ENTRUST: Argument Reframing with Language Models and Entailment

Tuhin Chakrabarty[1], Christopher Hidey [3] Smaranda Muresan[1,2]
Department of Computer Science, Columbia University [1]
Data Science Institute, Columbia University [2]
Google [3] *
{tuhin.chakr, smara}@cs.columbia.edu
chris.hidey@gmail.com

Abstract

Framing involves the positive or negative presentation of an argument or issue depending on the audience and goal of the speaker (Entman, 1983). Differences in lexical framing, the focus of our work, can have large effects on peoples' opinions and beliefs. To make progress towards reframing arguments for positive effects, we create a dataset and method for this task. We use a lexical resource for connotations to create a parallel corpus and propose a method for argument reframing that combines controllable text generation (positive connotation) with a postdecoding entailment component (same denotation). Our results show that our method is effective compared to strong baselines along the dimensions of fluency, meaning, and trustworthiness/reduction of fear.

1 Introduction

Public opinion has been shown to be significantly influenced by *framing effects.* Framing refers to the presentation of an issue, where even small changes may have outsized effects on beliefs (Chong and Druckman, 2007). For example, when asked about "welfare," the American public is largely against increasing spending (with only 20% in favor), but when asked about "assistance to the poor," 65% believe that the government is not spending enough (Rasinski, 1989).

While other research has focused on *syntactic* framing (Greene and Resnik, 2009) or *issue* framing (Hartmann, 2019), we focus specifically on *lexical* framing, distinguishing sentences by their *connotative* meaning even where they have the same *denotative* meaning. According to Frege (1892), two sentences with the same truth conditions may refer to the same entities or state of affairs (*"reference,"* also known as denotation) but be presented

Arg1	Alabama's Supreme Court Chief Justice was suspended... for ordering state probate judges not to grant *marriage licenses* to gay couples...
Rf Arg1	Alabama's Supreme Court Chief Justice was suspended... for ordering state probate judges not to grant *legal marriage equality* to gay couples...
Arg2	Every nation with *territorial claims* in the arctic is a member of NATO, except Russia.
Rf Arg2	Every nation with *sovereign competence* in the arctic is a member of NATO, except Russia.
Arg3	At this *dire* moment , we all need to amplify our voices in *defense* of free speech .
Rf Arg3	At this *crucial* moment , we all need to amplify our voices in *support* of free speech.
Arg4	It is difficult to think of any single act that would do more to restore America's *soft power* than the election of Obama to the presidency
Rf Arg4	It is difficult to think of any single act that would do more to restore America's *diplomatic credibility* than the election of Obama to the presidency

Table 1: Examples of arguments (Arg1, Arg2) with high partisan skew collocations (in red) (Webson et al., 2020) as well as appeal to fear or prejudice argument fallacies (Arg3, Arg4) (Da San Martino et al., 2019), along with reframed arguments as an attempt by our model **ENTRUST** to improve trustworthiness.

differently (*"sense"* or connotation). For example, "undocumented workers" and "illegal aliens" have the same denotation but different connotations (Webson et al., 2020).

The examples in Table 1 are instances of *lexical* framing, where word choice determines the difference in presentation (Mccombs and Ghanem, 2001). For example, Arg1 and Arg2 contain collocations (in red) that have a high partisan skew (Webson et al., 2020), while Arg 3 and Arg 4 are examples of appeal to fear or prejudice argument fallacies from propagandist news articles (Da San Martino et al., 2019). The goal is to reframe such arguments to be more trustworthy (e.g., less partisan, no appeal to fear fallacy).

Connotations may be distinguished along the

---

dimensions of politeness, sentiment, or tangibility, among others (Allaway and McKeown, 2020), but in our work we consider emotional association such as fear and trust. Appeal to *fear* is considered an argumentative fallacy (Walton, 2006; Thierer, 2012) and appears prominently in manipulative text such as propaganda (Da San Martino et al., 2019). On the other hand, arguments with *trusted* language align with the Aristotelian modes of persuasion, specifically ethos (Aristotle and Bartlett, 2019).

In our work, we leverage such a lexical resource for connotations (Allaway and McKeown, 2020) to reframe arguments to be more trustworthy (e.g., less partisan, no appeal to fear fallacy), while maintaining the same denotative meaning. While retrieve-and-replace methods perform well on other attribute transfer tasks such as sentiment (Li et al., 2018a; Sudhakar et al., 2019a), our task is more dependent on broader context within a sentence even though we are performing localized replacement. Thus, there are two main challenges we need to address: 1) the lack of a parallel dataset of negatively and positively framed arguments (naturally-occurring); and 2) a generation approach that can not only change the connotative meaning but also keep the same denotative meaning of the input argument.

We introduce our approach called **ENTRUST** : Argum**ENT R**eframing with lang**U**age model**S** and en**T**ailment, with the following contributions: 1) A Connotation-guided Masked Language Model approach to generate a parallel dataset of naturally occuring arguments and their reframings (Section 2); 2) A method for argument reframing that combines controllable text generation (connotative meaning associated with trust) and entailment (same denotative meaning) (Section 3); 3) An evaluation on two different tasks — reframing partisan arguments and appeal to fear/prejudice fallacies — showing that our method is preferred over a strong retrieval-based baseline (Sudhakar et al., 2019a) and state-of-the-art pretrained language model (Lewis et al., 2019), and it is close to human performance on several evaluation criteria such as fluency, meaning, trustworthiness/reduction in fear. Code, data, and models available at `https://github.com/tuhinjubcse/ArgReframingNAACL2021`

2 Automatic Parallel Data Creation

To facilitate the reframing of arguments, we require a large-scale parallel corpora of sentences with the same denotation but different connotative meaning.

