Multi-Task Learning for Improving Gender Accuracy in Neural Machine Translation

Carlos Escolano[1] and **Graciela O. Jiménez**[1] and **Christine Basta**[1 2] and **Marta R. Costa-jussà**[1]
[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
[2] Institute of Graduate Studies and Research, Alexandria University, Egypt
{carlos.escolano,christine.raouf.saad.basta,marta.ruizs}@upc.edu
{graciela.ojeda}@estudiantat.upc.edu

Abstract

Machine Translation is highly impacted by social biases present in data sets, indicating that it reflects and amplifies stereotypes. In this work, we study mitigating gender bias by jointly learning the translation, the part-of-speech, and the gender of the target language with different morphological complexity. This approach has shown improvements up to 6.8 points in gender accuracy without significantly impacting the translation quality.

1 Introduction

In recent years, the awareness about the bias present in Machine Translation (MT) systems has increased in the scientific community, especially gender bias. Gender is manifested differently in languages; gender bias problems occur when translating between languages with +various levels of morphology. There is bias when the system tends to translate according to gender roles, even when there is no ambiguity (Prates et al., 2020).

Surveys (Sun et al., 2019; Blodgett et al., 2020; Costa-jussà, 2019; Savoldi et al., 2021) have recently shown the great efforts carried out by scientists towards resolving the problem of gender bias in NMT. Tagging and additional context approaches (Vanmassenhove et al., 2018; Moryossef et al., 2019; Basta et al., 2020) have shown an improvement in translation accuracy when translating from English to languages with more complex morphology. Domain adaptation techniques have proved to impact the performance of translation in (Saunders and Byrne, 2020). Debiased pre-trained word embeddings have been leveraged in (Escudé Font and Costa-jussà, 2019) and have shown improvement in Spanish translations. Gender bias is mainly attributed to the already present bias in the data used to train MT systems (Savoldi et al., 2021; Costa-jussà, 2019). Furthermore, it has been shown that models trained on these data tend to amplify further this bias (Zhao et al., 2018). In this sense, research done aims to avoid or reduce this amplification, such as fine-tuning techniques with gender-balanced dataset corpus (Costa-jussà and de Jorge, 2020) or annotating the source language words of the training data with the gender of the target language words (Stafanovičs et al., 2020). These techniques have shown promising results regarding gender accuracy.

In line with reducing the amplification of data, we propose to obtain a system capable of predicting the part-of-speech (pos) and the gender of the words of the target language, besides the translation. We expect that, by having more information about the output words, the system maintains the quality of the translation and is able to better predict the gender based on the context without falling so much into the stereotype. In general, the proposed configurations outperformed the baseline NMT system on gender prediction accuracy by up to 6.8% while retaining average translation performance.

2 Bias statement

Nowadays, we live in a more globalized and connected world, which leads society to use MT tools to communicate with different nationalities. The fact that standard translators present a gender bias harms society, helping to perpetuate certain stereotypes and prejudices. An example is the tendency of specific systems to generalize the different professions carried out by men and women (Prates et al., 2020). It is significantly more visible when one tries to translate from a gender-neutral language, such as English, to another with grammatical gender, such as Spanish. In the first type, nouns have no grammatical gender, while in the other type, there is gender inflection in nouns, adjectives, verbs, etc. Therefore, when translating from

Proceedings of the 18th International Conference on Natural Language Processing, pages 12–17
Silchar, India. December 16 - 19, 2021.

English to Spanish, the system uses masculine or feminine inflections following stereotypes. Such stereotypical errors occur both when the context indicates gender explicitly and when it does not.

3 Background

In this section, we review the basics of multilinguality and multi-task learning in NMT, as well, as linguistic features and the framework to evaluate gender bias in MT. All these methodologies are later used in our proposed research.

