Contextual Text Style Transfer

Yu Cheng[1], **Zhe Gan**[1], **Yizhe Zhang**[2], **Oussama Elachqar**[2], **Dianqi Li**[3], **Jingjing Liu**[1]
[1]Microsoft Dynamics 365 AI Research [2]Microsoft Research
[3]University of Washington
{yu.cheng,zhe.gan,yizhe.zhang,ouelachq,jinjl}@microsoft.com, dianqili@uw.edu

Abstract

We introduce a new task, Contextual Text Style Transfer - translating a sentence into a desired style with its surrounding context taken into account. This brings two key challenges to existing style transfer approaches: (i) how to preserve the semantic meaning of target sentence and its consistency with surrounding context during transfer; (ii) how to train a robust model with limited labeled data accompanied by context. To realize high-quality style transfer with natural context preservation, we propose a Context-Aware Style Transfer (CAST) model, which uses two separate encoders for each input sentence and its surrounding context. A classifier is further trained to ensure contextual consistency of the generated sentence. To compensate for the lack of parallel data, additional self-reconstruction and back-translation losses are introduced to leverage non-parallel data in a semi-supervised fashion. Two new benchmarks, *Enron-Context* and *Reddit-Context*, are introduced for formality and offensiveness style transfer. Experimental results on these datasets demonstrate the effectiveness of the proposed CAST model over state-of-the-art methods across style accuracy, content preservation and contextual consistency metrics.[1]

1 Introduction

Text style transfer has been applied to many applications (e.g., sentiment manipulation, formalized writing) with remarkable success. Early work relies on parallel corpora with a sequence-to-sequence learning framework (Bahdanau et al., 2015; Jhamtani et al., 2017). However, collecting parallel annotations is highly time-consuming and expensive. There has also been studies on developing text style transfer models with non-parallel data (Hu et al., 2017; Li et al., 2018; Prabhumoye et al., 2018; Subramanian et al., 2018), assuming that disentangling style information from semantic content can be achieved in an auto-encoding fashion with the introduction of additional regularizers (e.g., adversarial discriminators (Shen et al., 2017), language models (Yang et al., 2018)).

Despite promising results, these techniques still have a long way to go for practical use. Most existing models focus on sentence-level rewriting. However, in real-world applications, sentences typically reside in a surrounding paragraph context. In formalized writing, the rewritten span is expected to align well with the surrounding context to keep a coherent semantic flow. For example, to automatically replace a gender-biased sentence in a job description document, a style transfer model taking the sentence out of context may not be able to understand the proper meaning of the statement and the intended message. Taking a single sentence as the sole input of a style transfer model may fail in preserving topical coherency between the generated sentence and its surrounding context, leading to low semantic and logical consistency on the paragraph level (see Example C in Table 4). Similar observations can be found in other style transfer tasks, such as offensive to non-offensive and political to neutral translations.

Motivated by this, we propose and investigate a new task - *Contextual Text Style Transfer*. Given a paragraph, the system aims to translate sentences into a desired style, while keeping the edited section topically coherent with its surrounding context. To achieve this goal, we propose a novel Context-Aware Style Transfer (CAST) model, by jointly considering style translation and context alignment. To leverage parallel training data, CAST employs two separate encoders to encode the source sentence and its surrounding context, respectively. With the encoded sentence and context embed-

[1] Code and datasets will be released at https://github.com/ych133/CAST.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2915–2924
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

dings, a decoder is trained to translate the joint features into a new sentence in a specific style. A pre-trained style classifier is applied for style regularization, and a coherence classifier learns to regularize the generated target sentence to be consistent with the context. To overcome data sparsity issue, we further introduce a set of unsupervised training objectives (e.g., self-reconstruction loss, back-translation loss) to leverage non-parallel data in a hybrid approach (Shang et al., 2019). The final CAST model is jointly trained with both parallel and non-parallel data via end-to-end training.

As this is a newly proposed task, we introduce two new datasets, *Enron-Context* and *Reddit-Context*, collected via crowdsourcing. The former contains 14,734 formal vs. informal paired samples from Enron (Klimt and Yang, 2004) (an email dataset), and the latter contains 23,158 offensive vs. non-offensive paired samples from Reddit (Serban et al., 2017). Each sample contains an original sentence and a human-rewritten one in target style, accompanied by its paragraph context. In experiments, we also leverage 60k formal/informal sentences from GYAFC (Rao and Tetreault, 2018) and 100k offensive/non-offensive sentences from Reddit (dos Santos et al., 2018) as additional non-parallel data for model training.

The main contributions of this work are summarized as follows: (i) We propose a new task - Contextual Text Style Transfer, which aims to translate a sentence into a desired style while preserving its style-agnostic semantics and topical consistency with the surrounding context. (ii) We introduce two new datasets for this task, Enron-Context and Reddit-Context, which provide strong benchmarks for evaluating contextual style transfer models. (iii) We present a new model - Context-Aware Style Transfer (CAST), which jointly optimizes the generation quality of target sentence and its topical coherency with adjacent context. Extensive experiments on the new datasets demonstrate that the proposed CAST model significantly outperforms state-of-the-art style transfer models.

