How To Build Competitive Multi-gender Speech Translation Models For Controlling Speaker Gender Translation

Marco Gaido[1], Dennis Fucci[1,2], Matteo Negri[1] and Luisa Bentivogli[1]

[1] *Fondazione Bruno Kessler*
[2] *University of Trento*

Abstract

When translating from notional gender languages (e.g., English) into grammatical gender languages (e.g., Italian), the generated translation requires explicit gender assignments for various words, including those referring to the speaker. When the source sentence does not convey the speaker's gender, speech translation (ST) models either rely on the possibly-misleading vocal traits of the speaker or default to the masculine gender, the most frequent in existing training corpora. To avoid such biased and not inclusive behaviors, the gender assignment of speaker-related expressions should be guided by externally-provided metadata about the speaker's gender.[1] While previous work has shown that the most effective solution is represented by separate, dedicated gender-specific models, the goal of this paper is to achieve the same results by integrating the speaker's gender metadata into a single "multi-gender" neural ST model, easier to maintain. Our experiments demonstrate that a single multi-gender model outperforms gender-specialized ones when trained from scratch (with gender accuracy gains up to 12.9 for feminine forms), while fine-tuning from existing ST models does not lead to competitive results.

Keywords

gender bias, gradient reversal, speech translation

1. Introduction

Spurred by growing concerns about fairness in language technologies, research on understanding and mitigating gender bias in automatic translation is gaining traction [1]. The bias of automatic systems is extremely evident when it comes to ambiguous sentences or expressions, where there are no explicit cues in the source content about the correct gender[1] assignment of a referent (e.g., en: *The doctor arrived* – it: *Il/La dottore/essa è arrivato/a*). In this setting, the state-of-the-art neural models often choose the masculine forms or perpetuate stereotypical assignments, as they reflect the condition statistically more likely based on their (biased) training data [2, 3].

This situation frequently occurs when the source language is genderless or employs notional gender, expressing gender in a limited set of parts of speech, and the target language follows a grammatical gender system, embedding gender distinctions throughout a broad inventory of parts of speech. Focusing on the case in which the source language is English, a notional gender language, and the target language is Italian, a grammatical gender language, a frequent instance of this condition is represented by first-person references, i.e. by the words and expressions referred to the speaker (e.g., en: *I am a young researcher* – it: *Sono una/un giovane ricercatrice/tore*). In this case, text-to-text machine translation (MT) models mostly output masculine forms, while direct (or end-to-end) speech-to-text translation (ST) systems partly rely on the biological cue of the speaker's vocal traits to assign gender [4, 5]. However, direct ST models are still largely biased toward producing masculine forms, and, most importantly, biological aspects are related to the sex rather than to the gender of an individual. Hence, their exploitation is not inclusive of all people, harming several groups such as transgenders [6].

As a solution, [7] proposed to leverage external metadata about the speaker's gender to control the gender assignment of words referred to the speaker. Specifically, they investigated two approaches: *i)* the development of two separate *gender-specialized models*, fine-tuned on gender-specific data as also proposed later in MT [8], and *ii)* a single *multi-gender model*, where the speaker gender is a tag fed to a single model as in multilingual systems [9]. While the second solution would be preferable (as the specialized solution involves the higher cost of maintaining two separate models), the experiments in [7] demonstrate that specialized models outperform the multi-gender approach by a large margin in terms of gender accuracy.

In light of the above, in this paper we address the following research questions: *i)* why do specialized models outperform multi-gender ones? *ii)* Can we build competitive multi-gender systems? Through experiments on English-Italian translation of TED talks, we show that the low accuracy of multi-gender models comes from the initialization with the weights of a gender-unaware ST

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

mgaido@fbk.eu (M. Gaido); dfucci@fbk.eu (D. Fucci); negri@fbk.eu (M. Negri); bentivo@fbk.eu (L. Bentivogli)

[1] Throughout the paper, we use the word *gender* to indicate the preferred linguistic expression of gender and not the gender identity.

system and the inability to override the behavior of the base ST model (i.e., the reliance on vocal cues) during the fine-tuning stage. We also try to address this problem with two solutions: *i)* a contrastive loss that penalizes the extraction of gender cues from speech input, and *ii)* altering vocal properties of training data to misalign gender cues with gender tags and gender translations.[2] Despite the slight improvements brought by these solutions in gender accuracy and overall translation quality, none of them effectively close the performance gap with the specialized solution. However, training multi-gender models from scratch yields competitive results, outperforming the specialized approach with gender accuracy gains of up to 12.9 points for feminine translations. Therefore, we recommend building multi-gender models from scratch, while building them on top of existing systems remains an open research question.

