Evaluating Gender Bias in Machine Translation for Low-Resource Languages

Walelign Tewabe Sewunetie[1], Atnafu Lambebo Tonja[2], Tadesse Destaw Belay[2], Hellina Hailu Nigatu[3], Gashaw Kidanu Gebremeskel[4], Zewdie Mossie [1], Hussien Seid[4], Eshete Derb[1], Seid Muhie Yimam[5]

[∀]Ethio NLP, [1]Debre Markos University, Ethiopia, [2]Instituto Politécnico Nacional, Mexico, [3]University of California, Berkeley, [4]Addis Abeba Science and Technology University, Ethiopia, [5]Universität Hamburg, Germany

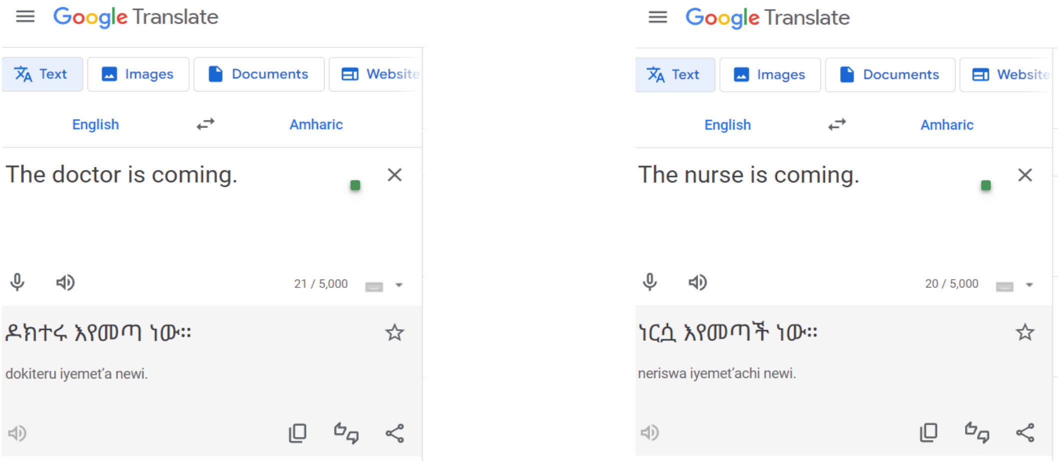
Abstract

While Machine Translation (MT) research has progressed over the years, translation systems still suffer from exhibiting biases, including gender bias. While an active line of research studies the existence and mitigation strategies of gender bias in machine translation systems, there is limited research exploring this phenomenon for low-resource languages. The limited availability of linguistic and computational resources confounded with the lack of benchmark datasets makes studying bias for low-resourced languages that much more difficult. In this paper, we construct benchmark datasets for evaluating gender bias in machine translation for three low-resourced languages: Afan Oromo (orm), Amharic (amh), and Tigrinya (tig). Building on prior work, we collected 2400 gender-balanced sentences parallelly translated into the three languages. From our human evaluations on the dataset we collected, we found that about 93% of Afan Oromo, 80% of Tigrinya, and 72% of Amharic sentences exhibited gender bias. In addition to providing benchmarks for improving gender bias mitigation research in the three languages, we hope the careful documentation of our work will help other low-resourced language researchers extend our approach to their languages.

1 Introduction

Machine Translation (MT) systems play a pivotal role in breaking down language barriers and facilitating cross-cultural communication. Gender bias poses a significant challenge, particularly in languages with limited linguistic resources. The imbalance within datasets used for MT training often results in gender-related disparities. In low-resource languages like Amharic, Tigrinya, and Afan-Oromo, professional names such as doctor, pilot, professor, etc., are mostly translated using male gender. As shown in Figure 1a in the sentence ”The doctor is coming” is the word ”doctor” is translated into male gender for the Amharic language. On the other hand, Figure 1b illustrates that the sentence ”The nurse is coming” is the word ”nurse” translated into only female gender for the Amharic language. Understanding and addressing gender bias in MT systems is vital for ensuring equitable and accurate communication across diverse linguistic communities.

