Reinforced Sequence Training based Subjective Bias Correction

Karthic Madanagopal
Texas A&M University, Texas, USA
karthic11@tamu.edu

James Caverlee
Texas A&M University, Texas, USA
caverlee@tamu.edu

Abstract

Subjective bias is ubiquitous on news sites, social media, and knowledge resources like Wikipedia. Many existing methods for subjective bias correction have typically focused on making one-word edits and have been trained over a single (often, noisy) domain. In contrast, we propose a novel reinforced sequence training approach for robust subjective bias correction. Three of the unique characteristics of the approach are: (i) it balances bias neutralization with fluency and semantics preservation through reinforcement learning, to broaden the scope to bias beyond a single word; (ii) it is cross-trained over multiple sources of bias to be more robust to new styles of biased writing that are not seen in the training data for a single domain; and (iii) it is used to fine-tune a large pre-trained transformer model to yield state-of-the-art performance in bias text correction task. Extensive experiments show that the proposed approach results in significant improvements in subjective bias correction versus alternatives.

1 Introduction

Objective writing is essential for many important communication venues like news, encyclopedias, scientific publications, and more. And yet, bias is seemingly ubiquitous whether due to malice or unintentional habits of the writer. This subjective writing not only expresses the writer's preferences and personal interpretations, but also can influence the reader's viewpoints on the topic (Greenstein and Zhu, 2014; Beukeboom and Burgers, 2017). Hence, much like modern spelling and grammar checkers, there is a need for effective methods to detect and neutralize biased language.

The goal of this *subjective bias correction* is to rewrite a source sentence s into a neutral sentence t that is clear, objective, and stereotype-free. Furthermore, the rewritten sentence t should preserve the original meaning of the source text s. Enabling such *subjective bias correction* is difficult. Many examples of bias are challenging to detect due to their subtle nature, or through sentence framing (rather than through subjective words), or through convoluted writing intend to obscure the truth.

Examples of Biased and Neutral Statements
Acetaminophen is sold over the counter as a pain medication.
Acetaminophen is the **most dangerous** over-the-counter pain medication.
A 2013 episode of This American Life presented a number of studies that verified that acetaminophen has killed more people than any other over-the-counter pain medication.
In the This American Life episode on acetaminophen, one segment described the **tragic** death of a five-month-old baby and thus **should** convince listeners that the Federal Drug Administration (FDA) must take immediate action.

Table 1: Examples of multi-word and multi-occurrence biased statements (2 and 4), and corresponding neutral statements (1 and 3). The words "most dangerous" in the second statement and the words "tragic" and "should" in the fourth statement express the author's opinion or belief. The third statement is neutral since it describes the subjective view expressed on a radio show. (NROC, 2022)

Encouragingly, previous research has begun to make progress in identifying prejudiced or emotive language used in factual statements (Recasens et al., 2013; Bhosale et al., 2013; Misra and Basak, 2016; Hube and Fetahu, 2018; Zhong et al., 2021; Madanagopal and Caverlee, 2022) and a few studies have begun to investigate subjective bias correction (Pryzant et al., 2019; Liu et al., 2021; Zhong et al., 2021). These approaches, however, typically face a number of key challenges:

Noisy and Limited Training Data. Many approaches rely on Seq2Seq models (encoder-decoder architecture) using a Wikipedia-derived Neutrality Corpus (WNC) (Recasens et al., 2013) based on edits to Wikipedia that have been flagged with a special "neutral point of view" tag. However, we find that over 5% of revisions in the WNC (e.g., as in (Pryzant et al., 2019)) are not related

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2585–2598

to bias mitigation, but rather content corrections, grammatical corrections, or typographical corrections. And more than 15% of the human-corrected neutral sentences with multi-occurrence bias still contain subjective bias. Training a bias correction approach with such noisy data can lead to models that focus on single word and single occurrence bias (missing examples like in Table 1), and result in inconsistent performance.

Training-Testing Mismatch. Second, existing approaches are primarily trained on maximizing the likelihood of each token of the target sequence (Pryzant et al., 2019; Zhong et al., 2021) by relying on the input sequence and previous ground-truth token, but testing is done on the entire input and output sequence. The training is conducted using token-level objective functions and tested using sentence-level evaluation metrics, such as BLEU. Further, although fluency and content preservation are used for evaluation, such objectives are not contained in training objectives. This training-testing criteria mismatch can lead to poor performance at inference time.

Lack of Robustness. Third, models trained on one domain tend to perform poorly on other domains, limiting their adoption. Recent studies have shown further fine-tuning of pre-trained models can improve domain generalization (Sun et al., 2020; Wang et al., 2020), but such methods generally need large volumes of training data from the target domain. Is it possible to generalize subjective bias correction models trained in one domain to multiple new domains even in the absence of rich training data?

To address these challenges, we propose a novel *reinforced sequence training* approach toward improving subjective bias correction. The overall approach first pre-trains a bias correction model using noisy training data (much like in previous approaches) but then fine-tunes the pre-trained model through reinforcement learning with a carefully designed cross-domain bias critic. This cross-domain bias critic aligns training and testing time objectives by giving equal importance to the quality of text generation and semantic content preservation through self-supervised reward-driven learning. That is, it balances bias neutralization with fluency and semantics preservation through reinforcement learning, to broaden the scope to bias beyond a single word or single occurrence within a sentence. Further, the proposed approach is cross-trained over multiple sources of subjective bias to be more robust to new styles of biased writing that are not seen in the training data for a single domain. And instead of using word-level cross-entropy loss during training, we directly optimize sentence-level task-based metrics through the policy gradient to gain significant improvement in performance.

In summary:

- We propose a reinforcement learning framework for improving subjective bias correction.
- We improve the generalizability of the model across multiple domains by adopting a cross-domain bias classifier-based reward, without the need for parallel data for domain adaptation.
- We align training and testing time objectives by using a novel reward-driven learning framework that goes beyond the traditional BLEU rewards, and guides the model to generate diverse bias-free sentences which are fluent and grammatically correct.
- We empirically demonstrate how the proposed reinforcement learning-based approach can fine-tune existing large pre-trained models like BART to perform efficient bias correction while preserving content semantics.

