Overview of the Shared Task on Machine Translation Gender Bias Evaluation with Multilingual Holistic Bias

Marta R. Costa-jussà[†], Pierre Andrews[†], Christine Basta[⋆], Juan Ciro[‡], Agnieszka Falenska[§], Seraphina Goldfarb-Tarrant[¶], Rafael Mosquera[‡], Debora Nozza[◇], Eduardo Sánchez[†]

[†]FAIR, Meta [⋆]Alexandria University [‡]Dynabench
[¶]Cohere / University of Edinburgh [§]Stuttgart University [◇]Bocconi University
[†]{costajussa,mortimer,esanchez}@meta.com [⋆]christine.basta@alexu.edu.eg
[‡]{rafael.mosquera,junciro}@mlcommons.org [◇]debora.nozza@unibocconi.it

Abstract

We describe the details of the Shared Task of the 5th ACL Workshop on Gender Bias in Natural Language Processing (GeBNLP 2024). The task uses Multilingual HolisticBias dataset to investigate the quality of Machine Translation systems on a particular case of gender robustness. We report baseline results as well as the results of the first participants. The shared task will be permanently available in the Dynabench platform.

1 Introduction

Gender bias poses challenges across various aspects of automatic translation. These challenges include preserving correct pronouns, understanding the correct gendered context, and relating adjectives and professions to the proper gender. The issue becomes even more complex when considering multilingual translation, especially for low-resource languages. The GeBNLP 2024 workshop aims to raise awareness of these challenges by introducing a dedicated shared task for investigating translation quality using the Multilingual HolisticBias dataset (Costa-jussà et al., 2023). This initiative seeks to foster a community-driven effort and long-term solutions toward improving gender representation in machine translation. We encourage researchers to contribute their expertise, not just for the workshop but for the ongoing pursuit of advancements in this field.

2 Motivation

The development of gender (Stanovsky et al., 2019; Renduchintala et al., 2021; Levy et al., 2021; Costa-jussà et al., 2022; Renduchintala and Williams, 2022; Savoldi et al., 2021; Alhafni et al., 2022; Attanasio et al., 2023) or demographic-specific (Costa-jussà et al., 2023) datasets has raised the interest in evaluating Natural Language Processing (NLP) models beyond standard quality terms. In Machine Translation (MT), gender bias is observed when translations show errors in linguistic gender determination despite the fact that there are sufficient gender clues in the source content for a system to infer the correct gendered forms. To illustrate this phenomenon, sentence (1) in Table 1 does not contain enough linguistic clues for a translation system to decide which gendered form should be used when translating into a language where the word for doctor is gendered. Sentence (2) in Table 1, however, includes a gendered pronoun which most likely has the word doctor as its antecedent. Sentence (3) in Table 1 shows two variations of the exact sentence differing only in the gender inflection.

(1)	I didn't feel well, so I made an appointment with my doctor.
(2)	My doctor is very attentive to *her* patients' needs.
(3)	Mi amig*a* es *una ama* de casa.
	Mi amig*o* es *un amo* de casa.
	[English: *My friend is a homemaker.*]

Table 1: Gender phenomenons' examples

Gender bias is observed when an MT system produces the wrong gendered form when translating sentence (2) into a language that uses distinct gendered forms for the word doctor. On the contrary, a single error in the translation of an utterance such as sentence (1) would not be sufficient to conclude that gender bias exists in the model; doing so would take consistently observing one linguistic gender over another. Finally, a lack of robustness would be shown if the translation quality differed for the two sentences in (3). It has previously been hypothesized that one possible source of gender bias in MT is gender representation imbalance in large training and evaluation data sets, e.g., Costa-jussà

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 399–404
August 16, 2024 ©2024 Association for Computational Linguistics

et al. (2022); Qian et al. (2022).

Our task goes beyond previous gender bias MT evaluation efforts, such as Stanovsky et al. (2019); Renduchintala et al. (2021); Levy et al. (2021); Costa-jussà et al. (2022); Renduchintala and Williams (2022); Savoldi et al. (2021); Alhafni et al. (2022); Attanasio et al. (2023), to name a few, mainly by increasing the number of languages and fairly comparing three main gender MT issues which are gender-specific, gender robustness, and unambiguous gender (see Section 4).

