A Causal Approach to Single-Attribute Controllable Text Generation

Ruineng Li

Abstract

Generating texts with required attributes is a major problem in natural language generation. Recent years, controllable text generation is emerging as a promising fundamental task for its potential usages in many real-world applications. However, current controllable text generation algorithms are based on a *non-identifiable* formulation, such that models are not guaranteed to find the correct causal diagram. This paper aims to take a very first step to address this commonly encountered non-identifiability issue, by tackling a single-attribute text rewriting problem, text style transfer on formality. We propose a non-autoregressive language model which reformalizes the attribute-transfer query and incorporates *do-calculus* as a solution. We demonstrate successful introduction of causality in text style transfer with competitive results compared to prior work on two formality style transfer datasets for all four metrics.

1 Introduction

Controllable Text Generation (CTG) is a task aiming to generate texts that have desired attributes such as topic, sentiment (Sudhakar et al., 2019), persona (Zhang et al., 2018), style (Reif et al., 2021) and so on. A common approach to accomplish this task is to leverage supervised data (e.g. ground truth texts, attribute information) as guidance to train language models (Hu et al., 2017; Brown et al., 2020). While such approaches have shown promising results in various CTG tasks, the query is indeed *non-identifiable*: CTG concerns about learning the conditional distribution $p(y|\boldsymbol{x})$, where $\boldsymbol{x}$ is the set of desired attributes. However, consider the case where $\boldsymbol{x} = (text, attribute)$ (i.e. a text attribute transfer task), there is no guarantee to always have $p(text, attribute) \neq 0$, leading to $P(y|\boldsymbol{x})$ non-identifiable in such cases.

Models learned in this way are non-identifiable. From a causal perspective, causal diagrams cannot be uniquely identified by a non-identifiable query (Galles and Pearl, 1995; Tian and Pearl, 2002). With a non-identifiable query, different causal diagrams might produce identical outputs for some cases, but not for the others. This means even well-trained non-identifiable models can still produce incorrect outputs sometimes, because possibly they are not learning the correct causal diagram. Existing approaches implicitly solve this problem by relying on pre-trained large language models with large-scale datasets (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020). The more cases a model has seen, the closer it is to the correct causal diagram. However, it is merely an expedient, for that models are inherently impossible to have encountered all possible inputs, and that data scarcity is still critical in many scenarios.

This paper resolves this problem from a causal perspective: we directly learn an identifiable query. More particularly, we take a very first step to explore this non-positivity issue on one of CTG tasks, that is, when there is only one attribute and $\boldsymbol{x} = (input, attribute)$, which refers to a text style transfer (TST) task. We introduce do-calculus (Pearl, 1995) to revisit the formulation of the TST problem. Experiments are conducted on formality transfer to verify our approach.

Formality style transfer (FST) is a subtask of TST, which aims to transfer informal speech to formal speech or vice versa, with scarce parallel-data available. Approaches to address this data scarcity problem in FST includes data augmentation with related tasks (Zhang et al., 2020), additional constraints such as auxiliary losses(Wang et al., 2020) and rule-based injection(Yao and Yu, 2021). Recently, methods based on large language models also (Reif et al., 2022; Lai et al., 2021; Chawla and Yang, 2020) also manage to leverage external knowledge from large datasets to resolve the data scarcity problem.

Experiments show that our method achieve competitive results as such methods on two formality transfer datasets for all four metrics. Besides, we also empirically show that incorporating causal inference makes the results more interpretable.

2 Preliminaries

2.1 Text Style Transfer

Text style transfer aims to transfer *(input, attribute)* pairs into *output* sentences which have the given *attribute*. Explicitly, an input text $x = (x_1, x_2, ..., x_n)$ is first fed into a language model to learn the representation $P(\boldsymbol{X}) = P(x_1, x_2, ..., x_n)$. Together with attribute s, we hope to generate a $P(\boldsymbol{Y}) = P(y_1, y_2, ..., y_n)$ where s can be recognized. A commonly used probabilistic paradigm (Dai et al., 2019; Zhang et al., 2022) is:

$$P(\boldsymbol{Y}|\boldsymbol{X}, s) = P(y_1, y_2, ..., y_n|x_1, x_2, ..., x_n, s), \tag{1}$$