However, addressing this issue requires an adequate dataset for evaluating gender bias, specifically tailored to low-resource languages. The absence of a standardized testing dataset dedicated to gender bias evaluation is an issue. This work aims to fill this gap by constructing a gold-test dataset for languages such as Amharic, Tigrinya, and Afan-Oromo. The methodologies developed in this research can subsequently be applied and scaled up for assessing gender bias in other low-resource languages. In detecting a thorough gender bias evaluation, we have carefully collected 2400 sentences, 1200 sentences for each gender. Establishing our gold test set dataset can provide a robust benchmark for gender bias evaluation in low-



The doctor is coming.

ዶክተሩ እየመጣ ነው።

dokiteru iyemet’a newi.

(a)

The nurse is coming.

ነርሷ እየመጣች ነው።

neriswa iyemet’achi newi.

(b)

Figure 1: Examples of Gender Bias of Google translation in English-Amharic pairs (Accessed January 28, 2024)

resource language translation, and we aspire to contribute to the broader goal of enhancing fairness and equity in MT systems. In this regard, this work presents the first gender bias test set dataset for MT in low-resource language.

In addition, this study investigates the perceptions of gender bias in commercial MT systems and evaluates the Google MT system as a use case in selected Ethiopian languages. Our analysis shows interesting differences in respondents’ perceptions of gender bias across these language communities. These findings underscore the detailed relationship between language, culture, and gender bias perception in MT systems, highlighting the need for adapted approaches to mitigate bias and enhance translation accuracy within specific linguistic contexts. Furthermore, this study investigates the performance of one Open-source MT model and one commercial model, namely, NLLB (Team et al., 2022), and Google MT using automatic evaluation metrics, such as SacreBleu (Post, 2018), and Chrf++ (Popović, 2017). The outcomes of this evaluation across various language pairs shed light on the efficacy and accuracy of MT systems in translating between English and the target languages. The evaluation shows diverse performance metrics across language pairs, with distinct variations in translation quality and effectiveness. These results underscore the importance of robust evaluation methodologies and metrics in assessing MT system performance and informing strategies for enhancing translation accuracy and efficiency across diverse linguistic contexts.

This introduction provides a foundation for exploring the subsequent sections investigating related work, gender bias test dataset preparation, gender bias evaluation techniques, gold dataset preparation, evaluation methodologies, and the findings.

2 Related work

Investigating bias in MT systems is an active body of work in the NLP community. Previous works in this space have relied on (1) curating benchmark datasets (e.g. (Wairagala et al., 2022; Cho et al., 2019)), (2) human evaluation schemes (e.g. (Stanovsky et al., 2019)), and (3) automatic evaluation schemes(e.g. (Savoldi et al., 2021)). In curating benchmark datasets, (Prates et al., 2020) prepared a gender-balanced dataset for evaluating gender bias in translation systems pertaining to occupation. Since different languages represent gender in various ways (Savoldi et al., 2021), evaluation and mitigation strategies might also have to account for such variation. For instance, (Cho et al., 2019) prepared test sets with gender natural pronouns used in the Korean language for investigating bias in Korean-English translation pairs.

In evaluating gender bias in MT, several works rely on automatic metrics. (Prates et al., 2020) found that Google Translate defaults to the male pronoun when translating job descriptions, particularly in relation to science, technology, engineering, and mathematics (STEM) professions. (Cho et al., 2019) introduces a new evaluation index, the Translation Gender Bias Index (TGBI), for measuring gender neutrality and evaluating Korean-English translation pairs. (Stanovsky et al., 2019) introduce an evaluation protocol that relies on co-reference resolution datasets and morphological analysis to automatically evaluate gender bias across eight target languages that use grammatical gender. (Wairagala et al., 2022) used the Word Embeddings Fairness Evaluation Framework (WEFE) to measure gender bias in MT systems built for Luganda-English translation. While automated measures allow us to capture a broader understanding of the phenomenon, they may limit the detail and depth of our analysis. The study by (Stanovsky et al., 2019) uses automatic and human evaluations in tandem, exploiting both the versatility of automated evaluation and the nuance and detail captured by human evaluation.

