Fine-tuning Neural Machine Translation on Gender-Balanced Datasets

Marta R. Costa-jussà[*] **and Adrià de Jorge**[*]
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu,adria.de.jorge@estudiantat.upc.edu

Abstract

Misrepresentation of certain communities in datasets is causing big disruptions in artificial intelligence applications. In this paper, we propose using an automatically extracted gender-balanced dataset parallel corpus from Wikipedia. This balanced set is used to perform fine-tuning techniques from a bigger model trained on unbalanced datasets to mitigate gender biases in neural machine translation.

1 Introduction

Misrepresentation of individual communities in current datasets is causing severe disruptions in artificial intelligence applications. Examples of this are a lower performance of speech recognizers for women than for men (Tatman, 2017), a lower accuracy in face recognition for Asian faces than American or European ones (Xiong et al., 2018) and an amplification of stereotypes in Neural Machine Translation (NMT) (Font and Costa-jussà, 2019). These challenges are at the core of natural language processing applications, and, in particular, many works are focusing on trying to solve gender biases (Costa-jussà, 2019). With this objective in mind, and in the specific context of NMT, we propose the use of balanced data sets to mitigate gender biases in a standard NMT system taking advantage of domain adaptation techniques.

Previous research in the area of NMT has proposed to either mitigate biases using debiased word embeddings (Font and Costa-jussà, 2019) and using contextual information (Basta et al., 2020) or evaluating and measuring the amount of bias present in the translation (Stanovsky et al., 2019). The closest work to ours is the one by Saunders and Byrne (2020) where authors generate a small gender-balanced dataset and use Elastic Weight Consolidation techniques to perform transfer learning and mitigate the consequences of training with unbalanced datasets. Differently from this one, we use a larger non-synthetic balanced dataset to perform fine-tuning on an unbalanced-dataset and evaluate the reduction of gender bias in the final translation.

2 Bias statement

As proposed in previous work (Blodgett et al., 2020), we formulate the bias statement of our work. Our work consists of studying the effects of using a gender-balanced dataset to mitigate gender biases in NMT. In the NMT context, we can define gender bias as incorrectly translating a gendered source word into a target word opposite gender, when no ambiguity exists. We can attribute this to datasets that are over-represented with a particular gender. As shown in previous work (Bolukbasi et al., 2016b), there are representational harms in word embeddings such as demeaning women's ability to work in tech, e.g., *man is to computer programmer as woman is to homemaker*. The main concern is that training a system on unbalanced data will perpetuate these biases: first, on the methods built on top of these datasets, and second, people that use these systems will learn incorrect associations between words, unknowingly perpetuating these social biases. A system trained on balanced data is a first step in eliminating this representational harm, as there is the same number of instances between genders.

[*] Equal contribution

Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pages 26–34
Barcelona, Spain (Online), December 13, 2020.

To avoid stereotypical bias in professions, the next step would be to have the same professional distribution between genders, which can be achieved by gender-swapping the initial dataset. That way, the model will equally represent genders. Beyond this, we point out the limitation of doing a binary representation of gender, not reflecting the LGTBQ+ community. Note that our work trains a word embedding and NMT model and does not aim to reflect reality. In the end, mitigating gender bias in artificial intelligence systems is a short-term solution that needs to be combined with higher-level long-term projects in challenging current social power, among other principles (D'Ignazio and Klein, 2018).

3 Gender Balanced Dataset

This section explains the procedure followed to obtain an English-Spanish gender-balanced dataset, which uses the available Gebiotoolkit (Costa-jussà et al., 2019), and it extracts (multi-)parallel corpus at the sentence level from the Wikipedia Biographies. The toolkit consists of 3 blocks: a corpus extractor, which provides a layer to transform, collect and select entries in the desired languages; (2) a corpus aligner, which finds the parallel sentences within a text and provides a quality check of the parallel sentences given a few restrictions; (3) a gender classifier which includes a filtering module that classifies the gender of the entry and outputs the final parallel corpus. Hereinafter, we refer to this dataset as Balanced. We quantify the amount of gender bias in the collected dataset due to gender bias in word embeddings. This quantification of bias is also compared to the case of word embeddings computed on the EuroParl corpus (Koehn, 2005).

