Gender Bias Mitigation for NMT Involving Genderless Languages

Ander Corral and **Xabier Saralegi**
Orai NLP Technologies, Basque Country, Spain
a.corral@orai.eus, x.saralegi@orai.eus

Abstract

It has been found that NMT systems have a strong preference towards social defaults and biases when translating certain occupations, which due to their widespread use, can unintentionally contribute to amplifying and perpetuating these patterns. In that sense, this work focuses on sentence-level gender agreement between gendered entities and occupations when translating from genderless languages to languages with grammatical gender. Specifically, we address the Basque to Spanish translation direction for which bias mitigation has not been addressed. Gender information in Basque is explicit in neither the grammar nor the morphology. It is only present in a limited number of gender specific common nouns and person proper names. We propose a template-based fine-tuning strategy with explicit gender tags to provide a stronger gender signal for the proper inflection of occupations. This strategy is compared against systems fine-tuned on real data extracted from Wikipedia biographies. We provide a detailed gender bias assessment analysis and perform a template ablation study to determine the optimal set of templates. We report a substantial gender bias mitigation (up to 50% on gender bias scores) while keeping the original translation quality.

1 Introduction

As the neural machine translation (NMT) field becomes more mature, there is a growing concern about the gender fairness of these systems (Stanovsky et al., 2019; Prates et al., 2020; Hovy et al., 2020; Savoldi et al., 2021). These data-driven approaches are trained on large real-world textual corpora which often exhibit implicit social gender stereotypes and biases. For example, Bolukbasi et al. (2016) noted that systems associate certain neutral occupations with males, such as doctor or programmer, and others with females, such as nurse or housekeeper. As a consequence, although not being required by the task, systems tend to inherit and amplify these social biases.

Several different solutions have been proposed to solve, or at least reduce, gender bias during the translation process: providing alternative masculine and feminine translations for some neutral words (Johnson, 2018); adding explicit gender information during training (Vanmassenhove et al., 2018; Stafanovičs et al., 2020; Saunders et al., 2020); removing bias from word embeddings (Font and Costa-Jussa, 2019); or fine-tuning on a small gender-balanced data set (Costa-jussà and de Jorge, 2020; Saunders et al., 2020). There have also been some efforts to construct some challenge sets to systematically assess gender bias (Stanovsky et al., 2019; Bentivogli et al., 2020).

Most of the previous work has focused on English as the source language which is then translated to languages with grammatical gender such as Spanish, French, German, etc. English is a notional gender language which encodes gender in a pronominal system (*he/she, his/her...*) (Savoldi et al., 2021). Consolidated evaluation benchmarks such as WinoMT (Stanovsky et al., 2019) or MuST-SHE (Bentivogli et al., 2020) are specially designed for English. However, for genderless languages such as Basque, existing previous work and benchmarks do not fully satisfy the requirements. For example, WinoMT uses pronominal references as a disambiguation signal for the correct inflection of occupations, which do not exist in Basque.

Gender information in Basque is explicit in neither the grammar nor the morphology. This fact implies that gender can only be determined when nouns correspond unequivocally to a female or a male, that is, person proper names or a limited number of gender-specific common nouns (e.g., *emakumea/gizona, aita/ama...*[1]), hereinafter referred to as gendered entities. Therefore, existing approaches and evaluation benchmarks need to be

[1]English translation: *woman/man, father/mother...*

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 165 - 176

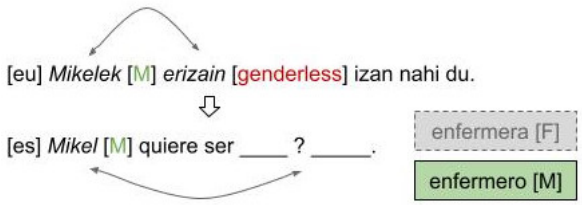

Figure 1: An illustrative example for the task of sentence-level agreement between gendered entities and occupations when translating from genderless languages to gendered languages. English translation: *Mikel wants to be a nurse.*

adapted to meet the requirements of a genderless language.

In this work we address the specific task of sentence-level gender agreement between gendered entities and occupations when translating from genderless languages to gendered languages (see example in Figure 1). We focus on the Basque to Spanish translation direction, a translation direction that presents the peculiarities described above and that has not been studied in the literature. The main contributions of the paper are the following:

- A template-based fine-tuning method with explicit gender signals to debias pre-trained systems involving genderless languages.
- A detailed experimentation to determine the source of gender bias for the task, including an in-depth ablation study of the template-based method and a comparison against fine-tuning on a gender-balanced Wikipedia biographies set.

2 Related work

A wide variety of publications warn about the lack of fairness some of the commercial MT systems have and how they might contribute to amplify and perpetuate social gender stereotypes due to their widespread use (Stanovsky et al., 2019; Prates et al., 2020; Hovy et al., 2020). In the case of the Basque-Spanish language pair, Salaberria et al. (2021) found a preference for the stereotyped translation of occupations according to their historically assigned role.