As the work by (Blodgett et al., 2020) argues, it is important first to articulate how bias in such systems can be harmful. Relying on the taxonomy of harms from prior work (Crawford, 2017; Barocas et al., 2017), we posit that understanding gender bias exhibited by MT systems would allow us to (1) uncover the representational harms the systems exhibit thereby understanding what power structures they uphold and (2) mitigate allocational harms that might result from deploying such systems in downstream applications (e.g. employment and job search).

One challenge in studying bias in machine-translated text is the diverse socio-cultural aspects that shape how gender is articulated among different groups and how stereotypes propagate in this diverse context. (Talat et al., 2022) have shed light on the difficulty of studying and mitigating bias across multicultural, multilingual groups. Such contexts require community-rooted efforts that thoroughly investigate how the culture and language are structured. In this work, we curate benchmark datasets for three low-resource languages through collaborations among native speakers. Based on previous works, (Renduchintala et al., 2021; Stanovsky et al., 2019), we build an automatic evaluation of the translation quality overall and human evaluations of gender bias in popular machine translation systems to understand the current landscape of translation systems for these languages.

3 Gold Gender Bias Test Dataset Preparation

3.1 Dataset Collection and Composition

The gold gender bias test dataset was thoughtfully crafted by combining sentences from public repositories (Sharma et al., 2022), aiming to a thorough examination of gender biases across these selected target languages. We first collected an English-centric dataset from a variety of publicly available sources such as SimpleGEN, [1] and winomt,[2] focusing on relevance and diversity. To maintain balance, a careful collection process ensures that for every gender-specific sentence, there is an equivalent counterpart. For example, if a sentence says, ”He is a doctor,” a corresponding sentence like ”She is a doctor” is included for gender parity.

However, these open-source datasets do not contain all professional names, even though they contain enough test datasets. For this reason, we used a crowdsourcing approach to collect enough datasets from various professions. For this approach, we first incorporated the major professional names currently used by the Ethiopian Civil Service Commission and the recent technological professional words by searching using GPT 3.5 LLM. Finally, through this process, we have collected 108 unique professional names. Figure 2 demonstrated the gender-balanced dataset of each professional name.

Then, we used paid freelancers for crowdsourcing and prepared a Google form containing clear and short instructions about the task. This crowdsourced dataset collection approach

[1] SimpleGEN: https://github.com/arendu-zz/SimpleGEN

[2] winomt: https://github.com/manandey/bias_machine_translation/tree/main/data/base/winomt

