Reducing Gender Bias in NMT with FUDGE

Tianshuai Lu
University of Zurich
tianshuai.lu@uzh.ch

Noëmi Aepli
University of Zurich
naepli@cl.uzh.ch

Annette Rios
University of Zurich
rios@cl.uzh.ch

Abstract

Gender bias appears in many neural machine translation (NMT) models and commercial translation software. Research has become more aware of this problem in recent years and there has been work on mitigating gender bias. However, the challenge of addressing gender bias in NMT persists. This work utilizes a controlled text generation method, Future Discriminators for Generation (FUDGE), to reduce the so-called *Speaking As* gender bias. This bias emerges when translating from English to a language that openly marks the gender of the speaker. We evaluate the model on MuST-SHE, a challenge set to specifically evaluate gender translation. The results demonstrate improvements in the translation accuracy of the feminine terms.

1 Introduction

When we talk about gender bias in neural machine translation (NMT), the first issue that comes to mind is stereotyping, e.g. associating the profession *doctor* with the male pronoun and *nurse* with the female pronoun. While this example does illustrate a clear instance of gender bias, Hardmeier et al. (2021) highlight that it is crucial to recognize that gender bias can manifest in various forms. It becomes essential to determine precisely what is considered harmful, the manner in which it is perceived as harmful, and the specific individuals or groups affected (Savoldi et al., 2021).

Current research on mitigating gender bias in MT often focuses on gender stereotypes (Stanovsky et al., 2019), translation errors due to speaker gender (Vanmassenhove et al., 2018), or pronoun translation (Loáiciga et al., 2017; Jwalapuram et al., 2020). Furthermore, the proposed methods are often only evaluated on BLEU (Papineni et al., 2002). However, BLEU evaluates on word level and is rather insensitive to specific linguistic phenomena that only affect a few words (Sennrich, 2017).

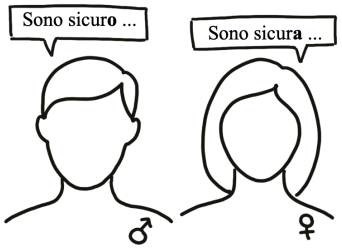

Figure 1: This work focuses on the *Speaking As* bias emerging when translating from English to a language that openly marks the **gender of the speaker** as here in Italian for instance: *sono sicuro/a ...* which translates to "I'm certain ..."

In this paper, we apply a controlled text generation method, Future Discriminators for Generation (FUDGE) (Yang and Klein, 2021), to mitigate the gender bias that arises when translating from English, a language that marks gender only on pronouns, to Italian, a language that openly marks the gender of the speaker in specific contexts. FUDGE has demonstrated its capabilities on many controlled text generation tasks, e.g. poetry couplet completion, topic-controlled language generation, and machine translation formality change. We further explore FUDGE's performance on a gender-controlled machine translation task.

Furthermore, instead of solely relying on BLEU

(Papineni et al., 2002) as an evaluation metric, we evaluate on MuST-SHE (Savoldi et al., 2022), a novel gender translation challenge set that was built on manually annotated test sets and is specifically designed to measure the translation accuracy of gendered expressions. FUDGE demonstrates improvements in several feminine gender terms' translation accuracy.[1]

2 Bias Statement

In this paper, we explore how to mitigate the mistranslation of feminine gender terms into masculine forms when translating from English to a language that openly marks the gender of the speaker. When an NMT system systematically assumes the gender of the speaker is male, this will cause representational harm, resulting in frequent translation errors for female speakers.

We borrow the systematic classification proposed by Dinan et al. (2020), which classifies gender bias into three dimensions: *Speaking About* (gender of the topic), *Speaking As* (gender of the speaker), and *Speaking To* (gender of the addressee). In this work, we focus on the *Speaking As* bias, which usually appears in first-person sentence translations.

Due to the limitations of annotated data sets, we can only experiment on sentences by male and female speakers. More in-depth research on reducing the representational bias towards non-binary speakers will be possible.

3 Related Work

Controlled Text Generation Some research focuses on fine-tuning a pre-trained model for a desired attribute. Ficler and Goldberg (2017) propose a framework for neural natural language generation (NNLG) controlling different stylistic aspects of the generated text. The method results in a class-conditional language model (CCLM), but it is difficult to separate the desired attribute from the generation model, i.e. the model is usually suitable for one task and needs retraining for another attribute of interest. Keskar et al. (2019) mitigate this issue by proposing a Conditional Transformer Language (CTRL) model that is conditioned on many factors including style, content, and task-specific behavior. However, this is quite expensive. Krause et al. (2021) suggest using discriminators to guide the decoding of LMs. Kumar et al. (2021) propose MUCOCO[2] where they formulate the decoding process as a continuous optimization problem that allows for multiple attributes.