Ideally, system bias could be mitigated by removing all the bias present in the training data. For example, by augmenting data samples with their corresponding counterfactual forms (Zhao et al., 2018; Zmigrod et al., 2019). Nevertheless, this task still poses some challenges for grammatical gender languages as it involves preserving the morpho-syntactic agreement of the whole sentence by gender swapping pronouns, adjectives, verbs, entities, etc., (Stafanovičs et al., 2020).

As a result, several alternative methods have been proposed to alleviate gender bias from MT systems. Vanmassenhove et al. (2018) add a special gender token to source sentences in order to improve morphological agreement between the uttered sentence and the gender of the speaker. Stafanovičs et al. (2020) annotate source words with the grammatical gender information of their corresponding target words. Basta et al. (2020) provide the system with discourse context by adding the previous sentence, and in the same direction, Moryossef et al. (2019) propose a method to guide a black-box model by appending some contextual gender unambiguous hints to source sentences, such as "... *she told them*". Other more complex approaches have targeted gender bias effects by directly equalizing genders in word embeddings (Escudé Font and Costa-jussà, 2019).

Another promising line of research addresses gender bias as a domain adaptation problem by fine-tuning a pre-trained biased system with a gender-balanced data set. Costa-jussà and de Jorge (2020) automatically collect gender-balanced parallel data from Wikipedia biographies by selecting an equal amount of examples for each gender. Choubey et al. (2021) generate gender filtered parallel data by forward-translating a monolingual corpus. Saunders and Byrne (2020) generate a small, trivial, gender-balanced set of synthetic examples by inflecting a single handcrafted template with an equal number of masculine and feminine entities. In Saunders et al. (2020) they further improve the method by adding explicit word-level gender tags.

Our proposed method uses a controlled gender-balanced set of examples (Section 3) to fine-tune a pre-trained model (Section 4.1). We further provide the system with explicit gender tags for proper gender inflections (Section 4.2). We assume that providing a stronger gender signal is better than letting the system infer the proper gender of each gendered entity. In order to annotate source words, Stafanovičs et al. (2020) propose a complex method involving morphological tagging and automatic alignment, and Saunders et al. (2020) rely on the proper coreference resolution. In contrast,

we directly annotate gendered entities and leave gender agreement to the system instead of directly providing gender inflection information for all the affected words. This assumption simplifies the annotation effort as only gendered entities lists are necessary to annotate the data.

3 Gender-balanced corpora

Existing previous work on gender bias has focused on high resource translation directions (English to Spanish, French, German, etc.). While consolidated benchmarks and data exist for these languages, they strongly rely on gendered pronouns which makes them difficult to adapt for Basque. Thus, we analyzed two different strategies to build gender-balanced corpora for the Basque to Spanish translation direction: a syntactically diverse set of handcrafted templates (Section 3.1), and real data extracted from Wikipedia biographies (Section 3.2).

3.1 Handcrafted templates

A Basque and Spanish native speaker manually designed a set of task-specific templates for the correct treatment of gender agreement at sentence-level between gendered entities and occupations. We argue that a single tiny template does not provide sufficient syntactic diversity, so we handcrafted a syntactically diverse set of 33 templates. Each of the templates has placeholders for an occupation and a gendered entity to help in the proper disambiguation of that occupation.

Saunders et al. (2020) reported that systems trained on single-entity templates tend to overgeneralize gender signals on multi-entity examples by indiscriminately applying the same gender to all the occupations regardless of the other entities' genders. For instance, the Basque source sentence "*Josean iragarlea zen eta Leirek idazkaria izan nahi zuen.*" [2] would be translated to "*Josean era adivino y Leire quería ser secretario.*" instead of producing the correct feminine form "*secretaria*". To address this issue, we also construct a set of 13 multi-entity templates with two gendered entities and their corresponding occupations.

We use an occupations list and a gendered entities list to automatically populate the templates. We slightly adapted the list of occupations from Salaberria et al. (2021) to obtain a set of 83 occupations in Basque and their respective translations for both genders in Spanish. The gendered entities list contains a set of 200 common Basque and Spanish person proper names. We further complemented that list with 14 gendered common nouns referring to humans in Basque (e.g., *emakumea/gizona, aita/ama...*[3]) by querying Basque WordNet (Pociello et al., 2011). We collected the same amount of gendered entities for each gender.

