Applying Intrinsic Debiasing on Downstream Tasks: Challenges and Considerations for Machine Translation

Bar Iluz[♣] **Yanai Elazar**[♢,♠] **Asaf Yehudai**[♣] **Gabriel Stanovsky**[♣,♢]
[♣] School of Computer Science, The Hebrew University of Jerusalem
[♢] Allen Institute for AI
[♠] University of Washington
bar.iluz@mail.huji.ac.il

Abstract

Most works on gender bias focus on intrinsic bias — removing traces of information about a protected group from the model's internal representation. However, these works are often disconnected from the impact of such debiasing on downstream applications, which is the main motivation for debiasing in the first place. In this work, we systematically test how methods for intrinsic debiasing affect neural machine translation models, by measuring the extrinsic bias of such systems under different design choices. We highlight three challenges and mismatches between the debiasing techniques and their end-goal usage, including the choice of embeddings to debias, the mismatch between words and sub-word tokens debiasing, and the effect of translating from English to different target languages. We find that these considerations have a significant impact on downstream performance and the success of debiasing.[1]

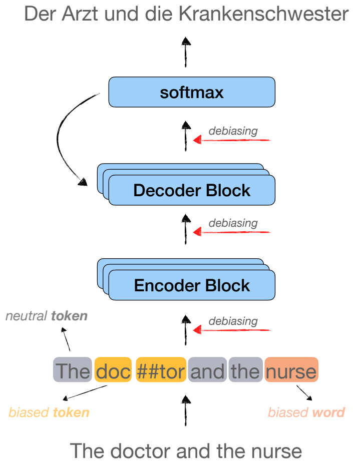

Figure 1: A schematic view of a neural machine translation system, highlighting different possibilities for applying intrinsic debiasing techniques. We examine three considerations: (1) where to apply the debiasing; (2) which tokens to apply the debiasing to (e.g. only gender-indicative words or the entire vocabulary); and (3) the effect of different target languages.

1 Introduction

Natural language processing models were shown to over-rely and over-represent gender stereotypes.[2] These can typically be found in their internal representation or predictions. For example, consider the following sentence:

coref

(1) The **doctor** asked the **nurse** to help *her* in the procedure.

Inferring that *her* refers to the nurse rather than the doctor may indicate that the model is biased. A useful distinction of model's biases was proposed by Goldfarb-Tarrant et al. (2021a) and Cao et al. (2022a): *Intrinsic bias* typically manifests in the geometry of the model's embeddings. For example, finding that stereotypically female occupations (e.g., "nurse", "receptionist") are grouped together in the embedding space, while stereotypically male occupations (e.g., "doctor", "CEO") are closer to each other (Gonen and Goldberg, 2019). *Extrinsic bias* on the other hand is measured in downstream tasks. For instance, in machine translation (MT), which is the focus of this work, a biased model may translate Example (1) to Spanish using a masculine inflection for the word "doctor", even though a human translator is likely to use a feminine inflection (Stanovsky et al., 2019). Intrinsic and extrinsic bias do not necessarily correlate (Goldfarb-Tarrant et al., 2021b; Cao et al., 2022b; Dawkins, 2021b), and biases might reoccur when applying debiased models on other tasks (Orgad et al., 2022).

In this work, we identify a gap in the literature between intrinsic bias mitigation and its influence on downstream tasks. Namely, while *extrinsic*

[1] We release our code at: https://github.com/bariluz93/intrinsic-debiasing-performance-on-NMT

[2] Throughout this work we refer to morphological gender, and specifically to masculine and feminine pronouns as captured in earlier work. We note that future important work can extend our work beyond these pronouns to e.g., neo-pronouns (Lauscher et al., 2022).

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14914–14921
November 12-16, 2024 ©2024 Association for Computational Linguistics

bias may affect human users in a variety of applications, debiasing techniques often focus only on *intrinsic measures*, aiming to obfuscate gender from pretrained embeddings (Bolukbasi et al., 2016; Elazar and Goldberg, 2018; Ravfogel et al., 2020). These approaches leave many unanswered questions when deploying them within a complex downstream model for specific tasks.

As shown in Figure 1, we systemically explore three fundamental challenges when integrating intrinsic debiasing techniques within complex open-source neural MT architectures. We find that different design choices lead to a wide difference in extrinsic bias as well as task performance.