Through extensive experiments, we show that the proposed approach results in significant improvements in subjective bias correction versus alternatives. Furthermore, evaluations on semantic similarity and fluency benchmarks show that this bias correction approach effectively removes subjective bias while maintaining semantic information in neutralized text.

2 Related Work

Detecting Language Bias. Many studies have focused on extracting bias lexicons or opinion words (Riloff and Wiebe, 2003; Liu et al., 2005; Wiebe et al., 2005; Appling, 2017). (Recasens et al., 2013) reduced the lexical ambiguity in detecting biased statements by using a combination of language models and bias lexicons. (Misra and Basak, 2016; Hube and Fetahu, 2018) developed deep learning models to detect bias in political speeches and Wikipedia. (Madanagopal and Caverlee, 2022) combined a cross-domain dataset and a contextualized language model to train a deep model for detecting biased language.

Linguistic Bias Correction. The majority of research to address bias correction in text has focused

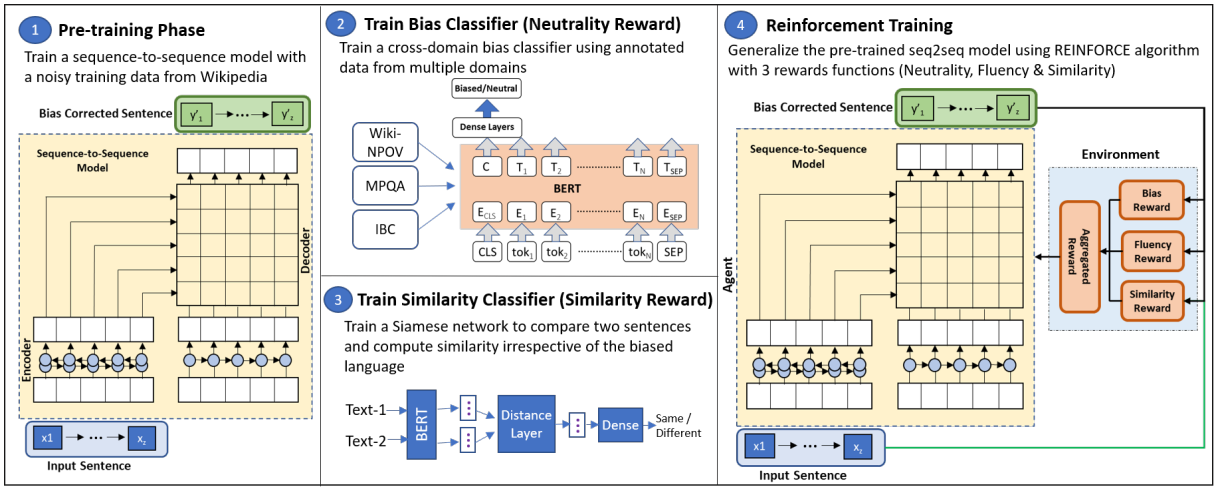

Figure 1: Illustration of the proposed Reinforced Sequence Training Approach for subjective bias correction. Task 1 is pre-training the agent (bias-correction-model) with noisy WNC data, Task 2&3 are building classifiers that will be used to compute rewards, and Task 4 is reinforced sequence training with 3 reward functions.

on demographic bias like gender bias (Manzini et al., 2019; Zhao et al., 2017, 2019; Bordia and Bowman, 2019; Wang et al., 2018). (Pryzant et al., 2019) were the first to address generic linguistic bias correction and later (Liu et al., 2021) worked on depolarizing political text. In both approaches, text segments (words or sentences) that are subjective or polarizing are first identified and then replaced with those that are semantically similar but less subjective. (Liu et al., 2021) studied generative adversarial networks to train bias correction models using non-parallel corpus. Pryzant proposed two joint embedding models to neutralize biased statements, but their research is restricted to addressing subjective bias that originated from single-word edits (Pryzant et al., 2019). Our proposed method extends from single-word to multi-word correction by generating fluent bias-free sentences.

Attribute Style Transfer. Subjective bias correction can be framed as a specialized attribute transfer task which aims at reconstructing an input text so that a linguistic attribute of interest is transferred to a desired value, such as text sentiment transfer (Jin et al., 2022; Lample et al., 2018). Most attribute transfer methods uses auto-encoders that rely on large parallel corpus (Rao and Tetreault, 2018; Briakou et al., 2021; Madaan et al., 2020; Prabhumoye et al., 2018; Pryzant et al., 2019; Jin et al., 2022). (Huang et al., 2019) used a dictionary based method to replace words in a sentence to create diverse paraphrases. (Lai et al., 2021) developed an efficient method to fine-tune a large pretrained model for text-attribute transfer on low-resource domains. (Chen et al., 2018) used cross-aligned auto-encoder trained on opposite ideology news data to generate flipped titles. Our proposed method is inspired from various attribute style transfer methods, but addresses the problem of using noisy parallel data for supervised style transfer for better generalization using reinforcement learning methods.

3 Reinforced Bias Corrector

In this section, we introduce the design of the proposed subjective bias correction approach (illustrated in Figure 1). Given a biased statement x that describes a facts f using a set of words $(x_1, x_1, ..., x_N)$, where N is the length of x. The aim of a bias correction system is to generate a statement $y = (y_1, y_2, ..., y_M)$, where M is the length of y. The generated statement y is not only expected to have a neutral tone, but also preserve the original semantic content expressed in statement x. Additionally, y is also expected to present the fact F in a grammatically correct and fluent language that is easy for a reader to comprehend.