2. Background

In this section, we introduce the basic concepts useful for understanding the rest of the paper. First, we provide an overview of the methods proposed in the literature to integrate language tags into neural multilingual translation models (§2.1), from which multi-gender models draw inspiration. Then, we present how gender information has been removed from neural representations through adversarial training in previous works (§2.2), from which we derive our solution presented in §3.1.

2.1. Tags Integration in Multilingual Models

State-of-the-art models in MT and ST are sequence-to-sequence models made of an encoder and an autoregressive Transformer decoder [10]. The autoregressive decoder predicts the next-token probability over a predefined vocabulary at every iteration by looking at the encoder output and at the previously generated tokens, which are pre-pended a special token named *beginning of sentence* (<*bos*>). Formally, the probability $p_V(y_t)$ over the vocabulary V at time step t is:

$$\text{softmax}(D(E(X); \text{<}\textit{bos}\text{>}, y_0, ..., y_{t-1})) \quad (1)$$

where *E* is the encoder, *D* is the decoder, *X* is the input sequence, and y_i is the token generated at the i-th time step.

While early attempts to build multilingual MT models were based on training dedicated encoders and decoders for each language [11, 12], nowadays the preferred solution is a model made of a single universal encoder and decoder where the language is represented as a tag pre-pended to the text [13, 14, 9]. In the case of one-to-many multilingual models, this means that the <*bos*> token is replaced with a token that indicates the language, so that Eq. 1 becomes:

$$\text{softmax}(D(E(X); \text{LID}, y_0, ..., y_{t-1})) \quad (2)$$

where *LID* is the identifier of the desired target language.

In direct ST, [15] demonstrated the effectiveness of this solution, also known as "target forcing", while [16] proposed other methods to integrate the language information into the architecture. Thanks to its simplicity and effectiveness, target forcing is currently the most widespread method to build multilingual ST systems [17, 18], also when using large pre-trained textual models such as mBART [19] to initialize the ST decoder [20, 21]. In line with this trend, [7] obtained their best multi-gender models with target forcing. As such, we build multi-gender models using target forcing with the *F* and *M* tags representing the two grammatical genders instead of the language identifiers.

2.2. Gender Information Removal

With the goal of fairer technology that does not rely on spurious cues reflecting stereotypical biases in the available data, researchers have tried to build systems that achieve "equalized odds" among different demographic groups [22]. Formally, this means that, given an attribute z representing the belonging to one of the Z demographic groups, the predicted probability $p(\hat{Y})$ of a fair system should be independent of the variable z, i.e. $p(\hat{Y}) = p(\hat{Y}|z), \forall z \in Z$. The variable z is named the *protected attribute* and in the context of gender bias literature represents the gender of the involved person. So in this work we consider $Z = \{F, M\}$.[3]

The first attempts to achieve equalized odds across genders in neural systems have focused on deep neural network (DNN) classifiers [23, 24, 25, 26]. In this line of work, the last hidden representation of the DNN is passed both to a linear layer that predicts the classification scores $\hat{Y}$ and to a linear layer (the *discriminator*) devoted to predicting the protected attribute z. The DNN is then trained in an adversarial manner [27], i.e. it is alternatively trained to *i)* predict z (while keeping the shared DNN freezed) and *ii)* predict $\hat{Y}$ while minimizing the ability to predict z (keeping the protected attribute classification layer freezer). As this training procedure is often unstable, similar practices based on minmax optimizations have been proposed [28], even with discriminators made of functions different from linear projections

[2] Our code is released open source under Apache 2.0 Licence at: https://github.com/hlt-mt/FBK-fairseq/

[3] Although this paper does not aim at perpetuating a binary vision of gender, in this work we limit to the feminine and masculine categories for the sake of simplicity, as the available benchmarks currently do not cover the non-binary case.

[29], or using more than one discriminator [30]. [31] also proposed methods to automatically extract the protected attributes in case they have not been provided.