3 Goals

The goals of the Multilingual HolisticBias task as part of the 5th ACL Workshop on Gender Bias in Natural Language Processing are:

- To investigate the quality of MT systems on a particular case of gender preservation for tens of languages.
- To examine and understand special gender challenges in translating in different language families.
- To investigate the performance of gender translation of low-resource, morphologically rich languages.
- To open to the community the first challenge of this kind.
- To generate up-to-date performance numbers in order to provide a basis for comparison in future research.
- To investigate the usefulness of multilingual and language resources.
- To encourage beginners and established research groups to participate and interchange discussions.

4 Multilingual HolisticBias Task

We propose to evaluate three cases of gender bias: gender-specific, gender robustness, and unambiguous gender translation.

4.1 Task 1: Gender-specific

In the English-to-X translation direction, we evaluate the capacity of MT systems to generate gender-specific translations from English neutral inputs (e.g., *I didn't feel well, so I made an appointment with my doctor*). This can be illustrated by the fact that MT models systematically translate neutral source sentences into masculine or feminine depending on the stereotypical usage of the word (e.g., *homemakers* into *amas de casa*, which is the feminine form in Spanish and *doctors* into *médicos*, which is the masculine form in Spanish).

4.2 Task 2: Gender Robustness

In the X-to-English translation direction, we compare the robustness of the model when the source input only differs in gender (masculine or feminine), e.g., Spanish *Mi amiga es una ama de casa* and *Mi amigo es un amo de casa* (*My friend is a homemaker*).

4.3 Task 3: Unambiguous Gender

In the X-to-X translation direction, we evaluate the unambiguous gender translation across languages and without being English-centric, e.g, Spanish-to-Catalan: *Mi amiga es una ama de casa* is translated into *La meva amiga és una mestressa de casa*.

4.4 Submission details

Data This task is based on Multilingual HolisticBias (Costa-jussà et al., 2023) – the first multilingual extension of HolisticBias (Smith et al., 2022) which covers tens of languages.

X Languages In addition to English, our challenge covers 26 languages: Modern Standard Arabic, Belarusian, Bulgarian, Catalan, Czech, Danish, German, French, Italian, Lithuanian, Standard Latvian, Marathi, Dutch, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Tamil, Thai, Ukrainian, Urdu

Evaluation The challenge is evaluated using automatic metrics: BLASER (Chen et al., 2022) and ChrF (Popović, 2015). Evaluation criteria are in terms of overall translation quality and difference in performance for masculine (m) and feminine (f) sets. Leaderboard ranking will be made using the following combination of BLASER and ChrF:

$$\mathrm{GES} = 20 \times \frac{\mathrm{average}(\mathrm{BLASER}_m, \mathrm{BLASER}_f)}{1 + |\mathrm{ChrF}_m - \mathrm{ChrF}_f|}$$

where $\mathrm{BLASER}_{m/f}$ and $\mathrm{ChrF}_{m/f}$ use masculine or feminine references. The metric is a percentage and it should be maximized. The numerator evaluates the semantic quality, and the denominator evaluates the difference in ChrF between using masculine or feminine references. We call this metric *Gender Equity Score* (GES).

Submission platform We use the Dynabench platform for all tasks.

Baseline systems We use open-source NLLB models: NLLB-600M and NLLB-3.3B (NLLB Team et al., 2022).

Participants This edition of the shared task received only one submission. The participants expanded the DAMA framework (*Debiasing Algorithm through Model Adaptation*, Limisiewicz et al. (2024)) to be applicable in the multilingual translation task. DAMA proposes a method for identifying and mitigating gender bias in language models. In the original paper, the researchers discovered that specific layers of LLaMA (Touvron et al., 2023) are responsible for gender bias and intervened on these layers by modifying their weights to nullify their effect. The shared task participants replicated the same intervention on ALMA-R (Xu et al., 2024), an MT-specific LLM that performs better than previous LLMs, including GPT-3.5. The findings showed that DAMA could reduce gender bias in translation without compromising quality in the overall domain. However, the suggested approach is susceptible to the introduction of bias in the prompts.

5 Results

This section reports results for the two baselines and the submitted system. We provide results only for Task 2 (gender robustness), which received the submission.

