StoryTrans: Non-Parallel Story Author-Style Transfer with Discourse Representations and Content Enhancing

Xuekai Zhu1∗, Jian Guan1†, Minlie Huang2† and Juan Liu2

1The CoAI group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
2Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan, 430072, China
{xuekaizhu,liujuan}@whu.edu.cn, j-guan19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Non-parallel text style transfer is an important task in natural language generation. However, previous studies concentrate on the token or sentence level, such as sentence sentiment and formality transfer, but neglect long style transfer at the discourse level. Long texts usually involve more complicated author linguistic preferences such as discourse structures than sentences. In this paper, we formulate the task of non-parallel story author-style transfer, which requires transferring an input story into a specified author style while maintaining source semantics. To tackle this problem, we propose a generation model, named StoryTrans, which leverages discourse representations to capture source content information and transfer them to target styles with learnable style embeddings. We use an additional training objective to disentangle stylistic features from the learned discourse representation to prevent the model from degenerating to an auto-encoder. Moreover, to enhance content preservation, we design a mask-and-fill framework to explicitly fuse style-specific keywords of source texts in generation. Furthermore, we constructed new datasets for this task in Chinese and English, respectively. Extensive experiments show that our model outperforms strong baselines in overall performance of style transfer and content preservation.

1 Introduction

Text style transfer aims to endow a text with a different style while keeping its main semantic content unaltered. It has a wide range of applications, such as formality transfer (Jain et al., 2019), sentiment transfer (Shen et al., 2017) and author-style imitation (Tikhonov and Yamshchikov, 2018).

Due to the lack of parallel corpora, recent works mainly focus on unsupervised transfer by self-reconstruction. Current methods proposed to disentangle styles from contents by removing stylistic tokens from inputs explicitly (Huang et al., 2021) or reducing stylistic features from token-level hidden representations of inputs implicitly (Lee et al., 2021). This line of work has impressive performance on single-sentence sentiment and formality transfer. However, it is yet not investigated to transfer author styles of long texts such as stories, manifesting in the author’s linguistic choices at the lexical, syntactic, and discourse levels.

In this paper, we present the first study on story author-style transfer, which aims to rewrite a story incorporating source content and the target author style. The first challenge of this task lies in imitating of author’s linguistic choices at the discourse level, such as narrative techniques (e.g., brief or detailed writing). As exemplified in Table 1, the generation text for the JinYong (JY) style not only rewrites some tokens to the martial arts style (e.g., “白光一闪” /“white cloud” to “白光一闪” /“light flashing”) but also adds additional events in detail and

Table 1: An example that transfers a vernacular story to the martial arts style of JY generated by StyleLM. The orange sentence indicates missing content in source text. The rewritten token is underlined. The red highlights are supplementary short phrases or plots to align with the target style. The English texts below the Chinese are translated versions of the Chinese samples.

*Equal contribution.
†Corresponding author

1JinYong is a Chinese martial arts novelist.
enrich the storyline (e.g., the red highlights). In contrast to the transfer of token-level features like formality, it is more difficult to capture the intersentence relations correlated with author styles and disentangle them from contents. The second challenge is that the author styles tend to be highly associated with specific writing topics. Therefore, it is hard to transfer these style-specific contents to another style. For example, the topic “talented man” hardly shows up in the novels of JY, leading to the low content preservation of such contents, as shown in the orange text in Table 1.

To alleviate the above issues, we propose a generation framework, named StoryTrans, which learns discourse representations from source texts and then combines these representations with learnable style embeddings to generate texts of target styles. Furthermore, we propose a new training objective to reduce stylistic features from the discourse representations, which aims to pull the representations derived from different texts close in the latent space. To enhance content preservation, we separate the generation process into two stages, which first transfers the source text with the style-specific content keywords masked and then generates the whole text by imposing these keywords explicitly.

To support the evaluation of the proposed task, we collect new datasets in Chinese and English based on existing story corpora. We conduct extensive experiments to transfer fairy tales (in Chinese) or everyday stories (in English) to typical author styles, respectively. Automatic evaluation results show that our model achieves a better overall performance in style control and content preservation than strong baselines. The manual evaluation also confirms the efficacy of our model. We summarize the key contributions of this work as follows:

I. To the best of our knowledge, we present the first study on story author style transfer. We construct new Chinese and English datasets for this task.

II. We propose a new generation model named StoryTrans to tackle the new task, which implements content-style disentanglement and stylization based on discourse representations, then enhances content preservation by explicitly incorporating style-specific keywords.

III. Extensive experiments show that our model outperforms baselines in the overall performance of style transfer accuracy and content preservation.

2 The codes and data are available at https://github.com/Xuekai-Zhu/storytrans_public

2 Related Work

2.1 Style Transfer

Recent studies concentrated mainly on token-level style transfer of single sentences, such as formality or sentiment transfer. We categorize these studies into three following paradigms.

The first paradigm built a style transfer system without explicit disentanglement of style and content. This line of work used additional style signals or a multi-generator structure to control the style. Dai et al. (2019) added an extra style embedding in input for manipulating the style of texts. Yi et al. (2020) proposed a style instance encoding method for learning more discriminative and expressive style embeddings. The learnable style embedding is a flexible yet effective approach to providing style signals. Such a design helps better preserve source content. Syed et al. (2020) randomly dropped the input words, then reconstructed input for each author separately, which obtained multiple author-specific generators. The multi-generator structure is effective but also resource-consuming.

