WIKIBIAS: Detecting Multi-Span Subjective Biases in Language

Yang Zhong[1], **Jingfeng Yang**[2], **Wei Xu**[2], **Diyi Yang**[2]
[1] The Ohio State University
[2] Georgia Institute of Technology
zhong.536@osu.edu
yjfllpyym@gmail.com
dyang888@gatech.edu
wei.xu@cc.gatech.edu

Abstract

Biases continue to be prevalent in modern text and media, especially subjective bias – a special type of bias that introduces improper attitudes or presents a statement with the presupposition of truth. To tackle the problem of detecting and further mitigating subjective bias, we introduce a manually annotated parallel corpus WIKIBIAS with more than 4,000 sentence pairs from Wikipedia edits. This corpus contains annotations towards both sentence-level bias types and token-level biased segments. We present systematic analyses of our dataset and results achieved by a set of state-of-the-art baselines in terms of three tasks: bias classification, tagging biased segments, and neutralizing biased text. We find that current models still struggle with detecting multi-span biases despite their reasonable performances, suggesting that our dataset can serve as a useful research benchmark. We also demonstrate that models trained on our dataset can generalize well to multiple domains such as news and political speeches.[1]

1 Introduction

People often rely on reference work like encyclopedias and textbooks to gather information, as such sources are designed to present facts fairly and objectively. Yet, bias is still pervasive in these sources. For instance, the sentence "*This album is arranged by many talented arrangers.*" is considered biased as the word *talented* inappropriately reflects the writer's positive opinion. As a result, methods that can automatically detect and reduce bias are in great demand, which could save human efforts and keep the quality of the reference work.

In this work, we study how to detect and further mitigate biases in language. Specifically, we focus on a particular type of bias, "subjective bias", in which the language is skewed towards an obvious feeling, with the presupposed or entailed proposition or considering opinions as truth. Contents with the subjective bias can make people be doubtful about the texts' reliability and possibly trigger social unrest with offensive language. Prior research has used the lexical and grammatical cues like lexicon-syntactic patterns (Wiebe and Riloff, 2005; Riloff and Wiebe, 2003) or various n-gram features (Murray and Carenini, 2009; Wilson and Raaijmakers, 2008; Wiebe et al., 1999) to classify sentences as either subjective or objective. For instance, in the encyclopedia domain, Recasens et al. (2013) constructed an automatic parallel corpus from Wikipedia revisions that violate the Neutral Point of View (NPOV) policy,which advocates for "*fairly presenting views with reliable sources and avoiding editor bias*" and introduced the task of identifying the bias-induced word in a statement. They further uncovered two types of subjective bias through linguistic analysis, which includes *framing* bias such as praising or perspective-specific words and *epistemological* bias related to presupposed/entailed propositions. Pryzant et al. (2020) extended such revision corpus and further proposed to transform the biased text into a neutral point of view, adding a third class of subjective bias, *demographic* bias, for texts with the presupposition of demographic categories like genders and races.

However, current corpora on subjective bias detection or mitigation tasks suffer from a set of issues. First, noises from automatically collected datasets (Recasens et al., 2013; Pryzant et al., 2020) are not neglectable. A pilot study conducted by Pryzant et al. (2020) on their Wikipedia Neutrality Corpus (WNC) demonstrated that over 5% of the revisions are not related to bias mitigation and thus wrongly labeled on the sentence level. Meanwhile, existing manually annotated corpora for subjectivity often suffer from the small dataset size in Wiebe et al. (1999) or limited annotation quality: annotator agreement from Hube and Fetahu (2019) falls

[1] Our code and data are publicly available at https://github.com/cs329yangzhong/WIKIBIAS.

Source Sentence: pre-edit (biased language)	Target Sentence: post-edit (neutral language)
It should be noted that [a] the nuclear-free zone act does not make building land-based nuclear power plants illegal, and there is **considerable** [b] support for nuclear power **in order to** [c] meet Kyoto emissions targets.	The nuclear-free zone act does not make building land-based nuclear power plants illegal, and there is **some business** support for **investigating** nuclear power, **which could help** [c] meet Kyoto emissions targets.
Anti-Americanism is a **claimed** [a] phenomenon of **subvert** [b] **ethnic discrimination** [c] and **overt irrational** [d] **hostility** [e] **toward** [f] the United States.	Anti-Americanism is a **global** [a] phenomenon of **discrimination** [c] and **criticism** [e] **of** [f] the United States.
However the term post-fascist has been used to describe **their belief** [a], owing to **apparent** [b] intellectual roots in **neo-fascist third positionism** [d].	However , the term 'post-fascist' has been used to describe **the beliefs of recent National Anarchists** [a], owing to **their** [b] intellectual roots **which lie partly** [c] in **third positionism** , **an ideology often considered to be neo-fascist** [d].

Table 1: Example sentence pairs in our manually annotated WIKIBIAS corpus with three fine-grained subjective bias types: **framing** , **epistemological** , **demographic** , and **not** bias. We annotate at the span-level to identify the corresponding pre- and post-edits, which are indicated by the same superscript characters (e.g., in row 1, the highlighted phrase *in order to* is changed to *which could help* during revision).

at 0.124 measured by Krippendorff's Alpha. Moreover, multiple edits are often needed when editing a subjectively biased framing into a neutral one. For instance, over 30% of Wikipedia revisions for NPOV justification contain two or more edits in the source side and a diverse set of modification strategies are involved. Existing work (Recasens et al., 2013; Pryzant et al., 2020) only focused on *single word* detection, presupposing a single word as the source of bias, and failing to utilize rich signals and resources of subjectively biased words or phrases as introduced in (Wiebe et al., 2004).

To address these problems, we introduce a high-quality manually annotated parallel corpus WIKIBIAS. It includes over 4,000 biased and neutralized sentence pairs, which cover both 1,525 single word and 2,068 multiple-word span annotations (building upon 53.5k non-identical word alignments with fine-grained bias types on the source sides. Samples of our corpus are shown in Table 1. We design an innovative two-stage annotation pipeline to help annotators accurately identify biased text segments, which obtains substantial agreement among different annotators. To the best of our knowledge, this is the first corpus on the multi-word multi-span subjective biased text understanding. Table 2 summarizes the key differences between WIKIBIAS and other previous datasets contributed for the subjective bias detection task.

Building on WIKIBIAS, we conduct a set of comprehensive analyses to better model subjectivity bias in text via three sub-tasks: bias classification, tagging biased segments and neutralizing biased text. We found that current state-of-the-art models still struggle with detecting multi-span biases despite their reasonable performances, suggesting that our dataset can serve as a useful benchmark. We also demonstrate that models trained on our dataset can generalize well to multiple domains such as news and political speeches.

