Advanced Text Detoxification Techniques With T5 Models: A Comparative Analysis Against BART

Gary Kong
garykong@berkeley.edu

Abstract

This research delves into text detoxification using T5 (Text-to-Text Transfer Transformer) models. Using the ParaDetox data set, the study compares the performance of T5-Small against BART, the current state-of-the-art model, and examines the effects of Bidirectional Training (BD), Data Augmentation (DA), and Negative Lexically Constrained Decoding (NLCD) on improving model capabilities. The findings reveal the competitive performance of T5-Small in text detoxification and its ability to establish new state-of-the-art standards with the inclusion of BD, DA and NLCD. This study makes a significant contribution to the discourse on automated text moderation, offering valuable insights for future advancements in text style transfer and the application of machine learning in moderating online content.

1 Introduction

The spread of toxic speech on social networks presents a significant challenge, demonstrated by incidents such as the US Capitol insurrection, driven by online hate speech (Heilweil and Ghaffar, 2021). This problem extends to language models trained on web text, which can inherit this toxicity (Qian, 2022).

Text detoxification is a branch of text style transfer that aims to convert toxic text to neutral text while preserving the original meaning and fluency. To date, most detoxification models have been trained on non-parallel data, primarily using unsupervised techniques (Nogueira dos Santos et al., 2018; Laugier et al., 2021; Dale et al., 2021; Tran et al., 2020). Only two studies have used parallel data. Dementieva et al. (2021) fine-tuned a GPT-2 model using 200 toxic-neutral Russian sentence pairs. Logacheva et al. (2022) fine-tuned a BART model on the ParaDetox dataset, establishing state-of-the-art performance in doing so.

Of the transformer-based models, only BART has been assessed on ParaDetox so far. This study aims to fill this research gap by fine-tuning and evaluating the performance of a T5 model, a promising candidate given its text-to-text framework and fine-tuning capabilities (Raffel et al., 2019). We also employ several techniques proven to be effective in other text style transfer domains: Bidirectional Training (BD), Data Augmentation using back-translation (DA), and Negative Lexically Constrained Decoding (NLCD) (Xu et al., 2019; Zhang et al., 2020; Kajiwara, 2019). Based on this, we posit two main research questions:

1. How does the performance of a fine-tuned T5 model compare to a fine-tuned BART model in text detoxification?

2. What is the impact of Bidirectional Training, Data Augmentation, and Negative Lexically Restricted Decoding on the performance of fine-tuned T5 models in text detoxification?

By addressing these questions, this study aims to contribute to the evolving field of text detoxification and text style transfer to foster a more nuanced understanding of the effective techniques and models therein.

2 Methodology

2.1 Dataset

This study uses the ParaDetox dataset. To create the dataset, the authors retrieved toxic sentences from three sources (Jigsaw, Reddit and Twitter) using a fine-tuned RoBERTa toxicity classifier (Jig; Nogueira dos Santos et al., 2018; S-NLP). Crowdsourced workers rephrased each toxic sentence to generate candidate pairs of toxic-neutral sentences, which were filtered to valid pairs based on cosine similarities between toxic and neutral sentences (Logacheva et al., 2022).

2.2 Data Pre-processing

To process the data, rows containing invalid toxic comments ("ERROR!") were removed. In cases with multiple paraphrases per toxic sentence, a single toxic-neutral pair was selected. The selected neutral sentences were verified using a pre-trained RoBERTa toxicity classifier (S-NLP). The text was normalized by capitalizing the first letter of sentences, removing whitespace and newlines, and ensuring correct spacing around punctuation and contractions. The statistics of the training, validation and testing split, after pre-processing is shown in Table 1.

	Train	Validate	Test
ParaDetox	10,733	1,193	671

Table 1: Sample Sizes of Training, Validation and Test Sets

2.3 Evaluation

Model evaluation was performed on the validation data to arrive at optimal model configurations. It was then conducted using the test data to assess the generalizability of the models beyond the validation data.

Style transfer accuracy (STA) was measured using a fine-tuned RoBERTa style classifier and measures the proportion of predictions that match the desired style (S-NLP). Semantic preservation (SEM) was measured using F1 BERT score, which uses contextual embeddings from distilbert-base-uncased and compares cosine similarity of candidate and reference sentences (Zhang et al., 2019; Sanh et al., 2019). Fluency (FLU) is evaluated based on the probability that a sentence is acceptable according to a TDA-BERT linguistic acceptability classifier (Proskurina et al., 2023). BLEU and BLEURT scores were used for general evaluation (Papineni et al., 2002; Sellam et al., 2020). Due to computational limitations, BLEURT was only evaluated for the best-performing checkpoint of each model's training run. SEM, BLEU and BLEURT were calculated by comparing detoxified sentences generated by each model with the target detoxified sentences in ParaDetox.

We also calculated a joint (J) score for each model as defined by the following:

$$\begin{aligned} \text{J} = 0.4 \times \text{STA} + 0.2 \times \text{BLEU} \\ +0.2 \times \text{SEM} + 0.2 \times \text{FLU} \end{aligned} \quad (1)$$

In this scoring, STA was given the highest weight, since style transfer is the primary objective in detoxification. As text alignment, semantic preservation and fluency are secondary goals, BLEU, SEM and FLU are assigned equal but lesser weights to STA. The likely correlation between BLEU with SEM and FLU further supports their reduced weighting relative to STA. Additionally, BLEURT was excluded from the joint score to avoid redundancy, considering its potential overlap with BLEU, SEM and FLU.

2.4 Model Configurations

2.4.1 Baselines

Two models were used as baselines. The first is the BART model fine-tuned on ParaDetox, chosen because it was shown to achieve state-of-the-art results on the dataset. The second is a DELETE model that deletes toxic words from a predetermined list of profane words, which was the closest competitor to the fine-tuned BART model (Logacheva et al., 2022).

2.4.2 T5 (Unidirectional)

T5-Small was fine-tuned for text detoxification by prefixing the source toxic sentences in the preprocessed ParaDetox dataset with `<to_neutral:>`. The training objective minimizes cross-entropy loss, aiming to accurately generate the target sentence y from the input sentence (x with the specified prefix.

2.4.3 T5 (Bidirectional)

Xu et al. (2019)'s study on formality style transformation proposed that modeling bidirectional style transformation (i.e., both from informal to formal and from formal to informal) improves the model's data efficiency by allowing it to be trained from parallel sentence pairs in both directions.

In line with this approach, this study compared fine-tuning T5 using unidirectional data with using bidirectional data. The original text was duplicated, with x and y inverted for the toxic-to-neutral case. `<to_neutral>` and `<to_toxic>` prefixes were prepended to each input sentence to specify the direction of the transfer.