3.1 Balanced Dataset Generation

We used the available Gebiotoolkit (Costa-jussà et al., 2019) to extract the Balanced dataset. Gebiotoolkit is a tool for extracting multilingual parallel corpora at the sentence level, together with document and gender information from Wikipedia biographies. In this sense, the collected data set is not synthetic. We can generate this dataset from any of the languages available on Wikipedia. In our case, we have selected the English-Spanish language pair, which have considerable differences at the morphological level, and exhibit gender bias issues in NMT (Font and Costa-jussà, 2019).

After extraction, the biographies dataset has approximately 27,000 female-related sentences and 47,000 male-related sentences. To have an equal probability of finding a male or female related sentence, we balanced the dataset by removing male-related samples until having the same amount of masculine and feminine instances. In total, we end up with 54,000 parallel sentences, and the word embedding model has a vocabulary size of 17,277 English words.

Similarly, the Europarl corpus has 2,007,758 parallel sentences, and its word embedding model has a vocabulary size of 87,033 English words.

3.2 Gender bias Analysis for the dataset

To evaluate the amount of bias in the Balanced dataset, we build word embeddings, which is a vectorization of words following the Word2Vec (Mikolov et al., 2013) technique, and we assume that the presence of bias in word embeddings is a kind of reflection of the biases in the dataset (Caliskan et al., 2017). We use 128 as the number of dimensions for these vectors, a minimum count of 5 to remove poorly represented words and a bidirectional window of 3 words, that is, given a word $x[n]$, its "context" is

$$x[n-3], \ldots, x[n], \ldots, x[n+3]$$

To perform the gender bias analysis of these words embeddings, we use the measures proposed in previous works (Bolukbasi et al., 2016b; Gonen and Goldberg, 2019a). Inspired by these previous studies, we make use of the following lists of words:

- Definitional List 8 pairs (he/she; boy/girl; father/mother; male/female; his/her; himself/herself; man/woman; son/daughter)

- Biased List, which contains 1000 words, 500 female-biased, and 500 male-biased. (e.g., diet for female and hero for male)
- Extended Biased List, extended version of Biased List (5000 words, 2500 female-biased, and 2500 male-biased)
- Professional List 319 tokens (e.g., accountant, surgeon)

3.2.1 Gender Direction and Direct Bias

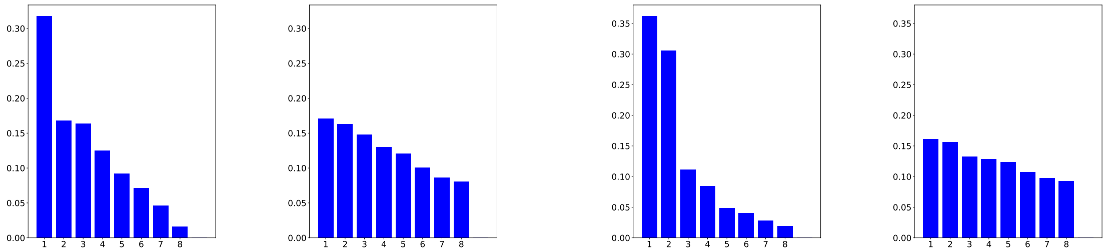


Figure 1: PCA Comparison between the gender base and a randomly generated base of 128 dimensions from Europarl (two graphs on the left) and Balanced (two graphs on the right) datasets.

Following the previous study (Bolukbasi et al., 2016a), we took the M gender pair difference vectors (Definitional List) and computed its principal components (PCs) to identify the gender subspace. We then generate a random base of M unit vectors of 128 dimensions for comparison. Figure 1 shows the PCA plots in both the gendered and the random vectors. In the EuroParl dataset, there is a clear dominance of one gender direction in the PCA from gender vectors. In the Balanced datasets, the supremacy is lower, but we can see that the 2 PCs from the left image (our gender base) explain almost 65% of the variance (information).