First, we explore different approaches to cope with discrepancies between different tokenization strategies. While intrinsic debiasing is largely performed over complete words from a fixed dictionary, modern MT requires mapping those onto sub-word elements determined via a data-dependant tokenizer. We find that debiasing only complete words outperforms a more naive debiasing of all sub-word tokens. Second, several word embedding tables could be debiased within an MT system. Therefore, a preliminary architectural question is which of them to debias. We explore various combinations, finding the optimal configuration depends on the intrinsic debiasing technique. Third, We explore the effects of debiasing a translation model over three target languages (Hebrew, German, and Russian). While all three encode morphological noun gender, they differ in script, typology, and morphology. We find that an important factor for debiasing efficiency is the number of words represented as single tokens, a property determined both by the language's morphological properties as well as its sampled distribution in the tokenizer training data.

Taken together, our results suggest that extrinsic debiasing involves many interdependent challenges which cannot be inferred from an intrinsic outlook. We hope our work will promote more research on combining intrinsic debiasing methods to downstream tasks to produce extrinsically fairer MT models.

2 Background

There is an abundance of debiasing methods in the field (Wang et al., 2021; Schick et al., 2021; Shen et al., 2021; Dev and Phillips, 2019; Dev et al., 2021; Kaneko and Bollegala, 2021; Shao et al., 2023; Kumar et al., 2020; Dev et al., 2020; Dawkins, 2021a). Most of them focus on intrinsic debiasing. We focus on three prominent methods, outlined below. Importantly, all of these methods learn a transformation that can be applied to arbitrary vectors, once the model has finished training, and all were tested mostly intrinsically.

Intrinsic debiasing methods. We experiment with three methods: (1) Hard-Debiasing (Bolukbasi et al., 2016) removes a gender subspace via a Principal Component Analysis (PCA) of predetermined word pairs which are considered as indicative of gender; (2) INLP (Ravfogel et al., 2020) learns the direction of the gender subspace rather than using a predefined list of words; and (3) LEACE (Belrose et al., 2023) which prevents all linear classifiers from detecting a guarded concept. A key difference between the methods is that Hard-Debiasing is non-linear and non-exhaustive, leaving stereotypical information after its' application (Gonen and Goldberg, 2019). In contrast, INLP and LEACE are linear and exhaustive; after applying INLP, stereotypical information can't be extracted with a specific linear classifier, and after applying LEACE, it can't be extracted with any linear classifiers.[3]

The effect of debiasing on NMT. Most related to our work, Escudé Font and Costa-jussà (2019) explored the impact of debiasing methods on an English-to-Spanish MT task. However, they tested the MT models only on simple synthetic data, while here we focus on complex data reflecting real biases, and explore various design choices.

3 Integrating Intrinsic Debiasing in MT

We examine debiasing methods within the popular encoder-decoder approach to MT, as shown in Figure 1. Next, we describe the different research questions addressed in our setup.

Which *embedding* to debias? An encoder-decoder model has multiple embedding tables that can be intrinsically debiased: (1) the input matrix of the encoder; (2) the input matrix of the decoder; and (3) the output of the decoder, usually before the softmax layer.[4] We employ different intrinsic

[3] Although a log-linear model can reconstruct some information under certain assumptions (Ravfogel et al., 2023).

[4] In a complex system, such as the transformer encoder-decoder architecture, the representations after each transformer layer and within each layer can be debiased as well. We leave the investigation of such debiasing to future work.

Language	Dataset Name	Dataset Size
Russian	newstest2019	1,997
German	newstest2012	3,003
Hebrew	TED dev	1,000

Table 1: Datasets used for evaluating different target languages. The Dataset Size describes the number of sentences in the dataset. Russian and German datasets are described in Choshen and Abend (2021)'s paper. The Hebrew dataset is based on the Opus TED talks dataset (Reimers and Gurevych, 2020).

debiasing techniques to each of these tables and evaluate their effect on downstream performance.

Which *words* to debias? Tokenization poses a challenge for extrinsic debiasing as it may introduce discrepancies between the intrinsically debiased elements (complete words) and the MT input model (sub-word tokens) (Iluz et al., 2023). We experiment with three different configurations: (1) *all-tokens*: debiases embeddings of all tokens in the model's vocabulary; (2) *n-token-profession*: debiases all embeddings of words that appear in a predefined set of professions from the WinoBias dataset (Zhao et al., 2018), even if they are split across multiple tokens, and (3) *1-token-profession*: debiases only the embeddings of the professions that align with the vocabulary of the debiasing technique, e.g., "nurse" is debiased only if it appears as a single token.

