Disentangled Learning with Synthetic Parallel Data for Text Style Transfer

Jingxuan Han[1], **Quan Wang**[3], **Zikang Guo**[1], **Benfeng Xu**[1]
Licheng Zhang[1] and **Zhendong Mao**[1,2*]
[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]MOE Key Laboratory of Trustworthy Distributed Computing and Service,
Beijing University of Posts and Telecommunications, Beijing, China
{hjx999222,gzk170401,benfeng,zlczlc}@mail.ustc.edu.cn
wangquan@bupt.edu.cn, zdmao@ustc.edu.cn

Abstract

Text style transfer (TST) is an important task in natural language generation, which aims to transfer the text style (e.g., sentiment) while keeping its semantic information. Due to the absence of parallel datasets for supervision, most existing studies have been conducted in an unsupervised manner, where the generated sentences often suffer from high semantic divergence and thus low semantic preservation. In this paper, we propose a novel disentanglement-based framework for TST named DisenTrans, where disentanglement means that we separate the attribute and content components in the natural language corpus and consider this task from these two perspectives. Concretely, we first create a disentangled Chain-of-Thought prompting procedure to synthesize parallel data and corresponding attribute components for supervision. Then we develop a disentanglement learning method with synthetic data, where two losses are designed to enhance the focus on attribute properties and constrain the semantic space, thereby benefiting style control and semantic preservation respectively. Instructed by the disentanglement concept, our framework creates valuable supervised information and utilizes it effectively in TST tasks. Extensive experiments on mainstream datasets present that our framework achieves significant performance with great sample efficiency.

1 Introduction

Text style transfer (TST) aims to modify sentence style with its semantics unchanged. The main task in TST is sentiment transfer, along with some other tasks like politeness, formality and humor transfer. TST has wide extensive applications, such as neutralizing offensive remarks (Nogueira dos Santos et al., 2018), data augmentation (Xu et al., 2019), and human-computer interaction (Li et al., 2016b).

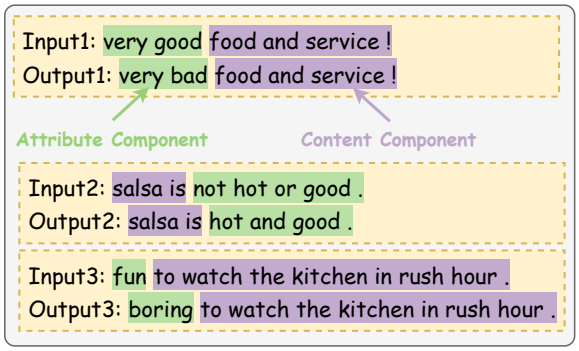

Figure 1: Three cases from the test set of sentiment transfer dataset Yelp, which can reflect the disentangled components.

The main challenge in TST is the lack of parallel data as supervised information, so the existing mainstream works design their frameworks in an unsupervised manner for training non-parallel data. These frameworks employ an additional style embedding into the Transformer architecture and design at least three losses to construct supervision, such as cycle consistency loss (Dai et al., 2019), self-construct loss (Ma and Li, 2021) and style loss (Lee et al., 2021). However, without parallel data as supervised information, the transferred sentences often exhibit high semantic divergence, which will consequently cause low semantic preservation and reduced human-like characteristics.

In this paper, we propose a disentanglement-based TST framework (**DisenTrans**), in which we create a disentangled data synthesis paradigm and construct a disentanglement learning method to learn from synthetic data. Specially, we employ a novel disentanglement concept to instruct our framework. This concept considers that a sentence is identified with a particular style mainly because some parts exhibit the characteristics of that style. The parts with these style characteristics are defined as the attribute component, while the remaining parts with content characteristics are defined as

*Corresponding author: Zhendong Mao.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15187–15201

the content component. Figure 1 shows three cases from the test set of the Yelp (Li et al., 2018) sentiment transfer dataset. The attribute components of these three cases are "very good", "not hot or good" and "fun" respectively. Obviously, the modification of attribute components determines style accuracy, while the modification of content components determines semantic preservation. These two components naturally exist in most of the sentences within the TST task and suggest that we can consider this task from a disentanglement perspective.

Inspired by the disentanglement concept, we create a special CoT prompting procedure for data synthesis. Initially, we prompt the LLM to explicitly identify the attribute components in the sentences. Subsequently, we query the LLM to transfer the original style by modifying these attribute components while keeping the rest unchanged. Therefore, the attribute components are accurately disentangled and original semantics are preserved to the greatest extent. Moreover, we design an error detection module to filter out synthetic sentences with incorrect style and inadequate semantic preservation, which ultimately minimizes the propagation of generated errors in LLMs. In general, this data synthesis paradigm synthesizes a certain amount of parallel data and corresponding attribute components with satisfactory quality.

With such synthetic data as supervised information, we develop a disentanglement learning method to address the TST task from two perspectives. Concretely, we design a meaningful contrastive loss, where the attribute components are regarded as positive examples and other style ones as negative examples for the original sentences. This loss pulls the representations of attribute components closer to the representations of the original sentences, enhancing the model's focus on style properties and ultimately improving the model's style control capability. In addition, we also employ a sequence-to-sequence loss with parallel data to constrain the semantic space of generated sentences, thereby enhancing the model's semantic preservation capability. By incorporating these two losses, we can achieve both substantial style control and semantic preservation, which demonstrate significant performance in the TST tasks. In addition, we also demonstrated in our experiments that this learning method allowed us to achieve better results than directly using LLM for the TST task.

In summary, our contributions are as follows:

- We are **the first** to create a CoT-based data synthesis paradigm in a disentangled manner for TST tasks. Combined with the error detection module, this paradigm synthesizes scarce parallel data and novelly alleviates the issue of limited supervised information in this field.
- We propose a disentanglement learning method with synthetic data, where we design two losses to achieve the TST task from two perspectives respectively.
- We validate the effectiveness of our framework on two widely used datasets and our approach outperforms multiple strong baselines. Besides, due to the synthetic parallel data for supervised information, we enhance sample efficiency in this field and require only about $1/10$ training samples compared to other methods.[1]

2 Approach

We propose a disentanglement-based framework for the TST task **(DisenTrans)**, where the disentanglement concept refers to separating the attribute and content components within a sentence and addressing this task from these two perspectives. In this framework, we primarily create a **disentangled data synthesis** paradigm to synthesize parallel data and corresponding attribute components. Subsequently, we construct a **disentanglement learning** method to learn from these synthetic data and achieve significant TST performance. We will formulate the TST task first, and then further elaborate on DisenTrans through the data synthesis paradigm and disentanglement learning method.

