POWERTRANSFORMER:
Unsupervised Controllable Revision for Biased Language Correction

Xinyao (Michelle) Ma[⋆ ◇] **Maarten Sap**[⋆ ◇] **Hannah Rashkin**[◇] **Yejin Choi**[◇†]

◇Paul G. Allen School of Computer Science & Engineering, University of Washington
†Allen Institute for Artificial Intelligence
Seattle, USA
{max36,msap,hrashkin,yejin}@cs.washington.edu

Abstract

Unconscious biases continue to be prevalent in modern text and media, calling for algorithms that can assist writers with bias correction. For example, a female character in a story is often portrayed as passive and powerless ("*She daydreams about being a doctor*") while a man is portrayed as more proactive and powerful ("*He pursues his dream of being a doctor*").

We formulate **Controllable Debiasing**, a new revision task that aims to rewrite a given text to correct the implicit and potentially undesirable bias in character portrayals. We then introduce POWERTRANSFORMER as an approach that debiases text through the lens of *connotation frames* (Sap et al., 2017), which encode pragmatic knowledge of implied power dynamics with respect to verb predicates. One key challenge of our task is the lack of parallel corpora. To address this challenge, we adopt an unsupervised approach using auxiliary supervision with related tasks such as paraphrasing and self-supervision based on a reconstruction loss, building on pretrained language models.

Through comprehensive experiments based on automatic and human evaluations, we demonstrate that our approach outperforms ablations and existing methods from related tasks. Furthermore, we demonstrate the use of POWER-TRANSFORMER as a step toward mitigating the well-documented gender bias in character portrayal in movie scripts.

1 Introduction

Narratives and news texts often reflect societal biases and stereotypes, such as the traditional gender role that women are passive and submissive (Lakoff, 1973; Fiske, 1993; Fast et al., 2016). The task of *controllable text revision*, i.e., rephrasing text to a targeted style or framing, can help correct for these biases by altering and equalizing the way

---

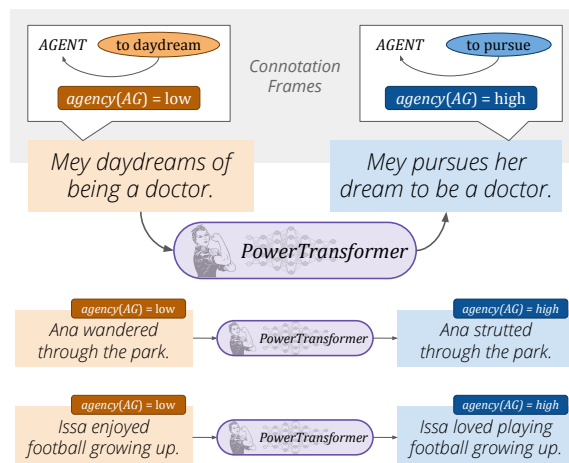⋆ Both authors contributed equally.

Figure 1: Examples of using connotation frames (Sap et al., 2017) for controllable revisions to portray characters with more agency and power. In the second example, "Ana strutted" implies that she is more active and decisive, compared to "Ana wandered" which portrays her as aimless and passive.

people are described. For example, automatically rewriting "*Mey daydreamed about being a doctor*" as "*Mey pursued her dream to be a doctor*" portrays Mey with more authority and decisiveness (Figure 1). Such controllable revision methods could be used to help reshape how gender roles are portrayed in media (e.g., through machine-in-the-loop writing systems; Clark et al., 2018).

To edit such biases out of text, a controllable rewriting model faces three key challenges. First, a model should be able to make edits beyond surface-level paraphrasing, as simple paraphrasing will often not adequately debias the underlying events described. For example, Mey's portrayal in Figure 1 carries both overt bias (the choice of action) and subtle bias (the framing of the action), both of which require rewriting to be adequately debiased. Second, a model's debiasing revisions should

be purposeful and precise and should not make unnecessary changes to the underlying meaning of the original text. Lastly, since parallel data does not exist, models must learn to revise and debias text without supervised data, thereby preventing straightforward machine translation-style modelling.

We formulate *Controllable Debiasing* as a new controllable text revision task that aims to correct the implicit and possibly unwanted bias against or towards a specific character portrayed in text (§2). As shown in Figure 1 (top), we study the portrayal biases through the lens of connotation frames of *power and agency* (Sap et al., 2017), which provide pragmatic knowledge about implied power and agency levels projected onto characters by a predicate.

We create POWERTRANSFORMER, an encoder-decoder model that rewrites sentences with a desired portrayal using agency connotation frames (§3). We combine a reconstruction and paraphrase objective into our model to overcome the lack of parallel supervised data, building off of the denoising autoencoder setup from Li et al. (2018a). To steer the revisions, we endow the model with connotation frame knowledge both at training time using control tokens, and at generation time using agency-based vocab boosting.

Our findings show that POWERTRANSFORMER is effective at rewriting sentences with desired agency connotations while only making minimal changes to their meaning, as measured through both human and automatic evaluations (§4). We also show that POWERTRANSFORMER significantly outperforms existing stylistic rewriting methods (Prabhumoye et al., 2018; Dathathri et al., 2020) on those metrics. Additionally, through ablations studies, we establish the usefulness of each component of the model, finding benefits from both the joint objective (47% gain in accuracy) and the agency scaling (12% gain in accuracy).

Finally, in §5, we apply Controllable Debiasing to a corpus of modern English movies (Gorinski and Lapata, 2015) as a step towards removing gender bias in character portrayal established by prior work (Sap et al., 2017). Using POWERTRANSFORMER, we revise the movie scripts and significantly increase the agency levels of female characters, thereby reducing the gender bias. Our findings show promise for using modern NLP tools to help mitigate societal biases in text. We release our preprocessed data and code at http://maartensap.com/controllable-debiasing.

2 Controllable Debiasing

Controllable Debiasing is a novel formalization of stylistic rewriting that aims to debias the portrayal of characters through controllable revision. To achieve the desired character portrayal, a system must be able to change the underlying meaning of events, unlike certain formalizations (e.g., politeness transfer; Rao and Tetreault, 2018) where full meaning preservation is required. Without this, systems run the risk of merely paraphrasing the biases in text. However, revisions must be precise and avoid unnecessary meaning changes, which can often occur in stylistic rewriting (e.g., reversing the sentiment of a review drastically changes its underlying meaning).