We randomly divided the handcrafted templates[4] into disjoint training and test sets, keeping 6 single entity templates and 3 multi-entity templates for testing purposes. For each gender, 20 proper names and 4 gendered terms are used to inject these test templates. A total amount of 3,120 single entity examples and 1,800 multi-entity examples were created, hereinafter referred to as *Templ_test* and *Multi_test* test sets. The rest, 27 and 10 templates respectively, are kept to create training data (*Templ_train* and *Multi_train*) and were injected in different ways as explained in Section 4.1. Some examples of the handcrafted templates are shown in Table 1.

3.2 Back-translated Wikipedia biographies

In order to generate a gender-balanced set of real data, we turned to Wikipedia biographies. We focused on the extraction of examples that present gender agreement between people and occupations for the Basque-Spanish language pair. Unlike the strategy proposed by Costa-jussà et al. (2019), we extract task specific examples from Spanish monolingual biographies which are then back-translated to Basque, instead of directly extracting parallel data. The reason behind this decision was that more task specific examples could be gathered from leveraging Spanish monolingual data only, as Basque biographies constrained the amount of examples that could be gathered.

We searched the Spanish Wikipedia (extracted using WikiExtractor[5]) for biographies of living persons using Petscan[6]. We found 160,641 biographies matching this criteria. Using the Wikidata API, we automatically detected the gender of these persons. We only extracted the first sentence from each biography, which generally includes examples

[2]English translation: *Josean was a fortune-teller and Leire wanted to be a secretary.*

[3]English translation: *woman/man, father/mother...*

[4]All the handcrafted templates and occupations and gendered entities lists are included in the supplementary material.

[5]https://github.com/attardi/wikiextractor

[6]https://petscan.wmflabs.org/

SINGLE-ENTITY TEMPLATE
eu: {entity}k {occupation} izan nahi du.
→ *Mikelek erizain izan nahi du.*
es: {entity} quiere ser {occupation}.
→ *Mikel quiere ser enfermero*
en: {entity} wants to be a {occupation}.
→ *Mikel wants to be a nurse.*

SINGLE-ENTITY TEMPLATE
eu: Nire lagun {entity} {occupation}a zela esan nizunean haserratu egin zinen.
→ *Nire lagun Ainara errementaria zela esan nizunean haserratu egin zinen.*
es: Cuando te dije que mi amigo|a {entity} era {occupation} te enfadaste.
→ *Cuando te dije que mi amiga Ainara era herrera te enfadaste.*
en: When I told you my friend {entity} was a {occupation} you got angry.
→ *When I told you my friend Ainara was a blacksmith you got angry.*

MULTI-ENTITY TEMPLATE
eu: {entity}k {occupation} izatea gustuko du, baina {entity2}k {occupation2} izatea gorroto du.
→ *Mikelek erizain izatea gustuko du, baina Ainarak errementari izatea gorroto du.*
es: A {entity} le gusta ser {occupation}, pero {entity2} odia ser {occupation2}.
→ *A Mikel le gusta ser enfermero, pero Ainara odia ser herrera.*
en: {entity} loves being a {occupation} while {entity2} hates being a {occupation2}.
→ *Mikel loves being a nurse while Ainara hates being a blacksmith.*

Table 1: Examples of the handcrafted templates for the gender agreement task between gendered entities and occupations. Along with the templates we provide an injected example and the corresponding English translation.

of gender agreement between persons and occupations. For example:

Elisabeth Rynell *es una* ***escritora*** *sueca que ha incursionado principalmente en los géneros de la novela y poesía.*[7]

Finally, the extracted examples were automatically back-translated to Basque with the baseline system described in Section 4. This process guarantees gender agreement is not altered during the translation process as Basque is a genderless language.

We obtained a final set of approximately 42,000 examples per gender with the required gender agreement (*Wiki_train*).

In addition, in order to have an in-domain test from Wikipedia, we manually created a disjoint test set by selecting 100 examples for each gender. In this case, translations were manually corrected to ensure their final quality. Hereinafter referred to as *Wiki_test* test set.

4 Experimentation

All the systems use the default configuration for the Transformer architecture (Vaswani et al., 2017) as implemented in the PyTorch version of the OpenNMT toolkit (Klein et al., 2017). We apply BPE tokenization (Sennrich et al., 2016) trained on 32,000 operations on the joint training data. Sentences larger than 100 tokens are discarded from the training set.

The baseline systems were trained on the Basque-Spanish portion (1.77M examples) of the Paracrawl (v8) data (Bañón et al., 2020). The gender-balanced systems are trained by fine-tuning the baseline system on the gender-balanced data sets described in Section 3. As in Costa-jussà and de Jorge (2020), to avoid catastrophic forgetting, where systems tend to forget about previous knowledge, we follow a mixed fine-tuning strategy (Chu et al., 2017). A weighted combination (10:1 ratio[8]) of general domain data from Paracrawl and task specific data, such as *Templ_train* or *Wiki_train*, is used during training and validation steps. For

[7] English translation: *Elisabeth Rynell is a Swedish writer who has mainly dabbled in the genres of novels and poetry.*

[8] Initial experiments showed that 10:1 ratio for general domain and task specific data respectively works well.

validation purposes, we concatenate 5,000 general domain examples and 1,000 task specific examples randomly extracted from the training data.