Such adversarial training procedures can be seen as an extension of the *gradient reversal* [32], where the training alternatively freezes the base model (to refine the discriminator) and the discriminator (inverting its loss to train the base model to be unable to discriminate). In fact, the gradient reversal layer, applied to the hidden representations before feeding them to the discriminator, is an identity function in the forward pass, while inverts the gradient in the backward pass, scaling it by a positive factor λ. By naming x the hidden representation, D the discriminator, and GRL the gradient reversal layer, this means that:

$$GRL(x) = x, \nabla D(GRL(x)) = -\lambda \nabla D(x) \tag{3}$$

where λ is a hyperparameter that can either be fixed or updated according to the following equation:

$$\lambda = \frac{2}{1 + e^{-\gamma p}} - 1 \tag{4}$$

where p is the ratio between the number of parameter updates performed and the total number of updates needed to complete the training.

3. Solutions for Multi-gender Models

To create multi-gender ST models that solely rely on the gender tag, ignoring spurious cues related to speakers' vocal traits, we test two approaches. First, we try to create gender-invariant encoder representations by adding a gradient-reverted discriminator on the speaker's gender (§3.1). Second, we manipulate the input audio by altering the speaker's pitch, so that the correlation between the gender tag (and output text) and the speaker's vocal traits is lost (§3.2).

3.1. Gradient Reversal

As seen in §2.1, the decoder of a multi-gender model has three inputs: the encoder output, a tag representing the speaker's gender, and the previously generated tokens. As we want the decoder to have the tag as the only source of information about the speaker's gender, we propose to create encoder outputs that do not convey any information regarding the speaker's gender by adding a gradient-reverted discriminator on top of the encoder, motivated by the success of this approach in MT with sentences where there is a single referent whose gender has to be determined [33]. The discriminator is made of two fully-connected layers with ReLU activation function [34], whose output is averaged over the temporal dimension to obtain a single vector representing the logit[4] of the discriminator.

Furthermore, we experiment with assigning dedicated class weights to the loss of the discriminator, as a countermeasure to the class imbalance between female and male speakers in the training data. Specifically, we assigned the weights (w_f, w_m) proportionally to the inverse of the frequency of each class (f_f, f_m) in the training data:

$$\begin{cases} w_f \propto \frac{1}{f_f}, w_m \propto \frac{1}{f_m} \\ w_f * \frac{f_f}{f_m + f_f} + w_m * \frac{f_m}{f_m + f_f} = 1 \end{cases}$$

resulting in $w_f = 1.4$, $w_m = 0.8$ in our case.

3.2. Audio Manipulation

Our second approach aims to break the correlation between the vocal characteristics of the speaker on one side and the gender tag and target translation on the other. To this aim we manipulate part of the training data using the *Opposite* pitch manipulation strategy by [35]. The amount of data that is manipulated at each iteration (epoch) is controlled with a hyperparameter, p, which determines the probability of altering an utterance, regardless of whether it is produced by a male or female speaker. The manipulation is performed by altering two crucial acoustic parameters distinguishing between male and female voices [36, 37]: $f0$ and formants. In particular, we first estimate the $\tilde{f0}$ median of the $f0$ contour of the considered speech segment. Then, we sample a new $\tilde{f0}'$ median of the desired output audio from a normal distribution whose mean and standard deviation depend on the target gender: for feminine voices, we use 250 Hz as the mean and 17 as the standard deviation so that the sampled value is between 199 Hz and 301 Hz with 99.7% probability; for masculine voices, the mean is 140 Hz and the standard deviation is 20 to obtain a 99.7% probability range within 80 Hz and 200 Hz. Once $\tilde{f0}$ and $\tilde{f0}'$ are defined, we compute a scaling factor α as the ratio $\tilde{f0}'/\tilde{f0}$. Lastly, the original $f0$ contour is scaled by the α factor, while the formants are scaled by 1.2 when converting from male to female voices, or by 0.8 otherwise. This perturbation is applied independently to each sample during each training epoch, so as to maximize the variability of the training data.

4. Experimental Settings

Our ST models are composed of a Conformer [38] encoder with 12 layers and a Transformer [10] decoder with 6 layers. We used the Conformer implementation by [39],

[4] The *logit* is the vector of raw predictions before a function (commonly, the softmax) that maps it into probabilities.