2 Related Work

2.1 Text Style Transfer

Text style transfer aims to modify an input sentence into a desired style while preserving its style-independent semantics. Previous work has explored this as a sequence-to-sequence learning task using parallel corpora with paired source/target sentences in different styles. For example, Jhamtani et al. (2017) pre-trained word embeddings by leveraging external dictionaries mapping Shakespearean words to modern English words and additional text. However, available parallel data in different styles are very limited. Therefore, there is a recent surge of interest in considering a more realistic setting, where only non-parallel stylized corpora are available. A typical approach is: (i) disentangling latent space as content and style features; then (ii) generating stylistic sentences by tweaking style-relevant features and passing them through a decoder, together with the original content-relevant features (Xu et al., 2018).

Many of these approaches borrowed the idea of adversarial discriminator/classifier from the Generative Adversarial Network (GAN) framework (Goodfellow et al., 2014). For example, Shen et al. (2017); Fu et al. (2018); Lample et al. (2018) used adversarial classifiers to force the decoder to transfer the encoded source sentence into a different style/language. Alternatively, Li et al. (2018) achieved disentanglement by filtering stylistic words of input sentences. Another direction for text style transfer without parallel data is using back-translation (Prabhumoye et al., 2018) with a de-noising auto-encoding objective (Logeswaran et al., 2018; Subramanian et al., 2018).

Regarding the tasks, sentiment transfer is one of the most widely studied problems. Transferring from informality to formality (Rao and Tetreault, 2018; Li et al., 2019) is another direction of text style transfer, aiming to change the style of a given sentence to more formal text. dos Santos et al. (2018) presented an approach to transferring offensive text to non-offensive based on social network data. In Prabhumoye et al. (2018), the authors proposed the political slant transfer task. However, all these previous studies did not directly consider context-aware text style transfer, which is the main focus of this work.

2.2 Context-aware Text Generation

Our work is related to context-aware text generation (Mikolov and Zweig, 2012; Tang et al., 2016), which can be applied to many NLP tasks (Mangrulkar et al., 2018). For example, previous work has investigated language modeling with context information (Wang and Cho, 2015; Wang et al., 2017; Li et al., 2020), treating the preceding sentences as context. There are also studies on response gen-

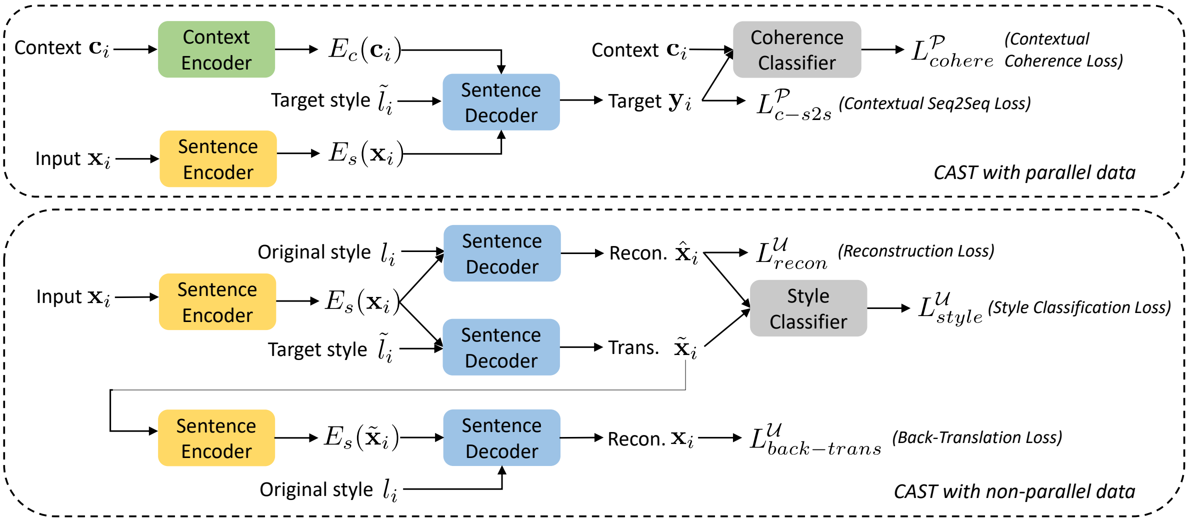
Context $\mathbf{c}_i$
Context Encoder
$E_c(\mathbf{c}_i)$
Target style $\tilde{l}_i$
Sentence Decoder
Input $\mathbf{x}_i$
Sentence Encoder
$E_s(\mathbf{x}_i)$
Context $\mathbf{c}_i$
Target $\mathbf{y}_i$
Coherence Classifier
$L^{\mathcal{P}}_{cohere}$ (Contextual Coherence Loss)
$L^{\mathcal{P}}_{c-s2s}$ (Contextual Seq2Seq Loss)
CAST with parallel data
Original style l_i
Input $\mathbf{x}_i$
Sentence Encoder
$E_s(\mathbf{x}_i)$
Target style $\tilde{l}_i$
Sentence Decoder
Recon. $\hat{\mathbf{x}}_i$
Trans. $\tilde{\mathbf{x}}_i$
$L^{\mathcal{U}}_{recon}$ (Reconstruction Loss)
Style Classifier
$L^{\mathcal{U}}_{style}$ (Style Classification Loss)
$E_s(\tilde{\mathbf{x}}_i)$
Recon. $\mathbf{x}_i$
$L^{\mathcal{U}}_{back-trans}$ (Back-Translation Loss)
Original style l_i
CAST with non-parallel data

