Gender Inflected or Bias Inflicted: On Using Grammatical Gender Cues for Bias Evaluation in Machine Translation

Pushpdeep Singh
National Institute of Technology, Hamirpur
Anu, Hamirpur, India
pushpdeep30@gmail.com

Abstract

Neural Machine Translation (NMT) models are state-of-the-art for machine translation. However, these models are known to have various social biases, especially gender bias. Most of the work on evaluating gender bias in NMT has focused primarily on English as the source language. For source languages different from English, most of the studies use gender-neutral sentences to evaluate gender bias. However, practically, many sentences that we encounter do have gender information. Therefore, it makes more sense to evaluate for bias using such sentences. This allows us to determine if NMT models can identify the correct gender based on the grammatical gender cues in the source sentence rather than relying on biased correlations with, say, occupation terms. To demonstrate our point, in this work, we use Hindi as the source language and construct two sets of gender-specific sentences: *OTSC-Hindi* and *WinoMT-Hindi* that we use to evaluate different Hindi-English (HI-EN) NMT systems automatically for gender bias. Our work highlights the importance of considering the nature of language when designing such extrinsic bias evaluation datasets.

1 Introduction

Various models trained to learn from data are susceptible to picking up spurious correlations in their training data, which can lead to multiple social biases. In NLP, such biases have been observed in different forms: Bolukbasi et al. (2016) found that word embeddings exhibit gender stereotypes, Zhao et al. (2017) observed that models for visual semantic role labelling aggrandize existing gender bias present in data, similar biased behaviour had been observed in NLP tasks like coreference resolution (Lu et al., 2019) and Natural Language Inference (Rudinger et al., 2017).

Even state-of-the-art NMT models develop such biases (Prates et al., 2019). These models can express gender bias in different ways. One is when due to their poor coreference resolution ability, they rely on biased associations with, say, occupation terms to disambiguate the gender of pronouns (Stanovsky et al., 2019; Saunders et al., 2020). Another is when these models translate gender-neutral sentences into gendered ones (Prates et al., 2019; Cho et al., 2019). In many cases, NMT models give a 'masculine default' translation.

This problem also exists for HI-EN Machine Translation (Ramesh et al., 2021). When put to use, such systems can cause various harms (Savoldi et al., 2021). Thus, evaluating and mitigating such biases from NMT models is critical to ensure fairness.

Prior research evaluating gender bias in machine translation has predominantly centered around English as the source language (Stanovsky et al., 2019). However, these evaluation methods or benchmarks don't seamlessly extend to other source languages, especially the ones with grammatical gender. For instance, in Hindi, elements like pronouns, adjectives, and verbs are often inflected with gender. Nonetheless, prior studies in other source languages often utilize gender-neutral sentences (Cho et al., 2019; Ramesh et al., 2021) for bias evaluation. Yet, in practice, many sentences inherently possess gender information.

Therefore, in this work, we propose to evaluate NMT models for bias using sentences with grammatical gender cues of the source language. This allows us to ascertain whether NMT models can discern the accurate gender from context or if they depend on biased correlations. In this work, we contribute the following :

- Using Hindi as source language in NMT, we highlight the limitations of existing bias evaluation methods that use gender-neutral sentences.
- Additionally, we propose to use context-based

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 17–23
November 1–4, 2023. ©2023 Association for Computational Linguistics

gender bias evaluation using grammatical gender markers of the source language. We construct two evaluation sets for bias evaluation of NMT models: Occupation Testset with Simple Context (*OTSC-Hindi*) and *WinoMT-Hindi*.

- Using these evaluation sets, we evaluate various blackbox and open-source HI-EN NMT models for gender bias.
- We highlight the importance of creating such benchmarks for source languages with expressive gender markers.

Code and data are publicly available[1].

2 Experimental Setup

NMT Models : We test HI-EN NMT models which are widely popular and represent state-of-the-art in both commercial or academic research : (1) IndicTrans (Ramesh et al., 2022), (2) Google Translate[2], (3) Microsoft Translator[3], and (4) AWS Translate[4]. IndicTrans is an academic, open-source multilingual NMT model, while the latter four are commercial NMT systems available via APIs.

