Learning from Bootstrapping and Stepwise Reinforcement Reward: A Semi-Supervised Framework for Text Style Transfer

Zhengyuan Liu, Nancy F. Chen
Institute for Infocomm Research, A*STAR, Singapore
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

Text style transfer is an important task in controllable language generation. Supervised approaches have pushed performance improvement on style-oriented rewriting such as formality conversion. However, challenges remain due to the scarcity of large-scale parallel data in many domains. While unsupervised approaches do not rely on annotated sentence pairs for each style, they are often plagued with instability issues such as mode collapse or quality degradation. To take advantage of both supervised and unsupervised paradigms and tackle the challenges, in this work, we propose a semi-supervised framework for text style transfer. First, the learning process is bootstrapped with supervision guided by automatically constructed pseudo-parallel pairs using lexical and semantic-based methods. Then the model learns from unlabeled data via reinforcement rewards. Specifically, we propose to improve the sequence-to-sequence policy gradient via stepwise reward optimization, providing fine-grained learning signals and stabilizing the reinforced learning process. Experimental results show that the proposed approach achieves state-of-the-art performance on multiple datasets, and produces effective generation with as minimal as 10% of training data.

1 Introduction

Text style transfer is a task in natural language generation, which aims to automatically control certain attributes during sentence paraphrasing, such as formality, sentiment, and humor (Rao and Tetreault, 2018; Li et al., 2018). Style transfer has many practical applications, such as altering emotions of spoken utterances, removing biases in transcripts, and conveying politeness in messages (Hovy, 1987). The key for a successful rewrite is to preserve the semantic content of the source sentence, while transforming it to a particular target style without sacrificing fluency and grammatical accuracy. Therefore, the performance of style transfer models is commonly assessed on both style accuracy and content preservation. When large-scale annotated sentence pairs are available, training neural sequence-to-sequence models via supervised learning shows impressive generation quality (Rao and Tetreault, 2018; Lai et al., 2021). However, in many use cases, it is unfeasible to adopt supervised approaches because parallel samples are unavailable. To address data insufficiency bottlenecks, various unsupervised approaches have been proposed for text style transfer, including learning disentangled representations of style and content (Shen et al., 2017) and adopting pairwise back-translation (Prabhumoye et al., 2018). Recently, reinforcement learning (RL) is introduced to develop unsupervised models such that rewards of content preservation and style conversion are used to optimize sequence generation (Luo et al., 2019; Gong et al., 2019). However, RL-based methods are often challenging to train in practice. For instance, the rewards have high variance during early stages when learning from scratch, which affects the training stability; and they cannot provide fine-grained learning signals as traditional token-level maximum likelihood estimation, since they are often calculated on the entire generated sequence (de Masson d'Autume et al., 2019). As a result, models are prone to mode collapse and often fail to produce acceptable generations in reality.

Herein, we propose a semi-supervised framework for text style transfer, and optimize it on training stability and signal fineness. Our semi-supervised model uses a small amount of parallel data for supervised learning, and gets further improvement by learning from a large amount of unlabeled data. In contrast to prior work that often relies on human-annotated parallel pairs like (Chawla and Yang, 2020), the approach we propose bootstraps the training process with automatically constructed pseudo parallel data. Two pseudo pair

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2633 - 2648

matching methods are investigated: a lexical-based strategy, which is straightforward by calculating the token-level overlap; and a semantic-based strategy, which uses semantic similarity as criteria and would have better general potential.

Furthermore, to obtain fine-grained signals for the RL-based sequence-to-sequence training process, we propose a stepwise reward re-weighting strategy. This is inspired by the observation that the style transfer weights are not uniform across tokens/spans in the source sentence: some tokens weigh more during attribute-guided text style transfer (Li et al., 2018). Therefore, instead of using the reward (e.g., style strength scores) calculated from the entire generated sentence (Luo et al., 2019; Lai et al., 2021), we use the token-level reward. Specifically, we extract attribute-related attentive scores from a pre-trained style discriminator, obtain a stepwise reward by re-weighting the sequence-level score, and utilize it as a fine-grained signal for policy gradient back-propagation.

We evaluate the proposed framework that incorporates both supervision and reward-based learning on three style transfer corpora (Section 4). Experiments show that our model achieves state-of-the-art performance. Particularly, the proposed model can produce reasonable generations with only 10% training data on the Yelp and Amazon corpora, and it also outperforms the supervised baselines when applying on the well-annotated GYAFC dataset.

2 Related Work

Neural Text Style Transfer The aim of text style transfer is to automatically convert text to a certain style while preserving the content (McDonald and Pustejovsky, 1985; Hovy, 1987). It has many applications, like persona-based dialogue generation (Niu and Bansal, 2018). Recently, neural sequence-to-sequence architectures becomes popular for this task. When parallel data are available, supervised training with cross-entropy loss is typically applied (Rao and Tetreault, 2018). However, annotated data are hard to obtain in many use cases, thus learning from non-parallel corpora has become an active research area. There are two approaches: (1) Disentangling style and content by learning a distinct representation for each element. For example, variational autoencoders are first used to transform a sentence into a low-dimension hidden state. Then the attribute-related latent representation is extracted to guide the decoder for target style generation (Shen et al., 2017; Fu et al., 2018; John et al., 2019); (2) Back translation, which uses cyclic reconstruction to improve content preservation (Zhang et al., 2018; Prabhumoye et al., 2018; Lample et al., 2019; Luo et al., 2019). For model optimization, some studies focus on applying reinforcement learning (RL), which defines a reward from a style classifier or a reward from back-translation to enhance style strength and content preservation (Gong et al., 2019; Luo et al., 2019; Wu et al., 2019; Sancheti et al., 2020). Recently, large-scale pre-trained language models are introduced to improve generation quality (Radford et al., 2019), and have been incorporated in both semi-supervised (Chawla and Yang, 2020) and supervised approaches (Lai et al., 2021). In this work, we use the BART (Lewis et al., 2020) as our language model backbone.

Pseudo Data Augmentation To tackle the data scarcity challenge in text style transfer, one solution is to build pseudo pairs from massive non-parallel data. Zhang et al. (2020b) proposed several augmentation methods for pre-training a Transformer-based model and fine-tuning on human annotations. Wang et al. (2019) proposed using harness-rule-based pre-processing, and joint training of bi-directional transfer and auto-encoder with two auxiliary losses (Wang et al., 2020). Jin et al. (2019) and Nikolov and Hahnloser (2019) constructed the pseudo corpora by iteratively matching via cosine similarity of sentence embeddings and hierarchical alignment. In this work, we use pseudo data as weak-supervision to bootstrap the training process, and further combine it with RL-based learning.

Attribute Salience Assessment In template-based and prototype editing methods for text style transfer, attribute marker detection is used to label the salient words and spans (Li et al., 2018). Aside from n-gram statistical features, neural attention-based methods train attribute-related classifiers, and consider words with attention weights higher than average as markers (Bahdanau et al., 2015; Xu et al., 2018; Sudhakar et al., 2019). Zhou et al. (2020) use the attribute salient scores as one of the model prediction output. To the best of our knowledge, we are the first to employ token-level attribute salience scores for reward re-weighting on policy gradient for sequence generation, and prior work only focuses on using attribute markers for text manipulation such as token replacement and template construction (Niu and Bansal, 2018).

