Automatically Neutralizing Subjective Bias in Text

Reid Pryzant,[1] **Richard Diehl Martinez,**[1] **Nathan Dass,**[1]
Sadao Kurohashi,[2] **Dan Jurafsky,**[1] **Diyi Yang**[3]
[1]Stanford University
{rpryzant, rdm, ndass, jurafsky}@stanford.edu
[2]Kyoto University
kuro@i.kyoto-u.ac.jp
[3]Georgia Institute of Technology
diyi.yang@cc.gatech.edu

Abstract

Texts like news, encyclopedias, and some social media strive for objectivity. Yet bias in the form of inappropriate subjectivity — introducing attitudes via framing, presupposing truth, and casting doubt — remains ubiquitous. This kind of bias erodes our collective trust and fuels social conflict. To address this issue, we introduce a novel testbed for natural language generation: automatically bringing inappropriately subjective text into a neutral point of view ("neutralizing" biased text). We also offer the first parallel corpus of biased language. The corpus contains 180,000 sentence pairs and originates from Wikipedia edits that removed various framings, presuppositions, and attitudes from biased sentences. Last, we propose two strong encoder-decoder baselines for the task. A straightforward yet opaque CONCURRENT system uses a BERT encoder to identify subjective words as part of the generation process. An interpretable and controllable MODULAR algorithm separates these steps, using (1) a BERT-based classifier to identify problematic words and (2) a novel *join embedding* through which the classifier can edit the hidden states of the encoder. Large-scale human evaluation across four domains (encyclopedias, news headlines, books, and political speeches) suggests that these algorithms are a first step towards the automatic identification and reduction of bias.

1 Introduction

Writers and editors of texts like encyclopedias, news, and textbooks strive to avoid biased language. Yet bias remains ubiquitous. 62% of Americans believe their news is biased (Gallup 2018) and bias is the single largest source of distrust in the media (Foundation 2018).

This work presents data and algorithms for automatically reducing bias in text. We focus on a particular kind of bias: *inappropriate subjectivity* ("subjective bias"). Subjective bias occurs when language that should be neutral and fair is skewed by feeling, opinion, or taste (whether consciously or unconsciously). In practice, we identify subjective bias via the method of Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013): using Wikipedia's *neutral point of view (NPOV)* policy.[1] This policy is a set of principles which includes "avoiding stating opinions as facts" and "preferring nonjudgemental language".

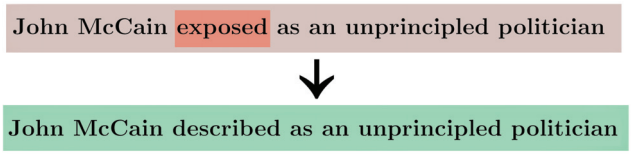

Figure 1: Example output from our MODULAR algorithm. "Exposed" is a factive verb that presupposes the truth of its complement (that McCain is unprincipled). Replacing "exposed" with "described" neutralizes the headline because it conveys a similar main clause proposition (someone is asserting McCain is unprincipled), but no longer introduces the authors subjective bias via presupposition.

For example a news headline like "John McCain exposed as an unprincipled politician" (Figure 1) is biased because the verb *expose* is a factive verb that presupposes the truth of its complement; a non-biased sentence would use a verb like *describe* so as not to presuppose the subjective opinion of the writer. "Pilfered" in "the gameplay is *pilfered* from DDR" (Table 1) subjectively frames the shared gameplay as a kind of theft. "His" in "a lead programmer usually spends *his* career" again introduces a biased and subjective viewpoint (that all programmers are men) through presupposition.

We aim to debias text by suggesting edits that would make it more neutral. This contrasts with prior research which has debiased *representations* of text by removing dimensions of prejudice from word embeddings (Bolukbasi et al. 2016; Gonen and Goldberg 2019) and the hidden states of predictive models (Zhao et al. 2018; Das, Dantcheva, and Bremond 2018). To avoid overloading the definition of "debias," we refer to our kind of text debiasing as *neutralizing* that text. Figure 1 gives an example.

We introduce the Wiki Neutrality Corpus (WNC). This is

[1]https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

Source	Target	Subcategory
A new downtown is being developed which will bring back...	A new downtown is being developed which **which its promoters hope** will bring back..	Epistemological
The authors' **exposé** on nutrition studies	The authors' **statements** on nutrition studies	Epistemological
He started writing books **revealing** a vast world conspiracy	He started writing books **alleging** a vast world conspiracy	Epistemological
Go is **the deepest** game in the world.	Go is **one of the deepest** games in the world.	Framing
Most of the gameplay is **pilfered from** DDR.	Most of the gameplay is **based on** DDR.	Framing
Jewish forces overcome Arab **militants**.	Jewish forces overcome Arab **forces**.	Framing
A lead programmer usually spends **his career** mired in obscurity.	Lead programmers often spend **their careers** mired in obscurity.	Demographic
The lyrics are about **mankind**'s perceived idea of hell.	The lyrics are about **humanity**'s perceived idea of hell.	Demographic
Marriage is a **holy union** of individuals.	Marriage is a **personal union** of individuals.	Demographic

Table 1: Samples from our new corpus. 500 sentence pairs are annotated with "subcategory" information (Column 3).

a new parallel corpus of 180,000 biased and neutralized sentence pairs along with contextual sentences and metadata. The corpus was harvested from Wikipedia edits that were designed to ensure texts had a neutral point of view. WNC is the first parallel corpus of biased language.