The baseline and the fine-tuned systems have been trained until convergence on the perplexity results on the validation set, stopping the training process if there was no improvement for 5 consecutive checkpoints. Validation is performed every 10,000 steps in the case of the baseline system whereas fine-tuning validation is performed every 1,000 steps.

We evaluate our systems using **BLEU** and **chrF++** scores from the sacreBLEU tool (Post, 2018). Additionally, we also provide **COMET** (Rei et al., 2020) scores[9], a metric which focuses on the semantic similarity by leveraging the recent breakthroughs in neural language modeling. These scores are computed on the test sets extracted from three publicly available corpora: **EiTB** (Etchegoyhen and Gete, 2020) a news domain data set, **EhuHac** (Sarasola et al., 2015) a collection of classic books, and **TED** (Reimers and Gurevych, 2020) comprising TED talks transcriptions. From each set, we randomly extracted 5,000 examples. Although BLEU, chrF++ and COMET metrics measure the overall translation quality of the systems, task specific metrics are required to evaluate gender bias with more precision. To that end, we measure the **accuracy** of the correctly translated and gender inflected occupations and we propose a new metric called **swap**. Swap is defined as the percentage of the occupations which are inflected with the opposite gender. Thus, errors are divided into unrecoverable errors where occupations are translated in a different way (errors) and gender swapped occupations (swap). A higher swap score means higher bias towards the opposite gender. These scores are computed on the task specific test sets mentioned in Section 3.

4.1 Gender bias assessment

We conducted a detailed experimentation to determine the source of gender bias in the agreement between gendered entities and occupations. We analyze four different strategies to inject different subsets of the gendered entities and occupations lists in the training templates in order to generate gender-balanced data:

- **Full** system uses all the available gendered entities and occupations, both training and test subsets, to inflect training templates. Therefore, the test subsets of the gendered entities and occupations are seen during training. We produce 772,896 training examples.

- **Unknown entities (Unk_ent)** system only uses the training subset of the entities to inflect training templates. There is no overlap between the entities used for fine-tuning the system and those for the test set. 693,880 training examples are produced.

- **Unknown occupations (Unk_occ)** system only uses the training subset of the occupations to inflect training templates, resulting in 633,216 training examples.

- **Unknown pairs (Unk_prs)** system only inflects templates with disjoint combinations of entities and occupations. For instance, if *Mikel-doctor* is present in the test, *Mikel-plumber* and *Jon-doctor* are seen during training. A total of 758,616 training examples are produced.

We remark that, in all the cases, training (27) and test (6) template sets are disjoint, and only single entity templates are used for fine-tuning.

In general terms, all the fine-tuned systems on gender-balanced data keep the baseline's translation quality on the general domains test sets (see Table 2). Specially, **Full** system performs at par with the baseline across all the test sets and metrics, except for the chrF++ score on the EhuHac test set.

In order to analyze bias effects, in Table 3 we report gender bias accuracy and swap scores. The baseline model shows a clear bias towards the masculine inflection of the occupations with significantly higher swap scores and lower accuracy scores for females. We note that a negative value for the swap difference means there exists masculine bias. Lower swap scores are obtained with the baseline system on the *Wiki_test*, suggesting stereotyped occupations from Wikipedia (Wagner et al., 2015) are being well inflected by the baseline.

Fine-tuning the baseline system on the **Full** set significantly drops the swap score on the *Templ_test* which indicates the fine-tuned system is able to correctly inflect the gender of the occupations for seen entities. In contrast, **Unk_ent** system shows higher swap scores. Despite the system is clearly less biased than the baseline, it is having

[9]The recommended model *wmt20-comet-da* was used and it already covers both Basque and Spanish.

System	EiTB			EhuHac			TED		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Baseline	**37.9**	**57.8**	**0.732**	**14.1**	**37.0**	**-0.149**	**24.2**	**48.6**	**0.462**
Full	**37.8**	**57.9**	**0.730**	**14.0**	36.8*	**-0.153**	**24.1**	**48.6**	**0.458**
Unk_ent	**37.8**	57.7*	0.725*	**14.0**	36.9*	**-0.154**	**24.1**	**48.6**	0.456*
Unk_occ	37.8*	**57.8**	0.725*	**14.0**	36.8*	**-0.152**	24.0*	48.5*	0.456*
Unk_prs	37.7*	57.7*	**0.727**	**14.0**	**36.9**	**-0.152**	**24.0**	**48.6**	0.456*

Table 2: BLEU, chrF++ and COMET scores for systems fine-tuned on gender-balanced data. * indicates statistically significant (p-value ≤ 0.05) differences by conducting paired bootstrap resampling with respect to the baseline. Best scoring systems are highlighted in bold.