Figure 1: Model architecture of the proposed CAST model for contextual text style transfer. Both training paths share the same sentence encoder and decoder. See Sec. 3 for details.

eration for conversational systems (Sordoni et al., 2015b; Wen et al., 2015), where dialogue history is treated as a context. Zang and Wan (2017) introduced a neural model to generate long reviews from aspect-sentiment scores given the topics. Vinyals and Le (2015) proposed a model to predict the next sentence given the previous sentences in a dialogue session. Sordoni et al. (2015a) presented a hierarchical recurrent encoder-decoder model to encode dialogue context. Our work is the first to explore context information in the text style transfer task.

3 Context-Aware Style Transfer

In this section, we first describe the problem definition and provide an overview of the model architecture in Section 3.1. Section 3.2 presents the proposed Context-Aware Style Transfer (CAST) model with supervised training objectives, and Section 3.3 further introduces how to augment the CAST model with non-parallel data in a hybrid approach.

3.1 Overview

Problem Definition The problem of contextual text style transfer is defined as follows. A style-labelled parallel dataset $\mathcal{P} = \{(\mathbf{x}_i, l_i), (\mathbf{y}_i, \tilde{l}_i), \mathbf{c}_i\}_{i=1}^{M}$ includes: (i) the i-th instance containing the original sentence $\mathbf{x}_i$ with a style l_i, (ii) its corresponding rewritten sentence $\mathbf{y}_i$ in another style $\tilde{l}_i$, and (iii) the paragraph context $\mathbf{c}_i$. $\mathbf{x}_i$ and $\mathbf{y}_i$ are expected to encode the same semantic content, but in different language styles (i.e., $l_i \neq \tilde{l}_i$). The goal is to transform $\mathbf{x}_i$ in style l_i to $\mathbf{y}_i$ in style $\tilde{l}_i$, while keeping $\mathbf{y}_i$ semantically coherent with its context $\mathbf{c}_i$. In practice, labelled parallel data may be difficult to garner. Ideally, additional non-parallel data $\mathcal{U} = \{(\mathbf{x}_i, l_i)\}_{i=1}^{N}$ can be leveraged to enhance model training.

Model Architecture The architecture of the proposed CAST model is illustrated in Figure 1. The hybrid model training process consists of two paths, one for parallel data and the other for non-parallel data. In the parallel path, a *Seq2Seq loss* and a *contextual coherence loss* are included, for the joint training of two encoders (Sentence Encoder and Context Encoder) and the Sentence Decoder. The non-parallel path is designed to further enhance the Sentence Encoder and Decoder with three additional losses: (i) a *self-reconstruction loss*; (ii) a *back-translation loss*; and (iii) a *style classification loss*. The final training objective, uniting both parallel and non-parallel paths, is formulated as:

$$\begin{aligned} L^{\mathcal{P},\mathcal{U}}_{final} = L^{\mathcal{P}}_{c-s2s} + \lambda_1 L^{\mathcal{P}}_{cohere} + \lambda_2 L^{\mathcal{U}}_{recon} \\ + \lambda_3 L^{\mathcal{U}}_{btrans} + \lambda_4 L^{\mathcal{U}}_{style} \,, \end{aligned} \quad (1)$$

where λ_1, λ_2, λ_3 and λ_4 are hyper-parameters to balance different objectives. Each of these loss terms will be explained in the following subsections.

3.2 Supervised Training Objectives

In this subsection, we discuss the training objective associated with parallel data, consisting of: (i) a contextual Seq2Seq loss; and (ii) a contextual coherence loss.

Contextual Seq2Seq Loss When parallel data is available, a Seq2Seq model can be directly learned for text style transfer. We denote the Seq2Seq model as (E, D), where the semantic representation of sentence $\mathbf{x}_i$ is extracted by the encoder E, and the decoder D aims to learn a conditional distribution of $\mathbf{y}_i$ given the encoded feature $E(\mathbf{x}_i)$ and style $\tilde{l}_i$:

$$L_{s2s}^{\mathcal{P}} = - \mathop{\mathbb{E}}_{\mathbf{x}_i, \mathbf{y}_i \sim \mathcal{P}} \log p_D(\mathbf{y}_i | E(\mathbf{x}_i), \tilde{l}_i) \,. \qquad (2)$$

However, in such a sentence-to-sentence style transfer setting, the context in the paragraph is ignored, which if well utilized, could help improve generation quality such as paragraph-level topical coherence.