2.2 The Non-Identifiability Problem in Current Style Transfer Approaches

A common pitfall lies in Equation 1: the positivity of $p(\boldsymbol{X}, s)$ cannot be guaranteed. Consider a sentiment transfer task where we transfer emotionally positive sentence $\boldsymbol{X}$ into a negative one, such that the input pair is $(\boldsymbol{X}, negative)$. However, if $\boldsymbol{X}$ conveys only positivity, this leads to $P(\boldsymbol{X}, negative) = 0$ and correspondingly, $P(\boldsymbol{Y}|\boldsymbol{X}, negative)$ *non-identifiable*. Intuitively, we cannot guarantee every $P(\boldsymbol{X}, s)$ larger than 0 since texts are sampled randomly. As a result, current style transfer algorithms might learn different causal diagrams for the same transfer query (i.e. for the same sentiment transfer task on the same dataset) since outputs for some $(\boldsymbol{X}, s)$ pairs can be undefined. Although there is no strict ground truth for a style transfer task, it is unknown how the model behaves for undefined cases, and whether the inclusion of such cases in the training process results in inaccurate causal diagrams and consequently yields uninterpretable incorrect outcomes.

3 Incorporating TST with Causality

The goal is then to learn an identifiable query instead of $P(\boldsymbol{Y}|\boldsymbol{X}, s)$. Our choice is to use a generative process known as a structural causal model (SCM) (Pearl, 2009), such that we can learn a causal effect $P(\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}, s))$. In fact, this is quite direct and intuitive: in essence, the goal of TST is

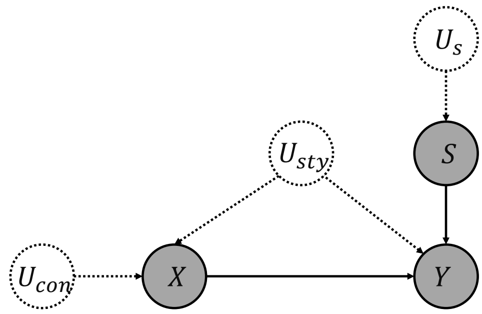

Figure 1: Causal graph for text style transfer. Input text X is constructed from content information in U_{con} and the information of its original attributes from U_{sty}. Y is transferred using conceptual information in X and the target attribute s.

to see how the change of s intervenes on the generation of $\boldsymbol{Y}$ given the content from $\boldsymbol{X}$. The SCM is constructed as Figure 1. We now start by grounding the TST problem in this causal framework.

3.1 SCM Formulation

The SCM is formulated in Figure 1. The proposed SCM encodes a 4-tuple $\langle V = \{X, Y, s\}, U = \{U_{con}, U_{sty}, U_s\}, \mathcal{F} = \{f_X, f_Y, f_s\}, P(U)\rangle$, where V is the set of observed variables, with X, Y, s as the input text, the target output text and the target attribute, respectively. $\mathcal{F}$ determines the generation process of X, Y, s such that $f_X(U_{con}, U_{sty}) \rightarrow X$, $f_s(U_s) \rightarrow s$, and $f_Y(X, s, U_{sty}) \rightarrow Y$. $P(U)$ represents a probability distribution over the unobserved variables. U contains the unobserved variables, U_{con} and U_{sty}, with U_{con} containing the content information in X while U_{sty} containing the style information in X. To perform style transfer, Y takes content information from X and style information from s. U_{sty} has spurious on Y because of the information of the original attribute in X. U_s contains the target attribute information.

We formulate the SCM so that we can next use causal inference tools to avoid the non-positivity issue. The problem lies in that X contains confounding features preventing Y from having the target attribute s. Some previous approaches propose style disentanglement (John et al., 2019; Cheng et al., 2020) techniques to extract content information from X. However, the effect of U_{sty} on Y is unobserved, meaning that the full information in U_{sty} cannot be captured by the model. But such an elimination can be conducted on an interventional level using causal inference. We now discuss

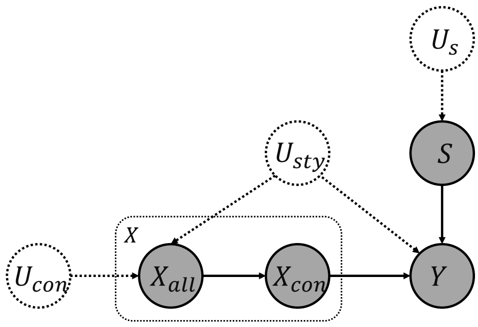

Figure 2: Causal graph for text style transfer after decomposition as described in Proposition 1 (Decomposition). Gray nodes denote observable variables while hollow nodes denote unobserved variables.

how to formulate the query as an identifiable causal estimand.