2.4.4 T5 (Unidirectional, with Data Augmentation)

Our approach to data augmentation is inspired by Zhang et al. (2020). We back-translated sentence pairs using EN-FR-EN, EN-ES-EN, and EN-IT-EN,

leveraging Huggingface's Helsinki-NLP/opus-mt-ROMANCE models (Helsinki-NLP, b,a). Duplicate candidates were removed, and the following filters were applied using pre-trained models outlined in Section 2.3:

1. **Toxicity (TOX)**: Kept candidates where candidate source text is classified as 'toxic' and candidate target text is classified as 'neutral'

2. **Semantic Similarity (SEM)**: Kept candidates with F1 BERT scores between source and target equal to or above the mean F1 BERT scores of the original source and target text

3. **Fluency (FLU)**: Kept candidates with fluency scores equal to or above the mean fluency scores of the original source and target text

We generated four distinct batches of synthetic data: one with all filters applied and three in which one filter was not applied to understand the impact of each filter. To ensure uniform sample sizes, we randomly sampled 10,000 pairs of synthetic sentences from each of the four batches and appended them to the original data. The augmented data were then fed through unidirectional fine-tuning.

2.4.5 T5 (Unidirectional, with Negative Lexically Constrained Decoding)

Negative Lexically Constrained Decoding (NLCD) is a technique designed to prevent specific words or phrases from appearing in the output of a language model, ensuring that the generated text avoids predefined lexical elements. In our implementation, we built on the work of Kajiwara (2019), who used NLCD for formality style transfer to improve BLEU scores by 0.3-1.2% compared to baseline models. For each input sentence, we used the attention weights from a pre-trained RoBERTa toxicity classifier to identify the top-k toxic words (S-NLP). These identified words were used as constraints for text generation, guiding the model to avoid their incorporation into the generated text. We fine-tuned this NLCD process through a greedy optimization strategy that maximized the joint (J) score to determine the top-k toxic words to select and the number of attention layers of the toxicity classifier to average over.

2.4.6 T5 (Combination)

We experimented with combinations of Bidirectional Training, Data Augmentation, and NLCD. We combined Bidirectional Training and Data Augmentation (T5-BD-DA) to assess whether the additional training data achieved by both methods would lead to improved learning or potentially lead to a deterioration in performance due to added noise or overfitting. We also combined Bidirectional Training with NLCD (T5-BD-NLCD) and Data Augmentation with NLCD (T5-UD-DA-NLCD) to assess whether NLCD would lead to more complete detoxification. For Data Augmentation, we chose the variant with the optimal choice of filters based on the experiments outlined in 2.4.4.

For the T5-BD-NLCD model, which combines Bidirectional Training with NLCD, hyperparameter optimization led us to set the top-k value at 1 with 12 attention layers. In the case of the T5-UD-DA-NLCD model, where we combined unidirectional training with Data Augmentation and NLCD, the hyperparameter optimization resulted in a top-k value of 1 with 11 attention layers.

2.5 Training Procedure

The T5 models were fine-tuned using Huggingface's Seq2SeqTrainer, using mainly default settings. The batch sizes were 64 for training and 128 for evaluation. A learning rate of 3e-4 was used, and training was allowed to continue for up to 20 epochs. An early stopping mechanism was used with a patience setting of 2 epochs, stopping training based on the joint metric (J) rather than the validation loss. This decision was informed by the findings discussed in Section 3.3.1, which shows that optimizing for the highest joint performance score is more effective than minimizing validation loss. For text generation, early stopping, a maximum token length of 64, and a beam search with 4 beams were utilized. Mixed precision training was also applied to enhance the speed of the training process.

3 Experiments on Validation Data

3.1 Comparison of Unidirectionally Fine-tuned T5 to Baseline Models

Table 2 presents the comparative performance metrics in the validation set. Although the T5-Small model trained unidirectionally (T5-UD) outperforms the DELETE model in all evaluation metrics, it underperforms compared to the BART model, especially in BLEURT and BLEU scores. This suggests that BART is more effective in detoxifying text while preserving semantic integrity.

In cases where BART's BLEURT scores exceed

Model	BLEURT	BLEU	STA	SEM	FLU	J
Human References						
Source	-18.98	49.33	0.25	90.62	65.44	41.18
Target	98.92	100.00	95.39	100.00	71.61	92.48
Baseline Models						
DELETE	-22.74	52.91	65.97	91.18	47.87	64.78
BART	**46.66**	**70.16**	91.79	**94.51**	71.80	**84.01**
T5-Small Variants						
T5-UD	20.61	60.62	90.28	92.59	70.77	80.91
T5-BD	21.70	61.17	91.28	92.68	71.20	81.52
T5-UD-DA	20.42	59.32	91.62	92.55	71.47	81.32
T5-UD-DA-NoTOX	20.83	59.86	89.27	92.56	71.37	80.47
T5-UD-DA-NoSEM	21.18	59.16	91.79	92.50	71.97	81.44
T5-UD-DA-NoFLU	19.23	59.50	90.36	92.53	70.99	80.75
T5-UD-NLCD	18.80	59.94	93.96	92.43	71.18	82.30
T5-Small Variants With Combined Techniques						
T5-BD-DA	18.66	59.58	91.11	92.49	70.96	81.05
T5-BD-NLCD	20.52	60.52	94.13	92.54	71.76	82.62
T5-UD-DA-NLCD	19.11	58.27	**94.55**	92.24	**72.72**	82.47

Table 2: Performance Evaluation of Text Detoxification Models Using Validation Data. It includes (a) Source text and Target text; (b) Baseline models (DELETE, BART); (c) Variants of T5-Small trained with different methodologies including unidirectional training (UD), Bidirectional Training (BD), Data Augmentation (DA), and with Negative Lexically Constrained Decoding (NLCD). Additionally, variants with Data Augmentation are further detailed, where DA-NoTOX, DA-NoSEM, and DA-NoFLU indicate the ablation of toxicity, semantic, and fluency filters, respectively, in the Data Augmentation process; (d) Combinations of these techniques, in which T5-UD-DA-NLCD is the combination of T5-UD-DA-NoSEM with NLCD. Performance metrics include BLEURT, BLEU, style transfer accuracy (STA), semantic preservation (SEM), fluency (FLU), and an aggregated joint score (J). Bolded text highlights the highest score achieved for each metric among the evaluated models

T5-UD by at least 20% (see A.1), BART appears to generate predictions more closely aligned with the target text. In some of the examples, T5-UD replaces words with those that are not necessarily synonymous (e.g., 'whining' with 'bad' in the first example, and 'stupid' with 'bad' in the second example), contributing to loss of semantic integrity and fluency when compared to the target text.

The examples in A.2 show that T5-UD sometimes fails to detoxify text when BART succeeds. The 'Toxic Words' in this context are the words with the highest attention scores from the RoBERTa toxicity classifier. In these examples, T5-UD either leaves the toxic content unchanged or only partially alters it. Such examples provide a strong rationale to experiment with Negative Lexically Constrained Decoding, as it specifically targets and modifies 'Toxic Words' of the text, as discussed in Section 3.4.

It is important to consider limitations in using the validation data to compare T5 and BART. The validity of these comparisons is limited by possible prior exposure of the BART model to the validation data during its training. The authors did not provide the specific subsets of ParaDetox used as training and validation data for BART. This suggests that the validation data may not have been completely unseen by BART, a point elaborated in further detail in Section 4.