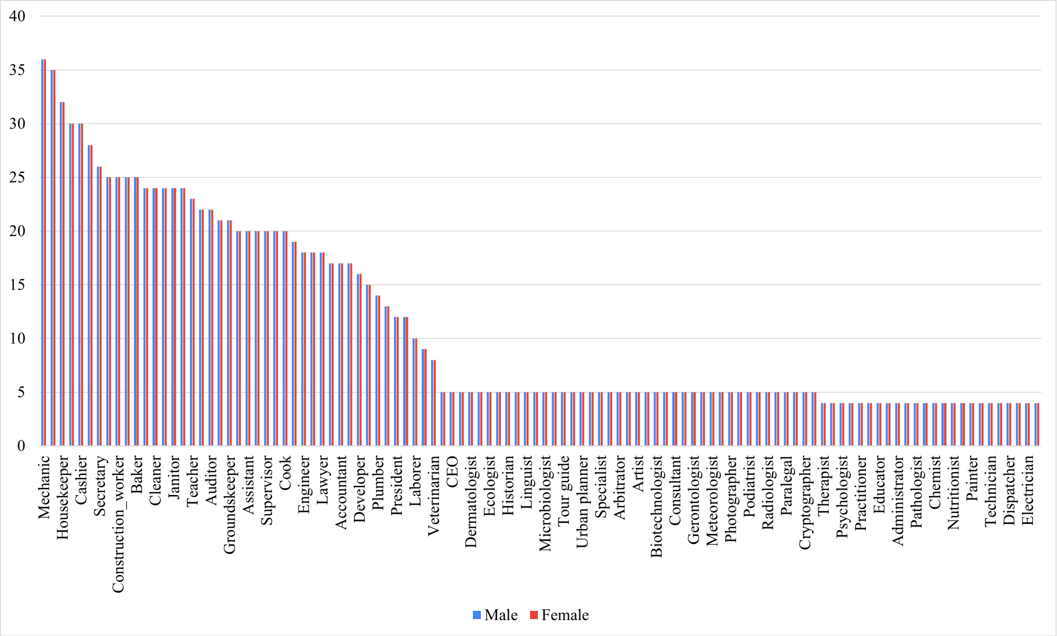


Figure 2: Illustration of professional names in our test dataset

aims to obtain English-centric data from various sources with specific criteria. One of the key considerations was to include both pronouns and occupations in the dataset. This ensured that each occupation was associated with different pronouns, such as ”he,” ”his,” and ”him” for the male, and ”she” and ”her” for the female gender. For this task, ten freelancers were involved and signed an incentive agreement first. Then, we collected the English dataset from SimpleGEN:130, winomt:192, and Crowdsource: 2078, a total of 2400 sentences.

3.2 Dataset Translation

The next task is to translate this collected dataset into three Ethiopian languages: Amharic, Afan Oromo, and Tigrinya. Likewise, we have used paid linguistic experts who were proficient in one of our target languages, then undertook the translation process to preserve linguistic accuracy and capture cultural differences specific to each target language.

To prevent boredom and errors, we engaged six language experts and fluent speakers and divided for each language pair, totaling eighteen individuals from various universities. Assigning one person to handle only 600 pairs of sentences would increase the quality. After the translation, we recruited two paid professional linguists and editors for each language pair for quality checking.

The dataset used in this research, referred to as the Gold Gender Bias Test Dataset (GG-BTD), comprises 2400 sentence pairs for each language pair, specifically English-Amharic, English-Afan Oromo, and English-Tigrinya, resulting in a total of 7200 sentence pairs. Within each language pair, the dataset maintains a comprehensive gender balance. Specifically, for each language pair, 1200 sentences represent male gender expressions, while the remaining 1200 sentences capture female gender expressions.

4 Evaluation Techniques

4.1 Automatic Evaluation

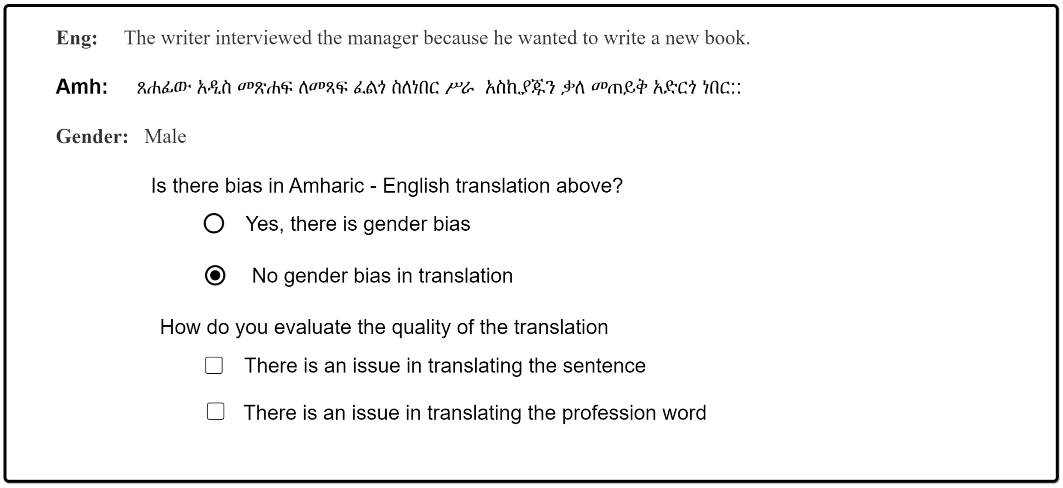
Different evaluation metrics are usually employed to automatically evaluate MT systems. These metrics are often based on word overlap and/or context similarity between references and model outputs. In our work, we employ both types of metrics to evaluate the quality of NLLB and Google MT that we consider in our study. Namely, we used SacreBleu (Post, 2018) and Chrf++ (Popović, 2017) machine translation evaluation metrics. We chose these MT evaluation metrics for several reasons. Firstly, these metrics are widely recognized and utilized in the field of MT research, ensuring compatibility and comparability with existing literature (Kadaoui et al., 2023).