Thus, to take advantage of the paragraph context $\mathbf{c}_i$, we use two separate encoders E_s and E_c to encode the sentence and the context independently. The outputs of the two encoders are combined via a linear layer, to obtain a context-aware sentence representation, which is then fed to the decoder to generate the target sentence. The model is trained to minimize the following loss:

$$L_{c-s2s}^{\mathcal{P}} = - \mathop{\mathbb{E}}_{\mathbf{x}_i, \mathbf{c}_i, \mathbf{y}_i \sim \mathcal{P}} \log p_D(\mathbf{y}_i | E_s(\mathbf{x}_i), E_c(\mathbf{c}_i), \tilde{l}_i) \,. \qquad (3)$$

Compared with Eqn. (2), the use of $E_c(\mathbf{c}_i)$ makes the text style transfer process context-dependent. The generated sentence can be denoted as $\tilde{\mathbf{y}}_i = D(E_s(\mathbf{x}_i), E_c(\mathbf{c}_i), \tilde{l}_i)$.

Contextual Coherence Loss To enforce contextual coherence (i.e., to ensure the generated sentence $\mathbf{y}_i$ aligns with the surrounding context $\mathbf{c}_i$), we train a coherence classifier that judges whether $\mathbf{c}_i$ is the context of $\mathbf{y}_i$, by adopting a language model with an objective similar to next sentence prediction (Devlin et al., 2019).

Specifically, assume that $\mathbf{y}_i$ is the t-th sentence of a paragraph $\mathbf{p}_i$ (i.e., $\mathbf{y}_i = \mathbf{p}_i^{(t)}$), and $\mathbf{c}_i = \{\mathbf{p}_i^{(0)}, \ldots, \mathbf{p}_i^{(t-1)}, \mathbf{p}_i^{(t+1)}, \ldots, \mathbf{p}_i^{(T)}\}$ is its surrounding context. We first reconstruct the paragraph $\mathbf{p}_i = \{\mathbf{p}_i^{(0)}, \ldots, \mathbf{p}_i^{(T)}\}$ by inserting $\mathbf{y}_i$ into the proper position in $\mathbf{c}_i$, denoted as $[\mathbf{c}_i; \mathbf{y}_i]$. Based on this, we obtain a paragraph representation $\mathbf{u}_i$ via a language model encoder. Then, we apply a linear layer to the representation, followed by a $\tanh$ function and a softmax layer to predict a binary label s_i, which indicates whether $\mathbf{c}_i$ is the right context for $\mathbf{y}_i$:

$$\mathbf{u}_i = \text{LM}([\mathbf{c}_i; f(\mathbf{y}_i)]) \qquad (4)$$
$$p_{\text{LM}}(s_i | \mathbf{c}_i, \mathbf{y}_i) = \text{softmax}\left(\tanh\left(\mathbf{W}\mathbf{u}_i + \mathbf{b}\right)\right) \,,$$

where LM represents the language model encoder, and $s_i = 1$ indicates that $\mathbf{c}_i$ is the context of $\mathbf{y}_i$. $f(.)$ is a softmax function with temperature τ, where the logits are the predicted network output with a dimension of vocabulary size. Note that since $\tilde{\mathbf{y}}_i$ are discrete tokens that are non-differentiable, we use the continuous feature $f(\tilde{\mathbf{y}}_i)$ to generates $\tilde{\mathbf{y}}_i$ as the input of the language model. We construct paired data $\{\mathbf{y}_i, \mathbf{c}_i, s_i\}_{i=1}^N$ for training the classifier, where the negative samples are created by replacing a sentence in a paragraph with another random sentence. After pre-training, the coherence classifier is used to obtain the contextual coherence loss:

$$L_{cohere}^{\mathcal{P}} = - \mathop{\mathbb{E}}_{\mathbf{x}_i, \mathbf{c}_i \sim \mathcal{P}} \log p_{\text{LM}}(s_i = 1 | \mathbf{c}_i, f(\tilde{\mathbf{y}}_i)) \,. \qquad (5)$$

Intuitively, minimizing $L_{cohere}^{\mathcal{P}}$ encourages $\tilde{\mathbf{y}}_i$ to blend better to its context $\mathbf{c}_i$. Note that the coherence classifier is pre-trained, and remains fixed during the training of the CAST model. The above coherence loss can be used to update the parameters of E_s, E_c and D during model training.

3.3 Unsupervised Training Objectives

For the contextual style transfer task, there are not many parallel datasets available with style-labeled paragraph pairs. To overcome the data sparsity issue, we propose a hybrid approach to leverage additional non-parallel data $\mathcal{U} = \{(\mathbf{x}_i, l_i)\}_{i=1}^N$, which are abundant and less expensive to collect. In order to fully exploit $\mathcal{U}$ to enhance the training of the Sentence Encoder and Decoder (E_s, D), we introduce three additional training losses, detailed below.

