9-2020

Mitigating Gender Bias in Neural Machine Translation Using Counterfactual Data

Alan Wong
The Graduate Center, City University of New York

Mitigating Gender Bias in Neural Machine Translation Using Counterfactual Data

by

Alan Wong

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2020

Mitigating Gender Bias in Neural Machine Translation Using Counterfactual Data

by

Alan Wong

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Gita Martohardjono Executive Officer

The City University of New York

ABSTRACT

Mitigating Gender Bias in Neural Machine Translation Using Counterfactual Data

by

Alan Wong

Advisor: Kyle Gorman

Recent advances in deep learning have greatly improved the ability of researchers to develop effective machine translation systems. In particular, the application of modern neural architectures, such as the Transformer, has achieved state-of-the-art BLEU scores in many translation tasks. However, it has been found that even state-of-the-art neural machine translation models can suffer from certain implicit biases, such as gender bias (Lu et al., 2019). In response to this issue, researchers have proposed various potential solutions: some have proposed approaches that inject missing gender information into models, while others have attempted modifying the training data itself. We focus on mitigating gender bias through the use of both counterfactual data augmentation and data substitution techniques, exploring how the two techniques compare when applied to different datasets, how gender bias mitigation varies with the amount of counterfactual data used, and how these techniques may affect BLEU score.

ACKNOWLEDGMENTS

To my advisor, Professor Kyle Gorman, whose thoughts and suggestions helped bring this work to fruition; to my family, for all of their support through the years; to Emily and Sean for making even the most chaotic evil random number generator bearable; and to all the other people I've met at the Graduate Center, for making the experience so memorable.

Contents

List of Tables

1 Introduction

In recent years, machine translation as a field has greatly benefited from modern developments in the field of deep learning, as well as advancements in the form of more powerful custom hardware such as graphics processing units (GPUs) and tensor processing units (TPUs). Constant advancements with novel neural architectures have further driven this progress. However, despite the incredible progress that has been made, there still exist systematic issues regarding bias within both the machine translation sub-field and natural language processing (NLP) as a whole (Sun et al., 2019). These biases, of course, include gender bias.

1.1 Bias issues and social desirability

Translation of texts has been performed, and valued, by society long before modern computational systems existed. The recent development of statistical and neural machine translation (NMT) systems has provided a breakthrough for the field, enabling many to access translation tools on demand—an incredible boon for society. Clearly, it is desirable for society as a whole to improve such tools as much as possible. However, these valuable tools are not without fault. They are still plagued by biases, many of which represent problems of social equity and fairness. Factors such as age, race, and gender, among many more, can find their way into computational translation systems. One example of where this is the case is the following: going from English to Spanish, *I am a nurse* might be translated as *Soy enfermera*. In Spanish, many occupational nouns have gendered forms marked by suffixes: typically *-o* for masculine forms and *-a* for feminine forms, although there exist others. In this example, *enfermera* has a feminine inflection despite the nurse potentially being male, translating to *I am a (female) nurse*—this essentially implies a world state where there are no male nurses. Clearly, this is an unfair assumption.

Yet the consequences of biases do not end there; NLP is far-reaching, extending its ap-

plications to many fields. In some fields, even small errors can be disastrous. Consider the case of medical applications. Recently, there has been work interested in translation between technical medical terminology and more vernacular language (Seiffe et al., 2020). If there were to be a biased version of such a translation system, it could easily result in miscommunication or misdiagnosis. For example, the model might translate a male patient's description of chest pain as angina, instead of (a symptom of) breast cancer. While hopefully such situations would not occur, using a biased model does introduce the possibility.

1.1.1 Gender bias in NLP

Machine translation models are typically trained on substantial amounts of parallel textual data; the number of parallel sentences can range anywhere from hundreds of thousands to hundreds of billions in some cases. While a large quantity of training data is typically beneficial for NLP models, it also carries a certain drawback: it is extremely difficult to establish what kinds of sources and biases a large collection of text data may contain. Due to the sheer scale, it is not feasible to individually examine sentences or to know definitively where any one sentence might have originated. Work by Bolukbasi et al. (2016) shows that even models trained on major, seemingly balanced sources such as Google News exhibit alarming amounts of gender bias. In other words, models easily pick up on any biases present in the data. The ability of data pipelines to exacerbate existing biases only adds to the problem. Rudinger et al. (2018) find that translation models overgeneralize on gender information, ultimately amplifying bias beyond what is already present in the data. These works show how even trace amounts of bias in the source data can easily be multiplied by NLP models.

The previous examples only examined the subfield of machine translation, but gender bias pervades many other areas of NLP as well. A large amount of recent work has examined gender bias in (contextual) word embeddings. Word embeddings are a form of geometric representation for words: a word is treated as an arrow pointing in a space, its direction

and length dependent on its linguistic characteristics. Each dimension in the space can be considered a characteristic (e.g., present tense on one axis, gender on another, and so on). Given multiple dimensions, words end up at specific points in space. As the dimensions are like characteristics, words that are associated with each other appear closer together.

Recent work has found that even the popular Embeddings from Language Models (ELMo) has inherited gender bias from the corpus it was trained upon: Chelba et al. (2014)'s One Billion Word Benchmark. Zhao et al. (2019) specifically note how ELMo has male pronouns occurring thrice as frequently as female ones and how male pronouns occur more often with all mentions of occupations. These word embedding biases, when applied to downstream tasks, transfer the gender bias as well. Other work by Chaloner and Maldonado (2019) has examined the potential for gender bias in word embeddings when trained on different domains of data. Testing four different-domain corpora, they find evidence of gender bias in all cases. More notably, they find that the type of gender bias exhibited differs between the domains. Some corpora transmit gender bias regarding career versus family life, while others involve strength versus weakness, etc. Although the exact methods employed differ, both Chaloner and Maldonado (2019) and Zhao et al. (2019) manage to debias word embeddings, essentially pushing gender biased words' positions in space away from their biases.