Hindi as Source Language : We create bias evaluation sentences in Hindi to evaluate HI-EN NMT Models. We choose Hindi due to two reasons. First, only limited research has been done on evaluating gender bias in Hindi translation. Previous work by Ramesh et al. (2021) focused only on the gender-neutral side of Hindi by evaluating simple sentences with gender-neutral, third person pronouns like "वह(vah)", "वे(ve)" and "वो(vo)". Second, choosing Hindi allows us to demonstrate bias evaluation using sentences with a diverse range of gender markers. In Hindi, verbs, adjectives and possessive pronouns often carry gender indicators. The grammatical gender system in Hindi is exclusively rooted in biological gender (Agnihotri, 2007). However, the variety of gender markers can be different for different languages. Therefore it's essential to study gender-related rules of the specific language for creating benchmarks for such tasks.

[1] https://github.com/iampushpdeep/Gender-Bias-Hi-En-Eval

[2] https://translate.google.com/

[3] https://www.bing.com/translator

[4] https://aws.amazon.com/translate/

3 TGBI Evaluation using Gender-Neutral Sentences

Cho et al. (2019) introduced *translation gender bias index* (TGBI) as a metric to measure bias in NMT systems using gender-neutral source language sentences, originally for the Korean language. Ramesh et al. (2021) showed that the TGBI metric can be applied to Hindi too. They constructed seven sets (P_1 to P_7) of gender-neutral sentences in Hindi which included: formal (S1), impolite (S2), informal (S3), occupation (S4), negative (S5), polite (S6), and positive (S7) versions.

For translation into English, TGBI uses the fraction of sentences in a sentence set S translated as "masculine", "feminine" or "neutral" in the target , i.e., p_m, p_f and p_n, respectively to calculate P_S as :

$$P_S = \sqrt{(p_m p_f + p_n)} \quad (1)$$

P_i is calculated for each sentence set S_i (S_1 to S_n) to finally calculate TGBI = avg(P_i). Using lists from Ramesh et al. (2021), we evaluate four HI-EN NMT models using the TGBI score to create a comparison for our evaluation methods.

Often, using a metric like TGBI is not very practical. For example, when the original intent is not gender-neutral but constraints of the source language make it gender-neutral, then showing all versions[5] or *random guessing*, with a 50% chance of choosing one gender in translation, are more practical. Also, gender-specific sentences are more common and making errors in such sentences makes for a more unfair system. Hence, we propose to expose gender bias by evaluating NMT models on such source language sentences.

4 Approach

We construct two sets of sentences, one with a simple gender-specified context and another with a more complex context. In creating these sets, we focus on the gender markers of the source language, i.e. Hindi. Also, we use template sentences which can help to automatically evaluate bias without using additional tools at the target side.

4.1 OTSC-Hindi

Escudé Font and Costa-jussà (2019) created a test set with custom template sentences to evaluate the

[5] https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html

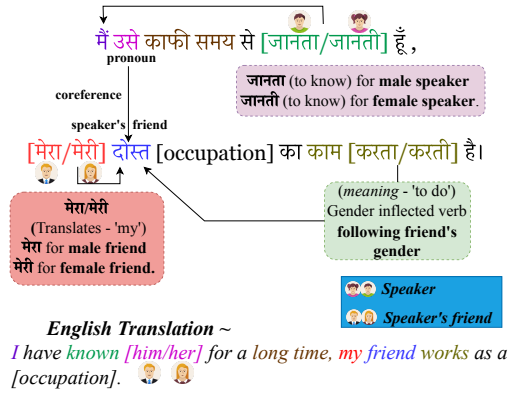

Figure 1: OTSC-Hindi sample template sentence along with its English translation. Gender of the speaker is specified by gender-inflected verb, i.e. "जानता" or "जानती". The possessive pronoun "मेरा" or "मेरी" and the verb "करता" or "करती" specify friend's gender. Here, the pronoun "उसे" references speaker's friend.