2 Construction of the WIKIBIAS Corpus

We create the new WIKIBIAS corpus by first extracting Wikipedia revisions where editors provide Neutral Point of View (NPOV) [2] justifications (Recasens et al., 2013; Yang et al., 2017; Zanzotto and Pennacchiotti, 2010; Pryzant et al., 2020) to construct automatically labeled data (WIKIBIAS-AUTO); then manually annotating sentences with fine-grained bias types at the span-level to create clean ground truth (WIKIBIAS-MANUAL). This is in contrast to the prior work on subjectivity that annotated only on the sentence-level (Wiebe et al., 1999; Hube and Fetahu, 2019, 2018). In particular, we design a two-stage human annotation methodology to handle sentences with *both* single- and multi-edits. We describe the details below.

2.1 Extracting and Filtering Wikipedia Edits

About 0.1% of revisions in Wikipedia are tagged with "NPOV" (or "POV-check", "POV-section", etc.) by editors to indicate that they have identified and rewritten biased content to achieve a more neutral tone. In total, we extracted 557,860 NPOV-related revisions from the Wikipedia revision history dump (dated 01/01/2021), out of the 691 million revisions that Wikipedia editors made between 2004 and 2021. We closely follow Pryzant et al. (2020)'s method[3] and apply a set of rules to filter out revisions that span across multiple blocks

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

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

Dataset	Domain	Covered Tasks	Annotation	Agreement	# Sentences
SUBJECTIVE (Wiebe et al., 1999)	News	Clas	annotators	high	1,004
NPOV-MANUAL (Recasens et al., 2013)	Wikipedia	Tag	crowd	medium	230
LANGUAGEBIAS (Hube and Fetahu, 2018)	Conservapedia	Clas	crowd	low	685
PHRASINGBIAS (Hube and Fetahu, 2019)	Wikipedia	Clas	crowd	low	4,952
WIKIBIAS-MANUAL (this work)	**Wikipedia**	**Clas;Tag;Gen**	**annotators**	**high**	**8,198**
WNC-WORD (Pryzant et al., 2020)	Wikipedia	Tag; Gen	automatic	–	111k
WIKIBIAS-AUTO (this work)	**Wikipedia**	**Clas;Tag;Gen**	**automatic**	**–**	**421k**

Table 2: Comparison of biased language detection datasets. **Clas**, **Tag** and **Gen** refer to sentence classification, tagging biased spans, and generation for neutralizing bias, respectively.

of text that contains only grammar error fix, involve either extremely dramatic (more than half words changed) or minimal (character-level Levenshtein distance is less than 4) changes, relate to table/punctuation or adding of references. To extract the sentence pairs from the collected revisions (68.5% contain multiple sentences), Pryzant et al. (2020) computed the pairwise BLEU of single sentences from the pre- and post-edited text and match the single sentence pairs with the highest score. In the end, we modified their post-processing script to remove duplicated revisions and keep the latest revisions for each pre-edited text based on the timestamp. We also removed duplicated revisions and keep the latest revisions for each pre-edited text based on the timestamp. We eventually acquired a parallel corpus of 214,987 sentence pairs of *pre* and *post*-NPOV edits.

After reserving 4,099 sentence pairs (randomly sampled) for human annotation (§2.2), we apply a rule-based method to extract modifications for the remaining 210,888 sentence pairs to construct the WIKIBIAS-AUTO. We pair up *pre* and *post*-edited text spans using a word diff extractor,[4] and clean with heuristic rules. More details can be found in Appendix C. We then treat edited spans in pre-edits as biased and assigned biased and neutral sentence-level labels for the sentence pairs respectively, similar to Pryzant et al. (2020). When evaluating on the 4,099 manually annotated sentence pairs, this heuristic method can obtain 87% accuracy for sentence-level labels, 84.7% precision, and 76.6% recall for extracting edited spans on the source side. We provide the statistics of WIKIBIAS-AUTO in Table 4.

[4]Following Pryzant et al. (2020), we use the simplediff package to compute a minimal diff at word level: https://github.com/paulgb/simplediff

	WIKIBIAS AUTO	WIKIBIAS MANUAL
Sentence level		
# of sent pair	210,888	4,099
# of biased sent	210,888	3,400
# of neutral sent	210,888	4,798
Span-level revisions		
# of source spans	286,156	5,148
# of unique source spans	153,598	3,804
average # of source spans	1.36	1.25
Source-side biased spans		
# of *framing* bias	–	2,654
# of *epistemological* bias	–	808
# of *demographic* bias	–	131
total number of spans	198,413†	3,593
average # of spans per input	0.94†	1.06
average length of spans	2.63†	2.93

Table 3: Statistics of our WIKIBIAS corpus with automatically (§2.1) and manually (§2.2) annotations.

2.2 Fine-grained Human Annotation

While most of these extracted revisions contain biased content as they were flagged by the editors as POV-related, our manual inspection on a preliminary subsample of 499 sentences pairs reveals that about 13% of them are not actually biased. Moreover, Wikipedia editors may make multiple changes to a sentence (see examples in Table 1). In contrast to previous work (Pryzant et al., 2020) that has discarded these sentences, we designed a two-stage annotation procedure to annotate them and include in our dataset. In particular, we introduce a simple but efficient step of word/phrase alignment, that has not been used before for annotating biased language, to tackle the difficulty in identifying biased spans in texts with multiple edits.

Recognizing Edited Spans via Word Alignment. For each pair of pre and post-edit sentences, we first visualize the using GoldAlign, an annotation tool from Gokcen et al. (2016), then ask two in-house annotators to highlight all word/phrase alignments

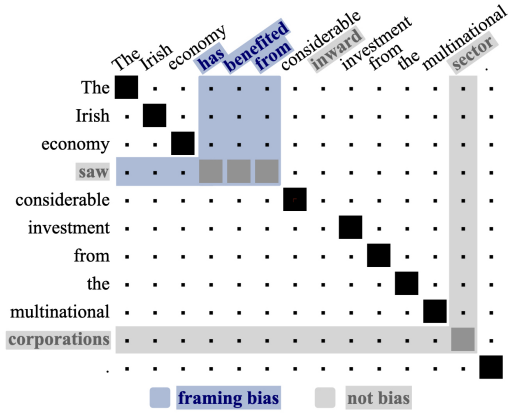

Figure 1: Word Alignment and Bias Annotation example for a pre-edited (top) and post-edited (left) sentence pair. Grey block ■ means non-identical word/phrase alignment. Three edits are extracted: *has benefited from* → *saw*, *inward* → `[NULL]` (Deletion), *sector* → *corporations* while the first is labelled as *framing* bias.

(see example in Figure 1). More specifically, we hire an undergraduate student and a high school intern, both undergo at least weeks of training sessions with the task description and examples to get them familiar with the task. We provide detailed guidelines to the annotators with an emphasis on identifying the modified spans and their post-edited counterparts can aid in the bias classification task on span level. Evaluations on a held-out task training data demonstrated that both students obtained equally high-quality annotations. In the end, we applied a post-processing script to extract non-identical word/phrase pairs from the alignment annotations. The words and phrases that are added or deleted by the Wikipedia editors are also extracted as they are aligned to a special symbol `[NULL]`.