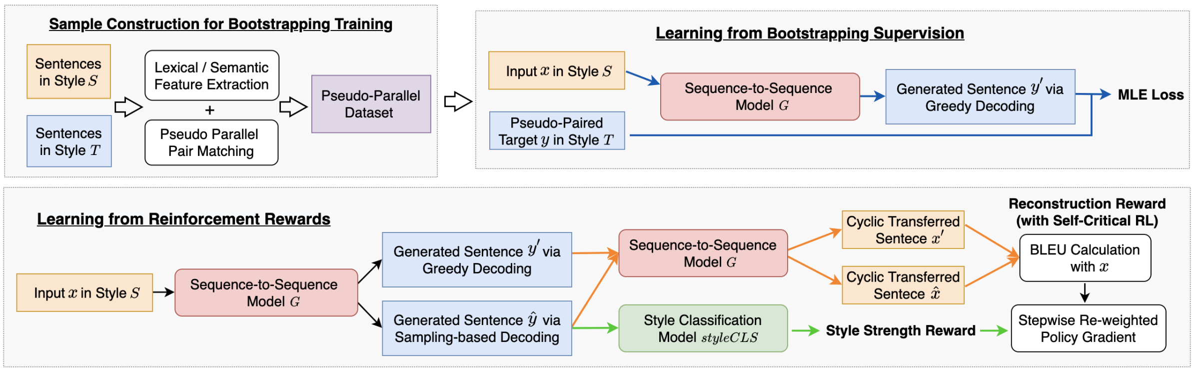

Figure 1: Overview of the proposed framework. Text samples in two different styles are in yellow and in blue. The sequence-to-sequence model is shared by style transfer and cyclic generation. The MLE loss, reconstruction reward, and style reward flows are in blue, yellow, and green arrow lines, respectively. See Algorithm 1 for training process.

3 Methodology

Define S as the source style and T as the target style (e.g., $S = negative$, $T = positive$). Let $\mathcal{D}_S$ and $\mathcal{D}_T$ be the two datasets which are comprised of sentences in each style respectively. The style transfer system, denoted as a text encoding-decoding model G, is to generate sentences in the target style. The goal is formulated to maximize $P(\boldsymbol{y}|\boldsymbol{x};\theta_G)$, where θ_G are the model parameters. In our setting, we make the rewriting bidirectional, i.e. it can be used to transfer source style to target style and verse versa. In this case, an additional input $\boldsymbol{c} \in \{S, T\}$ is fed to G specifying the style to which the sentence is to be converted. Hence, the objective is to maximize $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c};\theta_G)$.

3.1 Framework Overview

The overview of our proposed semi-supervised framework is shown in Figure 1. Given the non-parallel datasets $\mathcal{D}_S$ and $\mathcal{D}_T$, we use lexical or semantic features for pseudo parallel pair matching. The training process consists of two stages: (1) the generator model G is trained on the pseudo parallel samples, where cross-entropy loss over the target sentence tokens is used to optimize generated output probabilities, i.e. the **bootstrapping** step; (2) we incorporate reconstruction and style rewards to enhance attribute rewriting and content preservation, where reinforcement learning is used to optimize the generation, i.e. the **reward-based learning**. Moreover, the second stage can use pseudo parallel pairs as well as the non-parallel samples.

3.2 Pseudo Parallel Data Construction

To build the pseudo parallel data for bootstrapping, we investigate lexical similarity and semantic similarity for sentence matching.

Lexical Similarity In text style transfer, rewriting is often accomplished by changing a few words or phrases that are indicative of a particular attribute in the source sentence, namely attribute markers, while leaving the rest largely unaltered (Li et al., 2018). For example, *"Moving past the shape, they were dry and truly tasteless."*, a sentence with a negative sentiment style, can be transferred to a positive style by changing or replacing sentiment-specific words *"dry"* and *"tasteless"*, while keeping other words intact. This intuition has inspired the template-based and editing-based rewriting approaches (Li et al., 2018). Here we employ it for the lexical feature extraction. First, from unaligned corpora of two styled subsets (e.g., positive, negative), we identify attribute markers by sorting phrases that occur with far higher frequency within one attribute than the other (e.g., *"worst"* and *"very disappointed"* are negative markers). Second, for each sentence in the two subsets, we remove those markers, and regard the remaining words as its content-preserved spans. Then we match the content-preserved spans of style S to those of style T with the smallest Levenshtein editing distance (see examples shown in Table 1).

Semantic Similarity While the lexical features are straightforward and computationally-efficient, it may not generalize well in some tasks like formality conversion due to the ubiquitous span paraphrasing. Therefore, in this paper, we introduce semantic features for the pseudo data construction. While samples in different styles stand in different polarities, they are expected to be similar in the content-level semantic space. More specifically, for a sample i in style S, we match it to the closest sentence in style T in a semantic space. We use an unsupervised sentence representation

Source Sentence: if there were a way to put no stars, i would! **Lexical Match:** i'd give it more stars if i could. **Semantic Match:** love love love, if i could give you _num_ stars i would.
Source Sentence: the manager sat us at our table, and she seemed very angry. **Lexical Match:** the manager and employees are very nice. **Semantic Match:** the manager alice herself came by our table and greeted us as well.
Source Sentence: furthermore, i would rather drive _num_ minutes more to concord to race there. **Lexical Match:** furthermore, they have a nice bar that goes both indoor and outdoor. **Semantic Match:** i drive _num_ minutes to get here and it is definitely worth it!

Table 1: Pseudo parallel sentence pairs extracted from Yelp sentiment transfer dataset. Source sentences are from the negative polarity set, and are matched to sentences from the positive set.

model with contrastive learning (Gao et al., 2021), which achieves comparable performance to the supervised sentence embedding models, and calculate cosine similarity to measure the distance.[1] As shown in Table 1, the pseudo parallel data are similar at the semantic level, and they can be used as weak-supervision samples.

3.3 Learning with Supervision

With the pseudo parallel data, we can conduct supervised learning with token-level maximum likelihood estimation (MLE). In our framework, we use a sequence-to-sequence neural network. Since the large-scale pre-trained language models boost the performance of various downstream tasks, we use BART (Lewis et al., 2020) as the language backbone, which is a denoising autoencoder with strong language generation capability. Given a source sentence $\boldsymbol{x}$ and a reference sentence $\boldsymbol{y}$, the cross-entropy loss is calculated between the decoder's output and the reference sentence:

$$L_{\text{MLE}} = -\Sigma_i \log(p(\boldsymbol{y}_i|\boldsymbol{y}_{1:i-1}, \boldsymbol{x}, \boldsymbol{c}; \theta_G)) \quad (1)$$

Moreover, to avoid the generation becoming overfitting to the pseudo parallel data, we add the label smoothing on the cross-entropy loss (Müller et al., 2019), with the smoothing weight λ = 0.15.

3.4 Learning with Rewards

Upon the supervised learning from the pseudo parallel data, the model can be further improved by unsupervised learning from the massive unlabeled data. For the unsupervised stage, we adopt reinforcement learning, and use two rewards to enhance style rewriting and content preservation.

Reconstruction Reward Back translation has proved effective to improve content preservation, we feed the transferred sentence to model G for the backward rewriting, and calculate reconstruction reward on the cyclic generation. Here we measure the reward based on BLEU (Papineni et al., 2002) score as in (Sancheti et al., 2020) to foster content preservation, and adopt policy gradient (Sutton et al., 1999) with Self-Critical Sequence Training to reduce the variance (Rennie et al., 2017):

$$R_{cyclic} = \text{score}(G(\boldsymbol{y}'), \boldsymbol{x}) - \text{score}(G(\hat{\boldsymbol{y}}), \boldsymbol{x}) \quad (2)$$

where $\boldsymbol{x}$ is the backward target, $G(\hat{\boldsymbol{y}})$ is the back-translated output from greedy decoding generation $\hat{\boldsymbol{y}}$, and $G(\boldsymbol{y}')$ is the back-translated from sampling-based generation $\boldsymbol{y}'$ over a multi-nominal distribution. Noted that the score function can also be ROUGE and language model perplexity. The former is more suitable for summarization tasks; the latter needs additional computation.