Selection of naturally-occurring arguments. Since our goal is to re-write arguments, it is essential to identify an abundant source of naturally-occurring arguments. The *Change My View* sub-reddit, an argumentative discussion forum intended for persuasion on diverse topics, has been used extensively in computational argumentation research (Tan et al., 2016; Wei et al., 2016; Musi et al., 2018; Chakrabarty et al., 2019a,b; Hidey et al., 2017). We collect sentences from the same source and classify them as *claim*, *premise*, or *non-argument* using the fine-tuned BERT model released by Chakrabarty et al. (2019b). This results in 301,166 arguments labeled as premises. We consider only premises to create our parallel data because argumentative appeals occur within justifications (premises) for or against the speaker's claim.

Connotation-guided Masked Language Model. Allaway and McKeown (2020) provide a resource with words labeled for lexical connotations, using the aspects of *Social Value, Politeness, Impact, Factuality, Sentiment, Emotional Association*. For our work we only consider *Emotional Association*, although in future work our methods could be applied for other aspects. To create a parallel corpus, we use this lexical resource and the 301,166 automatically identified premises from *Change My View* to obtain candidate words within those premises for replacement. We match words from the premises to those that have entries in the dictionary with *emotional* connotations such as fear, trust, anticipation, and joy. To generate replacements for these words, we need to find substitutions that maintain denotative meaning while changing connotative meaning. We use the connotation dictionary to address the latter. However, to address the former, we need to provide only paraphrases that consider the context in which these words occur. We thus use a masked language model (MLM).

Masked language modeling approaches like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) can be considered *cloze* or "fill-in-the-blank" tasks, where the model uses the context surrounding a masked-out token to try to predict what the masked word should be. We borrow this framework (RoBERTa-large, in particular) to mask the candidate words we identified via the connotation lexicon. However, the rank of a predicted token from an MLM is based on the language model
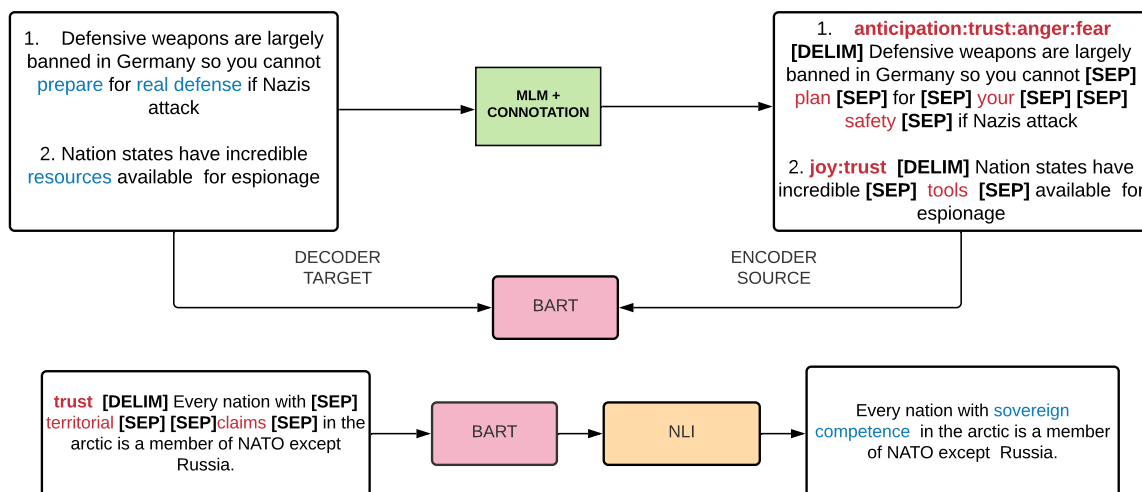
Figure 1: A schematic illustration of our system ENTRUST, where the top block shows our **training** process where we use MLM along with connotation resource to transform an original argument to an argument bearing different emotional connotation and use them to fine-tune BART. The block below shows the **inference** step (test time) where we use fine-tuned BART to reframe the argument containing *partisan collocation* along with a NLI component to ensure the same denotation with the input argument.

probability – it provides no information about lexical connotations. A premise re-written from MLM replacements may thus have the same connotative meaning. To avoid this scenario, we restrict the MLM replacements to be words with different connotations than the original masked word (i.e., different Emotional Association).

Our data creation process is depicted in Figure 1. In example 2, the word *"resources"* has the connotations *joy;trust* in our dictionary. The MLM generates the replacement *"tools,"* which we verify has a different connotation (*emotionally neutral*). For example 1, the words *"prepare," "real,"* and *"defense"* have the emotional connotations *anticipation*, *trust*, and *anticipation;anger;fear* respectively. These words are replaced with *plan*, *your*, and *safety*, using our MLM.