2.1 Problem Formulation

Considering a training corpus $\mathcal{D} = \left\{(x_i, s_i)_{i=1}^{T}\right\}$, x_i is an input sentence and s_i is its style label. The primary objective of TST is to acquire a model $\hat{x} = f_\theta(x, \hat{s})$, which takes an arbitrary natural language sentence x along with a desired style $\hat{s} \in \{s_j\}_{j=1}^{H}$ as input, and then generates a new sentence $\hat{x}$ that adheres to the desired style $\hat{s}$ while maintaining the semantic information of the original input sentence x. In this paper, we focus on the popular sentiment transfer task in TST to illustrate our approach, and it is also applicable to other transfer tasks.

[1]Please email Jingxuan Han with your affiliation and a short description of how you will use our synthetic parallel data, and we will then provide access to it.

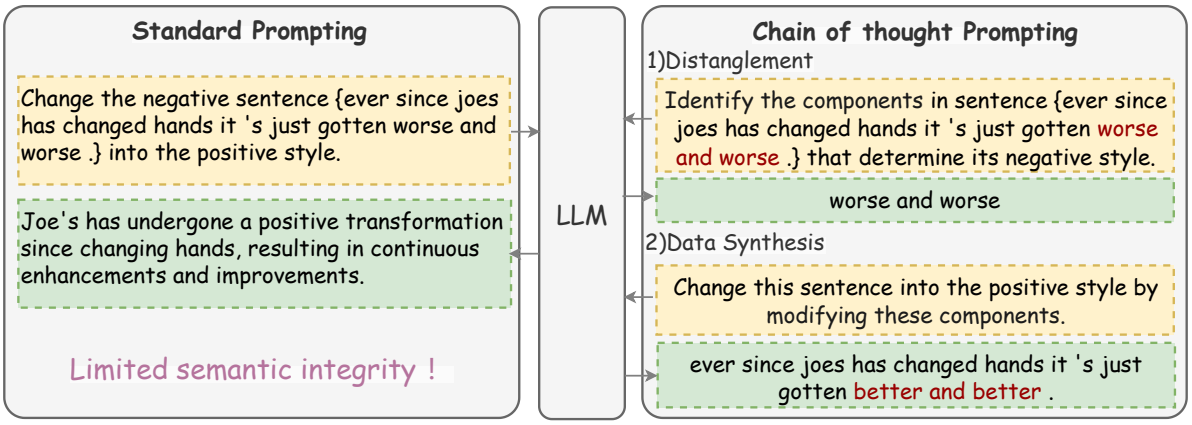

Figure 2: An example of data synthesis using ChatGPT with different prompting methods. The yellow box represents the input prompt for guiding the LLM, while the green box depicts the synthetic data generated by the LLM.

2.2 Disentangled Data Synthesis

We propose a disentangled data synthesis paradigm to obtain parallel data and corresponding attribute components, which are used as supervised information to constrain the outputs' semantic space and enhance the model's focus on attribute properties respectively. Specifically, we utilize the CoT prompting procedure to instruct the LLM to transfer the sentence to other styles while preserving its original semantics. Subsequently, we design an error detection module to filter the synthetic sentences and ensure their quality.

Chain-of-Thought Prompting When solving a complicated problem, it is typical to decompose the problem into intermediate steps and solve each before giving final answers (Wei et al., 2022). Inspired by this, we create a disentangled CoT prompting procedure for data synthesis, where the LLM is instructed to first disentangle the attribute components and then modify the style by revising these attribute components.

Figure 2 illustrates an example where ChatGPT performs a data synthesis procedure using two different prompting procedures, **standard prompting** and our **CoT prompting** procedures. We assume the original sentence $x =$ *"ever since joes has changed hands it's just gotten worse and worse"*. Under the standard prompting condition, we directly prompt the LLM for style transfer but obtain extremely limited semantic integrity. By contrast, under the CoT prompting condition, we primarily prompt the LLM to disentangle the attribute component $x^a =$ *"worse and worse"* and then transfer the style by revising x^a. Obviously, when employing the CoT prompting procedure to instruct LLM in a disentangled manner, we achieve significant style transformation and semantic preservation, thereby obtaining satisfactory synthetic data.

Error Detection Module There are a few errors inevitably during the CoT prompting process. Therefore, we designed an error detection module to ensure the precise style and high semantic preservation of synthetic data.

To ensure the style accuracy of synthetic data, we employ two open-source classifiers[2] on Hugging Face to filter out sentences with inaccurate styles. With BERT as the architecture, these two classifiers are fine-tuned on Yelp and IMDb datasets respectively. The reported accuracy is 97.0% and 91.0%. To achieve high semantic preservation, we measure the degree of semantic preservation with the BLEU score (Papineni et al., 2002) and utilize it to avoid synthetic sentences with low semantic preservation. Assuming that we aim to sample Y sentences and L_{max} is the max length, we divide the synthetic sentences into L_{max} groups based on their lengths. The l-th group contains k_l sentences, all of which have a length of l. We sample the top $Y \times \frac{k_l}{\sum_{l=1}^{L_{max}} k_l}$ sentences that have the highest BLEU scores for the l-th group. This sampling approach ensures that the model can learn transfer patterns of high semantic preservation across various sentence lengths.

By employing such data synthesis paradigm on the original dataset $\mathcal{D}$, we obtain a set of sentence pairs $\mathcal{W} = \left\{(x_i, x_i^a, \hat{x}_i, \hat{x}_i^a)_{i=1}^{Y}\right\}$ for training process, where x_i^a, $\hat{x}_i^a$ represent the attribute components in sentence x_i, $\hat{x}_i$ respectively. Y is the num-

[2]https://huggingface.co/textattack/bert-base-uncased-yelp-polarity, https://huggingface.co/JiaqiLee/imdb-finetuned-bert-base-uncased

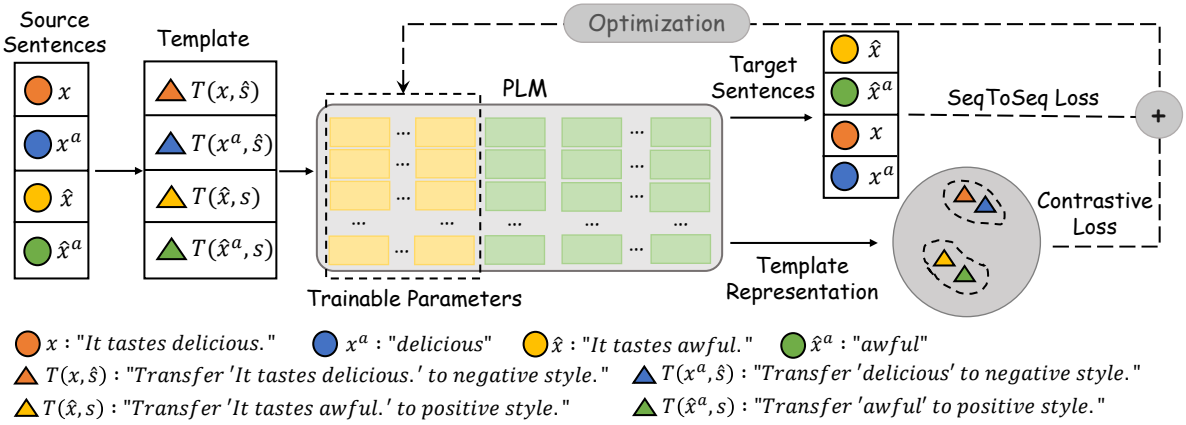

Figure 3: Disentanglement Learning Method.

ber of synthetic sentence pairs.