Style Classification Reward Aside from content preservation, we use a style strength reward to optimize the model. We train a Transformer model for the binary style classification, and use it to evaluate how well the transferred sentence $\boldsymbol{y}'$ matches the target style. The style reward is R_{style} defined as the classification score:

$$p(s_{style}|\boldsymbol{y}') = \text{softmax}(\text{styleCLS}(\boldsymbol{y}', \phi)) \quad (3)$$

where styleCLS denotes the style classifier, ϕ are the parameters of the classifier, which are fixed during the training of the generation framework. $\boldsymbol{y}'$ is the generated sentence by sampling from the multi-nominal distribution at each step. Then, the reward-based learning is conducted via Policy Gradient (Sutton et al., 1999) back-propagation:

$$R = \lambda_{cyclic} R_{cyclic} + \lambda_{style}(R_{style} - \gamma) \quad (4)$$

$$\nabla_{\theta_G} J = E[R \cdot \nabla_{\theta_G} log(P(\boldsymbol{y}'|\boldsymbol{x}, \boldsymbol{c}; \theta_G))] \quad (5)$$

where R is the sum of cyclic and style reward, $\boldsymbol{y}'$ is the generated sentence by sampling from the multi-nominal distribution at each step, θ_G are trainable parameters of the generator, the weight ratio λ are added on cyclic and style reward separately, and γ is a style reward penalty (see Table 9). The overall objectives for θ_G are the loss of the base model (Eq. 1) and the policy gradient of RL rewards (Eq. 5).

[1] Additionally, we observed that in some corpora like Amazon (Li et al., 2018), there are a number of samples labeled with incorrect style due to data noise, and the semantic approach is sensitive on this issue. Therefore, we use a style classifier to filter out the incorrectly clustered samples.

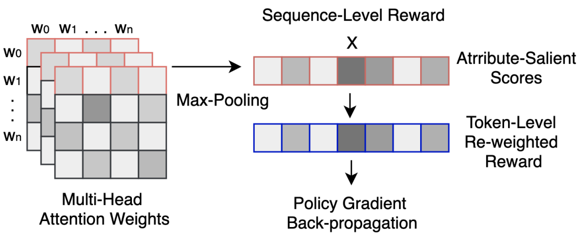

Figure 2: Our proposed stepwise reward re-weighting.

Stepwise Reward Re-weighting When applying reinforcement learning algorithms on sequence-to-sequence training, it is difficult for models to conduct end-to-end back-propagation due to the discrete nature of text. One of the common solutions is adopting policy gradient optimization (Sutton et al., 1999), where the rewards are generally calculated on the whole output sequence. Since all generated tokens obtain the same reward value, this coarse-grained signal is suboptimal for learning performance and stability (de Masson d'Autume et al., 2019). For instance, when positive sentiment is targeted, the output sentence *"I dislike this movie!"* will obtain a negative reward of style strength if its gold reference is *"I love this movie!"*. In this context, the word *"dislike"* should be punished more than the others in the sentence, but with sequence-level reward all words receive the same penalty. To address this drawback, we propose a solution by granulating the sequence-level reward with token-level salience scores, namely, stepwise re-weighting.

To re-weight the coarse-grained reward, we use the normalized attentive scores from the style classification model as the token-level attribute-salient scores. For the Transformer architecture, it is shown that heavily attended tokens correlate strongly with tokens that are indicative of the target style (Hewitt and Manning, 2019; Vig and Belinkov, 2019). Since the softmax linear layer is used over the attention stack of the first token $\langle s \rangle$ in a *'RoBERTa-base'* model, the attention weights of other input tokens that correspond to $\langle s \rangle$ are of special interest in identifying significant sentence tokens. We inspect the attentions computed by the Transformer with 12 multi-head layers, and empirically observed that the attention weights of top layers correlate strongly with salient tokens (see the visualization in Appendix Figure 4). Given the attention matrix A_i in the i-th multi-head layer, a_i^j represents the attention vector of the first token (e.g., $\langle s \rangle$, *"[CLS]"*) from the j-th attention head, which is normalized across all tokens. We max-pool A_i over all attention heads to form a_i, which represents the maximum extent to which each token was attended to by any head, and further max-pool the weights across the top-2 layers as the final stepwise attribute-salient scores (see layer selection in Section 5.2), which are in the range of (0, 1). Then sequence-level rewards are expanded to the token length n, and re-weighted by the stepwise scores (see Figure 2 and Algorithm 1), and the policy gradient is formulated as following:

$$\nabla_{\theta_G} J = E[\frac{1}{n}\sum_{t=1}^{n} R'_t \cdot \nabla_{\theta_G} log(P(\boldsymbol{y}'_t | \boldsymbol{y}'_{1:t-1}, \boldsymbol{x}, \mathbf{c}; \theta_G))] \tag{6}$$

Corpus	Train	Valid	Test
Yelp (Sentiment-Positive)	270K	2,000	500
Yelp (Sentiment-Negative)	180K	2,000	500
Amazon (Sentiment-Positive)	277K	985	500
Amazon (Sentiment-Negative)	278K	1,015	500
GYAFC E&M (Formality-Paired)	52.6K	2,877	1,416
GYAFC F&R (Formality-Paired)	51.9K	2,788	1,432

Table 2: Statistics of the style transfer datasets. The GYAFC Entertainment&Music (E&M) and Family&Relationships (F&R) are comprised of paired samples. For Yelp and Amazon, only their test sets include human-written parallel references.

4 Experiments

4.1 Experimental Datasets

For extensive experiments, in this paper, we select three representative text style transfer corpora: Yelp (business reviews), Amazon (product reviews), and Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Li et al., 2018; Rao and Tetreault, 2018). The training, validation, and test split are the same as previous work (Luo et al., 2019; Chawla and Yang, 2020), and their task types and statistics are shown in Table 2. In the non-annotated corpora Yelp and Amazon, human-written references are only available for the test set. Therefore, to build the pseudo parallel data described in the previous section, we filter out the sentence pairs with lexical or semantic similarity lower than a threshold, and remove sentences that are shorter than 5 words. The pseudo parallel set is used for the bootstrapping training (Section 3.3), and the rest samples are used for the unsupervised stage (Section 3.4).