The second paradigm disentangled the content and style explicitly in latent space, then combined the target style signal. Zhu et al. (2021) diluted sentence-level information in style representations. John et al. (2019) incorporated style prediction and adversarial objectives for disentangling. Lee et al. (2021) removed style information of each token with reverse attention score (Bahdanau et al., 2015), which is estimated by a pre-trained style classifier. This paradigm utilizes adversarial loss functions or a pre-trained estimator for disentanglement. And experiment results indicate that explicit disentanglement leads to satisfactory style transfer accuracy without explicit disentanglement.

The final paradigm views style as localized features of tokens in a sentence, which locates style-dependent words and replaces the target-style ones. Xu et al. (2018) employed an attention mechanism to identify style tokens and filter out such tokens. Wu et al. (2019) utilized a two-stage framework to mask all sentimental tokens and then infill them. Huang et al. (2021) aligned words of input and reference to achieve token-level transfer. To sum up, this paradigm maintains all word-level information, but it is hard to apply to the scenarios where styles are expressed beyond token level, e.g., author style.

Absorbing ideas from paradigm 1 and 2, we
apply explicit disentanglement by pulling close discourse representations, which is formulated into disentanglement loss. Furthermore, we design a fusion module to stylize the discourse representation.

### 2.2 High-Level Representation

Prior works captured the hierarchical structure of natural language texts by learning high-level representations. Li et al. (2015) and Zhang et al. (2019) proposed to learn hierarchical embedding representations by reconstructing masked version of sentences or paragraphs. Reimers and Gurevych (2019) derived semantical sentence embeddings by fine-tuning BERT (Devlin et al., 2019) on downstream tasks. Lee et al. (2020); Guan et al. (2021b) inserted special tokens for each sentence and devised several pre-training tasks to learn sentence-level representations. We are inspired to use a sentence order prediction task to learn high-level discourse representations.

### 2.3 Long Text Generation

In order to generate coherent long texts, recent studies usually decomposed generation into multiple stages. Fan et al. (2018); Yao et al. (2019) generated a premise, then transformed it into a passage. Tan et al. (2021) first produced domain-specific content keywords and then progressively refines them into complete passages. Borrowing these ideas, we adopted a mask-and-fill framework to enhance content preservation in text style transfer.

### 3 Methodology

#### 3.1 Task Definition and Model Overview

We formulate the story author-style transfer task as follows: assuming that $S$ is the set of all author-styles, given a multi-sentence input $x = (x_1, x_2, \cdots, x_T)$ of $T$ tokens and its author-style label $s \in S$, the model should generate a multi-sentence text with a specified author-style $\hat{s} \in S$ while keeping the main semantics of $x$.

As illustrated in Figure 1, we split the generation process into two stages. We first identify style-specific keywords $k = (k_1, k_2, \cdots, k_l)$ from $x$, and then mask them with special tokens $<mask>$. We denote the resulting masked version of $x$ as $x^m = (x_1^m, x_2^m, \cdots, x_T^m)$. In the first generation stage, we perform discourse representation transfer on $x^m$. In the second stage, we complete the masked tokens in the output of the first stage conditioned on $k$ in a style-unrelated manner.

Due to the lack of parallel data, typical style transfer models tend to optimize the self-reconstruction loss with the same inputs and outputs (Xiao et al., 2021; Lee et al., 2021). Obviously, training with only the self-reconstruction loss will make the model easily ignore the target style signals and simply repeat the source inputs. Therefore, in the first stage, we devise an additional training objective, to disentangle stylistic features from intermediate discourse representations $\{r_i\}_{i=1}^n$, where $n$ is the number of sentences. Then, we fused these style-independent discourse representations with the target style $\hat{s}$ as a discourse-
Figure 2: Illustration of loss functions during training for the first stage (a) and second stage (b). Enc, Fus, Dec and C denote the encoder, the fusion module, the decoder, and style classifier, respectively.

level guidance for the subsequent generation of the transferred text. As for discourse representations learning, we employ a sentence order prediction loss to capture inter-sentence discourse dependencies. And we use a style classifier loss to control the style of generated texts (Lee et al., 2021). In summary, the first-stage model is trained using the following loss function:

\[ \mathcal{L}_1 = \mathcal{L}_{\text{self}} + \lambda_1 \mathcal{L}_{\text{dis}} + \lambda_2 \mathcal{L}_{\text{sop}} + \lambda_3 \mathcal{L}_{\text{style}}, \]  

(1)

where \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) are adjustable hyperparameters. \( \mathcal{L}_{\text{self}}, \mathcal{L}_{\text{dis}}, \mathcal{L}_{\text{sop}} \) and \( \mathcal{L}_{\text{style}} \) are the self-reconstruction loss, the disentanglement loss, the sequence order prediction loss and the style classifier loss, respectively. Figure 2 shows the workflow of learning objects.

In the second stage, we use a denoising auto-encoder (DAE) loss to train another encoder-decoder model for reconstructing \( x \):

\[ \mathcal{L}_2 = - \sum_{t=1}^T \log P(x_t|x_{<t}, \{k_i\}_{i=1}^n, x^m). \]  

(2)

This stage is unrelated to author styles, and helps achieve better content preservation.

3.2 Discourse Representations Transfer

As described in Figure 2, we propose to learn discourse representations, and then reconstruct the texts from discourse representations. And we perform the disentanglement and stylizing operation based on discourse representations.