Multilingual NMT. Transformers (Vaswani et al., 2017) have advanced NMT, giving the ability to pay more attention to multilingual NMT. The primary approach behind multilingual is to have the same model architecture to translate different language pairs (Firat et al., 2016; Johnson et al., 2017). Previous studies explore different design approaches for the model architecture, either partial sharing with shared encoder (Sen et al., 2019), shared attention (Firat et al., 2016), task-specific attention (Blackwood et al., 2018), shared parameters (Zhu et al., 2020), full model sharing (Johnson et al., 2017) or independent encoder-decoders without sharing (Escolano et al., 2021; Lu et al., 2018). In this paper, we adopt this architecture without sharing.

Multi-task learning NMT. Multi-task learning (Caruana, 1997) trains the model on several co-related tasks. This training can lead to generalized improved performance and facilitate sharing representations (Ruder, 2017). In the NMT context, injecting linguistic knowledge has been successful when training NMT with related tasks (POS tagging, dependency parsing). This linguistic injection can lead to improving NMT generalization and translation quality, especially in low-resource scenarios (Kiperwasser and Ballesteros, 2018; Zaremoodi and Haffari, 2018; Eriguchi et al., 2017). Our work depends on adopting linguistic knowledge through training multi-tasks, besides NMT. We trained our NMT model with POS prediction and gender tagging tasks.

Linguistic Features. Different linguistic features can be utilized for words' classification, such as part-of-speech (POS) in morphology. POS refers to the lexical category of words, defining different linguistic categories depending on the shared morphological categories between these words. Universal Dependencies (UD)[1] is a framework for consistent grammar annotation across different human languages (Nivre et al., 2016). It considers a fixed list of 17 possible POS tags, e.g., noun, verb, adjective, adverb, and preposition. Furthermore, each word has morphological features, such as person, number, and gender. According to UD, depending on the language, there are four different possibilities for the gender of a word; feminine, masculine, neuter, and common (non-neuter). In this paper, we focus on the gender morphological feature of the words.

Gender Bias Analysis Framework. WinoMT (Stanovsky et al., 2019) is the first challenge test set used to evaluate gender bias in MT systems. The test set consists of 3888 sentences; 1826 male sentences, 1822 female sentences, and 240 neutral sentences. It is also distributed with 1584 anti-stereotype sentences, 1584 pro-stereotype sentences, and 720 neutral sentences. Each sentence contains two personal entities where one is a coreferent to a pronoun, and a golden gender is specified for this entity. Three metrics are used for assessment: accuracy (Acc.), which is measured by comparing the translated entity with the golden entity, ΔG and ΔS. ΔG is the difference between the correctly inflected masculine and feminine entities. ΔS is the difference between the inflected genders of the pro-stereotype and anti-stereotype entities. (Saunders and Byrne, 2020) also propose **M:F**, which is the ratio of hypotheses with masculine predictions to those with feminine predictions. ΔS can be skewed in low-accuracy systems; thus, **M:F** would be easier to interpret. Ideally, the absolute values of ΔS and ΔG should be closer to 0, and **M:F** should be closer to 1.

4 Proposed methodology

Previous work (Costa-jussà et al., 2020) has shown that the language-specific multilingual NMT architecture proposed by (Escolano et al., 2021) outperformed the universal shared encoder-decoder architecture (Johnson et al., 2017) on gender bias evaluations. We chose this language-specific architecture as the baseline for all our experiments.

Given parallel data for a set of languages $L = l_1, l_2, ..l_n$ with n languages and data for language pairs, the architecture consists of n encoders and n decoders, each of them specific for a single lan-

[1] https://universaldependencies.org/

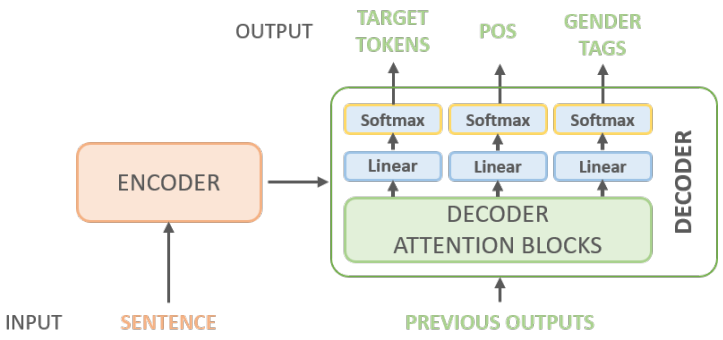

Figure 1: The model architecture, each task has its linear layer, softmax, and loss function

guage. Being l_i encoder used for all translation directions involving l_i as source language and l_i decoder for all directions involving l_i as target language. Using the same encoders and decoders on several translation directions enforces a common intermediate representation for all encoders in the system.