Reconstruction Loss The reconstruction loss aims to encourage E_s and D to reconstruct the input sentence itself, if the desired style is the same as the input style. The corresponding objective is similar to Eqn. (2):

$$L_{recon}^{\mathcal{U}} = - \mathop{\mathbb{E}}_{\mathbf{x}_i \sim \mathcal{U}} \log p_D(\mathbf{x}_i | E_s(\mathbf{x}_i), l_i) \,. \qquad (6)$$

Compared to Eqn. (2), here we encourage the decoder D to recover $\mathbf{x}_i$'s original style properties as accurate as possible, given the style label l_i. The self-reconstructed sentence is denoted as $\hat{\mathbf{x}}_i = D(E_s(\mathbf{x}_i), l_i)$.

Back-Translation Loss The back-translation loss requires the model to reconstruct the input sentence after a transformation loop. Specifically, the input sentence $\mathbf{x}_i$ is first transferred into the target style, i.e., $\tilde{\mathbf{x}}_i = D(E_s(\mathbf{x}_i), \tilde{l}_i)$. Then the generated target sentence is transferred back into its original style, i.e., $\hat{\mathbf{x}}_i = D(E_s(\tilde{\mathbf{x}}_i), l_i)$. The back-translation loss is defined as:

$$L^{\mathcal{U}}_{btrans} = - \mathop{\mathbb{E}}_{\substack{\mathbf{x}_i \sim \mathcal{U}, \tilde{\mathbf{x}}_i \sim \\ p_D(\mathbf{y}_i|E_s(\mathbf{x}_i), \tilde{l}_i)}} \log p_D(\mathbf{x}_i | E_s(\tilde{\mathbf{x}}_i), l_i) \,, \tag{7}$$

where the source and target styles are denoted as l_i and $\tilde{l}_i$, respectively.

Style Classification Loss To further boost the model, we use $\mathcal{U}$ to train a classifier that predicts the style of a given sentence, and regularize the training of (E_s, D) with the pre-trained style classifier. The objective is defined as:

$$L_{style} = - \mathop{\mathbb{E}}_{\mathbf{x}_i \sim \mathcal{U}} \log p_C(l_i | \mathbf{x}_i) \,, \tag{8}$$

where $p_C(\cdot)$ denotes the style classifier. After the classifier is trained, we keep its parameters fixed, and apply it to update the parameters of (E_s, D). The resulting style classification loss utilizing the pre-trained style classifier is defined as:

$$L^{\mathcal{U}}_{style} = - \mathop{\mathbb{E}}_{\mathbf{x}_i \sim \mathcal{U}} \Big[\mathop{\mathbb{E}}_{\hat{\mathbf{x}}_i \sim p_D(\hat{\mathbf{x}}_i | E_s(\mathbf{x}_i), l_i)} \log p_C(l_i | \hat{\mathbf{x}}_i) + \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}_i \sim p_D(\tilde{\mathbf{x}}_i | E_s(\mathbf{x}_i), \tilde{l}_i)} \log p_C(\tilde{l}_i | \tilde{\mathbf{x}}_i) \Big] \,. \tag{9}$$

4 New Benchmarks

Existing text style transfer datasets, either parallel or non-parallel, do not contain contextual information, thus unsuitable for the contextual transfer task. To provide benchmarks for evaluation, we introduce two new datasets: Enron-Context and Reddit-Context, derived from two existing datasets - Enron (Klimt and Yang, 2004) and Reddit Politics (Serban et al., 2017).

1) **Enron-Context** To build a formality transfer dataset with paragraph contexts, we randomly sampled emails from the Enron corpus (Klimt and Yang, 2004). After pre-processing and filtering with NLTK (Bird et al., 2009), we asked Amazon Mechanical Turk (AMT) annotators to identify informal sentences within each email, and rewrite them in a more formal style. Then, we asked a different group of annotators to verify if each rewritten sentence is more formal than the original sentence.

2) **Reddit-Context** Another typical style transfer task is offensive vs. non-offensive, for which we collected another dataset from the Reddit Politics corpus (Serban et al., 2017). First, we identify offensive sentences in the original dataset with sentence-level classification. After filtering out extremely long/short sentences, we randomly selected a subset of sentences (10% of the whole dataset) and asked AMT annotators to rewrite each offensive sentence into two non-offensive alternatives.

After manually removing wrong or duplicate annotations, we obtained a total of 14,734 rewritten sentences for Enron-Context, and 23,158 for Reddit-Context. We also limited the vocabulary size by replacing words with a frequency less than 20/70 in Enron/Reddit datasets with a special unknown token. Table 1 provides the statistics on the two datasets. More details on AMT data collection are provided in Appendix.