Model	BLEU (↑)	Gender Accuracy (↑)			
		1F	1M	1F-Tag M	1M-Tag F
Fine-tuning					
Specialized	**27.4**	73.3	92.5	80.9	56.1
Multi-gender	26.0	66.8	78.0	64.6	47.1
+ gradient reversal	26.7	60.7	85.9	77.0	43.3
+ gradient reversal weighted	26.3	62.4	83.5	77.7	45.9
+ audio manipulation (50%)	26.4	56.0	82.6	60.7	33.9
+ audio manipulation (80%)	26.3	69.3	81.1	69.0	47.7
Training from scratch					
Multi-gender	27.2	**84.0**	92.7	93.4	**69.0**
+ gradient reversal	24.9	70.9	**93.2**	**94.1**	58.7
+ gradient reversal weighted	24.2	75.8	92.6	92.8	63.5
+ audio manipulation (50%)	26.2	79.6	92.4	91.5	67.9
+ audio manipulation (80%)	25.7	81.7	92.6	91.0	65.3

Table 1
BLEU and gender accuracy scores for the specialized models (Specialized) and the multi-gender models (Multi-gender) both trained from scratch and fine-tuned, also with gradient reversal and audio manipulation.

which does not contain bugs related to the presence of padding. The embedding size was 512, and the dropout was set to 0.1. We optimized label-smoothed cross entropy using Adam. The learning rate followed the Noam scheduler with 25,000 warmup updates and a maximum value of $2e^{-3}$. We train for 50,000 updates and average the last 7 checkpoints.

We train our models on MuST-C [40], an ST corpus built from TED data, for which is also available the annotation of the gender of the speaker [7]. We extract 80 features with log mel-filterbank from the input audio and normalize them with cepstral mean and variance [41]. The target text is encoded into subwords with 8,000 BPE merge rules [42] learned on the training set. We evaluate on the MuST-SHE benchmark [4], which contains a section ("Category 1") dedicated to assessing the gender assignment of words referring to the speaker. We compute SacreBLEU[5] [43] on the whole MuST-SHE test set to evaluate the translation quality of our models and gender accuracy [7] on the feminine and masculine sections of "Category 1" to evaluate the ability of each model to correctly assign gender to words referring to the speaker.

Gradient Reversal. The loss of the auxiliary speaker-classification task is summed to the loss on the decoder output scaling it by a 0.5 factor. For the gradient reversal layer, we tested both fixed values of λ and controlling its value with γ. For fine-tunings, we set $\lambda = 10$, so as to give similar weight to the gender classification loss and the cross-entropy loss for the translation. When training from scratch, instead, despite different attempts the training is unstable and diverges unless lambda is set to a fixed, small value, where its contribution is negligible. We report results for $\lambda = 0.5$, which is the highest λ value for which the loss on the validation set does not explode during training.

Audio Manipulation. In our experiments, we tested two values (0.5 and 0.8) for the hyperparameter p, which controls the probability of manipulating a speech segment. In the first case, 50% of the data is manipulated, leading to a complete loss of correlation between gender tags and vocal traits (50% of the samples with the F tag would exhibit frequency characteristics typical of masculine voices, and 50% of the samples with the M tag would have frequency characteristics typical of feminine voices). In the second case, instead, the correlation between the gender tag and the vocal traits is negative, to counteract the patterns learned by a gender-unaware ST model. In any case, as the training data is imbalanced (70% of the samples are uttered by male speakers, and 30% by female speakers), and the manipulation probability is the same for segments uttered by male and female speakers, the gender imbalance in the training data is not mitigated.

5. Results

We investigate the performance of multi-gender models trained in two different ways: *i)* fine-tuning a base, gender-unaware ST model, and *ii)* training from scratch. In both cases, we study the effect of the introduction of the discriminator with gradient reversal and of the audio manipulation techniques. Table 1 presents BLEU and gender accuracy scores (separately for segments spoken by female (1F) and male (1M) speakers) for all the models, comparing them with the specialized models. To assess the inclusivity of our solution in cases where speakers exhibit vocal traits that do not conform with traditional gender perceptions, we also report gender accuracy for

[5] case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

tests where the gender tag is inverted compared to the original audio segment (1F-Tag M and 1M-Tag F). In these instances, the gender translation is expected to align with the gender tag, and we use the “wrong” reference of MuST-SHE, which swaps the speaker’s references to the opposite gender.