2.3 Disentanglement Learning

Based on synthetic data, we construct a disentanglement learning method to address the TST task from two distinct perspectives. The overall architecture is shown in Figure 3. For a sentence pair $(x, x^a, \hat{x}, \hat{x}^a)$, we enclose each sentence x into a transfer template T. Here, $T(x, \hat{s})$ denotes the template used to transfer the sentence x into the style $\hat{s}$ and $\hat{x}$ denotes its corresponding target sentence. In general, our goal is to train a model that takes a template as input and generates its corresponding target sentence effectively.

To save computational resources, we introduce a lightweight prefix-tuning network (Li and Liang, 2021), which tunes only the continuous prefix embedding with the parameters of pre-trained language models (PLMs) frozen. We prepend prefix tokens as "virtual tokens" before inputs and encode them into prefix embeddings with a trainable matrix P_θ (parameterized by θ) of dimension $L_{prefix} \times d_{model}$ in each layer, where L_{prefix}, d_{model} are the length of prefix tokens and PLM hidden size. Prefix parameters θ are the only trainable parameters. As follows, we design two losses to optimize the trainable prefix parameters.

Contrastive Loss We introduce the supervised contrastive loss (Khosla et al., 2020) to improve the model's focus on attribute properties, which finally benefits the style control. The sentence template and its corresponding attribute components are regarded as positive pairs, by which their representations will lie close together to improve the model's focus on attribute properties. The sentence template and other style templates serve as negative pairs, so their representations will lie far apart to make the template representations more distinguishable between different styles. Concretely, considering a template $T(x, \hat{s})$ in Figure 3, $P = \{T(x^a, \hat{s})\}$ is the positive set and $N = \{T(\hat{x}, s), T(\hat{x}^a, s)\}$ is the negative set. We prepend a special token *[StyleToken]* to each input template and take the hidden state of this token as the template representation z. The contrastive loss of template $T(x_i, \hat{s})$ can be defined as Eq.1, where z_i is the representation of template $T(x_i, \hat{s})$ and τ is the temperature.

$$\mathcal{L}_{con} = \frac{-1}{|P|} \sum_{p \in P} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in (N+P)} \exp\left(z_i \cdot z_a / \tau\right)} \tag{1}$$

SeqToSeq Loss We also utilize the loss derived from the sequence-to-sequence task as our loss function. We could constrain the output semantic space and take advantage of the abundant internal knowledge in PLM using this loss, which ultimately benefits the semantic preservation and human-likeness of the generated sentences. For the input template T and its corresponding output target sentence x, we perform gradient updates on the log-likelihood objective as Eq.2:

$$\mathcal{L}_{s2s} = \frac{1}{L} \sum_{t=1}^{L} -\log p\left(x_t \mid x_{<t}, T\right), \tag{2}$$

where L is the length of target sentence x.

Total Loss We optimize the prefix parameters with $\mathcal{L}_{s2s}$ and $\mathcal{L}_{con}$. The total loss of training steps is defined as Eq.3 and λ is a balancing parameter to balance $\mathcal{L}_{s2s}$ with $\mathcal{L}_{con}$.

$$\mathcal{L}_{total} = \mathcal{L}_{s2s} + \lambda \mathcal{L}_{con} \tag{3}$$

Dataset	Style	Train	Dev	Test	Vocab	Avg.Len
Yelp	Pos	270K	2K	0.5K	10K	8.9
	Neg	180K	2K	0.5K		
IMDb	Pos	180K	2K	1K	30K	18.5
	Neg	190K	2K	1K		

Table 1: Statistics of the Yelp and IMDb dataset.

3 Experimental Settings

3.1 Datasets

We employ the two most widely used datasets in the current research: the Yelp Review Dataset and the IMDb Movie Review Dataset, whose statistics of the two datasets are presented in Table 1.

Yelp Review Dataset (Yelp) The Yelp [3] dataset contains reviews of restaurants and businesses, each labeled to indicate positive or negative sentiment. In addition, it also includes human-annotated sentences for evaluating semantic preservation.

IMDb Movie Review Dataset (IMDb) The IMDb dataset, provided by the Style Transformer (Dai et al., 2019)[4], comprises movie reviews penned by online users, each tagged with either a positive or negative sentiment label. However, it lacks annotations by human evaluators.

3.2 Automatic Evaluation

We assess the effectiveness of our method based on two fundamental abilities: style control and semantic preservation, which are widely recognized as the most crucial factors in evaluating style transfer models (Xiao et al., 2021; Luo et al., 2021).[5]

Style Control Style accuracy (**S-ACC**) is a metric to assess the ability of style control. We trained two sentiment classifiers on the training set of Yelp and IMDb to evaluate all methods. To ensure a fair comparison, we utilize T5-base as the classifier architecture which is different from the error detection module. The classification performance is 98.0% and 93.3% respectively.

Semantic Preservation The Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002) serves as a metric to quantitatively assess the lexical-level similarity between two sentences, by which we can evaluate the degree of semantic preservation. Two BLEU scores are calculated by the Natural Language Toolkit (Bird et al., 2009) in our work. The **self-BLEU** score measures the similarity between output and input, while the **ref-BLEU** score measures the similarity between output and human reference.

G-score The G-score, calculated as the geometric mean of self-BLEU and S-ACC, is a holistic and comprehensive metric to evaluate the effectiveness of both style control and semantic preservation. This composite metric considers the delicate balance between maintaining the intended style and preserving the semantic information.

3.3 Human Evaluation

We performed human evaluations to assess the generated sentences. Concretely, we randomly sampled 200 outputs (100 per style) from each dataset, resulting in a total of 400 outputs per model. Three annotators were tasked with evaluating the generated sentences based on style control (SC), semantic preservation (SP), and fluency (FL), assigning scores ranging from 1 (Very Bad) to 5 (Very Good). The average score across the three annotators was used as the final evaluation metric.

3.4 Implementation Details

Data Synthesis Implementation We use ChatGPT[6] for data synthesis. For each dataset, we randomly select 100K original sentences to generate parallel synthetic sentences. Besides, we employ in-context learning to enhance synthetic quality, where we incorporate one $pos \rightarrow neg$ and one $neg \rightarrow pos$ specific example in the CoT prompt. The detailed CoT prompt is provided in Appendix A.1. When synthesizing 100k initial parallel sentences, we utilize the error detection module to sample high-quality synthetic sentences. For the Yelp and IMDb datasets, our goal is to sample **45K** and **40K** sentences respectively, while other baselines require **450K** and **370K** training samples.