5 Experiments

In this section, we compare our model with state-of-the-art baselines on the two new benchmarks, and provide both quantitative analysis and human evaluation to validate the effectiveness of the proposed CAST model.

5.1 Datasets and Baselines

In addition to the two new parallel datasets, we also leverage non-parallel datasets for CAST model training. For formality transfer, one choice is Grammarlys Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018), crawled and annotated from two domains in Yahoo Answers. This corpus contains paired informal-formal sentences without context. We randomly selected a subset of sentences (28,375/29,774 formal/informal) from the GYAFC dataset as our training dataset. For offensiveness transfer, we utilize the Reddit dataset. Following dos Santos et al. (2018), we used a pre-trained classifier to extract 53,028/53,714 offensive/non-offensive sentences from Reddit posts as our training dataset.

Table 2 provides the statistics of parallel and non-parallel datasets used for the two style transfer tasks. For the non-parallel datasets, we split them into two: one for CAST model training ('Train'), and the other for the style classifier pre-training.

Dataset	# sent.	# rewritten sent.	# words per sent.	# words per paragraph	# vocabulary
Reddit-Context	14,734	14,734	9.4	38.5	4,622
Enron-Context	23,158	25,259	7.6	25.9	2,196

Table 1: Statistics on Enron-Context and Reddit-Context datasets.

Formality Transfer							
Non-parallel	Train	Style classifier	Parallel	Train	Dev	Test	Coherence classifier
GYAFC	58k	12k	Enron-Context	13k	0.5k	1k	2.5k
Offensiveness Transfer							
Non-parallel	Train	Style classifier	Parallel	Train	Dev	Test	Coherence classifier
Reddit	106k	15k	Reddit-Context	22k	0.5k	1k	3.5k

Table 2: Statistics of the parallel and non-parallel datasets on two text style transfer tasks.

Similarly, for the parallel datasets, the training sets are divided into two as well, for the training of CAST ('Train/Dev/Test') and the coherence classifier, respectively.

We compare CAST model with several baselines: (i) Seq2Seq: a Transformer-based Seq2Seq model (Eqn. (2)), taking sentences as the only input, trained on parallel data only; (ii) Contextual Seq2Seq: a Transformer-based contextual Seq2Seq model (Eqn. (3)), taking both context and sentence as input, trained on parallel data only; (iii) Hybrid Seq2Seq (Xu et al., 2019): a Seq2Seq model leveraging both parallel and non-parallel data; (iv) ControlGen (Hu et al., 2017, 2018): a state-of-the-art text transfer model using non-parallel data; (v) MulAttGen (Subramanian et al., 2018): another state-of-the-art style transfer model that allows flexible control over multiple attributes.

5.2 Evaluation Metrics

The contextual style transfer task requires a model to generate sentences that: (i) preserve the original semantic content and structure in the source sentence; (ii) conform to the pre-specified style; and (iii) align with the surrounding context in the paragraph. Thus, we consider the following automatic metrics for evaluation:

Content Preservation. We assess the degree of content preservation during transfer, by measuring *BLEU* scores (Papineni et al., 2002) between generated sentences and human references. Following Rao and Tetreault (2018), we also use *GLEU* as an additional metric for the formality transfer task, which was originally introduced for the grammatical error correction task (Napoles et al., 2015). For offensiveness transfer, we include perplexity (*PPL*) as used in dos Santos et al. (2018), which is computed by a word-level LSTM language model pre-trained on non-offensive sentences.

Style Accuracy. Similar to prior work, we measure style accuracy using the prediction accuracy of the pre-trained style classifier over generated sentences (*Acc.*).

Context Coherence. We use the prediction accuracy of the pre-trained coherence classifier to measure how a generated sentence matches its surrounding context (*Coherence*).

The evaluation classifiers are trained separately from those used to train CAST, following (dos Santos et al., 2018). For formality transfer, the style classifier and coherence classifier reach 91.35% and 86.78% accuracy, respectively, on pre-trained dataset. For offensiveness transfer, the accuracy is 93.47% and 84.96%.

5.3 Implementation Details

The context encoder, sentence encoder and sentence decoder are all implemented as a one-layer Transformer with 4 heads. The hidden dimension of one head is 256, and the hidden dimension of the feed-forward sub-layer is 1024. The context encoder is set to take maximum of 50 words from the surrounding context of the target sentence. For the style classifier, we use a standard CNN-based sentence classifier (Kim, 2014).