Gender Debiasing A common method to mitigate gender bias is to attach gender tags as proposed by Vanmassenhove et al. (2018). In this case, gender information is integrated into the NMT systems via a tag on the source side. This approach achieves improvements for multiple language pairs. Given the original biased data set, Zhao et al. (2018) propose to construct an additional training corpus where all male entities are swapped for female entities and vice-versa. The goal of the augmentation is to mitigate the bias by training the model on gender-balanced data sets.

Gender Bias Evaluation Bemchmarks Zhao et al. (2018) introduce a benchmark, WinoBias, to measure gender bias in coreference resolution with entities corresponding to people referred to by their occupation. Another benchmark, WinoGender (Rudinger et al., 2018), is a Winograd schema-style (Levesque et al., 2012) set of minimal pair sentences that differ only by pronoun gender. Based on WinoBias and WinoGender, Stanovsky et al. (2019) compose a coreference resolution English corpus that contains sentences in which the subjects are in non-stereotypical gender roles. It is a standard test set to evaluate gender stereotyping in MT. In contrast, MuST-SHE (Savoldi et al., 2022) provides a fine-grained grammatical gender evaluation on word level and gender agreement level, which makes it more suitable to evaluate our model.

4 Method

In a controlled text generation task, it is usually nontrivial to retrain the model $\mathcal{G}$ to condition it on the new attribute a. Yang and Klein (2021) propose Future Discriminators for Generation (FUDGE), a flexible and modular way of conditioning the generative model $\mathcal{G}$ on the desired attribute a that only requires access to the output probabilities of $\mathcal{G}$. FUDGE achieves this by training a binary classifier that predicts at each time step t whether the attribute a will be satisfied in the complete sequence, based on the already generated tokens $y_0 - y_t$.

[1] Code and documentation for the experiments are available on https://github.com/tianshuailu/debias_FUDGE.

[2] The acronym for this algorithm stands for incorporating **mu**ltiple **co**nstraints through **c**ontinuous **o**ptimization.

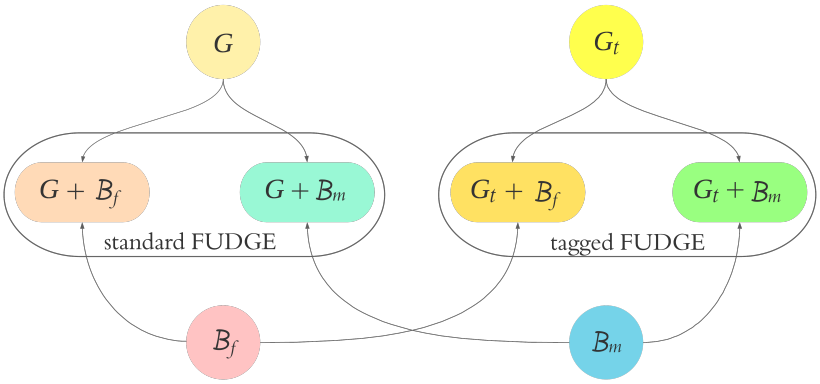

Figure 2: Illustration of four combinations between the underlying translation models $\mathcal{G}$ (translation model trained on original data sets), $\mathcal{G}_t$ (translation model trained on tagged data sets) and two classifiers $\mathcal{B}_f$ (feminine), $\mathcal{B}_m$ (masculine).

To see if gender tags improve FUDGE's performance, we have two underlying English–Italian translation models, $\mathcal{G}$ and $\mathcal{G}_t$. We train both models on the same sentence pairs, with the exception that $\mathcal{G}_t$'s data set includes gender tags on the English source side. The method of adding gender tags is inspired by Vanmassenhove et al. (2018). The desired attributes are feminine and masculine, hence we train two classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$. *Each* of them is combined with the two underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$, resulting in four combinations, as illustrated in Figure 2.

An advantage of FUDGE is the fact that it only needs access to the output logits of the generator model, meaning $\mathcal{G}$ and $\mathcal{G}_t$ can be directly combined with $\mathcal{B}_f$ and $\mathcal{B}_m$ without additional fine-tuning or modification. This allows us to directly use $\mathcal{G}$ and $\mathcal{G}_t$ as baselines.