We also define the task of *neutralizing* subjectively biased text. This task shares many properties with tasks like detecting framing or epistemological bias (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013), or veridicality assessment/factuality prediction (Saurí and Pustejovsky 2009; Marneffe, Manning, and Potts 2012; Rudinger, White, and Van Durme 2018; White et al. 2018). Our new task extends these detection/classification problems into a generation task: generating more neutral text with otherwise similar meaning.

Finally, we propose a pair of novel sequence-to-sequence algorithms for this neutralization task. Both methods leverage denoising autoencoders and a token-weighted loss function. An interpretable and controllable MODULAR algorithm breaks the problem into (1) detection and (2) editing, using (1) a BERT-based detector to explicitly identify problematic words, and (2) a novel *join embedding* through which the detector can modify an editors' hidden states. This paradigm advances an important human-in-the-loop approach to bias understanding and generative language modeling. Second, an easy to train and use but more opaque CONCURRENT system uses a BERT encoder to identify subjectivity as part of the generation process.

Large-scale human evaluation suggests that while not without flaws, our algorithms can identify and reduce bias in encyclopedias, news, books, and political speeches, and do so better than state-of-the-art style transfer and machine translation systems. This work represents an important first step towards automatically managing bias in the real world. We release data and code to the public.[2]

2 Wiki Neutrality Corpus (WNC)

The Wiki Neutrality Corpus consists of aligned sentences *pre* and *post*-neutralization by English Wikipedia editors (Table 1). We used regular expressions to crawl 423,823 Wikipedia revisions between 2004 and 2019 where editors provided NPOV-related justification (Zanzotto and Pennacchiotti 2010; Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013; Yang et al. 2017). To maximize the precision of bias-related changes, we ignored revisions where

- More than a single sentence was changed.
- Minimal edits (character Levenshtein distance < 4).
- Maximal edits (more than half of the words changed).
- Edits where more than half of the words were proper nouns.
- Edits that fixed spelling or grammatical errors.
- Edits that added references or hyperlinks.
- Edits that changed non-literary elements like tables or punctuation.

We align sentences in the *pre* and *post* text by computing a sliding window (size $k = 5$) of pairwise BLEU (Papineni et al. 2002) between sentences and matching sentences with the biggest score (Faruqui et al. 2018; Tiedemann 2008). Last, we discarded pairs whose length ratios were beyond the 95th percentile (Pryzant et al. 2017).

Data	Sentence pairs	Total words	Seq length (mean)	# revised words (mean)
Biased-full	181,496	10.2M	28.21	4.05
Biased-word	55,503	2.8M	26.22	1.00
Neutral	385,639	17.4M	22.58	0.00

Table 2: Corpus statistics.

Corpus statistics are given in Table 2. The final data are (1) a parallel corpus of 180k biased sentences and their neutral counterparts, and (2) 385k neutral sentences that were adjacent to a revised sentence at the time of editing but were not changed by the editor. Note that following Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013), the neutralizing experiments in Section 4 focus on the subset of WNC where the editor modified or deleted a single word in the source text ("Biased-word" in Table 2).

Table 1 also gives a categorization of these sample pairs using a slight extension of the typology of Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013). They defined **framing bias** as using subjective words or phrases linked with a particular point of view (using words like *best* or *deepest* or using *pilfered from* instead of *based on*), and **epistemological bias** as linguistic features that subtly (often via presupposition) modify the believability of a proposition. We add to their two a third kind of subjectivity bias that also occurs in our data, which we call **demographic bias**,

[2]https://github.com/rpryzant/neutralizing-bias

text with presuppositions about particular genders, races, or other demographic categories (like presupposing that all programmers are male).

Subcategory	Percent
Epistemological	25.0
Framing	57.7
Demographic	11.7
Noise	5.6

Table 3: Proportion of bias subcategories in Biased-full.

The dataset does not include labels for these categories, but we hand-labeled a random sample of 500 examples to estimate the distribution of the 3 types. Table 3 shows that while framing bias is most common, all types of bias are represented in the data, including instances of demographic bias.

2.1 Dataset Properties

We take a closer look at WNC to identify characteristics of subjective bias on Wikipedia.

Topic. We use the Wikimedia Foundation's categorization models (Asthana and Halfaker 2018) to bucket articles from WNC into a 44-category ontology,[3] then compare the proportions of NPOV-driven edits across categories. Subjectively biased edits are most prevalent in *history*, *politics*, *philosophy*, *sports*, and *language* categories. They are least prevalent in the *meteorology*, *science*, *landforms*, *broadcasting*, and *arts* categories. This suggests that there is a relationship between a text's topic and the realization of bias. We use this observation to guide our model design in Section 3.1.

Tenure. We group editors into "newcomers" (less than a month of experience) and "experienced" (more than a month). We find that newcomers are less likely to perform neutralizing edits (15% in WNC) compared to other edits (34% in a random sample of 685k edits). This difference is significant ($\tilde{\chi}^2$ p = 0.001), suggesting the complexity of neutralizing text is typically reserved for more senior editors, which helps explain the performance of human evaluators in Section 6.1.

3 Methods for Neutralizing Text

We propose the task of neutralizing text, in which the algorithm is given an input sentence and must produce an output sentence whose meaning is as similar as possible to the input but with the subjective bias removed.

We propose two algorithms for this task, each with its own benefits. A MODULAR algorithm enables human control and interpretability. A CONCURRENT algorithm is simple to train and operate.

We adopt the following notation:

- $\mathbf{s} = [w_1^s, ..., w_n^s]$ is a *source sequence* of subjectively biased text.
- $\mathbf{t} = [w_1^t, ..., w_m^t]$ is a *target sequence* and the neutralized version of $\mathbf{s}$.