Table 2 shows the results. We can notice that NLLB-3.3B performs better in terms of translation quality and GES. Note that in this case, higher GES shows that there is less translation quality variation when gender varies in the input. We observe that the difference across models differs across languages, with larger discrepancies in languages like Arabic or Thai and smaller in languages like German or Spanish. For a few languages, e.g., Catalan or Romanian, GES is higher for NLLB-600M. On average, NLLB-3.3B scores higher in GES by more than 0.5. The result is coherent with previous research that shows that by just increasing the translation quality of the model, gender robustness increases (Communication et al., 2023).

Finally, Table 3 shows the participant entry compared to the best baseline. We observe that the strongest baseline surpasses DAMA models in terms of translation quality (absolute ChrF or BLASER) and GES.

6 Final Remarks

This paper introduces the Multilingual HolisticBias Dynabench task[1] which has been launched in the context of the 5th ACL Worskhop on Gender Bias in NLP[2]. This task will remain open for participation. At the moment of the preparation of this paper, we have received a single participation which evaluates the mitigation strategy of DAMA (Limisiewicz et al., 2024) for gender robustness. We are also reporting strong baseline results with NLLB models for this particular task. However, we do not include baselines for gender-specification and unambiguous gender, which is left as further work.

We are looking forward to receiving more submissions in the near future. Also notice that an extension of the Multilingual HolisticBias dataset is currently going on and released (Tan et al., 2024).

Limitations

Our shared task shares the same limitations as the Multilingual HolisticBias dataset on which it is based (Costa-jussà et al., 2023).

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France. European Language Resources Association.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne,

[1]https://dynabench.org/tasks/multilingual-holistic-bias
[2]https://genderbiasnlp.talp.cat/