System	Test	Male		Female		$\bar{X}$ Swap	Δ Swap
		Acc	Swap	Acc	Swap		
Baseline	Templ_test	59.23	0.77	14.04	47.31	24.04	-46.54
	Multi_test	61.78	**6.17**	17.72	50.67	28.42	-44.50
	Wiki_test	**95.15**	**0.61**	65.82	25.95	13.00	-25.34
Full	Templ_test	**98.27**	0.96	**96.15**	**2.82**	**1.44**	**-1.86**
	Multi_test	76.22	23.72	**83.39**	16.61	20.17	7.11
	Wiki_test	93.94	1.21	67.09	24.68	12.69	-23.47
Unk_ent	Templ_test	93.72	5.51	83.33	15.71	10.61	-10.2
	Multi_test	67.39	32.56	76.83	23.11	27.83	9.45
	Wiki_test	93.33	**0.61**	67.09	**24.05**	**12.07**	-23.44
Unk_occ	Templ_test	62.95	**0.45**	58.91	6.41	3.43	-5.96
	Multi_test	56.28	12.56	55.61	**15.06**	**13.81**	**-2.50**
	Wiki_test	93.33	1.21	67.09	24.68	12.69	-23.47
Unk_prs	Templ_test	97.95	0.96	94.10	5.00	2.98	-4.04
	Multi_test	**76.56**	23.22	81.00	18.94	21.08	4.28
	Wiki_test	92.73	1.21	**67.72**	**24.05**	12.38	**-22.84**

Table 3: Accuracy and swap scores for systems fine-tuned on gender-balanced data. We report mean swap scores and swap differences for a better picture of the bias. Best scoring systems are highlighted in bold.

more difficulties inferring the gender of unseen entities. This behaviour is further corroborated on the *Wiki_test* as all the experiments show similar (slightly better) results of those obtained by the baseline system. A manual inspection of the translations showed that most of the swap errors are associated to foreign person names. In such cases, systems tend to provide the default masculine.

With respect to the lower accuracy scores in **Unk_occ**, most of the errors made were the result of the system not being able to produce the correct translation. Most of the time it produces correct but alternative translations for the given occupations such as *lechero/vendedor de leche (milkman)*. In any case, we remark that in these cases the correct gender is generally inflected: *lechera/vendedora de leche (milkwoman)* or *camarógrafo/el cámara (cameraman)*. This behavior blurs swap and accuracy results as the bias can not be automatically computed for those occupations.

Remarkably, accuracy and swap scores in **Unk_prs** obtains comparable enough results of those obtained in **Full**, which suggests the system does not require to see all the possible combinations of entities and occupations during training. Instead, in the light of the results obtained by the **Unk_ent** and **Unk_occ** systems, providing the system with all the gendered entities and occupations is more relevant than producing all their combinations.

In general terms, we can conclude that fine-tuning a pre-trained system with a mixed combination of gender-balanced examples and general domain data is useful to mitigate gender bias from NMT systems without a substantial drop in general domain translation quality.

Finally, we note that all the experiments show higher swap scores on the *Multi_test* test. A man-

ual inspection of the translations suggests that, as stated in (Saunders et al., 2020), systems simply learn to indiscriminately apply the same gender inflection to all the occupations when presented with multi-entity templates. This issue is addressed in Section 4.2.

4.2 Gender tagging entities

From the previous bias assessment section, we conclude that gender disambiguation for unknown entities is not obvious for the systems. Yet, it is essential to correctly inflect occupations with the corresponding gender. Therefore, in the line of the previous work by Saunders et al. (2020) and Stafanovičs et al. (2020), we propose using word level annotations to provide a stronger gender signal for gender disambiguation of gendered entities. (Stafanovičs et al., 2020) annotate all the source words with the grammatical gender information of their corresponding target words while (Saunders et al., 2020) add explicit word-level gender tags to the occupations that need to be inflected. In contrast to these methods, we only apply gender annotations to gendered entities and leave sentence-level gender agreement to the system. The main advantage of this approach is that it does not require any complex annotation step. During inference, a list of proper names and other gendered entities can be used to properly annotate the entities. This list can be dynamically updated without fine-tuning the whole system again.

We annotate each word in the source side of the **Full** set via source factors (Sennrich and Haddow, 2016) with three possible values (**Full_tag**): 1 for male entities, 2 for female entities and 0 for the rest of the words. For example,

Mikelek erizain izan nahi du.[10] → 1 0 0 0 0

These tags are then appropriately mapped to their corresponding subword tokens during the fine-tuning step.