[3]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Directory

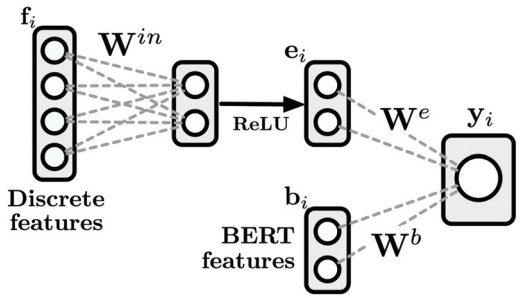

Figure 2: The detection module uses discrete features $\mathbf{f}_i$ and BERT embedding $\mathbf{b}_i$ to calculate logit y_i.

3.1 MODULAR

The first algorithm we are proposing is called MODULAR. It has two stages: BERT-based detection and LSTM-based editing. We pretrain a model for each stage and then combine them into a joint system for end-to-end fine tuning on the overall neutralizing task. We proceed to describe each module.

Detection Module The detection module is a neural sequence tagger that estimates p_i, the probability that each input word w_i^s is subjectively biased (Figure 2).

Module description. Each p_i is calculated according to

$$p_i = \sigma(\mathbf{b}_i\ \mathbf{W}^b + \mathbf{e}_i\ \mathbf{W}^e + b) \quad (1)$$

- $\mathbf{b}_i \in \mathcal{R}^b$ represents w_i^s's semantic meaning. It is a contextualized word vector produced by BERT, a transformer encoder that has been pre-trained as a masked language model (Devlin et al. 2019). To leverage the bias-topic relationship uncovered in Section 2.1, we prepend a token indicating an article's topic category (<arts>, <sports>, etc) to $\mathbf{s}$. The word vectors for these tokens are learned from scratch.
- $\mathbf{e}_i$ represents expert features of bias proposed by (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013):

$$\mathbf{e_i} = ReLU(\mathbf{f}_i\ \mathbf{W}^{in}) \quad (2)$$

  $\mathbf{W}^{in} \in \mathcal{R}^{f \times h}$ is a matrix of learned parameters, and $\mathbf{f}_i$ is a vector of discrete features[4].
- $\mathbf{W}^b \in \mathcal{R}^b$, $\mathbf{W}^e \in \mathcal{R}^h$, and $b \in \mathcal{R}$ are learnable parameters.

Module pre-training. We train this module using diffs[5] between the source and target text. A label p_i^* is 1 if w_i^s was deleted or modified as part of the neutralizing process. A label is 0 if the associated word was unchanged during editing, i.e. it occurs in both the source and target text. The loss is calculated as the average negative log likelihood of the labels:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\Big[p_i^* \log p_i + (1 - p_i^*)\log(1 - p_i)\Big]$$

[4]Such as lexicons of hedges, factives, assertives, implicatives, and subjective words; see code release.

[5]https://github.com/paulgb/simplediff

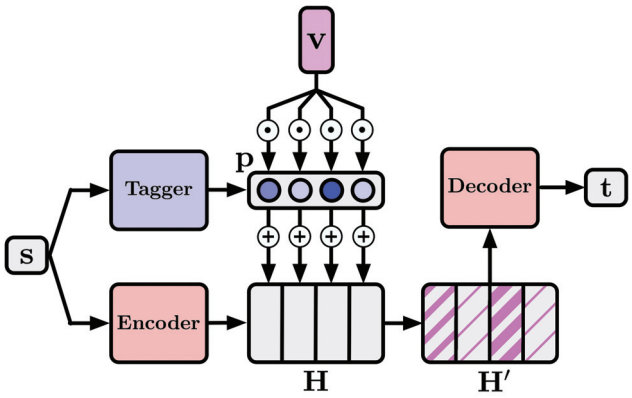

Figure 3: The MODULAR system uses *join embedding* $\mathbf{v}$ to reconcile the detector's predictions with an encoder-decoder architecture. The greater a word's probability, the more of $\mathbf{v}$ is mixed into that word's hidden state.

Editing Module The editing module takes a subjective source sentence $\mathbf{s}$ and is trained to edit it into a more neutral compliment $\mathbf{t}$.

Module description. This module is based on a sequence-to-sequence neural machine translation model (Luong, Pham, and Manning 2015). A bi-LSTM encoder turns $\mathbf{s}$ into a sequence of hidden states $\mathbf{H} = (\mathbf{h}_1, ..., \mathbf{h}_n)$ (Hochreiter and Schmidhuber 1997). Next, an LSTM decoder generates text one token at a time by repeatedly attending to $\mathbf{H}$ and producing probability distributions over the vocabulary. We also add two mechanisms from the summarization literature (See, Liu, and Manning 2017). The first is a copy mechanism, where the model's final output for timestep i becomes a weighted combination of the predicted vocabulary distribution and attentional distribution from that timestep. The second is a coverage mechanism which incorporates the sum of previous attention distributions into the final loss function to discourage the model from re-attending to a word and repeating itself.

Module pre-training. We pre-train the decoder as a language model of neutral text using the *neutral* portion of WNC (Section 2). Doing so expresses a data-driven prior about how target sentences should read. We accomplish this with a denoising autoencoder objective (Hill, Cho, and Korhonen 2016) and maximizing the conditional log probability of reconstructing a sequence $\mathbf{x}$ from a *corrupted* version of itself $\widetilde{\mathbf{x}}$ using noise model C ($\log p(\mathbf{x}|\widetilde{\mathbf{x}})$ where $\widetilde{\mathbf{x}} = C(\mathbf{x})$).