Additionally, SacreBleu and Chrf++ are known for their robustness and effectiveness (Puduppully et al., 2023) in assessing translation quality across different languages and translation systems. Their ability to capture detailed aspects of translation quality, such as fluency, adequacy, and fidelity to the source text, makes them suitable choices for our evaluation framework. Furthermore, both metrics are supported by well-established methodologies and have demonstrated consistent performance in benchmarking studies, giving us confidence in their reliability. However, these metrics evaluate only the overall translation accuracy, and there are no automatic metrics used to evaluate the gender bias of machine translation for low-resource languages.

4.2 Human Evaluation

However, to our knowledge, no metrics specifically assess the gender bias of machine translations. As a result, we relied solely on human-level evaluation techniques. We assessed the gender bias of currently available open-source LLM models and commercial machine translations. Among these, we chose those supporting all three languages (Amharic, Tigrinya, Afan Oromo), and we found only Google Translation and NLLB.

Given the high cost of human-level evaluation, we only evaluated Google Translation. For the human-level evaluation, first, we have developed the evaluation guidelines shown in the appendix 9.1, and we have used the Potato annotation tool (Pei et al., 2022) for evaluation. Figure 3 shows the Potato annotation tool GUI for human-label evaluation, which supports all modern browsers and can be accessed both from computers and mobile phones for manual evaluation annotation. Criteria included gender biases, translation quality, and the accuracy of professional name translations. For evaluation, eighteen paid linguistic experts per language were selected. To avoid subjectivity, we divided evaluators into three groups and made the evaluation into three phases; this implies each sentence is evaluated three times. This is good for taking the majority vote for result analysis.



Eng: The writer interviewed the manager because he wanted to write a new book.

Amh: ጸሐፊው አዲስ መጽሐፍ ለመጻፍ ፈልጎ ስለነበር ሥራ አስኪያጁን ቃለ መጠይቅ አድርጎ ነበር::

Gender: Male

Is there bias in Amharic - English translation above?

○ Yes, there is gender bias

● No gender bias in translation

How do you evaluate the quality of the translation

□ There is an issue in translating the sentence

□ There is an issue in translating the profession word

Figure 3: The Potato annotation GUI for the evaluation annotation.

For human evaluation, first, each sentence is evaluated (biased or not biased) by three different native-speaker experts by showing the English and translated versions of a language in parallel. After each sentence in each of the three languages is evaluated by three evaluators, we decide if a sentence is biased or not by taking the majority vote of the three evaluators.