5 Experimental Setup

5.1 Data

Europarl-Speaker-Information consists of Europarl (Koehn, 2005) tagged with speaker information, including the gender of the speaker. We chose Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) because it contains 44.5% first-person sentences, which makes it suitable for the kind of gender bias the experiments focus to reduce, i.e., *Speaking As*.

ParlaMint 2.1 is a multilingual set of 17 corpora containing parliamentary debates, including gender tags (Erjavec et al., 2021). In Italian, the adjectives and participles are marked with the gender of the speaker in certain grammatical contexts. In the full data set, the utterances where the gender of the speaker is marked are relatively sparse. Hence, we removed sentences that do not contain adjectives or participles for these experiments, since these cannot be marked for the gender of the speaker. The sizes of the original data set and the amount we used are shown in Table 1. In addition, to ensure balanced positive and negative class sizes, we used the same amount of utterances by female and male speakers to train the classifiers.

MuST-SHE v1.2 is a multilingual benchmark allowing for a fine-grained analysis of gender bias in Machine Translation and Speech Translation (Savoldi et al., 2022). MuST-SHE v1.2 contains 656 first-person sentences out of 1073, which makes it suitable for the evaluation of FUDGE.

Table 1 provides an overview of the three data sets along with the information on how they were used in our study. We used the English–Italian part of Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) to train and test the underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$. The monolingual Italian ParlaMint 2.1 corpus (Erjavec et al., 2021) was used to train and test the feminine and masculine classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$. Finally, the English–Italian parallel data from MuST-SHE v1.2 (Savoldi et al., 2022) was used to compare FUDGE and tagged FUDGE against the baselines.

5.2 Training

To get the underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$, we first trim the vocabulary of the pretrained mT5-small (Xue et al., 2021) from HuggingFace (Wolf et al., 2020)[3] to a smaller vocabulary of 25,000

[3] https://huggingface.co/google/mt5-small

		Europarl-Speaker-Information	ParlaMint 2.1	MuST-SHE v1.2
Type		en-it parallel	it monolingual	en-it parallel
#sentences	total	1.29M	996.5k	1095
	used	1.20M	91.6k	1073
M:F ratio	total	2:1	2.5:1	1:1
	used	2:1	1:1	1:1
Usage		train $\mathcal{G}$ and $\mathcal{G}_t$	train $\mathcal{B}_f$ and $\mathcal{B}_m$	evaluation

Table 1: An overview of the language type, gender ratio and the usage of the corpora. Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) English–Italian parallel data sets contain double the amount of utterances by male speakers compared to female speakers and were used to train and test the underlying translation models $\mathcal{G}$ and $\mathcal{G}_t$. ParlaMint 2.1 (Erjavec et al., 2021) Italian monolingual data sets were used to train and test the feminine and masculine classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$. MuST-SHE v1.2 (Savoldi et al., 2022) English–Italian parallel data sets were used to evaluate FUDGE and tagged FUDGE. Both ParlaMint and MuST-SHE data sets that were used for the experiment have an equal amount of utterances by female and male speakers.

	standard FUDGE		*tagged FUDGE*	
	$feminine$	$masculine$	$feminine$	$masculine$
$\lambda = 0$	**27.2**	**27**	**27.5**	**27.1**
$\lambda = 1$	27.1	**27.0**	27.3	26.9
$\lambda = 2$	27.0	26.8	27.2	26.9
$\lambda = 3$	26.9	26.7	27.0	26.7
$\lambda = 4$	26.5	26.6	26.6	26.5
$\lambda = 5$	26.2	26.4	26.2	26.5

Table 2: The BLEU scores of standard FUDGE and tagged FUDGE with both feminine and masculine classifiers, i.e. the four models illustrated in Figure 2. Each model was tested on λ ranging from 1 to 5. When $\lambda = 0$, the classifier does not contribute, hence the first row represents the BLEU scores of the baselines.

English and Italian subword entries.[4] We then fine-tune the trimmed mT5 on the English–Italian part of the Europarl-Speaker-Information (Vanmassenhove and Hardmeier, 2018) data set with adapted example scripts provided in the HuggingFace Transformers repository. $\mathcal{G}$ and $\mathcal{G}_t$ share model architecture and training hyperparameters.