Labeling Bias Type for Span Pairs. We then classify each non-identical word/phrase alignment into one of the following categories, following prior work (Recasens et al., 2013): (1) **framing bias** with the use of one-sided words or phrases containing a particular point of view; (2) **epistemological bias** which includes subtle linguistic features that can affect the believability of the texts; (3) **demographic bias** with word/phrase usage under presuppositions of a particular demographic factor (i.e., gender or religion); or (4) **no bias**.

We designed an annotation interface (see Appendix A.1 for a screenshot) using Label Studio (Tkachenko et al., 2020), and asked two more in-house annotators (both are native English speakers with college-level education) to label the type of bias at the span-level as shown in Table 1. We provided annotators with both the edited span pairs and the original sentences, taking into consideration the context dependent biases. The pilot study we conducted in the early stage of annotation shows that the proper extraction of span pairs can assist in identifying the fine-grained bias types. For example (Figure 1), knowing that the phrase "*in order to*" is replaced by "*which could help*" is helpful for annotators to determine that the the former presupposes the usefulness of the subject while the latter one behaves less determinate.

We ended up with the WIKIBIAS-MANUAL corpus that contains 4,099 sentence pairs. In total of 1,525 single- and 2,068 multiple- word spans are annotated as biased, of which 2,654 are classified as *framing*, 808 as *epistemological* and 131 as *demographic* biases. We derived the sentence-level labels from the span annotations. The pre-edited sentences are labeled as biased if one or more edited spans were classified as biased. Otherwise, both sentences are marked as neutral.

Annotation Agreement. Following previous work, we calculate the inter-annotator agreements for word/phrase alignment task by comparing one annotator against the gold arbitrated annotations on non-identical (non-trivial) alignments, which are 98.4/98.5/98.1 and 89.8/89.9/89.5 measured by Precision/Recall/F1 on the token-level and phrase-level respectively. The inter-annotator agreement is 0.712 for the fine-grained bias type classification and 0.734 for binary cases (all three types of biases vs. no bias) by Cohen's Kappa (Artstein and Poesio, 2008), suggesting a substantial agreement. To ensure the annotation quality, we constantly monitored annotators' agreement over 40 random examples in every batch of 200 instances for double annotation. Double-annotated contents with diverged opinions are further examined by the first author, followed by discussions with two annotators until all agreed.

3 Modeling Subjective Bias

Subjective biases shall be modeled differently for various applications. For instance, automatic bots of online media platforms may choose to flag and filter out biased sentences directly, for which classifying whether a sentence is biased is essential. When human editors work on an article, they might need some hints on potentially biased text snippets, as well as alternatives, where tagging biased segments or even generating a neutralized version becomes important. To this end, we propose three

different tasks on top of WIKIBIAS.

3.1 Sentence Classification

WIKIBIAS enables the development of classifiers to detect whether a sentence is biased or not on both coarse- and fine-grained level. We experiment with pre-trained language models and test how well they could pick up the nuance differences between biased and neutral sentences.

3.1.1 Binary Classification

Most prior work on bias detection (Hube and Fetahu, 2018, 2019; Pant et al., 2020) focus on predicting the presence of subjective bias in a sentence. We follow their setup. We also utilize the heuristically created WIKIBIAS-AUTO data with noisy labels (10% false positives for model training.

Experimental Setup. We trained multiple binary classifiers using different data splits: (1) use only human-annotated WIKIBIAS-MANUAL (i.e., $\text{Train}_{\text{manual}}$) data for training; (2) train on WIKIBIAS-AUTO (i.e., $\text{Train}_{\text{auto}}$) data. We additionally experimented with two methods from the literature for improving the performance with noisy labels: (3) finetune the model trained on noisy labels further using the clean data (Krause et al., 2016); (4) train on a filtered version of WIKIBIAS-AUTO, with top-5% and top-10% of automatically labeled "biased" instances with the lowest possibility removed (Li et al., 2017), utilizing a classifier trained on the original WIKIBIAS-AUTO.

Results. We observe that, as shown in Table 5, the incorporation of large noisy data improves the prediction. The model experiencing two-stage fine-tuning on $\text{Train}_{\text{auto}}$ and $\text{Train}_{\text{manual}}$ sets obtains the highest F1 and Accuracy. Although the model trained on clean data secures the highest precision, the low recall value suggests that the small $\text{Train}_{\text{manual}}$ set fails to fully cover the variants of biases. Meanwhile, removing low confidence "biased" samples from the training set brings improvements to recall and F1. In the end, we observe that the best baseline model achieves less than 70 F1, suggesting that baselines are still having trouble capturing biases on the sentence level.

3.1.2 Fine-grained Bias Type Classification

Initial analysis on the WIKIBIAS-MANUAL shows that 7% of the biased sentences contain more than one type of biases associated with multiple spans. We thus frame this task as multilabel classification where three binary classifiers predict the presence of each of the three subcategories (i.e., *framing*, *epistemological*, and *demographic*).

Dataset	Total (#sent)	biased	neutral	SLen
$\text{Train}_{\text{auto}}$	421,776	210,888	210,888	29.8
$\text{Train}_{\text{manual}}$	5,028	2,117	2,911	29.2
Dev	1,066	431	635	30.1
Test	2,104	852	1,252	30.1

Table 4: Data split and size for the experiments. The automatically constructed WIKIBIAS-AUTO corpus is used for training only ($\text{Train}_{\text{auto}}$). The manually annotated WIKIBIAS-MANUAL corpus is split into Train/Dev/Test set. SLen represent the average sentence length in terms of the number of tokens.

Experimental Setup. We fine-tuned BERT-base (Devlin et al., 2019) via the HuggingFace Transformers library (Wolf et al., 2020).[5] Pre-training on the binary task was explored with the hope to incorporate the inductive bias of binary prediction into the fine-grained setting. In detail, (1) we fine-tune a classifier with the BERT checkpoint and compare it to (2) the FINETUNED model with encoder copied from a BERT classifier fine-tuned on the binary task. (3) Similar to Ferracane et al. (2021), we use a HIERARCHICAL model with two classifiers to mimic the hierarchy of our label categories: the first binary classifier predicts the presence of bias while the second predicts the fine-grained label.

Results. We report macro-averaged F1, which gives equal weight to all classes, on the test set with an average of three runs (Table 6). Fine-grained prediction suffers from the imbalance of class labels. The improvement of 5.1 points on macro-F1 illustrates that pre-training the encoder with the binary task contributes to the fine-grained classification. However, in general, the models' performance is relatively low, which is primarily attributed to the incorrect prediction of *epistemological* and *demographic* bias. HIERARCHICAL obtains the highest macro-F1 and the per class results, showing the additional binary classifier helps to reduce the prediction error for *epistemological* bias.