Disentanglement Learning Implementation We have a small number of hyperparameters, the temperature τ and balancing parameter λ are 0.8 and 0.6 respectively. We select T5 as PLM to employ prefix tuning and L_{prefix} is 12. Our approach is trained on one NVIDIA A800 GPU.

[3] https://www.yelp.com/dataset

[4] https://github.com/fastnlp/style-transformer

[5] We don't adopt PPL like many studies in TST, which cannot evaluate the text fairly for the following reasons (Wang et al., 2022): (i) The PPL of short text is larger than long text, which goes against common sense. (ii) The repeated text span could damage the performance of PPL. (iii) The punctuation marks could affect the performance of PPL heavily.

[6] https://openai.com/blog/chatgpt, version: gpt-3.5-turbo.

Methods	Params	Yelp				IMDb		
		S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑	S-ACC↑	self-BLEU↑	G-score↑
Input Copy	–	1.4	21.8	100.0	11.8	2.8	100.0	16.7
StyleTrans (Dai et al., 2019)	1.8×10^7	90.0	17.3	46.0	63.6	75.0	66.9	70.8
DGST (Li et al., 2020)	3.2×10^7	88.0	18.7	54.5	69.3	70.1	**70.2**	70.1
DIRR (Liu et al., 2021c)	1.5×10^9	92.8	20.8	52.3	69.7	86.9	65.1	75.2
RACoLN (Lee et al., 2021)	1.0×10^7	86.9	20.0	56.3	69.9	80.8	68.2	74.2
LEWIS (Reid and Zhong, 2021)	1.1×10^8	86.3	19.4	53.0	67.6	N/A	N/A	N/A
ComPose (Liu et al., 2022)	7.7×10^8	79.4	18.2	51.6	64.0	N/A	N/A	N/A
CRF (Shuo, 2022)	1.2×10^8	86.7	20.2	53.5	68.1	83.0	59.0	70.0
DisenTrans (Ours)	$\mathbf{7.1 \times 10^6}$	**93.2**	**21.5**	**58.7**	**74.0**	**88.4**	67.0	**77.0**

Table 2: Automatic evaluation results. The ref-BLEU for the IMDb is not reported due to the absence of human references. Input Copy means an unmodified copy of the source sentence.

4 Results and Analysis

4.1 Baseline Description

We select several strongest TST methods including some SOTA methods like StyleTrans (Dai et al., 2019), RACoLN (Lee et al., 2021), and DIRR (Liu et al., 2021c) for comparison. (1) **StyleTrans** (Dai et al., 2019): a typical method that employs an additional style embedding and designs three losses to construct supervision. (2) **DGST** (Li et al., 2020): a novel dual-generator network that does not rely on any discriminators. (3) **DIRR** (Liu et al., 2021c): an RL-based approach to improving content preservation by leveraging a semantic similarity metric as the content reward. (4) **RACoLN** (Lee et al., 2021): a method that utilizes reverse attention to implicitly remove style tokens and fuse content information to style representation using conditional layer normalization. (5) **LEWIS** (Reid and Zhong, 2021): a coarse-to-fine editor that transforms text using Levenshtein edit operations. (6) **ComPose** (Liu et al., 2022): an efficient approach for composable text operations in the compact latent space of text. (7) **CRF** (Shuo, 2022): a probabilistic model that generates a tag sequence to modify the input sentences using programming search algorithms.

4.2 Automatic Evaluation Result

The automatic evaluation results are presented in Table 2. DisenTrans achieves competitive overall performance compared to these baseline methods with a limited number of training samples, as indicated by the G-score metric (**+4.1 for Yelp, +1.8 for IMDb**). Moreover, DisenTrans has a smaller amount of trainable parameters and does not introduce a more complex training process.

Concretely, DisenTrans demonstrates significant ability in style control, evident from its higher S-ACC metrics on both Yelp and IMDb datasets. Although DIRR achieved a comparable accuracy, its trainable parameters are about 1000x larger than ours and it cannot simultaneously do well in BLEU metrics. Furthermore, DisenTrans excels in semantic preservation and produces sentences with more human-like characteristics, as evidenced by its superior self-BLEU and ref-BLEU metrics, respectively. Notably, a low self-BLEU score may not directly reflect semantic preservation in the IMDb dataset, because sentences in the IMDb dataset are relatively long (Shuo, 2022). Besides, DisenTrans also exhibits good stability. We conducted multiple repetitions of experiments using different seeds and observed relatively low variance, which is presented in Appendix A.2.

Methods	Yelp			IMDb		
	SC	SP	FL	SC	SP	FL
StyleTrans	4.2	3.9	4.0	3.8	3.3	3.7
DGST	4.0	4.1	3.9	3.2	3.8	4.0
DIRR	4.4	4.5	4.0	3.8	3.6	**4.1**
RACoLN	4.1	4.6	4.2	3.7	**3.9**	4.0
LEWIS	4.0	3.9	3.8	N/A	N/A	N/A
ComPose	3.8	3.7	4.0	N/A	N/A	N/A
CRF	4.2	4.5	4.3	3.8	3.1	3.9
DisenTrans	**4.5**	**4.7**	**4.3**	**3.9**	3.8	**4.1**

Table 3: Human Evaluation. Each score represents the average score of three annotators, where SC, SP, FL represent style control, semantic preservation and fluency respectively.

4.3 Human Evaluation Result

We incorporate human evaluation for a detailed and reliable assessment of sentence quality. The results presented in Table 3 are generally consistent with the automatic evaluation, indicating that our model outperforms other methods in terms of style

Methods	S-ACC	ref-BLEU	self-BLEU	G-score
DisenTrans	93.2	21.5	58.7	74.0
(-)EDM	87.6	19.8	51.3	67.0
(-)ConLoss	90.3	21.1	60.5	73.9

Table 4: Ablation study of the error detection module and contrastive loss.

Methods	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
DisenTrans (two-stage data filtering)	93.2	21.5	58.7	74.0
DisenTrans (combined data filtering)	91.8	21.7	59.6	74.2

Table 5: Ablation study of the data filtering paradigm.

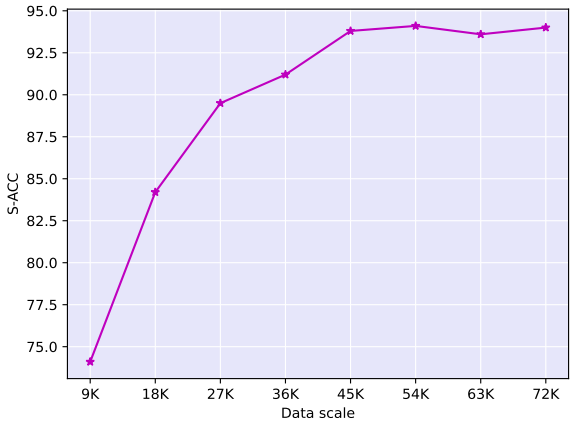

Figure 4: S-ACC performance with different data scales.