En-X		System	ChrF$_m$	ChrF$_f$	BLASER$_m$	BLASER$_f$	GES (↑)
arb_Arab	Modern Standard Arabic	NLLB-600M	0.4467	0.3486	4.0532	4.0175	74.2970
		NLLB-3.3B	0.5187	0.4099	4.2298	4.1852	76.8801
bel_Cyrl	Belarusian	NLLB-600M	0.2924	0.2852	3.7345	3.7167	73.1841
		NLLB-3.3B	0.3065	0.2922	3.7780	3.7566	73.9321
bul_Cyrl	Bulgarian	NLLB-600M	0.6376	0.6022	4.2443	4.2013	81.1622
		NLLB-3.3B	0.6573	0.6172	4.2883	4.2430	81.8014
cat_Latn	Catalan	NLLB-600M	0.6228	0.5254	4.3043	4.2536	78.3086
		NLLB-3.3B	0.6817	0.5743	4.3201	4.2664	77.7670
ces_Latn	Czech	NLLB-600M	0.4754	0.4600	4.2660	4.2263	83.0464
		NLLB-3.3B	0.4969	0.4778	4.3144	4.2719	83.6865
dan_Latn	Danish	NLLB-600M	0.6973	0.6547	4.2790	4.2446	81.8265
		NLLB-3.3B	0.7057	0.6612	4.3169	4.2795	82.5867
deu_Latn	German	NLLB-600M	0.4397	0.4794	4.3896	4.3490	84.1863
		NLLB-3.3B	0.4933	0.4528	4.4124	4.3697	84.5534
fra_Latn	French	NLLB-600M	0.6934	0.6581	4.4049	4.3807	84.6725
		NLLB-3.3B	0.7023	0.6656	4.4411	4.4155	85.2995
lit_Latn	Lithuanian	NLLB-600M	0.4794	0.4135	4.1266	4.1037	77.1694
		NLLB-3.3B	0.5336	0.4642	4.1645	4.1399	77.4397
lvs_Latn	Standard Latvian	NLLB-600M	0.4579	0.3986	3.9206	3.8685	73.7525
		NLLB-3.3B	0.5012	0.4416	4.0078	3.9543	75.2191
mar_Deva	Marathi	NLLB-600M	0.4797	0.4165	4.1501	4.1285	78.1608
		NLLB-3.3B	0.5256	0.4719	4.1966	4.1812	79.8954
nld_Latn	Dutch	NLLB-600M	0.5963	0.5590	4.3182	4.2791	82.9111
		NLLB-3.3B	0.6214	0.5836	4.3257	4.2845	83.0787
por_Latn	Portuguese	NLLB-600M	0.6122	0.5727	4.4257	4.3912	84.7750
		NLLB-3.3B	0.6372	0.5949	4.4750	4.4377	85.5887
ron_Latn	Romanian	NLLB-600M	0.5915	0.5562	4.3396	4.2989	82.4291
		NLLB-3.3B	05998.	0.5662	4.3788	4.3397	83.2860
rus_Cyrl	Russian	NLLB-600M	0.5483	0.5065	4.4017	4.3696	83.9947
		NLLB-3.3B	0.5635	0.5171	4.4679	4.4343	84.7446
slk_Latn	Slovak	NLLB-600M	0.6345	0.5453	4.3105	4.2475	79.0300
		NLLB-3.3B	0.6407	0.5474	4.3458	4.2775	79.3744
slv_Latn	Slovenian	NLLB-600M	0.5028	0.4531	4.0678	4.0138	76.7034
		NLLB-3.3B	0.5418	0.4963	4.1354	4.0832	77.2355
spa_Latn	Spanish	NLLB-600M	0.7543	0.6582	4.5410	4.4594	82.1978
		NLLB-3.3B	0.8024	0.6952	4.5801	4.4900	82.4332
swe_Latn	Swedish	NLLB-600M	0.6415	0.5876	4.2585	4.2226	80.4565
		NLLB-3.3B	0.6588	0.6034	4.3032	4.2652	81.1967
tam_Taml	Tamil	NLLB-600M	0.4309	0.4178	4.1646	4.1093	81.2719
		NLLB-3.3B	0.4488	0.4362	4.1792	4.1548	82.1098
tha_Thai	Thai	NLLB-600M	0.3335	0.4162	3.8589	3.8636	71.6030
		NLLB-3.3B	0.3833	0.3810	3.9551	3.9551	73.8386
ukr_Cyrl	Ukrainian	NLLB-600M	0.4166	0.4004	4.2594	4.2227	82.9483
		NLLB-3.3B	0.4640	0.4441	4.3106	4.2722	83.7743
urd_Arab	Urdu	NLLB-600M	0.3906	0.3489	4.1490	4.1026	79.2289
		NLLB-3.3B	0.4049	0.3632	4.1535	4.1071	79.3058
avg		NLLB-600M	0.5331	0.4962	4.2234	4.1842	80.1372
		NLLB-3.3B	0.5604	0.5112	4.2644	4.2245	**80.6534**

Table 2: Results for Task 2 Gender Robustness with NLLB-600M and NLLB-3.3B. Best averaged results in bold.

Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoari-

En-X		System	ChrF$_m$	ChrF$_f$	BLASER$_m$	BLASER$_f$	GES
ces_Latn	Czech	NLLB-3.3B	0.4969	0.4778	4.3144	4.2719	**83.6865**
		DAMA	0.4673	0.4489	4.1903	4.1484	81.6847
deu_Latn	German	NLLB-3.3B	0.4933	0.4528	4.4124	4.3697	**84.5534**
		DAMA	0.5175	0.4797	4.3832	4.3422	84.1084
rus_Cyrl	Russian	NLLB-3.3B	0.5635	0.5171	4.4679	4.4343	**84.7446**
		DAMA	0.4592	0.4114	4.2531	4.2214	81.1677

Table 3: Results from 2024 single entry participation compared to the strongest baseline (NLLB-3.3B). Best results in bold.

son, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2022. Evaluating gender bias in speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France. European Language Resources Association.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Gender bias in multilingual neural machine translation: The architecture matters.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomasz Limisiewicz, David Mareček, and Tomáš Musil. 2024. Debiasing algorithm through model adaptation. In *The Twelfth International Conference on Learning Representations*.

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, Loic Barrault, Gabriel Mejia Gonzalez, 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. 2022. No language left behind: Scaling human-centered machine translation.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. 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.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. 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.

Xiaoqing Ellen Tan, Prangthip Hansanti, Carleigh Wood, Bokai Yu, Christophe Ropers, and Marta R. Costa-jussà. 2024. Towards massive multilingual holistic bias.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.