Since the non-parallel corpus $\mathcal{U}$ contains more samples than the parallel one $\mathcal{P}$, we down-sample $\mathcal{U}$ to assign each mini-batch the same number of parallel and non-parallel samples to balance training, alleviating the 'catastrophic forgetting prob-

	Formality Transfer				Offensiveness Transfer			
Model	*Acc.*	*Coherence*	*BLEU*	*GLEU*	*Acc.*	*Coherence*	*BLEU*	*PPL*
Seq2Seq	64.05	78.09	24.16	10.46	83.05	80.28	17.22	140.39
Contextual Seq2Seq	64.28	81.25	23.72	10.37	83.42	81.69	18.74	138.42
Hybrid Seq2Seq	65.09	79.62	24.35	10.93	83.28	84.87	20.78	107.12
ControlGen	62.18	73.66	14.32	8.72	82.15	78.81	10.44	92.14
MulAttGen	63.36	72.97	15.14	8.91	82.71	78.45	11.03	92.56
CAST	**68.04**	**85.47**	**26.38**	**15.06**	**88.45**	**85.98**	**23.92**	**93.03**

Table 3: Quantitative evaluation results of different models on two style transfer tasks.

		Task: informal to formal transfer	**Context**
A	**Input**	I'm assuming that you'd set up be part of that meeting ?	I'll call him back to a meeting. [Input]. I asked him what sort of deals they're working on .
	ControlGen	I'm guessing that you would be set up that call ?	
	MulAttGen	I'm guessing that you would be set up that meeting ?	
	C-Seq2Seq	I am assuming that you would part of that person .	
	H-Seq2Seq	I am assuming that you would be part of that party ?	
	CAST	Am I correct to assume that you would attend that meeting ?	
B	**Input**	Do y'all interface with C/P .	Thanks . Can someone let the C/P know that the deals are good ? [Input]. If not deal confirmations could but they need the deal details .
	ControlGen	Do you compete with them ?	
	MulAttGen	Do you interface with them ?	
	C-Seq2Seq	Do we interface with them ?	
	H-Seq2Seq	Do we interface with them ?	
	CAST	Do you all interface with C/P ?	
		Task: offensive to non-offensive transfer	**Context**
C	**Input**	You are ugly .	With the glasses , [Input]. I don't need them because I never read . How do i look ?
	ControlGen	You bad guy !	
	MulAttGen	You are sad .	
	C-Seq2Seq	Have a bad day .	
	H-Seq2Seq	What a bad day !	
	CAST	You look not good .	

Table 4: Examples from the two datasets, where orange denotes the sentence to be transferred, and blue denotes the content that also appears in the context (**C-Seq2Seq**: Contextual Seq2Seq; **H-Seq2Seq**: Hybrid Seq2Seq).

lem' described in Howard and Ruder (2018). We train the model using Adam optimizer with a mini-batch size 64 and a learning rate 0.0005. The validation set is used to select the best hyper-parameters. Hard-sampling (Logeswaran et al., 2018) is used to back-propagate loss through discrete tokens from the pre-trained classifier to the model.

For the ControlGen (Hu et al., 2017) baseline, we use the code provided by the authors, and use their default hyper-parameter setting. For Hybrid Seq2Seq (Xu et al., 2019) and MulAttGen (Subramanian et al., 2018), we re-implement their models following the original papers.

5.4 Experimental Results

Formality Transfer Results on the formality transfer task are summarized in Table 3. The CAST model achieves better performance than all the baselines. Particularly, CAST is able to boost *GLEU* and *Coherence* scores with a large margin. Hybrid Seq2Seq also achieves good performance by utilizing non-parallel data. By incorporating context information, Contextual Seq2Seq also improves over the vanilla Seq2Seq model. As expected, ControlGen does not perform well, since only non-parallel data is used for training.

Offensiveness Transfer Results are summarized in Table 3. CAST achieves the best performance over all the metrics except for *PPL*. In terms of *Coherence*, Contextual Seq2Seq and CAST, that leverage context information achieve better performance than Seq2Seq baseline. Contextual Seq2Seq also improves *BLEU*, which is different from the observation in the formality transfer task. On *PPL*, CAST produces slightly worse performance than ControlGen and MulAttGen. We hypothesize that this is because our model tends to use the same non-offensive word to replace an offensive word, producing some untypical sentences, as discussed in dos Santos et al. (2018).

Qualitative Analysis Table 4 presents some generation examples from different models. We observe that CAST is better at replacing informal words with formal ones (Example B and C), and generates more context-aware sentences (Example A and C), possibly due to the use of coherence and

	Formality Transfer				Offensiveness Transfer			
Model	*Acc.*	*Coherence*	*BLEU*	*GLEU*	*Acc.*	*Coherence*	*BLEU*	*PPL*
CAST	**68.04**	**85.47**	**26.38**	**15.06**	**88.45**	**85.98**	**23.92**	**93.03**
w/o context encoder	65.35	82.9	23.98	14.17	84.15	80.96	20.54	127.02
w/o cohere. classifier	65.47	80.16	14.82	14.45	85.11	79.37	21.97	115.57
w/o both	62.19	74.47	15.88	10.46	72.69	78.15	13.14	147.31
w/o non-parallel data	60.19	75.49	13.5	9.88	70.84	78.72	10.53	151.08

Table 5: Ablation study of CAST on two style transfer tasks.