Fine-tuning. The fine-tuned models from a base ST system consistently yield lower scores compared to the specialized systems. The simple multi-gender model performs 1.4 BLEU points worse than the specialized models in terms of overall translation quality. However, when audio manipulation and especially gradient reversal techniques are employed during fine-tuning, the performance gap is reduced by up to half. Regarding gender accuracy, the multi-gender model achieves considerably lower scores than the specialized models, confirming previous findings from [7]. This indicates that a fine-tuned multi-gender model struggles to accurately follow the gender tag for gender translation. The accuracy gap is particularly high when the tag conflicts with vocal traits (-16.3 in 1F-Tag M, -9.0 in 1M-Tag F), where multi-gender models show below-chance accuracy for feminine forms, being below 50%. Both gradient reversal and audio manipulation techniques seem to further bias the model towards masculine forms. This likely indicates that the reduced ability to rely on the speakers’ vocal traits is not compensated by the model looking at the gender tag, rather it strengthens its tendency to default to the most frequent masculine forms. The only technique that consistently improves both masculine and feminine translations compared to the simple fine-tuned multi-gender model is the introduction of audio manipulation with high probability (80%). However, the gains (2.5 for 1F and 3.1 for 1M, and 4.4 for 1F-Tag M and 0.7 for 1M-Tag F) are limited and the gap with specialized models remains large.

Training from scratch. Unlike the fine-tuned models, the multi-gender models trained from scratch yield comparable or even higher results than the specialized models. This suggests that when an ST model is trained from scratch with gender tags, it learns to effectively follow them. Specifically, the simple multi-gender model trained from scratch achieves comparable translation quality to the specialized system (-0.2 BLEU) and significantly outperforms it in gender accuracy, with gains ranging from 0.2 (1M) to 12.5 (1F-Tag M). As in the fine-tuning case, neither gradient reversal nor audio manipulations increase the reliance of the model on the tag and the resulting models are more biased toward masculine forms. In fact, the multi-gender with gradient reversal reaches the highest accuracies in producing masculine forms (93.2 in 1M and 94.1 in 1F-Tag M), while suffering substantial drops in feminine accuracy. This effect is reduced when the weight of the F class is increased in the discriminator. In addition, in both cases the translation quality suffers a considerable drop. In the case of audio manipulation, the translation quality drop is lower (although still present), as well as the differences in terms of gender accuracy. They do not provide, though, any benefit compared to the simple multi-gender training.

In summary, the results demonstrate that the previous finding about the low performance of multi-gender models is due to the adoption of a fine-tuning strategy. In this setting, the model cannot effectively override the reliance on speakers’ vocal traits of the gender-unaware base ST model. In addition, techniques aimed at avoiding the exploitation of speakers’ vocal traits seem ineffective. However, training the multi-gender model effectively solves the problem and the model is capable of following the indication given by the gender tag, outperforming even the specialized strategy by up to 12.9 gender accuracy (1M-Tag F).

6. Conclusions

In this paper, we studied the effect of different training strategies to build multi-gender ST models, i.e. models that are informed of the gender of the speaker by an explicit gender tag. Focusing on English-Italian translations, we demonstrated that the low accuracy of multi-gender models shown by previous work stems from the their initialization with gender-unaware ST system weights and the inability of effectively overriding the reliance on vocal cues during fine-tuning. On the other hand, training multi-gender models from scratch proved to be an effective solution, outperforming the approach based on the creation of two gender-specialized models. As training from scratch is not always feasible, we also experimented with two methods to enhance the reliance on the gender tag in fine-tuned multi-gender models: penalizing the extraction of gender cues from speech input, and altering the vocal properties of the speakers in the training data to avoid the alignment between biological cues and gender tags and translations. While these solutions partially improved gender accuracy and overall translation quality in fine-tuned multi-gender models, they did not close the gap with specialized models. Therefore, further research is needed in this direction.

Acknowledgments

This work is part of the project “Bias Mitigation and Gender Neutralization Techniques for Automatic Translation”, which is financially supported by an Amazon Research Award AWS AI grant. We acknowledge the support of the PNRR project FAIR - Future AI Research

(PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, M. Turchi, Gender bias in machine translation, Transactions of the Association for Computational Linguistics 9 (2021) 845–874. URL: https://aclanthology.org/2021.tacl-1.51. doi:10.1162/tacl_a_00401.

[2] M. O. R. Prates, P. H. C. Avelar, L. C. Lamb, Assessing gender bias in machine translation: a case study with google translate, Neural Computing and Applications 32 (2020) 6363–6381. doi:10.1007/s00521-019-04144-6.