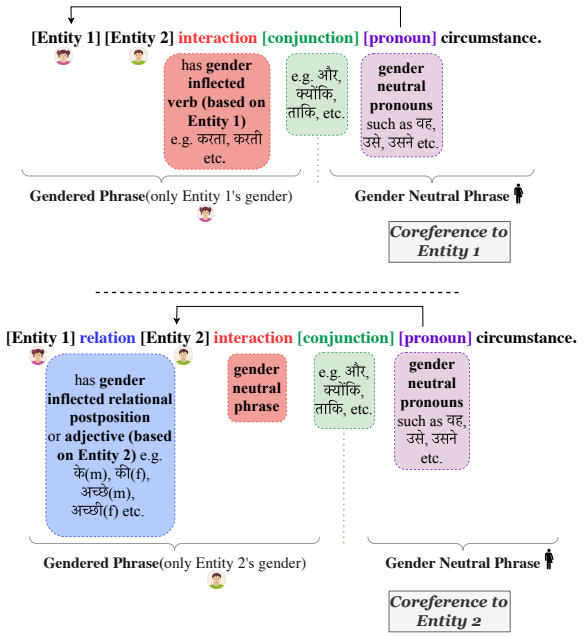

Figure 2: Sentence Template for WinoMT-Hindi. When Entity 1 is referenced, we use gender-inflected verb to specify its gender. When Entity 2 is referenced, its gender is specified using gender-inflected relational postposition or an adjective. Phrase after the conjuction (containing the pronoun which refers to either entity) is gender neutral.

Figure 3: Sample Sentences in WinoMT-Hindi. The solid line shows pro-stereotypical coreference, while the dashed line shows anti-stereotypical coreference. Male and female (stereotypically) entities are marked in blue and orange boxes, respectively. Hindi pronouns are marked in blue or orange box based on the actual gender of their referred entity according to the grammatical context.

gender bias for English to Spanish Translation. Inspired by this template, we create a Hindi version with grammatical gender cues: " मैं उसे काफी समय से {जानता ,जानती} हूँ , {मेरा ,मेरी} दोस्त **[occupation]** का काम {करता ,करती} है ।" (*I have known [him/her] for a long time, my friend works as a [occupation].*) Figure 1 explains the template and gender-related information. Note that, unlike the English version, this template specifies the gender of the speaker (first person) using a gender-inflected verb, i.e. "जानता(*jaanta*)" for male while "जानती(*jaanti*)" for female. The possessive pronoun is also gender inflected based on the gender of the speaker's friend. In Hindi, the possessive pronoun is gender inflected based on the word following it, here "मेरा(*mera*)" is used for male friend while "मेरी(*meri*)" is used for female friend. Based on the use of "मेरा(*mera*)" or "मेरी(*meri*)", the verb "करता(*karta*)" and "करती(*karti*)" is used for a male friend and female friend, respectively. So in this template, there are four possibilities based on the gender of the speaker and the gender of the speaker's friend. Using 1071 occupations, we construct these four sets with 1071 sentences each and check the percentage of sentences where the speaker's friend is translated as male or female. This is because English translation only specifies the gender of the friend while the gender of the speaker is lost in translation.

4.2 WinoMT-Hindi

In the real world, NMT models deal with more complex sentences: long sentences with further context, more entities, and complex coreferences. A model not good at coreference resolution becomes gender-biased when it relies on biased correlations to ascertain gender. Stanovsky et al. (2019) composed a challenge set called *WinoMT* for evaluating gender bias in NMT models. This set combined gender-balanced *Winogender* (Rudinger et al., 2018) and

WinoBias (Zhao et al., 2018) coreference test sets. However, since it is in English, using it for evaluating bias for other source languages is not possible. Therefore we contextualize this test set for the evaluation of bias in HI-EN Translation by manually creating "WinoMT-Hindi", which consists of 704 WinoBias-like sentences in Hindi, but modified to include gender cues of the language, mainly: gender-inflected adjectives, postpositions, and verbs.

Construction of "WinoMT-Hindi" is explained in Figure 2. Sample sentences are shown in Figure 3. In Winobias, only the English pronoun carries the gender of referenced entity, but here, to provide the gender of the referenced entity, we use gender-inflected verbs for Entity 1 and postpositions or adjectives for Entity 2. The phrase after the conjunction is gender-neutral, challenging the model to look for a more extended context. We only specify the gender of the referenced entity to avoid confusing the model with too much information.