4.2 Experiment Setup

The framework is implemented with Pytorch and Hugging Face Transformers[2]. The *'BART-base'*

[2]https://github.com/huggingface/transformers

Model	Accuracy	BLEU	G2	H2	BertScore
Cross Aligned (Shen et al., 2017)	75.3	17.9	36.7	28.9	68.3
Back Translation (Prabhumoye et al., 2018)	95.4	5.0	21.9	9.6	61.0
Style Embedding (Fu et al., 2018)	8.7	42.3	19.2	14.4	78.1
Multi-Decoding (Fu et al., 2018)	50.2	27.9	37.4	35.9	69.4
Unpaired (Xu et al., 2018)	64.9	37.0	49.0	47.1	73.7
Delete+Retrive (Li et al., 2018)	89.0	31.1	52.6	46.1	71.3
Template-Based (Li et al., 2018)	81.8	45.5	61.0	58.5	73.7
Unsupervised MT (Zhang et al., 2018)	95.4	44.5	65.1	60.7	80.8
DualRL (Luo et al., 2019)	85.6	55.2	68.7	67.1	84.1
IterativeMatch (Jin et al., 2019)	91.7	23.3	46.2	37.1	71.4
Deep Latent **w/ Language Models** (He et al., 2019)	85.2	46.4	62.8	60.0	76.4
Direct Rewards **w/ GPT-2** (Liu et al., 2021)	91.2	53.8	70.0	67.6	83.6
Only Lexical Pseudo Data Bootstrapping (30K Pairs)	81.3	26.5	46.4	39.9	72.1
Lexical Pseudo + Reward-Learning (30K)	81.1	50.4	63.9	62.1	82.1
Lexical Pseudo + Reward-Learning (100K)	86.2	59.4	**71.5**	**70.3**	**87.3**
Only Semantic Pseudo Data Bootstrapping (30K Pairs)	82.9	23.9	44.5	37.1	71.8
Semantic Pseudo + Reward-Learning (30K)	83.5	49.6	64.3	62.2	82.5
Semantic Pseudo + Reward-Learning (100K)	86.5	59.8	**71.9**	**70.7**	**87.1**

Table 3: Automatic evaluation scores on the Yelp sentiment style transfer task. Baseline results are reported with the model generations provided in published studies. Text examples are shown in Appendix Table 11.

model is selected as the generator G. For style classification, *'RoBERTa-base'* is used. We fine-tune models with AdamW (Kingma and Ba, 2015) with batch size 32; initial learning rates are all set at $2e^{-5}$. Style reward penalty γ is 0.2. Values for λ are set to 1.0 for style reward and 0.8 for cyclic reward. Beam search size is set at 6. Test results are reported with best validation scores (see Appendix Table 9 for environment and hyper-parameter setting details, and Algorithm 1 for the training process).

As previous work (Luo et al., 2019; He et al., 2020; Sancheti et al., 2020), we adopt the following evaluation metrics: (1) **Style Accuracy** is calculated via binary classification to measure the style strength of re-writing. While the *TextCNN* (Kim, 2014) is used in previous studies, we also adopt a Transformer *'RoBERTa-base'* classifier, where the reported scores are similar in our settings; (2) **BLEU** score is calculated on the prediction and human references to measure the content preservation; (3) We also compute the geometric mean (**G2**) and harmonic mean (**H2**) of style accuracy and BLEU score; (4) Since recent metrics with semantic similarity show better correlation with human judgments than traditional lexical measures. We also calculate **BertScore** between generation and references (Zhang et al., 2020a).

4.3 Results on Yelp Corpus

A number of representative unsupervised baseline models are selected for extensive comparison on the Yelp corpus: (1) models that adopt content-style disentanglement such as Cross Aligned (Shen et al., 2017) and Style Embedding (Fu et al., 2018); (2) models that adopt back-translation such as Unsupervised MT (Zhang et al., 2018), and Dual RL (Luo et al., 2019), and recent state-of-the-art models Deep Latent (He et al., 2019) and Direct Rewards w/ GPT-2 (Liu et al., 2021). For our semi-supervised framework, we first (1) apply vanilla supervised learning to assess the effectiveness of the pseudo parallel data construction; (2) bootstrap the model with 30K pseudo parallel pairs, then further train it via reward-based learning; (3) apply semi-supervised learning by bootstrapping the model with 30K pseudo parallel pairs, and using 70K non-parallel samples for the reward-based training. As shown in Table 3, vanilla supervised training on the 30K pseudo parallel data lead to favorable scores of style accuracy, though they do not perform well in terms of BLEU scores, as the pseudo pairs emphasize style converting rather than content preservation. Further training with rewards improves both the style accuracy and BLEU score, and models with both lexical and semantic pseudo data produce comparable results with only 30k samples. Performance is further improved by using additional non-parallel data (70k samples), where our models outperform state-of-the-art baselines significantly.

4.4 Results on Amazon Corpus

For the Amazon sentiment transfer corpus, we adopt the same training strategies described in Section 4.3. Aside from unsupervised models, we also select the semi-supervised model Semi-LM-MMI

Model	Accuracy	BLEU	G2	H2	BertScore
Cross Aligned (Shen et al., 2017)	74.1	0.4	5.4	0.8	55.3
Style Embedding (Fu et al., 2018)	43.3	10.0	20.8	16.2	68.1
Multi-Decoding (Fu et al., 2018)	68.3	5.0	18.4	9.3	18.2
Template-Based (Li et al., 2018)	68.7	27.1	43.1	38.9	85.5
Delete+Retrieve (Li et al., 2018)	48.0	22.8	33.1	30.9	83.7
Word-level Conditional GAN (Lai et al., 2019)	77.4	6.7	22.7	12.3	-
Semi-LM-MMI **w/ BART-Large** (Chawla and Yang, 2020)	68.9	28.6	44.4	40.4	-
Direct Rewards **w/ GPT-2** (Liu et al., 2021)	68.3	38.6	51.3	49.3	72.1
Only Lexical Pseudo Data Bootstrapping (30K Data)	79.8	16.4	36.1	27.2	63.3
Lexical Pseudo + Reward-Learning (30K)	71.2	36.1	50.6	47.9	73.4
Lexical Pseudo + Reward-Learning (100K)	73.1	46.3	**58.1**	**56.6**	78.4
Only Semantic Pseudo Data Bootstrapping (30K Data)	81.2	10.3	28.9	18.2	60.5
Semantic Pseudo + Reward-Learning (30K)	72.3	35.5	50.6	47.6	72.7
Semantic Pseudo + Reward-Learning (100K)	74.1	45.4	**58.0**	**56.3**	78.1

Table 4: Automatic evaluation scores on the Amazon sentiment style transfer task. Baseline results are calculated and reported with the model generations provided in published studies. See examples in Appendix Table 12.

	E&M Domain				F&R Domain			
Model	Accuracy*	BLEU	G2	H2	Accuracy*	BLEU	G2	H2
Human Reference (Rao and Tetreault, 2018)	81.5	100.0	90.2	89.8	80.5	100.0	89.7	89.2
Rule-Based (Rao and Tetreault, 2018)	29.7	72.4	46.4	42.1	82.1	65.8	73.4	73.1
Hybrid Annotations (Xu et al., 2019)	28.8	69.2	44.6	40.6	34.8	74.3	50.8	47.3
Semi-LM-MMI **w/ BART-Large** (Chawla and Yang, 2020)	30.4	76.5	48.2	43.5	30.6	79.9	49.4	44.2
Rewarded **BART-Large** (Lai et al., 2021)	75.1	76.5	75.7	75.7	74.6	79.2	76.8	76.8
Only Labeled Data Supervision (Full)	75.0	71.2	73.1	73.1	73.7	72.5	73.1	73.1
Labeled Data + Reward-Learning (30K)	75.7	71.4	73.5	73.4	72.4	74.4	73.3	73.4
Labeled Data + Reward-Learning (Full)	82.2	71.0	**76.3**	**76.2**	80.5	74.2	**77.3**	**77.2**

Table 5: Automatic evaluation scores on the GYAFC formality transfer task of baselines and our framework. Baseline results are reported with the generations provided as in (Chawla and Yang, 2020). *The style accuracy is calculated with a fine-tuned *'RoBERTa-base'* model (see Appendix for the result with *TextCNN* classifier).

w/ BART (Chawla and Yang, 2020), which adopted a language model-based discriminator for maximizing token-level conditional probabilities for training. Due to label noise in online-crawled data, the style accuracy for all models becomes lower than those trained on Yelp, and the classifier precision is only 86% (see Table 4). We also observed that the lexical similarity of pseudo parallel pairs is smaller than Yelp samples, and results in lower BLEU scores, especially when we apply supervised training on the 30K pseudo parallel data. On the other hand, content preservation largely benefits from the reward-based learning. Unsurprisingly, after bootstrapping, training with rewards significantly improves the generation quality, and our framework achieves state-of-the-art performance. Moreover, bootstrapping with lexical-based and semantic-based pseudo data resulted in a similar final performance with reward learning.