[3] W. I. Cho, J. W. Kim, S. M. Kim, N. S. Kim, On Measuring Gender bias in Translation of Gender-neutral Pronouns, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 173–181. URL: https://www.aclweb.org/anthology/W19-3824. doi:10.18653/v1/W19-3824.

[4] L. Bentivogli, B. Savoldi, M. Negri, M. A. Di Gangi, R. Cattoni, M. Turchi, Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6923–6933. URL: https://www.aclweb.org/anthology/2020.acl-main.619.

[5] M. Gaido, M. A. Di Gangi, M. Negri, M. Turchi, On Knowledge Distillation for Direct Speech Translation, in: Proceedings of CLiC-IT 2020, Online, 2021. URL: http://ceur-ws.org/Vol-2769/paper_28.pdf.

[6] L. Zimman, Transgender language, transgender moment: Toward a trans linguistics, in: K. Hall, R. Barrett (Eds.), The Oxford Handbook of Language and Sexuality, 2020. doi:10.1093/oxfordhb/9780190212926.013.45.

[7] M. Gaido, B. Savoldi, L. Bentivogli, M. Negri, M. Turchi, Breeding Gender-aware Direct Speech Translation Systems, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3951–3964. URL: https://www.aclweb.org/anthology/2020.coling-main.350. doi:10.18653/v1/2020.coling-main.350.

[8] P. K. Choubey, A. Currey, P. Mathur, G. Dinu, GFST: Gender-filtered self-training for more accurate gender in translation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1640–1654. URL: https://aclanthology.org/2021.emnlp-main.123. doi:10.18653/v1/2021.emnlp-main.123.

[9] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, Google's multilingual neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351. URL: https://aclanthology.org/Q17-1024. doi:10.1162/tacl_a_00065.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, USA, 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[11] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 866–875. URL: https://aclanthology.org/N16-1101. doi:10.18653/v1/N16-1101.

[12] O. Firat, B. Sankaran, Y. Al-onaizan, F. T. Yarman Vural, K. Cho, Zero-resource translation with multi-lingual neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 268–277. URL: https://aclanthology.org/D16-1026. doi:10.18653/v1/D16-1026.

[13] R. Sennrich, B. Haddow, A. Birch, Improving Neural Machine Translation Models with Monolingual Data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009. doi:10.18653/v1/P16-1009.

[14] T.-L. Ha, J. Niehues, A. Waibel, Toward multilingual neural machine translation with universal encoder and decoder, in: Proceedings of the 13th International Conference on Spoken Language Translation, International Workshop on Spoken Language Translation, Seattle, Washington D.C, 2016. URL: https://aclanthology.org/2016.iwslt-1.6.

[15] H. Inaguma, K. Duh, T. Kawahara, S. Watanabe, Multilingual end-to-end speech translation, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 570–577. doi:10.1109/ASRU46091.2019.9003832.

[16] M. A. Di Gangi, M. Negri, M. Turchi, One-to-many multilingual end-to-end speech translation, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 585–592. doi:10.1109/ASRU46091.2019.9004003.

[17] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, J. Pino, Fairseq S2T: Fast speech-to-text modeling with fairseq, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Suzhou, China, 2020, pp. 33–39. URL: https://aclanthology.org/2020.aacl-demo.6.

[18] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, M. Post, The Multilingual TEDx Corpus for Speech Recognition and Translation, in: Proc. Interspeech 2021, 2021, pp. 3655–3659. doi:10.21437/Interspeech.2021-11.

[19] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: https://aclanthology.org/2020.tacl-1.47. doi:10.1162/tacl_a_00343.

[20] D. Liu, T. Binh Nguyen, S. Koneru, E. Yavuz Ugan, N.-Q. Pham, T. Nam Nguyen, T. Anh Dinh, C. Mullov, A. Waibel, J. Niehues, KIT's multilingual speech translation system for IWSLT 2023, in: Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), Association for Computational Linguistics, Toronto, Canada (in-person and online), 2023, pp. 113–122. URL: https://aclanthology.org/2023.iwslt-1.6.

[21] E. Gow-Smith, A. Berard, M. Zanon Boito, I. Calapodescu, NAVER LABS Europe's multilingual speech translation systems for the IWSLT 2023 low-resource track, in: Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), Association for Computational Linguistics, Toronto, Canada (in-person and online), 2023, pp. 144–158. URL: https://aclanthology.org/2023.iwslt-1.10.