This method relies on parallel data only, without any additional linguistic information. Our proposed modification to this architecture (see Figure 1) adds a new linear projection layers to the decoders for each additional tagging task $tag_k \in T$. Each task focuses on different linguistic aspects that may improve gender representation. It is known that multitasking allows models, by inductive bias (Ruder, 2019), to learn a representation that contains features useful for all the involved tasks, improving its generalization capabilities. For each translation direction, the loss is computed as follows:

$$L(t, y', y) = L_{tr}(y', y) + \sum_{k=1}^{K} L_{tag}(t_k, tag_k(y)) \quad (1)$$

Where y' are the translation logits, y is the reference target sentence, L_{tr} is the cross-entropy loss for the translation task, t is the set of tagging logits for each of the K tagging tasks, L_{tag} is the cross-entropy loss over each tagging task and $tag_k(y)$ is the tagging function over the reference target.

5 Experimental Framework

In this section, we describe the details about the data and model parameters involved in our experiments.

5.1 Data and preprocessing

We have used *Europarl* dataset as training data between all 12 possible language pairs between Spanish, German, English, and French. For each pair, approximately 2 million sentences were available, for a total amount of 24 million. As validation and evaluation data for NMT results, newstest2012 and newstest2013 (Bojar et al., 2013). Only results with English as the source are provided to match the Gender Bias evaluation framework. All data has been preprocessed by applying tokenization, punctuation normalization, and true-casing using standard *Moses* (Koehn et al., 2007) scripts and tokenized at BPE subword level (Sennrich et al., 2016) with 32 thousand steps using subword-nmt framework[2].

Linguistic features have been extracted using the *Stanza* framework (Qi et al., 2020) at word level. For split words, a tag is repeated. For Gender bias evaluation, WinoMT dataset has been used. All data has been preprocessed following the same pipeline depicted for NMT.

5.2 Model Parameters

All models are implemented on *fairseq*'s (Ott et al., 2019) [3] Transformer (Vaswani et al., 2017) with 6 attention layers on both encoder and decoder, 512 embedding size, 8 attention heads and 2048 feed-forward size. Each translation direction has approximately 60 million parameters. POS tag and Gender predictions only account for 8704 and 2048 additional parameters compared to the baseline system. All models have been trained using no further improvement of the validation loss as an early stopping criterion.

[2] https://github.com/rsennrich/subword-nmt

[3] https://github.com/pytorch/fairseq version 0.6

	en-es	en-fr	en-de	es-en	es-fr	es-de	fr-en	fr-es	fr-de	de-en	de-es	de-fr	Avg
Baseline	**29.48**	**29.56**	21.5	**27.38**	**30.25**	19.61	26.03	29.04	18.96	24.04	**24.95**	**25.24**	**25.50**
+POS	29.06	29.27	21.4	27.16	29.79	19.43	25.92	28.89	18.91	**24.24**	24.84	24.99	25.32
+Gender	29.38	**29.56**	**21.81**	27.11	30.05	**19.84**	26.13	29.09	18.99	24.07	24.86	25.16	**25.50**
+POS&Gender	29.12	29.37	21.59	26.92	29.9	19.48	**26.54**	**29.2**	**19.24**	24.23	24.91	24.96	25.45

Table 1: Results in terms of BLEU for different language pairs for baseline, the baseline trained with POS tagging task, the baseline trained with gender tagging task and the baseline trained with aforementioned tasks together.