We take the definition of gender bias (Bolukbasi et al., 2016b), where they define the gender bias of a word $\overrightarrow{w}$ by its projection on the gender direction g. The higher the magnitude of the projection onto the previously defined base, the more biased the word is. We use the lists of neutral professions in (Zhao et al., 2018) to compute the direct bias of our Balanced dataset as follows.

$$\frac{1}{|N|}\sum_{\omega \epsilon N} |cos(\overrightarrow{\omega}, g)| \tag{1}$$

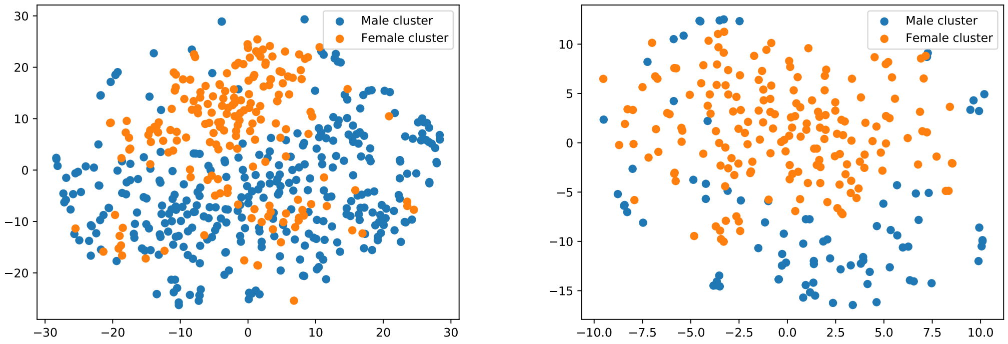
After filtering by words in our word embeddings model, we get N=147 for the Europarl dataset and N=140 for the Balanced dataset. Direct bias is 0.23 for the EuroParl, and 0.10 for the Balanced dataset[1]. This measure confirms that most words still have some of its information alongside the gender direction. These results are higher of what is reported in Bolukbasi's work (although it is not directly comparable). Having a lower N may interfere in the direct bias measure. We use the PCA analysis to measure the gender bias in the word embeddings. The extracted PCs could be used to debias such embeddings (Bolukbasi et al., 2016b), but we are not using them in current work.

3.2.2 Clustering

The clustering measure wants to evaluate if stereotypically-gendered words (Biased List) are easy to cluster based on their word embedding representations. The higher the clustering accuracy, the more bias the words embeddings have. We use *Scikit learn* (Pedregosa et al., 2011) toolkit to perform an unsupervised k-means clustering classification (with 2 clusters).

Figures 2a and 2b show the tSNE projections of the vectors for both Europarl and Balanced datasets, respectively. The clustering model trained with the Europarl aligns with gender with an accuracy of

[1] Find words used in https://github.com/adridjs/thesis2020/tree/master/genderbias/data. Files with *pca_professions* suffix.



(a) Clustering on EuroParl

(b) Clustering on Balanced

Figure 2: tSNE projection after K-means clustering on Balanced and EuroParl datasets.

77.67% and Balanced dataset word embeddings aligns with gender with an accuracy of 68.47%. Note that not all the words in the Biased List appear in fact, we were only able to use 512 words and 263 words[2] (out of 1000) from the original Biased List, in the Europarl and Balanced cases, respectively,

3.2.3 Classification

We want to know if we can classify stereotypically-gendered words (Extended Biased List) into masculine or feminine based solely on their word embedding representations. We build an RBF-kernel SVM classifier to discover if the model can generalize its predictions into other stereotypically-gendered words. We evaluate on the EuroParl and Balanced corpus.

We start from the Extended Biased List of the 5000 most-biased words in (Gonen and Goldberg, 2019b) according to the original bias (2,500 from each gender). As in previous experiments, in our datasets, this list is reduced to 1277[3], which are the words that we can find in both of our datasets. We then split these into train and test sets, drawing a 20% (255 words) for the train set and 1022 for testing the model's performance. The classifier's accuracy for the Europarl dataset is 80.59%, and the Balanced dataset is 73.28%.