For our new rewriting task of changing portrayal bias, we focus on connotation frames that measure the *power* and *agency* ascribed to characters through the actions they take. Connotation frames (Rashkin et al., 2016; Sap et al., 2017) distill implicit relations between a verb, its agent, and its theme. In this work, we use the positive, neutral, and negative agency dimensions, where agency is defined as the capacity to intentionally make changes or act upon one's environment (Dennett, 1989). For example, illustrated in Figure 1, "X pursued Y" implies that X has positive agency.[1] Using machine-in-the-loop writing systems (e.g., Ghazvininejad et al., 2016, 2017; Clark et al., 2018, Textio[2]), models trained on this task could help authors write news, stories, or movies that portray characters in less biased ways, and thereby help mitigate the negative effects of stereotypical portrayals in media (Behm-Morawitz and Mastro, 2008; Field et al., 2019).

3 POWERTRANSFORMER

We present a new approach for Controllable Debiasing called POWERTRANSFORMER, which addresses two key challenges: the paucity of parallel supervised data for training and the difficulty of incorporating fine-grained control for steering the agency of the output. Our approach (Figure 2) jointly learns to reconstruct partially masked story

---

[1]Future work could explore using the power dimension instead of agency, or alternative operationalizations of biases, e.g., Social Bias Frames (Sap et al., 2020) or *regard* towards minorities as introduced by Sheng et al. (2019).
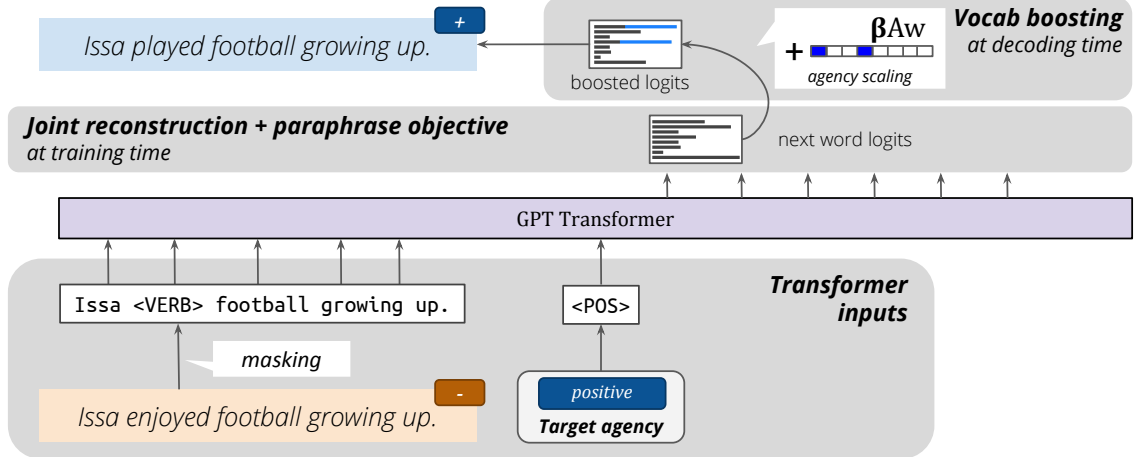
[2]https://textio.com/

Figure 2: Overview of the full POWERTRANSFORMER model. An input sentence is masked for verb tokens indicative of agency. Masked inputs and target agency are used as GPT inputs. We use a joint objective using both paraphrase data and masked input sentences for training. At decoding time, we employ a vocab boosting technique to steer generations towards the target agency.

sentences while also learning to paraphrase from an external corpus of paraphrases (§3.2). At generation time, we also include a boosting method for fine-grained steering towards the desired agency level as described in §3.3.

3.1 Model Overview

POWERTRANSFORMER is an encoder-decoder style model with an OpenAI-GPT transformer model (Radford et al., 2018) as the base. The input sentence $\mathbf{x}$ is converted to a sequence of byte pair encodings (BPE) $\{x_1, ..., x_n\}$, and given to the encoder after being scrubbed of its agency markers as described below. To steer the model, we also give the encoder the target agency t, which we represent as one of three special tokens $\{<\text{Pos}>, <\text{Equal}>, <\text{Neg}>\}$.[3]

3.2 Joint Objective

We train our model on both a reconstruction and a paraphrasing task, for which inputs are masked and paraphrased versions of the output, respectively.

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{para}} \qquad (1)$$

Masking and Reconstructing Inspired by the delete-retrieve-generate model from Li et al. (2018a), this objective teaches the model to recover masked out agency-associated verbs in sentences. We first assign an agency level to an input sentence by counting verbs in the agency lexicon from Sap

et al. (2017).[4] Then, we mask out all verbs indicative of the agency level, replacing them with a special $<\text{VERB}>$ token. In this setup, the target output is the original sentence $\mathbf{x} = \{x_1, ..., x_n\}$, with the masked sentence $\hat{\mathbf{x}}$ and the target agency level t as inputs. During training, we minimize the cross entropy of the target output sentence given the inputs:

$$\mathcal{L}_{\text{recon}} = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i | x_{<i}, \hat{\mathbf{x}}, t) \qquad (2)$$

Paraphrasing To go beyond reconstructing sentences, we add a paraphrasing objective using an out-of-domain paraphrase corpus (§4.1). We extract agency levels for each sentence and its paraphrase and mask out the agency verbs in the input, using the same methods as described above. Here, the inputs are the masked sentence $\hat{\mathbf{x}}$ and the target agency t, while the target output $\mathbf{y} = \{y_1, ..., y_m\}$ is the paraphrase. As with reconstruction, we minimize the cross entropy of the target output given the inputs:

$$\mathcal{L}_{\text{para}} = -\frac{1}{m} \sum_{i=1}^{m} \log p(y_i | y_{<i}, \hat{\mathbf{x}}, t) \qquad (3)$$

3.3 Controlled Decoding with Vocab Boosting

We employ a vocab-boosting technique during generation to encourage models towards generating with the desired agency, inspired by Ghosh

---

[3]In earlier experiments, we also provided the original agency as an input to the model during training and decoding, but found that it made little difference in performance.