Some other areas of NLP that have struggled with gender bias are the subfields of sentiment analysis and text classification. Bhaskaran and Bhallamudi (2019) found statistically significant differences in mean sentiment scores (i.e., probability of positive sentiment) between sentences which mention stereotypically male and female occupations. In each of their three models, the inequality persisted. In two models, females were determined to have a higher mean sentiment score than males, and in the third model males had a higher mean sentiment score. Working on abusive language detection, Park et al. (2018) measured the extent of gender bias in abusive language corpora and applied strategies for effectively mitigating it, determining that models trained on the corpora were biased towards detecting the

abuse of women. As seen through these examples, gender bias is quite pervasive in NLP.

1.2 Mitigating bias

While it is ideal to diminish bias as much as possible, being able to do so first requires one be able to identify and measure the bias in question. Unfortunately, this is where the task becomes much more challenging. Quantifying gender bias itself is difficult without first using some sort of downstream task, such as training word embeddings or translation models. Even after training these sorts of tasks, measuring it is difficult. Despite the problem gender bias poses, there are currently very few benchmarks for measuring it. Zhao et al. (2018) introduced the WinoBias set, a collection of anaphora resolution sentences focused on gender bias and Rudinger et al. (2018) introduced the Winogender set, another collection of the same sort. They are comprised of English sentences that contain a pair of often-stereotyped professions (e.g., nurses and doctors) and a gendered pronoun that ambiguously references one of them (i.e., requires anaphora resolution). Stanovsky et al. (2019) later built upon their work, concatenating them to form the WinoMT set. Without consistent means by which to measure bias, it becomes more difficult to recognize and treat. It also becomes more difficult to evaluate whether a bias mitigation method is truly working; to a certain degree, there is an equivalency problem between researchers' work.

1.2.1 Dataset domains and adaptation

One perspective with which researchers have viewed the bias problem is that of datasets comprising discrete, topical domains. In this view gender bias, among other forms, is primarily attributable to the original source of the data containing fewer or more stereotyped representations of a particular gender. By using such sources for training data, the gender bias present in the source will be passed along to the resulting models. Reagle and Rhue (2011) show that even encyclopedic data sources such as Wikipedia and Britannica, which many

would assume to be balanced, contain consistent gender bias. They show that Britannica lacks the coverage of female figures that Wikipedia has, but that the entries that are missed by Wikipedia are disproportionately of female figures. In order to overcome those biases using this perspective, it is necessary to treat the problem as one of domain adaptation.

In terms of those who have attempted to address this problem accordingly, Saunders and Byrne (2020) forego attempts at direct dataset modification and instead apply transfer learning techniques. They find success in initializing model training on a better-balanced handcrafted dataset (i.e., domain) and then resuming training on an in-domain dataset known to have gender bias. In this way, they effectively initialize model parameters from a balanced state, successfully adapting the domain.

1.2.2 Direct injections

While many researchers concentrate their efforts on mitigating bias by using different domain data or adding anti-biased sentences, an alternative approach taken by some is to account for bias by directly injecting the necessary gender data or corrections. For example, Vanmassenhove et al. (2018) insert speaker gender information tags (available in Europarl source files) into the English source side of bilingual Europarl data. A resulting sentence may appear as *FEMALE Madam President...*, as opposed to just *Madam President...*. They find that injecting this kind of gender feature into a neural machine translation system can significantly improve translation quality.

Taking the latter approach, later work by Moryossef et al. (2019) finds success in injecting gender information after model training. Leveraging a monologue from the "Sarah Silverman: A Speck of Dust" comedy show, they are able to ascertain line-by-line gender information. With this, they use a black-box approach to provide the missing gender information for translations without the need to train or retrain the original translation model. In doing so, they improve translation quality, as measured by BLEU score. While both methods

prove to be effective, they also require having access to the gender information—that is not always possible, however.

1.2.3 Counterfactual use

A popular approach other researchers have taken is targeting the data that models are trained on in order to reduce gender bias. A specific strategy utilized is that of counterfactual data augmentation, which inserts counterfactuals—sentences where gendered components are reversed, or countered—into the training data, alongside the originals. An example of a Spanish sentence and its counterfactual is shown in Table 1. The reasoning behind using counterfactual data is that it can specifically target gender bias by removing existing predispositions in the training data. If training data becomes gender balanced, it becomes less likely for models trained on it to pick up on one gender bias as opposed to its inverse. This kind of technique is particularly helpful when translating from languages like English (which lacks morphosyntactic gender marking outside of personal pronouns) to languages such as Spanish (which marks morphosyntactic gender on adjectives, determiners, and nouns using affixes), mitigating gender bias without sacrificing grammaticality (Zmigrod et al., 2019).

Empirically, the work of multiple researchers has also found merit in both the theory behind and application of counterfactual data. Zhao et al. (2018) have provided evidence that NLP systems can be successfully "cued" in, essentially trained, to ignore biases. Additionally, it has been shown by Lu et al. (2019) that counterfactual data augmentation can outperform more tradition means of mitigating gender bias, such as word embedding debiasing, providing an excellent case for its use. Of course, they also show that counterfactual data augmentation can reduce gender bias without severe loss of accuracy.

While counterfactual data augmentation appears to be a useful technique, there have also been those who suggest that the duplication effect it causes may be problematic. This refers to how the augmented counterfactual sentences are highly similar to the originals, only

Original:	El chico es bueno.
Original Gloss:	The boy is good.
Counterfactual:	La chica es buena.
Counterfactual Gloss:	The girl is good.

Table 1: Example of counterfactual data for Spanish.

differing by the reinflected words. In light of that possibility, they introduce the technique of counterfactual data substitution, where the counterfactuals are not used in addition to the originals, but as a replacement (Hall Maudslay et al., 2019). The debate over the potential effects of duplication and the merits of substitution over augmentation inspire a large component of the experiments presented later in this work.

1.3 Evaluation metrics

Despite how detrimental bias is to machine translation systems and natural language processing in general, there exist few tools by which to precisely measure it. This issue is only exacerbated when targeting a specific form of bias, such as gender bias. The difficulty of measuring bias can in part be attributed to the intrinsic difficulty of recognizing biases in the first place. Typically biases manifest as unintended products of the training data. In the case of gender bias, one sentence referring to a male programmer is innocuous enough, but when the same phenomenon happens to occur within thousands of the training sentences, there are undesirable consequences for the final model's output. Given the massive scale at which commercial-grade translation models are trained and produced, these individual cases in the data—which do not constitute the whole of gender bias in and of themselves—can combine to result in a model that never creates a hypothesis with female programmers.