3.2 Query Formulation - An identifiable Estimand

We now propose a decomposition for later leveraging causal inference tools to eliminate the spurious effect:

Proposition 1 (Decomposition) *We first decompose each X as $X = \{X_{all}, X_{con}\}$ with X_{all} containing spurious factors, and X_{con} containing causal factors.*

X_{all} contains low-level semantic information including the original attribute, which might be confounding with s. X_{con} contains conceptual information of X, which is necessary for content preservation in Y. The decomposition blocks confounding information in X_{all} flowing into Y, such that our model can perform style transfer correctly. The decomposed SCM is shown in Figure 2 for clearer illustration.

The spurious effect is now between X_{all} and Y through U_{sty}. To eliminate such spurious effect, do-calculus (Pearl, 2009) is necessary to be introduced to perform $\text{do}(X_{con})$ to intervene the causal effect of X_{all} on X_{con}.

Proposition 2(interventional query) *We use $P(Y|do(X,s))$ as a surrogate for $P(Y|X,s)$ to solve the non-identifiability issue.*

With Proposition 1 (Decomposition) and Proposition 2(interventional query) we can now have the following derivation:

$$
\begin{aligned}
& P(Y|\text{do}(X,s)) \\
=& P(Y|\text{do}(X_{all}, X_{con}, s)) && \text{(Assumption 1)} \\
=& P(Y|\text{do}(X_{all}, X_{con}), s) && \text{(rule 2(Pearl, 2009))} \\
=& P(Y|\text{do}(X_{con}), s) && \text{(rule 3 (Pearl, 2009))} \\
=& \sum_{X'_{all}} P(Y|X_{con}, s, X'_{all}) P(X'_{all}) \\
& && \text{(Backdoor Criterion)}
\end{aligned}
$$

where X'_{all} is the low-level semantic information from another random sample and other notions remain the same. We then leverage the conditional independence between X_{all} and s to further guarantee the identifiability of our query:

$$
\begin{aligned}
& \sum_{X'_{all}} P(Y|X_{con}, s, X'_{all}) P(X'_{all}) \\
=& \sum_{X'_{all}} P(Y|X_{con}, s, X'_{all}) P(X'_{all}|s) \\
=& \sum_{X'_{all}} P(Y|X_{con}, s, X'_{all}) \sum_{X'_{con}} P(X'_{all}, X'_{con}|s) \\
=& \sum_{x'} P(Y|X_{con}, s, X'_{all}) P(X'|s)
\end{aligned}
$$

This means instead of sampling $P(X'_{all})$, we directly sample X' from samples of target attribute s. The positivity of $P(X_{con}, s, X'_{all})$ can then be guaranteed for the following reasons:

1. $P(X_{con}, s) > 0$, for that X_{con} only contains conceptual information and can be combined with any attribute s.

2. $P(X_{con}, X'_{all}) > 0$, for that both X_{con} and X'_{all} comes from the same domain, such that their conceptual information share similarity. For example, if both samples are product reviews, their targets are both commenting on their purchases.

3. $P(s, X'_{all}) > 0$, for that we sample X'_{all} from samples which have attribute s.

The query $P(Y|\text{do}(X,s))$ can now be approximated using an identifiable estimand. We now construct neural network models to satisfy the estimand of in the following discussion.

3.3 Network Modeling

To properly use the causal estimand, token-level causal effects need to be eliminated, such that we can treat X and Y as two integrals to perform causal inference.