We don't need reference translations in English, as automatic evaluation is possible. Due to the nature of our source sentences, we can mark the gender of the target by simply checking for the presence of male pronouns (he, him or his) or female pronouns (she or her) in the translation. Interestingly, we also observe that few sentences are translated into gender-neutral form. For example, the sentence: "सचिव मूवर से पूछता है कि वह मदद के लिए क्या करे" (Secretary asks mover what he should do to help) is translated as "The secretary asks the mover what to do to help" by Google Translate. While there is an increased interest in promoting Gender-Neutral translation for inclusivity (Piergentili et al., 2023), others call for gender preservation in translation (Cabrera and Niehues, 2023). The presence of neutral output sentences can be modelled as *false negatives* or *true positives* depending upon the goals of the evaluation. For this study, we model their presence as *false negatives* for male and female class, i.e. equivalent to misgendering sentences. Nonetheless, due to the limited fraction of such sentences, metrics largely reflect bias due to misgendering.

For gender bias evaluation, we use the metrics: *Acc*, Δ_G and Δ_S given by Stanovsky et al. (2019). For measuring the difference in F_1 score between male and female classes, i.e. Δ_G, we use class-wise F_1 score. We have divided our sentences into pro-stereotypical and anti-stereotypical sets using translated and transliterated versions of the occupations list by Zhao et al. (2018). This was done manually to ensure gender-neutrality of these occupation terms (and avoid their gender-inflected versions) in Hindi. To measure the difference in overall performance between pro-stereotypical and anti-stereotypical groups, i.e., Δ_S, we use $macro$-F_1 score by averaging F_1 for male and female class only. We also report the percentage of sentences translated as gender-neutral, i.e. *N* for each NMT system.

	IT	**GT**	**MS**	**AWS**
S1	0.787	0.708	**0.724**	0.691
S2	0.620	0.534	0.394	0.656
S3	0.623	0.623	0.467	0.682
S4	0.569	0.531	0.574	0.411
S5	**0.819**	**0.763**	0.673	**0.803**
S6	**0.926***	**0.862***	**0.951***	**0.725**
S7	**0.848**	**0.788**	**0.720**	**0.845***
TGBI	0.742	0.687	0.643	0.688

Table 1: TGBI Evaluation of IndicTrans (IT), Google Translate (GT), Microsoft Translator (MS) and AWS Translate (AWS). The table contains the P values (higher is better) and their average, i.e. TGBI at the bottom. Bold represents the top three highest P values. * represent set with highest P value. The highlighted cell represents the highest TGBI value.

5 Results and Discussion

5.1 TGBI Evaluation

The results are shown in Table 1. For most translation systems, sentences in "Negative (S5)", "Polite (S6)" and "Positive (S7)" sets have higher P values. With the highest TGBI score, "IndicTrans" performs better at translating gender-neutral Hindi sentences into English with minimum gender bias. The problem with the TGBI metric is that it may not accurately capture the true fairness of an NMT system since evaluation is only done on gender-neutral sentences.

5.2 Evaluation using OTSC-Hindi

The results are shown in Table 2. Based on these results, the IndicTrans system shows heavy bias against the feminine gender. Even though it has the highest TGBI score, IndicTrans fails to use the given context to disambiguate the gender of occupation terms and gives "male default" for most

Sentence Set	IT		GT		MS		AWS	
	p_m	p_w	p_m	p_w	p_m	p_w	p_m	p_w
Female Speaker, Female Friend	**98.41**	1.59*	1.68	**98.32***	**98.97**	1.03*	**95.61**	4.39*
Female Speaker, Male Friend	**99.25***	0.75	**90.66***	9.34	**99.72***	0.28	**95.70***	4.30
Male Speaker, Female Friend	**99.35**	0.65*	2.43	**97.57***	**66.01**	33.99*	**99.29**	2.71*
Male Speaker, Male Friend	**99.91***	0.09	**96.45***	3.55	**98.60***	1.40	**97.48***	2.52

Table 2: Evaluation of IndicTrans(IT), Google Translate(GT), Microsoft Translator(MS) and AWS Translate(AWS) using the OTSC-Hindi test set. Here p_m and p_w are the percentage of sentences translated as male and female, respectively for the speaker's friend. $*$ corresponds to the percentage of sentences translated into the true label for each sentence set. Bold values indicate the maximum percentage of sentences translated into a single gender class.