Task	Aspects	CAST vs. Contextual Seq2Seq			CAST vs. Hybrid Seq2Seq			CAST vs. ControlGen		
		win	lose	tie	win	lose	tie	win	lose	tie
Formality Transfer	Style Control	**57.1**	28.3	14.6	**46.9**	26.1	28.0	**72.1**	12.6	25.3
	Content Preservation	**59.7**	22.1	18.2	**50.4**	20.8	28.2	**68.8**	14.5	17.7
	Context Consistence	**56.4**	23.1	20.5	**51.5**	19.7	28.8	**70.1**	10.6	19.3
Offensiveness Transfer	Style Control	**58.6**	25.3	16.1	**50.1**	29.2	20.3	**54.8**	19.9	25.3
	Content Preservation	**62.3**	26.5	11.2	**54.0**	17.5	28.5	**53.1**	30.2	16.7
	Context Consistence	**60.1**	32.4	17.5	**55.3**	24.9	20.8	**58.1**	35.8	16.7

Table 6: Results of pairwise human evaluation between CAST and three baselines on two style transfer tasks. Win/lose/tie indicate the percentage of results generated by CAST being better/worse/equal to the reference model.

style classifiers. We also observe that the exploitation of context information can help the model preserve semantic content in the original sentence (Example B).

Ablation Study To investigate the effectiveness of each component of CAST model, we conduct detailed ablation studies and summarize the results in Table 5. Experiments show that the context encoder and the coherence classifier play an important role in the proposed model. The context encoder is able to improve content preservation and style transfer accuracy, demonstrating the effectiveness of using context. The coherence classifier can help improve the coherence score but not much for style accuracy. By using these two components, our model can strike a proper balance between translating to the correct style and maintaining contextual consistency. When both of them are removed (the 4th row), performance on all the metrics drops significantly. We also observe that without using non-parallel data, the model performs poorly, showing the benefit of using a hybrid approach and more data for this task.

Human Evaluation Considering the subjective nature of this task, we conduct human evaluation to judge model outputs regarding content preservation, style control and context consistency. Given an original sentence along with its corresponding context and a pair of generated sentences from two different models, AMT workers were asked to select the best one based on these three aspects. The AMT interface also allows a neutral option, if the worker considers both sentences as equally good in certain aspect. We randomly sampled 200 sentences from the test set, and collected three human responses for each pair. Table 6 reports the pairwise comparison results on both tasks. Based on human judgment, the quality of transferred sentences by CAST is significantly higher than the other methods across all three metrics. This is consistent with the experimental results on automatic metrics discussed earlier.

6 Conclusion

In this paper, we present a new task - Contextual Text Style Transfer. Two new benchmark datasets are introduced for this task, which contain annotated sentence pairs accompanied by paragraph context. We also propose a new CAST model, which can effectively enforce content preservation and context coherence, by exploiting abundant non-parallel data in a hybrid approach. Quantitative and human evaluations demonstrate that CAST model significantly outperforms baseline methods that do not consider context information. We believe our model takes a first step towards modeling context information for text style transfer, and will explore more advanced solutions e.g., using a better encoder/decoder like GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019), adversarial learning (Zhu et al., 2020) or knowledge distillation (Chen et al., 2019).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2019. Distilling the knowledge of BERT for text generation. *CoRR*, abs/1911.03829.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*.

Zhiting Hu, Haoran Shi, Zichao Yang, Bowen Tan, Tiancheng Zhao, Junxian He, Wentao Wang, Lianhui Qin, Di Wang, et al. 2018. Texar: A modularized, versatile, and extensible toolkit for text generation. *arXiv preprint arXiv:1809.00794*.

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

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *ICLR*.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2019. Domain adaptive text style transfer. *arXiv preprint arXiv:1908.09395*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.

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

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *NeurIPS*.

Sourab Mangrulkar, Suhani Shrivastava, Veena Thenkanidiyoor, and Dileep Aroor Dinesh. 2018. A context-aware convolutional natural language generation model for dialogue systems. In *SIGDIAL*.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. *IEEE Spoken Language Technology Workshop (SLT)*.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

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 *NAACL*.

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

Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *EMNLP-IJCNLP*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NeurIPS*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015a. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015b. A neural network approach to context-sensitive generation of conversational responses. In *NAACL*.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*.

Jian Tang, Yifan Yang, Samuel Carton, Ming Zhang, and Qiaozhu Mei. 2016. Context-aware natural language generation with recurrent neural networks. *arXiv preprint arXiv:1611.09900*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Tian Wang and Kyunghyun Cho. 2015. Larger-context language modelling. *arXiv preprint arXiv:1511.03729*.

Wenlin Wang, Zhe Gan, Wenqi Wang, Dinghan Shen, Jiaji Huang, Wei Ping, Sanjeev Satheesh, and Lawrence Carin. 2017. Topic compositional neural language model. *arXiv preprint arXiv:1712.09783*.

Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. *arXiv preprint arXiv:1805.05181*.

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

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

Hongyu Zang and Xiaojun Wan. 2017. Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.