[4]For sentences that have multiple verbs, we assign the agency level that the most verbs in the sentence have (e.g., a sentence with two positive agency verbs and one negative agency verb will be assigned positive agency).

	Type	# Instances	Pos	Neutral	Neg
ROC	train	10721	3834	4151	2736
	dev	1803	633	710	460
	test	899	325	350	224
Para.	train	45000	16410	14153	14437
	dev	10000	3645	3328	3127

Table 1: Statistics for our main story sentences dataset (ROC) and for the external paraphrase corpus (Para.).

et al. (2017). At each decoding timestep i, we rescale the unnormalized token probabilities (logits $l_i \in \mathbb{R}^V$, where V is the vocabulary size) to boost the likelihood of predicting words with the target agency. The next token probabilities are then computed using the "boosted" logits:

$$P(y_i | y_{<i}, x, t) \propto \text{softmax}(l_i + \beta \cdot Aw) \quad (4)$$

where A is a $\mathbb{R}^{V \times 3}$ matrix that represents a 3-dimensional {positive, equal, and negative} agency embedding for each token in the vocabulary, w is a $\mathbb{R}^3$ one-hot vector denoting the target agency for the output, and β is a scalar hyperparameter representing the boosting strength. We create A manually using the verbs in the agency lexicon (Sap et al., 2017).[5] Used only at decoding time, this method effectively increases the likelihood of using a word with the target agency level.

4 Controllable Debiasing Experiments

In this section, we describe three experiments for investigating POWERTRANSFORMER performance. First, we evaluate performance of our full model and ablated baselines, using automatic metrics to quantify the effectiveness of each modelling component (§4.4). Next, we compare our full model to baselines from related work (§4.5). Lastly, given the limitations of automated metrics for evaluating generations (Liu et al., 2016; Mir et al., 2019), we obtain human judgments of model performance through crowdsourcing (§4.6). We additionally include examples of generations in Table 4.

4.1 Datasets

In our experiments, we use a dataset of short stories for the reconstruction task and a parallel corpus of

paraphrases for both paraphrase and reconstruction tasks. We show data statistics in Table 1, with additional preprocessing details in Appendix A.

ROC story corpus The main focus of our study is controllable revision of story sentences; therefore, we select sentences from the ROC story corpus (ROC Mostafazadeh et al., 2016). After extracting agency levels for all sentences from the training stories, we sample roughly equal amounts of all three agency levels, and randomly split sentences into training, development, and test sets.[6]

Paraphrase corpus As additional training data, we use the corpus of automatically aligned paraphrases of TV subtitles (Creutz, 2018, Para.). As with the ROC story corpus, we extract agency levels for each sentence and its paraphrase, then sample roughly equal amounts of pairs with all different sentence-paraphrase agency combinations (further details in §A.2). We randomly split the data into 45k train and 10k dev. instances (Table 1).[7]

4.2 Metrics

In addition to human evaluations, we also use a variety of automated evaluation metrics to characterize different aspects of performance. We measure the accuracy of the change in agency by comparing the target agency level with that of the output (extracted using the connotation frames lexicon). As a measure of meaning preservation, we use BERT-score F1 metrics (Zhang et al., 2020) to compare the semantic similarity of the input sentence with the machine output.

As additional metrics, we measure the fluency, the repetitiveness, and diversity of the output. Following previous work (Dai et al., 2019), we measure fluency as perplexity (*PPL*) of the output sentence using a pre-trained GPT model that has not been fine-tuned for this task. As an additional metric of potential text degeneration, we compute the fraction of output sentences that have a bigram that is repeated two or more times (*w/ rep*). Finally, we compute the fraction of generations that are unique with respect to the rest of the output, to ensure diverse, input-specific generations (*unique*).

---

[5] Since our model operates on BPE tokens, we manually set the first BPE token of every tense of every verb to the desired agency. We also experimented with learning A from data, but found no improvement over manually setting it.

[6] We use a 80:13:7 train, development, test ratio.

[7] Since this is just additional training data, we do not test our models on this corpus, but do use the dev. set for selecting some hyperparameters.

Ablations using the Development Set

POWERTRANSFORMER variants	Main Metrics		Additional Metrics		
	Agency Acc (↑)	**Meaning** BertScore (↑)	**Fluency** PPL (↓)	**Repetition** w/ Rep (↓)	**Diversity** Unique (↑)
(*ParaOnly+noBoost*)	.30	.95	**58.76**	.002	.54
(*ParaOnly+Boost*)	.42	.90	76.25	**.001**	.59
(*Joint+noBoost*)	.77	**.96**	70.61	.007	.87
(*Joint+noBoost*)+*SupplyVerb*	.77	**.96**	94.54	.004	.92
FULL = (*Joint+Boost*)	**.89**	**.96**	76.78	.015	**.99**

Table 2: Ablation study results on the development set. We present separate metrics for evaluating the change in agency, the meaning preservation, fluency, repetitiveness and diversity of the output (bolding the best performance). (↑) indicates that higher is better and (↓) indicates that lower is better.

4.3 Experimental Setup

We randomize ROC story and paraphrase data, and use OpenAI GPT LM as our pretrained model. For decoding, we use top-p=0.4 nucleus sampling (Holtzman et al., 2020), and a boosting strength of β=5 (hyperparameters and details in §B.1).

4.4 Investigating Effectiveness of Approach

We first establish our model's effectiveness at Controllable Debiasing on our dev. set, and investigate the importance of various components in our approach through ablation analyses. For qualitative analyses, we also show example revisions in Table 4 (and Table 6 in the appendix).

4.4.1 Ablated Baselines

We first investigate the importance of the reconstruction objective, by comparing our joint objective model (*Joint*) with a model trained with just the paraphrasing objective (without masking, *ParaOnly*). Then, to quantify the effect of boosting, we compare models with (*Boost*) and without (*noBoost*) agency-specific vocab boosting. Note that *ParaOnly+noBoost* is equivalent to a GPT-based encoder-decoder model, similar to seq2seq frameworks commonly used in paraphrasing tasks (Cao et al., 2017; Li et al., 2018b; Prakash et al., 2016).