3.2.4 Discussion

The accuracy reported in Europarl, and Balanced datasets are not comparable since both have different total and vocabulary words. We know that the word embedding representation changes when having more word repetitions. The results in absolute terms tend to report less bias in the Balanced dataset compared to the Europarl dataset. Moreover, clustering and classification results in absolute terms are lower than the ones noted in previous studies (Gonen and Goldberg, 2019b).

4 Use of Domain Adaptation techniques for Gender Bias Mitigation

In this section we use the gender-balanced dataset described in the previous section to mitigate the gender bias present in a standard MT system. We build the NMT system using the standard Transformer (Vaswani et al., 2017) on a large dataset. Our idea is to use fine-tuning techniques with the balanced dataset on this baseline system.

4.1 Methodology

The idea is that we have a parent translation model trained with unbalanced data, and we want to learn a child model taking advantage of the balanced dataset. To avoid catastrophic forgetting, where the child model forgets everything learned from the parent, we use the mix fine tunning strategy. This strategy, which consists of initializing the child model with the parent model and train it on a percentage of the

[2] Words used can be found in https://github.com/adridjs/thesis2020/tree/master/gender$_{b}ias/dataFileswithclustering\_wordssuffix$.

[3] Words used can be found in https://github.com/adridjs/thesis2020/blob/master/gender$_{b}ias/data/svm_{w}ords.txt$

unbalanced data set concatenated with the entire balanced data set, has been proven to mitigate the catastrophic forgetting problem (Chu and Dabre, 2019).

We train the parent model with large datasets. We then fine-tune it with 3 types of datasets: Balanced, a Mix of the Large and Balanced dataset, having different proportions of the large dataset into it, and Concat, which contains the entire Large and Balanced datasets (see Figure 3).

4.2 Experimental Framework

Generic Training Data To train the parent model, we used the English-Spanish EuroParl corpus (Koehn, 2005), which contains parallel data from the European Parliament's proceedings. We extract a part of the corpus that consists of 2 million parallel sentences. We applied a preprocessing step that consisted of tokenizing, truecasing, and filtering. We performed all these steps using scripts from the well-known Moses (Koehn et al., 2007) scripts.

Parameters We train the network for an undefined number of epochs until convergence with an early stopping policy. That policy consists of setting a *patience*, which means that if the validation loss does not improve in *patience* epochs, stop the training. We established that to 5 as it gives good results empirically. We used 512 embeddings dimension, 6 layers in the encoder and decoder, 8 attention heads. We used a batch size of 16, a dropout of 0.1, and a learning rate of 0.001. We optimized with Adam.

Architecture We use the Transformer (Vaswani et al., 2017) as baseline NMT model architecture, an encoder-decoder architecture based on attention-based mechanisms that boost the performance in NMT tasks compared to RNNs or CNNs architectures.

4.3 Fine-tuning

The baseline model is fine-tuned with a dropout to 0.3. This is used as a regularization technique together with the mixed fine-tuning approach to handle the catastrophic forgetting problem. This is the only modified hyperparameter between the baseline training and fine-tuning steps.

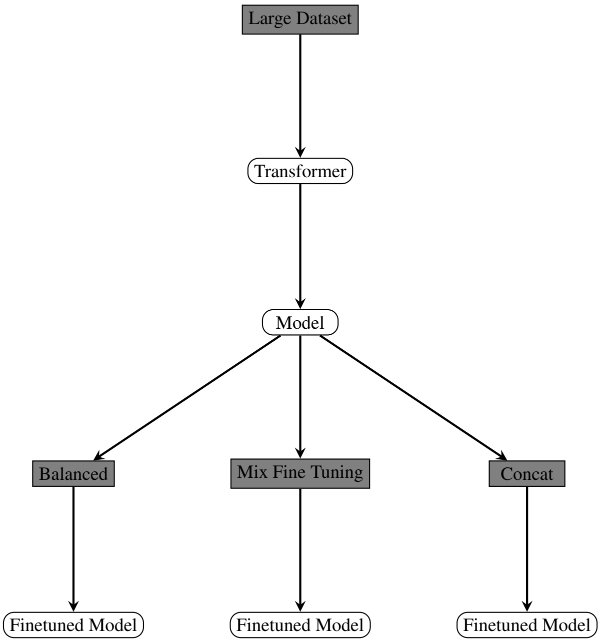


Figure 3: NMT training pipeline. The gray boxes represent the corpus used to train the model that they are pointing to.