Proposition 3(Non-autogressive Modeling) *All text representation is encoded using a non-autoregressive (NAR) (Devlin et al., 2019; Shen et al., 2020) manner, which can be formalized as:*

$$P(Y|X, s) = \prod_{i=1}^{N} P(y_i|X, i, N, s), \quad (2)$$

By using an NAR approach, positional information and contextual information can be encoded globally, making each token representation sufficient to make independent inference (Gu et al., 2018).

We can now learn causal effects between X and Y on sentence-level. The estimation for $P(X'|s)$ is straightforward since we only need to sample X' from samples of target attribute s. We then describe how to construct $P(Y|X_{con}, s, X'_{all})$.

Constructing X_{con}, s and X_{all} X_{con}, s and X_{all} are learned by separate modules. X first goes through a pre-trained language model to obtain its text representation. X is then put through an attention module (Vaswani et al., 2017) to construct X_{con}, such that conceptual information can be extracted. As for X_{all}, we use a single projection layer combined with an average pooling layer in order to keep all low-level semantics. s will be learned using an embedding layer (Devlin et al., 2019). The above process can be written as:

$$\boldsymbol{X_{enc}} = \text{BartEncoder}(x^1, x^2, ..., x^n), \quad (3)$$
$$X_{con} = \text{Attn}(X^{enc}), \quad (4)$$
$$s_{emb} = \text{embedding}(s), \quad (5)$$
$$X_{all} = \text{AvgPool}(\text{Proj}(X_{enc})), \quad (6)$$

where $\boldsymbol{X} = (x^1, x^2, ..., x^n)$. Then, we combine the learned X_{con}, s and X_{all} together to perform inference of Y.

Causal Prompt Tuning We borrow the prompt tuning idea from (Li and Liang, 2021) to combine the information of X_{con}, s, X'_{all} together, which is to reform the decoder input as:

$$input_{dec} = \langle s \rangle s^{enc} \langle /s \rangle X_{all}^{enc'} \langle /s \rangle \boldsymbol{x_{con}^{enc}} \langle s \rangle, \quad (7)$$

where $\langle s \rangle$ and $\langle /s \rangle$ denote the start token and the end token of the pretrained model. We separate the elements by special tokens such that the model learns they are functionally different.

Learning Objectives Our loss function consists of four parts: a style transfer loss $\mathcal{L}_{style}$, a cross-entropy-based content preservation loss $\mathcal{L}_{CE}$, a bleu-based (Papineni et al., 2002) content preservation loss $\mathcal{L}_{bleu}$, and a back-door adjustment loss.

The style transfer loss utilizes a pretrained TextCNN (Kim, 2014) classifier to penalize on transferred sentences which don't have the target attribute. The loss can be formalized as:

$$\mathcal{L}_{style} = -\mathbb{E}_{(x,s)\sim\mathcal{D}}[\log p_{cls}(s|\hat{y})], \quad (8)$$

where p_{cls} is the conditional distribution computed by the pretrained TextCNN cls. $\hat{y}$ is the generated output of (x, s) using soft sampling.

The content preservation loss includes two parts: a cross-entropy based loss, and a bleu-based loss. The cross-entropy one can be written as:

$$\mathcal{L}_{CE} = -\mathbb{E}_{(x,s)\sim\mathcal{D}}[\log p_{\Theta}(\boldsymbol{y}|\boldsymbol{x}, s)], \quad (9)$$

where Θ denotes the parameter set of the entire network describe in Section 3.3. However, the cross-entropy loss is limited since it penalizes on generated sentences which contain the same tokens as the ones in the reference outputs, but in different positions. We leverage the bleu-based loss from (Lai et al., 2021) to resolve this issue:

$$\mathcal{L}_{bleu} = bleu(\hat{\boldsymbol{y}}, \boldsymbol{y}) - bleu(\boldsymbol{y^s}, \boldsymbol{y}), \quad (10)$$

where $\hat{\boldsymbol{y}}$ is the same as above, $\boldsymbol{y^s}$ is randomly sampled from the distribution.