We treat the original premises as the "target" and the connotation-guided MLM generated premises as the "source" for our method of argument reframing detailed in the next section (Figure 1). While this process provides us with a parallel dataset for reframing, we enhance the source-side of the data to provide additional control during generation. Motivated by the work of Schiller et al. (2020), which used aspect as a "control code" (Keskar et al., 2019) for argument generation, we also prepend the emotional associations of the replaced words. Using the connotations from the lexical resource, we add all listed emotions as control codes by sepa-

rating them with a special token ("[DELIM]") (the top right block of Figure 1). During inference, we thus have more control over the emotion of the words we are generating (in our case we specifically use trust as the control code). For additional control, we also insert demarcator tokens ("[SEP]") at the boundary of the words we aim to replace to provide our generative model with a better signal on what to replace or rewrite. While the downside is that we need to identify spans for replacement at test/inference time, our experiments will show that using collocations or fear words makes it unnecessary.

By using the lexical connotation resource we do not have to rely on a separate module/tagger based approach like that of Pryzant et al. (2020) to find biased or problematic words that may introduce additional noise during training. Our parallel data has 271,022 pairs for training and 30,114 for validation on which perplexity is evaluated.

3 Method for Argument Reframing

As our goal is to change *connotation* while maintaining *denotation*, we divide our approach to rewriting arguments into two primary tasks: 1) generating the appropriate lexical substitutions while being pertinent to the context; 2) ensuring that rewritten arguments reflect the desired emotional association while maintaining the same denotative meaning as the input.

Source	trust <V>I suppose we could argue that they're much better at *soft power* than Nazi Germany or the USSR, but come on
BART	I suppose we could argue that they're much better at *military strength* than ....
BART + NLI	I suppose we could argue that they're much better at *diplomatic communication* than ......

Table 2: Generation from fine-tuned BART without control for entailment can sometime contradict the input thereby failing to maintain the same denotative meaning

3.1 Controllable Text Generation

BART (Lewis et al., 2019) is a pre-trained model combining bidirectional and auto-regressive transformers that achieves state-of-the-art results in several text generation tasks. It is implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right auto-regressive decoder. In principle, the pre-training procedure has two stages: (1) text is corrupted with an arbitrary noising function, and (2) a transformer-to-transformer model is learned to reconstruct the original text. Because BART has an auto-regressive decoder, it can be directly fine-tuned for most sequence generation tasks. Here, the encoder input is a sequence of words, and the decoder generates outputs auto-regressively. We refer the reader to (Lewis et al., 2019) for further details.

For our task, we fine-tune BART on our parallel data, where the reframed argument using MLM & connotation dictionary is the encoder source and the original argument is the decoder target (Figure 1). The emotional connotations added to the source via the special token DELIM (see Section 2) act as a "control code" for generation. Moreover, for lexical framing, subtle differences in word choices matter the most. By explicitly using special tokens ([SEP]) in our parallel data during fine-tuning, the BART model learns what to edit, instead of editing random words in the sentence, a common issue often found in attribute transfer models (Li et al., 2018a; Sudhakar et al., 2019a). At test time, therefore, we can ensure the model reframes a desired content span. All hyper-parameters are mentioned in the **Appendix A**.

Post fine-tuning at the decoding step, we use a top-k sampling strategy (Fan et al., 2018) to reframe arguments conditioned on a input argument and a target emotion.

3.2 Post-decoding NLI

Our task is challenging in comparison to traditional text attribute transfer tasks as we need to maintain the same denotative meaning as the input. While in most cases BART is able to generate content which is semantically similar to the input, it sometimes contradicts the input. For example, Table 2 shows that BART changes *soft power* to *military strength*. Here the denotative meaning changes. To control for this, we introduce an additional post-processing step.

We generate multiple outputs by varying the value of k (between 5 and 50) while conducting top-k sampling. We then calculate the entailment scores of these outputs with the input argument respectively using a RoBERTa (Liu et al., 2019) model fine-tuned on the Multi-NLI dataset (Williams et al., 2018) and then select the output having the best entailment score. We also experimented with other methods for incorporating entailment during decoding based on prior work (Section 8), but found these techniques to be less effective than our method. As pre-trained sequence-to-sequence language models are good at copying input and generating natural-sounding output, we hypothesize that our approach will better allow us to change connotative meaning without affecting fluency and denotation. In contrast, approaches such as "vocab boosting" (Ghosh et al., 2017) increase the logits of key connotative words, which would necessarily decrease the probabilities of functional words and words necessary for maintaining denotative meaning. Other approaches such as reinforcement learning (Pasunuru and Bansal, 2017) may further decrease these desired qualities, while trying to maximize another objective.

4 Evaluation Tasks and Test Data

To evaluate our methods for argument reframing we need to look beyond our automatically labeled data. We consider two tasks: 1) reframing an argument that contains partisan language to a less partisan argument; and 2) reframing an appeal to fear or prejudice fallacy to an argument without this fallacy.

Recently Webson et al. (2020), proposed resources and methods to disentangle denotation and connotation in vector spaces. They evaluate their methods on a sample of around 300 collocations from vocabulary of Congressional records (Gentzkow et al., 2019) and Hyper-partisan News

INP1	It would be *dangerous , suicidal folly* for infidels to pretend that ramadan is not the month of jihad
HUM1	It would be *counterproductive and unreasonable* for infidels to .... jihad
INP2	Trump backs away from further *military confrontation* with Iran
HUM2	Trump backs away from further *military engagement* with Iran

Table 3: INP1 and INP2 are test data instances where INP1 is a Appeal to Fear example while INP2 is an argument containing partisan collocation

(Kiesel et al., 2019) that occur at least 100 times and have high partisan skew. We use these words to filter arguments from the subreddits *Change-MyView* and *Politics*. Some of these collocations include phrases such as *abortion providers, investment vehicles, broken system, soft power,* and *territorial claims*. We randomly sample 100 such arguments to benchmark the performance of our model and further use towards human evaluation.