4.5 Results on GYAFC Corpus

In recent work, it is shown that style transfer models trained on parallel data can benefit from additional reward-based learning (Lai et al., 2021). Here we conduct additional experiments to assess our semi-supervised framework on the GYAFC formality transfer corpus with well-annotated data. We evaluate the proposed model on the *informal*-to-*formal* task as previous work (Chawla and Yang, 2020), and compare them with strong baselines. As shown in Table 5, while the baselines show impressive BLEU scores on the formality transfer task, our framework outperforms them significantly in terms of style accuracy, approaching upper-bound human performance. Moreover, compared with the contemporary supervised work (Lai et al., 2021), which also introduced additional RL-based optimization, our model still achieves higher G2 and H2 scores. The examples shown in Appendix Table 13 demonstrate that our approach generates sentences with accurate formality paraphrasing.

4.6 Human Assessment

Additionally, we conducted a human evaluation on Yelp, Amazon and GYAFC datasets. Following previous work (Chawla and Yang, 2020; Liu et al., 2021), we evaluated the generated sentences from three aspects: style transfer strength (*Style*), text flu-

Model	Yelp Data Accuracy	BLEU	G2	H2	Amazon Data Accuracy	BLEU	G2	H2
Sequence-Level Reward (30K Data)	85.1	26.5	47.5	40.4	78.4	19.0	38.5	30.5
Stepwise Reward (30K Data)	81.1	50.4	**63.9**	**62.1**	71.2	36.1	**50.6**	**47.9**
Sequence-Level Reward (100K Data)	84.8	35.3	54.7	49.8	81.4	21.9	42.2	34.5
Stepwise Reward (100K Data)	86.2	59.4	**71.5**	**70.3**	73.1	46.3	**58.1**	**56.6**

Table 6: Ablation study on the proposed stepwise reward on the Yelp and Amazon dataset. *Sequence-level* denotes the reward is calculated on the whole sequence, without the stepwise re-weighting.

Layer No.	Accuracy	BLEU	G2	H2
Layer-12	78.5	46.2	60.2	58.1
Layer-11	81.1	45.5	**60.7**	**58.2**
Layer-10	84.2	38.7	57.0	53.0
Layer-9	72.3	43.8	56.2	54.5
Layer-8	76.1	44.5	58.1	56.1
Layer-7	70.3	41.6	54.0	52.2

Table 7: Layer selection for the proposed stepwise reward re-weighting. The Yelp sentiment transfer dataset and the semantic-based matching are used. We conduct experiments on the last 6 Transformer layers of the style classifier.

Train Size	Accuracy	BLEU	G2	H2
1,000	62.9	31.6	44.5	42.0
5,000	68.2	36.8	50.0	47.8
10,000	73.3	43.6	56.5	54.6
15,000	76.1	45.5	58.8	56.9
30,000	83.5	49.6	64.3	62.2

Table 8: Results from different pseudo sample sizes using the proposed framework. The Yelp sentiment transfer dataset and semantic-based matching are used.

ency (*Fluency*), and content preservation (*Content*), separately. The three aspects are rated with range [1, 5], then their average value is calculated and reported as *Mean* (see Table 14 in Appendix). For each corpus, we randomly selected 80 test samples and compared the outputs of representative and previous state-of-the-art models. Each candidate was rated by three linguistic experts, and we report the average scores. Our model achieves better overall performance when considering all three evaluation metrics on each dataset. Moreover, we observe that leveraging the pre-trained language models such as BART and GPT-2 is beneficial for the text fluency.

5 Analysis

To extensively assess the effectiveness of the proposed methods, we conduct the following in-depth analyses.

5.1 Ablation Study on Stepwise Re-weighting

We conduct an ablation experiment to assess the effectiveness of stepwise reward re-weighting. As shown in Table 6, the performance degrades significantly without the stepwise reward re-weighting, especially the BLEU score. In particular, we observed that when removing stepwise optimization, the generator was prone to mode collapse. In one manifestation of mode collapse, the model appended a limited set of phrases to the source sentences, resulting in generation with disfluency and low diversity. It demonstrates that token-level reward optimization provides finer-granularity for policy gradient of sequence-to-sequence training. This approach can also be potentially extended to other text generation tasks.

5.2 Attention Layer Selection for Stepwise Reward Re-weighting

We utilize attentive scores from the top-2 multi-head layers for stepwise reward re-weighting. To study the effect of layer selection, we compared the results using attention scores extracted from different Transformer layers in the style classifier described in Section 3.4. As shown in Table 7, the performance shows an overall increasing trend from the 7-th to the 12-th layer, and we obtained better results with the top layers. In scores of lower layers, we found that the model tended to assure content preservation rather than style accuracy. This is consistent with the observations from recent linguistic probing and model interpretation studies (Hewitt and Manning, 2019; Xu et al., 2020): the information modeled in the Transformer-based networks, especially the pre-trained language backbones, is represented in a hierarchical manner, and the higher layers provide more effective information on scoring the span importance for text classification (see visualization in Appendix Figure 4).

5.3 Bootstrapping Sample Size

We investigate the effect of different pseudo parallel sample sizes. As shown in Table 8, the result shows that the evaluation result by automatic metrics becomes acceptable when training reaches 10K samples. Results comparable to state-of-the-art are achieved with merely 30K data (10% of the Yelp training set). We speculate that the relatively weak performance with 10K samples is because the BART model uses a denoising autoencoding paradigm (Lewis et al., 2020), which is trained to reconstruct the input sentence, and style strength of sentence rewriting is strongly affected in this low resource scenario.

Additionally, we conduct an ablation study on the bootstrapping step, and the result shows that with the same training sample size, the generation performance (considering both style accuracy and content preservation) obtained significant improvement by adding the bootstrapping learning stage (see Appendix Table 15).

6 Conclusions

In this paper, we proposed a framework for text style transfer taking advantage of both supervised and unsupervised paradigms. The training process is bootstrapped with supervision guided by automatically constructed pseudo parallel data. Both lexical-based and semantic-based sentence matching proved effective. Moreover, the stepwise reward re-weighting significantly improved the generation performance, and is a generic design that can be easily extended. Experimental results showed that the proposed approach achieved state-of-the-art performance in multiple datasets, while producing reasonable generation even with minimal training data (10% of original size).

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore. We thank Ai Ti Aw for the insightful discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354.

Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, and Jack Rae. 2019. Training language gans from scratch. *Advances in Neural Information Processing Systems*, 32.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 663–670.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 3168–3180.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *Proceedings of Ninth International Conference on Learning Representations*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4129–4138.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. Imat: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3097–3109.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. 2019. Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3579–3584. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral Ruiz, and Malvina Nissim. 2021. Thank you bart! rewarding pre-trained models improves formality style transfer. In *Proceedings of the ACL-IJCNLP 2021*. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4262–4273.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122.

David D McDonald and James Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *Second Conference of the European Chapter of the Association for Computational Linguistics*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Nikola I Nikolov and Richard Hahnloser. 2019. Large-scale hierarchical alignment for data-driven text rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 844–853.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the NAACL 2018*, pages 129–140.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6833–6844.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of EMNLP 2019*, pages 3269–3279.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. 1999. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the EMNLP-IJCNLP 2019*, pages 3573–3578.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249, Barcelona, Spain (Online).