Currently, a majority of machine translation models are evaluated using Bilingual Evaluation Understudy (BLEU) scores, which are metrics that compare a model's hypothesized translation with a reference translation (Papineni et al., 2002). More specifically, BLEU

scores compute a value from 0 (poor translation) to 1 (good translation) by using a modified n-gram precision and a sentence brevity penalty. However, this commonly used metric is also problematic for many reasons, most notably its failure to correlate with actual human judgments and its reliance on typically unreported parameters (Callison-Burch et al., 2006; Post, 2018). Moreover, BLEU scores are based on how well hypothesized translations can mimic target translations. If gender bias is exhibited in the target translation, BLUE scores will decrease for a non-gender biased model—the non-biased model's translations will not match the biased target translations, causing a loss in BLEU score.

Taking the earlier example of *I am a nurse*, both translations of the sentence in Spanish, *Soy enfermera* and *Soy enfermero* are well-formed and correct. However, if only one of those translations is treated as the target translation in the test set, then a translation model's BLEU score is penalized for not providing that answer. Given the likelihood of corpora being gender biased, this would mean penalization for not adopting gender biases. In short, if the test set itself codifies gender bias, then the very tool meant to serve as a measure of success instead becomes a barrier to it. It will be strictly impossible to debias models past what the test set allows. BLEU scores provide no way of accounting for these situations, further highlighting their inadequacy.

If one foregoes BLEU scores however, there are few mainstream measures for evaluating machine translation in general, let alone its biases. In response to this, many researchers have either proposed or developed specific challenge sets (i.e., evaluation sets). These test sets, when translated by a trained machine translation model, are meant to produce a more specific measure of some bias. For gender bias specifically, there exists the WinoMT challenge set by Stanovsky et al. (2019), which is a combination of the Winogender and Winobias test sets introduced by Rudinger et al. (2018) and Zhao et al. (2018), respectively. Importantly, the WinoMT set is only in English, is evenly split between male and female referencing anaphora resolutions, and is only 3,888 sentences in length.

Due to it being the only substantive gender bias set currently available, the WinoMT set is the one that is used for the evaluation of the experiments in this work. Given how specific the test set must be, they are, by necessity, typically hand-crafted. As a result, the test sets are usually quite small in size, particularly when considering the normal scope of natural language processing applications. Stemming from this small size are possible issues regarding domain mismatch (as might be the case for the anaphora-resolution heavy nature of this set), where a translation model trained specifically with more of such sentences will perform better than an otherwise superior translation model. Furthermore, generated sentences are rarely an adequate substitute for real data. These factors together do threaten the validity of the WinoMT set, to an extent. They raise questions about how applicable the results of such specific tests are—will a given metric correlate with human judgments, or other tests for measuring gender bias? At this point, it is impossible to truly say. Further discussion of this is relegated to Chapter 4. As a secondary metric, BLEU scores are utilized simply to verify the effect that counterfactual data may or may not have on them.

1.4 The present work

Having reviewed the current work regarding gender bias mitigation in neural machine translation, there are various angles at which to tackle the problem. The present work, however, examines the effectiveness of applying counterfactual data to the training data; this takes the form of both counterfactual data augmentation and counterfactual data substitution. As a primary objective, the effectiveness of augmentation is determined and then compared with that of substitution. Secondly, we experiment with varying amounts of counterfactual data augmentation and substitution, in order to gain an idea of the optimal quantity for mitigating gender bias. In all cases the evaluation set used is the aforementioned WinoMT set, with the primary metric being gender accuracy. The purpose of using such a set is to maintain some sort of standard for evaluating gender bias. As it is scientifically interesting to

examine what kind of impact counterfactual data augmentation and substitution may have on conventional metrics like BLEU score, a third experiment examines how the different translation models (unmodified, augmented, and substituted) compare in terms of BLEU on a separate test set.

Following this section is Chapter 2, which details the datasets, software tools, and methods used for this work's experiments. That is followed by Chapter 3, which provides a detailed account of the experiments and their results. Chapter 4 delves into a discussion of the findings, noting aspects of the work which may be improved, as well as issues that can only currently be mentioned. They are followed by a short section on potential future work. Lastly, Chapter 5 provides a conclusion to the work, reiterating the key findings.

2 Materials and Methods

This chapter details the information regarding the datasets used for the experiments, as well as the means by which that data was preprocessed, the tools used for the experiments, information about the utilized Transformer parameters and hyperparameters, and the primary evaluation metrics used to judge the experiments.

2.1 Data

The English-Spanish data for the experiments came from three corpora. The first was the Wikipedia corpus version 1.0,[1] consisting of translated sentences from the Wikipedia website (Wołk and Marasek, 2014). Wikipedia is a free, multi-lingual, and collaborative encyclopedic website. The sentences are taken from Wikipedia pages that exist in both English and Spanish versions. The second was the Global Voices corpus version 2017q3,[2] hereby abbreviated Global, a corpus comprised of news stories gathered and compiled by CASMACAT from the Global Voices website. The headlines and stories are written by their news team and report on events from across the globe. The translation of the original articles into other languages is performed by volunteers. The third and final corpus was the Tatoeba corpus version 20190709.[3] Tatoeba is a collection of human translated sentences taken from the popular website of the same name. The sentences and translations are contributed by site members in an open-collaboration fashion. All three corpora were obtained from OPUS and are freely accessible online (Tiedemann, 2012). For the English-Spanish (EN-ES) language pair, the Tatoeba corpus is smallest in size, with the Global corpus in the middle, and Wikipedia corpus as the largest. Basic information regarding the dataset sizes, median word length (without punctuation), and mean sentence length (without punctuation) is provided in Table 2.

[1] `http://opus.nlpl.eu/Wikipedia.php`
[2] `http://opus.nlpl.eu/GlobalVoices.php`
[3] `http://opus.nlpl.eu/Tatoeba.php`

Corpus	Sentence Pairs	Words (Mil.)	Median Word Len.	Mean Sent. Len.
Wikipedia	1,741,038	70.90	EN: 5, ES: 5	EN: 23.81, ES: 23.66
Global	693,544	27.38	EN: 5, ES: 5	EN: 21.06, ES: 22.12
Tatoeba	203,272	2.73	EN: 4, ES: 4	EN: 7.25, ES: 6.92

Table 2: Corpora word and sentence information.