3.1 RL and Bias Correction

Given these goals, a natural approach to perform subjective bias correction is to train an encoder-decoder with attention network using a parallel corpora similar to many natural language generation tasks (Bahdanau et al., 2014; Zhao et al., 2019; Chollampatt and Ng, 2018; Pryzant et al., 2019; Konstas et al., 2017; Li et al., 2018; Gupta et al., 2018; Vaswani et al., 2017). The encoder converts the input sentence into a fixed-length vector that will be used by the decoder to generate the output sequence, word by word. The entire network will

be trained with the parallel data (x, y) to minimize the cross-entropy loss (CE) that is given by:

$$L_{CE} = \sum_{t=1}^{T} log\pi_\theta(y_t|y_{t-1}, s_t, c_{t-1}, x) \quad (1)$$

Since minimizing L_{CE} makes the model generate text close to the human-reference edits, the training data needs to be of high-quality. Since the only large parallel corpus available for bias correction is noisy, such likelihood optimized models have limited performance (Pryzant et al., 2019; Zhong et al., 2021). Additionally, these models do not produce the best results for metrics during evaluation because the training is not optimized to generate text that addresses evaluation metrics (such as BLEU or Content Similarity). By aligning the training and testing objective, we can improve the performance of bias correction in a source domain (Wikipedia), though this does not guarantee generalizability over other domains.

Hence, we propose to improve the performance and generalizability of bias correction models by fine-tuning the Seq2Seq model using Reinforcement Learning (RL). In this method, instead of maximizing the probability that the predicted word is close to the ground truth, the parameters of the agent are optimized to maximize long-term rewards. We can formulate the RL engine for bias correction as composed of an agent (A), action (a), policy (pi), and reward (R). The agent is our bias correction model with parameters θ that observes the current state (encoder's output) at time t and takes an action a (predict the next word $\hat{y}_t$) by using a policy (π). The reward (R) is a feedback returned to the bias correction model from the system by evaluating the quality of generated text.

3.2 REINFORCE Training

In reinforced bias correction task, the agent's behavior is controlled primarily through the neutrality score that measures the degree of subjective bias in the generated text $\hat{y}$. So, the objective of maximizing the expected scalar reward $R : \hat{y} \rightarrow [0, 1]$ is given by:

$$J(\theta) = E_{\pi_\theta}(\hat{y}/x)[R(\hat{y})] \quad (2)$$

Since the reward is the discrete function of the model's output, the RL objective is non-differentiable with respect to the model parameter θ,), which makes it difficult to back-propagate the error signals from the critic to the generator. This issue can be addressed through policy gradient or Q-learning based methods. In our initial experiments, we considered several RL methods, including Q-learning, but these models did not converge even after training for a long time (more than 48 hours). This is mainly due to the extremely large action space $O(W^T)$, where W is the number of words in the vocabulary (10^4) and T is the sentence length.

In contrast, the REINFORCE-based method showed good performance and converged faster than other methods. Hence, we focus our discussion here on REINFORCE as our primary policy gradient method. In REINFORCE, the expected reward is approximated using a sampling method and the model is trained using stochastic gradient ascent (Williams, 1992), which can be formulated as:

$$\nabla_\theta J(\theta) = E_{\pi_\theta(\hat{y}/x))}[R(\hat{y})\nabla_\theta log\pi_\theta(\hat{y}/x)] \quad (3)$$

where, π_θ is a policy that generates a probability of picking a word as output. Even though RL training is promising, efficiently applying it for real world problems remains a challenge due to extremely large sequence space (50000^{50} for a sentence with length of 50 and vocabulary of 50000) (Mnih et al., 2015). To achieve the RL objective, a sequence model is pre-trained first and then fine-tuned using REINFORCE algorithm using different rewards. The REINFORCE method optimizes the best policy directly by modifying the parameters of the model based on the observed rewards. Since it directly optimizes the return, it tends to be more stable in converging to a good behavior

4 RL Rewards and Policy Extensions

In our subjective bias correction task, the key testing objectives are: (i) reduce subjective bias; (ii) preserve content expressed in the source text; and (iii) the generated text needs to be grammatically correct fluent. Based on the above objectives, the designed reward function $R(x, \hat{y})$ consist of three rewards:

$$R(x, \hat{y}) = \alpha * R_N(\hat{y}) + \beta * R_S(x, \hat{y}) + \gamma * R_F(\hat{y}) \quad (4)$$

where, $R_N(\hat{y})$ is the neutrality reward for the bias corrected text $\hat{y}$, $R_S(x, \hat{y})$ is the semantic similarity reward computed between the input text x and output text $\hat{y}$, and $R_F(\hat{y})$ is the fluency reward for the output text $\hat{y}$. Here α, β, $\gamma > 0$ and represents weights for the respective rewards.

Neutrality Reward. The central objective of this first reward is to remove any subjective bias that is contained in the input sentence x. The neutrality reward $R_N(\hat{y})$ computes the presence or absence of biased tone in the output sentence $\hat{y}$. It is not a relative score that is computed between the input text x and output text $\hat{y}$, rather an absolute score that is computed purely on the output text $\hat{y}$. Because of which, there is a possibility that the model might remove a chunk of text that is biased and output a partial sentence that is neutral. But the combination of neutrality reward R_N and semantic similarity reward R_S ensures the corrected text carries all the information expressed in an unbiased or neutral tone. To compute the neutrality reward, a RoBERTa based binary classifier is used that takes the sentence as input and produces a probability score that the input text is biased or not (Madanagopal and Caverlee, 2022). More details about this classifier is available on our previous work on detecting linguistic bias (Madanagopal and Caverlee, 2022).

Semantics Similarity Reward. The semantic similarity reward $R_S(x, \hat{y})$ is used to ensure that all the information that is expressed in the input sentence is retained in the output sentence $\hat{y}$. Standard semantic similarity metrics like BERTScore cannot be used for this task, because the tone or slant change will affect the similarity score significantly. To create a subjectivity-free similarity reward, we created a Siamese network based on BERT and fine tuned using Wikipedia-derived parallel corpus (Mueller and Thyagarajan, 2016).

Fluency Reward. The fluency reward encourages the model to generate a sentence $\hat{y}$ that is grammatically correct or natural sounding. There are various methods to compute the grammatical correctness of a sentence such as Perplexity score (PPL) (Meister and Cotterell, 2021). Large pretrained models are used to compute the PPL scores, where grammatically incorrect sentences will have high PPL and grammatically correct sentence will yield a lower PPL score (Meister and Cotterell, 2021). But, the fluency reward needs to be a normalized value between 0 and 1. We used the Corpus of Linguistic Acceptability (CoLA) that contains 10,657 English sentences that are labelled for grammatical and syntactic correctness (Warstadt et al., 2018). We then built a binary classifier using RoBERTa and used the classifier result as the fluency reward for measuring the correctness of the generated text $\hat{Y}$.