3.2 Tagging of Biased Language Spans

To extract the biased spans from given sentences, we frame it as a sequence tagging task using the BIO scheme. We also experiment with a joint model in a multi-task learning fashion, aiming at

[5] Implementation Details in Appendix D.1

Train Data	P	R	F1	Acc
Standard Dataset				
$\text{Train}_{\text{manual}}$	**70.2**	38.6	52.1	68.1
$\text{Train}_{\text{auto}}$	63.6	67.1	65.2	71.8
$\text{Train}_{\text{auto}}$ ♦	68.0	63.9	65.8	**73.0**
Variations of Train$_{auto}$				
$\text{Train}_{\text{auto}}$ - 5% positive	61.6	68.5	65.0	69.9
$\text{Train}_{\text{auto}}$ - 10% positive	62.0	**72.6**	**66.3**	70.0

Table 5: Binary classification result on test set with different training data, reported on average of three runs. ♦ means the model is further fine-tuned on $\text{Train}_{\text{manual}}$.

Model	macro-F1	class-level F1		
		F	*E*	*D*
BERT	33.9	56.3	22.0	24.2
FINETUNED	39.0	62.1	20.5	35.2
HIERARCHICAL	41.0	61.0	26.5	35.8

Table 6: Macro and class-level F1 (*Framing*, *Epistemological*, and *Demographic* bias) on test set, averaged across three runs.

learning inter-relations between the segment tagging and the sentence classification tasks.

Biased Segment Tagging. We experiment with multiple baselines (Table 7), including (1) a BiLSTM-CNN-CRF model (Ma and Hovy, 2016), (2) a $\text{BERT}_{\text{Atten}}$ baseline which extracts words/phrases receiving high self-attention scores in the BERT encoder fine-tuned for the binary classification task (§3.1.1), (3) a DETECTOR model from (Pryzant et al., 2020) which labels the word with highest predicted probability, and (4) a fine-tune BERT tagging model in which we use the base size checkpoint as the encoder and a linear layer to predict token labels. Prior work (Recasens et al., 2013; Pryzant et al., 2020) demonstrated that linguistic features can assist in the detection of subjective bias. Thus, (5) we incorporate the linguistic features into the BERT-based tagging model. We concatenate the contextualized BERT embedding of each token with the encoded discrete linguistic features[6] and use a two-layer feed-forward network for final prediction (BERT-LING). We also apply our best BERT-LING model to relabel the large $\text{Train}_{\text{auto}}$ dataset, aiming at removing apparent noises that could be easily detected with the model.

Joint Sentence Classification and Tagging. We deploy a model to jointly learn sentence-level classification and token-level segmentation of bias. More specifically, we utilize a BERT tagging model with an additional sentence classifier. The model is trained on $\text{Train}_{\text{auto}}$ through a joint loss term. We then assign different weights for the classification loss of biased sentences, the classification loss of neutral sentences, and the tagging loss of biased sentences, trading off on the contribution of each task. We also add the Joint Model-LING, where we incorporate in the linguistic features.

[6]I.e., lexicons of hedges (Thompson, 2005), factive verbs (Hooper, 1975), and subjective clues (Wilson et al., 2005).

Model	Tagging		Classification	
	EX F1	P F1	F1	Acc
Tagging				
BILSTM-CNN-CRF*	32.7	36.4	–	–
$\text{BERT}_{\text{ATTEN}}$	29.8	37.3	–	–
DETECTOR	26.2	35.9	–	–
BERT*	35.3	42.5	–	–
BERT	47.5	55.4	–	–
BERT-LING	47.9	56.4	–	–
BERT-LING♦	47.9	56.5	–	–
BERT-LING†	**48.3**	**56.8**	–	–
Classification				
BERT	–	–	65.2	71.8
BERT♦	–	–	65.7	**73.0**
Joint Classification and Tagging Models				
JOINT MODEL	47.0	55.0	66.3	71.2
JOINT MODEL-LING	47.7	56.0	**67.0**	71.9
Upper Bound	83.8	85.8	95.3	92.8

Table 7: Tagging results. * indicates that model is fine-tuned on $\text{Train}_{\text{manual}}$ only while all others are trained on $\text{Train}_{\text{auto}}$. ♦ indicates further fine-tuning on $\text{Train}_{\text{manual}}$. † indicates training on the relabelled $\text{Train}_{\text{auto}}$ with labels predicted by the best BERT-LING♦ model. Results are averaged over 3 runs.

Results We report the phrase-level Exact Match and Partial Match F1 on the WIKIBIAS-MANUAL test set in Table 7. We also estimate the human upper bound by reporting the average performance of two annotators over the double-annotated test set. More specifically, for each individual annotator, we obtain the span annotations following the steps in §2.2 and further derive the sentence-level labels if at least one span in the pre-edit sentence is marked as biased.

We first observe that the incorporation of large noisy data improves the prediction. The injection of linguistic features boosts the performance and re-filtering of the noisy labels with the trained model provides further performance gain. The state-of-the-art baselines still struggle with multi-span detection, with significantly worse performance comparing to the estimated human upper bond. Thus, our corpus can serve as a useful research benchmark for future studies. Manual inspections on tagging results suggest that models mainly failed in detecting spans with content-dependent bias and

preserving the completeness of phrases. The joint model achieves worse performance on the segment tagging task which is mainly attributed to the lower recall, while obtains a slight performance gain on the classification task.

3.3 Text Generation for Neutralizing Bias

Bias neutralization can also be viewed as a text generation problem (Pryzant et al., 2020). In this section, we experiment with multiple generation baselines over WIKIBIAS, including Source Copy (directly copy input as output), LSTM and attention based seq2seq model (Luong et al., 2015), CopyNet (Gu et al., 2016), Transformer (Vaswani et al., 2017), pre-trained BART (Lewis et al., 2020) as well the MODULAR model in Pryzant et al. (2020) as baselines. All models are trained on $\text{Train}_{\text{auto}}$ except for the off-the-shelf MODULAR model, which was trained on WNC corpus and could provide comparisons between multi-span based generation and single-word edit oriented generation.

Automatic Evaluation. To evaluate the generated sentences, we compared them with neutralization references based on three generation related metrics: BLEU (Papineni et al., 2002), Sent Acc (the percentage of generated sentences that exactly match with the references) as well as Acc (the neutralization success rate using our best-performed classifier). We report statistical significance with bootstrap resampling and a 95% confidence level (Koehn, 2004; Efron and Tibshirani, 1994).

As shown in Table 8, CopyNet improves the performance of other unpretrained Seq2Seq in terms of BLEU and Sent Acc, because the models still retain most words in the original sentence despite the modified multi-word spans. Pre-trained BART model outperforms all other models on generating the same sentence as the references, although BLEU of BART does not outperform CopyNet. The inconsistent trend of BLEU and Sent Acc indicates that neither automatic metric is perfect enough to measure the naturalness of debiased results. We also observe a huge gap on Acc (15 points) between MODULAR model and all others. We suspect that generation models equipped only with single-word bias detection might not pick up the complete multi-word biased spans, thus fail to generate high-quality sentence neutralization.