How does debiasing affect different *languages*? We experiment with three target languages that encode morphological gender for nouns, representing different typological features: (1) Hebrew, a Semitic language with abjad script, (2) Russian, a Slavic language with a Cyrillic script, and (3) German, a Germanic language with Latin alphabet.

4 Experiments

4.1 Experimental Setup

MT model. We make use of OPUS-MT (Tiedemann and Thottingal, 2020),[5] a transformer-based MT model built of 6 self-attention layers and 8 attention heads in the encoder and the decoder. The model was trained on Opus,[6] an open-source web-text dataset, which uses SentencePiece tokenization (Kudo and Richardson, 2018).

Target Language	German	Hebrew	Russian
no-debiasing	57.7	45.6	41.0
n-token-profession	60.9	48.3	41.0
1-token-profession	**61.9**	**48.4**	**41.2**

Table 2: Accuracy on different target languages when varying the tokens across two debiasing strategies: INLP and Hard-Debiasing. Presenting results for applying (1) the baseline (no-debiasing), (2) *n-token-profession*, debiasing tokens corresponding to professions that are tokenized into one or more tokens, and (3) *1-token-profession*, debiasing only professions that are tokenized into a single token. For brevity, each cell presents the the best performing choice of embedding table and debiasing method.

Metrics and evaluation data. For extrinsic debiasing measurement, we employ the automatic accuracy metric from Stanovsky et al. (2019), assessing the percentage of instances where the target entity retains its original gender from the English sentence, using morphological markers in the target language. We focus on the performance on the *anti-stereotypical* set of 1,584 sentences from WinoMT (Stanovsky et al., 2019). These consist of anti-stereotypical gender role assignments, such as the female doctor in Example 1, or the female lawyer in the following example "The lawyer looked into illegal accusations against the cashier, because she needed to understand the case." In addition, we approximate the translation quality before and after debiasing using BLEU (Papineni et al., 2002) on several parallel corpora described in Table 1, and manually evaluate the translations to corroborate our findings.

4.2 Results

Debiasing *1-token-profession* professions outperforms other approaches. Table 2 shows the gender translation accuracy when applying debiasing methods on different tokens.[7] For the three tested languages, debiasing only professions that are tokenized into single tokens improved the gender prediction the most. This hints that the sub-word tokens that compose a profession word do not hold the same gender information as the whole word.

The optimal embedding table to debias depends on the debiasing method. Table 3 shows the im-

[5] https://github.com/Helsinki-NLP/Opus-MT

[6] https://opus.nlpl.eu

[7] Excluding results for debiasing all tokens, as it led to garbled translations where automatic debiasing measures are irrelevant.

Embedding Table	Baseline	Hard-Debiasing	INLP	LEACE
Encoder Input	48.1	**49.6**	43.2	43.4
Decoder Input	48.1	48.0	50.0	**53.8**
Decoder Output	48.1	48.0	**50.7**	**53.8**

Table 3: Opus MT's gender prediction accuracy with intrinsic debiasing methods applied on different embedding tables. Each cell is averaged across our target languages (*de, he, ru*). Bold numbers represent best per debiasing method. The accuracy is measured by Stanovsky et al. (2019)'s method on their WinoMT dataset

provement in gender prediction averaged across languages when applied on different embedding tables. Hard-Debiasing improves gender prediction only when debiasing the encoder's inputs, while INLP and LEACE improves gender prediction accuracy the most when applied to the decoder output. This may be explained by INLP's and LEACE's linearity, which therefore works best at the end of the decoder, after all nonlinear layers, while Hard-Debiasing employs a non-linear PCA component.[8]

Results vary between languages. Debiasing has a positive impact on the accuracy of gender translation in both German and Hebrew, with German improving by 3.7 points and Hebrew by 2.8 points. In contrast, Russian did not see as much improvement (Table 2). The difference may be due to Russian's relatively rich morphology (e.g., it has 7 cases compared to 4 in German (Dryer and Haspelmath, 2013)), resulting in much fewer single-token professions (59% in Russian compared to 65% in Hebrew, and 83% in German).

LEACE and Hard-Debiasing do not significantly harm BLEU scores. Figure 2 shows the relationship between the difference in the gender prediction and the difference in BLEU. Hard-Debiasing and LEACE both have a small negative effect to the BLEU scores, while in comparison, INLP presents a trade-off between the improvement in gender prediction and the translation quality according to BLEU scores. This shows that INLP removes information which is important for the translation model, while LEACE (which was proved to be the minimal transformation needed to remove gender information) and Hard-Debiasing indeed preserve more of the information. In terms of the gender prediction accuracy, the best setting of Hard-Debiasing is when applied to the encoder, while INLP and LEACE improve the gender prediction the most when applied to the decoder outputs. LEACE performs better than INLP when applied on the decoder as it was designed to prevent all linear classifiers from detecting the guarded concept, while INLP learns to obfuscate only one linear classifier.