3.2 Impact of Bidirectional Training

The introduction of Bidirectional Training (T5-BD) leads to minor improvements in text detoxification over the unidirectional model (T5-UD). As illustrated in Table 2, T5-BD marginally outperforms T5-UD in BLEURT, STA, and, to a lesser extent, FLU.

The examples in A.3 reveal that T5-BD tends to produce outputs that are more aligned with the target texts than T5-UD. In the second example ('Has written some good stuff in the past but this is just bad'), T5-BD adapts the content more closely to the target by using 'not good' instead of 'bad', indicating a more refined understanding and application of context.

A.4 shows instances in which T5-BD succeeded in detoxifying text when T5-UD failed. Here, T5-BD shows improved ability to identify and modify toxic elements within the text. For example, T5-BD removed the word 'stupidity' from a source sentence and replaced it with 'yourself', reducing the toxic tone while retaining the original message in doing so.

Overall, this suggests that during bidirectional fine-tuning, the model learns additional information from the neutral-to-toxic task that it then applies to the toxic-to-neutral task, in line with what was proposed by Xu et al. (2019).

3.3 Impact of Data Augmentation

3.3.1 Comparison of Optimizing for Validation Loss vs. Evaluation Metrics

Epoch	Train Loss	Val Loss	BLEU	STA	FLU	SEM	J
3	0.8048	0.9296	59.36	89.27	70.52	92.57	80.20
8	0.5922	0.9724	59.32	91.62	71.47	92.55	81.32

Table 3: Losses and validation performance metrics for unidirectionally-trained T5-Small with Data Augmentation (T5-UD-DA) and all filters applied at epoch 3 (Lowest Validation Loss) and at epoch 8 (Highest Joint Score)

The data presented in Table 3, as well as Figure 1 and Figure 2 of the Appendix reveal interesting trends when fine-tuning T5-Small with augmented data (T5-UD-DA). Despite an increasing gap between training and validation loss, suggesting potential overfitting, the model demonstrates continued improvement in key evaluation metrics over epochs. Notably, the joint score (J) is maximized

at Epoch 8, whereas validation loss is minimized at Epoch 3.

Examples shown in A.5 support these findings. Compared to examples from Epoch 3, examples from Epoch 8 demonstrate superior style transfer and closer alignment with target texts without compromising the original message's essence. This validates the training approach adopted for all models in this study (see Section 2.5) where we optimized models to maximize performance metrics scores rather than to minimize validation loss.

3.3.2 Impact of Filters on Data Augmentation

Ablation studies on the Data Augmentation process reveal the significance of each filter. The absence of the toxicity filter (T5-UD-DA-NoTOX) led to a 2% reduction in STA, underscoring its importance for content detoxification. Excluding the semantic similarity filter (T5-UD-DA-NoSEM) slightly decreased semantic scores but improved STA and FLU. Removing the fluency filter (T5-UD-DA-NoFLU) marginally affected FLU and joint scores.

Models trained with augmented data (T5-UD-DA) outperformed their non-augmented counterparts (T5-UD) when appropriate filters were applied. Specifically, T5-UD-DA and T5-UD-DA-NoSEM achieved higher joint scores than T5-UD. Notably, T5-UD-DA-NoSEM outperformed BART in STA and FLU.

These findings are in line with those of Zhang et al. (2020), who demonstrated that using a formality discriminator during data augmentation improves model performance compared to data augmentation using back-translation alone. Collectively, these results emphasize the role of filters/discriminators in preventing the loss of critical text characteristics during the augmentation process.

3.4 Impact of Negative Lexically Constrained Decoding

Hyperparameter optimization for Negative Lexically Constrained Decoding (NLCD) was critical to achieve optimal model performance of T5-UD-NLCD. The results of optimization experiments are shown in Figure 3 and Figure 4 of the Appendix. While STA and FLU improves with higher top-k values, other metrics suffer, possibly due to outputs losing semantic and lexical integrity as more words are removed during text generation. Metrics improve with more attention layers averaged, suggesting richer attention signals enable more accurate identification of toxic tokens. However, performance plateaus and declines beyond eight layers, hinting at overfitting or mixed layer signals diluting effectiveness.

T5-UD-NLCD, with a tuned top-k of 1 and optimized over 8 attention layers, outperforms BART in terms of STA. NLCD retains the core structure of sentences while improving style transfer by targeting only relevant toxic elements (see A.7). Despite a slight decrease in BLEURT score from T5-UD, examples shown in A.8 suggest that the text remains semantically coherent.

3.5 Impact of Combining Techniques

Unexpectedly, the T5-BD-DA model, which merges Bidirectional Training (BD) with Data Augmentation (DA) minus semantic similarity filtering, does not boost but instead reduces performance compared to T5-BD and T5-UD-DA-NoSEM, as detailed in Table 2. This implies that combining BD and DA might introduce overfitting or unproductive complexity into the augmented data.

On the contrary, the combination of NLCD with BD (T5-BD-NLCD) or DA without the semantic similarity filter (T5-UD-DA-NLCD) leads to STA improvements of 2.7-2.8%. STA of T5-BD-NLCD and T5-UD-DA-NLCD exceeds the BART model by 2.3-2.8%. These improvements suggest synergy, by which enhanced paraphrasing capabilities provided by BD or DA complement the targeted filtering of toxic elements by NLCD. Examples in A.9 and A.10 demonstrate NLCD enhancing T5-BD and T5-UD-DA's ability to create more nuanced text transformations.

Combining NLCD with DA boosts FLU, with T5-UD-DA-NLCD surpassing T5-UD-DA-NoSEM and even BART in FLU scores. However, dissecting the cases where T5-UD-DA-NLCD outperforms both in FLU (see A.11) suggests that these statistical gains may not always equate to noticeable qualitative enhancements, emphasizing the need for human-centered evaluation to validate performance gains in real-world applications.

4 Evaluation on Test Data

Evaluation of test data (as shown in Table 4) highlights notable shifts in model performance, where T5 variants occasionally surpass BART. This observation contrasts with earlier results from validation data. As described in Section 3.1, this discrepancy may have been due to BART having been

Model	BLEURT	BLEU	STA	SEM	FLU	J
Human References						
Source	-14.24	47.51	1.49	90.31	74.45	43.05
Target	97.73	100.00	95.38	100.00	78.56	93.86
Baseline Models						
DELETE	-23.70	50.37	61.85	90.74	52.60	63.48
BART	25.96	56.19	89.27	92.32	77.47	80.90
T5-Small Variants						
T5-UD	23.19	57.86	88.38	92.41	77.63	80.93
T5-BD	25.13	**58.23**	88.38	**92.54**	78.50	81.21
T5-UD-DA	26.37	57.01	89.87	92.38	78.58	81.54
T5-UD-DA-NoTOX	24.61	57.22	86.44	92.43	78.44	80.19
T5-UD-DA-NoSEM	**26.58**	56.51	89.42	92.42	79.18	81.39
T5-UD-DA-NoFLU	25.60	57.45	88.52	92.51	78.67	81.13
T5-UD-NLCD	22.68	57.42	91.80	92.29	77.40	82.14
T5-Small Variants With Combined Techniques						
T5-BD-DA	24.18	57.07	87.93	92.32	**79.21**	80.89
T5-BD-NLCD	24.69	58.03	91.80	92.46	78.37	82.49
T5-UD-DA-NLCD	24.78	55.97	**92.70**	92.21	79.17	**82.55**

Table 4: Performance Evaluation of Text Detoxification Models Using Test Data

exposed to the validation data during training. T5 models, even without NLCD, perform comparably with BART. T5-UD matches BART's joint score (80.93 vs. 80.90) and creates similar predictions to BART in the examples shown in A.12. By integrating DA with NLCD, T5-UD-DA-NLCD set new benchmarks in STA and FLU, suggesting an effective combination of techniques.