Balanced We hypothesize that fine-tuning on a corpus balanced in gender will improve the accuracy in gendered translations. We use the corpus extracted by Gebiotoolkit as reported in Section 3 - balanced in gender - to test this hypothesis. Note that the Balanced data is from a different domain than the training and test data.

Mix This approach is building a dataset based on a mix of EuroParl and Balanced datasets. We study the influence on gender bias and NMT performance by having more or less in-domain data fed in the fine-tuning step. More percentage means more EuroParl data.

Concat This approach consists of concatenating the whole EuroParl corpus with the Gender-Balanced biographies dataset.

Note that Balanced and Concat could also lie into the mix fine-tuning strategy, being 0% and 100%, respectively, the percentage of sentences from the Europarl corpus.

4.4 Results

Our findings are presented from Table 1 to 4. We report two baseline models: one trained with the EuroParl corpus and another trained with the concatenated dataset (Base-Concat) composed by EuroParl and Gebiotoolkit dataset. We report an evaluation in terms of translation performance and an evaluation in terms of gender bias accuracy.

4.4.1 Translation Evaluation

We use BLEU (Papineni et al., 2002) to evaluate the performance of our translation models on the WMT13 test set (*newstest2013*)[4]. The second baseline shows an increment of 1.5 points in the English-to-Spanish model and almost the same increment in the reversed model.

Translation Performance		
Corpus	en2es	es2en
EuroParl	26.87	25.50
Base-Concat	28.37	26.91
FT-Balanced	27.51	27.80
FT-Mix 5%	28.51	28.71
FT-Mix 10%	28.52	28.76
FT-Mix 20%	28.72	28.78
FT-Mix 30%	**28.76**	28.95
FT-Mix 40%	28.61	**29.05**
FT-Concat	28.68	28.29

Table 1: BLEU results for the different trained systems.

All the fine-tuned models surpass these two baselines, except the Balanced in the English-to-Spanish model. The best performance achieved in the English-to-Spanish model is the one where 30% of EuroParl data is present, while in the Spanish-English model the proportion is of 40%. We get final improvements over the best baseline system (Base-Concat) of up to 2 BLEU points.

4.4.2 Gender Bias Evaluation

We use the gender bias evaluation pipeline from (Stanovsky et al., 2019), also known as WinoMT, to evaluate the gender bias in these models. The dataset consists of 3,888 sentences. In each of these sentences, a primary entity that is coreferent with a pronoun, and a secondary entity tries to trick the translation system. The scripts provided by the authors extracted the grammatical gender of the primary entity from each translation by automatic word alignment and followed by morphological analysis. Then, it compares the translated primary entity with the annotated gender. The objective is to have a translation where the primary entity's gender matches the gold annotated one.

[4] http://www.statmt.org/wmt13/

General Bias For the general bias measures, the best performance is achieved with FT-Concat, getting 49.8% accuracy at identifying the correct gender when translating into Spanish, which is an improvement of 2.5% points concerning the highest baseline, which is the Base-Concat.

General Gender Bias				
Corpus	Acc.	F-Score		Δ_g
		M	F	
EuroParl	46.6%	59.8%	31.3%	28.5
Base-Concat	47.3%	60.3%	32.4%	27.9
FT-Balanced	48.3%	60.4%	33.8%	26.6
FT-Mix 5%	47.5%	60.2%	32.0%	28.2
FT-Mix 10%	47.9%	60.4%	32.6%	27.8
FT-Mix 20%	48.2%	60.7%	33.3%	27.4
FT-Mix 30%	48.8%	60.8%	35.2%	25.6
FT-Mix 40%	49.0%	**61.1%**	35.5%	25.6
FT-Concat	**49.8%**	59.9%	**41.7%**	**18.2**

Table 2: Accuracy in the General WinoMT test set. F1-Score for masculine and feminine scores, and difference in performance between masculine and feminine scores (Δ_g)

Pro-Stereotypical Bias In this setup, FT-Concat performs much better than any other model. Its accuracy is 10 points higher than the best baseline system. Its F-score differences are also the lowest, meaning less bias than in any different trained model.