Our C is similar to (Lample et al. 2018). We slightly shuffle $\mathbf{x}$ such that x_i's index in $\widetilde{\mathbf{x}}$ is randomly selected from $[i-k, i+k]$. We then drop words with probability p. For our experiments, we set $k = 3$ and $p = 0.25$.

Final System Once the detection and editing modules have been pre-trained, we join them and fine-tune together as an end to end system for translating $\mathbf{s}$ into $\mathbf{t}$.

This is done with a novel *join embedding* mechanism that lets the detector control the editor (Figure 3). The join embedding is a vector $\mathbf{v} \in \mathcal{R}^h$ that we add to each encoder hidden state in the editing module. This operation is gated by the detector's output probabilities $\mathbf{p} = (p_1, ..., p_n)$. Note that the same $\mathbf{v}$ is applied across all timesteps.

$$\mathbf{h}'_i = \mathbf{h}_i + p_i \cdot \mathbf{v} \tag{3}$$

We proceed to condition the decoder on the new hidden states $\mathbf{H}' = (\mathbf{h'}_1, ..., \mathbf{h}'_n)$ which have varying amounts of $\mathbf{v}$ in them. Intuitively, $\mathbf{v}$ is enriching the hidden states of words that the detector identified as subjective. This tells the decoder what language should be changed and what is safe to be be copied during the neutralization process.

Error signals are allowed to flow backwards into both the encoder and detector, creating an end-to-end system from the two modules. To fine-tune the parameters of the joint system, we use a token-weighted loss function that scales the loss on neutralized words (i.e. words unique to $\mathbf{t}$) by a factor of α:

$$\mathcal{L}(s,t) = -\sum_{i=1}^{m} \lambda(w_i^t, s) \log p(w_i^t | s, w_{<i}^t) + c$$

$$\lambda(w_i^t, s) = \begin{cases} \alpha & : \ w_i^t \notin s \\ 1 & : \ \text{otherwise} \end{cases}$$

Note that c is a term from the coverage mechanism (Section 3.1). We use $\alpha = 1.3$ in our experiments. Intuitively, this loss function incorporates an inductive bias of the neutralizing process: the source and target have a high degree of lexical similarity but the goal is to learn the structure of their *differences*, not simply copying words into the output (something a pre-trained autoencoder should already have knowledge of). This loss function is related to previous work on grammar correction (Junczys-Dowmunt et al. 2018), and cost-sensitive learning (Zhou and Liu 2006).

3.2 CONCURRENT

Our second algorithm takes the problematic source $\mathbf{s}$ and directly generates a neutralized $\hat{\mathbf{t}}$. While this renders the system easier to train and operate, it limits interpretability and controllability.

Model description. The CONCURRENT system is an encoder-decoder neural network. The encoder is BERT. The decoder is the same as that of Section 3.1: an attentional LSTM with copy and coverage mechanisms. The decoder's inputs are set to:

- Hidden states $\mathbf{H} = \mathbf{W}^H \mathbf{B}$, where $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_n) \in \mathcal{R}^{b \times n}$ is the BERT-embedded source and $\mathbf{W}^H \in \mathcal{R}^{h \times b}$ is a matrix of learned parameters.
- Initial states $\mathbf{c}_0 = \mathbf{W}^{c0} \sum \mathbf{b}_i / n$ and $\mathbf{h_0} = \mathbf{W}^{h0} \sum \mathbf{b}_i / n$. $\mathbf{W}^{c0} \in \mathcal{R}^{h \times b}$ and $\mathbf{W}^{h0} \in \mathcal{R}^{h \times b}$ are learned matrices.

Model training. The CONCURRENT model is pre-trained with the same autoencoding procedure described in Section 3.1. It is then fine-tuned as a subjective-to-neutral translation system with the same loss function described in Section 3.1.

4 Experiments

4.1 Experimental Protocol

Implementation. We implemented nonlinear models with Pytorch (Paszke et al. 2017) and optimized using Adam

Method	BLEU	Accuracy	Fluency	Bias	Meaning
Source Copy	91.33	0.00	-	-	-
Detector (always delete biased word)	92.43*	38.19*	-0.253*	-0.324*	1.108*
Detector (predict substitution from biased word)	92.51	36.57*	-0.233*	-0.327*	1.139*
Delete Retrieve (ST) (Li et al. 2018)	88.46*	14.50*	-0.209*	-0.456*	1.294*
Back Translation (ST) (Prabhumoye et al. 2018)	84.95*	9.92*	-0.359*	-0.390*	1.126*
Transformer (MT) (Vaswani et al. 2017)	86.40*	24.34*	-0.259*	-0.458*	0.905*
Seq2Seq (MT) (Luong, Pham, and Manning 2015)	89.03*	23.93	-0.423*	-0.436*	1.294*
Base	89.13	24.01	-	-	-
+ *loss*	90.32*	24.10	-	-	-
+ *loss* + *pretrain*	92.89*	34.76*	-	-	-
+ *loss* + *pretrain* + *detector* (MODULAR)	93.52*	**45.80***	-0.078	**-0.467***	0.996*
+ *loss* + *pretrain* + *BERT* (CONCURRENT)	**93.94**	44.87	**0.132**	-0.423*	**0.758***
Target copy	100.0	100.0	-0.077	-0.551*	1.128*

Table 4: Bias neutralization performance. ST indicates a style transfer system. MT indicates a machine translation system. For quantitative metrics, rows with asterisks are significantly different than the preceding row. For qualitative metrics, rows with asterisks are significantly different from zero. Higher is preferable for *fluency*, while lower is preferable for *bias* and *meaning*.