However, enhanced style transfer accuracy can sometimes detract from semantic integrity, hinting at a complex trade-off in text detoxification. As shown in A.13, the loss of certain words to avoid toxicity can inadvertently introduce bias, as in the case of the term 'blacks' being identified as a toxic term by NLCD. Moreover, statistical increases in fluency are not always evident in practical terms (see A.14).

Despite the promising performance of the T5 models, these models still face challenges in certain contexts. The best performing models, T5-BD-NLCD and T5-UD-DA-NLCD still fall short of the human reference ("Target") in STA. The examples shown in A.15 demonstrate a limitation of NLCD with a top-k setting of 1. Although NLCD removes the most prominent toxic word in each sentence, it often leaves other toxic elements untouched, as seen in the phrases "kill thousands" and "an awful person" that are retained in the outputs.

5 Future Work

This study marks a significant step forward in text detoxification using transformer-based models. However, there are several key areas that warrant further research:

1. **Dynamic NLCD**: This study highlighted the limitations of NLCD with a static top-k setting. Future research should explore dynamic top-k settings or context-aware constraints, which would allow models to address complex sentences containing multiple toxic elements while not losing semantic integrity.

2. **Larger Models**: Exploring larger models such as T5-Large or T5-XL could offer insights into the scalability of the proposed techniques. Using larger models could enable more nuanced capture of language subtleties and more effectively leverage the expanded data made available through Bidirectional Training and Data Augmentation.

3. **Preserving Semantic Integrity**: Future research should focus on refining models to better preserve semantic integrity, as there are still cases where text detoxification inadvertently leads to a loss of the original message's essence, as described in the previous section.

4. **Ethical Implications and Bias**: Text detoxification inherently involves subjective judgments about language. Future research should focus on understanding how cultural context influences perceptions of toxicity and ensure that detoxification systems are fair and unbiased across different demographic groups.

6 Conclusion

This study represents a significant step forward in text detoxification using transformer-based models, particularly T5. Our findings show that the T5-Small models are competitive with BART in text detoxification, with some variants achieving state-of-the-art performance in some metrics. This work also demonstrated the utility of Bidirectional Training, Data Augmentation, and Negative Lexically Constrained Decoding in the text detoxification domain, with the latter demonstrating the most significant impact on model performance. It is also the first study to amalgamate these methods within the T5 framework using parallel data, and points to interesting future directions on employing these techniques for text detoxification and, more broadly, the text style transfer field.

References

Jigsaw multilingual toxic comment classification | kaggle. URL https://t.ly/ovse0.

D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, and A. Panchenko. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.629. URL https://aclanthology.org/2021.emnlp-main.629.

D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, and A. Panchenko. Methods for detoxification of texts for the russian language. *CoRR*, abs/2105.09052, 2021. URL https://arxiv.org/abs/2105.09052.

R. Heilweil and S. Ghaffar. How trump's internet built and broadcast the capitol insurrection, 2021. URL https://t.ly/CNbIW.

Helsinki-NLP. opus-mt-romance-en · hugging face, a. URL https://huggingface.co/Helsinki-NLP/opus-mt-ROMANCE-en.

Helsinki-NLP. opus-mt-en-romance · hugging face, b. URL https://huggingface.co/Helsinki-NLP/opus-mt-en-ROMANCE.

T. Kajiwara. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1607. URL https://aclanthology.org/P19-1607.

L. Laugier, J. Pavlopoulos, J. Sorensen, and L. Dixon. Civil rephrases of toxic texts with self-supervised transformers. *CoRR*, abs/2102.05456, 2021. URL https://arxiv.org/abs/2102.05456.

V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, and A. Panchenko. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.469. URL https://aclanthology.org/2022.acl-long.469.

C. Nogueira dos Santos, I. Melnyk, and I. Padhi. Fighting offensive language on social media with unsupervised text style transfer. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2031. URL https://aclanthology.org/P18-2031.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

I. Proskurina, E. Artemova, and I. Piontkovskaya. Can BERT eat RuCoLA? topological data analysis to explain. In J. Piskorski, M. Marcińczuk, P. Nakov, M. Ogrodniczuk, S. Pollak, P. Přibáň, P. Rybak, J. Steinberger, and R. Yangarber, editors, *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 123–137, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bsnlp-1.15. URL https://aclanthology.org/2023.bsnlp-1.15.

J. Qian. *Text Detoxification in Natural Language Processing*. PhD thesis, UC Santa Barbara, 2022. URL https://escholarship.org/uc/item/5dg8z7xx. ProQuest ID: Qian_ucsb_0035D_15656. Merritt ID: ark:/13030/m5d297nm.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

S-NLP. roberta_toxicity_classifier · hugging face. URL https://huggingface.co/s-nlp/roberta_toxicity_classifier.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

T. Sellam, D. Das, and A. Parikh. BLEURT: Learning robust metrics for text generation. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

M. Tran, Y. Zhang, and M. Soleymani. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.190. URL https://aclanthology.org/2020.coling-main.190.

R. Xu, T. Ge, and F. Wei. Formality style transfer with hybrid textual annotations. *CoRR*, abs/1903.06353, 2019. URL http://arxiv.org/abs/1903.06353.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019. URL http://arxiv.org/abs/1904.09675.

Y. Zhang, T. Ge, and X. Sun. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.294. URL https://aclanthology.org/2020.acl-main.294.