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883.

Hu Xu, Lei Shu, S Yu Philip, and Bing Liu. 2020. Understanding pre-trained bert for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 244–250.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 979–988.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 2020*.

Yi Zhang, Tao Ge, and Xu Sun. 2020b. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation.

Chulun Zhou, Liang-Yu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the ACL 2020*, pages 7135–7144.

A Appendix

Algorithm 1 Training process of the proposed semi-supervised text style transfer framework.

```
Given non-labeled datasets D_S and D_T in two different styles S and T, construct pseudo parallel dataset D_pseudo with sentence pairs matched with lexical-based or semantic-based similarity
Pre-train a binary style classifier styleCLS on the two datasets D_S and D_T
Pre-train the text style transfer model G_θ using pseudo-parallel sentence pairs in dataset D_pseudo, with MLE loss (Eq. 1).
for each iter i = 1, 2, ..., M do
    Sample sentence x of source style S from D_S
    Generate sentence y' of target style T via model G_θ by greedy decoding
    Generate sentence ŷ of target style T via model G_θ by sampling-based decoding
                                   ▷ Reconstruction Reward Calculation (Content Preservation)
    Given y', generate back-translated sentence x' of source style S via model G_θ by greedy decoding
    Given ŷ, generate back-translated sentence x̂ of source style S via model G_θ by greedy decoding
    Compute reconstruction reward R_cyclic based on BLEU scores of the pair [x, x'] and the pair [x, x̂], following Self-Critical Sequence Training (Eq. 2)
                                   ▷ Style Reward Calculation (Style Strength)
    Compute style reward R_style of generated sentence ŷ using the style classifier styleCLS
                                   ▷ Stepwise Reward Re-weighting
    Compute the stepwise re-weighting values by max-pooling attentive scores from style classifier styleCLS on the generated sentence ŷ
    Expand R_style and R_cyclic from 1-D (sequence level) to 2-D (token level), and re-weight R_style with stepwise values
    Compute the total stepwise reward R' by adding R_style and R_cyclic, based on Eq. 4
    Update θ using reward R' based on Eq. 6
end for
```

Environment Details	
Sequence Generator	BART-Base (12-layer, 768-hidden, 16-heads, 139M parameters).
Style Classifier	RoBERTa-base (12-layer, 768-hidden, 12-heads, 125M parameters).
GPU Model	Single Tesla A100 with 40 GB memory; CUDA version 11.0.
Library Version	Pytorch==1.8.1; Transformers==4.8.2.
Computational Cost	Average 5 hours training time for one round. Average 3 rounds for each reported result (calculating mean of the result scores).
Hyper-parameter	**Setting Detail**
Learning Rate and Batch Size	We set the learning rate and batch size according to regular language model fine-tuning strategy (Lewis et al., 2020).
Beam Search Size	We evaluated models on beam search sizes from 3 to 10, and 6 provided the best balance of performance and inference speed.
Style Reward Penalty γ (Eq. 4)	(1) In our experiment, we observed that the style reward R_{style} values given by the style classifier were up to 0.9 (indicating a high level of style transfer strength), while the cyclic reconstruction reward R_{cyclic} values were at a lower level (average was 0.5). Therefore, we added the γ to adjust the R_{style} to the same level of R_{cyclic}. (2) We evaluated values from 0.1 to 0.4 (0.1 as step), and empirically set the γ at 0.2. Training without the penalty γ did not produce significantly degraded results.
λ_{cyclic} and λ_{style} (Eq. 4)	We evaluated both values with 1.0 +/- 0.2, and empirically set λ_{cyclic} at 1.0, λ_{style} at 0.8. Setting at 1.0 by default did not produce degraded results.
Sequence-Level & Stepwise Reward	For the comparison of using sequence level and stepwise rewards, we run experiments with the aforementioned parameter setting.
Combination of lexical and semantic pseudo-parallel data	In our pilot experiment, we tried to combine both lexical and semantic pseudo-parallel data, but this did not bring any improvement on the Yelp and Amazon. Presumably this is because the semi-supervised model only requires weak supervision from the pseudo-parallel data, and either the lexical and semantic data can provide sufficient information at the bootstrapping training stage.

Table 9: The detailed environment settings and search strategy of training parameters in our experiment. It is worth mentioned that our proposed semi-supervised approach with bootstrapping strategy and stepwise reward re-weighting is targeted to tackle the unstable learning issue of RL-based models.

Model	E&M Domain Accuracy*	BLEU	G2	H2	F&R Domain Accuracy*	BLEU	G2	H2
Human Reference (Rao and Tetreault, 2018)	58.7	100.0	76.6	73.9	51.4	100.0	71.6	67.8
Rule-Based (Rao and Tetreault, 2018)	11.4	72.4	28.7	19.6	52.1	65.8	58.5	58.1
Hybrid Annotations (Xu et al., 2019)	10.4	69.2	26.8	18.0	8.75	74.3	25.4	15.6
Semi-LM-MMI **w/ BART** (Chawla and Yang, 2020)	10.6	76.5	28.4	18.6	9.68	79.9	27.8	17.2
Rewarded **BART-Large** (Lai et al., 2021)	52.8	76.5	63.5	62.4	45.9	79.2	60.2	58.1
Only Labeled Data Supervision (Full)	55.2	71.2	62.6	62.1	47.3	72.5	58.5	57.2
Labeled Data + Reward-Learning (30K)	55.3	71.4	62.8	62.3	45.2	74.4	57.9	56.2
Labeled Data + Reward-Learning (Full)	58.1	71.0	**64.2**	**63.9**	50.3	74.2	**61.0**	**59.9**

Table 10: Automatic evaluation scores on the GYAFC formality style transfer task of baseline models and our framework. Baseline results are reported with the model generations provided in published studies (Chawla and Yang, 2020). * The style accuracy is calculated with a *TextCNN* classifier.

Model	**Text**
Source Sentence	ever since joes has changed hands it 's just gotten worse and worse .
Human Reference	ever since joes has changed hands it 's gotten better and better .
Cross Aligned (Shen et al., 2017)	i recommend that has out to it 's always great and fun .
Delete+Retrieve (Li et al., 2018)	ever since joes has changed hands it 's just so good !
DualRL (Luo et al., 2019)	ever since dedicated has changed hands it 's just gotten better and better .
IterativeMatch (Jin et al., 2019)	dominos has gotten better and better .
Deep Latent w/ LMs (He et al., 2019)	just since their sausages has changed it 's just gotten worse and worse .
Direct Rewards w/ GPT-2 (Liu et al., 2021)	ever since joes has changed hands it 's just gotten better and better .
Bootstrapping + Reward-Learning (Ours)	ever since joes has changed hands it 's just gotten better and better .
Source Sentence	no , i 'm not at a scottsdale club .
Human Reference	this was a great club.
Cross Aligned (Shen et al., 2017)	great , i 'm so at a local business .
Delete+Retrieve (Li et al., 2018)	this is a great place to get a scottsdale club .
DualRL (Luo et al., 2019)	great job .
IterativeMatch (Jin et al., 2019)	i 'm so glad i found this place .
Deep Latent w/ LMs (He et al., 2019)	great food , great service at a scottsdale club .
Direct Rewards w/ GPT-2 (Liu et al., 2021)	great , nice and a scottsdale club .
Bootstrapping + Reward-Learning (Ours)	great , i 'm at a scottsdale club .
Source Sentence	french toast plate was good , mom said , but eggs were cold .
Human Reference	french toast plate was good , mom said , eggs were hot .
Cross Aligned (Shen et al., 2017)	their food tasted was good , juicy , and fries are very clean .
Delete+Retrieve (Li et al., 2018)	french toast plate was good , mom said , but eggs were amazing !
DualRL (Luo et al., 2019)	french toast plate was good , mom said , but eggs were delicious .
IterativeMatch (Jin et al., 2019)	the food was delicious and the eggs were fresh .
Deep Latent w/ LMs (He et al., 2019)	wow !
Direct Rewards w/ GPT-2 (Liu et al., 2021)	french toast plate was good , mom said , with amazing eggs are warm .
Bootstrapping + Reward-Learning (Ours)	french toast plate was good , mom said , but eggs were amazing .
Source Sentence	however , it turned out to be nothing like i thought it would .
Human Reference	this turned out exactly how i thought it would .
Cross Aligned (Shen et al., 2017)	however , it right out to be great , it is the place .
Delete+Retrieve (Li et al., 2018)	it turned out to be nothing like i thought it was so good !
DualRL (Luo et al., 2019)	however , it turned out to be nothing extraordinary it would thought it would
IterativeMatch (Jin et al., 2019)	it turned out i worried about nothing .
Deep Latent w/ LMs (He et al., 2019)	loved it !
Direct Rewards w/ GPT-2 (Liu et al., 2021)	although , it turned out to be great with i thought it will .
Bootstrapping + Reward-Learning (Ours)	however , it turned out to be great like i thought it would .