As a final comparison, we implement a model variant that more closely mirrors the delete-retrieve-generate paradigm (Li et al., 2018a) by adding a "retrieve" step in which we concatenate transformer input with a verb retrieved from the verb agency lexicon that is most similar to the masked out verb (*SupplyVerb*).[8]

4.4.2 Results

In Table 2, our results show that the full model (*Joint+Boost*) yields text revisions with the most accurate target agency and the most meaning preservation. In general, we find that both the joint objective and vocab boosting (*Boost*) substantially increase the target agency accuracy, as also illustrated in examples (d) and (e) in Table 4. However, unsurprisingly, vocab boosting also slightly lowers fluency, yielding higher perplexities than models' non-boosted counterparts. Our results also show that using the joint objective with boosting increases the diversity of output, but causes marginally more repetition of bigrams.

Counterintuitively, our ablations show that supplying a verb to the model as an explicit retrieval step (*SupplyVerb*) does not improve the agency or meaning metrics and actually hurts the fluency of the output (as measured by higher perplexities). Upon qualitative investigation (Table 6 in the appendix), the retrieved verb is often related to a different word sense of the masked verb, breaking the grammaticality of the sentence.

4.5 Comparison with External Approaches

To further validate our approach, we compare against two baselines from related style transfer and stylistic generation tasks. As these models were designed for binary style transfer, we only report our baseline and model results on the positive and negative agency portions of our data.

---

[8] We retrieve a verb from the Sap et al. (2017) lexicon that has the target agency and is most similar to the masked out verb, where similarity is defined as cosine distance between word embeddings using GloVe 300-d embeddings (Pennington et al., 2014).

Test Set Comparisons (pos-to-neg and neg-to-pos set)

	Main Metrics		Additional Metrics		
	Agency Acc (↑)	**Meaning** BertScore (↑)	**Fluency** PPL (↓)	**Repetition** w/ rep (↓)	**Diversity** unique (↑)
PPLM (Dathathri et al., 2020)	.13	.95	106.12	.053	**1.00**
BST (Prabhumoye et al., 2018)	**.88**	.83	**91.22**	.053	0.79
POWERTRANSFORMER	.86	**.96**	95.19	**.015**	**1.00**

Table 3: Performance of different re-writing methods on the neg-to-pos and pos-to-neg subsets of the test set (bolding the best performance). We evaluate the change in agency and the meaning preservation. As secondary metrics, we include fluency, repetitiveness, and diversity of the output.

4.5.1 Baselines

BST We compare to the backtranslation style transfer model from Prabhumoye et al. (2018). This model first translates input sentences to a pivot language (preserving the meaning but losing language-specific style), then relies on style-specific decoder-translators for generating the output sentence. We include set-up details in §B.3.

PPLM Recent work in controllable generation has introduced PPLM, a new plug-and-play technique with promising results for decoding stylistic text (Dathathri et al., 2020). This method operates on an underlying neural language model at decoding time. It uses backpropagation from a stylistic discriminator to update the past and present hidden representations to be more consistent with the targeted style or domain. We adapt the approach to controllable revision by replacing the base language model with an autoencoder trained on a reconstruction objective, described in detail in §B.2.

4.5.2 Results

We present results in Table 3. Our experiments show that POWERTRANSFORMER performs better than the baselines overall. Specifically, while the BST revisions obtain slightly higher accuracy on the output agency levels, these revisions have the both the lowest diversity and meaning preservation, suggesting the model ignores the input (Table 4). PPLM shows opposite trends, yielding the lowest accuracy with high meaning preservation and high diversity of generations. Illustrated in Table 4, this model often makes less purposeful and less concise alterations.

4.6 Evaluating with Human Judgements

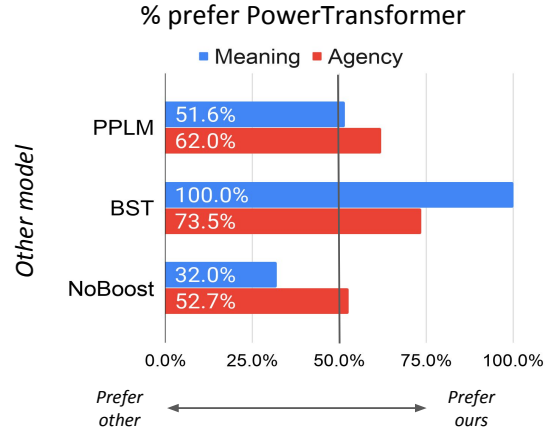To validate our automatic evaluations, we collect human judgments of the controllable revisions

Figure 3: Human judgements of target agency and meaning preservation in POWERTRANSFORMER vs. three other model variants. Selection rates >50% indicate preference towards our model.

from several baselines and POWERTRANSFORMER (*Joint+Boost*).

4.6.1 Human Evaluation Task

We design a head-to-head[9] crowdsourcing task on Amazon Mechanical Turk where we ask raters to compare two outputs from different models given the same input sentence and target agency (see Figure 5 in the appendix). We first ask them to judge whether either output is gibberish, then, in two questions, choose which revision has better targeted agency and which better preserves the meaning of the original sentence. For consistency, each pair is rated by three judges. To ensure the quality of our evaluations, we selected workers who could reliably distinguish high from low agency sentences in a qualification task (see Figure 6 in

---

[9]We use head-to-head evaluations as those have been shown to be more reliable than scale-rating evaluations (Kiritchenko and Mohammad, 2017).