In addition, we test our models on propaganda techniques employed in news articles with an *Appeal to Fear or Prejudice* (Da San Martino et al., 2019). There are a total of 182 sentence-level text fragments labeled as Appeal to Fear or Prejudice in the dataset released by Da San Martino et al. (2019). We classify these 182 fragments as claims/premises/non-argument and randomly sample 50 premises. Our goal is to reduce the fallacious nature of the argument without changing the denotative meaning.

As our training distribution is different from these two datasets, these tasks and test sets allow us to better test the generalization capabilities of our models. Furthermore, almost none of the collocations introduced by Webson et al. (2020) appear in the connotation dictionary of Allaway and McKeown (2020), which helps us avoid the risk of mimicking replacements from our training data.

For both of these tasks, we ask humans to generate reframings based on our input test data for comparison and benchmark. We recruit two experts with argumentation and journalism background (not authors of this paper) to reframe arguments. For Appeal to Fear the instructions given were to make it less fallacious by reducing the fear and rephrasing the argument (HUM1 in Table 3), while for arguments with partisan collocation the human was instructed to change the collocation so as to make it trustworthy (HUM2 in Table 3).

5 Experimental Setup

To compare the quality of the reframed arguments, we benchmark our ENTRUST model against human performance and four baseline systems described below. For the data containing collocations from (Webson et al., 2020), because we know they represent partisan language the ideal goal is to reframe them. For Appeal to Fear or Prejudice data we reframe words which portray an emotion of fear based on the popular NRC Emotion Lexicon (Mohammad and Turney, 2013).

5.1 Baseline Systems

As argument reframing is a new task, we adapt several baselines that have been used for other generation tasks and also compare with human-generated reframings.

Bart wIthout Demarcator and ENtailment (BART$_{D+EN}^{w/o}$): This is the pre-trained BART model fine-tuned on our parallel data without explicitly adding signals on what to edit or reframe and without post-processing based on entailment scores. This experiment helps us understand if BART learns to adapt to the emotional connotations and can automatically edit partisan collocations or words inducing fear without control.

Bart without EnTAilment (BART$_{EN}^{w/o}$): This is the pre-trained BART model fine-tuned on our parallel data with explicit signals ([SEP] token) but without the NLI component as a post-processing tool. This experiment helps us understand how well BART learns to adapt to the emotional connotations without altering the denotative meaning once guided with what to reframe.

Lexical Replacement (LEXREP): We use a similar method employed for our parallel data creation. We rely on Masked Language Models for lexical substitutions. Because our goal is to reframe arguments to be trustworthy we prefer substitutions which have a connotation of trust in the resource by Allaway and McKeown (2020). In case we cannot find the substitution in the connotation dictionary we honor default MLM predicted infilling.

Generative Style Transformer (GST): We use the state of art for text style transfer by Sudhakar et al. (2019a), which is a part of a larger "*Delete Retrieve Generate*" framework (Li et al., 2018a). To maintain parity with other baselines, instead of letting the model delete attribute keywords we delete

System	Partisan Task	Appeal to Fear Task
BART$_{D+EN}^{w/o}$	64.1	38.5
BART$_{EN}^{w/o}$	91.9	43.1
GST	86.4	38.3
LEXREP	92.4	44.3
ENTRUST	92.9	**44.5**
HUMAN	**93.9**	41.6 *

Table 4: Semantic Similarity of reframed arguments with input arguments. (*) Here human did not restrict themselves to just lexical framing, so automated metrics might penalize them for more reframing.

System	Fluency	Meaning	Trust ↑	Fear↓
INPUT	-	-	3.24	3.36
BART$_{D+EN}^{w/o}$	2.78	2.56	2.60	3.01
BART$_{EN}^{w/o}$	3.39	3.00	3.13	2.58
LEXREP	3.38	3.00	3.08	2.54
GST	2.14	1.81	2.01	2.44
ENTRUST	3.51	3.30	3.52	**2.39**
HUMAN	**3.72**	**3.63**	**3.71**	2.59

Table 5: Fluency and Meaning Preservation scores given by human judges on a scale of (1-5) for reframed arguments with respect to input arguments. Fluency and Meaning Preservation ratings are for all arguments in test set, while Trust ratings are for arguments with Partisan collocation (higher scores better), and Fear ratings for Appeal To Fear or Prejudice ones only (lower scores better).

the partisan collocations or fear related words from the arguments as the first step, followed by the usual retrieve and generate steps. Our training data for this method includes only arguments labeled with their attribute (e.g., positive or negative). Arguments containing lexical connotations catering to trust are positive, while those not catering to trust are negative.

5.2 Evaluation Criteria

Automatic evaluation. One important criterion is to measure if the reframed arguments are faithful to the input. Even though we are changing the argument for connotations, it should still maintain the same denotative meaning as the input. To this end we calculate **Semantic Similarity** with our input using SENTENCE BERT(SBERT) (Reimers and Gurevych, 2019).