For the two classifiers $\mathcal{B}_f$ and $\mathcal{B}_m$, we use the same amount of filtered sentences by female speakers and male speakers, i.e. 45,800 sentences each.[5] The architecture of the classifier is a 3-layer causal LSTM (Hochreiter and Schmidhuber, 1997) with a hidden dimension of 512. The FUDGE classifiers use the same vocabulary as the generation models (trimmed mT5-small). While it is not mandatory, we choose to initialize the embeddings in the classifier using the pre-trained mT5-small. Alternatively, embeddings can be initialized randomly or using another pre-training method. To train $\mathcal{B}_f$, sentences by female speakers are the positive class, whereas in training $\mathcal{B}_m$, sentences by male speakers are the positive class.

5.3 Evaluation

For evaluation, we use SacreBLEU (Post, 2018)[6] to calculate the BLEU scores. Furthermore, we use

[4]We tokenize 4.5 million English and Italian sentences with the mT5-small tokenizer and keep the 25k most frequent subwords as the trimmed vocabulary.

[5]Filtering based on part of speech (POS): We kept only sentences that contain adjectives and/or participles since those are the only POS that can be marked for the gender of the speaker.

[6]For reproducibility reasons, the version signature is "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1´´

the MuST-SHE challenge set (Savoldi et al., 2022) to assess the models' performance at two levels of granularity, i.e. word-level parts-of-speech (POS) gender evaluation and chain-level gender agreement evaluation. Both POS and agreement chain annotations are on the Italian side.

For word-level evaluation, MuST-SHE performs a fine-grained qualitative analysis of the system's accuracy in producing the target gender-marked words. MuST-SHE computes the accuracy as the proportion of gender-marked words in the references that are correctly translated by the system. An upper bound of one match for each gender-marked word is applied to prevent rewarding over-generated terms.

For agreement-level evaluation, MuST-SHE inspects the agreement chain coverage and translation accuracy. Each agreement chain is composed of several agreement terms. The agreement chain is in coverage only when all the terms appear in the translation (regardless of their gender forms). Then MuST-SHE further evaluates the accuracy of the in-coverage chains. Either the agreement is not respected (*No*), or it is respected with the correct gender (*Correct*) or wrong gender (*Wrong*).

6 Results

6.1 BLEU

Table 2 shows the BLEU scores of standard FUDGE and tagged FUDGE with both feminine and masculine classifiers, i.e. the four models illustrated in Figure 2. The hyperparameter λ determines how much weight is accorded to the classifier's predictions during inference. We test each model with λ ranging from 1 to 5. When $\lambda = 0$, the classifier does not contribute, hence the first row in Table 2 represents the BLEU scores of the baselines. The baselines have the highest BLEU scores for utterances by both female speakers and male speakers. With the increase of λ's value, the BLEU score either does not change or decreases.

6.2 MuST-SHE Gender Translation Evaluation

Word-level Gender Evaluation Table 3 displays the **word-level** feminine and masculine form open-class POS accuracy of standard FUDGE and tagged FUDGE with λ ranging from 1 to 5. The first rows ($\lambda = 0$) display the accuracy scores of the two baselines. *Adj-des* denotes descriptive adjectives. As shown in Table 3a, For both standard and tagged FUDGE, the accuracy of all three feminine form open-class words improves with the increase of λ, while both the baselines and FUDGE maintain high accuracy on masculine form open-class POS, as displayed in Table 3b.

Chain-level Gender Agreement Evaluation Table 4 shows the feminine and masculine **gender agreement** evaluation results of standard FUDGE and tagged FUDGE with λ ranging from 1 to 5, again, the first rows are the accuracy of the baselines. As shown in Table 4a, for the feminine agreement chains, the tagged baseline has more correct agreement chains and less percentage of no agreements than the standard baseline. With the increase of λ, standard FUDGE has more correct agreement chains than tagged FUDGE and a lower percentage of wrong or no agreements. Table 4b illustrates that both the baselines and FUDGE are quite accurate on the masculine agreement chains.

7 Discussion

7.1 BLEU

The first row in Table 2 demonstrates that the tagged baseline improves more on utterances by female speakers, indicating that the advantage of adding a gender tag to the English source side is more noticeable for sentences by female speakers. This result is somewhat expected since there are more utterances by male speakers in the training data, as shown in Table 1, i.e. the model is more likely to produce masculine forms by default.

On the other hand, the BLEU scores of both standard and tagged FUDGE decreases with the increase of λ. Since the classifiers were trained on a relatively small amount of data compared to the generation models, their fluency and grammaticality is not as good. Giving the classifiers more weight during generation while correcting for gender mistakes also makes the output less fluent compared to the mT5-small baselines.