Human Evaluation. We also perform a human evaluation on Amazon Mechanical Turk over 100 random sentence pairs for each model. Following Pryzant et al. (2020), for each sentence pair (randomized order), we collect 3 judgments on three criteria: Fluency, Meaning preservation, and Bias.[7] Table 8 shows that the pre-trained BART model with multi-span edit information outperformed all others in bias mitigation while maintaining text fluency and preserving the meaning. In contrast, single-word edit-based model MODULAR fails to neutralize the bias and suffers from the loss of information by dropping off a single word, a frequent strategy utilized in Pryzant et al. (2020).

Error Analysis. We examine 100 generation results produced by BART and MODULAR model and compared to the references, observing several error types: (1) No change (30%), (2) Reinforcing Bias (12%) where generated contents become more biased due to improper modification. For instance, BART changes "*himself or herself*" to "*himself*", which reinforces the demographic bias related to gender. In another example, BART model change the word "*Sadly*" to "*However*", making negative point of view more explicit. (3) Noise (10%) in which generated contents successfully mitigate the bias, but do not match with the references.

4 Generalization to Out-Of-Domain Data

To demonstrate the out-of-domain generalizability of our tagging model, we perform inferences on three out-of-domain datasets: (1) Ideological Books Corpus (**IBC**) (Sim et al., 2013; Iyyer et al., 2014) which consists of partisan books and magazine article; (2) **News** headlines of partisan news articles identified as biased according to `mediabiasfactcheck.com`; and (3) **Political speeches** of the first and third 2020 presidency election debates between Donald Trump and Joe Biden. All three sets of corpora can be separated into two groups based on their partisan identifications (Liberal/Democratic vs. Conservative/Republican). Examples of extracted spans are shown in Table 9.

Qualitative Results. We find that: (1) Our tagging model can extract meaningful multi-word phrases, as well as subtle metaphor phenomena. For instance, "*out of thin air*" in the last row of Table 9 carries the subjective bias of sudden/mysterious appearing. Interesting metaphors such as "*but there are some bad apples*" would never be detected by a single-word tagger. (2)

[7] Fluency and bias had scales of -2 to 2, Meaning was evaluated on a scale from 0 (identical) to 4 (totally different).

Method	Automatic Evaluation			Human Evaluation		
	BLEU↑	Sent Acc↑	Acc↑	Fluency↑	Bias↓	Meaning↓
SOURCE COPY	80.10	0.00	–	–	–	–
LSTM	82.12*	15.26*	68.20*	0.090	-0.367*	0.943*
TRANSFORMER	81.34*	15.49	65.96*	**0.119***	-0.211*	0.989*
COPYNET	**82.95***	16.31	65.96	-0.030	-0.507*	**0.577***
BART	82.22	**17.84**	**75.35***	0.017	**-0.588***	0.753*
MODULAR†	80.36*	13.76*	51.04*	-0.007	-0.313*	1.074*
TARGET COPY	100.0	100.0	80.63	0.023	-0.578*	1.074*

Table 8: Bias neutralization generation results on the test set. All models are trained on the noisy $\text{Train}_{\text{auto}}$ data and † means we used the off-the-shelf model released by their authors. For automatic metrics, rows with asterisks are significantly different than the preceding row. For human evaluation, rows marked with * are significantly different from 0 (according to a t-test with $p < 0.05$). ↑ / ↓ means higher/lower score is preferred for the corresponding metric.

Corpus	F1	Extracted multi-word spans
BIDEN	21.7	*they have a plan* $^{+}$, *but there are some bad apples*, *totally thoroughly discredited* $^{-}$, *being ripped down*
TRUMP	14.5	*because Obamacare is no good* $^{-}$, *very powerfully*, *tremendous* $^{+}$, *very big, incredibly, huge, big stuff*
NEWS	38.0	*exposes trumps dirty little apprentice lie* $^{-}$, *frustrated hypocrite* $^{-}$, *barbaric trumpcare* $^{-}$, *creepy*
NEWS	15.4	*huge scandal* $^{-}$, *with this mighty act* $^{+}$, *trump triumph* $^{+}$, *seriously wrong*, *stealing from* $^{-}$
IBC	25.5	*as skillfully as anyone, slightly more legitimate, less-beloved but more dogged,* *extraordinary* $^{+}$, *seize*
IBC	18.0	*it should be obvious that* $^{+}$, *frivolous lawsuits is killing the goose that lays the golden egg, out of thin air*

Table 9: Samples of frequent multi-word phrases extracted by our tagging model from each corpus with manual annotation on polarity of stance. The second column refers to the partial matching F1 based on 50 manually annotated samples from each corpus. Text colors in the first column refer to the opinions leaning towards U.S. political parties **Liberal/Democratic** or **Conservative/Republican**. Colored Boxes refer to the target of *Republican* or *Democratic* party respectively and $+/-$ signs illustrate whether the phrase is supported or against the stance of the target (i.e., *totally irresponsible* $^{-}$ illustrates that the speaker uses this phrase to criticize the work of Republican Party).

The extracted phrases from the speeches domain cover the signature words of the speaker without in-domain knowledge. "*have a plan*" is prevalent in 2020's presidency debates and signature words "*tremendous*" and "*very powerfully*" of Donald Trump have also been captured. (3) The model can tight the connection between subjective bias with research over stance detection, especially in the formal text domains (Thomas et al., 2006; Walker et al., 2012; Chakrabarty et al., 2019; Lawrence and Reed, 2020). With our subjective bias tagger, complete verb phrases or noun phrases can be obtained, which naturally eases the extraction of topics and opinions, two necessary components for stance detection problem. For instance, "*because Obamacare is no good*" span can sufficiently illustrate the opinion of Trump that is against the prior healthcare policy. Meanwhile, "*frustrated hypocrite*" can indicate the left-wing media's dislike of the Republican governor's behavior.

Human Evaluation. We sampled 50 sentences per corpus for human annotations. For each sentence, 3 qualified Turkers were asked to pick the biased spans without length constraints. We consider a span receiving more than one annotator vote the gold label. The second column in Table 9 shows that our model performs well on news headlines, as the annotated spans are mostly single or short multi-word spans given the relative short context. In contrast, low agreements are obtained in the speech domain. Manual inspections reveal that our model tends to tag phrases including subjective pronouns such as "*I*" and "*we*", which are informing signals in the Wikipedia domain for expressing subjective opinions, but under-perform in speech transcripts.

5 Related Work

Detection of Subjective Bias. The study of detection of subjectivity can be dated back to 1990s, when pioneers start noticing the subjectivity genre on document level classification (Karlgren and Cutting, 1994; Kessler et al., 1997). Later, works like (Bruce and Wiebe, 1999; Hatzivassiloglou and Wiebe, 2000) bring people's attention to the subjectivity on sentence level. There is a long line of research focusing on sentence classification utilizing methods based on linguistic features or handcrafted rules (Riloff and Wiebe, 2003; Wiebe and Riloff,

2005; Pang and Lee, 2004; Lin et al., 2011; Murray and Carenini, 2009; Yang et al., 2017), then neural models (Morstatter et al., 2018; Hube and Fetahu, 2018; Pant et al., 2020; Hube and Fetahu, 2019). Work of Recasens et al. (2013) and Pryzant et al. (2020) on detecting biased language over single-word edit is closely related to our work, but we study the biased language on a broader scale to cover multi-word spans.