(Kingma and Ba 2014) as configured in (Devlin et al. 2019) with a learning rate of 5e-5. We used a batch size of 16. All vectors were of length $h = 512$ unless otherwise specified. We use gradient clipping with a maximum gradient norm of 3 and a dropout probability of 0.2 on the inputs of each LSTM cell (Srivastava et al. 2014). We initialize the BERT component of the tagging module with the publicly-released `bert-base-uncased` parameters. All other parameters were uniformly initialized in the range $[-0.1, 0.1]$.

Procedure. Following Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013), we train and evaluate our system on a subset of WNC where the editor changed or deleted a single word in the source text. This yielded 53,803 training pairs (about a quarter of the WNC), from which we sampled 700 development and 1,000 test pairs. For fair comparison, we gave our baselines additional access to the 385,639 *neutral* examples when possible. We pretrained the tagging module for 4 epochs. We pretrained the editing module on the *neutral* portion of our WNC for 4 epochs. The joint system was trained on the same data as the tagger for 25,000 steps (about 7 epochs). We perform interference using beam search and a beam width of 4. All computations were performed on a single NVIDIA TITAN X GPU; training the full system took approximately 10 hours.

Evaluation. We evaluate our models according to five metrics. BLEU (Papineni et al. 2002) and accuracy (the proportion of decodings that exactly matched the editors changes) are quantitative. We report statistical significance with bootstrap resampling and a 95% confidence level (Koehn 2004; Efron and Tibshirani 1994).

We also hired fluent English-speaking crowdworkers on Amazon Mechanical Turk for qualitative evaluation. Workers were shown the Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013) and Wikipedia definition of a "biased statement" and six example sentences, then subjected to a five-question qualification test where they had to identify subjectivity bias. Approximately half of the 30,000 workers who took the qualification test passed. Those who passed were asked to compare pairs of original and edited sentences (not knowing which was the original) along three criteria: fluency, meaning preservation, and bias. Fluency and bias were evaluated on a Semantic Differential scale from -2 to 2. Here, a semantic differential scale can better evaluate attitude oriented questions with two polarized options (e.g., "is text A or B more fluent?"). Meaning was evaluated on a Likert scale from 0 to 4, ranging from "identical" to "totally different". Inter-rater agreement was fair to substantial (Krippendorff's alpha of 0.65 for fluency, 0.33 for meaning, and 0.51 for bias)[6]. We report statistical significance with a t-test and 95% confidence interval.

4.2 Wikipedia (WNC)

Results on WNC are presented in Table 4. In addition to methods from the literature we include (1) a BERT-based system which simply predicts and deletes subjective words, and (2) a system which predicts replacements (including deletion) for subjective words directly from their BERT embeddings. All methods appear to successfully reduce bias according to the human evaluators. However, many methods appear to lack fluency. Adding a token-weighted loss function and pretraining the decoder help the model's coherence according to BLEU and accuracy. Adding the detector (MODULAR) or a BERT encoder (CONCURRENT) provide additional benefits. The proposed models retain the strong effects of systems from the literature while also producing target-level fluency on average. Our results suggest there is no clear winner between our two proposed systems. MODULAR is better at reducing bias and has higher accuracy, while CONCURRENT produces more fluent responses, preserves meaning better, and has higher BLEU.

Table 5 indicates that BLEU is more correlated with fluency but accuracy is more correlated with subjective bias reduction. The weak association between BLEU and human evaluation scores is corroborated by other research (Cha-

[6] Rule of thumb: k $<$ 0 "poor" agreement, 0 to .2 "slight", .21 to .40 "fair", .41 to .60 "moderate", .61 - .80 "substantial", and .81 to 1 "near perfect" (Gwet 2011).

Metric	Fluency	Bias	Meaning
BLEU	0.65	0.34	0.16
Accuracy	0.56	0.52	0.20

Table 5: Spearman correlation (R^2) between quantitative and qualitative metrics.

ganty, Mussman, and Liang 2018; Mir et al. 2019). We conclude that neither automatic metric is a true substitute for human judgment.

4.3 Real-world Media

To demonstrate the efficacy of the proposed methods on subjective bias in the wild, we perform inference on three out-of-domain datasets (Table 6). We prepared each dataset according to the same procedure as WNC (Section 2). After inference, we enlisted 1800 raters to assess the quality of 200 randomly sampled datapoints. Note that for partisan datasets we sample an equal number of examples from "conservative" and "liberal" sources. These data are:

- The Ideological Books Corpus (IBC) consisting of partisan books and magazine articles (Sim et al. 2013; Iyyer et al. 2014).
- Headlines of partisan news articles identified as biased according to mediabiasfactcheck.com.
- Sentences from the campaign speeches of a prominent politician (United States President Donald Trump).[7] We filtered out dialog-specific artifacts (interjections, phatics, etc) by removing all sentences with less than 4 tokens before sampling a test set.

Overall, while MODULAR does a better job at reducing bias, CONCURRENT appears to better preserve the meaning and fluency of the original text. We conclude that the proposed methods, while imperfect, are capable of providing useful suggestions for how subjective bias in real-world news or political text can be reduced.

5 Error Analysis

To better understand the limits of our models and the proposed task of bias neutralization, we randomly sample 50 errors produced by our models on the Wikipedia test set and bin them into the following categories:

- **No change.** The model failed to remove or change the source sentence.
- **Bad change.** The model modified the source but introduced an edit which failed to match the ground-truth target (i.e. the Wikipedia editor's change).
- **Disfluency.** Errors in language modeling and text generation.
- **Noise.** The datapoint is noisy and the target text is not a neutralized version of the source.