4.1 Cross-domain linguistic bias critic

In classic REINFORCE learning scenarios, to minimize the distinction between human reference edit and the machine correction, the bias critic is trained with the same training data used for agent's pretraining (Wu et al., 2018). Since some of the ground-truth data still contains bias, using similar approach can adversely affect the agent's behaviour. Also, our objective is to learn beyond what is available in the training corpus so that it can efficiently handle new style of writing and all forms of subjective bias outside of Wikipedia domain. Hence, we investigate in this paper, the potential of using a cross-domain bias classifier as the bias critic. The cross-domain bias classifier is trained by leveraging annotated datasets from other domains that are rich in subjectivity and apply recent deep transformer models like BERT in order to more robustly model factual statements. The cross-domain bias classifier is evaluated to have an accuracy of 89% and coincide with human judgement for bias detection. There are two advantages to using a cross-domain bias classifier: First, the cross-domain dataset used for training the binary classifier does not need a parallel dataset. Second, the cross-domain aspect lets the classifier learn new subjective writing styles from other subjectivity rich domains and assign rewards that are of high quality, which in turn makes the reinforcement more efficient.

Generally, the most effective method to approximate the policy gradient is either multinomial sampling of the softmax-normalized outputs or beam search. Both the objectives can be trained either sequentially or simultaneously. Our proposed method uses a sequential training, where the agent is first trained with supervised methods and then fine-tuned with reinforced methods.

4.2 Delayed and Sparse Reward

Typically, in a RL setting, rewards are assigned as soon as the action is taken by the agent. But in our method, the reward is observed only at the end of generating a complete sentence. Which means one reward for multiple actions taken by the agent. Due to this sparsity of reward, the agent receives same reward for all steps taken to generate a complete sentence. This makes the RL training inefficient because the agent has no clue which action led to bad results. To address this issues, Reward shaping strategy (Ng et al., 1999) is used to compute intermediate rewards as suggested by

(Bahdanau et al., 2014):

$$R(y_t) = R(y_{1:t}) - R(y_{1:t-1}) \quad (5)$$

For simple rewards like BLEU, it is easy to compute the intermediate rewards with partial sentences. But the reward functions used in the proposed approach only works if the complete sentence is available. We used the modified reward shaping method used by Hongyu et al. (Gong et al., 2019), where the agent is made to generate the entire text at each step and the generated text is used to compute the reward.

5 Experimental Setup

This sections provide an overview of the datasets, the baseline bias correction models, the evaluation metrics, and the details of their implementation.

5.1 Datasets

For pre-training the agent (Seq2Seq or Transformer), we harvested Wikipedia-derived Neutrality Corpus (WNC) (204k parallel sentences) following the instructions of (Recasens et al., 2013), but included multi-word edits (See the Appendix). For training a cross-domain bias classifier, we used the dataset assembled in a prior study (Madanagopal and Caverlee, 2022). For evaluation, we used WIKIBIAS (2,117 biased and 2,911 neutral), a high-quality dataset that is manually curated by (Zhong et al., 2021).

5.2 Baselines

We consider three baselines models:

Delete Biased Word(s) uses a bias tagger to identify subjective words in a sentence and removes them. More details on the bias tagger is available in the Appendix.

Join Embedding Model uses a denoising autoencoder and a token-weighted loss function to automatically neutralize subjective bias in text (Pryzant et al., 2019). We retrained their best performing CONCURRENT model.

OpenNMT Transformer is a widely used medium transformer model implemented in the OpenNMT-py library that contains 12 heads, 768 d_{model} size and 3072 d_{ff} size (Klein et al., 2017).

5.3 Models

We consider two classes of models for the proposed *reinforced sequence training*: (i) a vanilla Seq2Seq model; and (ii) a pre-trained transformer model.

Reinforced Seq2Seq Model (RL-Seq2Seq): The first model is an attention-based Seq2Seq model similar to (Bahdanau et al., 2014). It is pretrained with the WNC corpus and the pre-trained model is used as an agent in the proposed reinforced sequence training approach with three reward functions. The following three variations were trained and evaluated:

RL-Seq2Seq$_B$: Reinforced sequence-to-sequence model with base bias reward (WNC).

RL-Seq2Seq$_{CDB}$: Reinforced sequence-to-sequence model with cross-domain bias reward.

RL-Seq2Seq$_{CDB+FL+SIM}$: Reinforced sequence-to-sequence model with cross-domain bias, fluency and semantic content preservation reward.

Reinforced Transformer Model (RL-Trans): The second model is the BART pre-trained transformer model (Lewis et al., 2019), which performs well on paraphrasing tasks with good fluency and content preservation (Lewis et al., 2019; Lai et al., 2021). We downloaded a pre-trained BART model (base) and further pre-trained on the noisy WNC corpus. The further pre-trained BART model is used as one of the baseline models, and further fine-tuned, resulting in three variations akin to the Seq2Seq ones: (RL-Trans$_B$, RL-Trans$_{CDB}$, RL-Trans$_{CDB+FL+SIM}$).

