Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information

Christine Basta **Marta R. Costa-jussà** **José A. R. Fonollosa**
Universitat Politècnica de Catalunya
{christine.raouf.saad.basta,marta.ruiz,jose.fonollosa}@upc.edu

Abstract

Gender bias negatively impacts many natural language processing applications, including machine translation (MT). The motivation behind this work is to study whether recent proposed MT techniques are significantly contributing to attenuate biases in document-level and gender-balanced data. For the study, we consider approaches of adding the previous sentence and the speaker information, implemented in a decoder-based neural MT system. We show improvements both in translation quality (+1 BLEU point) as well as in gender bias mitigation on WinoMT (+5% accuracy).

1 Introduction

Gender bias is negatively affecting Natural Language Processing (NLP) (Costa-jussà, 2019; Sun et al., 2019). Gender bias clearly appears in word embeddings, associating certain neutral professions with males (*programmer*) and others with females (*housekeeper*) (Bolukbasi et al., 2016).

This bias has been demonstrated in Neural Machine Translation (NMT), where translations seem to ignore the context and translate professions with their stereotyped genders (Font and Costa-jussà, 2019; Stanovsky et al., 2019). This occurs due to the fact that NMT systems generally work on a sentence by sentence basis. Several approaches have been proposed to output different gendered translations (Kiritchenko and Mohammad, 2018), add gender information in the process of training (Vanmassenhove et al., 2018), and use debiased word embeddings (Font and Costa-jussà, 2019). Other approaches focused on measuring gender bias in translation systems (Prates et al., 2020; Stanovsky et al., 2019). Finally, the work by Costa-jussà et al. (2019) presented a non-synthetic gender-balanced data set, which can be considered to evaluate NMT.

The main contribution of this work is using existing NMT contextual methodologies, both context of the previous sentence (Junczys-Dowmunt, 2019) and speaker identification (Vanmassenhove et al., 2018), in a prominent and competitive NMT architecture (Fonollosa et al., 2019). These approaches are explicitly tested for the purpose of mitigating gender bias while improving the translation quality. The architecture in our experiments uses only the decoder part of the popular Transformer (Vaswani et al., 2017; He et al., 2018; Fonollosa et al., 2019); thus, reduces training parameters and simplifies the model.

2 Methodology: adding context and speaker id in a decoder-based NMT model

This study uses the following two recent proposed methodologies to improve the accuracy of NMT. While these methodologies are not new, we are adding them on a different baseline (Fonollosa et al., 2019) and testing specifically on a gender-balanced data sets. We describe the baseline system and the techniques as follows and examples are shown in Table 1.

Neural Machine Translation with joint source-target self-attention. The current state of the art is the encoder-decoder architecture using the Transformer (Vaswani et al., 2017) that avoids recurrence completely and gives better translations depending on the stacked self-attention and fully connected layers between encoder and decoder. An alternative to this architecture is based on the simplified architecture by Fonollosa et al. (2019) [1]. This model, instead of having both encoder-decoder, only uses the decoder block and it adopts the idea of language modeling for translation task. The joint source-target representations are learnt in the early layers. Positional embeddings are applied to the source and target

[1] https://github.com/jarfo/joint

independently. There are also language embeddings representing the language of the source and the target separately. Different from the self attention in normal transformers, a locally constrained attention is proposed by the authors to attend only to a token's locality, to form a reduced receptive field.

Adding the previous context sentence (PreSent): Concatenating two sentences with a separator token. This method adopts the idea of increasing the context (Junczys-Dowmunt, 2019).

Incorporating the speaker gender identification (SpeakerId): Incorporating the information of the gender of the speaker in NMT by adding the gender tag before each sentence (Vanmassenhove et al., 2018). This approach is specially helpful when translating from a less inflected language to a more inflected one, e.g., from English to Spanish.

Methods	Examples
Baseline	I have only done this once before.
+PreSent	I have only done this once before. <sep> This is not a joke.
+SpeakerId	MALE I have only done this once before.

Table 1: Methodologies examples

Methods	EuroParl	GeBio
Baseline	44.01	36.34
+PreSent	**45.10**	**36.55**
+SpeakerId	44.18	36.51

Table 2: BLEU results (best in bold).

3 Experimental Framework, Results and Discussion

Data and Parameters: Spanish is a highly-gendered morphological language compared to English, associating gender to professions and adjectives. That is why the language pair (EN-ES) has been used from the proposed data in Vanmassenhove et al. (2018).The size of the EN-ES dataset is considered moderate with 1,419,507 number of sentences. We have used two test datasets: a random set of EuroParl (2000 sentences) and the gender-balanced set from wikipedia biographies (GeBioCorpus) (Costa-jussà et al., 2019) that contains 1000 sentences from male bios and 1000 sentences from female bios. The gender of the main character in the biography article is used as the gender tag. The model is built on top of fairseq[2] library and the parameters are customized as follows: Adam optimizer, 30K training steps, 14 layers, 512 as embedding dimensionality, feedforward expansion of dimensionality 2048 and 8 attention heads, based on best performing parameters from previous work (Fonollosa et al., 2019).

BLEU results (Table 2). These results have been acquired by testing Europarl test set and GeBioCorpus. Adding the previous sentence has higher impact in Europarl (+1.09) than in GeBioCorpus (+0.21) due to the fact that documents in GeBioCorpus are not coherent (all sentences belong to the same document but some sentences may not be contiguous). Adding the gender tag shows exactly the same effect in GeBioCorpus than in Europarl (+0.17), even if the speaker identification is not from the same nature in the EuroParl and in GeBioCorpus. In the former, the speaker identification comes from the speaker, whereas in the latter, it comes from the biography main character.