Methods	S-ACC	ref-BLEU	self-BLEU	G-score
DisenTrans (Ours, 45K)	93.2	21.5	58.7	74.0
StyleTrans (450K)	90.0	17.3	46.0	63.6
StyleTrans (45K)	82.7	9.1	39.0	56.8
RACoLN (450K)	86.9	20.0	56.3	69.9
RACoLN (45K)	81.4	19.8	55.9	67.5

Table 6: The comparison with two typical baselines under the same data scale.

control and particularly in semantic preservation. Besides, the fluency score proves our ability to generate fluent outputs. The human evaluation for data synthesis and the Inter Annotator Agreement are reported in Appendix A.3.

4.4 Ablation Study

To validate the effects of individual components in DisenTrans, we conduct comprehensive ablation studies on the Yelp dataset and these results remain consistent when applied to the IMDb.

Ablating Error Detection Module To assess the impact of the Error Detection Module (EDM), we conduct training without its incorporation. Concretely, we utilize the same quantity of data directly generated by the LLM for the experiment and the results are shown in Table 4. Obviously, we would obtain suboptimal performance without the EDM to enhance the quality of synthetic data, which suggests the importance of the EDM.

Ablating Data Filtering Paradigm In our two-stage data filtering paradigm, we first filter the original synthetic sentences based on accuracy to obtain first-stage sentences with high classification accuracy. Then, we further filter first-stage sentences based on semantic similarity to obtain two-stage sentences with both high classification accuracy and semantic similarity. For further exploration, we employ a unified metric for data filtering and develop a combined data filtering paradigm for comparison. Specifically, we employ the G-score as the unified metric and select top-performing sentences across different length groups. The results in Table 5 show that the combined data filtering can also achieve comparable performance.

Ablating Contrastive Loss To explore the impact of contrastive loss, we optimize our network with only SeqToSeq loss while maintaining consistent hyperparameters. The results in Table 4 demonstrate that the contrastive loss effectively enhances the model's capability to style control. Moreover, it also illustrates the substantial influence of the SeqToSeq loss on semantic preservation.

Ablating Data scale We chose training data of different scales to investigate the impact of our sample efficiency. In more detail, we conduct experiments with different data scales while maintaining consistent hyperparameters. Figure 4 illustrates the variation curve of the S-ACC metric, which indicates that as the data scale reaches a certain point, the model's performance converges. Therefore, there's no requirement to use a larger data scale, which would sacrifice more computational resources to achieve less improvement.

We also retrained two typical baselines with 45K training samples on the Yelp dataset for a detailed comparison. The results in Table 6 distinctly illustrate a performance decline in these baselines when restricted to just 45K training samples. DisenTrans demonstrates significant performance using the same number of samples (45K). Therefore, while baselines require up to 450K training samples for effective performance, DisenTrans requires only up to 45K.

LLM vs DisenTrans We directly evaluate the performance of ChatGPT on the Yelp test set with

Methods	Negative → Positive	Positive → Negative
Input 1	the food is worse than you find in the freezer section at walmart .	suzanne and her staff were excellent !
Human	the food is way better than you find in the freezer section at walmart .	suzanne and her staff were horrible .
Ours	the food is better than you find in the freezer section at walmart .	suzanne and her staff were horrible .
RACoLN	the food is great than you find in the world section at walmart .	take-out and her staff were awful !
Lewis	the food here is even better than you find in the freezer section at walmart .	suzanne and her staff were soooo rude !
ComPose	the food is amazing as you find in the department at walnut grill .	suzanne and her staff were incorrect !
CRF	the food is than you find in the freezer section at eye !	suzanne and her staff were rude !
Input 2	i may just post pictures to prove their shoddy work .	i can honestly say i am so glad we will be moving to az .
Human	i may just post pictures to prove their high quality work .	i can honestly say i am very unhappy we will be moving to az !
Ours	i may just post pictures to prove their great work .	i can honestly say i am so disappointed we will be moving to az .
RACoLN	i absolutely just post pictures to prove their discounted work .	i can honestly say i am so glad we will be moving to az .
Lewis	i may just post pictures to prove their quality of work .	i can not say i am so sure we will be back to az .
ComPose	i will post some fabulous work pictures to their craftsmanship .	i can honestly say i am so glad we will not be moving to this place .
CRF	i may just pictures to prove their quality work .	i can only say i am so glad we will be going to az .
Input 3	as soon as they delivered i was like ugh .	love my cut and color and sage is amazing !
Human	as soon as they delivered i was in awe .	hate my cut and color and sage is awful !
Ours	as soon as they delivered i was like delightful .	hate my cut and color and sage is horrible !
RACoLN	as soon as they delivered i was like heaven .	avoid my cut and color and meat is gross !
Lewis	as soon as they delivered i was like u rock !	i hate my cut and color and sage is horrible !
ComPose	as soon as they delivered i was like wow .	asked my cut and color and mine broke apart .
CRF	as soon as they delivered i was like ugh happy .	my cut and color and sage is terrible !

Table 7: Case Study from the Yelp dataset. The red portion indicates the modified portions in comparison to the inputs. We have selected four of the most recent studies, with "Human" representing the human reference sentence.

Methods	S-ACC	ref-BLEU	self-BLEU	G-score
DisenTrans	93.2	21.5	58.7	74.0
Standard Prompting	84.2	14.3	25.1	46.0
CoT Prompting	86.0	18.7	49.1	65.0

Table 8: Ablation study of the disentanglement learning. Standard prompting and CoT Prompting represent directly querying ChatGPT to transfer texts without disentanglement learning.

Methods	G-score	Params	Hours/Epoch	GPU
DisenTrans(Prefix-tuning)	74.0	7.1×10^6	2	$1 \times A800$
Fine-tuning	74.4	2.9×10^9	5	$4 \times A800$

Table 9: Ablation study of the prefix tuning.

two prompting procedures. As shown in Table 8, although the CoT prompting procedure significantly enhances the semantic preservation performance, it still falls short compared with DisenTrans which benefits from our disentanglement learning method. Therefore, we require our learning method to assist our model in better performance, rather than directly using LLM to generate style transfer texts.

Prefix-tuning vs Fine-tuning We utilize a lightweight prefix-tuning network in DisenTrans. To verify the effectiveness of our prefix-tuning network, we carried out fine-tuning comparative experiments. The results are displayed in Table 9. DisenTrans employs prefix-tuning to achieve comparable performance while requiring significantly fewer parameters and shorter training duration. Consequently, we choose prefix-tuning in DisenTrans to reduce the computational resources.

4.5 Case Study

To better understand the characteristics of different methods, we sampled several output sentences from Yelp, as shown in Table 7. Compared with other methods, DisenTrans always revises the fewest words which are possibly the attribute components to transfer sentence style. Such transformation aligns more closely with human properties and results in greater semantic preservation. Although some methods may modify the same number of words in certain cases, DisenTrans exhibits a more accurate style and smoother semantic coherence. Interestingly, it even preserves better semantic information than human reference in some cases.