Debiasing Generation. Generating debiased text can be viewed as a stylistic transferring task. Supervised approaches with parallel corpus have been shown to be effective across multiple styles (Xu et al., 2012; Hu et al., 2017; Reddy and Knight, 2016; Xu et al., 2015; Rao and Tetreault, 2018). More recently, pipeline-based or stepwise approaches (Li et al., 2018; Leeftink and Spanakis, 2019; Madaan et al., 2020) focuses on first localizing the style to a fixed portion of the word, then generating replacement based on target style. Pryzant et al. (2020) adopts a similar approach by incorporating the localized style attribute into a joint-embedding and enforces the text generation model to pay attention to the modifications.

6 Conclusion

In this work, we contribute the first manually annotated parallel corpus of over 4,000 sentence pairs for the task of subjective bias detection. This corpus covers multiple-word span annotations with fine-grained bias type on the source side and sentence level bias type. We perform the first systematic study for the detection of multi-span biased language. Experiments results on three tasks: classification, tagging, and generation demonstrated the usefulness of our corpus with state-of-the-art baselines. We also conclude a set of challenges that current models struggled with. In the future, we plan to generalize our models to more domains for bias detection, mitigation, and neutralization.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. We would like to thank NVIDIA for providing GPU computing resources. We also thank Sarah Flanagan, Kenneth Koepcke, Yulu Qin, and Panya Bhinder for their help with data annotation. DY is supported in part by funding from Google and Amazon.

7 Ethical Considerations

The collected dataset aims at helping detect and further mitigate subjective biases, such as Wikipedia and books, thus keeping the contents fair and unbiased. Our dataset was originally extracted from Wikipedia's revision history. As a free online encyclopedia, Wikipedia grants users the rights to copy and reuse contents under the copyleft licenses: Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA)[8] and GNU Free Documentation License (GFDL)[9].

Regarding dataset manual annotations, three undergraduate students and one high school student are involved in the in-house annotation task. Payment assignments are based on self-reported working hours, and the price item was set to ensure that workers were paid ($10∼$13 per hour) beyond the minimum wage. We kept the annotators' demographic information confidential and only release the final format of the dataset. The contents of this dataset are writing in a formal style and in English. Parallel sentence pairs (before and after revision) are included with human-annotated labels. We assign both token-level labels, indicating whether a word/phrase contains bias as well as a sentence-level label that reflects the statement's neutrality. To guarantee the dataset's quality and avoid potential problems brought by the annotators, thorough training sessions and discussions with domain experts were performed at the early stage. Periodic discussions on annotations results and embedded double-annotated questions were also included for quality control.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2933–2943.

[8] https://creativecommons.org/licenses/by-sa/3.0/us/

[9] http://www.gnu.org/licenses/fdl-1.3.html

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? Subjective acts and intents in conversational discourse. *arXiv preprint arXiv:2104.04470*.

Ajda Gokcen, Evan Jaffe, Johnsey Erdmann, Michael White, and Douglas Danforth. 2016. A corpus of word-aligned asked and anticipated questions in a virtual patient dialogue system. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3174–3179.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1631–1640.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics*.

Joan Bybee Hooper. 1975. *On assertive predicates.*, volume 4. Syntax and Semantics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1587–1596.

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

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pages 195–203.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1113–1122.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38.

Diederik P Kingma and Jimmy Ba. 2015. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2676–2686.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics, System Demonstrations*, pages 67–72.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Wouter Leeftink and Gerasimos Spanakis. 2019. Towards controlled transformation of sentiment in sentences. *arXiv preprint arXiv:1901.11467*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1865–1874.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. 2017. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918.

Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1153–1161.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1064–1074.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):1–18.

Gabriel Murray and Giuseppe Carenini. 2009. Predicting subjectivity in multimodal conversations. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1357.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271–278.

Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. 2020. Towards detection of subjective bias using contextualized word embeddings. In *Companion Proceedings of the Web Conference 2020*, pages 75–76.

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.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 480–489.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 129–140.

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*, pages 1650–1659.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

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.

Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Geoff Thompson. 2005. Ken hyland, metadiscourse: Exploring interaction in writing. *Language in Society - LANG SOC*, 37.

Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, and Nikolai Liubimov. 2020. Label Studio: A swiss army knife of data labeling and annotation tools. Open source software available from https://github.com/heartexlabs/label-studio.

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

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 812–817.

Janyce Wiebe, Rebecca Bruce, and Thomas P O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 246–253.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Theresa Wilson and Stephan Raaijmakers. 2008. Comparing word, character, and phoneme n-grams for subjective utterance recognition. In *Ninth Annual Conference of the International Speech Communication Association*, pages 1614–1617.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2899–2914.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.

Fabio Massimo Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–36.