Other advantages (Table 4 and Table 5). The outputs of our systems have been manually analysed to study the impact of explored techniques on final translation. We report translation examples where both PreSent and SpeakerId techniques are **helping towards named entity disambiguation (in terms of gender)**. Furthermore, we show other translation examples where adding PreSent shows an **improvement towards morphological agreement and quality on translation style**.

Evaluating on WinoMT (Table 3 and Figure 1). This step is carried out by translating the WinoMT dataset and evaluating the translation by Stanovsky et al. (2019) system that depends on extracting the gender of entities of the translated sentences. These entities are evaluated against the gold annotations provided by the original English dataset. The evaluation is performed on three aspects, the whole WinoMT dataset, and subsets of both pro-stereotypical and anti-stereotypical sentences. An example of anti-stereotypical sentences is *The developer argued with the designer because she did not like the design.*, where the *developer* is a *female* entity. An example of pro-stereotypical sentences is *The CEO helped the nurse because he wanted to help.*, where the *CEO* is a *male* entity. As shown in Figure 1, the systems are performing better on the pro-stereotypical portion of WinoMT than on the anti-stereotypical

[2]https://github.com/pytorch/fairseq

one. The accuracy (Acc), shown in Table 3 and Figure 1, indicates that the methodology PreSent performs best compared to the other approaches in this paper (baseline or SpeakerId). The PreSent detects the gender more correctly than the others whether pro-stereotyped or anti-stereotyped, and its accuracy reaches 61% with 12.2% difference between f1-scores of males and females. This accuracy improves over the best results presented in the original paper (Stanovsky et al., 2019), where the best accuracy is 59.4% with 15.4% difference between f1-scores of males and females. It is important to notice that WinoMT is a test set that does not contain information at the level of document and without speaker identification, so translation with our methodologies is done without this information. Therefore, adding the information of the previous sentence makes the system more robust and it does not mind that we are doing inference without this information.

Methods	**Acc.**	$\triangle G$
Baseline	56.0	18.7
+PreSent	**61.0**	**12.2**
+SpeakerId	52.5	22.2

Table 3: WinoMT evaluation results. Acc. indicates gender accuracy (% of instances the translation had the correct gender), $\triangle G$ denotes the masculine/feminine difference in F1 score. In bold, best results.

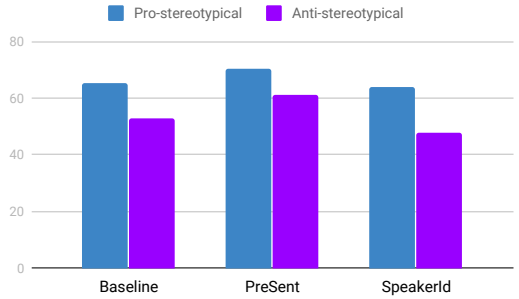

Figure 1: Acc.% on gender translation with respect to pro-stereotypical entities and anti-stereotypical entities in WinoMT.

	Named Entity Disambiguation
Source	María del Carmen Pérez ...is **a Spanish Egyptologist , curator and researcher**.
Baseline	María del Carmen Pérez ...es **un egipcio pintor , curador e investigador español**.
+PreSent	María del Carmen Pérez ...es **una ciudadana española egipcia , curadora e investigadora**.
	Better dealing with articles
Source	Míriam Hatibi ... is **a Catalan data analyst and activist**.
Baseline	Míriam Hatibi ... es **un analista de datos catalán y un activista**.
+PreSent	Míriam Hatibi ... es **una analista y activista catalana en materia de datos**.
	Better style of translations
Source	Helena Maleno Garzón ... is a Spanish human rights **defender , journalist, researcher , documentalist and writer**.
Baseline	Helena Maleno Garzón ... es **un defensor** de los derechos humanos español, **periodista, investigador , documentalista y escritor**.
+PreSent	Helena Maleno Garzón ....es **una defensora** española de los derechos humanos, **periodista, investigadora , documental y escritora**.

Table 4: Baseline vs PreSent Examples from GeBioCorpus.

	Named Entity Disambiguation
Source	Bianca Maria Piccinino ... is an **Italian writer , journalist and television hostess**.
Baseline	Bianca Maria Piccinino ... es **un escritor italiano , periodista y centro** de televisión.
+SpeakerId	Bianca Maria Piccinino ... es **una escritora italiana , periodista y anfitriona** de televisión.

Table 5: Baseline vs SpeakerId Examples from GeBioCorpus.

Acknowledgments

This work is supported in part by the Catalan Agency for Management of University and Research Grants (AGAUR) through the FI PhD Scholarship. This work is also supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, contract TEC2015-69266-P (MINECO/FEDER,EU) and contract PCIN-2017-079 (AEI/MINECO).

References

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. In *Proceedings of Advances in Neural Information Processing Systems 29*, pages 4349–4357.

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, Pau Li Lin, and Cristina España-Bonet. 2019. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. In *Proceedings of 12th Language Resources and Evaluation Conference (LREC)*.

José A. R. Fonollosa, Noe Casas, and Marta R. Costa-jussà. 2019. http://arxiv.org/abs/1905.06596 Joint source-target self attention with locality constraints.

Joel Escudé Font and Marta R. Costa-jussà. 2019. https://doi.org/10.18653/v1/W19-3821 Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First ACL Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Proceedings of Advances in Neural Information Processing Systems*, pages 7944–7954.

Marcin Junczys-Dowmunt. 2019. https://doi.org/10.18653/v1/W19-5321 Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. https://doi.org/10.18653/v1/S18-2005 Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2020. Assessing gender bias in machine translation – a case study with google translate. *Neural Computing and Applications, 32*, page 6363–6381.

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*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. https://doi.org/10.18653/v1/P19-1159 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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.