Additionally, as noted in Saunders et al. (2020) and corroborated in Section 4.1, in cases where multiple entities are present in a sentence, systems tend to overgeneralize gender signals by applying the same gender to all the occupations. Accordingly, we add the *Multi_train* set (see Section 3.1) during the fine-tuning step to help the **Full_tag** system better handle these cases.

Likewise, we follow the same gender tagging strategy on the Wikipedia biographies set (*Wiki_train*), described in Section 3.2, to assess whether using real data extracted from Wikipedia is a feasible approach. We remark that due to the characteristics of the biographies it is not possible to extract multiple entity examples. We fine-tune the baseline system with (**Wiki_tag**) and without gender tags (**Wiki**).

Overall, all the systems keep the baseline's translation quality in terms of BLEU, chrF++ and COMET for the general domain test sets, either fine-tuned with templates or with real Wikipedia data (see Table 4).

Moreover, Table 5 shows gender bias accuracy and swap scores for the gender tagged systems along with their untagged version. **Full_tag** considerably outperforms **Full**. Despite the swap difference on the *Templ_test* is slightly higher for **Full_tag**, the total swap score is lower. Adding unambiguous gender tags to the entities provides a stronger signal that helps reducing gender bias from the system. We report a substantial improvement on the *Multi_test*, which further encourages the use of a stronger gender signal via gender tags. Remarkably, **Full_tag** obtains perfect scores on the *Multi_test*, showing that providing multi-entity templates during training helps mitigating the gender signal overgeneralization issue.

Fine-tuning on *Wiki_train* also helps improving the baseline system and adding gender tags further improves those results. As expected, **Wiki_tag** obtains the best bias reduction results on the in-domain *Wiki_test*, as most of the occupations overlap between training and test data. Notice that the Wikipedia occupations set is potentially small and closed. As reported by Costa-jussà and de Jorge (2020) using real gender-balanced data instead of manually created templates can also contribute to reduce bias, although the results we achieved are not as good as those obtained with the template-based version. However, we note that systems fine-tuned on *Wiki_train* still perform poorly when multiple entities are present, showing a clear tendency towards masculine overgeneralization.

4.3 Templates ablation study

In this section we report the results of the template ablation experiments conducted to determine the optimal amount of templates needed in order to reduce the manual effort to build them. We focused on their complexity too, as generating simpler templates might be easier without a strong knowledge

[10] English translation: Mikel wants to be a nurse.

System	EiTB			EhuHac			TED		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Baseline	**37.9**	**57.8**	**0.732**	14.1	**37.0**	**-0.149**	**24.2**	**48.6**	**0.462**
Full	**37.8**	**57.9**	**0.730**	14.0	36.9*	**-0.153**	**24.1**	**48.6**	**0.458**
Full_tag	**37.8**	**57.8**	**0.729**	14.0	36.9*	**-0.150**	**24.1**	**48.6**	**0.457**
Wiki	**38.0**	**57.9**	**0.733**	**14.2***	**37.0**	**-0.143**	**24.1**	**48.6**	**0.461**
Wiki_tag	**37.9**	**57.8**	**0.731**	14.1	**36.9**	**-0.147**	**24.1**	**48.6**	**0.461**

Table 4: BLEU, chrF++ and COMET scores for the gender tagging systems compared to their untagged versions. * indicates statistically significant (p-value ≤ 0.05) differences by conducting paired bootstrap resampling with respect to the baseline. Best scoring systems are highlighted in bold.

System	Test	Male		Female		$\bar{X}$ Swap	Δ Swap
		Acc	Swap	Acc	Swap		
Baseline	Templ_test	59.23	0.77	14.04	47.31	24.04	-46.54
	Multi_test	61.78	6.17	17.72	50.67	28.42	-44.5
	Wiki_test	95.15	0.61	65.82	25.95	13.00	-25.34
Full	Templ_test	98.27	0.96	96.15	2.82	1.44	**-1.86**
	Multi_test	76.22	23.72	83.39	16.61	20.17	7.11
	Wiki_test	93.94	1.21	67.09	24.68	12.69	-23.47
Full_tag	Templ_test	**99.49**	**0.06**	**96.86**	**2.12**	**1.09**	-2.06
	Multi_test	**100.00**	**0.00**	**100.00**	**0.00**	**0.00**	**0.00**
	Wiki_test	95.15	**0.00**	84.81	5.70	2.79	-5.70
Wiki	Templ_test	57.37	2.24	20.38	40.13	21.19	-37.89
	Multi_test	60.11	7.00	19.78	49.22	28.11	-42.22
	Wiki_test	**95.76**	**0.00**	76.58	15.82	7.74	-15.82
Wiki_tag	Templ_test	62.95	0.32	37.50	25.32	12.82	-25.00
	Multi_test	60.39	4.56	17.67	50.00	27.28	-45.44
	Wiki_test	**95.76**	**0.00**	**92.41**	**1.90**	**0.93**	**-1.90**

Table 5: Accuracy and swap scores for the gender tagged systems compared to their untagged versions on the task specific test sets. We report mean swap scores and swap differences for a better picture of the bias. Best scoring systems are highlighted in bold.

about the language.