Discourse Representations Suppose that \( x^m \) consists of \( n \) sentences, we insert a special token \( \langle \text{Sen} \rangle \) at the end of each sentence in \( x^m \) (Reimers and Gurevych, 2019; Lee et al., 2020; Guan et al., 2021b). Let \( r_n \) denote the hidden state of the encoder at the position of the \( n \)-th special token, \( \{r_i\}_{i=1}^n = \text{Encoder}(x^m) \). And \( z_n \) is the output of the fusion module corresponding to \( r_n \). Previous studies have demonstrated that correcting the order of shuffled sentences is a simple but effective way to learn meaningful discourse representations (Lee et al., 2020). As shown in Figure 1, we feed \( z_n \) into a pointer network (Gong et al., 2016) to predict orders. During training, we shuffled the original sentence order and feed the perturbed text into the encoder for calculating \( \mathcal{L}_{\text{sop}} \).

Fusion Module To provide signals of transfer direction, we concatenate the learned discourse representations \( \{r_i\}_{i=1}^n \) with the style embedding \( s \) and fuse them using a multi-head attention layer, as illustrated in Figure 1. To capture discourse-level features of texts with different author-styles, we set each style embedding to a vector with the same dimension as \( r_i \). Formally, we derive the style-aware discourse representations \( \{z_i\}_{i=1}^{n+1} \) as follows:

\[ \{z_i\}_{i=1}^{n+1} = \text{MHA}(Q = K = V = \{s \| \{r_i\}_{i=1}^n \}), \]  

(3)

where \( \text{MHA} \) is the multi-head attention layer, \( Q/K/V \) is the corresponding query/key/value, \( \| \) is the concatenation operation. Then, the decoder gets access to \( \{z_i\}_{i=1}^{n+1} \) through the cross-attention layer, which serves as a discourse-level guidance for generating the transferred texts. Then, we feed \( \{z_i\}_{i=1}^{n+1} \) into the decoder.

Pointer Network Following Logeswaran et al. (2018); Lee et al. (2020), we use a pointer network to predict the original orders of the shuffled sentences. The each position probability of sentence order is formulated as follows:

\[ p_i = \text{softmax}(\{z_i\}_{i=1}^n W_{z_i}^T), \]  

(4)

where \( p_i \) is predicted probabilities of sentence \( i \), \( W \) is a trainable parameter.

3.3 First-Stage Training Objectives

Self-Reconstruction Loss We formulate self-reconstruction loss as follows:

\[ \mathcal{L}_{\text{self}} = - \sum_{t=1}^T \log P(x_t^m|x_{<t}, \{r_i\}_{i=1}^n, s), \]  

(5)

where \( s \) is the learnable embedding of \( s \). During inference, we replace \( s \) with the embedding of the target style \( \hat{s} \) (i.e., \( \hat{s} \)), to achieve the style transfer.
Disentanglement Loss. We disentangle the style and content on discourse representations. Inspired by prior studies on structuring latent spaces (Gao et al., 2019; Zhu et al., 2021), we devise an additional loss function $\mathcal{L}_{\text{dis}}$ to pull close discourse representations from different examples in the same mini-batch, corresponding to different author styles. $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{self}}$ work as adversarial losses and lead the model to achieve a balance between content preservation and style transfer. We derive $\mathcal{L}_{\text{dis}}$ as follows:

$$L_{\text{dis}} = \frac{1}{2b} \sum_{i=1}^{b} \sum_{j=1}^{b} || \hat{r}_i - \hat{r}_j ||_2^2,$$

where $b$ is the size of mini-batch.

Sentence Order Prediction Loss. We formulate $\mathcal{L}_{\text{sop}}$ as the cross-entropy loss between the golden and predicted orders as follows:

$$L_{\text{sop}} = -\frac{1}{n} \sum_{i=1}^{n} o_i \log(p_i),$$

where $o_i$ is a one-hot ground-truth vector of correct sentence position, and $p_i$ is predicted probabilities.

Style Classifier Loss. We expect the transferred text to be of the target style. Hence we train a style classifier to derive the style transfer loss as follows:

$$L_{\text{style}} = -E_{x^m \sim \text{Decoder}}[\log P_C(s|\hat{x}^m)],$$

where $P_C$ is the conditional distribution over styles defined by the classifier. We train the classifier on the whole training set with the standard cross-entropy loss. Then, we freeze the weights of style classifier for computing $L_{\text{style}}$. On the other hand, we follow Lee et al. (2021); Dai et al. (2019) to use soft sampling to allow gradient back-propagation.

3.4 Content Preservation Enhancing

As aforementioned, author styles have a strong correlation with contents. It is difficult to transfer such style-specific contents to other styles directly. Since we train the model in an auto-encoder manner, it will have no idea how to transfer those content representations that have never seen other style embeddings during training. To address the issue, we propose to mask the style-specific keywords in the source text and perform style transfer on the masked text in the first generation stage. Then, we fill the masked tokens in the second stage.

We follow Xiao et al. (2021) to use a frequency-based method to identify the style-specific keywords. Specifically, we extract style-specific keywords by (1) obtaining the top-10 words with the highest TF-IDF scores from each corpus, (2) retaining only people’s names, place names, and proper nouns, (3) and filtering out those words with a high frequency in all corpora. We denote the resulting word set as $D^s$ for the corpus with the style $s$. We extract the style-specific keywords $k$ from the text $x$ by selecting the words that are in $D^s$. We detail above operation and explain it in Appendix A.