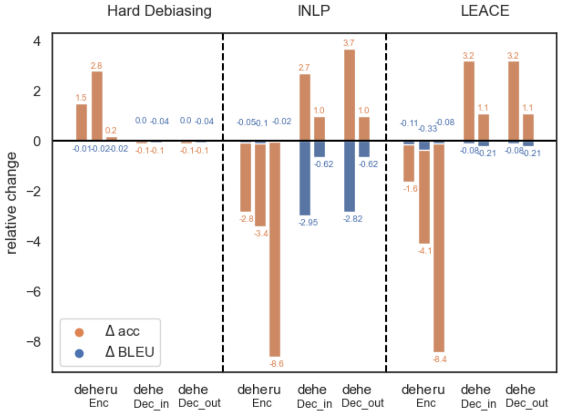

Figure 2: The relation between gender prediction accuracy difference (orange) and the BLEU difference (blue) between the original model (without any intervention) and the debiased model. The left part presents the results with Hard-Debiasing, INLP in the middle, and LEACE on the right. For each method, we present the results per each location (Encoder, Decoder-input, and Decoder-output), as well as each language).

Finally, all results are statistically significant with p-value < 0.05, see Appendix C for details.

Human evaluation shows that gender prediction is indeed improved with Hard-Debiasing. We manually annotate a portion of the translation to assess how well the automatic gender prediction metrics estimate real bias. We annotate the configuration of Hard-Debiasing which changes the translations the most compared to other methods: applying Hard-Debiasing to the encoder's input with the *1-token-profession* paradigm. Out of the 1,584 sentences in the dataset, 184 (11%) changed after the debiasing. 32% of the sentences that changed after the debiasing corrected the profession's gender prediction. These numbers are somewhat higher than what the automatic metrics suggest (26% improvement on the same setup). See Appendix A for additional details.

5 Conclusions and Future Work

We systematically explore different challenges and design choices when integrating intrinsic debiasing methods within complex machine translation sys-

[8] We tested debiasing all 8 combinations of the three embedding tables, but this did not change our findings.

tems. We find that it is better to debias only words representative of gender and correspond to single tokens. Furthermore, it is important to couple the debiasing method with the specific embedding table (e.g., encoder versus decoder), and that different target languages lead to vastly different results. Our findings do not suggest the "correct" approach to addressing bias in machine translation. Instead, they highlight the complexity of mitigating gender bias in large model architectures.

Dawkins et al. (2024) studied the effect of debiasing different transformer layers in BERT's next sentence prediction task. They found that interventions at the model's inner layers, including attention mechanisms, led to improvements in intrinsic bias mitigation. In future work, it would be interesting to test how debiasing the attention and inner layers of an MT model affects bias. Finally, Antoniak and Mimno (2021) evaluate the impact of different gender pairs defining the gender direction in various debiasing methods (e.g., Hard-Debiasing). They noticed that bias measurements can be affected by biases in the pairs themselves like cultural and cognitive biases. Future work can explore gender pair selection for the debiasing methods both in source and target languages, and examine how different pairs influence extrinsic bias.

Limitations

Our work explores the integration of debiasing within a complex machine translation system. As such, the space of possible combinations to explore is very large, including the embedding table to debias, the choice of target languages, their corresponding test corpora, the debiasing method to explore and their hyperparameter settings, and more. We systematically explore a subset of these options, which may hinder the generalizability of our specific results, e.g., which tokenization scheme works best. We encourage future work to re-examine our findings in other settings and possibly refine or amend them, while our main takeaway is the broader set of considerations which should be taken into account when debiasing complex, real-world systems. Additionally, to solve this task, machine translation systems need to also improve their coreference resolution abilities, which we did not examine here (Yehudai et al., 2023). Additionally, our work focuses on gender bias, but certain debiasing techniques are broad and can be used for other protected attributes, thus we aspire that our work will pave the way for exploring other attributes in future works.

Acknowledgements

We thank anonymous reviewers for their valuable comments. This work was partially supported by the Israeli Ministry of Science and Technology (grant no. 2088).

References

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *ArXiv preprint*, abs/2306.03819.