We sorted all the training templates, both single entity templates and multi-entity templates, according to their complexity. Word counts are used as an indicator of their complexity. We wanted to analyze the simplest scenario with just one single entity template and one multi-entity template (**1_1**), as this is the case in (Saunders and Byrne, 2020). Additionally, we analyzed scenarios with different numbers of single entity templates and multi-entity templates, hereinafter referred to as **2_2**, **5_5**, **10_10**, **20_10**[11]. To analyze the influence of the complexity, for all the combinations we produced a simple version (**S**), which comprises the less complex templates and a complex version (**C**) including the most complex ones. All these ablation experiments were compared against the baseline system and the **Full_tag** system fine-tuned with all the handcrafted templates possible (27 single entity and 10 multi-entity templates). All the experiments use gender tags.

In general terms, all the systems comply with the requirement of keeping the baseline's translation quality for the general domain test sets (see Table 6). We therefore focus on the task specific metrics as shown in Figure 2.

All the systems, even for **S_1_1**, significantly improve swap scores when compared to the baseline. Mean swap curves show a clear descending trend which suggests that having a more syntactically diverse set helps generalizing gender signals. Remarkably, from **S_10_10** and **C_10_10** systems on, the curve tends to converge, showing little improvement with more templates. This is an interesting

[11]Names indicate the number of single entity and multi-entity templates respectively

System	EiTB			EhuHac			TED		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Baseline	**37.7**	**57.8**	**0.732**	**14.1**	**37.0**	**-0.149**	**24.2**	**48.6**	**0.462**
S_1_1	**37.8**	**57.8**	**0.728**	**14.0**	36.9*	**-0.148**	23.9*	48.5*	**0.458**
C_1_1	37.7*	**57.7**	**0.728**	13.9*	36.8*	**-0.154**	**24.1**	**48.7**	**0.462**
S_2_2	**37.8**	**57.8**	**0.727**	**14.0**	**36.9**	**-0.152**	**24.0**	48.5*	**0.459**
C_2_2	37.7*	57.7*	**0.727**	**14.0**	36.8*	**-0.155**	23.9*	48.5*	0.456*
S_5_5	37.7*	57.7*	0.725*	**14.0**	**36.9**	**-0.152**	24.0*	48.5*	0.456*
C_5_5	**37.8**	**57.8**	**0.731**	**14.0**	36.8*	**-0.149**	23.9*	**48.4**	0.454*
S_10_10	**37.8**	**57.8**	**0.727**	**14.0**	**36.9**	**-0.153**	**24.2**	**48.7**	**0.461**
C_10_10	**37.8**	**57.8**	**0.729**	**14.0**	36.8*	**-0.152**	**24.0**	48.5*	0.456*
S_20_10	**37.8**	**57.7**	**0.727**	**14.0**	36.8*	**-0.153**	24.0*	**48.6**	**0.461**
C_20_10	**37.8**	57.7*	**0.726**	**14.0**	36.8*	**-0.150**	23.9*	48.5*	0.455*
Full_tag	**37.8**	**57.8**	**0.729**	**14.0**	36.9*	**-0.150**	**24.1**	**48.6**	**0.457**

Table 6: BLEU, chrF++ and COMET scores for the template ablation experiments. * indicates statistically significant (p-value ≤ 0.05) differences by conducting paired bootstrap resampling with respect to the baseline. Best scoring systems are highlighted in bold.

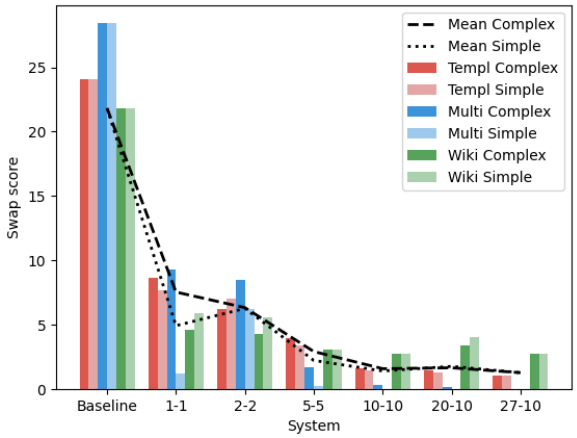

Figure 2: Total swap scores for the template ablation experiments. Dashed and doted lines represent the mean swap values for the three test sets (*Templ_test*, *Multi_test* and *Multi_test*).

insight, as there seems to be a limit in the amount of templates, which considerably reduces the manual effort to create templates.