[7]Transcripts from www.kaggle.com/binksbiz/mrtrump

IBC Corpus

Method	Fluency	Bias	Meaning
MODULAR	-0.041	-0.509*	0.882*
CONCURRENT	-0.001	-0.184	0.501*
Original	**Activists** have filed a lawsuit...		
MODULAR	**Critics of it** have filed a lawsuit...		
CONCURRENT	**Critics** have filed a lawsuit...		

News Headlines

Method	Fluency	Bias	Meaning
MODULAR	-0.46*	-0.511*	1.169*
CONCURRENT	-0.141*	-0.393*	0.752*
Original	Zuckerberg **claims** Facebook can...		
MODULAR	Zuckerberg **stated** Facebook can...		
CONCURRENT	Zuckerberg **says** Facebook can...		

Trump Speeches

Method	Fluency	Bias	Meaning
MODULAR	-0.353*	-0.563*	1.052*
CONCURRENT	-0.117	-0.127	0.757*
Original	This includes **amazing** Americans like...		
MODULAR	This includes Americans like...		
CONCURRENT	This includes **some** Americans like...		

Table 6: Performance on out-of-domain datasets. Higher is preferable for *fluency*, while lower is preferable for *bias* and *meaning*. Rows with asterisks are significantly different from zero

Error Type	Proportion (%)	Valid (%)
No change	38	0
Bad change	42	80
Disfluency	12	0
Noise	8	87

Table 7: Distribution of model errors on the Wikipedia test set. We also give the percent of errors that were valid neutralizations of the source despite failing to match the target sentence.

The distribution of errors is given in Table 7. Most errors are due to the subtlety and complexity of language understanding required for bias neutralization, rather than the generation of fluent text. These challenges are particularly pronounced for neutralizing edits that involve the replacement of factive and assertive verbs. As column 2 shows, a large proportion of the errors, though disagreeing with the edit written by the Wikipedia editors, nonetheless succeeded in neutralizing the source.

Examples of each error type are given in Table 9 (two pages away). As the examples show, our models have have a tendency to simply remove words instead of finding a good replacement.

6 Algorithmic Analysis

We proceed to analyze our algorithm's ability to detect and categorize bias as well as the efficacy of the proposed join embedding.

Method	Accuracy
Linguistic features	0.395*
Bag-of-words	0.584*
+*Linguistic features* (Recasens, 2013)	0.617
BERT	0.744*
+*Linguistic features*	0.752
+*Linguistic features* + *Category* (MODULAR detector)	**0.759**
CONCURRENT encoder	0.745
Human	0.543*

Table 8: Performance of various bias detectors. Rows with asterisks are statistically different than the preceding row.

6.1 Detecting Subjectivity

Identifying subjectivity in a sentence (explicitly or implicitly) is prerequisite to neutralizing it. We accordingly evaluate our model's (and 3,000 crowdworker's) ability to detect subjectivity using the procedure of Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013). We use the same 50k training examples as Section 4 (Table 8). For each sentence, we select the word with the highest predicted probability and test whether that word was in fact changed by the editor. The proportion of correctly selected words is the system's "accuracy". Results are given in Table 8.

Note that CONCURRENT lacks an interpretive window into its detection behavior, so we estimate an upper bound on the model's detection abilities by (1) feeding the encoder's hidden states into a fully connected + softmax layer that predicts the probability of a token being subjectively biased, and (2) training this layer as a sequence tagger according to the procedure of Section 3.1.

The low human performance can be attributed to the difficulty of identifying bias. Issues of bias are typically reserved for senior Wikipedia editors (Section 2.1) and untrained workers performed worse (37.39%) on the same task in (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013) (and can struggle on other tasks requiring linguistic knowledge (Callison-Burch 2009)). CONCURRENT's encoder, which is architecturally identical to BERT, had similar performance to a stand-alone BERT system. The linguistic and category-related features in the MODULAR detector gave it slight leverage over the plain BERT-based models.

6.2 Join Embedding

We continue by analyzing the abilities of the proposed join embedding mechanism.

Join Embedding Ablation The join embedding combines two separately pretrained models through a gated embedding instead of the more traditional practice of stripping off any final classification layers and concatenating the exposed hidden states (Bengio et al. 2007). We ablated the join embedding mechanism by training a new model where the pretrained detector is frozen and its pre-output hidden states $\mathbf{b}_i$ are concatenated to the encoder's hidden states before decoding. Doing so reduced performance to 90.78 BLEU and 37.57 Accuracy (from the 93.52/46.8 with the join embedding). This suggests learned embeddings can be a high-performance and end-to-end conduit between sub-modules of machine learning systems.

Join Embedding Control We proceed to demonstrate how the join embedding creates controllability in the neutralization process. Recall that MODULAR relies on a probability distribution $\mathbf{p}$ to determine which words require editing (Equation 3). Typically, this distribution comes from the detection module (Section 3.1), but we can also feed in user-specified distributions that force the model to target particular words. This can let human advisors correct errors or push the model's behavior towards some desired outcome. We find that the model is indeed capable of being controlled, letting users target specific words for rewording in case they disagree with the model's output or seek recommendations on specific language. However, doing so can also introduce errors into downstream language generation (Table 9, next page).