In the second stage, we train another model to fill the mask tokens in outputs of the first stage conditioned on the identified style-specific keywords in source inputs. During training, we concatenate the keywords in $k$ with a special token (key) and feed them into the encoder paired with $x^m$, as shown in Figure 1. The training object is formulated as Equation 2. During inference, the decoder generates the transferred text $\hat{x}$ conditioned on the output of the first stage $\hat{x}^m$ in an auto-regressive manner.

4 Experiments

4.1 Datasets

We construct stylized story datasets in Chinese and English, respectively. The Chinese dataset consists of three styles of texts, including fairy tales from LOT (Guan et al., 2021a), LuXun (LX), and JinYong (JY). Specifically, LuXun writes realism novels while JinYong focuses on martial arts novels. These texts of different styles have a gap in lexical, syntactic, and semantic levels. Samples of different styles are detailed in Appendix C.

In our experiments, we aim to transfer a fairy tale to the LX or JY style. The English dataset consists of tales from Shakespeare (ROC) and JinYong’s novels while JinYong focuses on martial arts novels. These texts of different styles have a gap in lexical, syntactic, and semantic levels. Samples of different styles are detailed in Appendix C.

We set those words appearing in at least 10% samples in a corpus as high-frequency words.
of two styles of texts, including everyday stories from ROCStories (Mostafazadeh et al., 2016) and fragments from Shakespeare’s plays. We expect to transfer a five-sentence everyday story into the Shakespeare style. The statistics of datasets are shown in Table 2. The more details are described in Appendix B.

4.2 Implementation

We take LongLM\textsubscript{BASE} (Guan et al., 2021a) and T5\textsubscript{BASE} (Raffel et al., 2020) as the backbone model of both generation stages for Chinese and English experiments, respectively. Furthermore, the fusion module and pointer network consist of two and one layers of randomly initialized bidirectional Transformer blocks (Vaswani et al., 2017), respectively. We conduct experiments on one RTX 6000 GPU. In addition, we build the style classifier based on the encoder of LongLM\textsubscript{BASE} and T5\textsubscript{BASE} for Chinese and English, respectively.

We set $\lambda_1/\lambda_2/\lambda_3$ in Equation 1 to 1/1/1, the batch size to 4, the learning rate to 5e-5, the maximum sequence length of the encoder and decoder to 512 for both generation stages in the Chinese experiments. And the hyper-parameters for English experiments are the same except that $\lambda_1/\lambda_2/\lambda_3$ are set to 0.5/0.5/0.5 and the learning rate to 2.5e-5. More implementation details are presented in Appendix D.

4.3 Baselines

Since no previous studies have focused on story author-style transfer, we build several baselines by adapting short-text style transfer models. For a fair comparison, we initialize all baselines using the same pre-trained parameters as our model. Specifically, we adopt the following baselines:

**Style Transformer:** It adds an extra style embedding and a discriminator to provide style transfer rewards without disentangling content from styles (Dai et al., 2019).

**StyleLM:** This baseline generates the target text conditioned on the given style token and corrupted version of the original text (Syed et al., 2020).

**Reverse Attention:** It inserts a reverse attention module on the last layer of the encoder, which aims to negate the style information from the hidden states of the encoder (Lee et al., 2021).

4.4 Automatic Evaluation

**Evaluation Metrics** Previous works evaluate style transfer systems mainly from three aspects including style transfer accuracy, content preservation, and sentence fluency. A good style transfer system needs to balance the contradiction between content preservation and transfer accuracy (Zhu et al., 2021; Niu and Bansal, 2018). We use a joint metric to evaluate the overall performance of models. On the other hand, previous studies usually use perplexity (PPL) of a pre-trained language model. However, in our experiments, we found that the PPL of model outputs is lower than human-written texts, suggesting that PPL is not reliable for evaluating the quality of stories. Therefore, we evaluate the fluency through manual evaluation.

Specifically, we adopt the following automatic metrics: **(1) Style Transfer Accuracy:** We use two variants of style transfer accuracy following Krishna et al. (2021), absolute accuracy (a-Acc) and relative accuracy (r-Acc). We train a style classifier and regard the classifier score as the a-Acc. And r-Acc is a binary value to indicate whether the style classifier score the output higher than the input (1/0 for a higher/lower score). We train the classifier by fine-tuning the encoder of LongLM\textsubscript{BASE} and T5\textsubscript{BASE} on the Chinese and English training set, respectively. The classifier achieves a 99.6% and 99.41% accuracy on the Chinese and English test sets, respectively. **(2) Content Preservation:** We use BLEU-$n$ ($n=1,2$) (Papineni et al., 2002) and BERTScore (BS) (Zhang et al., 2020) between generated and input texts to measure their lexical and semantic similarity, respectively. And we report recall (BS-R), precision (BS-P) and F1 score (BS-F1) for BS. **(3) Overall:** We use the geometric mean of a-ACC and BLEU/BS-F1 score (BL-Overall/BS-Overall) to assess the overall performance of models (Krishna et al., 2020; Lee et al., 2021).

**Results on the Chinese Dataset** We show the overall performance and individual metrics results in Table 3. In terms of overall performance, StoryTrans outperforms baselines, illustrating that StoryTrans can achieve a better balance between style transfer and content preservation.

