Gender Bias in Machine Translation: A Survey

1st Semester of 2022-2023

Dana Dăscălescu dana.dascalescu@s.unibuc.ro

Miruna-Andreea Zăvelcă miruna.zavelca@s.unibuc.ro

Abstract

Detrimental effects of gender bias on the translation quality of machine translation systems have been shown in numerous studies. Consequently, identifying, reducing, and evaluating gender bias in machine translation systems has become a current area of interest in natural language processing. In this survey, we give a comprehensive review of articles and their reproducibility on gender bias in machine translation, including their theoretical and practical contributions. First, we identify and describe three generic approaches for eliminating gender bias from translations. Then, we present a multi-perspective categorization of several approaches for evaluating neural machine translation systems before they reach production, as well as the importance of ensuring that detrimental biases are eliminated prior to being shown to the end user. Finally, we discuss the current limitations of eliminating gender bias and envision several promising directions for future research. Our code is available on GitHub.

1 Introduction

Due to the increasing demand for multilingual communication and real-time translations, data availability, technological advancements that lead to cost-effectiveness and speed, and its applicability in several areas, machine translation has become a rapidly growing field. Despite the many breakthroughs in the domain, these systems are still highly susceptible to the introduction and perpetuation of unintended gender bias, which is inadequately reflected in their translation (Frank et al., 2004; Moorkens, 2022).

Compared to the extensive resources and efforts devoted to improving state-of-the-art translation quality through the use of word-overlap automatic metrics, little attention has been paid to eliminating the inherent gender bias in these systems. In addition, the field of gender bias in Machine Translation lacks coherence, which results in the absence of a consistent framework, hindering future research in this area. In recent years, however, there has been a growing emphasis on identifying, interpreting, and eliminating gender bias in machine translation systems with the goal of developing more accurate, fair, and inclusive models that benefit all individuals (Figure 1), in response to numerous concerns about the societal impact of NLP tools raised both within (Zhao et al., 2018a; Prates et al., 2020; Bender et al., 2021; Troles and Schmid, 2021) and outside (Dastin, 2018; Feast, 2020) the scientific community. Studies in this field have included the creation of benchmarking datasets for bias mitigation (Rudinger et al., 2018a; Stanovsky et al., 2019; Sakaguchi et al., 2021), the development of algorithms for reducing bias (Bolukbasi et al., 2016a; Elaraby et al., 2018a; Basta et al., 2020; Saunders and Byrne, 2020; Kim et al., 2019; Zhao et al., 2018b), and the establishment of evaluation measures for assessing the degree of bias present in the learned representations of models (Dixon et al., 2018; Park et al., 2018; Cho et al., 2019; Vanmassenhove et al., 2018; Stanovsky et al., 2019; Gonen and Webster, 2020; Hovy et al., 2020).

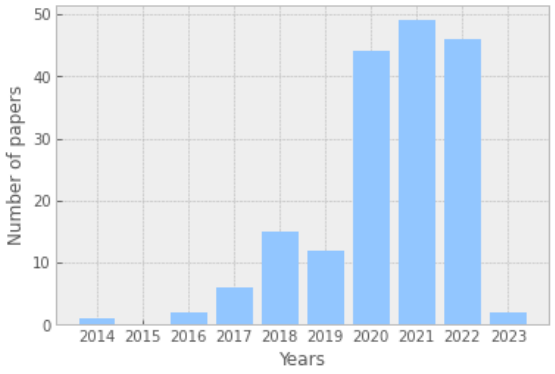

Figure 1: Number of papers per year

The feminist approach to technology has long acknowledged the influence of gender bias on ma-

chine translation systems, particularly how it silences women's voices and hinders progress in many areas (Tallon, 2019; Monti, 2020), although the field of machine translation itself has only begun to pay attention to this issue. It wasn't until 2004 that the issue of gender in machine translation was first tackled from a linguistic and software engineering perspective within the scientific community (Frank et al., 2004). This marked a significant effort to address gender-related issues in language technology and to stress the importance of improving the quality of translations by making them gender-appropriate. Many years passed before the seminal study (Bolukbasi et al., 2016a) brought to the forefront of the discourse within the scientific community the issue of overt sexism in word embeddings and the potential for perpetuation of long-attending prejudicies and inequities between men and women in Machine Translation. The ground-breaking research was the first to empirically demonstrate the existence of gender bias in word embeddings trained on a large corpus, specifically in the relation of male-associated terms with career-related words and female-associated terms with family and domestic-related words. The authors also proposed an algorithm for debiasing static word embeddings. Another study, (Caliskan et al., 2017) which explores human-like semantic biases, similarly found that word embeddings trained on text data exhibit gender biases. Multiple studies (Prates et al., 2018; Hovy et al., 2020) have also shown a tendency toward male defaults in translations, which is only a mirror of controversial societal asymmetries, lending weight to these arguments.