A Annotation Interfaces

A.1 Word/phrase Classification Interface

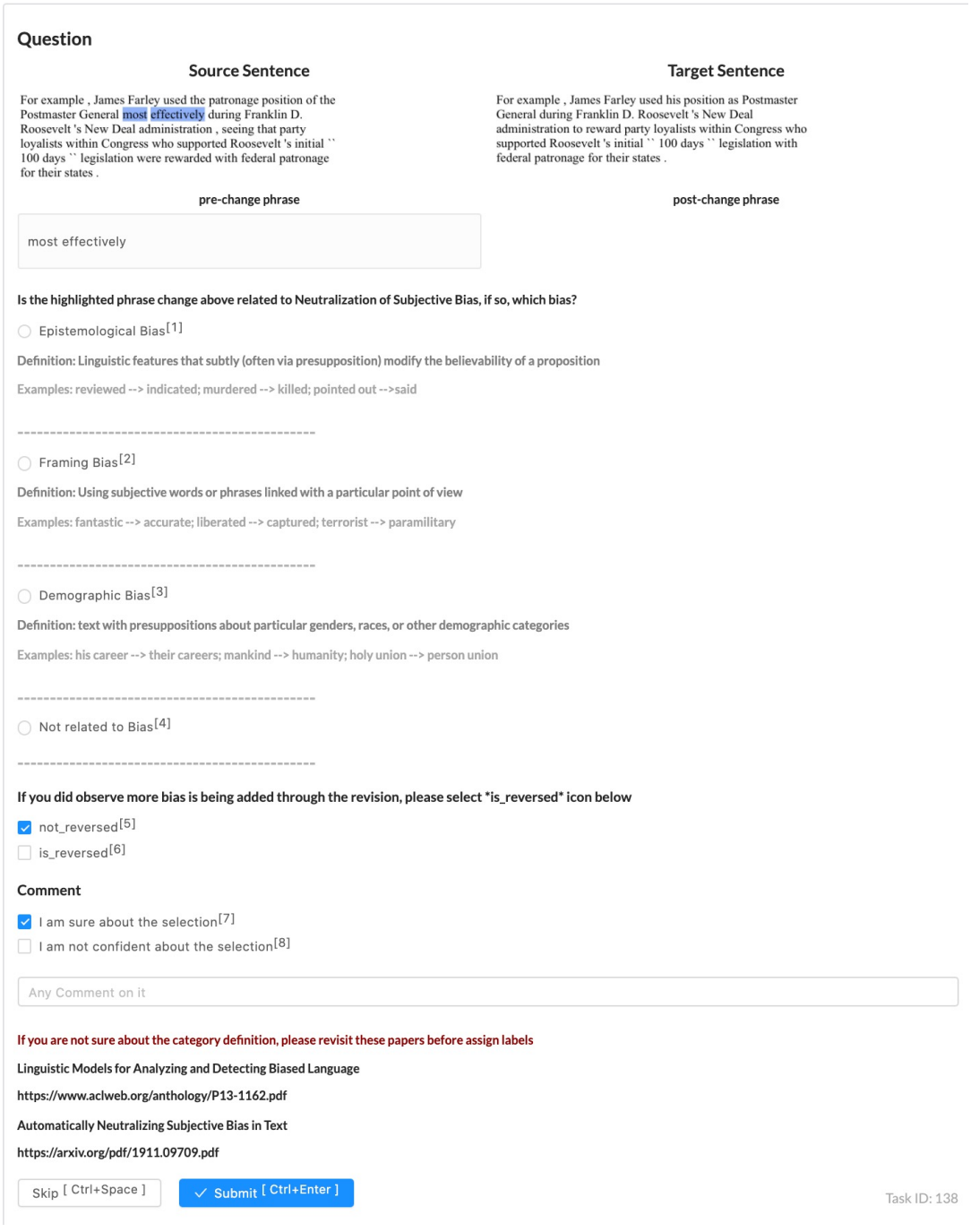

Figure 2: Annotation interface for Bias Classification of word/phrase edits.

A.2 Generation Classification Interface

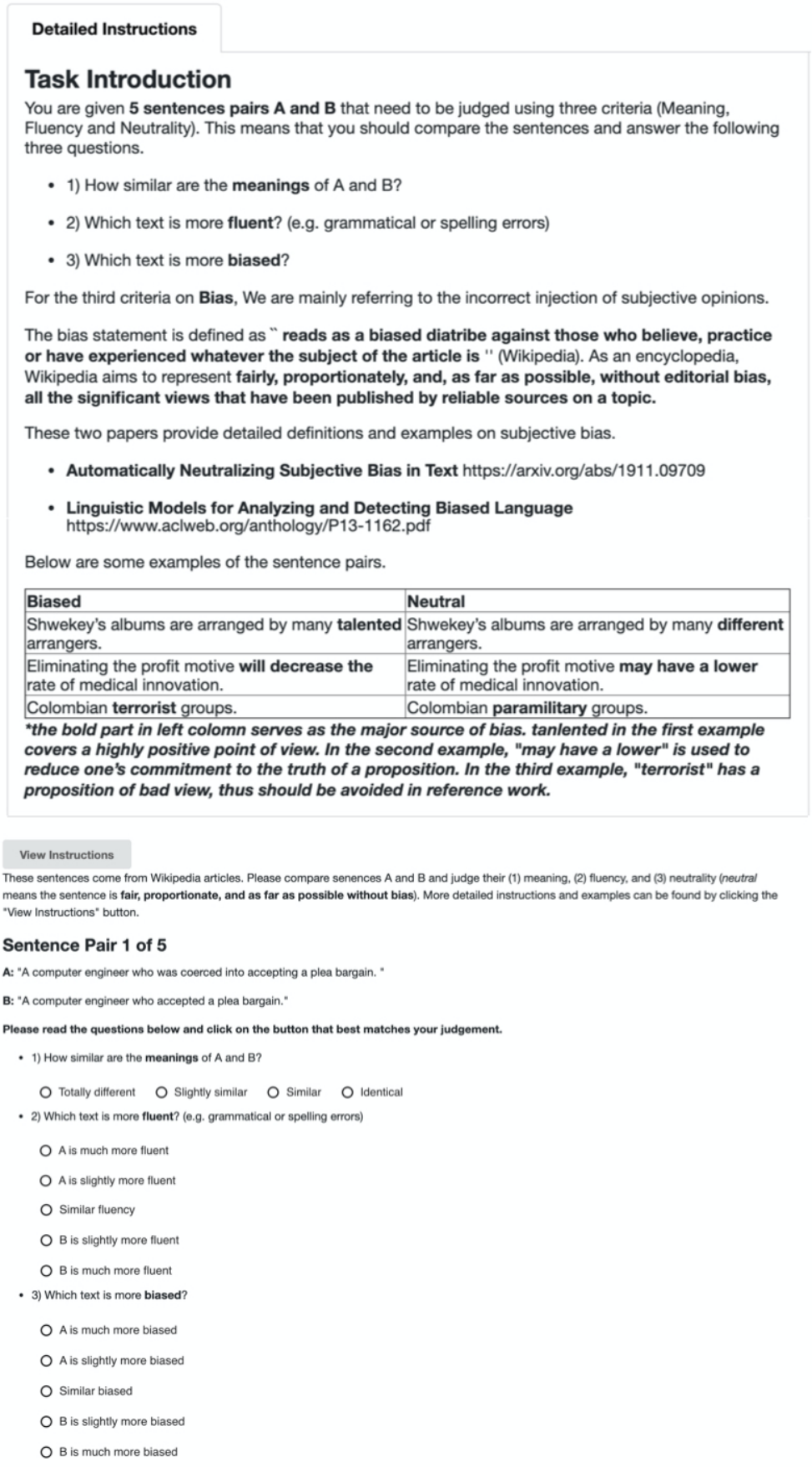

Figure 3: Annotation guidelines for the evaluation on text generation results along with a example question.

B Annotation Details

B.1 Manual Annotation Training

For both stages of the annotation task, annotators were asked to read the definition and task description with examples carefully, and then had discussions with the authors to share their understanding of the material. The annotators were then trained on a subset of the WIKIBIAS dataset (499 sentence pairs) with detailed instructions. For instance, for the bias type classification task, annotators were encouraged to leave comments that support their selection. Meanwhile, our annotation interface (Figure 2) provided the definition and multiple examples of each bias. The training set was split into 4 batches, each took 5 days to annotate and 2 days for discussion and revision on labels. The whole training process took 30 days until reasonable agreement was achieved, with each annotator having 4 rounds of discussions with the authors. We release the annotation guidelines with our source code and dataset.

B.2 Amazon M-Turk Details

To assess the quality of different generation models, we set up tasks on assessing the quality of sentence pairs.

We settle on a task design as follows: Annotators are told that we are collecting their judgments of the quality difference between a sentence pair on three perspectives: Fluency, Neutrality, and Meaning. we then show an instruction page and guide them through 4 practical trials with true answers. They then are asked to annotate on a qualification test set with 5 trial items. Qualified annotators who passed the test (we asses the annotators' results with gold answers and filter out annotators with incomplete submission or error rate above 20%) are then asked to continue with the main trail items. In the end, 100 annotators passed the test.