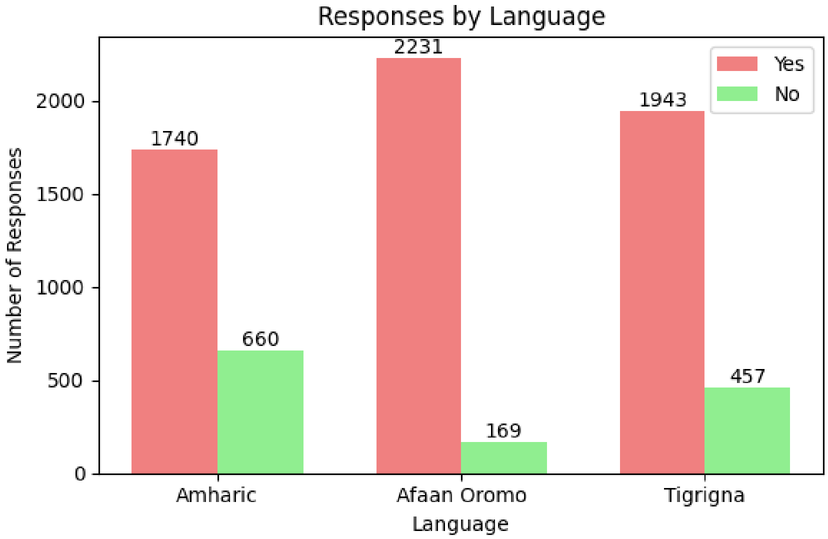


Figure 4: Illustration of the Google Translation Gender Bias test dataset human evaluation result. ”Yes” and ”no” are the answers to the question, ”Is there bias in the translation?”. ”Yes” means the sentence is biased when translated to a specified language. ”No” is no biased in the translated sentence; the sentence is correctly translated.

5 Result and Analysis

Figure 4 visualizes the distribution of responses across three language categories. For each language, the number of ”Yes” (biased sentences) and ”No” (not biased) responses is displayed using grouped bars. The lightcoral bars represent the number of ”Yes” responses, while the lightgreen bars represent the number of ”No” responses.

In general, Afan Oromo received the highest number of ”Yes” responses, followed by Tigrinya and Amharic. Conversely, Amharic received the highest number of ”No” responses, followed by Tigrinya and Afan Oromo. Figure 4 provides a clear comparison of responses across different language categories, allowing for insights into the distribution of responses within each language. It presents the gender bias across various language groups, delineating respondents’ perceptions regarding the presence or absence of gender bias within each language category.

The data in Table 1 underscores the disparate perceptions of gender bias among respondents across different linguistic backgrounds. Particularly notable is the significantly higher percentage (92.96%) of Afan Oromo respondents who perceive gender bias compared to other language groups, with only 7.04% indicating otherwise. Similarly, in the Amharic group, approximately 72.50% of respondents perceived gender bias, contrasting with 27.50% who did not. Likewise, in the Tigrinya group, the majority (80.96%) perceived gender bias, while 19.04% expressed no bias. These findings reveal distinct patterns of perception regarding gender bias across language groups, suggesting potential implications for addressing and understanding gender bias within these communities.

Table 1 outlines translation issues across languages, categorized into ”Translating the sentence issue” and ”Professional word translation issue.” Amharic records the highest instances of sentence translation issues at 1429, followed by Tigrinya with 936, and Afan Oromo with 918. Regarding professional word translation, Afan Oromo leads with 612 instances, trailed by Tigrinya at 475, and Amharic at 258. Interestingly, Tigrinya exhibits the fewest reported issues overall, with 619 respondents indicating no translation issues, compared to 510 for Amharic and 421 for Afan Oromo. Conversely, Amharic shows the highest incidence of respondents facing both types of issues at 203, followed by Afan Oromo at 449, and Tigrinya at 370. This data underscores the diverse challenges faced in translation across languages and

	Amharic	Tigrinya	Afan Oromo
There is an issue in translating the sentence	1429	936	918
There is an issue in translating the profession	258	475	612
No issue	510	619	421
Both issues	203	370	449
Total	2400	2400	2400

Table 1: Translation Issues by Language

provides valuable insights for enhancing translation quality and addressing language-specific obstacles.

Table 2: Automatic Evaluation Results

Language	NLLB		Google MT	
	SacreBleu	Chrf++	SacreBleu	Chrf++
Eng- Amh	3.48	23.73	16.13	47.97
Amh- Eng	21.87	50.76	-	
Eng- orm	4.85	34.85	22.96	56.71
orm- Eng	17.80	41.63	-	
Eng- tig	3.89	18.52	16.00	38.00
tig- Eng	20.01	43.91	-	

Table 2 presents the evaluation results for NLLB and Google's translation models in selected language pairs. The table is divided into rows representing different language pairs and columns representing the specific evaluation metrics. Each language pair is evaluated in both translation directions (e.g., Eng-Amh and Amh-Eng), providing insights into machine translation systems' translation quality and performance across various linguistic contexts.