5.4 Training

Training is in four steps: (1) pre-train the Seq2Seq model using the noisy Wikipedia NPOV data with an MLE objective; (2) train a cross-domain bias classifier using a large pre-trained language model like RoBERTa; (3) train a siamese network to compute semantic similarity between sentences without subjective tone; and (4) fine-tune the pre-trained Seq2Seq model using reinforcement learning with three reward components. A combination of training data sampling and new domain data (non-parallel) is used for reinforced fine-tuning. For fluency reward, we used the CoLA dataset (Warstadt et al., 2018) and trained a RoBERTa model with an accuracy of 92%. The hyperparameters used for pre-training the Seq2Seq model and fine-tuning of the reinforcement model are presented in the Appendix. For semantic similarity evaluation, the BERTScore model is downloaded from HuggingFace.[1] All code is implemented using PyTorch and

[1] https://huggingface.co/spaces/evaluate-metric/bertscore

Model	Neutrality↑	BLEU↑	BLEURT↑	PPL↓	BERT-Score↑	Bias↓	Fluency↑	Content↓
Source Copy	9.47	90.83	32.47	33.47	100.00	-	-	-
Biased word removal	38.19*	92.03*	36.24*	49.21*	85.24*	-0.179*	-0.127*	1.28*
Join Embedding	45.80*	93.94*	52.92*	31.87*	87.58*	-0.213*	0.011*	1.23*
OpenNMT Transformer	49.47*	87.82*	61.24*	34.57*	88.45	-0.205*	0.034*	1.14*
Seq2Seq	45.14*	93.23*	54.27*	39.47*	86.14*	-0.217*	0.008*	1.29*
$RL - Seq2Seq_{CDB+FL+SIM}$	63.72*	**94.52**	61.27*	26.62*	92.97*	-0.475*	0.013*	**1.06***
Transformer	52.51*	91.17*	60.84	22.78*	83.23*	-0.418*	**0.169***	1.12*
$RL - Trans_{CDB+FL+SIM}$	**66.48***	89.68*	**66.66***	**23.97**	**94.65***	**-0.492***	0.164*	1.19*
Target Copy	92.14	100	100	31.68	98.44	-	-	-

Table 2: Bias neutralization performance. RL-Seq2Seq and RL-Trans models are trained using our reinforced sequence training approach. For quantitative metrics, rows with asterisks (*) are significantly different than the preceding row. ↑ / ↓ means higher/lower score is preferred for the corresponding metric.

Model	Neutrality↑	BLEU↑	BLEURT↑	PPL↓	BERT-Score↑
$RL - Seq2Seq_{B}$	59.08	94.08	59.24	32.58	87.74
$RL - Seq2Seq_{CDB}$	63.65	94.23	61.24	31.24	85.23
$RL - Seq2Seq_{CDB+FL+SIM}$	63.72	**94.52**	61.27	26.62	92.97
$RL - Trans_{B}$	61.16	89.37	66.51	22.89	84.50
$RL - Trans_{CDB}$	**66.51**	89.11	65.84	22.57	82.21
$RL - Trans_{CDB+FL+SIM}$	66.48	89.68	**66.66**	**23.97**	**94.65**

Table 3: Performance comparison Reinforced sequence models on different reward schemes. ↑ / ↓ means higher/lower score is preferred for the corresponding metric.

Model	News Articles			Academics			Conservapedia		
	Bias↓	Fluency↑	Content↓	Bias↓	Fluency↑	Content↓	Bias↓	Fluency↑	Content↓
Delete Biased	-0.39	-0.77	1.6	-0.41	-0.13	1.53	0.25	-0.35	1.22
Join Embedding	-0.015	0.11	1.58	-0.93	-0.73	2.12	-0.31	-0.37	1.25
OpenNMT Transformer	-0.07	0.05	1.64	-0.61	0.48	1.65	0.08	-0.38	1.69
$RL - Seq2Seq_{CDB+FL+SIM}$	**-1.38**	0.59	**1.14**	-0.92	0.41	1.45	-0.42	0.43	**1.09**
$RL - Trans_{CDB+FL+SIM}$	-0.96	**0.72**	1.33	**-1.01**	**0.55**	**1.11**	**-0.63**	**0.56**	1.22

Table 4: Human evaluation of subjective bias correction performance across 3 different domains.

trained on a Google Cloud Platform with NVIDIA Tesla P100 GPU.

5.5 Evaluation Metrics

The following set of automated metrics are used to evaluate the quality of bias correction models:

Neutrality: Similar to previous work (Luo et al., 2019; He et al., 2020), the neutrality score is computed using a binary classifier that is pre-trained to evaluate subjective bias across various domains; on human reference dataset it has an accuracy of 89% (Madanagopal and Caverlee, 2022). The text generated by each model is sent to the binary classifier and the probability score that it belongs to the neutral category is used as the neutrality score.

BLEU: A commonly used metric that measures the similarity of machine corrected text and human reference correction through n-gram precision counting (Papineni et al., 2002).

BLEURT: A trained metric that uses transformer models to evaluate the quality of natural language generation models (Sellam et al., 2020).

BERTScore: Uses contextual language models such as BERT and computes semantic distance between candidate and reference sentences (Zhang* et al., 2020).

PPL: To evaluate the grammatical correctness and fluency of the machine generated text, we computed the perplexity score (PPL) using the large pre-trained language model GPT-2. The perplexity score PPL is computed directly on the generated text with no reference text.

6 Experimental Results

We present the results of the experiments through both automated and human judgement evaluation.

6.1 Automated Evaluation

Table 2 shows our proposed model's performance in comparison with the selected baseline models. Among the baseline models, the OpenNMT transformer model performed better with a neutrality score of 49.47 and BERTScore of 88.45, and the Join Embedding model had the best BLEU be-

cause of its copy network nature. But our two proposed reinforced sequence models performed significantly better than the baseline models on all metrics, especially the neutrality score (18% increase). In terms of BLEU score, the RL-Seq2Seq model had the best score of 94.52. Since the RL-Trans generated more diverse text compared to the human-edits, it's BLEU score was relatively low. But its BERTScore being high confirmed the RL-Trans model's ability to retain the information conveyed in the source sentence to the target sentence with high neutrality score. Similarly the perplexity score for the RL-Trans model indicates high fluency. This might be due to the fact the pre-trained BART model is trained on large diverse corpus. Overall the RL-Trans had the best performance on all automated metrics except BLEU.