We sample 100 sentence pairs for each individual model. For each pair, we recruit 3 individual annotators to do the work. We include the annotation task interface and guidelines in Figure 3. We also restrict annotators whose IP address is in the U.S., who have a minimum HIT approval rating of 98% and a minimum of 1,000 HITs approved. We do not collect specific demographic information of the annotators. The price item was set to ensure that workers were paid ($10 - $13 per hour) beyond the minimum wage.

C Rule-based System

Given a parallel sentence pair, we utilize the diffs[10] as a starting point. In detail, the package returns a list of edit tuples.[11]

We apply different rules for varying scenarios. For sentence pair with one single-/multi- word phrase change, we match nearby edit in the extracted diffs with "-" and "+" signs as *substitution* edit pairs and leave else as one *deletion* and one *addition*. This is inspired by the observation that people would replace the old word/phrase with a new one in the same location. Note that we also apply several cleaning rules to filter out non-bias-related modifications such as spell correction.

For sentence pairs with multiple word/phrase changes, similar to the single edit extraction, we first aim at extracting all *substitution* cases. However, due to the complexity of the multiple changes, even neighboring changes can be non-related. We also find that several phrase pairs are broken into multiple pieces due to the duplicated prepositions and determinants. To handle such cases, we first parse the raw output of the diffs and reconnect the disjoint pieces into complete continuous phrases. We then use a constituency parser (Kitaev and Klein, 2018) to check whether two candidate changes belong to the same type of sub-tree. For the remaining changes, we greedily compute the similarities between the edit pairs in the pre and post-edited sentence, then utilize a threshold tuned on the dev set to construct more *substitutions*. In the end, we label the remaining without alignments as *deletion* or *addition* accordingly.

D Implementation Details

All our experiments are run on NVIDIA TITAN X GPUS. BERT-based models pre-trained on $\text{Train}_{\text{auto}}$ take on average of 2 hours for each epoch and 5 mins per epoch for $\text{Train}_{\text{manual}}$ fine-tuning.

D.1 Classification

For classification tasks, we use `bert-base-uncased` model and Adam (Kingma and Ba, 2015) for optimization. We utilize the sentence representations embedded in the [CLS] token, then project it with a weight matrix $W \in R^{dx2}$ and We

[10] Following the work of (Pryzant et al., 2020), we use the simplediff package to extract diffs

[11] i.e. [("=", [The Irish economy]), ("-", [has benefited from]), ("+", saw) ...] in Figure 1.

jointly fine-tune the language model and classification parameters. Each model is fine-tuned with a maximum of 3 epochs, batch size of 16, learning rate of 2e-5, gradient clip of 1.0, and no weight decay. We set the maximum sequence length 128. We save the checkpoint after each epoch and pick the model with best performance on dev set for final evaluation. We trained the model which only used $\text{Train}_{\text{manual}}$ for 5 epochs. For the two step fine-tuning, We further fine-tuned the pre-trained models on $\text{Train}_{\text{manual}}$ with 3 epochs.

D.2 Tagging

For BiLSTM-CNN-CRF, we kept most parameters consistent with the original paper[12] (Ma and Hovy, 2016) with a grid search on learning rate between [1e-3, 1e-2, 5e-3] and batch size between [10,16,32]. The reported results are experimented with a learning rate of 1e-3 and batch size of 16. For DETECTOR model, following the setup in (Pryzant et al., 2020), we trained the tagging model[13] on a portion of the WIKIBIAS-AUTO corpus which only covers single-word edit and report results with the selection of top-1 possible word based on token possibility. We implemented all BERT tagging models with `bert-base-cased` checkpoint and optimized with Adaw (Loshchilov and Hutter, 2019). We used a learning rate searched in [3e-5, 5e-5], a warmup rate of 0.1, a batch size of 16 and trained each model for 3 epochs. We trained the model which only used $\text{Train}_{\text{manual}}$ for 5 epochs. For the two step fine-tuning, We further fine-tuned the pre-trained models on $\text{Train}_{\text{manual}}$ with 3 epochs. For the Joint Model, We tuned the weights of classification losses for positive and negative instances. We searched from the combination of [(1,1), (0.5, 0.5), (0.6, 0.4), (0.7, 0.3)] using the dev set and report the result on the test set with the best setting $\alpha = 1$ and $\beta = 1$. For $\text{BERT}_{\text{Attention}}$, we use encoder from the best performed classifier (§3.1.1). For each layer in layers 9-12, we look at the attention scores aggregated towards each token and pick the target tokens based on a threshold tuned on dev set as the candidate for tagging. We aggregate overall 12 heads. We further experiment with 4 different methods of computing the attention. The first two are token_count and word_count, where we sum up the times a token/word obtains the highest attention score from the other tokens. Besides the counts, we also try to directly employ the attention score, either using the average score out of 12 heads or the sum of the scores.

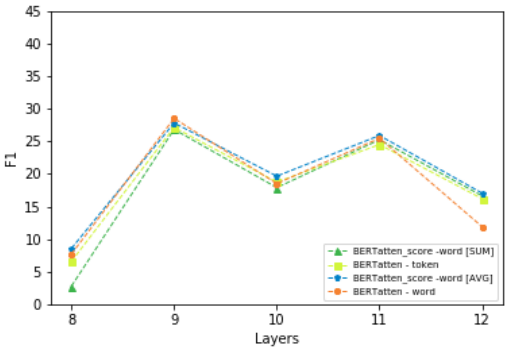

Figure 4: Test set Exact Match F1 of $\text{Bert}_{\text{Attention}}$ models with different layers

Previous work (Clark et al., 2019) shows that heads often attend to "special" tokens, so we excluded special tokens such as ['CLS'] and ['SEP'] as well as ending period from the candidates pool. We examined on layer 9-12 with the observation that layers below layer 8 gave much poorer performance. This is in consistent with previous work's finding that different layers of BERT capture diverse perspectives of information in the text, while higher level tend to cover more semantic information. As shown in Figure 4, we report the performance of the 9th layer's word-count based method in Table 7.

D.3 Generation

When we use generation models for neutralizing bias, we adapted OpenNMT (Klein et al., 2017) for LSTM and Attention-based Seq2seq and CopyNet baselines. We also used fairseq (Ott et al., 2019) to implement Transformer and BART model. For Seq2Seq model, we use default setting in OpenNMT and a SGD optimizer with a learning rate of 0.5. For Seq2Seq model, we use the default setting in OpenNMT and a SGD optimizer with a learning rate of 0.5. For CopyNet, we reuse the attention as copy attention, and we also use a SGD optimizer with a learning rate of 1. For BART model, we used `BART-large` and an Adam optimizer. We use a polynomial leaning rate scheduler with 500 warmup steps and 3e-5 max learning rate. We also use 0.1 dropout and 0.1 label smoothing. The setting of Transformer is the same as BART except that Transformer architecture is randomly initialized.

[12] https://github.com/XuezheMax/NeuroNLP
[13] https://github.com/rpryzant/neutralizing-bias