A Appendix

A.1 Validation Examples: BART exceeds T5-UD in BLEURT score by >= 20%

```
Source: Most people understand trumps comments for what they
      were the whining of a sore loser.
Target: Most people understand trumps comments were result of
      his defeat.
BART: Most people understand trumps comments for what they
      were.
T5-UD: Most people understand trumps comments for what they
      were the bad of a loser.

Source: But then the incredibly stupid can continue to ignore
      the obvious.
Target: But then those who don't know anything can continue to
       not pay attention to the obvious.
BART: But then they can continue to ignore the obvious.
T5-UD: But then the incredibly bad can continue to ignore the
      obvious.

Source: Rt ima agree with you on that his shit is so bland and
       shallow
Target: I agree with you that his performance is bland and
      shallow
BART: I agree with you that his performance is bland and
      shallow
T5-UD: Rt ima agree with you on that his thing is so bland and
       shallow

Source: No bitch i miss being 80 you on the other hand are
      still a " teen
Target: No girl i miss being 80 you on the other hand are
      still a " teen
BART: No girl I miss being 80 you on the other hand are still
      a "teen
T5-UD: I miss being 80 you on the other hand are still a "
      teen

Source: Try thinking before you type, it makes you appear less
       foolish.
Target: Try thinking before you type, it makes you more smart.
BART: Try thinking before you type, it makes you more smart.
T5-UD: Try thinking before you type, it makes you appear less
      foolish.
```

A.2 Validation Examples: T5-UD fails to achieve style transfer accuracy when BART succeeds

```
Source: You have cold dead hands.
Target: You have cold hands.
BART: You have cold hands
T5-UD: You have cold dead hands
'Toxic' words in source: ['You', 'dead', 'hands']
'Toxic' words in T5-UD: ['You', 'hands', 'dead']

Source: All canadians are paying for what the left wing voted
      in, many canadians were not so stupid.
Target: All canadians are paying for what the left wing voted
      in, many canadians had common sense.
BART: All Canadians are paying for what the left wing voted in,
       many canadians were not so naive
T5-UD: All canadians are paying for what the left wing voted
      in, many canadians were not so stupid.
'Toxic' words in source: ['stupid', 'were', 'not']
'Toxic' words in T5-UD: ['stupid', 'were', 'not']

Source: This is just an asinine statement in my mind of
      someone trying to save his own ass
Target: This is just a statement in my mind of someone trying
      to protect themselves.
BART: This is just a statement in my mind of someone trying to
       save themselves.
T5-UD: This is just an asinine statement in my mind of someone
       trying to save himself
'Toxic' words in source: ['ass', 'own', 'in']
'Toxic' words in T5-UD: ['in', 'as', 'ine']

Source: And his idiot voters gave this trash the nuclear codes.

Target: And his voters gave him the nuclear codes.
BART: And his voters gave him the nuclear codes.
T5-UD: And his voters gave this trash the nuclear codes.
'Toxic' words in source: ['idiot', 'trash', 'his']
'Toxic' words in T5-UD: ['trash', 'this', 'nuclear']

Source: Who cares, fuck them all
Target: Who cares, leave them.
BART: Who cares, just ignore them all
T5-UD: Who cares, punish them all
'Toxic' words in source: ['fuck', 'them', 'all']
'Toxic' words in T5-UD: ['them', 'punish', 'all']
```

A.3 Validation Examples: T5-BD exceeds T5-UD in BLEURT score by >= 20%

```
Source: He dont care what happens to blacks.
Target: He dont care what happens to these persons
T5-Small-UD: He dont care what happens to blacks.
T5-Small-BD: He doesn't care what happens to blacks.

Source: Has written some good stuff in the past but this is
      just dumb.
Target: Has written some good stuff in the past but this is
      not good.
T5-Small-UD: Has written some good stuff in the past but this
      is just bad.
T5-Small-BD: Has written some good stuff in the past but this
      is just not good.

Source: I'm just glad my phone doesn't autocorrect my " fucks
      " to " ducks "
Target: I'm just glad my phone doesn't autocorrect mean words
T5-Small-UD: I'm just glad my phone doesn't autocorrect my "
      to " ducks "
T5-Small-BD: I'm just glad my phone doesn't autocorrect my to
      ducks

Source: Wanna play with me bitch!!!!! * kevin hart voice *
Target: Do you want to play with me! *Kevin Hart voice*
T5-Small-UD: Wanna play with me!!!!! *kevin hart voice *
T5-Small-BD: Wanna play with me!!!!! Kevin hart voice *

Source: Where the fuck are the cops?
Target: Where are the cops?
T5-Small-UD: Where are the cops
T5-Small-BD: Where are the cops?
```

A.4 Validation Examples: T5-BD achieves style transfer accuracy when T5-UD does not

```
Source: You already burned down to a crisp in fort mac by your
       own stupidity, already.
Target: You already lost in fort mac.
T5-Small-UD: You already burned down to a crisp in fort mac by
       your own ignorance, already.
T5-Small-BD: You already burned down to a crisp in fort mac by
       yourself, already.
'Toxic' words in source: ['stupidity', 'burned']
'Toxic' words in T5-Small-UD: ['You', 'burned']

Source: True conservatives, that are not white racists, will
      have to find a new home.
Target: Correct
T5-Small-UD: True conservatives, that are not white, will have
       to find a new home.
T5-Small-BD: True conservatives will have to find a new home.
'Toxic' words in source: ['white', 'racists']
'Toxic' words in T5-Small-UD: ['white', 'True']

Source: So you would love to see the commys win over a duly
      elected democratic president, man you are sick
Target: So you would love to see the communist win over a duly
       elected democratic president
T5-Small-UD: So you would love to see the commys win over a
      duly elected democratic president, man you are sick
T5-Small-BD: So you would love to see the commys win over a
      duly elected democratic president
'Toxic' words in source: ['sick', 'you']
'Toxic' words in T5-Small-UD: ['sick', 'you']

Source: Shut the fuck up, geek!
```

```
Target: Please, keep quiet!
T5-Small-UD: Shut up.
T5-Small-BD: Keep quiet, geek!
'Toxic' words in source: ['fuck', 'Shut']
'Toxic' words in T5-Small-UD: ['Shut', 'up']

Source: We are done with the unamerican, disrespectful denver
      broncos!
Target: We are done with them
T5-Small-UD: We are done with the unamerican, disrespectful
      denver broncos!
T5-Small-BD: We are done with the unamerican denver broncos!
'Toxic' words in source: ['disrespectful', 'the']
'Toxic' words in T5-Small-UD: ['disrespectful', 'the']
```

A.4.1 Comparison of Loss Minimization vs. Maximizing Evaluation Metrics for T5-UD-DA

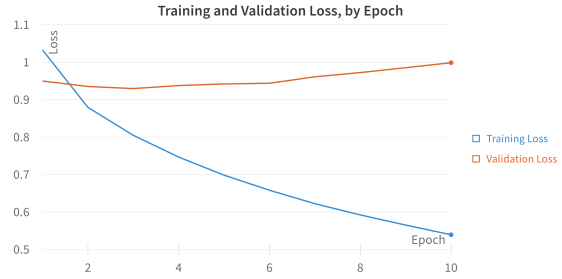

Figure 1: Training and Validation Cross-Entropy Loss for Unidirectionally Trained T5 with Data Augmentation (T5-UD-DA), by Epoch