2.2 Preprocessing

In terms of data-preprocessing, both the English and Spanish components of the corpora were tokenized using the Moses tokenizer. This was done through the use of the `sacremoses` Python library version 0.0.43, a Python port of the original implementation. The tokenized data for both languages was then split into training and validation sets using a matching random seed to preserve the parallel nature of the sentences. 80% of the dataset was allocated to the training set, while the remaining 20% was allocated to the validation set. This ratio was used for all of the datasets. The sizes of the resultant partitions of the baseline datasets are described in Table 3.

Prior to training, the data was also subjected to the byte pair encoding (BPE) subword algorithm, a method of data compression where certain components of words are replaced with shorter ones that do not otherwise occur in the data (Sennrich et al., 2016). The `sentencepiece` library's implementation of this algorithm (version 0.1.91) was utilized for this purpose. A vocabulary size parameter of 16,000 was utilized to train the models used for the byte pair encoding of each of the data partitions, although recent work has also shown the success of using a low number of merge operations (0–4k) for training BPE models when used for Transformers (Ding et al., 2019). The test sets were preprocessed similarly to the training and validation data prior to evaluation.

Corpus	Partition	Sentence Pairs
Wikipedia	Train	1,449,143
	Validation	362,286
Global	Train	554,836
	Validation	138,709
Tatoeba	Train	162,618
	Validation	40,655

Table 3: Number of sentence pairs per preprocessed baseline partition.

2.3 Modeling and tools

2.3.1 Counterfactual production

Due to the scarcity of datasets containing information on gender morphology (specifically oppositely gender-inflected sentences), it was necessary to generate counterfactual data using more of a heuristic method. This was accomplished using a three step procedure, the first of which was morphological analysis using the `spaCy` library (Honnibal and Montani, 2017). This library was used to determine the morphological tags (i.e., features) of each word in the corpora. Only words that were tagged with the correct part of speech (noun, determiner, adjective) and a gender (in this case, masculine or feminine) proceeded to the next step.

The second step involved the application of handwritten morphological rules. For Spanish, nouns, determiners, and adjectives were targeted for opposite-gender inflection. Simple morphological rules were applied to nouns and adjectives using Python (e.g., the ending *-o* of *chico* would be dropped and replaced with *-a*, forming *chica*). The determiners, forming a closed set, were replaced using a dictionary mapping to their opposites where applicable. Only in cases where the nouns were reinflected into valid words were determiners and adjectives subsequently inflected to agree. Only for Spanish sentences that underwent changes were the corresponding English sentences altered to match (e.g. *he* changed to *she*).

In the third step, the `pyspellcheck` library was used. The library utilizes word frequency

lists based on the OpenSubtitles corpus and an algorithm for spell checking proposed by Peter Norvig to determine the correctness of a target spelling. In the context of these experiments, it was used to verify that a proposed, oppositely-inflected word actually appeared in common usage of the language with some degree of frequency. It should be noted that specific gender inflections on certain words may carry subjective sociopolitical notions, or not be considered "prescriptively correct" (e.g., the standard Spanish word for *president* is *presidente*, which is grammatically masculine, but the grammatically feminine *presidenta* is also in common use, albeit considered by some to be incorrect). The use of a word frequency based tool was meant to partially circumvent these notions: as long as the word, as inflected, appeared in the wild with sufficient frequency, it was considered a valid word.

2.3.2 Transformer training

To train and evaluate the models for these experiments, the `fairseq` sequence modeling toolkit was used (Ott et al., 2019). The Transformer architecture was used, with the parameters and hyperparameters generally based upon those of the single GPU base model as described by Vaswani et al. (2017). This included using the Adam optimizer with β values of .9 and .98 (Kingma and Ba, 2015), a warm-up period of 4,000 steps for reducing label smoothed cross entropy (Szegedy et al., 2016), label smoothing with $\epsilon = .1$, 6 encoder hidden layers, 6 decoder hidden layers, and 8 attention heads. The few deviations were a higher dropout value of .3, encoder-decoder dimensionalities of 256 each, and a hidden layer size of 1024. An important point is that the random seed utilized was set to 1 for all experiments. There has been research on the matter suggesting that choice of random seed can influence the learning of language models in significant ways, suggesting that the same could be the case here (Dodge et al., 2020). However, I did not attempt to experiment optimizing for the random seed during this work. This is addressed later, in Chapter 4.

Training on the data continued until such a point that the validation set loss failed to

decrease for five consecutive epochs. Every epoch was saved as a checkpoint, with only the best checkpoint (as determined by a minimized validation loss value) being evaluated on the test set. A beam search width of 5 was used for decoding during evaluation.

2.4 Evaluation metrics

In terms of the test set, the data used was the challenge set proposed by Stanovsky et al. (2019), which is comprised of sentences containing frequently gender biased professions that require anaphora resolution (e.g., *The developer argued with the designer because she did not like the design*). For evaluating the quality of test set translations, a gender accuracy value was utilized: accuracy increased when the proposed translation's morphological genders correctly matched the gold targets. This was done using the tool provided by Stanovsky et al., built upon the `fast_align` unsupervised word aligner (Dyer et al., 2013).

Although acceptable translations can often differ by a few words, in this case the focus was solely on gender accuracy. Regardless of how close the translation might otherwise have been, if the incorrect gender was produced, the translation was marked as incorrect. For this reason, and others stated earlier in the introduction, little emphasis is placed on metrics such as BLEU. However, in the interest of showing the potential impacts of counterfactual data augmentation and substitution on conventional metrics such as BLEU, it is used to gauge performance on a second test set: the test set for the 2010 shared task on translation for European languages, which consists of 1,199 lines of text in both English and Spanish.

3 Experiments

This chapter conveys the details and results of this work's experiments, which are broadly separated into three groups.