5 Related Work

Text Style Transfer TST task has proven to be quite challenging due to the discrete nature of language (Jin et al., 2022). Most approaches are unsupervised owing to the difficulty of obtaining parallel data (Shang et al., 2019). Previous work can mainly be categorized into two groups.

The first group attempts to identify and revise the words with stylistic attributes in the early stage of TST (Rao and Tetreault, 2018a; Li et al., 2018; Wu et al., 2019). These methods employed attention mechanisms and directly revised words with style information in sentences according to attention scores. However, most of these approaches suffer from flat attention distribution due to the limitations of the network's capacity, where similar at-

tention scores are assigned to tokens and numerous tokens without stylistic attributes would be revised. Moreover, due to the lack of a high-level understanding of semantics, the generated sentences are not sufficiently authentic and logical. (Li et al., 2018) employ the editing technique which involves token deletion through simple counts and target word retrieval via TF-IDF weighted overlap.

The second group focuses on revising an entangled representation of input (Dai et al., 2019; Huang et al., 2020; Kashyap et al., 2022; Liu et al., 2021a), which is mainstream in TST currently. These methods employ a style embedding or encoder and create many losses to construct training supervision. ARAE (Kashyap et al., 2022) proposes the incorporation of two collaborative loss functions with the adversarially regularized autoencoder framework. SA-MLM (Narasimhan et al., 2023) performs TST by using a style-masked input and performs a simple same-style reconstruction task with a Transformer Encoder block.

Large Language Model Recently, there has been a growing trend of research related to LLMs (Sun et al., 2023; Lu et al., 2023), and our primary focus is on data synthesis. For text classification tasks, (Kurakin et al., 2023) generates private synthetic data by DP-finetuned LLM to privately estimate predictive models. In the Computational Social Science field, (Veselovsky et al., 2023) study three strategies to increase the faithfulness of synthetic data generated by LLMs: grounding, filtering, and taxonomy-based generation. For Human-Computer Interaction research, (Hämäläinen et al., 2023) employ LLMs to generate open-ended questionnaire responses about experiencing video games. However, there is also no suitable CoT method available to assist LLMs in synthesizing the supervised data in the TST field.

6 Conclusion

In this work, we propose a novel disentanglement-based framework for text style transfer (DisenTrans), which is instructed by a disentanglement concept. Concretely, we first create a CoT prompting procedure to synthesize parallel data and corresponding disentangled information for supervision. Then we develop a disentanglement learning method with such synthetic data, where contrastive loss and sequence-to-sequence loss are designed to improve the focus on attribute properties and constrain the semantic space, thereby benefiting the TST tasks from style control and semantic preservation respectively. Experiments on two widely used datasets demonstrate that DisenTrans achieves superior performance with great sample efficiency.

Limitations

Since there is a lack of multi-attribute datasets in the existing literature, the efficiency of our approach to multi-attribute TST tasks has not been validated. As part of our future work, we intend to apply our data synthesis paradigm to create datasets for multi-attribute style transfer using large language models. This will help us confirm the capability of our method in tackling multi-attribute transfer tasks.

Ethics Statement

The text style transfer task finds extensive application in the domain of controlled text generation. However, due to the various corpus of different styles, the development in the field of text style transfer may bring about potential risks. The style transfer model can be misused to generate false information, which could result in unreliable social media, news and other online platforms. Moreover, misuse of style transfer applications could raise concerns related to privacy and copyright, such as unauthorized content transformation and abuse of others' work. Therefore, with the development of the text style transfer, it is essential to remain vigilant and ensure that its applications are ethical and lawful, which ultimately mitigates potential adverse effects. In our data synthesis paradigm, ChatGPT synthetic data might contain toxic contents. To avoid toxic content, we intend to introduce a detoxification model to our error detection module.

Acknowledgements

We would like to express our sincere gratitude to the professional reviewers for their suggestions and comments. This work is supported by the National Science Fund for Excellent Young Scholars under Grant 62222212, the National Natural Science Foundation of China under Grant 62376033, and the University Synergy Innovation Program of Anhui Province GXXT-2022-037. We thank Sudha Rao and Joel Tetreault for providing the GYAFC dataset to our third author, Hal Bush (Zikang Guo).

References

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *International Conference on Applications of Natural Language to Information Systems*, pages 47–61. Springer.

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 Xuan-Jing 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.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2213–2223.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann, and Soujanya Poria. 2022. So different yet so alike! constrained unsupervised text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–431.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

Dongkyu Lee, Zhiliang 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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

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

X Li, G Chen, C Lin, and R Li. 2020. Dgst: a dual-generator network for text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7131–7136. Association for Computational Linguistics (ACL).

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597.

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2022. Composable text controls in latent space with odes. *arXiv preprint arXiv:2208.00638*.

Ruibo Liu, Chongyang Gao, Chenyan Jia, Guangxuan Xu, and Soroush Vosoughi. 2021a. Non-parallel text style transfer with self-parallel supervision. In *International Conference on Learning Representations*.

Ruibo Liu, Chongyang Gao, Chenyan Jia, Guangxuan Xu, and Soroush Vosoughi. 2021b. Non-parallel text style transfer with self-parallel supervision. In *International Conference on Learning Representations*.

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

pages 4262–4273, Online. Association for Computational Linguistics.

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints. *arXiv preprint arXiv:2302.09185*.

Ruikun Luo, Guanhuan Huang, and Xiaojun Quan. 2021. Bi-granularity contrastive learning for post-training in few-shot scene. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1733–1742.

Yun Ma and Qing Li. 2021. Exploring non-autoregressive text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9267–9278.

Sharan Narasimhan, Pooja H, Suvodip Dey, and Maunendra Sankar Desarkar. 2023. On text style transfer via style-aware masked language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 362–374, Prague, Czechia. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

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

Sudha Rao and Joel Tetreault. 2018a. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Sudha Rao and Joel Tetreault. 2018b. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944.

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 *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 4937–4946.

Yang Shuo. 2022. Tagging without rewriting: A probabilistic model for unpaired sentiment and style transfer. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 293–303.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. *arXiv preprint arXiv:2310.14542*.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2022. Perplexity from plm is unreliable for evaluating text quality. *arXiv preprint arXiv:2210.05892*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive learning for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2510–2521.

Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Z. Xu, T. Chalasani, K. Ghosal, S. Lutz, and A. Smolic. 2019. Stada: Style transfer as data augmentation. pages arXiv–1909.