In terms of style accuracy, StoryTrans achieves the best style transfer accuracy (a-Acc) in LX and comparable performance in JY. The bad performance of baselines indicates the necessity to perform explicit disentanglement beyond the token level. In addition, manual inspection shows that Style Transformer tends to copy the input, accounting for the highest BLEU score and BERTScore.
This means Style Transformer only takes the target style signals as noise, which may result from the stylistic features existing in the contents. StyleLM and Reverse Attention get better transfer accuracy than Style Transformer by removing such stylistic features from the contents. Moreover, Reverse Attention obtains better style accuracy but worse content preservation than StyleLM. Therefore, re-weighting hidden states allows better control over style than deleting input words explicitly.

In terms of content preservation, StoryTrans outperforms Reverse Attention. Additionally, StyleLM achieves better performance in content preservation, benefiting from inputting noisy versions of golden texts. But without disentanglement, it can’t strip style information. This leads to a lower overall performance than StoryTrans. As for Style Transformer, the results demonstrate that only an attention-based model hardly removes style features in overwhelming tokens information, leading to degenerate into an auto-encoder.

### Results on the English Dataset

Similarly, StoryTrans achieves the best overall performance on the English dataset, showing its effectiveness and generalization. And StoryTrans outperforms baselines significantly in terms of style transfer accuracy. As for content preservation, Style Transformer and Reverse Attention degenerate into an auto-encoder, and tend to copy the input even more than their performance on the Chinese dataset.

Table 3: Automatic evaluation results on the test set of the Chinese and English datasets. Bold numbers indicate best performance. ZH-LX/ZH-JY is the Chinese author LuXun/JinYong, respectively. EN-SP is the English author Shakespeare. StoryTrans achieves the best overall performance (BL/BS-Overall), with a good trade-off between style accuracy (r/a-Acc) and content preservation (BLEU-1/2 and BS-P/R/F1).

| Target Styles | Models       | r-Acc | a-Acc | BLEU-1 | BLEU-2 | BS-P | BS-R | BS-F1 | BL-Overall | BS-Overall |
|---------------|-------------|-------|-------|--------|--------|------|------|-------|------------|------------|
| ZH-LX         | Style Transformer | 65.84 | 0.13  | 82.53  | 77.17  | 96.92 | 96.51 | 96.70 | 2.96       | 3.26       |
|               | StyleLM     | 97.80 | 33.33 | 39.43  | 19.66  | 77.71 | 75.02 | 76.30 | 31.38      | 50.42      |
|               | Reverse Attention | 98.49 | 42.93 | 20.98  | 6.70   | 65.38 | 63.39 | 64.35 | 24.37      | 52.55      |
|               | Story Trans | 97.96 | 59.94 | 32.19  | 14.44  | 68.53 | 70.48 | 69.45 | 57.38      | 64.52      |
| ZH-JY         | Style Transformer | 46.77 | 0.13  | 83.24  | 77.85  | 96.82 | 96.97 | 97.15 | 3.23       | 3.55       |
|               | StyleLM     | 79.97 | 51.16 | 36.72  | 18.03  | 74.20 | 75.19 | 74.62 | 37.41      | 61.78      |
|               | Reverse Attention | 94.51 | 66.39 | 21.15  | 6.32   | 64.05 | 65.08 | 64.54 | 30.19      | 65.45      |
|               | Story Trans | 84.39 | 62.96 | 30.71  | 14.5   | 68.76 | 71.09 | 70.16 | 37.12      | 66.46      |
| EN-SP         | Style Transformer | 0.34  | 0.01  | 99.88  | 99.88  | 87.10 | 95.43 | 90.78 | 3.31       | 3.16       |
|               | StyleLM     | 57.93 | 3.44  | 37.05  | 19.40  | 84.72 | 90.53 | 87.30 | 9.85       | 17.32      |
|               | Reverse Attention | 20.68 | 0.01  | 96.90  | 96.16  | 86.93 | 95.27 | 90.61 | 3.29       | 3.15       |
|               | Story Trans | 88.62 | 52.41 | 32.20  | 12.71  | 81.77 | 87.51 | 84.31 | 34.31      | 66.47      |

This means Story Transformer only takes the target style signals as noise, which may result from the stylistic features existing in the contents. StyleLM and Reverse Attention get better transfer accuracy than Style Transformer by removing such stylistic features from the contents. Moreover, Reverse Attention obtains better style accuracy but worse content preservation than StyleLM. Therefore, re-weighting hidden states allows better control over style than deleting input words explicitly.

**Table 4: Ablation study results on English datasets. (-) indicates removing the component in proposed model. CE denote content enhancing, which means removing the second stage. More ablation results shown in Appendix E.**

|                  | Proposed Model | r-Acc | a-Acc | BLEU-1 | BLEU-2 | BS-P | BS-R | BS-F1 | BL-Overall | BS-Overall |
|------------------|----------------|-------|-------|--------|--------|------|------|-------|------------|------------|
|                  | r-Acc | a-Acc | BLEU-1 | BLEU-2 | BS-P | BS-R | BS-F1 | BL-Overall | BS-Overall |
| (-) L_{dis}      | 75.86 | 31.37 | 33.49  | 14.52  | 82.38 | 88.07 | 84.92 | 27.44      | 51.61      |
| (+) L_{style}    | 50.68 | 7.93  | 45.00  | 23.79  | 84.38 | 89.16 | 86.5  | 16.51      | 26.19      |
| (+) L_{sop}      | 78.96 | 38.96 | 39.45  | 19.20  | 82.92 | 88.62 | 85.47 | 33.80      | 57.70      |
| (+) CE           | 92.41 | 73.10 | 21.62  | 6.09   | 79.73 | 86.12 | 82.59 | 31.82      | 77.70      |