7 Related Work

Subjectivity Bias. The study of subjectivity in NLP was pioneered by the late Janyce Wiebe and colleagues (Bruce and Wiebe 1999; Hatzivassiloglou and Wiebe 2000). Several studies develop methods for highlighting subjective or persuasive frames in a text (Rashkin et al. 2017; Tsur, Calacci, and Lazer 2015), or detecting biased sentences (Hube and Fetahu 2018; Morstatter et al. 2018; Yang et al. 2017; Hube and Fetahu 2019) of which the most similar to ours is Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013), whose early, smaller version of WNC and logistic regression-based bias detector inspired our study.

Debiasing. Many scholars have worked on removing demographic prejudice from *meaning representations* (Manzini et al. 2019; Zhao et al. 2017; 2018; Bordia and Bowman 2019; Wang et al. 2018, inter alia). Such studies begin with identifying a direction or subspace that capture the bias and then removing this bias component to make representations fair across attributes like gender and age (Bolukbasi et al. 2016; Manzini et al. 2019). For instance, Bordia and Bowman (2019) introduced a regularization term for the language model to penalize the projection of the word embeddings onto that gender subspace, while Wang et al. (2018) used adversarial training to squeeze directions of bias out of hidden states.

Neural Language Generation. Several studies propose stepwise or modular procedures for text generation, including sampling from a corpus (Guu et al. 2018) and identifying language ripe for modification (Leeftink and Spanakis 2019). Most similar to us is Li et al. (2018) who localize a text's style to a fraction of its words. Our MODULAR detection module performs a similar localization in a soft manner, but our steps are joined by a smooth conduit (the join embedding) instead of discrete logic. There is also work related to our CONCURRENT model. The closest is Dun, Zhu, and Zhao (2019), where a decoder was attached to BERT for question answering, or Lample et al. (2018), where machine translation systems are initialized to LSTM

Error Type	Source, Output, then Target
No change	Existing hot-mail accounts were **upgraded** to outlook.com on April 3, 2013. Existing hot-mail accounts were **upgraded** to outlook.com on April 3, 2013. Existing hot-mail accounts were **changed** to outlook.com on April 3, 2013.
Bad change	His **exploitation** of leased labor began in 1874 and continued until his death in 1894... His **actions** of leased labor began in 1874 and continued until his death in 1894... His **use** of leased labor began in 1874 and continued until his death in 1894...
Disfluency	Right before stabbing a cop, flint attacker shouted one thing that **proves** terrorism is still here. Right before stabbing a cop, flint attacker shouted one thing that **may may** terrorism is still here. Right before stabbing a cop, flint attacker shouted one thing that **may prove** terrorism is still here.
Noise	...then whent to war with him in the Battle of **Bassorah**, and ultimately left that battle. ...then whent to war with him in the Battle of **Bassorah**, and ultimately left that battle. ...then whent to war with him in the Battle of **the Camel**, and ultimately left that battle.

Revised Word	Source, Output, then Target
Magnificent	After a dominant performance, Joshua...with a **magnificent** seventh-round knockout win. After a dominant performance, Joshua...with a seventh-round knockout win. After a dominant performance, Joshua...with a seventh-round knockout win.
Dominant	Jewish history is...interacted with other **dominant** peoples, religions and cultures. Jewish history is...other peoples, religions and cultures. Jewish history is...other peoples, religions and cultures.

Selected Word	Output
(input) merely misapplied	In recent years, the term has often been **misapplied** to those who are **merely** clean-cut. In recent years, the term has often been **misapplied** to those who are clean-cut. In recent years, the term has often been *shown* to those who are **merely** clean-cut.
(input) assassination massacre	He was responsible for the **assassination** of Carlos Marighella, and for the Lapa **massacre**. He was responsible for the *killing* of Carlos Marighella, and for the Lapa **massacre**. He was responsible for the assassination of Carlos Marighella, and for the Lapa *incident*.
(input) desperately scandal	Paul Ryan **desperately** searches for a new focus amid Russia **scandal**. Paul Ryan searches for a new focus amid Russia **scandal**. Paul Ryan **desperately** searches for a new focus amid Russia.

Table 9: **Top**: examples of model errors from each error category. **Middle**: the model treats words differently based on their context; in this case, "dominant" is ignored when it accurately describes an individual's winning performance, but deleted when it describes a group of people in arbitrary comparison. **Bottom**: the MODULAR model can sometimes be controlled, for example by selecting words to change, to correct errors or otherwise change the model's behavior.

and Transformer-based language models of the source text.

8 Conclusion and Future Work

The growing presence of bias has marred the credibility of our news, educational systems, and social media platforms. Automatically reducing bias is thus an important new challenge for the Natural Language Processing and Artificial Intelligence community. This work represents a first step in the space. Our results suggest that the proposed models are capable of providing useful suggestions for how to reduce subjective bias in real-world expository writing like news, books, and encyclopedias. Nonetheless our scope was limited to single-word edits, which only constitute a quarter of the edits in our data, and are probably among the simplest instances of bias. We therefore encourage future work to tackle broader instances of multi-word, multi-lingual, and cross-sentence bias. Another important direction is integrating aspects of fact-checking (Mihaylova et al. 2018), since a more sophisticated system would be able to know when a presupposition is in fact true and hence not subjective. Finally, our new join embedding mechanism can be applied to other modular neural network architectures.

9 Acknowledgements

We thank the Japan-United States Educational Commission (Fulbright Japan) for their generous support. We thank Chris Potts, Hirokazu Kiyomaru, Abigail See, Kevin Clark, the Stanford NLP Group, and our anonymous reviewers for their thoughtful comments and suggestions. We gratefully acknowledge support of the DARPA Communicating

with Computers (CwC) program under ARO prime contract no. W911NF15-1-0462 and the NSF via grant IIS-1514268. Diyi Yang is thankful for support by a grant from Google.