[22] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 3323–3331.

[23] A. Beutel, E. H. Chi, J. Chen, Z. Zhao, Data decisions and theoretical implications when adversarially learning fair representations, 2017. URL: https://arxiv.org/pdf/1707.00075.pdf.

[24] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 335–340. URL: https://doi.org/10.1145/3278721.3278779. doi:10.1145/3278721.3278779.

[25] Y. Elazar, Y. Goldberg, Adversarial removal of demographic attributes from text data, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 11–21. URL: https://aclanthology.org/D18-1002. doi:10.18653/v1/D18-1002.

[26] L. Gao, H. Zhan, A. Chen, V. Sheng, Mitigate gender bias using negative multi-task learning, 2022. URL: https://doi.org/10.21203/rs.3.rs-2024101/v1. doi:10.21203/rs.3.rs-2024101/v1.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (2020) 139–144. URL: https://doi.org/10.1145/3422622. doi:10.1145/3422622.

[28] S. Ravfogel, M. Twiton, Y. Goldberg, R. D. Cotterell, Linear adversarial concept erasure, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 18400–18421. URL: https://proceedings.mlr.press/v162/ravfogel22a.html.

[29] S. Ravfogel, F. Vargas, Y. Goldberg, R. Cotterell, Adversarial concept erasure in kernel space, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6034–6055. URL: https://aclanthology.org/2022.emnlp-main.405.

[30] X. Han, T. Baldwin, T. Cohn, Diverse adversaries for mitigating bias in training, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2760–2765. URL: https://aclanthology.org/2021.eacl-main.239. doi:10.18653/v1/2021.eacl-main.239.

[31] S. Shao, Y. Ziser, S. B. Cohen, Erasure of unaligned attributes from neural representations, Transactions of the Association for Computational Linguistics 11 (2023) 488–510. URL: https://aclanthology.

org/2023.tacl-1.29. doi:10.1162/tacl_a_00558.

[32] Y. Ganin, V. Lempitsky, Unsupervised Domain Adaptation by Backpropagation, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 1180–1189. URL: https://proceedings.mlr.press/v37/ganin15.html.

[33] E. Fleisig, C. Fellbaum, Mitigating Gender Bias in Machine Translation through Adversarial Learning, 2022. arXiv:2203.10675.

[34] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Omnipress, Madison, WI, USA, 2010, p. 807–814.

[35] D. Fucci, M. Gaido, M. Negri, M. Cettolo, L. Bentivogli, No Pitch Left Behind: Addressing Gender Unbalance in Automatic Speech Recognition through Pitch Manipulation, in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 2023.

[36] R. O. Coleman, A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice, Journal of Speech & Hearing Research 19(1) (1976) 168–180. doi:https://doi.org/10.1044/jshr.1901.168.

[37] J. M. Hillenbrand, M. J. Clark, The role of f0 and formant frequencies in distinguishing the voices of men and women, Attention Perception & Psychophysics 71(5) (2009) 1150–1166. doi:10.3758/APP.71.5.1150.

[38] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in: Proceedings of the 21st Annual Conference of the International Speech Communication Association, International Speech Communication Association, Shanghai, China (Online), 2020, pp. 5036–5040. doi:10.21437/Interspeech.2020-3015.

[39] S. Papi, M. Gaido, A. Pilzer, M. Negri, When Good and Reproducible Results are a Giant with Feet of Clay: The Importance of Software Quality in NLP, 2023. arXiv:2303.16166.

[40] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, M. Turchi, Must-c: A multilingual corpus for end-to-end speech translation, Computer Speech & Language 66 (2021) 101155. URL: https://www.sciencedirect.com/science/article/pii/S0885230820300887. doi:https://doi.org/10.1016/j.csl.2020.101155.

[41] O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Communication 25 (1998) 133–147. URL: https://www.sciencedirect.com/science/article/pii/S0167639398000338. doi:https://doi.org/10.1016/S0167-6393(98)00033-8.

[42] M. A. Di Gangi, M. Gaido, M. Negri, M. Turchi, On target segmentation for direct speech translation, in: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, Virtual, 2020, pp. 137–150. URL: https://aclanthology.org/2020.amta-research.13.

[43] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.