3.1 Group 1: Augmentation vs. substitution

The first experiments focused on whether or not there was a substantive difference in gender bias mitigation when using full counterfactual data augmentation as opposed to full counterfactual data substitution.[1] After substitution and augmentation, the modified datasets were preprocessed in the same manner described in Chapter 2. Each of the datasets was then used for training and the resulting models evaluated on the WinoMT set. Table 4 shows the results, comparing the original (henceforth baseline) dataset performances on the WinoMT set with the augmented and substituted dataset performances for each corpus. The absolute changes in accuracy between the baseline and modified datasets are presented in parentheses, next to the absolute accuracy values.

In terms of gender accuracy, we found that the model trained on the augmented Tatoeba dataset performed the best with a gender accuracy score of 60.9%. Also important to note is how the use of substitution on the same dataset produced a decreased gender accuracy of 41.9%. This suggested a possible interaction between dataset size and modification technique, where small datasets may benefit more from augmentation. Although the model trained on the Wikipedia dataset still performed decently at approximately 43% accuracy for both techniques, the model trained on the Global dataset appeared unable to approach the other two in terms of gender accuracy gain: it was the only case where gender accuracy decreased using full augmentation and full substitution.

One finding from the changes in gender accuracy was that the Wikipedia dataset appeared to consistently benefit the most from counterfactual data modification. While the

[1] Here, "full" refers to how every sentence that qualified for opposite-gender reinflection was changed and used in the modified datasets.

Corpus	Baseline	Substitution	Augmentation
Wikipedia	36.0	42.7 (+ 6.7)	43.4 (+ 7.4)
Global	54.8	42.8 (−12.0)	43.5 (−11.3)
Tatoeba	43.2	41.9 (− 1.3)	60.9 (+17.7)

Table 4: Gender accuracy scores for base and modified corpora.

Tatoeba dataset showed a large increase for augmentation (+17.7%), it also displayed a decrease with substitution (−1.3%). Conversely, the Global dataset showed more dramatic, negative changes with its gender accuracy falling by 12 and 11.3 using substitution and augmentation, respectively. Wikipedia, however, improved using both techniques.

Considering the results thus far, it appeared that augmentation was the more effective technique when applied at full scale. Full augmentation showed a greater potential for gender accuracy improvement across the datasets, with the only decrease being smaller than the one associated with substitution for that dataset. However, these results also raised questions about whether one data modification technique might be more effective at non-full scale and whether full scale is actually the optimal choice.

3.2 Group 2: Amount of counterfactual data

The next set of experiments investigated how much counterfactual data was optimal, and which modification technique would be optimal at which values. Generally, it is preferable to use natural data, as opposed to handcrafted or computer-generated data. There are computational and/or monetary costs associated with data creation, as well as potential effects on test metrics stemming from its use. Therefore, utilizing only the amount of counterfactuals strictly necessary to control gender bias is ideal: this minimizes any potential drawbacks of using too much counterfactual data. Using probabilistic values sweeping across {0, .25, .5, .75. 1.0}, augmentation and substitution were performed on the datasets. Here, 0 refers to the baseline and 1.0 refers to the full augmentation and substitution from the first exper-

Corpus	Modification	0	.25	.50	.75	1.0
Wikipedia	Augmentation	36.0	38.2	44.4	38.7	43.4
	Substitution	36.0	35.5	42.2	40.5	41.9
Global	Augmentation	54.8	47.5	56.0	53.3	43.5
	Substitution	54.8	44.0	56.2	46.4	42.8
Tatoeba	Augmentation	43.2	47.1	53.3	55.0	60.9
	Substitution	43.2	45.7	45.6	44.0	41.9

Table 5: Gender accuracy scores for corpora modified at different rates.

iment. At .25, sentences that qualified for opposite gender reinflection were modified at a rate of 25%, at .50 a rate of 50%, and at .75 a rate of 75%. The augmented and substituted datasets were preprocessed before being used for training, all with the same hyperparameters, parameters, and architectures as detailed in Chapter 2. The results are summarized in Table 5.

Here, we saw that the best overall performance was obtained by using the Tatoeba corpus augmented at a rate of 1.0. We observed an increase in gender accuracy of 19% compared to using full substitution and of 17.7% in comparison to using the baseline Tatoeba dataset. In terms of the best substitution rate for the Tatoeba dataset, a rate of .25 was optimal, producing a score of 45.7%. Interestingly, the Tatoeba dataset was the only one where gender accuracy consistently increased with augmentation rate.

For the Global dataset, the best performance was achieved with substitution at a rate of .50, although augmentation at the same rate produced a gender accuracy score just 0.2% lower. A peculiar characteristic of this dataset was how the baseline was relatively high, and how gender bias was only mitigated at the .50 rate level. At all other rates, the bias appeared to have increased (i.e., gender accuracy decreased).

On the Wikipedia dataset, augmentation at a rate of .50 performed best. Substitution on this dataset was optimal at the same rate. While the starting gender accuracy on the Wikipedia dataset was the lowest of the three datasets, both augmentation and substitution appeared to near-constantly (with the exception of substitution at .25) improve the results.

Corpus	Mod.	0	.25	.50	.75	1.0
Wikipedia	Aug.	12.55	14.39 (+1.84)	13.59 (+1.04)	12.49 (− .06)	10.72 (−1.83)
	Sub.	12.55	10.78 (−1.77)	11.18 (−1.37)	8.93 (−3.62)	6.51 (−6.04)
Global	Aug.	16.45	16.77 (+ .32)	16.80 (+ .35)	14.26 (−2.19)	12.65 (−3.80)
	Sub.	16.45	16.67 (+ .22)	13.05 (−3.40)	9.11 (−7.34)	8.74 (−7.71)
Tatoeba	Aug.	4.71	4.77 (+ .06)	5.72 (+1.01)	5.87 (+1.16)	5.17 (+ .46)
	Sub.	4.71	4.80 (+ .09)	3.51 (−2.00)	2.92 (−1.79)	2.39 (−2.32)

Table 6: BLEU scores for corpora modified at different rates.

In terms of trends across all three datasets, we observed that augmentation generally performed better than substitution. We found that augmentation performed at a rate of .50 seemed to mitigate the most gender bias. While augmentation at .25 and .75 also seemed to perform decently, using no augmentation and using full augmentation were on average lower-performing options. For substitution, the optimal rate also appeared to be .50, with .75 being the next-best rate.