Table 6: Human evaluation results on Chinese for transfer direction in LX and JY. \( \kappa \) denotes Fleiss’ kappa (Fleiss, 1971) to measure the inter annotator agreement (all are moderate or substantial). The scores marked with \(*\) indicate that StoryTrans outperforms the baselines significantly with \( p<0.01 \) (sign test).

| Models       | LX     | JY     | \( \kappa \) | \( \kappa \) |
|--------------|--------|--------|-------------|-------------|
| Style Transformer | 1.02   | 1.20** | 1.20**      | 0.89        |
| StyleLM     | 1.61   | 1.99   | 1.58 0.20   | 1.79 1.92 0.23 |
| Reverse Attention | 1.09   | 1.25   | 1.21 0.21   | 1.05 1.25 0.20 |
| Story Trans | 1.84** | 1.94   | 1.87 0.24   | 2.48** 1.99 0.20 |

Results on Ablation Study

As shown in Table 4, we observe a significant drop in transfer accuracy without \( L_{dis} \) or \( L_{style} \). \( L_{dis} \) works by disentangling stylistic features from the discourse representations, while \( L_{style} \) exerts direct supervision on styles of generated texts. Without \( L_{sop} \), model can hardly capture discourse-level information and keeps more source tokens, leading to higher BLEU scores and lower accuracy. When removing the second stage, the lowers BLEU scores show the benefit of the mask-and-fill framework for content preservation.
柯里教授独自攀登了上去,到达了山脉最高峰。只听得脚步声似乎越走越近,其后似乎还发出一些喘息之声。

郭靖双手撑在墙上,远远望去,柯里和教授站住了,只见那道门缓缓闭上,大雾随即散开,他迈开了步子,继续前进,脚步声音也随之消失。柯里先生 ... Chinese test set.

Table 5: Cases generated by different models, which are transferred from the fairy tale style (ZH) to the JY style and from StoryTrans and three baseline models. Then, we hire three Chinese native speakers to evaluate the mask-and-fill framework.

4.6 Case Study
Table 5 shows the cases generated by StoryTrans and the best baseline. StyleLM inserts many unrelated sentences, which overwhelm the original content and impair the coherence, further leading to the content loss of sentences 3 and 4. On the contrary, StoryTrans supplement several short phrases or plots (e.g., “纵身跃起” / “hurriedly jumped up”) to enrich the storyline and maintain the main content. Furthermore, StoryTrans can rewrite most sentences with the target style and maintain source semantics. In addition, StyleLM tends to discard the source entities and use words which is specific in the target style (e.g., “郭靖” / “Guo Jing”), while StoryTrans dose not, suggesting the necessity of the mask-and-fill framework.

4.7 Stylistic Feature Visualization
We follow Syed et al. (2020) to define several stylistic features and visualize the features of the golden texts and generated texts on the Chinese test set. The stylistic features include the type and number of punctuation marks, the number of sentences, and
the number of words. As shown in Figure 3, the texts generated by Reverse Attention and StyleLM have similar stylistic features to source texts. In contrast, StoryTrans can better capture different stylistic features and transfer source texts to specified styles. More details are in Appendix F.

5 Conclusion

In this paper, we present the first study for story author-style transfer and analyze the difficulties of this task. Accordingly, we propose a novel generation model, which explicitly disentangles the style information from high-level text representations to improve the style transfer accuracy, and achieve better content preservation by injecting style-specific contents. Automatic evaluations show StoryTrans outperform baselines on the overall performance. Further analysis shows StoryTrans has a better ability to capture linguistic features for style transfer.

Limitations

In style transfer, content preservation and style transfer are adversarial. Long texts have richer contents and more abstract stylistic features. We also notice that content preservation is the main disadvantage of StoryTrans in automatics evaluation results. Case studies also indicate that StoryTrans can maintain some entities and the relations between entities. However, strong discourse-level style transfer ability endangered content preservation. In contrast, baselines such as Style Transformer have better content preservation but hardly transfer the style. We believe that StoryTrans is still a good starting point for this important and challenging task.

During preliminary experiments, we also manually inspected multiple author styles besides Shakespeare, such as Mark Twain. However, we found that their styles are not as obvious as Shakespeare, as shown in the following example. Therefore, we only selected authors with relatively distinct personal styles for our transfer experiments. In future work, we will expand our research and choose more authors with distinct styles for style transfer. For example, the style distinction between the following examples is not readily apparent.

- "A Double Barreled Detective Story" by Mark Twain: You will go and find him. I have known his hiding-place for eleven years; it cost me five years and more of inquiry.

Ethics Statement

We perform English and Chinese experiments on public datasets and corpora. Specifically, English datasets come from ROCCstories and Project Gutenberg. Moreover, Chinese datasets include the LOT dataset and public corpora of JY and LX. Automatic and manual evaluation demonstrate that our model outperforms strong baselines on both Chinese and English datasets. In addition, our model can be easily applied to different baselines by substituting specific pre-trained language models.

As for manual evaluation, we hired three native Chinese speakers as annotators to evaluate generated texts and did not ask about personal privacy or collect the personal information of annotators. We pay 1.8 yuan (RMB) per sample in compliance with Chinese wage standards. Considering it would cost an average of 1 minute for an annotator to score a sample, the payment is reasonable.

Acknowledgments

This work was supported by the NSFC projects (Key project with No. 61936010 ). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2020GQG0005.

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, ICLR 2015.

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

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of
the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1814–1823, Hong Kong, China. Association for Computational Linguistics.