Human evaluation. We use Amazon Mechanical Turk to evaluate on a total of 900 utterances, 750 generated from 5 systems and 150 utterances generated by humans. We proposed a set of 3 criteria to evaluate the generated output: (1) **Fluency (F)** ("How fluent and grammat-ical are the utterances?"), (2) **Meaning Preservation (M)** ("How well does the reframed argument capture the same denotative meaning as the input argument?"), (3) **Trustworthiness/Presence of Fear(T/PF)**. For the 100 input arguments reflecting partisan view we ask Turkers to rate reframed arguments based on trustworthiness with respect to the input. For the 50 Appeal to Fear or Prejudice fallacies we ask Turkers to rate reframed arguments based on presence of fear (the intention behind this being that we want to rank systems which portray the least amount of fear). In both of these ratings we still ask Turkers to keep into account the denotative meaning (i.e., making it trustworthy or less fallacious at the expense of meaning alterations should be scored lower). We hired 40, 25, 39 (23 and 16) Turkers for the three separate tasks respectively. The computed IAA using Krippendorff's alpha for Fluency, Meaning Preservation , Trust-Worthiness and Presence of Fear is 0.62, 0.65, 0.51, 0.46, respectively.

6 Results

Automatic Evaluation. As can be seen in Table 4 our model ENTRUST maintains the denotative meaning with the input better than other systems ($p < .001$ using approximate randomization tests) and only marginally behind humans when it comes to arguments with partisan collocations. For Appeal to Fear or Prejudice our system maintains better denotative meaning than all systems except LEXREP ($p < .001$). The automatic metric somewhat penalizes humans for changing more content than just targeted words; this unreliability is a known issue with automated metrics (Novikova et al., 2017) and strongly implies a need for human evaluation.

Human Evaluation. Table 5 shows the results of our human-based evaluations. For fluency, meaning preservation, trustworthiness, and reduction of fear the ENTRUST model is better than all the baselines ($p < .001$ using approximate randomization tests). It is further encouraging to see that the entailment step helps us maintain better denotative meaning (See Table 5 Col3: Row 4 vs Row 7). For Presence of Fear, Turkers often rate our ENTRUST model to be the least fearful, including slightly when compared to reframings of an expert. We hypothesize this is because the human judges found it difficult to completely remove fear while keeping the denotative meaning (indeed the

Original Argument	System	reframed Argument	F	M	T↑/PF↓
It is difficult to think of any single act that would do more to restore America's *soft power* than the election of Obama to the presidency	BART$_{D+EN}^{w/o}$	It is difficult to think of any single act that would do more to restore America's *soft power* than the election of Obama to the presidency	3.7	2.3	3.3
	BART$_{EN}^{w/o}$	It is difficult to think of any single act that would do more to restore America's *moral authority* than the election of Obama to the presidency	3.7	3.3	3.0
	LEXREP	It is difficult to think of any single act that would do more to restore America's *moral standing* than the election of Obama to the presidency	3.7	3.3	2.7
	GST	*Hated* it is difficult to think of any single act that would do more to restore America's *economy* than the election of Obama to the presidency	1.7	1.7	2.7
	ENTRUST	It is difficult to think of any single act that would do more to restore America's *diplomatic credibility* than the election of Obama to the presidency	**4.3**	**3.7**	**5.0**
	HUMAN	It is difficult to think of any single act that would do more to restore America's *political and economic influence* than the election of Obama to the presidency	4.0	3.3	4.0
Or he can *seize* this opportunity and use his *government's* diplomatic *influence* to *defend* the life of an australian citizen , whose *courageous* public service is recognised by countless people across the world	BART$_{D+EN}^{w/o}$	Or he can *seize* this opportunity and use his *governments's* diplomatic *influence* to *defend* the life of an australian citizen , whose courageous *military* service is recognised by countless people across the world	2.7	3.0	2.3
	BART$_{EN}^{w/o}$	Or he can *honor* this opportunity and use his *nation's* diplomatic *resources* to *honor* the life of an australian citizen , whose *unwavering* public service is recognised by countless people across the world	4.0	4.7	**1.3**
	LEXREP	Or he can *take* this opportunity and use his *nations's* diplomatic *resources* to *defend* the life of an australian citizen , whose courageous *military* service is recognised by countless people across the world	**4.7**	4.0	2.0
	GST	Or he can *do* this opportunity and use his diplomatic *expertise* to *change* the life of an australian citizen , whose public service is recognised by countless people across the world	1.7	2.0	2.0
	ENTRUST	Or he can *honor* this opportunity and use his *nation's* diplomatic *resources* to *vindicate* the life of an australian citizen , whose *unwavering* public service is recognised by countless people across the world	**4.7**	**4.7**	**1.3**
	HUMAN	Or he can *pick up* this opportunity and use his government's diplomatic influence to defend the life of an Australian citizen, whose *actions have been publicly recognized as highly relevant at an international level*.	3.7	4.0	2.3

Table 6: Examples of generated outputs from different systems (with human reframed argument as references) for arguments containing partisan collocations and appeal to fear, respectively. We show average scores (over three annotators) on a 1-5 scale with 1 denotes the worst and 5 be the best. More examples in Appendix

humans scores slightly better on meaning comparing with our system ENTRUST). Sometimes, an ungrammatical generation or a reframing which change the meaning will contain less fear (rating 1 meaning no fear at all). However, to avoid this we explicitly asked Turkers to rate those samples as moderate so as to not bias the overall results.