3.3 Group 3: Effects on BLEU score

The third group of experiments examined the degree to which BLEU scores were affected by counterfactual data augmentation and substitution. Having established the usefulness of counterfactual data augmentation and substitution for reducing gender bias, it is also important to show that they do not inflict adverse harm to conventional evaluation metrics such as BLEU score. Using the English and Spanish test sets from the 2010 Shared Task on machine translation for European languages, BLEU scores were calculated by comparing each model's hypothesized English-to-Spanish translations with the gold targets. Prior to evaluation, the same preprocessing applied to the other datasets was applied to the test set. The results are shown in Table 6. The absolute changes in BLEU between the baseline and modified models are presented in parentheses next to the absolute BLEU scores.

We found that the use of counterfactual data augmentation at an appropriate rate, implied by these experiments to be between .25 and .50, did not adversely affect BLEU score.

In fact, the application of this technique at those rates consistently resulted in increased BLEU scores. On the other hand, counterfactual data substitution appeared to have a small negative effect on BLEU, with the impact worsening as the rate of substitution increased. However, as shown by the BLEU scores for Tatoeba and Global at .25 substitution, it seems possible for BLEU scores to increase using substitution as well. It might be the case that a small rate of substitution is optimal—a rate at or less than .25, depending on the dataset.

Overall, the experimental findings help to emphasize the relative safety of applying counterfactual data modification techniques. Granted an appropriate amount is used, even those focused on maximizing BLEU scores can simultaneously mitigate gender bias in their machine translation models. A summary of the best rates and methods for BLEU scores is provided in Table 7.

3.4 Summary of overall findings

Having performed these experiments, we found that the overall best method for gender bias mitigation varied depending on the dataset. For the Wikipedia corpus, the best mitigation was generally achieved by using data augmentation at a rate of .50. For the Global corpus, data substitution at a rate of .50 was best, only slightly beating augmentation at the same rate. And for the Tatoeba data, augmentation at a rate of 1.0 was best. These findings are encapsulated in Table 8.

Broadly, we found that augmentation was preferable over substitution, although both appeared to mitigate gender bias to some degree. Although augmentation generally performed better, using high rates of it (.75 or 1.0) generally seemed to decrease both gender accuracy and BLEU score. The one exception to this was with the small Tatoeba dataset. At lower rates, both gender accuracy and BLEU increased. Likewise, while substitution appeared less effective, its application at lower rates (such as .25) increased gender accuracy without substantial negative impact on BLEU score. With substitution too, we observed the

Corpus	Baseline BLEU	Best BLEU	Best BLEU Method
Wikipedia	12.55	14.39	25% Augmentation
Global	16.45	16.80	50% Augmentation
Tatoeba	4.71	5.87	75% Augmentation

Table 7: Comparison of baseline and best BLEU scores.

Corpus	Baseline Acc.	Best Acc.	Best Acc. Method
Wikipedia	36.0	44.4	50% Augmentation
Global	54.8	56.2	50% Substitution
Tatoeba	43.2	60.9	100% Augmentation

Table 8: Comparison of baseline and best gender accuracy scores.

same trend of both gender accuracy and BLEU decreasing as the rate increased. From these experiments, we conclude that the application of counterfactual data at a rate between .25 and .50 is generally recommendable. At these rates, a balance is struck between mitigating gender bias and maximizing BLEU score.

4 Discussion

Within this chapter, we reiterate the findings of the experiments and the implications that those findings may have on future work. We then proceed to an analysis of our work, identifying areas that could be improved upon as well as some potential threats to validity. We end with a note on future directions this work could proceed in.

4.1 General findings and implications

Based on the results of the experiments, we see the value of using counterfactual data augmentation and substitution techniques on datasets for training neural translation models. We find that a partial application of counterfactual data (around a rate of .50) is helpful for reducing gender biased translation output without causing harm to conventional metrics like BLEU score. We also find that augmentation disproportionately improves the gender accuracy of smaller datasets. These findings suggest that gender bias, and likely other forms of bias, can be successfully mitigated by addressing the data utilized for model training. Used in conjunction with post-training corrections and injections, it is likely that the biases in neural machine translation systems can be significantly reduced going forward.

4.2 Limitations

While this work aimed to avoid major threats to validity, there remain a few areas that must be acknowledged; while some can be immediately addressed, others can only be mentioned at the present time.

4.2.1 Limitations that can be addressed

The most immediate limitation of this work is also one that can be easily addressed: this work only evaluates one language pair in one direction. Both English and Spanish are incredibly common, high-resource languages. There is no guarantee that counterfactual data

augmentation and substitution techniques will work using other languages or in other language directions, assuming that suitably sized datasets are even available for those languages. Among many potential reasons, it is possible that gender bias may not be exhibited when translating between languages that both have gender agreement morphology—if both languages have a similar method for marking gender morphology (e.g., adding a suffix), then a translation model could likely map those features one to one. Then, assumptions about gender would not need to be made and gender bias would not be present. Similarly, there may be difficulties when it comes to translating languages with different writing conventions (e.g., right to left writing). Expanding the experiments to cover a more diverse array of languages would resolve these issues.

A secondary limitation lies within the datasets themselves—they all have different lengths, domains, and styles of writing. Taken from a translation website, the Tatoeba sentences are relatively short, syntactically simple, and sanitized. The Global sentences are from news titles, and likely disproportionately utilizes certain words and phrases in those titles. The Wikipedia sentences are from an encyclopedic website, which implies the use of more expository language. These facets are reflected in features such as their median word length and average sentence length, shown previously in Table 2. Importantly, none of these datasets necessarily includes or focuses on anaphora resolution or generic professions, and as such are intrinsically at odds with the style of the evaluation set—there is certainly domain mismatch. For example, Tatoeba does not have any Spanish translations with *chiropractor* or *acrobat*, Global doesn't have any news headlines involving a *mason*, etc. If the translation models had been trained on a set such as the GAP corpus by Webster et al. (2018), which is at least similar to the WinoMT set in its focus on ambiguous pronouns (e.g., *McFerran's horse farm was named Glen View. After* ***his*** *death in 1885, John E. Green acquired the farm.*), things might have been different. Similarly, the OpenSubtitles corpus (which the `pyspellcheck` tool relied upon) may have suffered from domain mismatch as well. There

may have been valid inflected words not present in the OpenSubtitles corpus, causing the tool to reject them. Utilizing transfer learning or experimenting with differently sized and domained datasets for training are potential solutions for these issues.