In general terms, systems fine-tuned on simpler templates perform at par or even better than the ones trained on more complex templates. This indicates that generating complex and syntactically rich templates is not worth the effort. Also, results suggest that multi-entity templates present a strong signal which solves the overgeneralization issue for systems with 10 or more templates. Thus, it emphasizes the hypothesis that templates can be easily adapted to small specific tasks with little effort.

System	Templ	Multi	Multi
S_1_1	7.72	1.22	5.88
S_10_10	**1.44**	**0.03**	**2.79**
S_10_10_limit	1.70	0.25	3.41

Table 7: Mean swap scores for the **S_10_10_limit** system compared against its complete version (**S_10_10**) and **S_1_1**. The system was fine-tuned on only 32,200 examples as **S_1_1**. Best scoring systems are highlighted in bold.

Finally, we tested whether having more templates improves the results because of the syntactically diverse templates or just because of the mere fact that more training examples are generated. To that end, as we observed some convergence with 10 templates, we fine-tuned the baseline on randomly selected 32,200 examples, namely **S_10_10_limit**, that is, the same amount of templates used in **S_1_1** (Table 7). **S_10_10_limit** performs slightly worse than **S_10_10** and it clearly outperforms **S_1_1**. This suggests that the improvement comes to a greater extent from syntactic diversity rather than from a higher amount of templates.

5 Conclusions

In this work we addressed gender bias mitigation from an already pre-trained system. In particular, we focused on the specific task of sentence-level gender agreement between gendered entities and occupations when translating from genderless languages to gendered languages.

The proposed template-based fine-tuning strategy with explicit gender tags helps mitigating gender bias from NMT systems. We proved that the mixed fine-tuning strategy using a weighted combination of general domain and task specific data is beneficial to overcome catastrophic forgetting and keep the original translation quality.

We demonstrated that adding explicit gender tags to gendered entities provides a stronger gender signal and helps the system to gender inflect occupations correctly. At inference, entities can be easily annotated by using a list of proper names and other gendered entities, which can be dynamically updated without fine-tuning the system again.

Our results on the Basque to Spanish translation direction showed substantial bias mitigation and confirmed that handcrafted templates are suitable to create task specific training examples, to the point of improving the results obtained by using gender-balanced real examples extracted from Wikipedia. The ablation study showed that with little manual effort a set of useful templates can be created for gender bias mitigation. Therefore, the proposed method can be applied to other language pairs.

Limitations

The ablation study in Section 4.3 showed that with little manual effort a set of useful templates could be created for gender bias mitigation. In this sense, the proposed method still requires some linguistic knowledge about the languages involved in order to manually create the templates. Some of the entities and the occupations list should be adapted to the new language pair too. We acknowledge that this requirement can be a limiting factor for a massive deployment of our method. However, we believe that some challenges in NMT require a prior linguistic knowledge of the languages at hand in order to detect the possible errors and flaws and to provide a solution or mitigation response.

Furthermore, our work focuses on the Basque to Spanish translation direction as an example of the translation direction from a genderless language to a language with explicit grammatical gender. Although, the proposed gender tagging method does not rely on Basque or Spanish exclusive linguistic features, we believe that adding additional language pairs would have shown a broader picture of our method's potential. We leave this discussion for future work where different language families could be analyzed.

Finally, it must be noted that we approach the task using a binary gender representation schema. This decision should not be interpreted as a denial of a more complex reality.

Acknowledgments

This work has been partially funded by the Basque Government (TANPER, Hazitek grant no. ZL-2021/00904)

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4555–4567, Online. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In Proceedings of the The Fourth Widening Natural Language Processing Workshop, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 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, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

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.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short

Papers), pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Marta R Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2019. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. arXiv preprint arXiv:1912.04778.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Thierry Etchegoyhen and Harritxu Gete. 2020. Handle with care: A case study in comparable corpora exploitation for neural machine translation. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 3792–3800. European Language Resources Association.

Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1686–1690.

Melvin Johnson. 2018. Providing gender-specific translations in google translate. Accessed: 2022-02-17.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. 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, Florence, Italy. Association for Computational Linguistics.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the basque wordnet. Language resources and evaluation, 45(2):121–142.

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

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Ander Salaberria, Jon Ander Campos, Iker García, and Joseba Fernandez de Landa. 2021. Itzulpen automatikoko sistemen analisia: Genero alborapenaren kasua. In Fourth Conference for Basque Researchers.

Ibon Sarasola, Pello Salaburu, and Josu Landa. 2015. Hizkuntzen arteko corpusa (hac).

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. Transactions of the Association for Computational Linguistics, 9:845–874.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In Proceedings of the Fifth Conference on Machine Translation, pages 629–638, Online. Association for Computational Linguistics.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In Proceedings of the international AAAI conference on web and social media, volume 9, pages 454–463.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

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