7 Analysis and Discussion

As can be seen in Table 6, the ENTRUST model accurately captures *diplomatic credibility* as an alternate for soft power which is encouraging as soft power is measured through *culture, diplomacy,*

education, business/innovation, and government.[1] The BART$_{D+EN}^{w/o}$ model often fails to reframe anything, which shows the importance of adding [SEP] tokens as explicit supervision so that the model knows what to edit. The GST model fails at both grammaticality and meaning preservation, which makes it harder to judge its trustworthiness and ability to ameliorate fearful appeal. Finally, ENTRUST reframings are not static. Table 8 shows that for the same collocation of *targeted killing*, the reframings are different, contingent on the context. This goes on to prove that our model not only

---

[1] https://en.wikipedia.org/wiki/Soft_power

An Iranian government official seemed to suggest that President Trump's properties could be potential targets in retaliation for the US *targeted killing* of Iranian general Qassem Soleimani.
An Iranian government official seemed to .... suggest that President Trump's properties .... in retaliation for the US *involvement in the execution* of Iranian .....
A federal appeals court ordered the U.S. Department of Justice to turn over key portions of a memorandum justifying the government's *targeted killing* of people linked to terrorism, including Americans
A federal appeals court ordered the U.S. Department of Justice to turn .... justifying the government's *extrajudicial execution* of people ... terrorism, ...

Table 7: Different reframed arguments by ENTRUST model based on same collocation

generalizes to unseen test data, but can produce novel, grammatical and meaningful edits based on context.

8 Related Work

The effects of lexical framing have been studied for social and political issues, although our work is the first to use lexical framing in generation for positive framing effects (less partisan, no appeal to fear fallacy). Demszky et al. (2019) and Tyagi et al. (2020) study political polarization and how this manifests in differences in word choice among different groups; KhudaBukhsh et al. (2020) provide an interpretable framework using machine translation between groups to generate differences. While these works encourage computational approaches to reframe arguments for better lexical choice, these approaches do not control for denotation or connotation and thus may cause differences in word choice to result in a change in meaning. The most similar work to ours is that of Pryzant et al. (2020), who use a corpus of Wikipedia edits to train a model for debiasing, which includes framing. However, in their work communicative intent is left implicit; the corpus is only labeled for types of debiasing, which includes framing at a high level and not the connotations involved. Thus, their model only learns lexical differences, whereas our model is controllable.

While our focus is on lexical framing, other work has investigated the identification of other types of frames and their effects. Greene and Resnik (2009) studied *syntactic* framing, finding a link between implicit sentiment and syntactic packagings. Previous studies have also involved *emphasis* framing – Ding and Pan (2016) find that emphasizing aspects of products given personal information is more

effective for content selection in advertisements. Other research has involved *issue* framing – Ajjour et al. (2019) and Hartmann et al. (2019) study how arguments are framed in debates (e.g., in terms of economics or safety). Nguyen (2013) and Field et al. (2018) study "agenda-setting" for news and congressional debates and August et al. (2018) for study recruitment. Cano-Basave and He (2016) and Musi and Aakhus (2019) leverage *semantic* frames for distant labeling and analysis of arguments in political debates, respectively, and find, for example, that evidence and reasoning are among the most common. However, these approaches have focused on identification rather than generation.

Finally, our work is also related to style transfer and controllable generation. Much of the work in "style transfer" has referred to changing the sentiment of a statement, which changes the truth condition and thus the denotative meaning. Sentiment is often explicitly marked and thus approaches such as deleting and replacing lexical markers are effective (Li et al., 2018b; Sudhakar et al., 2019b), although our experiments showed the difficulty of applying these techniques to our task. To control text generation by limiting contradictions, Pasunuru and Bansal (2017) use an entailment score as a reward in Reinforcement Learning, ensuring that a generated text is logically implied by the ground-truth text. Holtzman et al. (2018) utilize a discriminative model trained on SNLI (Bowman et al., 2015) to complement an RNN generator and guide the decoding process to improve contradictions in generation. Although we experimented with both of these approaches, including the approach of Holtzman et al. (2018) with MNLI to account for entailment in text generation, none of them yielded better results than our method. Other approaches have explored "vocab boosting" (Ghosh et al., 2017) for tasks such as de-biasing (Ma et al., 2020), which involves increasing the values of certain words; however, as these values are on the simplex, the softmax function necessarily decreases the values of other logits which are key to fluency such as function words.

9 Conclusion

Our experiments showed that our approach is effective in reframing partisan arguments and appeals to fear for increased trustworthiness. We provided a method for creating a dataset using a lexical resource for connotations and masked lan-

guage modeling. We used this dataset to fine-tune a controllable text generation model for the task of changing connotative meaning and used a model trained for natural language inference to maintain the denotative meaning. Our evaluations found that our approach generalized to two different tasks and data sets. In future work, we plan to directly incorporate the role of stance in framing (for arguments and counter-arguments). We also plan to expand our work to generating concessions (Musi, 2018), where the goal is for the speaker to portray some point of agreement in a positive light before disagreeing.

10 Ethics

Our data is collected from Reddit and we understand and respect user privacy. Our models are fine-tuned on sentence level data obtained from user posts. These do not contain any explicit detail which leaks information about a users name, health, negative financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs, sexual orientation, trade union membership, alleged or actual commission of crime.