dir.		Input	Model	Revised Sentence (out)	Agency(out)
	(a)	After the party I headed home.	PPLM	after the party my classmate, Kayla and the Tgirls of the Universe.	=
			BST	please 's , i have a word of this .	−
			POWERT$_{Joint+NoBoost}$	after the party i stayed home.	−
			POWERT$_{Joint+Boost}$	after the party i stayed home.	−
	(b)	A friend asked me to watch her two year old child for a minute.	PPLM	a friend asked me to watch her two year old child for a minute.	+
			BST	l didn 't have a word of this , you 're .	−
			POWERT$_{Joint+NoBoost}$	a friend needed me to watch her two year old child for a minute.	−
			POWERT$_{Joint+Boost}$	a friend needed me to watch her two year old child for a minute.	−
	(c)	After filling in the data it looked quite sharp.	PPLM	before filling the last question it it it it looked quite sharp. before filling the last question it it	+
			BST	when the 't you want a word ?	−
			POWERT$_{Joint+NoBoost}$	after analyzing in the data it looked quite sharp.	=
			POWERT$_{Joint+Boost}$	after seeing in the data it seemed quite sharp.	−
	(d)	Allie was failing science class.	PPLM	allie was failing science grade.	−
			BST	do you want me ?	+
			POWERT$_{Joint+NoBoost}$	allie was failing science class.	−
			POWERT$_{Joint+Boost}$	allie was taking science class.	+
	(e)	Darla wanted a soft drink.	PPLM	darla wants a hard hard drink.	−
			BST	don 't take me a man .	+
			POWERT$_{Joint+NoBoost}$	darla ordered a soft drink.	+
			POWERT$_{Joint+Boost}$	darla ordered a soft drink.	+
	(f)	Clint paused on the trail.	PPLM	clint was on the trail.	=
			BST	don 't you want me ,	−
			POWERT$_{Joint+NoBoost}$	clint hiked on the trail.	=
			POWERT$_{Joint+Boost}$	clint walked on the trail heading down.	+

agency (+ → −) spans rows (a)-(c); *agency (− → +)* spans rows (d)-(f).

Table 4: Example sentences from our dev. set, along with their revisions from various models and the achieved agency levels (Agency(out)). Examples (a)-(c) should be rewritten from high to low agency, and (d)-(f) from low to high agency. Confirming our quantitative results in Tables 2 and 3, POWERTRANSFORMER (*Joint+Boost*) is the most effective at making purposeful and precise changes to the input sentences to alter their agency while minimally changing their meaning. Revisions from more models are listed in Table 6 (in the appendix).

the appendix).

For this evaluation, we generate three revisions–one for each target agency level–for a random subset of 100 test examples. We compare the output of our full POWERTRANSFORMER model with two external baselines (PPLM and BST). For further comparison, we also include the most competitive ablated baseline from Table 2 (i.e., *Joint+noBoost*).

4.6.2 Results

In Figure 3, we show the percentages of times in which POWERTRANSFORMER was preferred over the three baseline models.[10] Percentages >50% indicate a preference towards POWER-TRANSFORMER.

Overall, the sentence revisions by POWER-TRANSFORMER are preferred over all of the baselines in obtaining the desired agency level. For

meaning preservation, our model is always selected over BST, mirroring BertScores in Table 3. The difference is less stark when comparing to PPLM which sometimes makes no changes or irrelevant changes to the input sentence, and reversed when comparing to the ablated *noBoost*.

Additionally, BST revisions were marked as gibberish substantially more than those by other models (63% vs. 3-7%). While this seemingly contradicts BST's low perplexity scores, this is in line with previous work showing automatic fluency metrics can favor degenerate, bland, or repetitive language (Holtzman et al., 2020).

5 Gender Bias in Movies

As a proof-of-concept of Controllable Debiasing, we investigate whether gender biases in portrayals of movie characters can be mitigated using POW-ERTRANSFORMER.

---

[10]Judgments in our evaluation task had an average pairwise agreement of 75% (Krippendorf's α=.52).

5.1 Movie Scripts Corpus

We draw our data from the 767 modern English movie scripts by Gorinski and Lapata (2015), focusing on the narrations which describe characters and their actions (as opposed to the character's dialogue utterances). Described in further detail in §C in the appendix, we automatically extract characters and assign them a binary[11] gender (man, woman) using a list of highly gendered names (e.g., "Sarah", "William") and a list of gendered words (e.g., "waiter," "waitress"). Following previous work (Ramakrishna et al., 2017; Sap et al., 2017), we assign narration sentences to characters if their name appears in them.

Our corpus contains 16,763 characters from 767 different English movies. Of those characters, 68% are inferred to be men and only 32% to be women,[12] consistent with known gender skews in movie characters (Google, 2017). This bias in representation is also present at the narrative level. Specifically, female characters are only mentioned in $n_{narr,f} = 27$ narrations on average, compared to $n_{narr,m} = 34$ narrations for male characters (Cohen's $|d| = 0.13$, $p < 0.001$). Similarly, compared to their male counterparts, female characters are described in significantly fewer words ($n_{words,f} = 329$, $n_{words,m} = 435$, $|d| = 0.14$, $p < 0.001$) and with fewer verbs ($n_{verbs,f} = 41$, $n_{verbs,m} = 54$, $|d| = 0.13$, $p < 0.001$).

5.2 Debiasing Portrayal in Movies

Given the known bias that female characters are portrayed with less agency (Sap et al., 2017), our goal is to re-balance their agency levels to be more on par with those of male characters. Therefore, we revise only the sentences describing female characters to have higher agency, using POWERTRANS-FORMER. Then we extract connotation frames of agency for revised script sentences, and aggregate per character. Shown in Figure 4, revisions successfully increase the instances of positive agency of female characters, and decrease their negative agency or passiveness.

We further examine the change in gender association of positive and negative agency, to verify the
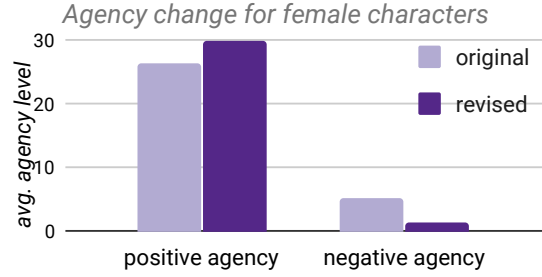

Figure 4: Average agency levels (i.e., number of agency verbs) for female characters in original and revised scripts. POWERTRANSFORMER can revise the portrayals of female characters in movies to give them higher positive agency and lower negative agency.

effectiveness of Controllable Debiasing. We first count all the positive and negative agency verbs used to describe characters (in original or rewritten sentences). Following Sap et al. (2017), we then fit a logistic regression model to quantify the association between character's gender with their agency levels, controlling for their number of words, verbs, and narrations. For better interpretation of the β coefficients, we z-score all the continuous variables.