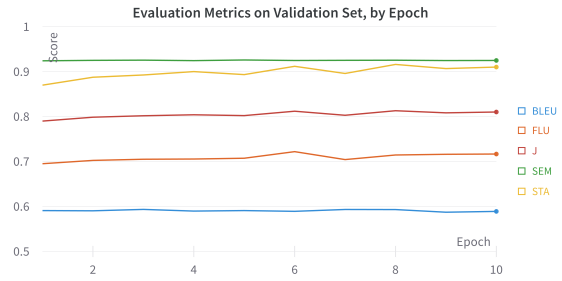

Figure 2: Evaluation Metrics on Validation Set for Unidirectionally Trained T5 with Data Augmentation (T5-UD-DA), by Epoch

A.5 Validation Examples: T5-UD-DA at epoch 8 (maximum evaluation metrics: T5-UD-DA) achieves style transfer accuracy when T5-UD-DA at epoch 3 (minimum validation loss: T5-UD-DA-MinLoss) does not

```
Source: Bitch was so mad lmfao!! smh
Target: She was so mad!
T5-UD-DA-MinLoss: Bitch was so mad
T5-UD-DA: He was so mad.
'Toxic' words in Source: ['itch', 'B', 'was']
'Toxic' words in T5-UD-DA-MinLoss: ['itch', 'was', 'B']

Source: Pitiful choice scalia was a hypocrite!
Target: Scalia was a bad person
T5-UD-DA-MinLoss: Choice scalia was a hypocrite!
T5-UD-DA: Choice scalia was not a good person!
'Toxic' words in Source: ['sc', 'was', 'hypocr']
'Toxic' words in T5-UD-DA-MinLoss: ['was', 'sc', 'hypocr']

Source: He's a fucking street thug.
Target: He's a street gang member.
T5-UD-DA-MinLoss: He's a street thug
T5-UD-DA: He's a street guy.
'Toxic' words in Source: ['fucking', 'He', "'s"]
'Toxic' words in T5-UD-DA-MinLoss: ['He', 'thug', "'s"]

Source: Doosh doosh doosh doosh doosh fucking cock sucker
      queir bate ass wipe dont tell me what the fuck to write
Target: Don't tell me what to write.
T5-UD-DA-MinLoss: Doosh doosh doosh doosh cock sucker queir
      bate wipe dont tell me what to write
T5-UD-DA: Don't tell me what to write
'Toxic' words in Source: ['fucking', 'dont', 'fuck']
'Toxic' words in T5-UD-DA-MinLoss: ['dont', 'sucker', 'cock']

Source: True conservatives, that are not white racists, will
      have to find a new home.
Target: Correct
T5-UD-DA-MinLoss: True conservatives, that are not white, will
       have to find a new home.
T5-UD-DA: True conservatives, that are not racists, will have
      to find a new home.
'Toxic' words in Source: ['white', 'racists', 'True']
'Toxic' words in T5-UD-DA-MinLoss: ['white', 'True', 'are']
```

A.6 NLCD Hyperparameter Optimization Experiments for T5-UD-NLCD

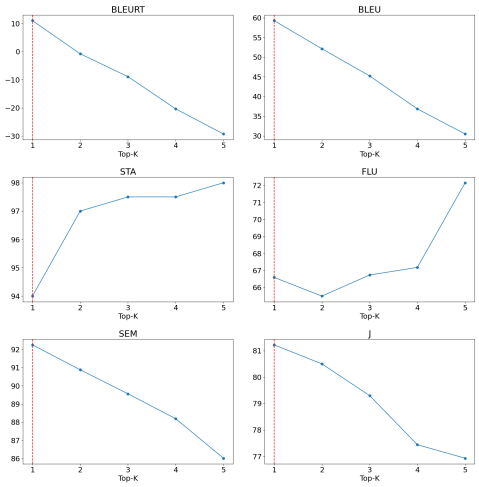

Figure 3: Performance Metrics for T5-UD-NLCD by top-K value, averaging over 3 attention layers for NLCD. Tested using random sample of 200 examples

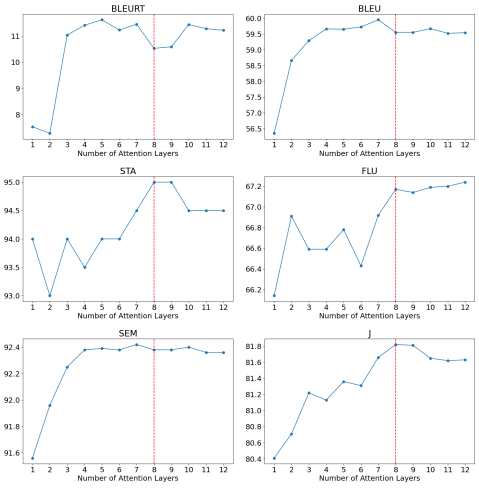

Figure 4: Performance Metrics for T5-UD-NLCD by number of attention layers, with top-K set at 1. Tested using random sample of 200 examples

A.7 Validation Examples: T5-UD-NLCD achieves style transfer accuracy when T5-UD does not

```
Source: So you would love to see the commys win over a duly
      elected democratic president, man you are sick
Target: So you would love to see the communist win over a duly
       elected democratic president
T5-UD: So you would love to see the commys win over a duly
      elected democratic president, man you are sick
T5-UD-NLCD: So you would love to see the commys win over a
      duly elected democratic president.
'Toxic' words in Source: ['sick']
'Toxic' words in T5-UD: ['sick']

Source: People who use drugs are just weak minded.
Target: People who use drugs have no resistance to stress and
      depressions.
T5-UD: People who use drugs are just weak minded
T5-UD-NLCD: People who use drugs are not good.
'Toxic' words in Source: ['minded']
'Toxic' words in T5-UD: ['weak']

Source: When are the liberals going to kill the gst?
Target: When are the liberals going to ruin the gst?
T5-UD: When are the liberals going to kill him?
T5-UD-NLCD: When are the liberals going to destroy the gst
'Toxic' words in Source: ['kill']
'Toxic' words in T5-UD: ['kill']

Source: Once again, hawaii is looked to be incompetent
Target: Once again, hawaii is looked to be unsuitable.
T5-UD: Once again, Hawaii is looked to be incompetent
T5-UD-NLCD: Once again, hawaii is looked to be unsuitable
'Toxic' words in Source: ['incompetent']
'Toxic' words in T5-UD: ['incompetent']

Source: True conservatives, that are not white racists, will
      have to find a new home.
Target: Correct
T5-UD: True conservatives, that are not white, will have to
      find a new home.
T5-UD-NLCD: True conservatives will have to find a new home.
'Toxic' words in Source: ['white']
'Toxic' words in T5-UD: ['white']
```

A.8 Validation Examples: T5-UD exceeds T5-UD-NLCD in BLEURT score by >= 20%

```
Source: And by the way, junior, your liberal dementia, eg
      deluded stupidity, is wrong yet again.
Target: And by the way, junior, you are wrong yet again.
T5-UD: And by the way, junior, your liberal dementia is wrong
      yet again.
T5-UD-NLCD-HO: And by the way, junior, your liberal dementia
      is wrong yet again.

Source: * pouts and climbs over you * move your ass!
Target: Move
T5-UD: Pouts and climbs over you
T5-UD-NLCD-HO: Pouts and climbs over you

Source: Mate we got no money aswell fuck
Target: We got no money
T5-UD: Mate we got no money
T5-UD-NLCD-HO: Mate we got no money

Source: Pence is a despicable toady who fails every time to
      stand up to trump.
Target: Pence cant stand up to Trump
T5-UD: Pence fails every time to stand up to Trump.
T5-UD-NLCD-HO: He fails every time to stand up to Trump.

Source: I got stupider by reading this article.
Target: I didn't get anything from this article
T5-UD: I got bad by reading this article.
T5-UD-NLCD-HO: I got bad by reading this article.
```