Table 5 illustrates an example of overcorrection: This sentence is uttered by a female speaker, but the translation of the English word, *medium*, *mezzo*, is a noun, not an adjective. However, with high enough λ, FUDGE overcorrects this to a feminine adjective form, *media*.

7.2 MuST-SHE Gender Translation Evaluation

Word-level Gender Evaluation As displayed in Table 3a, for both standard FUDGE and tagged

	standardFUDGE			*taggedFUDGE*		
	Verbs	*Nouns*	*Adj-des*	*Verbs*	*Nouns*	*Adj-des*
baseline	27.4	11.4	35.4	27.3	13.5	36.3
$\lambda = 1$	43.7	12.8	42.9	39.5	13.2	45.7
$\lambda = 2$	60.6	13.2	61.2	56.3	**20.5**	55.1
$\lambda = 3$	62.1	10.8	55.1	63.6	14.3	61.7
$\lambda = 4$	70.1	11.8	61.2	**67.1**	15.4	64.6
$\lambda = 5$	**71.0**	**17.1**	**61.4**	62.9	19.0	**66.0**

(a) The **feminine** form open-class **POS** accuracy

	standardFUDGE			*taggedFUDGE*		
	Verbs	*Nouns*	*Adj-des*	*Verbs*	*Nouns*	*Adj-des*
baseline	87.8	**97.6**	94.3	94.4	**97.6**	94.1
$\lambda = 1$	91.4	96.3	**94.4**	94.5	97.5	92.2
$\lambda = 2$	92.9	97.5	94.2	95.8	97.5	91.7
$\lambda = 3$	94.1	97.4	94.1	93.1	97.5	92.2
$\lambda = 4$	**96.9**	97.5	94.1	**97.0**	97.3	**96.1**
$\lambda = 5$	96.6	97.5	92.0	95.5	97.5	91.8

(b) The **masculine** form open-class **POS** accuracy

Table 3: The feminine and masculine form open-class **POS** accuracy of standard FUDGE and tagged FUDGE with λ ranging from 1 to 5. The first row displays the accuracy scores from the baselines. *Adj-des* denotes descriptive adjectives.

FUDGE, the accuracy of all three open-class words improves, especially for verbs and descriptive adjectives. The classifier helps with the translation of gender-marked terms. As shown in Table 6, the gender of the speaker is female, meaning that the word *sure* needs to be translated into the feminine form *certa* or *sicura*. Both the standard baseline and the tagged baseline translate *sure* to the masculine form *sicuro*, while FUDGE corrects it to *sicura*.

The accuracy of nouns improves with FUDGE, however, the overall accuracy on nouns is much lower than on verbs and descriptive adjectives. A possible explanation is that participles and adjectives refer to the speaker more commonly, and are thus marked with the gender of the speaker, whereas cases where a speaker refers to themselves with a noun (e.g. *I'm a doctor*) are much less frequent in our data sets consisting of parliamentary sessions. Nouns in many cases refer to someone other than the speaker, and thus do not necessarily match the gender of the speaker.

Chain-level Gender Agreement Evaluation The first row of Table 4a shows the agreement chains percentage of the baselines. The tagged baseline performs slightly better than the standard baseline. With the increase of λ, FUDGE improves the percentage of correct agreement chains and reduces the percentage of wrong agreement chains. Furthermore, standard FUDGE performs better than tagged FUDGE.

8 Conclusion

We explore controlled text generation in the context of gender bias by utilizing Future Discriminators for Generations (FUDGE) (Yang and Klein, 2021) in combination with a pre-trained model, mT5-small. Our experiments show that baseline models generally work well on masculine forms since those are much more frequent in the training data Table 1. However, a targeted evaluation reveals that the baselines tend to mistranslate feminine forms. Controlled generation with FUDGE can correct this considerably. Moreover, we observe a trade-off between fluency and gender bias. This is attributed to the fact that our FUDGE classifiers are trained on a relatively small amount of data compared to the generation models. As a con-

	$standardFUDGE$			$taggedFUDGE$		
	$Correct \uparrow$	$Wrong \downarrow$	$No \downarrow$	$Correct \uparrow$	$Wrong \downarrow$	$No \downarrow$
$baseline$	45.5	36.4	18.2	48.6	37.1	**14.3**
$\lambda = 1$	52.8	33.3	13.9	45.7	34.3	20.0
$\lambda = 2$	57.9	28.9	**13.2**	52.6	31.6	15.8
$\lambda = 3$	52.8	27.8	19.4	**56.7**	**27.0**	16.2
$\lambda = 4$	57.1	20.0	22.9	51.3	32.4	16.2
$\lambda = 5$	**63.6**	**18.2**	18.2	44.7	34.2	21.1