6.2 Evaluation of Reward Function

To understand the effect of each reward on the performance of the bias correction model, we conducted an ablation study as shown in Table 3. Base bias rewards are based on a bias classifier trained solely on Wikipedia dataset, while cross-domain bias rewards are trained on a dataset containing biased statements from various domains like political speech and product reviews. In terms of the neutrality score, we observed significant improvement in the bias correction performance by the use of cross-domain bias reward. It could be because pre-training data is noisy, so the output still contains biased statements that the cross-domain bias classifier is able to detect more subtle forms of bias and multiple occurrences of bias. In terms of BLEU, the RL-Seq2Seq model with all rewards performed much better. The RL-Trans model's BLEU score was low which is due to the diversity in the text generated relative to the source text. This shows BLEU is not a good metric for comparing bias correction accuracy. So we use BLEURT which accounts for the n-gram accuracy with embeddings. The reinforced transformer model with all rewards performed the best. One interesting observation is that the improvement in RL-Trans using fluency and content similarity reward is small relative to RL-Seq2Seq. The large pre-trained BART model already performs well on fluency and content preservation, so adding rewards has little impact. All in all, the rewards were more effective at achieving the testing objective and also produced high-quality text.

6.3 Multi-Occurrence Bias Evaluation

We performed a separate study to investigate the performance of reinforced bias correction in addressing multiple instances of bias within a single sentence. A set of 331 sentences were selected from the WIKIBIAS corpus that contain more than one instance of subjective bias. The output of each model is then analyzed with the help of a bias tagger to understand how much of the bias still exists or if new bias is introduced. RL-Seq2Seq model trained with all 3 rewards had the best performance (16%) of addressing multiple occurrence of bias in a single sentence (See Table 5).

Model	Neutrality	% biased
Biased Word Removal	34.06	30.85
Join Embedding	50.56	37.32
OpenNMT	58.85	41.62
RL-Seq2Seq	67.33	**16.17**
RL-Trans	**67.15**	18.24

Table 5: Performance evaluation of addressing multi-occurrence bias. % biased represents the percentage of sentences that contains at least one instance of biased chunk after bias correction (lower the better).

6.4 Human Judgement

With a sample of 100 sentences collected for each of the experimented models, we conducted a human evaluation study with 10 judges to determine if the bias correction models generate human-like sentences. See the Appendix for more detail. A majority of the human judgement results correlated with the automatic evaluation. The Reinforced-Transformer model was preferred by a majority of the judges for its neutral tone and language fluency. The Reinforced Seq2Seq model was shown to preserve the same information as the source text. Since the pre-trained transformer model generates text that is diverse from the original input sentence, it might have read slightly different for a human. But the BERTScore showed the Reinforced-Transformer model to have better performance in content preservation.

The Kappa statistic was used to compute the inter-annotator agreement for human evaluation task. The average kappa values for the individual aspects are (i) Bias neutralization: 0.72; (ii) content presentation: 0.75; and (iii) fluency: 0.85. Due to the complex nature of subjective bias detection, the inter-annotator agreement for that task is the lowest (0.72). On the other hand, Since most of the annotators are well-versed in assessing the gram-

matical correctness of a sentence, the kappa value for fluency was high (0.85).

6.5 Cross-domain Model Performance

Finally, we explore the generalizability of the proposed bias correction model by performing inferences on three datasets that are manually curated outside the training domain (Wikipedia): (1) News Headlines (Media Cloud[1]); (2) Academic (National Geographic[2]); and (3) Conservapedia[3]. Thirty factual sentences were collected from all three selected domains, processed through the bias correction models and evaluated through human judgement on the basis of neutrality, fluency and content preservation. Overall, both the proposed models performed very well in all three domains. In terms of neutralization effort, the RL-Seq2Seq model performed the best on the News domain (See Table 4). This is because most of the bias in news are related to framing bias (perspective-specific words). But in the Academic domain and Conservapedia domain, the RL-Trans performed better. Relatively, the transformer model performs better on addressing both framing as well as epistemological bias.

7 Conclusion and Future Work

We proposed a hybrid method to improve the performance of Seq2Seq and transformer-based bias correction models by incorporating semi-supervised training strategy that uses a supervised pre-training using noisy data and reinforced fine-tuning using high-quality cross-domain data. The proposed training method is able to successfully alleviate the exposure bias in MLE optimized sequence models and address the in-stability issue in reinforcement learning methods. It also shows that by carefully designing the reward function with respect to the testing objective, high quality results can be obtained. Specifically, our method was able to generate text with high neutrality. The text generated by our method is more fluent and retains more semantic information relative to previous methods. In this study, we explored only a limited range of rewards using simple aggregations, but a greater range of rewards and scaling could be explored for better domain adaptation.

[1]https://mediacloud.org/

[2]https://education.nationalgeographic.org/

[3]https://www.conservapedia.com

Limitations

Although the proposed Reinforced Seq2Seq framework presents an intriguing framework for automatically correcting subjective bias, we also find that – like previous research (Williams, 1992) on similar reinforcement-style learning regimens – it has the following limitations: (i) after each policy run, the trajectories are discarded by the update process, making it inefficient. In some cases, the collected trajectory may not accurately represent the policy, so the gradient estimate becomes uncertain. Further, whether a trajectory reinforces good or bad actions depends entirely on the final output, so the reward assignment can be unclear.

In this study, we examine a specific form of subjective bias that manifests at the sentence level. However, taking into account the context (paragraph) of the statement could change the viewpoint. Although this study evaluates the Reinforced Seq2Seq framework on a real-world subjective bias correction task, further testing is needed in order to include more challenging bias types, as well as other target model architectures.

The work presented here offers a promising approach for improving bias correction models using reinforcement learning methods, a field of high impact but under-explored to date. We hope that evaluating the models across different domains will inspire further work on building robust, nuanced, and fair bias correction models.

References

CJ Hutto Dennis Folds Scott Appling. 2017. Computationally detecting and quantifying the degree of bias in sentence-level text of news stories.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Camiel J Beukeboom and Christian Burgers. 2017. Linguistic bias. In *Oxford research encyclopedia of communication*.

Shruti Bhosale, Heath Vinicombe, and Raymond Mooney. 2013. Detecting promotional content in wikipedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1851–1857.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! xformal: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International conference on natural language generation*, pages 79–88.

Shamil Chollampatt and Hwee Tou Ng. 2018. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. *arXiv preprint arXiv:1903.10671*.