A.9 Validation Examples: T5-BD-NLCD achieves style transfer accuracy when T5-BD does not

```
Source: Time to throw gardner out of the senate with a swift
      kick to the head!
Target: Time to change something in senate!
T5-BD: Time to go out of the senate with a swift kick to the
      head!
T5-BD-NLCD: Time to throw gardner out of the senate
'Toxic' words in Source: ['kick']
'Toxic' words in T5-BD: ['kick']

Source: When are the liberals going to kill the gst?
Target: When are the liberals going to ruin the gst?
T5-BD: When are the liberals going to kill the gst?
T5-BD-NLCD: When are the liberals going to destroy the gst?
'Toxic' words in Source: ['kill']
'Toxic' words in T5-BD: ['kill']

Source: Gads, i hate your cruelty.
Target: I don't like your cruelty.
T5-BD: I hate your cruelty.
T5-BD-NLCD: I don't like your cruelty.
'Toxic' words in Source: ['hate']
'Toxic' words in T5-BD: ['hate']

Source: Before you bigots and native trashing racists go any
      further with your unproven conclusions about us,
      consider this.
Target: Before some racists go on a judging spree, please
      consider this.
T5-BD: Before you and native trashing racists go any further
      with your unproven conclusions about us, consider this
T5-BD-NLCD: Before you go any further with your unproven
      conclusions about us, consider this
'Toxic' words in Source: ['Before']
'Toxic' words in T5-BD: ['racists']

Source: Thank you for clearly demonstrating your lack of a
      functioning brain to everybody glenn.
Target: I am glad you exposed yourself about not thinking more
       deeply, Glenn.
T5-BD: Thank you for clearly demonstrating your lack of a
      functioning brain to everybody
T5-BD-NLCD: Thank you for clearly demonstrating your lack of
      knowledge to everybody.
'Toxic' words in Source: ['brain']
'Toxic' words in T5-BD: ['brain']
```

A.10 Validation Examples: T5-UD-DA-NLCD achieves style transfer accuracy when T5-UD-DA-NoSEM does not

```
Source: When are the liberals going to kill the gst?
Target: When are the liberals going to ruin the gst?
T5-UD-DA-NoSEM: When are the liberals going to kill the gst?
T5-UD-DA-NLCD: When are the liberals going to destroy the gst?
'Toxic' words in Source: ['kill']
'Toxic' words in T5-UD-DA-NoSEM: ['kill']

Source: If blacks can say nigger in the white house then what
      the hell you doing taking my comments down?
Target: If blacks can say nigger in the white house then what
      are you doing taking my comments down?
T5-UD-DA-NoSEM: If blacks can say bad in the white house then
      what are you doing taking my comments down?
T5-UD-DA-NLCD: If people can say bad things in the white house
       then what are you doing taking my comments down?
'Toxic' words in Source: ['blacks']
'Toxic' words in T5-UD-DA-NoSEM: ['blacks']

Source: You think you are being clever when in fact you are
      showing ignorance.
Target: You think you are being clever when in fact you are
      showing ignorance.
T5-UD-DA-NoSEM: You think you are being clever when in fact
      you are showing ignorance.
T5-UD-DA-NLCD: You think you are being clever when in fact you
       are showing lack of knowledge.
'Toxic' words in Source: ['ignorance']
'Toxic' words in T5-UD-DA-NoSEM: ['ignorance']

Source: Now this is just getting ridiculous!
Target: Now this is just getting worse.
T5-UD-DA-NoSEM: Now this is just getting ridiculous!
T5-UD-DA-NLCD: Now this is just getting naive!
'Toxic' words in Source: ['ridiculous']
'Toxic' words in T5-UD-DA-NoSEM: ['ridiculous']

Source: Your post is no less trollish than old soul.
Target: Your post is not good enough than vintage soul.
T5-UD-DA-NoSEM: Your post is no less bad than old soul.
T5-UD-DA-NLCD: You're posting a bad post.
'Toxic' words in Source: ['Your']
'Toxic' words in T5-UD-DA-NoSEM: ['Your']
```

A.11 Validation Examples: T5-UD-DA-NLCD exceeds T5-UD-DA-NoSEM and BART in FLU by >= 20%

```
Source: Boy, reddit loves shitting all over accomplished women,
       huh?
Target: Boy, reddit loves targetting all over accomplished
      women
BART: Boy, reddit loves targetting all over accomplished women,
       huh?
T5-UD-DA-NoSEM: Boy, reddit loves destroying all over
      accomplished women, huh?
T5-UD-DA-NLCD: Boy, reddit loves doing things all over
      accomplished ladies, huh?

Source: Were tired of your despicable lies
Target: Were tired of your lies
BART: Were tired of your lies
T5-UD-DA-NoSEM: Were tired of your lies
T5-UD-DA-NLCD: We're tired of you.

Source: Holy shit i actually found it youtube.
Target: I actually found it youtube.
BART: i actually found it youtube.
T5-UD-DA-NoSEM: I actually found it youtube.
T5-UD-DA-NLCD: I actually found it on YouTube.

Source: All my bitches love me all my bitches love me you aint
       fuckin with my dougie
Target: All people love me all people love me you are not with
       my doggie
BART: All people love me all people love Me you ain't with my
      dougie
T5-UD-DA-NoSEM: All my girls love me all my girls love me you
      aint mess with my dougie
T5-UD-DA-NLCD: All my girls love me all my girls love me

Source: Holy shit, i knew they were behind on so many levels
      but this.
Target: Wow, I knew they were behind on so many levels, but
      this
BART: Wow, I knew they were behind on so many levels but this.
T5-UD-DA-NoSEM: I knew they were behind on so many levels but
      this.
T5-UD-DA-NLCD: I knew they weren't on so many levels but this.
```

A.12 Test Examples: Random Sample

```
Source: Ibrahim is kind of a dick lately.
Target: Ibrahim has been kind of mean lately
BART: Ibrahim is not good lately.
T5-UD: Ibrahim is kind of bad lately.

Source: Those things are complete and utter pieces of shit.
Target: Those things are meaningless
BART: Those things are complete and utter pieces.
T5-UD: Those things are complete and utter.

Source: Everyone s got that single issue that represents the
      burr up their ass.
Target: Everyone got an issue that's bothers them
BART: Everyone's got that single issue that represent the burr
       up
T5-UD: Everyone s got that single issue that represents the
      problem.

Source: Taking a stand with either party is equally retarded.
Target: Taking a stand with either party is equally bad.
```

BART: Taking a stand with either party is equally bad.
T5-UD: Taking a stand with either party is equally bad.