Second, although we use language models trained on data collected from the Web, which have been shown to have issues with bias and abusive language (Sheng et al., 2019; Wallace et al., 2019), the inductive bias of our models should limit inadvertent negative impacts. Unlike model variants such as GPT, BART is a conditional language model, which provides more control of the generated output. We have two levels of control on our generation approach: lexical replacements via connotations associated with trust and an entailment method that aims to keep the same denotation of the original argument. While dual-use concerns are certainly possible here, we think that open-sourcing this technology will help to generate arguments with more balanced and trusted language that are less targeted towards partisanship or appeals to fear.

Finally, while there may be concerns about building generative models for persuasion, social scientists distinguish persuasion from manipulation based on two aspects: dissimulation and constraint (Nettel and Roque, 2012). Dissimulation involves concealing intention, which requires hiding information, whereas constraint involves removing options from the audience and forcing them to accept the conclusion. Our work on reframing arguments does not aim to hide information about a

topic or present it as the only choice, but aims to provide the same argument using more balanced and trusted language. We achieve this by two key components of our technology: controllable text generation (connotation associated with trust) and entailment model to ensure same denotation.

The technology should be used responsibly, particularly making sure the generation is controllable for trust and positive emotion and that the entailment component is used for ensuring the same denotation with the original argument.

Finally we pay the Turkers at a rate of 15$/hour, complying with minimum wage standards in most places.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2020. A unified feature representation for lexical connotations.

Aristotle and R. Bartlett. 2019. Aristotle's "art of rhetoric".

Tal August, Nigini Oliveira, Chenhao Tan, Noah A. Smith, and Katharina Reinecke. 2018. Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. *PACMHCI*, 2(CSCW):22:1–22:19.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Kathleen McKeown. 2019a. Imho fine-tuning improves claim detection. *arXiv preprint arXiv:1905.07000*.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, and Alyssa Hwang. 2019b. 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 (EMNLP-IJCNLP), pages 2926–2936.

Dennis Lee Chong and James N. Druckman. 2007. Framing theory.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5640–5650.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

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

Tao Ding and Shimei Pan. 2016. Personalized emphasis framing for persuasive message generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1432–1441, Austin, Texas. Association for Computational Linguistics.

R. Entman. 1983. Framing: Toward clarification of a fractured paradigm. *Journal of Communications*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

G. Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

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 *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

Frederik Hartmann. 2019. Predicting historical phonetic features using deep neural networks: A case study of the phonetic system of Proto-Indo-European. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 98–108, Florence, Italy. Association for Computational Linguistics.

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. Issue framing in online discussion fora. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1401–1407, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom M Mitchell. 2020. We don't speak the same language: Interpreting polarization through machine translation. *arXiv preprint arXiv:2010.02339*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: A simple approach

to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018b. 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, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

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

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. *arXiv preprint arXiv:2010.13816*.

M. Mccombs and Salma I. Ghanem. 2001. The convergence of agenda setting and framing.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Elena Musi. 2018. How did you change my view? a corpus-based study of concessions' argumentative role. *Discourse Studies*, 20:270 – 288.

Elena Musi and Mark Aakhus. 2019. Framing fracking: Semantic frames as meta-argumentative indicators for knowledge-driven argument mining of controversies.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2018. Changemyview through concessions: Do concessions increase persuasion? *arXiv preprint arXiv:1806.03223*.

Ana Laura Nettel and Georges Roque. 2012. Persuasive argumentation versus manipulation. *Argumentation*, 26(1):55–69.

Viet-An Nguyen. 2013. Lexical and hierarchical topic regression. In *NIPS 2013*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1904.01038*.

Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. *arXiv preprint arXiv:1708.02300*.

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*, volume 34, pages 480–489.

K. Rasinski. 1989. The effect of question wording on public support for government spending. *Public Opinion Quarterly*, 53:388–394.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Aspect-controlled neural argument generation. *arXiv preprint arXiv:2005.00084*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019a. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019b. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Adam Thierer. 2012. Technopanics, threat inflation, and the danger of an information technology precautionary principle. *IO: Productivity*.

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M Carley. 2020. A computational analysis of polarization on indian and pakistani social media. *arXiv preprint arXiv:2005.09803*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

D. Walton. 2006. Fundamentals of critical argumentation. In *Critical reasoning and argumentation*.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Do "Undocumented Workers" == "Illegal Aliens"? Differentiating Denotation and Connotation in Vector Spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105, Online. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

A Hyper-Parameters and Other Experimental Settings

1. **No of Parameters:** For Connotation guided MLM and Entailment we use RoBERTa large model (355M). For generation we use the BART large checkpoint (400M parameters) and use the implementation by FAIRSEQ (Ott et al., 2019) [2].

2. **No of Epochs:** We fine-tune pre-trained BART for 20 epochs for ENTRUST model and save best model based on validation perplexity.

3. **Training Time:** Our training time is 80 minutes for BART.

4. **Hardware Configuration:** We use 4 RTX 2080 GPU

5. **Training Hyper parameters:** We use the same parameters mentioned in the github repo

where BART was fine-tuned for CNN-DM summarization task with the exception of MAX-TOKENS (size of each mini-batch, in terms of the number of tokens.) being 1024 for us. For entailment we rely on AllenNLP roberta checkpoint that is finetuned on MNLI [3]

6. **Decoding Strategy & Hyper Parameters:** For decoding we reframe from our models using a top-k random sampling scheme (Fan et al., 2018). At each timestep, the model generates the probability of each word in the vocabulary being the likely next word. We randomly sample from the k = 5 to k=50 for most likely candidates from this distribution and re-rank them based on entailment scores.