Jingjing Gong, Xinchu Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. arXiv preprint arXiv:1611.04953.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021a. Lot: A benchmark for evaluating chinese long text understanding and generation.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021b. Long text generation by modeling sentence-level and discourse-level coherence. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6379–6393, Online. Association for Computational Linguistics.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. Advances in neural information processing systems, 15.

Fei Huang, Zikai Chen, Chen Henry Wu, Qihan Guo, Xiaoyan Zhu, and Minlie Huang. 2021. NAST: A non-autoregressive generator with word alignment for unsupervised text style transfer. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1577–1590, Online. Association for Computational Linguistics.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6554–6561.

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, Florence, Italy. Association for Computational Linguistics.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2021. Few-shot controllable style transfer for low-resource settings: A study in indian languages. arXiv preprint arXiv:2110.07385.

Kalpesh Krishna, John Wieeting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 737–762, Online. Association for Computational Linguistics.

Dongkyu Lee, Zhihong Tian, Lanqing Xue, and Nevin L. Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 93–102, Online. Association for Computational Linguistics.

Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1551–1562, Online. Association for Computational Linguistics.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1106–1115, Beijing, China. Association for Computational Linguistics.

Lajanugen Logeswaran, Hongkak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In Thirty-second aaai conference on artificial intelligence.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 839–849, San Diego, California. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. Transactions of the
What do we mean by "style-specific content"?

We detail how we extract style-specific contents and explain how they are used from the following three aspects:

**What do we mean by “style-specific content”?**

We refer to "style-specific content" as those mainly used in texts with specific styles and should be retained after style transfer. For example, "Harry Potter" and "Horcrux" are style-specific since they are used only in J.K. Rowling-style stories. When transferring J.K. Rowling-style stories to another author's work, if we want to keep the style of the original text, we need to extract and preserve style-specific content.
other styles, style-specific tokens shouldn’t be changed. However, existing models tend to drop style-specific tokens since they are not trained to learn these tokens conditioned on other styles.

**How do we extract “style-specific contents”?** We extract style-specific contents by (1) obtaining top-10 salient tokens using TF-IDF, (2) reserving only people names (e.g., "Harry Potter"), place names (e.g., "London"), and proper nouns (e.g., "Horcrux"), and (3) filtering out high-frequency tokens in all corpus (e.g., "London") since these tokens can be learned conditioned on every style. We regard the remaining tokens as style-specific contents.

As mentioned before, we employ the TF-IDF algorithm on the corpus to obtain rough style-specific contents for different styles, respectively. The reason for using TF-IDF: it is necessary to ensure that the extracted tokens are salient to the story plots. We extract style-specific tokens from the salient tokens using the second and third steps. Then, we use a part-of-speech tagging toolkit (e.g., NLTK) to identify function words and prepositions to retain people’s names, place names, and proper nouns. Note that the frequency is an empirical value observed from datasets. However, the TF-IDF algorithm chooses the important words corresponding to the special style based on word frequency. There may be some style-unrelated words that are important to the content. Therefore, we need to filter out style-unrelated words. Concretely, we use Jieba/NLTK (Bird et al., 2009) to collect the word frequency for Chinese/English dataset in data pro-processing. In addition, these data are public corpora, and we also check the information for anonymization.

Regarding to limitation of modern language models, the length of samples is also limited. We set the max length as 384 and 90 for Chinese and English, respectively. Each sample has 4 sentences at least. We choose above length to balance the data length of different styles. Additionally, we filtered the texts which are too long to generate or too short to unveil author writing style. As Figure 4 shows, texts in the Chinese dataset spans a diverse range of length.

**C Different Style Samples**

In process of constructing datasets, we try to collect different author corpus who have a gap in writing styles. As shown in Table 8, the JY-style texts mostly describe martial arts actions and construct interesting plots, while the LX-style texts focus on realism with profound descriptive and critical significance. And the fairy tales differ from these texts in terms of topical and discourse features. In the English datasets, the Shakespeare-style texts are flamboyant and contain elaborate metaphors and ingenious ideas, which the everyday stories are written in plain language and without rhetoric.

Then, the "style-specific contents" will be filled in the second stage, as shown in Figure 1.

**B Data Pre-Processing**

Due to lack of stylized author datasets, we collected several authors’ corpus to construct new datasets. As for Chinese, we extracted paragraphs from 21 novels of LuXun (LX) and 15 novels of JinYong (JY), and fairy tales collected by Guan et al. (2021a). On the other hand, the English dataset consists of everyday stories from ROCStories (Mostafazadeh et al., 2016) and fragments from Shakespeare’s plays. Each fragment of Shakespeare’s plays comprises multiple consecutive sentences and as long as samples in ROC-Stories. We collect the Shakespeare-style texts from the Shakespeare corpus in Project Gutenberg under the Project Gutenberg License. We use Jieba/NLTK (Bird et al., 2009) for word tokenization for the Chinese/English dataset in data pro-processing. In addition, these data are public corpora, and we also check the information for anonymization.

5https://www.gutenberg.org
6https://www.gutenberg.org/policy/license.html

---

4https://github.com/fxsjy/jieba

---

14814
Once, when Professor Curry was climbing the highest peak of the Cairngorm Mountains alone, he heard some huge footsteps and panting behind him. He turned around in one breath, then there was a loud bang in front of him, and at the same time, there was a sound of panting and shouting from a distance.