	Spanish				French				German			
	Acc↑	ΔG↓	ΔS↓	M:F↓	Acc↑	ΔG↓	ΔS↓	M:F↓	Acc↑	ΔG↓	ΔS↓	M:F↓
Baseline	57.7	15.1	18.5	2.85	48.8	23.6	9.5	3.99	62.7	9.9	12.4	2.3
+POS	63.7	7.8	15.5	2.27	51.8	17.5	7.4	3.13	**68.3**	**3.5**	8.8	**1.749**
+Gender	58.2	15.2	**7.2**	2.97	49	21.6	**3.8**	3.52	61.3	8.8	**6.4**	2.06
+POS&Gender	**64.5**	**7.1**	13.3	**2.16**	**54.8**	**14.3**	13.8	**2.8**	65.3	5.5	8	1.95

Table 2: WinoMT Results for the three languages Spanish, French and German

6 Results

In this section, we explore the improvements achieved by multi-task training, whether regarding the general translation accuracy or the gendered results.

Translation results. Table 1 show the translation performance of all proposed configurations. Looking at their average performance, we observe that the addition of tagging tasks does not significantly impact the average performance, with a difference of less than 0.2 between all systems.

When looking at individual directions, we can see that training gender tagging task with NMT in English-to-German has improved. We can argue that German is a higher morphological language than English, and the gender tagging task helps the system inject more knowledge about gender in German, leading to better translation accuracy, up to 0.31 for the English-to-German pair compared to the baseline. Training more than one task seems to be beneficial when the translation is between high morphological language pairs like French-to-Spanish and French-to-German where French, German, and Spanish are all gendered high morphological languages. French-to-English seems to also benefit from the POS and Gender tagging tasks together.

WinoMT results. WinoMT helps us investigate how the gender-biased entities and professions translated. The framework investigates the translations when English is translated to higher morphological languages; therefore, we show the WinoMT results from English to Spanish, German and French.

Spanish seems to benefit from the POS and Gender tagging tasks together, where the accuracy increased by 6.8 over the baseline with a lower difference between the correct translated masculine entities and the correct translated feminine entities; of ΔG 7.1. This is assured by the lower ratio of male vs. female predictions (**M:F**) of 2.16. French also has better accuracy having both tasks trained together, with 54.8 accuracy, which increases by six over the baseline. The ΔG is also lower in this case, with a value of 14.3. In both languages, the difference between stereotyped and non-stereotyped translations did not improve. The improvements are more related to the general accuracy.

Multitask training of gender seems to impact the stereotyped translations in the three languages; however, the general accuracy was not impacted that much by training the gender tagging task.

POS tagging also appears to help disambiguation of gender from English to German, giving higher gender accuracy reaching 68.3 with a low difference of male and female correct predictions; ΔG of 3.5 and **M:F** of 1.749.

7 Conclusions

In this paper, we have proposed and analyzed the use of multi-task learning in multilingual NMT. Learning linguistic tagging simultaneously as multilingual helps mitigate gender bias while maintaining the average translation performance over the tested languages. More than the methodology that we are proposing, which is simple and effective, we would like to encourage the community to evaluate their methodologies not only in terms of translation quality, but also in terms of social bias mitigation.

Acknowledgements

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

References

Christine Basta, Marta Ruiz Costa-Jussà, and José Adrián Rodríguez 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. Association for Computational Linguistics.

G. Blackwood, Miguel Ballesteros, and T. Ward. 2018. Multilingual neural machine translation with task-specific attention. *ArXiv*, abs/1806.03280.

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.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020. Gender bias in multilingual neural machine translation: The architecture matters. *CoRR*, abs/2012.13176.

Marta R. Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. 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, Florence, Italy. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association for Computational Linguistics*, 6:225–240.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. 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.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo,

Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Marcelo Prates, Pedro Avelar, and Luís Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem.

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

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. 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.

Gabriel Stanovsky, Noah Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. pages 1679–1684.

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.

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Poorya Zaremoodi and Gholamreza Haffari. 2018. Neural machine translation for bilingually scarce scenarios: a deep multi-task learning approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1356–1365, New Orleans, Louisiana. 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.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.