Shane Greenstein and Feng Zhu. 2014. *Do Experts Or Collective Intelligence Write with More Bias?: Evidence from Encyclopædia Britannica and Wikipedia*. Harvard Business School.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

Shaohan Huang, Yu Wu, Furu Wei, and Zhongzhi Luan. 2019. Dictionary-guided editing networks for paraphrase generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6546–6553.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion Proceedings of the The Web Conference 2018*, pages 1779–1786.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural amr: Sequence-to-sequence models for parsing and generation. *arXiv preprint arXiv:1704.08381*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you bart! rewarding pre-trained models improves formality style transfer. *arXiv preprint arXiv:2105.06947*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A transformer-based framework for neutralizing and reversing the political polarity of news articles. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.

Karthic Madanagopal and James Caverlee. 2022. Improving linguistic bias detection in wikipedia using cross-domain adaptive pre-training. In *Companion Proceedings of the Web Conference 2022*, pages 1301–1309.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.

Arkajyoti Misra and Sanjib Basak. 2016. Political bias analysis.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287.

NROC. 2022. Recognizing objective and subjective language. *NROC developmental English foundations*.

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.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically neutralizing subjective bias in text. *arXiv preprint arXiv:1911.09709*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. *Advances in neural information processing systems*, 28.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A Olshausen, and Trevor Darrell. 2020. Fully test-time adaptation by entropy minimization.

Tianlu Wang, Jieyu Zhao, Kai-Wei Chang, Mark Yatskar, and Vicente Ordonez. 2018. Adversarial removal of gender from deep image representations. *arXiv preprint arXiv:1811.08489*, 3.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Adversarial neural machine translation. In *Asian Conference on Machine Learning*, pages 534–549. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. Wikibias: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1799–1814.

Appendix A. Data Pre-processing Details

We developed a parallel corpus for bias correction by analyzing Wikipedia articles histories with Neutral Point of View (NPOV) tags. Our data harvesting approach was similar to Recasens et. al (Recasens et al., 2013) and Pryzant et. al (Pryzant et al., 2019), but with some minor changes. Following (Recasens et al., 2013), we analyzed the revision histories for each Wikipedia article and downloaded sentences that were argued for NPOV issues. Sentences that had NPOV tags before the revision were considered as biased sentences and the sentences whose NPOV tags were removed after edits were considered as unbiased sentences. We ignored revisions that were related to missing references, misspellings and punctuation.

Corpus Name	Biased	Unbiased	Total
NPOV	32,541	75,024	107,565
MPQA	8,575	42,282	50,857
IBC	3,726	600	4,062

Table 6: Corpus statistics for each dataset used for training cross-domain bias classifier.

From the original NPOV corpus, we extracted sentences that had NPOV or peacock tags in their content before the edit. Pryzant et. al only considered single word edits for their bias correction model. Since the objective of this research is to correct bias that is induced by single word and multiword, we expanded the corpus by modifying the harvest function of pryzant et. al. Also, our method uses the latest dump from Wikipedia which contains new biased sentences that were not considered in the previous study. Additionally, some data cleanups were done to make this model not sensitive to noun phrases in the text. We replaced all noun phrases, but retained honorifics because some of the gender biases were introduced through honorifics. The numbers mentioned in the text were also replaced with NUM tag. A total of 408,738 sentences were extracted for our study. The NPOV corpus will help our bias detection model to learn common patterns that are used by Wikipedia editors for imposing subjective views. The details of the dataset are provided at Table 6.

Appendix B. Implementation

For Seq2Seq model, a multi-layer recurrent neural network based encoder, and an attention-based decoder was used(Vinyals et al., 2015; Bahdanau et al., 2014). Both the models are designed with 3 LSTM layers with 256 units at each layer. The hyper-parameters are tuned based on the noisy Wikipedia data. The pre-trained GloVe embedding is used as the embedding layer. The model is trained with Adam optimizer with the following hyper-parameters batch size = 32, learning rate = $5e^{-}5$ and a learning rate decay factor = 0.99. Based on the training and testing data statistics, the maximum sentence length is set to 30 for both input and output. Glove vectors is used as the pre-trained embedding. For pre-trained transformer model, we used the BART model (Lewis et al., 2019) with 139M parameters and fine tuned with Adam optimizer (learning rate $3e-5$).

Algorithm 1 REINFORCE Algorithm

Input: Input sequences (X), the output sequences (Y), and a pre-trained policy (θ)
Output: Trained policy with REINFORCE

Training Steps:

while not converged **do**
Select a batch of size N from X and Y
Sample N full sequence of actions:
$\{y_1, ..., y_M\}$
Observe the sequence reward and calculate the baseline r_b.
Calculate the loss
Update the parameters of network
end while

Testing Steps:

for batch of input and output sequences X and Y **do**
Use the trained model and sample the output
Evaluate the model using a performance metric, e.g. BLEU
end for

Appendix C. Bias Tagger

The bias tagger takes an input sentence and identifies a sequence of words as biased based on the context of use. Some of the subjective bias correction methods we developed needs additional information like biased words in the sentence to efficiently rewrite a polarized sentence, so a sequence tagging based bias tagger model was developed. By comparing the word edits between the sentences in parallel text generated for bias correc-

tion, we constructed a bias tagging dataset using a BIO format. Since state of the art sequence tagging models are BERT based (Devlin et al., 2018), we developed a BERT based sequence tagger by adding a dropout layer and a classification layer at the end of BERT model with a cross entropy loss. Pretrained RoBERTa model (Liu et al., 2019) was downloaded and fine-tuned for bias tagging. The bias tagger model had an accuracy of 95% with a recall of 92% on the validation set.