Table 11: Examples of human references and generated sentences on the Yelp corpus from representative baseline models and our proposed framework. The text style is converted from negative to positive.

Model	Text
Source Sentence	it makes a buzzing sound when devices are plugged in.
Human Reference	it makes a useful buzzing sound when devices are plugged in.
Cross Aligned (Shen et al., 2017)	it s a nice , and easy to clean out .
Style Embedding (Fu et al., 2018)	it makes a bit different , while but num_extend mode .
Template-Based (Li et al., 2018)	it makes a buzzing sound when devices are plugged in and use it to charge my .
Delete+Retrieve (Li et al., 2018)	it makes a buzzing sound when the devices are plugged in .
Direct Rewards w/ GPT-2 (Liu et al., 2021)	it makes a cooking faster than devices are plugged in .
Bootstrapping + Reward-Learning (Ours)	it makes a great sound when devices are plugged in .
Source Sentence	it was not as good as our much cheaper model .
Human Reference	its a great as before .
Cross Aligned (Shen et al., 2017)	it s not not worth the phone and very well .
Style Embedding (Fu et al., 2018)	it was worth it size but at least my product , .
Template-Based (Li et al., 2018)	it was not as good as our much cheaper model and works just .
Delete+Retrieve (Li et al., 2018)	as using the much cheaper model as it is also much cheaper .
Direct Rewards w/ GPT-2 (Liu et al., 2021)	it was excellent as our much cheaper model .
Bootstrapping + Reward-Learning (Ours)	it was as good as our much cheaper model .
Source Sentence	i received the wrong color and it shreds easily .
Human Reference	i received the right color and it works well.
Cross Aligned (Shen et al., 2017)	i bought the phone and it s easy to .
Style Embedding (Fu et al., 2018)	i received the fact that and quickly is no clean .
Template-Based (Li et al., 2018)	i received the wrong color and it shreds easily to order more .
Delete+Retrieve (Li et al., 2018)	i received the wrong color and it looks very nice ! he would highly recommend it easily .
Direct Rewards w/ GPT-2 (Liu et al., 2021)	i received the best cooking efficiently .
Bootstrapping + Reward-Learning (Ours)	i received the right color and it shreds easily .
Source Sentence	i am actually afraid to open the remaining jars .
Human Reference	I look forward to opening the remaining jars.
Cross Aligned (Shen et al., 2017)	i have to say and the other ones .
Style Embedding (Fu et al., 2018)	i am actually used the right over a container .
Template-Based (Li et al., 2018)	i am actually afraid to open the remaining jars highly recommend .
Delete+Retrieve (Li et al., 2018)	i am actually afraid to open the remaining jars this is great .
Direct Rewards w/ GPT-2 (Liu et al., 2021)	i am actually faster cooking than items .
Bootstrapping + Reward-Learning (Ours)	i am actually happy to open the remaining jars .

Table 12: Examples of human references and generated sentences on the Amazon corpus from representative baseline models and our proposed framework. The text style is converted from negative to positive.

Model	Text
Source Sentence	my dad likes action,my mom likes romance,but for me i like comedy.
Human Reference	My father likes action, my mother likes romance, but for me I prefer comedy.
Rule-Based (Rao and Tetreault, 2018)	My dad likes action , my mom likes romance , but for me I like comedy .
Hybrid Annotations (Xu et al., 2019)	My father likes action , my mother likes romance , but I like comedy .
Semi-LM-MMI w/ BART-large (Chawla and Yang, 2020)	My dad likes action , my mom likes romance , but for me I like comedy .
Rewarded BART-Large (Lai et al., 2021)	My dad likes action , my mom likes romance , but for me I like comedy .
Labeled Data + Reward-Learning (Ours)	My father likes action, my mother likes romance, but for me I prefer comedy.
Source Sentence	I want to be on TV!
Human Reference	I would like to be on television.
Rule-Based (Rao and Tetreault, 2018)	I want to be on television !
Hybrid Annotations (Xu et al., 2019)	I want to be on television .
Semi-LM-MMI w/ BART-large (Chawla and Yang, 2020)	I want to be on TV .
Rewarded BART-Large (Lai et al., 2021)	I would like to be on television.
Labeled Data + Reward-Learning (Ours)	I would like to be on television.
Source Sentence	BUT IT IS OKAY TO KISS ON THE FIRST DATE.
Human Reference	It is okay to kiss on the first date.
Rule-Based (Rao and Tetreault, 2018)	However, it is okay to kiss on the first date.
Hybrid Annotations (Xu et al., 2019)	It is okay to kiss on the first date .
Semi-LM-MMI w/ BART-large (Chawla and Yang, 2020)	It is okay to kiss on the first date .
Rewarded BART-Large (Lai et al., 2021)	However, it is acceptable to kiss on the first date.
Labeled Data + Reward-Learning (Ours)	However, it is acceptable to kiss on the first date.
Source Sentence	The same guy you wanna be in a relationship with?
Human Reference	Do you want to be in a relationship with the same man?
Rule-Based (Rao and Tetreault, 2018)	The same man with whom you would like to be in a relationship?
Hybrid Annotations (Xu et al., 2019)	The same guy you want to be in a relationship with ?
Semi-LM-MMI w/ BART-large (Chawla and Yang, 2020)	The same guy you want to be in a relationship with ?
Rewarded BART-Large (Lai et al., 2021)	The same man you want to be in a relationship with ?
Labeled Data + Reward-Learning (Ours)	Is this the same man you want to be in a relationship with?

Table 13: Examples of human references and generated sentences on the GYAFC corpus from representative baseline models and our proposed framework. The text style is converted from informal to formal.