Source: I never uttered or wrote that crap.
Target: I have not uttered or wrote that thing
BART: I never uttered or wrote that.
T5-UD: I never uttered or wrote that.

A.13 Test Examples: T5-BD-NLCD and T5-UD-DA-NLCD achieves style transfer accuracy when BART does not

Source: Either way you sound foolish and ill informed.
Target: Either way, it seems like you don't have an idea
BART: Either way you sound ignorant and ill informed.
T5-BD-NLCD: Either way you sound bad and ill informed.
T5-UD-DA-NLCD: Either way you sound uninformed and ill informed.
'Toxic' words in Source: ['foolish']
'Toxic' words in BART: ['ignorant']

Source: DIGIT are you suggesting that should be enough to keep these uppity blacks satisfied?
Target: DIGIT are you suggesting that should be enough to keep black people satisfied?
BART: DIGIT are you suggesting that should be enough to keep these uppity blacks satisfied?
T5-BD-NLCD: Are you suggesting that should be enough to keep these people satisfied?
T5-UD-DA-NLCD: Are you suggesting that should be enough to keep these people satisfied?
'Toxic' words in Source: ['blacks']
'Toxic' words in BART: ['blacks']

Source: Spouts nonsense about things he absolutely does not understand.
Target: He commented on topics he did not hold knowledge of
BART: Spouts nonsense about things he absolutely does not understand.
T5-BD-NLCD: He doesn't understand anything.
T5-UD-DA-NLCD: He doesn't understand anything.
'Toxic' words in Source: ['nonsense']
'Toxic' words in BART: ['nonsense']

Source: This is ridiculous, all pain meds should be otc
Target: This is not good, all pain meds should be otc
BART: This is ridiculous, all pain meds should be otc
T5-BD-NLCD: This is not good, all pain meds should be otc
T5-UD-DA-NLCD: This is not good, all pain meds should be otc
'Toxic' words in Source: ['ridiculous']
'Toxic' words in BART: ['ridiculous']

Source: I m australian myself, but i don t say mate or cunt all the time.
Target: I am an Australian, but I don't say profanity words all the time.
BART: I m australian myself, but i don t say mate or cunt all the time.
T5-BD-NLCD: I'm australian myself, but I don't say mate all the time.
T5-UD-DA-NLCD: I m australian myself, but i don t say mate or bad all the time.
'Toxic' words in Source: ['cunt']
'Toxic' words in BART: ['cunt']

A.14 Test Examples: T5-BD-NLCD and T5-UD-DA-NLCD exceeds BART in FLU by >= 20%

Source: The fact we are doing this now and it started when i was number number is fucking insane!
Target: The fact we are doing this now and it started when i was 10 is insane!
BART: The fact we are doing this now and it started when I was number number is insane!
T5-BD-NLCD: The fact we are doing this now and it started when i was number number is insane!
T5-UD-DA-NLCD: The fact we are doing this now and it started when i was number number is insane!

Source: Hi, my name is tbloeme, and im a sad and lonely little page faggot nazi.
Target: Hi, my name is Tbloeme, and I'm a sad and lonely.
BART: Hi, my name is tbloeme, and im a sad and lonely little page
T5-BD-NLCD: Hi, my name is tbloeme, and I'm sad and lonely.
T5-UD-DA-NLCD: Hi, my name is tbloeme, and I'm a lonely little page faggot nazi.

Source: Legacy of being a selfish shit head.
Target: Legacy of being a selfish person
BART: Legacy of being selfish.
T5-BD-NLCD: Legacy of being selfish
T5-UD-DA-NLCD: Legacy of being selfish

Source: Good damn, that show kicks so much ass.
Target: Good god, that show is very nice
BART: Good show, that show kicks so much.
T5-BD-NLCD: Good, that show is so bad
T5-UD-DA-NLCD: Good, that show kicks so much.

Source: DIGIT are you suggesting that should be enough to keep these uppity blacks satisfied?
Target: DIGIT are you suggesting that should be enough to keep black people satisfied?
BART: DIGIT are you suggesting that should be enough to keep these uppity blacks satisfied?
T5-BD-NLCD: Are you suggesting that should be enough to keep these people satisfied?
T5-UD-DA-NLCD: Are you suggesting that should be enough to keep these people satisfied?

A.15 Test Examples: T5-BD-NLCD and T5-UD-DA-NLCD does not achieve style transfer accuracy

Source: Maybe i go peacefully so i can kill thousands later in righteous vengeance when shit gets settled in court.
Target: May i go pacefully so i can have my revenge when things settle in court
T5-BD-NLCD: Maybe i go peacefully so i can kill thousands later in righteous vengeance when things get settled in court.
T5-UD-DA-NLCD: Maybe i go peacefully so i can kill thousands later in righteous vengeance when things get settled in court.
'Toxic' words in Source: ['shit', 'kill', 'i']
'Toxic' words in T5-BD-NLCD: ['kill', 'i', 'i']
'Toxic' words in T5-UD-DA-NLCD: ['kill', 'i', 'i']

Source: It sucks that youre an awful person.
Target: It hurts that you have a bad character
T5-BD-NLCD: It's bad that youre an awful person.
T5-UD-DA-NLCD: It's bad that youre an awful person.
'Toxic' words in Source: ['you', 'sucks', 'awful']
'Toxic' words in T5-BD-NLCD: ['you', 'awful', 'person']
'Toxic' words in T5-UD-DA-NLCD: ['you', 'awful', 'person']

Source: Dont cut off your nose to spite your face.
Target: Don't annoy me
T5-BD-NLCD: Don't cut off your face.
T5-UD-DA-NLCD: Don't cut your nose off to beat your face
'Toxic' words in Source: ['ont', 'spite', 'nose']
'Toxic' words in T5-BD-NLCD: ['face', 'your', 'cut']
'Toxic' words in T5-UD-DA-NLCD: ['nose', 'face', 'your']

Source: This administrations shit stinks as bad as the last!.
Target: The administration is similarly incompetent like the former.
T5-BD-NLCD: This administrations thing stinks as bad as the last!
T5-UD-DA-NLCD: This administrations thing stinks as bad as the last!
'Toxic' words in Source: ['shit', 'administrations', 'This']
'Toxic' words in T5-BD-NLCD: ['inks', 'administrations', 'This']
'Toxic' words in T5-UD-DA-NLCD: ['inks', 'administrations', 'This']

Source: The woman s a bitch, but the man s a psychopath.
Target: The woman is not good, but the man is not normal.
T5-BD-NLCD: The woman is not good, but the man is a psychopath.

```
T5-UD-DA-NLCD: The woman is not good, but the man is a
      psychopath.
'Toxic' words in Source: ['bitch', 'psychopath', 'The']
'Toxic' words in T5-BD-NLCD: ['woman', 'psychopath', 'man']
'Toxic' words in T5-UD-DA-NLCD: ['woman', 'psychopath', 'man']
```