Appendix D. Policy Gradient

The goal of policy gradient methods is to update the probability distributions of actions in such a way that actions with higher expected rewards have a higher probability value for an observed state. The objective function for policy gradients is given by:

$$J(\theta) = E[\sum_{t=0}^{T-1} r_{t+1}] \tag{6}$$

where r_{t+1} is the reward received by performing action a_t at state s_t; $r_{t+1} = R(s_t, a_t)$, where R is the reward function. The policy is optimized by taking the gradient ascent based on the partial derivative of the objective based on the policy parameter theta:

$$\theta \leftarrow \theta + \frac{\partial}{\partial\theta} J(\theta) \tag{7}$$

The objective function can be expanded as:

$$J(\theta) = E[\sum_{t=0}^{T-1} r_{t+1}|\pi_\theta] \tag{8}$$

If $P(s_t, a_t|\tau)$ represents the probability of s_t, a_t occurring given the trajectory τ, then objective function is:

$$J(\theta) = \sum_{t=i}^{T-1} P(s_t, a_t|\tau) r_{t+1} \tag{9}$$

Then the policy gradient is given as:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta log\pi_\theta(a_t, s_t) \sum_{t'=t+1}^{T} \gamma^{t'-t-1} r_{t'} \tag{10}$$

where $\gamma \epsilon [0, 1]$ is the discount factor that helps to weight immediate rewards more than future rewards.

Appendix E. Human Judgement

A total of 100 sentences were collected from the model output and evaluated using 10 judges. The sentence selected for human judgement contains both biased and unbiased statements.Additionally, we ensured that the biased sentences collected contained both multi-occurrence bias as well as all three types of bias. The judges selected for this evaluation has computational linguistic background, but they don't have background information on what models were developed and their characteristics. Before presenting the data, a detailed definition of what is subjective bias with examples was presented to the user. For each example, a detailed note is provided to explain why the selected sentence is biased. Reference edits varied based on the aspect being evaluated. The text generated is evaluated for three aspects:

- **Neutrality**: The user is presented with the original biased statement and bias corrected results from one of the models, and asked to rate which of the statement is more biased on a scale of -2 to 2 (-2 is for original text is more biased, 0 is same biased and 2 for generated text is more biased). When presenting sentences, the placement of biased and corrected sentences is randomised.
- **Fluency**: A combination of the text generated by various models along with ground truth is presented to the user in pairs and asked which one is more fluent. The fluency of a sentence is rated in a scale of -2 to 2 (-2 for ground-truth is more fluent, 0 if both have same fluency and 2 if generated text is more fluent).
- **Content preservation**: For content preservation, the original sentence and model corrected sentence is presented to the user and asked whether the second sentence contains the same information as the first sentence. The content preservation is rated on a scale of 0 to 4 (0 representing the content is totally different and 4 being very similar).

Appendix F. Examples of System Output - Qualitative Comparison

See Table 8

Biased Statements	Neutral Statements
Less sophisticated believers in biblical inerrancy **may fail** to allow for the possibility of transcription errors or translation errors.	**A portion of** believers in biblical inerrancy **fail** to allow for the possibility of transcription errors or translation errors.
Released on May 16 , 2002 , Attack of the Clones was **generally perceived** as a **slight improvement** upon the feeble The Phantom Menace, **though not at all on par** with the original Star Wars trilogy.	Released on May 16 , 2002 , Attack of the Clones was **perceived** as an **improvement** upon the feeble The Phantom Menace, **though not on par** with the original Star Wars trilogy.
Only a tiny proportion of these companies have **so far** grown into multinationals : ARM, Autonomy Corporation and AVEVA are the **most obvious examples** , and more recently CSR has seen rapid growth due to the uptake of Bluetooth.	**A proportion** of these companies have **grown** into multinationals: ARM, Autonomy Corporation and AVEVA are the **few examples**, and more recently CSR has seen rapid growth due to the uptake of Bluetooth.

Table 7: Sample of biased and its corresponding neutralized statements extracted from Wikipedia. Spans highlighted are the output of our bias tagger by type such as **Epistemological Bias**, **Framing Bias** and **Demographic Bias**. **Neutralized** spans are highlighted in the second column. Neutralized spans are manually identified, not by bias tagger.

Examples of Bias Corrections	
Source	**She wore a beautiful dress which she had made herself.**
Reference	She wore a dress which she had made herself.
RL-Seq2Seq	The dress she wore was one she had made herself.
RL-Trans	The dress she wore was made by her
Source	**Harvington is a lively village with a number of amenities including a shop , farm shop and Post Office.**
Reference	Harvington is a village with a number of amenities including a shop , farm shop and Post Office.
RL-Seq2Seq	Harvington has a number of amenities including a shop , farm shop and Post Office.
RL-Trans	The village of Harvington has a number of amenities, including a shop, farm shop, and a post office.
Source	**It is known as one of the only places to get an authentic New York City style egg cream.**
Reference	It is known as one of the places to get an authentic New York City style egg cream.
RL-Seq2Seq	It is known for selling authentic New York City style egg cream.
RL-Trans	The restaurant is famous for its authentic New York City egg cream.
Source	**He supports the failed war in Iraq.**
Reference	He supports the war in Iraq.
RL-Seq2Seq	He is in favor of the Iraq war.
RL-Trans	The war in Iraq is supported by him.
Source	**They promote the false perception that evolution is a theory in crisis.**
Reference	They promote their minority perception that evolution is a theory in crisis.
RL-Seq2Seq	They promote a minority perception of evolution as a theory in crisis.
RL-Trans	They promote the perception that evolution is in crisis among minorities.
Source	**Less sophisticated believers in biblical inerrancy may fail to allow for the possibility of transcription errors or translation errors.**
Reference	Some believers in biblical inerrancy don't allow for the possibility of transcription errors or translation errors.
RL-Seq2Seq	Some believers in biblical inerrancy deny the possibility of transcription errors or translation errors.
RL-Trans	There are some believers in biblical inerrancy who don't accept the possibility of transcription errors or translation errors.
Source	**As a professional librarian he was exceptionally erudite and influenced by a wide range of authors , including Dante Alighieri , Miguel de Cervantes , Franz Kafka , H.G**
Reference	As a librarian he was exceptionally erudite and influenced by a wide range of authors , including Dante Alighieri , Miguel de Cervantes , Franz Kafka , H.G .
RL-Seq2Seq	As a librarian he was erudite and influenced by a wide range of authors , including Dante Alighieri , Miguel de Cervantes , Franz Kafka , H.G .
RL-Trans	As a librarian, he was well-educated and influenced by a variety of authors, among them Dante Alighieri, Miguel de Cervantes, Franz Kafka and H.G. Wells.

Table 8: Sample model outputs