A Appendix

A.1 Implementation Appendix

Data Synthesis Implementation We employ the ChatGPT API (version: gpt-3.5-turbo) to synthesize the parallel sentences along with their corresponding components, incurring a cost of approximately 30$. Our implementation of this API was facilitated through OpenAI's Python library. The detailed prompt is *"Identify the components in sentence {Original Sentence} that determine its {Original Style} style. Then, change this sentence into the {Oppisite Style} style by modifying these components. For instance, if the positive sentence is {along with cops , the goat is one of keaton 's two funniest shorts .}, the style component would be {funniest} and the negative revision would be {along with cops , the goat is one of keaton 's two worst shorts .}. Similarly, if the negative sentence is {worst oysters i had so far .}, the style component would be {worst} and the positive revision would be {best oysters i had so far .}."*

Disentanglement Learning Implementation T5 adopts an encoder-decoder framework, to which we introduced prefix embeddings in every layer of both the encoder and decoder. The contrastive loss operates on the encoder's output, exclusively optimizing the prefix parameters of the encoder. In contrast, the sequence-to-sequence loss is based on the decoder's output, leading to the optimization of the prefix parameters for both the encoder and decoder. For supervised contrastive learning, we label each sentence, ensuring that sentences and their corresponding attribute components share the same label. Take an instance, for the sentence pairs $\mathcal{W} = \{(x_1, x_1^a, \hat{x}_1, \hat{x}_1^a), (x_2, x_2^a, \hat{x}_2, \hat{x}_2^a)\}$ in a batch where size is 2, the assigned label $\mathcal{L} = \{(0, 0, 1, 1), (2, 2, 3, 3)\}$. Such sentence pairs and corresponding labels will be used to calculate the supervised contrastive loss.

Human Evaluation Implementation We hired three human annotators to evaluate our method along with other recent methods. Here we provide details about the annotators who were tasked with assigning scores to the generated sentences. The average sentence length in the IMDb dataset is twice that of the Yelp dataset, making the evaluation of the IMDb dataset more challenging. Consequently, considering the different complexities between the two datasets, annotators will receive a payment of 0.05$ for each sentence in the Yelp dataset and 0.1$ for each sentence in the IMDb dataset. We randomly selected 200 outputs (100 per style) from each dataset, resulting in a total of 1.8K sentences from the Yelp dataset and 1.4K sentences from the IMDB dataset. This process incurred a cost of 230$ for each annotator.

A.2 Experiment Appendix

The performance on formality style transfer We conducted experiments on the formality style transfer task using the GYAFC dataset (Rao and Tetreault, 2018b), which is created from the Yahoo Answers L6 corpus. This corpus includes approximately 5.2k parallel training data and 1.0k test data in the Family & Relationships domain. We developed a formality classifier on the training set to calculate the accuracy metric. Based on the T5-base architecture, this classifier achieved an accuracy of 92.0%. Due to the availability of parallel data, there was no need for synthesizing additional parallel data. Instead, our focus was on effectively disentangling these parallel sentences for disentangled learning. The results are presented in Table 10, which indicates that our approach achieved significant performance in this specific task. Additionally, Figure 5 presents two disentangled instances from the GYAFC dataset, demonstrating our method's ability to disentangle attribute components in the sentences of this task.

Methods	BLEU↑	S-ACC↑	G-score↑
Semi-Supervised (Shang et al., 2019)	81.4	85.1	83.2
LEWIT (Babakov et al., 2023)	75.7	84.2	79.8
DisenTrans	80.2	87.3	83.7

Table 10: The performance on formality style transfer task.

Methods	ref-BLEU↑	S-ACC↑
Probabilistic (He et al., 2019)	10.8	81.4
P-R (Suzgun et al., 2022)	21.9	78.0
DisenTrans	20.3	86.4

Table 11: The performance on personal writing style transfer task.

The performance on personal writing style transfer In complex style transfer tasks, style and content cannot be completely disentangled within a small number of sentences. To explore the effectiveness of our method in these tasks, we utilized the Shakespeare personal writing style dataset (Xu et al., 2012) for experiments. This transfer task

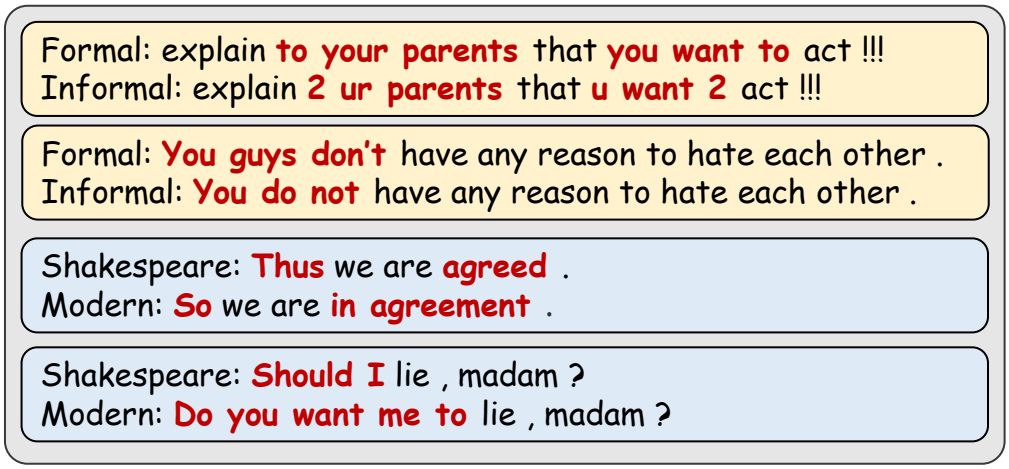

Figure 5: Disentangled instances on two transfer tasks. The yellow box represents the formality transfer task, while the blue box denotes the personal writing style transfer task. The red part is the attribute component in the sentence.

is one of the most complex transfer tasks in TST, even for human manual operations. The Shakespeare dataset contains 18k parallel training data and 1.4k test data. We trained a Shakespeare classifier on the training set to calculate the accuracy metric. With T5-base as the architecture, the classifier achieves an accuracy of 87.0%. The results in Table 11 demonstrate that DisenTrans sill achieves comparable performance in this complex task, which implies that our approach does not strictly require complete disentanglement for every sentence. Although a small number of sentences cannot be disentangled, our method is effective for most sentences, which is beneficial for this task. Additionally, we provide two disentangled instances in Figure 5. The instances showcase our method's ability to disentangle attribute words in the sentences of this complex task.

The performance of Open-source LLMs We selected Llama-2-70b[7] as the open-source LLM for data synthesis and replicated the experiments with consistent settings. The results in Table 12 demonstrate that leveraging an open-resource LLM can also produce comparable performance. Therefore, ChatGPT serves merely as one of the tools for data synthesis in our research and we do not depend on it heavily. We prefer to use ChatGPT for data synthesis primarily because open-source LLMs require extensive GPU resources (8*A100).

The performance of LLM+CoT+Error Detection Module We conducted experiments to investigate the effectiveness of the LLM+CoT+Error Detection Module (EDM). Specifically, we used EDM to filter the test results of LLM+CoT and considered the correctly filtered results as a test plus set (about 860 test samples). Subsequently, we evaluated the LLM+CoT and our model on this set. The results in Table 13 show that the LLM+CoT+EDM achieves sub-optimal performance.

