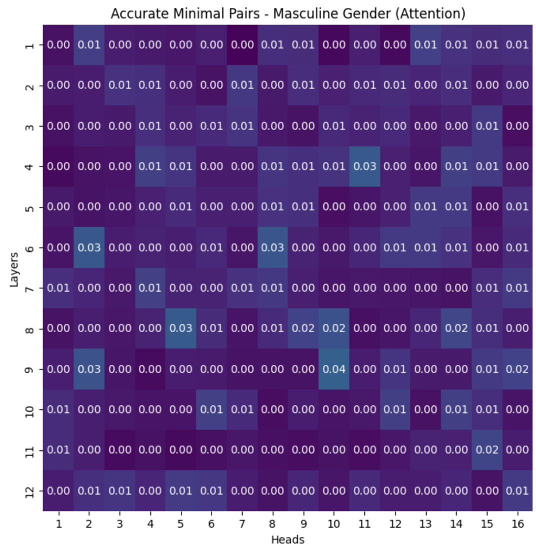
Figure 6: Heatmaps of average cross-attention weights to the the gender cue when generating the profession noun across accurate minimal pairs. A standardized colormap is applied.



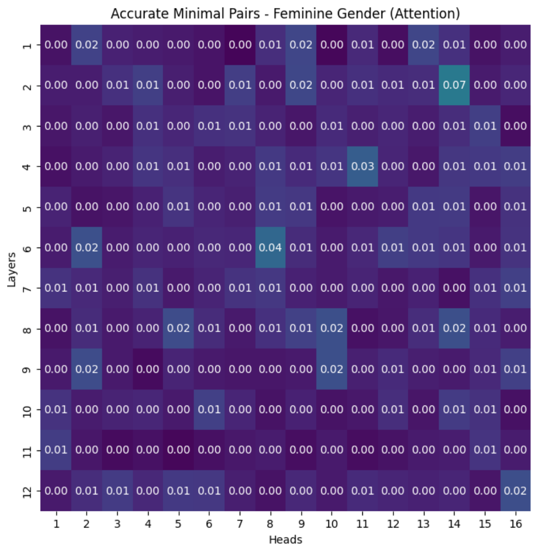
(a) OPUS-MT



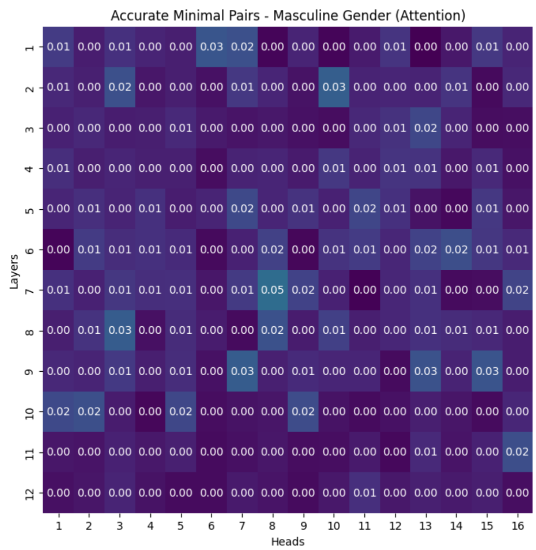
(b) OPUS-MT



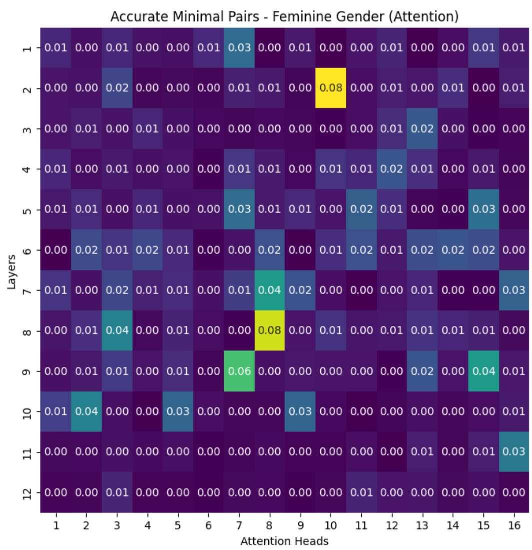
(c) NLLB-200



(d) NLLB-200



(e) mBART



(f) mBART

Figure 7: Heatmaps of average cross-attention weights to the gender cue when generating the profession noun across accurate minimal pairs. Each row contrasts masculine (left) *vs.* feminine (right) referents. A standardized colormap is applied.