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 *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022a. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022b. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2021. Transition based graph decoder for neural machine translation. *ArXiv preprint*, abs/2101.12640.

Hillary Dawkins. 2021a. Marked attribute bias in natural language inference. *arXiv preprint arXiv:2109.14039*.

Hillary Dawkins. 2021b. Second order winobias (sowinobias) test set for latent gender bias detection in coreference resolution. *arXiv preprint arXiv:2109.14047*.

Hillary Dawkins, Isar Nejadgholi, Daniel Gillis, and Judi McCuaig. 2024. Projective methods for mitigating gender bias in pre-trained language models. *arXiv preprint arXiv:2403.18803*.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. Oscar: Orthogonal subspace correction and rectification of biases in word embeddings. *arXiv preprint arXiv:2007.00049*.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021a. 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.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021b. 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. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Bar Iluz, Tomasz Limisiewicz, Gabriel Stanovsky, and David Mareček. 2023. Exploring the impact of training data distribution and subword tokenization on gender bias in machine translation. *ArXiv preprint*, abs/2309.12491.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

H Orgad, S Goldfarb-Tarrant, and Y Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. corr, abs/2204.06827, 2022. doi: 10.48550. *ArXiv preprint*, abs/2204.06827.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2021. Contrastive learning for fair representations. *ArXiv preprint*, abs/2109.10645.

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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. 2021. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750, Online. Association for Computational Linguistics.

Asaf Yehudai, Arie Cattan, Omri Abend, and Gabriel Stanovsky. 2023. Evaluating and improving the coreference capabilities of machine translation models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 980–992, Dubrovnik, Croatia. 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.

A Human Evaluation

The evaluation proposed by Stanovsky et al. (2019) is automatic and thus bound to introduce some noise. To mitigate this concern, we perform a human evaluation which complements the automatic metric and provides a more thorough analysis of the results. The evaluation was applied to translations into Hebrew which were debiased according to the *1-token-profession* words (see Section 3), when specifically applied on encoder's input. We only analyze the sentences that changed in any way compared to the original (non debiased) translations. There were 184 out of 1584 sentences which changed (11%). 32% of the changed sentences improved the gender prediction (i.e the gender of a profession was incorrect in the original translation and correct in the debiased translation), 3% of the changed sentences were detracted (i.e the gender of a profession was correct in the original translation and incorrect in the debiased translation), and in the rest 65% the change did not affect the gender of the profession. In 23% out of the 65% of unaffected sentences, the gender of the other profession in the sentence was changed, for example in the sentence "The salesperson talked to the auditor because she was worried about the audit" the gender of salesperson did not change in the debiased translations, but the gender of the auditor was changed from male to female form. Note that this is not a mistake since the gender of auditor is unknown in the source sentence. This shows that the debiasing method affects a larger amount of professions which are not counted in the improvement of the model.

B Manual Annotation

To define the gender direction in the target language for both debias methods, we needed the translations of the 10 representative gender word pairs for each language. To get those pairs, we asked a native speaker of each of these languages to translate them into their language. In the case of a pair that is irrelevant to the target language (like Mary and John which are common male and female names in English but not in other languages), we asked them to adapt the pair to represent gender pairs in their language. The set of professions that we debias was also translated into the target languages by three native speakers in each language. The professions annotations were taken from Iluz et al. (2023). The translations of the 10 pairs were collected for four languages, German, Hebrew, Russian, and Spanish. [9]

C Statistical Significance

In order to determine the statistical significance of our findings, we employed McNemar's test, as recommended by Dror et al. (2018). McNemar's test is designed for models with binary labels, therefore it is suitable to test the gender bias scores where each sentence is classified as correct if the gender is accurately identified in the translation and incorrect otherwise. The null hypothesis for this test states that the marginal probability for each outcome is equal between the two algorithms being compared, indicating that the models are identical. In our case, the two models being compared are the original translation model and the debiased version. When concatenating results per debias method, we get that the results of Hard Debias are significant with p-value of 3.01E-07, and the results of INLP are significant with p-value of 9.65E-06. When comparing the results per embedding table to debias, we get that debiasing the encoder inputs is significant with p-value of 5.42E-10, debiasing the decoder inputs is significant with p-value of 0.016 and debiasing the decoder outputs is significant with p-value of 0.014. finally when concatenating all the results, we get that comparing the outputs of a debiased model to a the original model, the results are significant with p-value of 0.01.

[9] link for the 10 pairs datasets will be released upon publication.