We confirm that indeed, Controllable Debiasing using POWERTRANSFORMER can reverse the bias in portrayal in movies. In original scripts, male characters were portrayed with significantly higher positive agency ($\beta_{pos} = 1.2$, $p < 0.001$) and lower negative agency ($\beta_{neg} = -0.3$, $p < 0.001$) than female characters. However, our model successfully reverses this gender bias, portraying women with significantly more positive agency ($\beta'_{pos} = -62.6$, $p < 0.001$) and significantly less negative agency ($\beta'_{neg} = 8.7$, $p < 0.001$).

Our findings on movie scripts show the promise of using Controllable Debiasing to successfully mitigate gender biases in portrayal of characters, which could be extended to other domains (e.g., news or fiction, Field and Tsvetkov, 2019; Fast et al., 2016). Additionally, future work could consider alternative views of portrayal biases (e.g., "regard" or bias directed at different demographic groups; Sheng et al., 2019; Sap et al., 2020), or use more holistic views of gender roles (e.g., "masculine default" cultures; Cheryan and Markus, 2020).

6 Related Work

Controllable Debiasing is a new formalization of the unsupervised stylistic rewriting task, contrasting with supervised approaches which benefit from

---

[11]Note that gender is a social construct that goes beyond the man-woman binary (Lorber et al., 1991), however more inclusive analyses (e.g., with non-binary genders) are not possible given the limited information about the individuals mentioned in our data.

[12]There were 2597 characters for which the gender could not be inferred.

parallel corpora (e.g., Xu et al., 2012, 2015; Rao and Tetreault, 2018; Pryzant et al., 2020). In unsupervised settings, a majority of work has dealt with the dearth of parallel data by using encoder-decoder setups paired with discriminators to disentangle style from content and steer generations (e.g., Shen et al., 2017; Zhang et al., 2018; Fu et al., 2018; Yang et al., 2018; Niu and Bansal, 2018; Romanov et al., 2019; Dai et al., 2019; John et al., 2019) or backtranslation setups (Prabhumoye et al., 2018; Lample et al., 2018). In contrast, Li et al. (2018a) introduce a modular approach (later adapted to transformer models by Sudhakar et al., 2019) that relies on drop-in replacement of attribute markers followed by language correction. POWER-TRANSFORMER improves on this approach with an additional out-of-domain paraphrasing objective.

While a majority of related existing stylistic rewriting work defines style as sentiment (e.g., on reviews), a notable exception is Nogueira dos Santos et al. (2018), who use stylistic rewriting to make text less hateful or offensive. Similar in spirit, Controllable Debiasing is a novel formalization that aims to address and revise social biases expressed in text, but using the nuanced implications distilled in connotation frames of power and agency instead of binary offensiveness.

Our work also draws inspiration from controllable generation methods (e.g., Koncel-Kedziorski et al., 2016; Hu et al., 2017; Ficler and Goldberg, 2017). While those methods steer the generation output to contain desired attributes, controllable revision is constrained to revise an input sentence in addition to generating with desired attributes.

7 Conclusion

We introduce a new text revision task of Controllable Debiasing, to help debias the portrayal of characters through the lens of connotation frames of power and agency. To this end, we create POWERTRANSFORMER, a transformer-based encoder-decoder trained on a joint reconstruction and paraphrasing objective. Our approach demonstrates promising results to revise sentences with targeted power and agency, and outperforms ablations and baselines on both automatic and human evaluations. Finally, as a case study, we show the feasibility for Controllable Debiasing at debiasing the portrayal of characters in movie scripts. Our findings highlight the potential of neural models as a tool for editing out social biases in text.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL*.

Elizabeth Behm-Morawitz and Dana E Mastro. 2008. Mean girls? the influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs. *Journalism & Mass Communication Quarterly*, 85(1):131–146.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase. In *AAAI*.

Sapna Cheryan and Hazel Rose Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *IUI*.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *LREC*. Corpus available at http://urn.fi/urn:nbn:fi:lb-201804191.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *ACL*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Daniel Clement Dennett. 1989. *The intentional stance*. MIT press.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *ICWSM*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *EMNLP Workshop on Stylistic Variation*.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *ICWSM*.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *ACL*.

Susan T Fiske. 1993. Controlling other people. the impact of power on stereotyping. *American psychologist*, 48(6):621–628.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *EMNLP*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *ACL Demonstrations*.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *ACL*.

Google. 2017. Using technology to address gender bias in film. https://www.google.com/about/main/gender-equality-films/index.html.

Philip Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *NAACL*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *ACL*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *ACL*.

Rik Koncel-Kedziorski, Ioannis Konstas, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2016. A theme-rewriting approach for generating algebra word problems. In *EMNLP*.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-Attribute text rewriting. In *ICLR*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL*.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018b. Paraphrase generation with deep reinforcement learning. In *EMNLP*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Judith Lorber, Susan A Farrell, et al. 1991. The social construction of gender. *Newbury Park*, 5.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *NAACL*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*. Corpus available at https://www.cs.rochester.edu/nlp/rocstories/.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *TACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through Back-Translation. In *ACL*. Code available at https://github.com/shrimai/Style-Transfer-Through-Back-Translation.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *COLING*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *AAAI*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished.

Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *ACL*.

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 *NAACL*.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *ACL*.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In *NAACL*.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*. Connotation Frames downloaded from http://maartensap.com/movie-bias/.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from Non-Parallel text by Cross-Alignment. In *NeurIPS*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *EMNLP*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. In *EMNLP*.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. Unpublished.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *TACL*.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *NeurIPS*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Ye Zhang, Nan Ding, and Radu Soricut. 2018. SHAPED: Shared-Private Encoder-Decoder for text style adaptation. In *NAACL*.

A Additional data description

A.1 ROC story corpus

This English corpus originally contains 100,000 five-sentence stories written by crowdworkers about realistic everyday scenarios. We select the data for our task by first extracting agency levels for each sentence, filtering out those with indeterminable agency. Additionally, we filter out sentences with four or more verbs, to prevent the sentence masking from deleting too many content words.