[7]https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

Methods	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
DisenTrans (with ChatGPT)	93.2	21.5	58.7	74.0
DisenTrans (with Llama2-70b)	92.8	21.2	56.2	72.2

Table 12: The performance of Open-source LLMs.

Hyperparameters	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
LLM+CoT+EDM	89.6	20.9	46.4	64.5
DisenTrans	95.6	22.3	58.5	74.8

Table 13: The performance of LLM+CoT+EDM.

The comparison with other data augmentation methods In the TST field, to the best of our knowledge, there is really limited work corresponding to data augmentation. To further evaluate the effectiveness of our approach, we chose LaMer (Liu et al., 2021b) for comparison, which generates parallel sentences from each style for data augmentation based on scene graphs and large-scale LMs. The results are presented in Table 14. DisenTrans achieves better performance than LaMer, which suggests that our method achieves more effective data augmentation.

Ablating Sampling Method We conducted the ablation study to quantify the importance of sam-

Methods	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
LaMer (Liu et al., 2021b)	90.1	-	40.6	60.5
DisenTrans	93.2	21.5	58.7	74.0

Table 14: The comparison with other data augmentation methods.

pling high BLEU sentences based on sentence length intervals. Specifically, we sampled synthetic data directly in descending BLEU score order, while keeping other settings unchanged for the experiment. The results in Table 15 revealed a notable decline in the S-ACC metric. Furthermore, upon observing the sampled data, we noticed that there is a lack of short sentences in our training samples, which we think is a contributing factor to the decreased performance. Therefore, it is important to sample based on sentence length intervals to ensure that sentences of all lengths are included in the sampling process.

Hyperparameters	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
(-)Length Interval	82.0	20.9	61.1	70.8
DisenTrans	93.2	21.5	58.7	74.0

Table 15: Ablation study of the sampling method.

Explore Training Stability To explore the training stability of DisenTrans, we select ten seeds for multiple experiments under the same hyperparameters and calculate the variance of each metric. The results are presented in Table 16, which depicts that our method has achieved impressive performance with strong stability.

Methods	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
StyleTrans	$90.0_{1.2}$	$17.3_{0.4}$	$46.0_{1.7}$	$63.6_{1.6}$
RACoLN	$86.9_{1.4}$	$20.0_{0.8}$	$56.3_{1.5}$	$69.9_{1.5}$
DIRR	$92.8_{1.8}$	$20.8_{1.9}$	$52.3_{1.4}$	$69.7_{1.6}$
DisenTrans	$93.2_{1.6}$	$21.5_{0.3}$	$58.7_{0.9}$	$74.0_{1.2}$

Table 16: Exploration of the training stability on the Yelp dataset. The number in the lower right corner represents the variance.

Ablating Prefix Length To investigate the impact of prefix lengths, we conduct ablation studies with different prefix lengths. The results are presented in Table 17. It can be observed that as the prefix length reaches a certain value, the model converges, and there is no further improvement in performance with further increases in the prefix length. A longer prefix leads to more trainable parameters, but it also has a negligible impact on inference speed since the attention computation across the entire prefix is parallelized on GPUs.

Prefix Length	G-score	Params
$L_{prefix} = 6$	73.3	3.5×10^6
$L_{prefix} = 12$	74.0	7.1×10^6
$L_{prefix} = 18$	74.3	1.1×10^7
$L_{prefix} = 24$	74.4	1.4×10^7
$L_{prefix} = 30$	74.7	1.8×10^7
$L_{prefix} = 40$	74.6	2.4×10^7
$L_{prefix} = 50$	74.4	2.9×10^7

Table 17: Ablation study of the prefix length.

Explore Diversity We have incorporated an assessment of diversity to demonstrate that our explicit controlled modification does not lead to less diversity in the output. Concretely, we select the diversity metric Distinct (Li et al., 2016a) to quantify the diversity of baselines on the Yelp dataset. The results in Table 18 indicate that our method achieves comparable diversity with better semantic preservation and style control abilities

Method	Dist-1↑	Dist-2↑	Dist-2↑
StyleTrans	0.154	0.550	0.733
DIRR	0.141	0.492	0.667
RACoLN	0.150	0.543	0.734
LEWIS	0.136	0.480	0.667
ComPose	0.160	0.559	0.747
CRF	0.139	0.498	0.673
DisenTrans	0.157	0.551	0.737

Table 18: The performance of diversity.

Ablating Balance Parameter λ To investigate the impact of varying balance parameters, we select different λ to conduct experiments. The results in Table 19 show that our performance is insensitive to λ. We select $\lambda = 0.6$ according to the highest G-score metric.

Hyperparameters	S-ACC↑	ref-BLEU↑	self-BLEU↑	G-score↑
$\lambda = 0.4$	91.8	22.5	59.4	73.8
$\lambda = 0.6$	93.2	21.5	58.7	74.0
$\lambda = 0.8$	92.4	22.1	58.4	73.5
$\lambda = 1.0$	92.9	21.6	58.0	73.4

Table 19: Ablation study of the balance parameter.

A.3 Evaluation Appendix

Human Evaluation For Data Synthesis We employed three annotators for human evaluation to ensure the satisfactory quality of our synthesized data. For each dataset, we randomly selected 1000 synthesized sentences (500 for positive style, 500 for negative style) for quality assessment. The scores (on a scale from 0 to 5) and inter-annotator agreement (quantified by Fleiss' Kappa coefficient) are presented in Table 20. The scores demonstrate the superior quality of our synthesized data. The high consistency among human annotators indicates the acceptability of our human evaluation.

	Yelp			IMDb		
	SC	SP	FL	SC	SP	FL
Human Evaluation Score	4.9	4.8	4.6	4.7	4.9	4.7
Fleiss' Kappa coefficient	0.798	0.775	0.632	0.778	0.782	0.641

Table 20: Human evaluation for data synthesis. The SC, SP, and FL correspond to style accuracy, semantic preservation, and fluency respectively.

Inter Annotator Agreement We calculate the Fleiss' Kappa coefficient to measure the inter-annotator agreement score for each human evaluation metric. The results are shown in the Table 21. The high consistency among human annotators indicates the acceptability of our human evaluation.

Methods	Yelp			IMDb		
	SC	SP	FL	SC	SP	FL
StyleTrans	0.773	0.742	0.658	0.754	0.732	0.651
DGST	0.787	0.797	0.687	0.747	0.786	0.674
DIRR	0.792	0.775	0.667	0.779	0.752	0.653
RACoLN	0.801	0.753	0.662	0.794	0.743	0.665
LEWIS	0.751	0.732	0.668	N/A	N/A	N/A
ComPose	0.763	0.742	0.652	N/A	N/A	N/A
CRF	0.812	0.794	0.698	0.765	0.781	0.685
DisenTrans	0.805	0.788	0.705	0.785	0.775	0.692

Table 21: The inter-annotator agreement score for human evaluation. The SC, SP, and FL correspond to style accuracy, semantic preservation, and fluency respectively.