A third limitation was that there was no application of animacy detection in this work. Using animacy detection, it would have been possible to only reinflect animate words such as man, woman, officer, etc. Considering how in Spanish only animate nouns have gender pairs, this could have resulted in better counterfactuals than the ones produced by the comparatively greedy approach that was used—it is likely some undesired forms of nouns slipped past `spaCy` and `pyspellcheck`. Inanimate nouns typically cannot have their genders reinflected and the application of rules that are only appropriate for animate nouns can result in errors (e.g., *libro* changed to *libra*, where the meaning has changed from *book* to *pound*, the currency). Cutting down the number of possible candidates for reinflection by restricting them to animate nouns decreases the chances of such errors. Sentences where the nouns were reinflected had their determiners and adjectives changed to match, so applying animacy detection would eliminate many of the wrongfully created counterfactuals. Importantly, work done by Jahan et al. (2018) achieved classification of animacy at rates over 90%, confirming the relative safety of using animacy detection models.

4.2.2 Other limitations to acknowledge

While there are easily remediable limitations to this work, there are also limitations that can (at the present time) only be acknowledged. These are name issues in the counterfactual data, computational resource issues, random seed and ordering issues, morphological rule issues, and evaluation issues.

A specific issue that may have affected the counterfactuals is that of proper names. In both the English and Spanish data, no attempts were made to change any proper names present. While this is fine when it comes to words referring to specific places, it quickly

becomes an issue in the case of entity names. Living creatures are typically given names, which often carry certain sociocultural connotations involving gender. However, it is no easy task to swap one name out for another; outside of a few cases (e.g., Daniel and Daniela), one typically can't produce the opposite or counterpart of a name. In the context of this work's experiments, some counterfactuals may have had certain names erroneously attached to contrasting genders (e.g., *Gabriella* to *he*). If the counterfactuals only involved one language, it might be possible to address this by substituting stereotypically male and female names, as done in recent work by Hall Maudslay et al. (2019). But as there are two languages with names being mapped between them, to do the same would not be appropriate here.

Another issue was that of computational resources. Given more expansive resources with respect to hardware, it would likely have been possible to achieve more satisfactory results in terms of both gender accuracy and BLEU score. This would be due to the ability to parallelize across graphics or tensor processing units, which would in turn allow for deeper Transformer models, use of different floating point precisions, and for larger batch sizes to fit within their memory. All of these have been said to potentially improve the performance of Transformer models (Popel and Bojar, 2018). Moreover, those resources would greatly improve the speed at which experiments would be possible, allowing for quicker experimentation and development. However, resources were bottle-necked to a certain degree during this work's experiments due to external factors.

Computational randomization, in the form of random seeds, were also a potential issue for the present work. Utilized for probabilistic augmentation and substitution of counterfactual data, splitting into training and validation sets, and for training the Transformer models, there undoubtedly would have been a more optimal choice of random seed in each of the three situations. For example, Reimers and Gurevych (2017) find statistically significant differences in the performance of state-of-the-art named entity recognition models simply by changing the random seed. Schluter and Varab (2018) find natural language inference

models greatly affected by different permutations of the training data (i.e., one of the typical applications of random seeds). This means there was a potential validity issue in the gender bias findings stemming, to some degree, from the variation caused by the choice of random seed. However, given the sheer number of possible seeds, it is not possible to remedy this issue. While it would have been possible to improve the results by trying out an assortment of random seeds, there would always exist some random seed superior to the one currently being used. As such, they were held constant and otherwise ignored in this work.

Regarding the handwritten morphological rules used in this work, there are likely issues regarding exceptions or their over/under-application. Currently, however, there are few available resources that can deal with this. Databases such as UniMorph[1] are currently incomplete, even for languages as prevalent as Spanish. There is also a scarcity of datasets with oppositely gender-inflected sentences (e.g., a set with *La chica es buena* corresponding to *El chico es bueno*, and so forth), which limits the possibility of supervised machine learning for gender morphology. While there are tools that can perform morphological analysis, such as FreeLing[2] and `spaCy`, such resources typically do not extend to reinflection, with fewer still that can deal with morphological gender. There has been work done with respect to this problem, such as Zmigrod et al. (2019) who utilize Markov Random Fields (RMFs) for the task. While they manage to achieve promising results (accuracy scores ranging from 87 to 90 for form-level morphology), their RMF model is an unsupervised approach, is not publicly-available, and requires additional parameter specification and training. Considering the alternatives, handwritten rules seemed reasonably effective and computationally simple: sufficient for this research, despite their apparent flaws.

The last major threat to validity involves the form of evaluation. As discussed in the introduction, measuring bias is difficult at best. Utilizing a single gender accuracy score on a

[1] `https://unimorph.github.io/`
[2] `http://nlp.lsi.upc.edu/freeling/`

single, highly specific test set is not at all a comprehensive way to measure gender bias. This is especially true given how the test set has not been tested for its correlation with other bias tests that may exist. Unfortunately, the addition of BLEU scores as a metric is not particularly helpful either. At the current time there is little recourse to using gender accuracy on a WinoMT-style set, which is why this issue can only be acknowledged. Unfortunately, common methods for measuring gender bias in corpora, such as counting pronoun usage and examining word associations, don't apply particularly well when evaluating machine translation models. Translations in a target language are meant to mirror the meaning of the source language. In the case of the pronouns, unless there was some ambiguity (e.g., anaphora resolution in the WinoMT set) target pronouns would just correspond to the source pronouns, so macro-level statistics on the target pronouns would just mirror the source. Similarly, gendered associations wouldn't just spontaneously manifest in the target translations. Perhaps if additional test sets for gender bias were made available and their correlations measured with each other, we could produce more accurate and holistic evaluations of gender bias. Until then, these kinds of test sets must suffice as proxies for measuring gender bias.