A.2 Paraphrase corpus

This corpus contains paraphrases of spoken dialogue extracted from movie and TV subtitles.[13] OpusParcus was created by automatically aligning the subtitles sentences using several probabilistic metrics, including likelihood under a round-trip translation paraphrasing model (Bannard and Callison-Burch, 2005) and pointwise mutual information. For our paraphrasing dataset, we apply the same filtering as with the ROC story corpus to the English portion of the OpusParcus training corpus and select the top 10% highest scoring paraphrases using the PMI scoring from the original paper. We extract agency levels for each pair of paraphrases, and select pairs to obtain roughly equal number of agency-level pairs (i.e., 1/9th positive-neutral, 1/9th positive-negative, etc.) We preprocess the text by stripping any leading periods and commas.

B Experimental details

We use the Hugging Face (Wolf et al., 2019) implementation of OpenAI's GPT model (117M parameters; Radford et al., 2018). our final setup uses AdamW (Loshchilov and Hutter, 2019) as our optimizer with a learning weight of 1e-5, batch size of 4 and maximum sequence length of 64. In preliminary results, we find that β=5 aptly steers the generation while avoiding repetition issues.

B.1 POWERTRANSFORMER details

All the experiments are performed on NVIDIA TITAN card and use the model hyperparameters listed in Table 5.

B.1.1 POWERT$_{ParaOnly+None}$

We train this model for 10 epochs with each epoch taking approximately an hour. The learning rate

Hyperparameter	Value
Vocabulary Size	40486
Maximum Sequence Length	64
Training Batch Size	4
Embedding Size	768
# Attention Heads	12
# Attention Layers	12

Table 5: POWERTRANSFORMER hyperparameters.

is 1e-5 with AdamW optimizer, which is tuned manually in the [1e-6, 1e-3] range for 7 times. We use $p = 0.4$ for nucleus sampling and p is tuned manually in the [0.4, 0.9] range for 5 values.

B.1.2 POWERT$_{ParaOnly+Static}$

The POWERT$_{ParaOnly+Static}$ loads the trained model from POWERT$_{ParaOnly+None}$ and add re-scaling to the logits. The re-scaling factor, β was tuned manually tuned in the [0, 10] range. We try 8 βs and use 5 in the final model. We use the same p as POWERT$_{ParaOnly+None}$

B.1.3 POWERT$_{Joint+None}$

Similar to POWERT$_{ParaOnly+None}$, we train this model for 10 epochs with each epoch taking approximately an hour. The learning rate is 1e-5 with AdamW optimizer, which is tuned manually in the [1e-6, 1e-3] range for 7 times. We use the same p as POWERT$_{ParaOnly+None}$

B.1.4 POWERT$_{Joint+Static}$

The POWERT$_{Joint+Static}$ loads the trained model from POWERT$_{Joint+None}$ and add re-scaling to the logits. The re-scaling factor, β was tuned manually tuned in the [0, 10] range. We try 8 βs and use 5 in the final model. We use the same p as POWERT$_{ParaOnly+None}$

B.2 PPLM details

The PPLM decoding method can be used on top of any model, but their original codebase is for use with a pre-trained language model rather than a model for paraphrasing or style transfer. We augment their techniques for this task by replacing the base model in their code with a denoising autoencoder that was trained to reconstruct the input sentence. The denoising autoencoder was implemented using the base GPT2 model (to fit with their code library and be similar size to our model). It was trained on our ROC only training data with a

---

[13]From http://www.opensubtitles.org

reconstruction objective. In order to denoise the autoencoder, we randomly "dropout" about 50% of the tokens from the context by replacing them with mask tokens. This autoencoder is trained to reconstruct input sentences, but when used with the PPLM decoding method, the input gets dynamically updated to decode a sentence that is similar in meaning but more likely to have a positive/negative agency according to a discriminator that is trained on top of the autoencoder. The PPLM decoding method also has hyperparameters that control the strength of the target label. If set too high, then the output could be degenerate. We manually set the hyperparameters to be as strong possible without producing degenerate text, using a subset of the dev. set as a guide.

B.3 Backtranslation details

We use the code provided by Prabhumoye et al. (2018) for running this baseline. After lowercasing all the negative and positive agency examples in our training data (ROC and OpusParcus), we translate to French using the machine translation model provided in the code base. This baseline requires training a style classifier (agency) and two decoders (one for each agency level). Since the classifier essentially re-learns the agency lexicon, we do not search for hyperparameters, and simply set a learning rate of 5, and 6 epochs. For training the decoders, we perform grid search to find the best hyperparameters. We experiment with a learning rates of {0.5, 1, 2, 5}, {2, 3, 5} epochs, a classification-loss weight of {0.5, 1, 2}, and a word-loss weight of {0.5, 1, 2}, and select the configuration with the best word-level accuracy on the dev. set. We use SGD with a batch size of 64 for all experiments, and refer the reader to the code base for other default parameters.

C Gender Bias in Movies

C.1 Extracting gender from characters

The movie scripts mention characters in all caps, making it easy to identify and extract them. We then cross reference the name (or, description for unnamed characters, e.g., "the doorman") with a list of gendered names[14] and gendered words (e.g., "waitress," "policeman," "police woman"). To allow for better rewriting using our model, we split

the narratives into sentences (using NLTK's sentence tokenizer Bird et al., 2009), and assign each sentence to a character if their name appears in the sentence.

---

[14]http://www.cs.cmu.edu/Groups/AI/util/areas/nlp/corpora/names/0.html

Task

Original Sentence:

Alex loves football.

Revisions:

Revision A:

Alex loves watching
football.

Easy to understand ◉
Some grammar errors ○
Impossible to understand ○

Revision B:

Alex loves to play
football.

Easy to understand ◉
Some grammar errors ○
Impossible to understand ○

Q1: Which of these portrays the main person so they have the **highest agency** (regardless of meaning preservation)?
If there are multiple characters in the sentence, usually the ones referred to by pronouns (he, she, etc.) are the main characters.

○ **Revision A** Alex loves watching football.

○ **Revision B** Alex loves to play football.

Q2: Which do you think is **closer in meaning** to the original sentence (regardless of agency change)?
Pick the sentence that has the general events and measing closest to the original.