A.1 Role of Context in re-framing

As mentioned ENTRUST reframings are not static. Table 8 shows that for multiple same collocations the reframings are different, contingent on the context. This supports our claim that our model not only generalizes on unseen test data, but can produce novel, grammatical and meaningful edits based on context. We notice that for the first example, ENTRUST believes scientific proof leads to be better credibility instead of using the broad umbrella term experts. For the same collocation of *leading expert* we see that given the premise talks about a neurological basis related to birth , ENTRUST reframes it to reflect to *leading medical fact* as the term "expert" can be ambiguous, while people are generally more likely to believe medical facts.

Statistically based-wise though, the number of violent crime is more prevalent in warm areas,*leading experts* to believe a correlation between warm and people being more aggressive.
Statistically based-wise though, the number of violent crime is more prevalent in warm areas,*leading scientific proof* to believe a correlation between warm and people being more aggressive.
The *leading experts* all agree that there's a neurological basis & there is significant evidence showing it's caused during fetal neurological development
The *leading medical facts* all agree that there's a neurological basis & there is significant evidence showing it's caused during fetal neurological development
We need leaders who recognize our *moral obligation* for health equality
We need leaders who recognize our *moral accountability* for health equality
We all have a *moral obligation* to the next generation to leave America's natural resources in better condition than when we inherited them
We all have a *personal oath* to the next generation to leave America's natural resources in better condition than when we inherited them

Table 8: Different reframed arguments by ENTRUST model based on same collocation

Original Argument	System	reframed Argument	F	M	T↑/PF↓
AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. said *bilateral relations* are in crisis	BART$_{D+EN}^{w/o}$	AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. said *bilateral relations* are in *fact stable*	3.7	2.0	2.3
	BART$_{EN}^{w/o}$	AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. said *diplomatic relations* are in crisis	4.7	3.3	4.3
	LEXREP	AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. said *that negotiations* are in crisis	4.0	2.3	3.0
	GST	AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. *law said* are in crisis	1.3	1.3	1.3
	ENTRUST	AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. said *diplomatic ties* are in crisis	**5.0**	**4.7**	**4.7**
	HUMAN	AIPAC called on the Obama administration to take steps to defuse tension with Israel, while Israel's ambassador to the U.S. said *relations* are in crisis	4.7	3.7	4.0
Kentucky Senate President Robert Stivers on Wednesday a federal judge to withhold action against Davis for her defiance of the ruling by denying *marriage licenses* to gay couples, until the Kentucky General Assembly can act	BART$_{D+EN}^{w/o}$	Kentucky Senate President Robert Stivers on Wednesday asked a federal judge to withhold action against Davis for her defiance of the *law* ruling by denying *marriage licenses* to gay couples, until the Kentucky General Assembly can act.	4.3	3.7	3.0
	BART$_{EN}^{w/o}$	Kentucky Senate President Robert Stivers on Wednesday asked a federal judge to withhold action against Davis for her defiance of the ruling by denying *legal marriage equality* to gay couples, until the Kentucky General Assembly can act.	**4.7**	4.0	4.0
	LEXREP	Kentucky Senate President Robert Stivers on Wednesday asked a federal judge to withhold action against Davis for her defiance of the ruling by denying *wedding services* to gay couples, until the Kentucky General Assembly can act.	**4.0**	1.3	2.3
	GST	Kentucky Senate President Robert Stivers on Wednesday asked a federal judge to withhold action against Davis for her defiance of the ruling by denying *the* gay couples, until the Kentucky General Assembly can act.	1.7	1.3	2.7
	ENTRUST	Kentucky Senate President Robert Stivers on Wednesday asked a federal judge to withhold action against Davis for her defiance of the ruling by denying *legal marriage authorization* to gay couples, until the Kentucky General Assembly can act.	4.3	**4.7**	**4.3**
	HUMAN	Kentucky Senate President Robert Stivers on Wednesday asked a federal judge to withhold action against Davis for her defiance of the ruling by denying *marriage paperwork* to gay couples, until the Kentucky General Assembly can act.	**4.7**	3.0	3.7
A campaign of hate laced with blatant anti-semitic overtones	BART$_{D+EN}^{w/o}$	A campaign of hate laced with genuine anti-semitic overtones	**4.3**	**4.7**	2.3
	BART$_{EN}^{w/o}$	A campaign of trust laced with obvious anti-semitic overtones	2.0	1.7	**2.0**
	LEXREP	A campaign of intimidation laced with strong anti-semitic overtones	3.3	3.0	3.0
	GST	A campaign of *injuring rollercoaster* laced with anti - semitic overtones	2.3	1.0	3.0
	ENTRUST	A campaign of *prejudice* laced with *genuine* anti - semitic overtones	3.7	4.0	**2.0**
	HUMAN	A campaign with *clear* anti-semitic overtones which *does not promote peace*	2.7	2.7	3.7

Table 9: Examples of generated outputs from different systems (with human reframed argument as references) for arguments containing partisan collocations and appeal to fear, respectively. We show average scores (over three annotators) on a 1-5 scale with 1 denotes the worst and 5 be the best.