2 Understanding Bias and its Impact on Machine Translation Systems

Sun et al., 2019 separates gender bias in four categories (Table 1). As such: "denigration refers to the use of culturally or historically derogatory terms; stereotyping reinforces existing societal stereotypes; recognition bias involves a given algorithm's inaccuracy in recognition tasks; and underrepresentation bias is the disproportionately low representation of a specific group." (Sun et al., 2019).

It is important to understand how bias works and the impact it has in order to be able to mitigate it. There has been a lot of work done on debiasing in recent years, but also a lot of backlash regarding currently used methods. While Levesque, 2014 condemns neural networks for relying on easy-to-learn shortcuts or "cheap tricks", (Gonen and Goldberg, 2019) states a clear position regarding debiasing as a whole, prominent from title: *Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them*. It has also been debated whether using methods that deprive systems of some knowledge is the right direction toward developing fairer language models (Nissim et al., 2019; Goldfarb-Tarrant et al., 2021)

3 A Categorization of Approaches for Reducing Bias

3.1 Dataset alignment

Bias starts from the dataset. While long term we aspire for more balanced datasets, it might take some time before this aspiration will become the norm. (Costa-jussà and de Jorge, 2020) propose a gender-balanced dataset built from Wikipedia biographies. While this generally improves the generation of feminine forms, the approach is not as effective as it does not account for stereotypes that arise from the qualitative different ways in which men and women are portrayed (Wagner et al., 2021).

An alternative approach implies altering already existing datasets so that the model can suppress this bias. According to (Savoldi et al., 2021) we can do this by gender tagging the sentences or adding context. We propose a third different category, gender swapping.

3.1.1 Gender tagging

It was hypothesised (and proven) that integrating gender information into NMT systems improves the gender referential makings of a translation, whether the tagging is made at the sentence-level (Vanmassenhove et al., 2018; Elaraby et al., 2018b) or at the word-level (Stafanovičs et al., 2020; Saunders et al., 2020).

(Vanmassenhove et al., 2018) uses additional metadata in order to prepend a gender tag (M or F) to each input sentence. This proves especially useful for translating sentences from first person English, where there are usually no gender markings, to any language that has them – "I am a nurse" will become either "Je suis un infirmier" or "Je suis une infirmière" in French, depending on the speaker's gender. However, metadata might not always be

Task	Example of Representation Bias in the Context of Gender	D	S	R	U
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017; Zhao et al., 2017; Rudinger et al., 2018b)		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017)			✓	✓
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).		✓		
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).		✓	✓	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016b)	✓	✓	✓	✓

Table 1: Following the talk by (Crawford, 2017), we categorize representation bias in Machine Translation into the following four categories: (D)enigration, (S)tereotyping, (R)ecognition, (U)nder-representation. Table from (Sun et al., 2019)

available or easy to procure, and an automatic annotation may introduce additional bias.

(Elaraby et al., 2018b) defines a set of cross-lingual rules based on POS tagging in an English-Arabic parallel corpus, yet this approach would not be feasible in realistic conditions, as it requires reference translations for the gender tagging.

(Saunders et al., 2020) has a particularly interesting approach on word-level gender tags as it explores non-binary translations in an artificial dataset. The dataset contains neutral tags, having the gendered reflections replaced with placeholders.

3.1.2 Adding context

A more accessible approach implies expanding the context of our dataset. As such, multiple methods have been suggested.

(Basta et al., 2020) adopts a generic technique, concatenating each sentence with the preceding one. This slightly reduces bias as it expands the chances of coreference resolution.

(Stanovsky et al., 2019) uses a heuristic morphological tagger to extract the gender of the target entity from the source and from the translation. This is not added *per se* to the model, but is used for identifying and adding pro-stereotypical adjectives – *fighting bias with bias*. Thus, the sentence "The doctor asked the nurse to help her in the operation" becomes "The *pretty* doctor asked the nurse to help her in the operation". The authors wanted to test whether mixing signals ("doctor" biases towards a male translation, while "pretty" has female inflections) corrects the model.

Lastly, (Sharma et al., 2022) expands knowledge through relevant context using a template based on morphological taggers. The template is used greedily from a set of 87 possibilities in the form of "The occupation in the following sentence is excellent at m/f-pos-prn job.".

3.1.3 Gender swapping

Gender swapping is characterized by swapping all genders in a sentence and adding it to the dataset. While it might seem like a simple approach for creating a balanced dataset, it has the risk of creating nonsensical sentences (swapping "she gave birth" to "he gave birth") or removing relevant bias (women and men tend to express themselves differently) (Madaan et al., 2018). This subject has been tested by (Zhao et al., 2018a; Lu et al., 2018; Kiritchenko and Mohammad, 2018).