4.3 Future work

Moving forward, there are many natural extensions to this work. Building off of some of the previously mentioned problems, some future work could include extensions to more language pairs or finding some way to improve the counterfactual data.

This work looked solely at the English-Spanish language pair in a few datasets, going from English to Spanish. However, Spanish is just one of many languages in the world which exhibits gender marking morphology and there are many more sources of language data available. A natural extension is to test the potential effectiveness of counterfactuals across many of the other world languages (many with more complex morphology or with fewer natural language processing resources), so as to determine whether the same gender

bias mitigation can be achieved. Similarly, additional training of models on different domain datasets of various sizes is another possible direction.

As previously mentioned, the counterfactual data produced during this work is by no means flawless or without error. Having native speakers available to produce a gold set of oppositely gender-inflected sentences would help a great deal with future research in this area. Crowd sourcing verification of produced counterfactual data could also be immensely helpful. On the computational side, the use of more advanced tools, like Markov Random Fields, to produce the sentences may be beneficial. Similarly, the development of open-source software for the task of gender reinflection is also an excellent option. Of course, using the above in some combination would be interesting to consider as well.

5 Conclusion

In this thesis, we investigated the efficacy of counterfactual data augmentation and substitution techniques, both individually and comparatively, for mitigating the implicit gender biases present in neural machine translation models. Furthermore, we tested data augmentation and substitution techniques at various probabilistic levels, finding an appropriate amount of data modification for reducing those biases while simultaneously minimizing potential computational and evaluatory costs. Lastly, we investigated the effects of counterfactual data augmentation and substitution on BLEU scores.

We found that overall, data augmentation proved to be the significantly more effective method for inserting counterfactual data. This was shown by the generally higher gender accuracy scores obtained by training on augmented data: in 10 of the 12 experiments spanning the three datasets, augmentation achieved higher gender accuracy scores than substitution. In those 10 cases, the augmentation gender accuracy scores were at minimum 2% better than substitution, and at maximum 19% better. Conversely, the cases where substitution beat augmentation were by margins of just 1.8% and .02%.

We also found the optimal rate of data augmentation for the English to Spanish language pair to be somewhere in the range of .25 to .50. The comparative ineffectiveness of both augmentation and substitution at the .75 and 1.0 rates suggested that modification at those levels was too much to be helpful. While the optimal value will likely vary depending on the dataset being used, the previously listed values are broadly supported across the corpora used in this work, providing them a basis to serve as a general recommendation.

Lastly, we found that the use of counterfactual data augmentation and substitution has a minor positive effect on BLEU score when utilized at lower rates. In the best case of the Wikipedia dataset augmented at a rate of .25, the BLEU score increased by 1.84 points. Moreover, across all three datasets, BLEU scores consistently increased when using counterfactual data augmentation at rates of .25 and .50. This provided evidence for the safety

of their use: application of counterfactual data augmentation at lower rates simultaneously mitigated gender bias while increasing BLEU scores.

Given the findings of this work's experiments, counterfactual data modification appears to be an effective technique for mitigating the gender biases present in neural machine translation models. We are optimistic that the application of these techniques will significantly improve the social equity and quality of machine translation output in the future.

A Appendix A

Table 9: Example WinoMT translations before and after augmentation.

Occupation and True Gender:	developer, female
English Source:	The developer argued with the designer because she did not like the design.
Spanish Output:	(El desarrollador → La desarrolladora) argumentó con la diseñadora porque no le gustó el diseño.
Occupation and True Gender:	manager, female
English Source:	The manager suggested the cashier get a new job because she was cruel.
Spanish Output:	(El gerente → La gerente) sugirió que la cajera obtienen una nueva trabaja porque era cruel.
Occupation and True Gender:	chief, female
English Source:	The chief encouraged the writer because she likes the book.
Spanish Output:	(El jefe → La jefe) animó a la escritora porque le gusta el libro.
Occupation and True Gender:	secretary, male
English Source:	The secretary asked someone to sign in so that he could update the guest log.
Spanish Output:	(La secretaria → El secretario) pidió a alguien que firmara para que pudiera actualizar el registro de invitados.
Occupation and True Gender:	farmer, female
English Source:	The farmer knows the editor because she likes the book.
Spanish Output:	(El granjero → La granjera) conoce al editor porque le gusta el libro.

References

Bhaskaran, J. and Bhallamudi, I. (2019). Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Chaloner, K. and Maldonado, A. (2019). Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2014). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *arXiv:1312.3005 [cs]*. arXiv: 1312.3005.

Ding, S., Renduchintala, A., and Duh, K. (2019). A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. (2020). Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv:2002.06305 [cs]*. arXiv: 2002.06305.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Hall Maudslay, R., Gonen, H., Cotterell, R., and Teufel, S. (2019). It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Jahan, L., Chauhan, G., and Finlayson, M. (2018). A New Approach to Animacy Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1–12, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *3rd International Conference for Learning Representations*, San Diego, California, USA.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2019). Gender Bias in Neural Natural Language Processing. *arXiv:1807.11714 [cs]*. arXiv: 1807.11714.

Moryossef, A., Aharoni, R., and Goldberg, Y. (2019). Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Park, J. H., Shin, J., and Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Popel, M. and Bojar, O. (2018). Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70. Number: 1.

Post, M. (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. Association for Computational Linguistics.

Reagle, J. and Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0):21. Number: 0.

Reimers, N. and Gurevych, I. (2017). Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (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. Association for Computational Linguistics.

Saunders, D. and Byrne, B. (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. Association for Computational Linguistics.

Schluter, N. and Varab, D. (2018). When data permutations are pathological: the case of neural natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4935–4939, Brussels, Belgium. Association for Computational Linguistics.

Seiffe, L., Marten, O., Mikhailov, M., Schmeier, S., Möller, S., and Roller, R. (2020). From Witch's Shot to Music Making Bones - Resources for Medical Laymen to Technical Language and Vice Versa. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6185–6192, Marseille, France. European Language Resources Association.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA. IEEE.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. page 5.

Vanmassenhove, E., Hardmeier, C., and Way, A. (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. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Wołk, K. and Marasek, K. (2014). Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs. *Procedia Technology*, 18:126–132.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.