References

Asthana, S., and Halfaker, A. 2018. With few eyes, all hoaxes are deep. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):21.

Bengio, Y.; Lamblin, P.; Popovici, D.; and Larochelle, H. 2007. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, 153–160.

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 *Advances in neural information processing systems*, 4349–4357.

Bordia, S., and Bowman, S. R. 2019. Identifying and reducing gender bias in word-level language models. In *NAACL 2019 Student Research Workshop*.

Bruce, R. F., and Wiebe, J. M. 1999. Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering* 5(2):187–205.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of EMNLP*, 286–295.

Chaganty, A. T.; Mussman, S.; and Liang, P. 2018. The price of debiasing automatic metrics in natural language evaluation. In *Proceedings of ACL*.

Das, A.; Dantcheva, A.; and Bremond, F. 2018. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dun, P.; Zhu, L.; and Zhao, D. 2019. Extending answer prediction for deep bi-directional transformers. *32nd Conference on Neural Information Processing Systems (NIPS)*.

Efron, B., and Tibshirani, R. J. 1994. *An introduction to the bootstrap*. CRC press.

Faruqui, M.; Pavlick, E.; Tenney, I.; and Das, D. 2018. Wikiatomicedits: A multilingual corpus of wikipedia edits for modeling language and discourse. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Foundation, O. S. 2018. Indicators of news media trust. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/216/original/KnightFoundation_Panel4_Trust_Indicators_FINAL.pdf.

Gallup. 2018. Americans: Much misinformation, bias, inaccuracy in news. https://news.gallup.com/opinion/gallup/235796/\\americans-misinformation-bias-\\inaccuracy-news.aspx.

Gonen, H., and Goldberg, Y. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Guu, K.; Hashimoto, T. B.; Oren, Y.; and Liang, P. 2018. Generating sentences by editing prototypes. *Transactions of the Association of Computational Linguistics* 6:437–450.

Gwet, K. L. 2011. On the krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October* 2(2011):2011.

Hatzivassiloglou, V., and Wiebe, J. M. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING 2000*, 299–305.

Hill, F.; Cho, K.; and Korhonen, A. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Hube, C., and Fetahu, B. 2018. Detecting biased statements in wikipedia. In *The Web Conference*, 1779–1786. International World Wide Web Conferences Steering Committee.

Hube, C., and Fetahu, B. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 195–203. ACM.

Iyyer, M.; Enns, P.; Boyd-Graber, J.; and Resnik, P. 2014. Political ideology detection using recursive neural networks. In *Proceedings of ACL*, 1113–1122.

Junczys-Dowmunt, M.; Grundkiewicz, R.; Guha, S.; and Heafield, K. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of NAACL*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *International Conference for Learning Representations (ICLR)*.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Phrase-based & neural unsupervised machine translation.

Leeftink, W., and Spanakis, G. 2019. Towards controlled transformation of sentiment in sentences. *International Conference on Agents and Artificial Intelligence (ICAART)*.

Li, J.; Jia, R.; He, H.; and Liang, P. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of NAACL*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*.

Manzini, T.; Lim, Y. C.; Tsvetkov, Y.; and Black, A. W. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *NAACL 2019*.

Marneffe, M.-C. d.; Manning, C. D.; and Potts, C. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics* 38(2):301–333.

Mihaylova, T.; Nakov, P.; Marquez, L.; Barron-Cedeno, A.; Mohtarami, M.; Karadzhov, G.; and Glass, J. 2018. Fact checking in community forums. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mir, R.; Felbo, B.; Obradovich, N.; and Rahwan, I. 2019. Evaluating style transfer for text. In *Proceedings of NAACL*.

Morstatter, F.; Wu, L.; Yavanoglu, U.; Corman, S. R.; and Liu, H. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing* 1(2):5.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, 311–318.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. *Association for Computational Linguistics (ACL)*.

Pryzant, R.; Chung, Y.; Jurafsky, D.; and Britz, D. 2017. Jesc: Japanese-english subtitle corpus. *11th edition of the Language Resources and Evaluation Conference (LREC)*.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*, 2931–2937.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL*, 1650–1659.

Rudinger, R.; White, A. S.; and Van Durme, B. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 731–744.

Saurí, R., and Pustejovsky, J. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation* 43(3):227.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. *Proceedings of ACL*.

Sim, Y.; Acree, B. D.; Gross, J. H.; and Smith, N. A. 2013. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*, 91–101.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.

Tiedemann, J. 2008. Synchronizing translated movie subtitles. In *Language Resources and Evaluation Conference (LREC)*.

Tsur, O.; Calacci, D.; and Lazer, D. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of ACL*, 1629–1638.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, T.; Zhao, J.; Chang, K.-W.; Yatskar, M.; and Ordonez, V. 2018. Adversarial removal of gender from deep image representations. *arXiv preprint arXiv:1811.08489*.

White, A. S.; Rudinger, R.; Rawlins, K.; and Van Durme, B. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4717–4724.

Yang, D.; Halfaker, A.; Kraut, R.; and Hovy, E. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Conference on Empirical Methods in Natural Language Processing*, 2000–2010.

Zanzotto, F. M., and Pennacchiotti, M. 2010. Expanding textual entailment corpora from wikipedia using co-training. In *The People's Web Meets NLP Workshop (COLING)*, 28–36.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhou, Z.-H., and Liu, X.-Y. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Transactions of IEEE*.