3.2 Model Debiasing

- Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques (Escudé Font and Costa-jussà, 2019)

3.2.1 Debiasing Static Word Embeddings

- Man is to Computer Programmer as Woman is to Homemaker (Bolukbasi et al., 2016b)

- Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods (Zhao et al., 2018a)

- Adversarial Removal of Demographic Attributes from Text Data (Elazar and Goldberg, 2018)

3.2.2 Debiasing Contextual Word Embeddings

- Gender Bias in Contextualized Word Embeddings (Zhao et al., 2019)

- Debiasing Pre-trained Contextualised Embeddings (Kaneko and Bollegala, 2021)

3.3 Debiasing through External Components via Inference-time

Instead of directly debiasing the Machine Translation model, we can intervene with an external component at inference time. This approach has the benefit that we don't have to retrain time-consuming models, but comes with the drawbacks specific to maintaining and handling a separate model that needs to be integrated with the previous results.

According to (Savoldi et al., 2021), debiasing through external components can be split into three categories: black-box injection, lattice re-scoring and gender re-inflection.

3.3.1 Black-box Injection

Black-box injection doesn't take into account anything related to the dataset or the models' bias. (Moryossef et al., 2019) attempts to control the production of feminine and plural references by adding a prepended construction ("*she* said to *them*") to the source sentences and then removing it from the output.

3.3.2 Lattice Re-scoring

Unlike the previous approach, lattice re-scoring is a post-processing technique based on analyzing the dataset and producing a second lattice with differently scored gender-marked words. In (Saunders and Byrne, 2020) the gender marked words are mapped to all their possible inflectional variants and the sentences corresponding to the paths in the lattice are re-scored with a gender debiased model. Then, the highest probability sentences are chosen as the new output. This lead to an increase in the accuracy of gender form selection but a decrease in the generic translation quality.

3.3.3 Gender Re-inflection

Finally, gender re-inflection implies changing first person references into masculine / feminine forms. (Alhafni et al., 2020) feeds to the component the preferred gender of the speaker together with the translated Arabic sentence, while (Habash et al., 2019) attempts two approaches: i) a two-step system that first identifies the gender of 1st person references in an MT output, and then re-inflects them in the opposite form and ii) a single-step system that always produces both forms from an MT output.

4 Limitations

There are a number of limitations that can arise when addressing gender bias. Some of the main limitations include:

1. **Data availability:** Gender bias can be difficult to detect and measure, particularly if data on gender is not collected or is not easily accessible.

2. **Selection bias:** Studies may be affected by selection bias, which occurs when the sample of participants is not representative of the population being studied. This can make it difficult to generalize findings to other groups.

3. **Confounding variables:** Gender bias may be confounded with other factors, such as socioeconomic status or race, making it difficult to isolate the effects of gender.

4. **Lack of consensus on how to measure bias:** There is currently no consensus on how to measure or define gender bias, which can lead to inconsistent findings and conclusions.

5. **Implicit bias:** People may have implicit biases, which are unconscious attitudes or beliefs that can affect their behavior and decision-making. These biases are difficult to detect and measure.

6. **Difficulty in disentangling bias from structural inequalities:** Gender bias is often intertwined with structural inequalities, such as discrimination and lack of representation in certain fields, making it hard to separate the effects of bias from those of inequality.

7. **Limited understanding of intersectionality:** Gender bias often interacts with other forms of

bias, such as race and class, making it important to consider intersectionality when studying gender bias.

5 Conclusions and Future Work

In this paper we present a short overview over debiasing methods currently used in literature, together with their benefits and drawbacks.

Ethical Statement

This study relies on data that includes sexism and hate speech. The examples provided for identifying gender bias include instances of sexism which might be disturbing for certain individuals. Reader discretion is advised. The authors vehemently oppose the use of any derogatory language against women.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns.

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.

Kate Crawford. 2017. The trouble with bias.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification.

Laura Douglas. 2017. Ai is not just learning our biases; it is amplifying them. *Medium*.

Mostafa Elaraby, Ahmed Y Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018a. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018b. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. 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.

Josh Feast. 2020. 4 ways to address gender bias in ai.

Anke Frank, Christiane Hoffmann, and Margaret Shay Strobel. 2004. Gender issues in machine translation.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. *arXiv preprint arXiv:2004.14065*.

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, Florence, Italy. Association for Computational Linguistics.

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.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. 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? *arXiv preprint arXiv:1910.00294*.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems.

Hector J. Levesque. 2014. On our best behaviour. *Artificial Intelligence*, 212:27–35.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing.

Nishtha Madaan, Sameep Mehta, Taneea S. Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In *FAT*.

Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468. Routledge.

Joss Moorkens. 2022. Ethics and machine translation. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:121.

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.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2019. Fair is better than sensational:man is to doctor as woman is to doctor.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2018. Assessing gender bias in machine translation – a case study with google translate.

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*, 32:6363–6381.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018a. 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*, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018b. 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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

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.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts.

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, Florence, Italy. 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.

Tina Tallon. 2019. A century of "shrill": How bias in technology has hurt women's voices. *The New Yorker*.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives.

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.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2021. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):454–463.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.