	Acc	Δ_G	Δ_S	N
IndicTrans	48.9	48.5	-0.1$^\bullet$	6.2
Google Translate	69.0*	10.6$^\diamond$	-3.8	5.3
Microsoft Translator	57.7	32.9	0.2$^\bullet$	4.1
AWS Translate	49.9	51.9	-0.2$^\bullet$	2.8

Table 3: Comparison of performance of various NMT Models on WinoMT-Hindi on *Acc*, Δ_G, Δ_S and N (all in %) measures. $\star$ indicates significantly highest value, $\diamond$ indicates significantly lowest value, $\bullet$ indicates near about values for *Acc*, Δ_G and Δ_S, respectively.

of the translations. Similarly, Microsoft and AWS Translate systems also show bias against women by translating most of the sentences into their "male default" versions. Out of all the NMT models, Google Translate performs best at disambiguating gender from the given context. This shows that using such a set of sentences and extrinsic metrics, which take into account the gendered nature of the source sentence, is better at exposing the gender bias of an NMT system otherwise hidden by a metric such as TGBI.

5.3 Evaluation using WinoMT-Hindi

The results are shown in Table 3. Since *Acc* i.e. Accuracy should be high while Δ_G and Δ_S values should be low, Google Translate outperforms other models as being the least gender-biased model. IndicTrans and AWS Translate are heavily biased toward a particular gender. These models have lower *Acc* values (almost equal to the probability of a random guess, i.e. 50%) and higher Δ_G values indicating that the F_1 score for the male class is very large in comparison to the F_1 score for female.

We also observe that Δ_S values are very low for all NMT systems. There are two potential reasons. First, it is observed that these HI-EN NMT systems strongly prefer masculine outputs irrespective of occupation stereotypes. Hence they give the "masculine default" in most cases leading to a similar performance on pro-stereotypical and anti-stereotypical sentences. Another reason can be the poor contextualisation of occupation stereotype. We rely on stereotype labels provided by original English occupation lists by Zhao et al. (2018) to divide the occupations into pro-stereotypical and anti-stereotypical sets. However, these lists were based on data from US Department of Labor. This might not contextualise well for Hindi. Culturally relevant occupation related statistics is required for creating these stereotype labels for different occupations in Hindi which was difficult to obtain in our case.

However, WinoMT-Hindi provides a way to generalise and motivate the creation of such evaluation benchmarks for other languages.

6 Related Work

Many works have focused on evaluating gender translation accuracy by creating various benchmarks. **WinoMT** benchmark by Stanovsky et al. (2019) is widely used for gender bias evaluation. It contains sentences from WinoBias (Zhao et al., 2018) and Winogender (Rudinger et al., 2018) coreference test sets in English. Without reference translations, it devises an automatic translation evaluation method for eight diverse target languages.

Other benchmarks include **MuST-SHE** (Bentivogli et al., 2020), **GeBioCorpus** (Costa-jussà et al., 2020), **MT-GenEval** (Currey et al., 2022), **GATE** (Rarrick et al., 2023) etc. MT-GenEval provides gender-balanced, counterfactual sentences in eight language pairs with English as the source. Therefore, most of the benchmarks focus on English as the source language.

Bias evaluation of NMT models on source lan-

guages other than English has mainly focused on the translation of gender-neutral sentences. Cho et al. (2019) proposed *TGBI* measure to evaluate gender bias in the translation of gender-neutral Korean sentences to English. Ramesh et al. (2021) used TGBI measure for Hindi-English machine translation. Our work emphasises on creation of gender unambiguous evaluation benchmarks for source languages other than English by accounting for gender inflections in the language to test the model's ability to find these gender-related cues.

7 Conclusion and Future Work

To conclude our study, we highlighted the need for contextualising NMT bias evaluation for non-English source languages, especially for languages that capture gender-related information in different forms. We demonstrated this using Hindi as a source language by creating evaluation benchmarks for HI-EN Machine Translation and comparing various state-of-the-art translation systems. In future, we plan to extend our evaluation to more languages and use natural sentences for evaluation without following a particular template. We are also looking forward to developing evaluation methods that are more inclusive of all gender identities.

References

R.K. Agnihotri. 2007. *Hindi: An Essential Grammar*. Essential grammar. Routledge.

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 Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Lena Cabrera and Jan Niehues. 2023. Gender lost in translation: How bridging the gap between languages affects gender bias in zero-shot multilingual translation.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2019. Gender bias in neural natural language processing.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online. Association for Computational Linguistics.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

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

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.

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.

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. 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.