(a) **Feminine** gender agreement chain accuracy

	$standardFUDGE$			$taggedFUDGE$		
	$Correct \uparrow$	$Wrong \downarrow$	$No \downarrow$	$Correct \uparrow$	$Wrong \downarrow$	$No \downarrow$
$baseline$	91.1	3.6	5.4	**96.2**	**0.0**	3.8
$\lambda = 1$	94.5	1.8	3.6	94.4	1.9	3.7
$\lambda = 2$	94.4	1.9	3.7	94.2	1.9	3.8
$\lambda = 3$	94.4	1.9	3.7	94.4	1.9	3.7
$\lambda = 4$	**96.5**	**0.0**	**3.5**	**96.2**	**0.0**	3.7
$\lambda = 5$	92.3	**0.0**	7.7	94.7	1.8	**3.5**

(b) **Masculine** gender agreement chain accuracy

Table 4: The feminine and masculine **gender agreement** evaluation results of standard FUDGE and tagged FUDGE with λ ranging from 1 to 5. The first row displays the accuracy scores from the baselines. *Correct* denotes the agreement is respected with the correct gender, *Wrong* denotes the agreement is respected but with the wrong gender, and *No* denotes the agreement is not respected. The numbers represent the percentage of each case.

EN	The internet is a **medium** ...
Ref	Internet è un **mezzo** ...
FUDGE	Internet è un **media** ...
BASE	Internet è un **medio** ...

Table 5: An overcorrection example of tagged FUDGE when $\lambda = 4$ on a sentence by a female speaker. The correct translation of the English word *medium* should be the masculine noun *mezzo*. The baseline uses a wrong masculine noun *medio*, which refers to "middle finger". FUDGE overcorrects *medio* with a feminine noun *media*, means "average value".

EN	I am **sure** you will agree ...
Ref	Sono **certa** che sarà d'accordo ...
FUDGE	Sono **sicura** che lei concorderà ...
BASE	Sono **sicuro** che lei concorderà ...

Table 6: A correct translation example of tagged FUDGE with $\lambda = 3$ on a sentence by a female speaker. FUDGE translates the English word *sure* into the correct feminine form, *sicura*, while the baseline generates the masculine form, *sicuro*.

sequence, assigning greater weight to their predictions leads to a reduction in fluency and a decrease in BLEU scores. Ideally, the classifiers should be trained on more data. If this is not an option, FUDGE needs to be carefully balanced to obtain improvements without harming the fluency of the translations.

Acknowledgements

This work was funded by the EU's Horizon 2020 Programme as part of the project EASIER under grant agreement number 101016982 and the Swiss National Science Foundation project no. 191934).

References

Dinan, Emily, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceed-*

ings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 314–331, Online, November. Association for Computational Linguistics.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Roberto Bartolini, Andrea Cimino, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. 2021. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI.

Ficler, Jessica and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark, September. Association for Computational Linguistics.

Hardmeier, Christian, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. 2021. How to write a bias statement: Recommendations for submissions to the workshop on gender bias in nlp.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11.

Jwalapuram, Prathyusha, Shafiq Joty, and Youlin Shen. 2020. Pronoun-targeted fine-tuning for NMT with hybrid losses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2267–2279, Online, November. Association for Computational Linguistics.

Keskar, Nitish Shirish, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15.

Krause, Ben, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Kumar, Sachin, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. In Beygelzimer, A., Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*.

Levesque, Hector J., Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Loáiciga, Sharid, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16, Copenhagen, Denmark, September. Association for Computational Linguistics.

Papineni, Kishore, 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, July. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Rudinger, Rachel, 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, June. Association for Computational Linguistics.

Savoldi, Beatrice, 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.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland, May. Association for Computational Linguistics.

Sennrich, Rico. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April. Association for Computational Linguistics.

Stanovsky, Gabriel, 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, July. Association for Computational Linguistics.

Vanmassenhove, E. and C. Hardmeier. 2018. Europarl datasets with demographic speaker information. In *EAMT 2018 - Proceedings of the 21st Annual Conference of the European Association for Machine Translation*.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Yang, Kevin and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online, June. Association for Computational Linguistics.

Zhao, Jieyu, 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, June. Association for Computational Linguistics.