The result shows that the Google MT system outperformed the NLLB model when using English as the source language in both evaluation metrics. This shows that translating English sentences into the target Ethiopian language is challenging for the model. On the other hand, the Google MT system showed better results compared to the NLLB model when translating English sentences into target Ethiopian languages. We observed better performance results when using English as the target language than when using it as the source language in the NLLB model. From this, we can see that for low-resource languages, publicly available MT models like NLLB are struggling to predict the correct translation when using English as the source language.

6 Conclusion and Future Works

In this paper, we presented a methodology for evaluating gender bias in machine translation systems. Addressing gender bias in machine translation is crucial for creating more equitable and accurate language processing systems. First, we have prepared the first gold test dataset in MT for low-resource Ethiopian languages: Amharic, Tigrinya, and Afan Oromo. With this test dataset, we did a human-level gender bias evaluation of the Google transition for the given language pairs. The evaluation result shows that 92.96% of Eng-Oro, 80.96% of Eng-Tig, and 72.50% of Eng-Amh language pairs translations have a gender bias. In addition, we used the automatic evaluation to measure the translation quality of the currently available translation tools that support Amharic, Tigrinya, and Afan Oromo languages. Our findings highlight the need for further research and development efforts to mitigate gender bias and promote gender-inclusive language translation. We observed that this work can be scaled up and used as a benchmark for other low-resource languages. In future work, we will develop an automatic gender bias evaluation method that can be used for low-resource languages.

In addition, We will prepare a gender-balanced dataset for the given language, and we will fine-tune the currently available MT tools.

7 Limitations

The cost and time constraints limit our work to only three language pairs. The sources of gender biases in NLP are different such as the nature of the language gender unbalanced occupational names in the dataset, and gender unbalanced pronouns in the dataset. This work only focuses on solving unbalanced occupational names and pronouns and is limited to sentence level.

8 Acknowledgment

This work is supported by sub-grant GRAIN, which has provided essential financial support for the completion of this research project. We are grateful for their contribution.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In 9th Annual conference of the special interest group for computing, information and society, 2017.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. ACL Anthology, pp. 5454–5476, July 2020. doi: 10.18653/v1/2020.acl-main.485.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On measuring gender bias in translation of gender-neutral pronouns. arXiv preprint arXiv:1905.11684, 2019.

Kate Crawford. The trouble with bias., 2017.

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. arXiv preprint arXiv:2308.03051, 2023.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. POTATO: The portable text annotation tool. In Wanxiang Che and Ekaterina Shutova (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 327–337, Abu Dhabi, UAE, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.33. URL https://aclanthology.org/2022.emnlp-demos.33.

Maja Popović. chrf++: words helping character n-grams. In Proceedings of the second conference on machine translation, pp. 612–618, 2017.

Matt Post. A call for clarity in reporting bleu scores. arXiv preprint arXiv:1804.08771, 2018.

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

Ratish Puduppully, Raj Dabre, Ai Ti Aw, and Nancy F Chen. Decomposed prompting for machine translation between related languages using large language models. arXiv preprint arXiv:2305.13085, 2023.

Adithya Renduchintala, Denise Díaz, Kenneth Heafield, Xian Li, and Mona Diab. Gender bias amplification during Speed-Quality optimization in Neural Machine Translation. ACL Anthology, pp. 99–109, August 2021. doi: 10.18653/v1/2021.acl-short.15.

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

Shanya Sharma, Manan Dey, and Koustuv Sinha. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 1968–1984, Abu Dhabi, United Arab Emirates, Dec 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.143.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. arXiv preprint arXiv:1906.00591, 2019.

Zeerak Talat, Aurelie Neveol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. ACL Anthology, pp. 26–41, May 2022. doi: 10.18653/v1/2022.bigscience-1.3.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejía González, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672, 2022. doi: 10.48550/arXiv.2207.04672.