Table 7: More Chinese cases generated by baselines, which are transferred from the fairy tale style to the JY style. The number before the sentences indicate their corresponding sentences in the source text in semantics. The table shows how different techniques perform on this task.

| Authors | Texts |
|---------|-------|
| JY      | 杨贵朝着马鞍上的马鞍，双腿一夹，小红马便飞快地向前奔跑。郭某不敢怠慢，便催着小红马顺着小路向前跑去。 |
| LX      | 戍《新序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序》篇以《旧序}
observed on a validation set, which was created by pre-extracting 5% of the training data for this purpose.

E More Ablation Study Results

To explore the effect of the proposed component, we also conduct more ablation studies on Chinese datasets. As shown in Table 9, the ablation of $L_{\text{dis}}$ leads to better style accuracy, which show the different trends comparing with English dataset. We conjecture that $L_{\text{dis}}$ aims to maintain the content and reduces style information. Without $L_{\text{dis}}$, the powerful $L_{\text{style}}$ leads the StoryTrans to degenerate to style conditional language model. Furthermore, the ablation of $L_{\text{style}}$ also confirms the powerful ability of style control as in previous paper. And we find that when removing $L_{\text{dis}}$, the model loses the ability to transfer at the discourse level and has only learned token-level copy.

F Style Analysis of Transferred Texts

In order to investigate whether our StoryTrans indeed rephrase the expression of texts, we employ surface elements of text to show author writing styles. And the surface element are associated with statistical observations. For example, the small average length of sentences show the author preference to write a short sentence, and more question marks indicate the author accustomed to using questions. To this end, we use the number of (1) commas, (2) colons, (3) sentences in a paragraph, (4) question mark (5) left quotation mark, (6) right quotation mark, and (7) average number of words in a sentence to quantify surface elements into a 7 dimension vector. Then we leverage the t-SNE to visualize the golden texts and transferred texts. As shown in Figure 3, different style distribute separately across the style space. This proves JY, LX and fairy tale in Chinese dataset have a gap in writing style. And Figure 5 shows the transferred texts fall in golden texts in style space, indicating StoryTrans successfully transferred the writing style.

G More Details of Manual Evaluation

In addition to automatic evaluation, we conduct manual evaluation on generated texts. As mentioned before, we require the annotators to score each aspect from 1 (the worst) to 3 (the best). As for payment, we pay 1.8 yuan (RMB) per sample in compliance with Chinese wage standards. Our annotators consist of undergraduate students who are experienced in reading texts written in the styles of the respective authors (JY and LX). To ensure they fully understand the evaluation metrics, we conducted case analyses with them. Our scoring rubric assigns 1, 2, or 3 points to the transferred text based on the proportion of sentences meeting the following criteria (1/3, 2/3, or 3/3):

- Style Accuracy: whether the transferred text conforms to the corresponding style.
- Content Preservation: whether the source content, such as character names, are retained.
- Coherence: whether the sentences in the transferred text are semantically connected.

And we compute the final score of each text by averaging the scores of three annotators.

As illustrated in manual evaluation, we observe that the results mainly conform with the automatic evaluation. Our StoryTrans obtained the highest score on the style accuracy in both transferred directions by a sign test compared to the other baselines, showing its stable ability of style control. Moreover, in terms of content preservation, the score
Table 9: More ablation study results on Chinese datasets. (-) indicates removing the component in proposed model.
ZH-LX/ZH-JY is the Chinese author LuXun/JinYong, respectively.

of StoryTrans is comparable with StyleLM and slightly higher than Reverse Attention, demonstrating that StoryTrans can keep the main semantics of input. In terms of coherence, the score of StoryTrans is also comparable with baselines, showing some room for improvement. As discussed before, Style Transformer tends to copy the input, leading to the highest performance in content preservation and coherence. In summary, human evaluation depicts the strength of StoryTrans not only on style control but also on overall performance, indicating a balance of these metrics.

H More Case Studies

We show more cases in Table 7. Comparing source text with Style Transformer, Style Transformer copies the input and only changes little tokens. This result also confirms with highest BLEU and BERTScore in automatic results. Like StyleLM, Reverse Attention also incorporates some target author content into generated texts. However, Reverse Attention inserts too much content that overwhelms original plots. Furthermore, some critical entities (e.g., character name, "柯里教授" /"Professor Curry" → "柯镇恶" /"Ke Zhen’e") are revised to the similar word on in target author corpus. To maintain the story coherence, these important entities should stay the same. In summary, the token-level transfer may destroy the essential plots and damage the coherence.
ACL 2023 Responsible NLP Checklist

A  For every submission:

✓ A1. Did you describe the limitations of your work?
   6

✓ A2. Did you discuss any potential risks of your work?
   6

✓ A3. Do the abstract and introduction summarize the paper’s main claims?
   1

✗ A4. Have you used AI writing assistants when working on this paper?
   Left blank.

B  ✓ Did you use or create scientific artifacts?
   4.2

✓ B1. Did you cite the creators of artifacts you used?
   4.2

✓ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
   Appendix B

✓ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
   Appendix B

✓ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
   Appendix B

✓ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
   4.1 and Appendix B

✓ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
   We counted the details of the dataset and discussed the details in Section 4.1

C  ✓ Did you run computational experiments?
   4.4

✓ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
   4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?  
Appendix B

D ✓ Did you use human annotators (e.g., crowdworkers) or research with human participants?

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?  
Appendix B and Appendix F

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?  
Not applicable. We do not submit the protocol to an ethics review board because our country has not yet established an ethical committee at the national level.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?  
4.5 and Appendix F