○ **Revision A** Alex loves watching football.

○ **Revision B** Alex loves to play football.

Submit

Figure 5: Screenshot of the human evaluation annotation task.

Full Instructions　　(Expand/Collapse)

Instructions

Thanks for participating in this qual task! Your job is to:

- Read a pair of sentences
- Select which ones portray the ***main character*** with the highest agency vs. the lowest agency.

What is agency

Agency: The agency level is how active, decisive, or powerful the main person in the sentence is. For example, someone with **high** agency is:
- actively participating in events
- has a lot of power or ability to shape their own future
- pro-active in making their own decisions

Background

We are trying to test out a few automatic systems for automatically generating sentences, and want to see how they portray characters / people in sentences. Machines are not as good at understanding nuanced concepts like *agency*, so your help is crucial and very much appreciated!

Examples

Sentence	Agency Level	Explanation
Alex answered a phone call.	low agency	Alex picked up the phone but did not actively initiate the conversation.
Alex waited around all day while the TV played.	low agency	Alex was not actively participating in actions.
Alex received a book from their friend.	low agency	Alex is portrayed passively receiving things not actively asking for the book.
Alex calls their friend.	high agency	Alex initiated a conversation.
Alex did most of the work by themselves.	high agency	Alex is taking charge of the situation.
Alex took a book from the friend.	high agency	Alex is actively participating in borrowing the book.

Task

Pair 1

> **Sentence A:** Yolanda hates roller coasters.
>
> **Sentence B:** she decided to go and the la and the de

1) First, let's rate how understandible each of these sentences are:

Q1: Which of these sentences are too **ungrammatical/difficult to understand**?

- ☐ **Sentence A** Yolanda hates roller coasters.
- ☐ **Sentence B** she decided to go and the la and the de

2) Now, let's rank them in terms of agency level:

Q2: Which of these portrays the main person so they have the **highest agency**?

- ◯ **Sentence A** Yolanda hates roller coasters.
- ◯ **Sentence B** she decided to go and the la and the de

[Submit]

Figure 6: Screenshot of the qualification task and its instructions. In the real task, workers rated three pairs of sentences, but only one is shown here.

ΔAgency	Input	Model	Output	Agency(out)
(a) $+ \rightarrow -$	After the party I headed home.	PPLM	After the party my classmate, Kayla and the Tgirls of the Universe.	=
		BST	please 's , i have a word of this .	−
		POWER$T_{ParaOnly+NoBoost}$	after the party i headed home.	+
		POWER$T_{ParaOnly+Boost}$	after the party i headed home.	+
		POWER$T_{Joint+SupplyVerb}$	after the party i faced home.	−
		POWER$T_{Joint+NoBoost}$	after the party i stayed home.	−
		POWER$T_{Joint+Boost}$	after the party i stayed home.	−
(b) $+ \rightarrow -$	A friend asked me to watch her two year old child for a minute.	PPLM	A Friend asked me to watch her two year old child for a minute.	+
		BST	l didn 't have a word of this , you 're .	−
		POWER$T_{ParaOnly+NoBoost}$	a friend asked me to watch her two year old child for a minute.	+
		POWER$T_{ParaOnly+Boost}$	a friend asked me to watch her two year old child for a minute.	+
		POWER$T_{Joint+SupplyVerb}$	a friend told me to watch her two year old child for a minute.	+
		POWER$T_{Joint+NoBoost}$	a friend needed me to watch her two year old child for a minute.	−
		POWER$T_{Joint+Boost}$	a friend needed me to watch her two year old child for a minute.	−
(c) $+ \rightarrow -$	After filling in the data it looked quite sharp.	PPLM	Before filling the last question it it it it looked quite sharp. Before filling the last question it it	+
		BST	when the 't you want a word ?	−
		POWER$T_{ParaOnly+NoBoost}$	after filling in the data it looked quite sharp.	+
		POWER$T_{ParaOnly+Boost}$	after filling in the data it seemed quite sharp.	+
		POWER$T_{Joint+SupplyVerb}$	after putting in the data it looked quite sharp.	=
		POWER$T_{Joint+NoBoost}$	after analyzing in the data it looked quite sharp.	=
		POWER$T_{Joint+Boost}$	after seeing in the data it seemed quite sharp.	−
(d) $- \rightarrow +$	Allie was failing science class.	PPLM	Allie was failing science grade.	−
		BST	do you want me ?	+
		POWER$T_{ParaOnly+NoBoost}$	allie was failing science class.	−
		POWER$T_{ParaOnly+Boost}$	allie was failing science class.	−
		POWER$T_{Joint+SupplyVerb}$	allie was ignoring science class.	+
		POWER$T_{Joint+NoBoost}$	allie was failing science class.	−
		POWER$T_{Joint+Boost}$	allie was taking science class.	+
(e) $- \rightarrow +$	Darla wanted a soft drink.	PPLM	darla wants a hard hard drink.	−
		BST	don 't take me a man .	+
		POWER$T_{ParaOnly+NoBoost}$	darla wanted a soft drink.	−
		POWER$T_{ParaOnly+Boost}$	darla wanted a soft drink.	−
		POWER$T_{Joint+SupplyVerb}$	darla got a soft drink.	+
		POWER$T_{Joint+NoBoost}$	darla ordered a soft drink.	+
		POWER$T_{Joint+Boost}$	darla ordered a soft drink.	+
(f) $- \rightarrow +$	Clint paused on the trail.	PPLM	clint was on the trail.	
		BST	don 't you want me ,	−
		POWER$T_{ParaOnly+NoBoost}$	clint paused on the trail.	−
		POWER$T_{ParaOnly+Boost}$	clint stopped on the trail.	+
		POWER$T_{Joint+SupplyVerb}$	clint walked on the trail.	+
		POWER$T_{Joint+NoBoost}$	clint hiked on the trail.	=
		POWER$T_{Joint+Boost}$	clint walked on the trail heading down.	+

Table 6: Full version of Table 4. Example revisions from various models for sentences from the dev. set. Columns are: the target change in agency from the original to the target agency, the input sentence, the model, generated output, and the actual agency level of the output measured by the connotation frame lexicon.