Eric Peter Wairagala, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. Gender bias evaluation in luganda-english machine translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pp. 274–286, 2022.

9 Appendix

9.1 Appendix: Human-level Evaluation Guideline

Hello everyone,

We are excited to invite you to participate in an important evaluation task aimed at assessing gender bias in Google Translation from English into Amharic, Afan Oromo, and Tigrinya. As well as, to evaluate the quality of the overall translation, you are asked to evaluate the translation issue of the whole sentence and whether there is an issue with professional name translation only. As an evaluator, your valuable insights will help us ensure that translations accurately reflect gender inclusivity and professionalism. By carefully reviewing each sentence pair and considering both gender specification and professional terminology, you will play a pivotal role in enhancing translation quality. Your diligent efforts in evaluating 400 sentences will contribute to creating more inclusive and accurate translations. Thank you for your time and cooperation in this endeavor. Let's work together to promote fairness and accuracy in translation.

Evaluation Task: Gender Bias in Google Translation from English into Amharic, Afan Oromo, and Tigrinya

1. Login Credentials: Use the provided username and password to access the evaluation platform.

2. Accessing the Task: Open the designated link on your preferred device, whether mobile or computer.
3. Evaluation Procedure:
   - Reviewing Sentences: Carefully examine each provided sentence in English alongside its translation into Amharic, Afan Oromo, or Tigrinya.
   - Identifying Gender Bias: Determine the presence of gender bias by considering two factors:
     - Gender Section: Assess whether the translated gender (female or male) aligns with the gender specified in the original sentence.
     - Professional Words: Check if professional terms are translated with the same gender as provided in the original sentence.
   - Selecting Response: Choose ”Yes, there is gender bias” if bias is detected, or ”No, gender bias in translation” if not.
   - Evaluate the quality of translation: Select the first check box “There is an issue in translating the sentence” if there is an issue in overall translation or/and select the second check box “There is an issue in translating the profession word”.
   - Moving to Next Sentence: Click the ”Next” button after making your assessment to proceed to the next set of sentences.
4. Total Sentences: The evaluation task consists of 400 sentences to be assessed.
5. Completion and Compensation: Upon completing the evaluation of all 400 sentences, compensation will be provided according to the prearranged agreement.

We appreciate your dedication and cooperation in contributing to this evaluation task. Your feedback is crucial for improving translation quality and mitigating gender bias.

9.2 Appendix: List of Pronouns in English, Amharic, Tigrinya, Afan Oromo

In Figure 5, we give a list of the pronouns in English, Amharic, Tigrinya, and Afan Oromo.

English	Amharic	Tigrinya	Afaan Oromoo
I	**እኔ** (əne)	**ኣነ** (anä)	ana, na
We	**እኛ** (əñña)	**ንሕና** (nəḥəna)	nu
You (M. sg.)	**አንተ** (antä)	**ንስኻ** (nəssəxa)	
You (F. sg.)	**አንቺ** (anči)	**ንስኺ** (nəssəxi)	
You (sg.)			si
You (R)	**እርስዎ** (ərswo)		
You (F, R)		**ንሰን/ንስኽን** (nsen/nskhn)	
You (M, R)		**ንሶም/ንስኹም** (nsom/nskhum)	
You (pl.)	**እናንተ** (ənnantä)		isin
You (M. pl.)		**ንስኻትኩም** (nəssəxatkum)	
You (F. pl.)		**ንስኻትክን** (nəssəxatkən)	
He	**እሱ** (əssu)	**ንሱ** (nəssu)	isa
She	**እሷ** (əsswa)	**ንሳ** (nəssa)	isii, ishii, isee, ishee
S/he (R)	**እሳቸው** (əssaččäw)		
She (R)		**ንሰን** (nsen)	
He (R)		**ንሶም** (nsom)	
They	**እነሱ** (ənnässu)		isaan
They (M.)		**ንሳቶም** (nəssatom)	
They (F.)		**ንሳተን** (nəssatän)	

Figure 5: Pronouns in English, Amharic, Tigrinya, and Afan Oromo