Pro-stereotypical Gender Bias				
Corpus	Acc.	F-Score		Δ_g
		M	F	
EuroParl	53.5%	67.7%	35.9%	31.8
Base-Concat	56.2%	69.1%	38.8%	30.3
FT-Balanced	59.3%	70.0%	47.7%	22.3
FT-Mix 5%	57.3%	69.3%	43.2%	26.1
FT-Mix 10%	57.8%	69.4%	44.1%	25.3
FT-Mix 20%	58.2%	69.9%	44.6%	25.3
FT-Mix 30%	58.9%	70.3%	46.0%	24.3
FT-Mix 40%	59.0%	70.8%	45.5%	25.3
FT-Concat	**66.3%**	**74.1%**	**62.0%**	**12.1**

Table 3: Accuracy in the WinoMT test set. Pro-Stereotypical translations.

Anti-Stereotypical Bias Lastly, the best model performance is obtained on the FT-Mix40% model, which has an accuracy of 45% (lowest for all the setups). The minimum F-score difference is 28,9%, which is very high (also competitive to the commercial reference systems reported in the original paper (Stanovsky et al., 2019). In general, we can see that in this setup, the models do not perform very well. This reveals that the systems are still biased, as we have low anti-stereotypical and high pro-stereotypical translation performance.

5 Conclusions

The motivation of our work lies in the hypothesis that the use of a gender-balanced dataset can diminish the gender bias in NMT systems.

For doing so, we first report an analysis of this Balanced dataset in terms of gender bias by using a word embedding evaluation set of measures. This analysis shows that this Balanced dataset only encodes

Anti-stereotypical Gender Bias				
Corpus	Acc,	F-Score		Δ_g
		M	F	
EuroParl	44.3%	57.1%	28.2%	28.9
Base Concat	39.0%	52.3%	21.5%	30.8
FT-Balanced	43.1%	56.7%	22.9%	33.8
FT-Mix 5%	43.1%	56.6%	23.5%	33.1
FT-Mix 10%	43.4%	57.1%	23.2%	33.9
FT-Mix 20%	44.1%	**57.4**%	24.7%	32.7
FT-Mix 30%	44.3%	57.1%	26.6%	30.5
FT-Mix 40%	**45.0%**	**57.4%**	**28.6%**	**28.8**
FT-Concat	44.5%	57.0%	26.3%	30.7

Table 4: Accuracy in the WinoMT test set. Anti-Stereotypical translations.

a small amount of bias when compared in absolute terms with other more massive datasets. Note that we can use these representations by downstream applications with the ability to have little gender bias.

Then, after this analysis, we use fine-tuning techniques to reduce gender bias in a standard MT system. Results show that even if our balanced dataset is from a different domain than the training and the test of the MT system, it does improve the translation quality (up to 2 BLEU points), and it can mitigate the gender bias in a significant amount (up to a 12.5% accuracy).

Acknowledgments

This work is supported in part by the Catalan Agency for Management of University and Research Grants (AGAUR) through an Industrial Ph.D. Grant, the Spanish Ministerio de Ciencia e Innovación and by the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947657).

References

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA, July. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of Advances in Neural Information Processing Systems 29*, pages 4349–4357.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora necessarily contain human biases. *Science*, 356:183–186.

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *CoRR*, abs/1906.07978.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2019. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. In *Proceedings of 12th Language Resources and Evaluation Conference (LREC)*.

Marta R. Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

Catherine D'Ignazio and Lauren Klein. 2018. Data feminism. MIT Press.

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 ACL Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August.

Hila Gonen and Yoav Goldberg. 2019a. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT (1)*, pages 609–614. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019b. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

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, July. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online, July. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhangyang Xiong, Zhongyuan Wang, Changqing Du, Rong Zhu, Emily Xiao, and Tao Lu, 2018. *An Asian Face Dataset and How Race Influences Face Recognition: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part II*, pages 372–383. 09.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *CoRR*, abs/1809.01496.