I. Scoring result on the Yelp corpus

Model	Style	Fluency	Content	Mean
Delete+Retrieve (Li et al., 2018)	3.25	2.72	2.86	2.94
IterativeMatch (Jin et al., 2019)	3.40	2.88	2.69	2.99
Direct Rewards w/ GPT-2 (Liu et al., 2021)	3.51	3.15	3.18	3.28
Bootstrapping + Reward-Learning (Ours)	3.49	3.29	3.25	**3.34**

II. Scoring result on the Amazon corpus

Model	Style	Fluency	Content	Mean
Template-Based (Li et al., 2018)	2.78	2.36	2.55	2.56
Delete+Retrieve (Li et al., 2018)	2.94	3.08	2.73	2.91
Direct Rewards w/ GPT-2 (Liu et al., 2021)	3.20	3.23	2.21	2.88
Bootstrapping + Reward-Learning (Ours)	3.31	3.28	3.12	**3.23**

III. Scoring result on the GYAFC corpus

Model	Style	Fluency	Content	Mean
Hybrid Annotations (Xu et al., 2019)	2.56	3.15	3.13	2.95
Semi-LM-MMI w/ BART (Chawla and Yang, 2020)	3.12	3.47	3.22	3.27
Rewarded BART-Large (Lai et al., 2021)	3.36	3.60	3.33	3.43
Labeled Data + Reward-Learning (Ours)	3.37	3.67	3.37	**3.47**

Table 14: Human evaluation are conducted on the Yelp, Amazon, and GYAFC style transfer datasets. Following previous work (Chawla and Yang, 2020; Liu et al., 2021), we evaluated the generated sentences from three aspects: style transfer strength (**Style**), text fluency (**Fluency**), and content preservation (**Content**), separately. The three aspects are rated with range [1, 5], then their average value is calculated and reported as **Mean**. For each corpus, we randomly selected 80 test samples and compared the outputs of representative and previous state-of-the-art models. Each candidate was rated by three linguistic experts, and we report the average scores. Our model achieves better overall performance when considering all three evaluation metrics on each dataset. Moreover, we observe that leveraging the pre-trained language models such as BART and GPT-2 is beneficial for the text fluency.

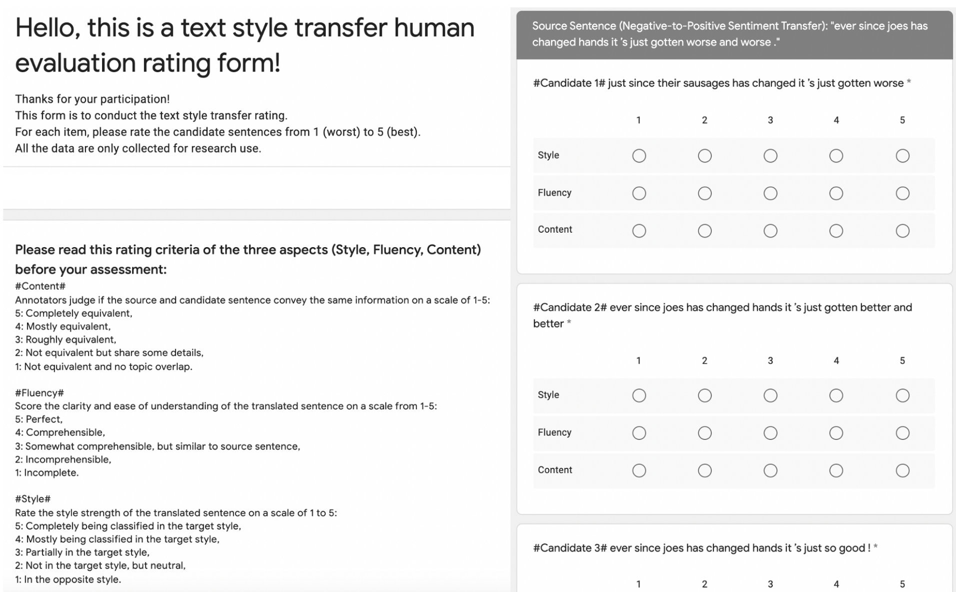

Figure 3: Rating interface for the human evaluation. Text candidates are shuffled for each sample.

Model	Yelp Data Accuracy	BLEU	G2	H2	Amazon Data Accuracy	BLEU	G2	H2
Lexical Pseudo + Reward-Learning (30K)	81.1	50.4	63.9	62.1	71.2	36.1	50.6	47.9
Pure Reward Learning (30K)	70.8	41.3	54.0	52.1	61.2	26.1	39.9	36.5
Lexical Pseudo + Reward-Learning (100K)	86.2	59.4	71.5	70.2	73.1	46.3	58.1	56.6
Pure Reward Learning (100K)	75.5	46.1	58.9	57.2	65.6	26.5	41.6	37.7

Table 15: Ablation study of the proposed bootstrapping on the Yelp and Amazon datasets. Models are running in a RL-based unsupervised manner, and we used the same data sizes as the experiments in Table 3 and Table 4.

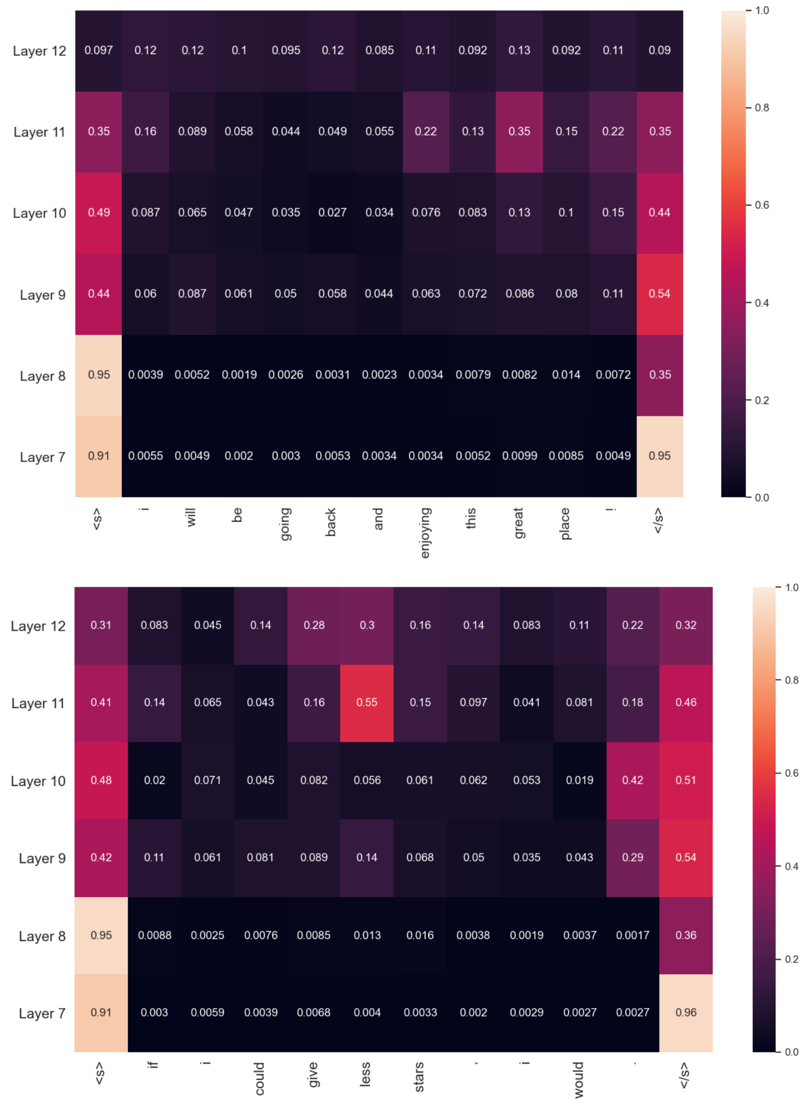

Figure 4: Attention heatmap examples of the attention scores with layer-level max-pooling. The *'RoBERTa-base'* model is fine-tuned on the Yelp data for style classification. The higher scores denotes higher attention weights on the tokens, and the top layers (especially the 11-th layer) shows better attribute-specified correlation. At the token level, the attention values and the max-pooled step-wise values described in Section 3.4 are all in the range of (0, 1).