The back-door adjustment loss ensures X_{all} to contain low-level semantic information in X. This means the information in X_{all} and X should be similar. Besides, since X_{all} should contain only low-level semantics, we want its distribution to be less proxy. We leverage a KL-divergence penalty to achieve this goal:

$$\begin{aligned} L_{bd} =& \mathbb{E}_{(x,s)\sim\mathcal{D}}||\boldsymbol{x^{enc}} - \boldsymbol{x_{all}}||_2^2 \\ &+ KL[p_{\Theta}(\boldsymbol{x_{all}}|\boldsymbol{x})||p(\boldsymbol{x_{all}})] \end{aligned} \quad (11)$$

For each input text, we randomly sample in total λ back-door samples containing target attribute in a single iteration. This means, for every single iteration, a sample goes through the network for λ

times. We take the average loss scores into total loss computation. Finally, the total loss is:

$$\mathcal{L}_{total} = \sum_{i=1}^{\lambda}(\mathcal{L}_{style} + \mathcal{L}_{CE} + \mathcal{L}_{bleu} + \mathcal{L}_{bd}) \quad (12)$$

3.4 Workflow

We use two well-known NAR models, BART(Lewis et al., 2020) and T5(Raffel et al., 2020) as the pre-trained language model. For every iteration, the target attribute is embeded as s_{emb} using Equation 5, and each input text X is first encoded as X_{enc} using Equation 3. Then X_{con} and X_{all} are obtained using Equation 4 and 6 separately. We then combine these outputs using Equation 7 and put through the decoder and the subsequent generation head of the pretrained model. The model is trained by the loss described in Equation 12. Details of the proposed algorithm are given in Algorithm 1 for clearer illustration.

Algorithm 1 Causal Style Transfer Training

Require: $\mathcal{D} = \{x_i, s_i, y_i\}_{i=1}^{N}$, dataset size N, # of back-door samples λ,
  while $i < N$ **do**
    sample (x_i, s_i, y_i) from $\mathcal{D}$
    sample (x'_j, s'_j, y'_j) from samples of the target attribute s_j for λ times to obtain $\{(x'_j)\}_{j=1}^{\lambda}$
    while $j < \lambda$ **do**
      Train the model to obtain $X_{con}, s_{emb}, X'_{all}$ using Equation 3,4,5,6
      Train $P(Y|X_{con}, s, X'_{all})$ using Equation 7
    end while
    Calculate the average loss using Equation 12
  end while

4 Experiments

4.1 Experimental Settings

Datasets Experiments are conducted on the two domains of Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018), which are Family & Relationships (F&R), and Entertainment & Music (E&M). Each domain has approximately 100K informal-formal sentence pairs. We follow the data processing process as (Lai et al., 2021) for fair comparison.

Evaluation Metrics We consider both the style strength and content preservation for the evaluation of style transfer quality. For style strength, we use a pre-trained TextCNN (Kim, 2014) classifier to evaluate the transfer accuracy following (Lai et al., 2021). The classifier has an accuracy of 89.2% on E&M and 90.1% on F&R. We use the classifier to see whether the transferred sentences can be recognized as having the target attribute. For content preservation, we consider two metrics: BLEU(Papineni et al., 2002) and BLEURT(Sellam et al., 2020). Both metrics compute the similarity between transferred sentences and human references. We use BLEURT in addition to BLEU because it is reported to have better alignments with human evaluation. Both BLEU[1] and BLEURT[2] are computed using official packages for fair comparison. Besides, we also include the geometric mean (GM) of all three metrics mentioned above to evaluate the overall transfer quality of our model.

Training Setups We fine-tune BART and T5 using AdamW (Loshchilov and Hutter, 2019), with the learning rate set as 5×10^{-5} and the batch size as 32. For the hyper-parameter λ, we use $\lambda = 5$ in our experiments. Further discussion on the choice of λ is given in section 4.4. Other parameter settings follow the configuration of the pretrained BART and T5.

Baselines We compare our model to in total five baselines: PBMT-combined (Rao and Tetreault, 2018), NMT-combined(Rao and Tetreault, 2018), Bi-direction FT(Niu et al., 2018), tkBART(Lai et al., 2021), tkGPT(Lai et al., 2021). Results from unsupervised approaches are not included in the baseline, for that supervised methods all significantly outperform unsupervised approaches. We use officially reported results from these baselines for fair comparison. As for models which are lacking in officially reported results, we use their official codes to run the models to obtain results.

4.2 Results Analysis

The overall results are shown in Table 1. For simplicity, we denote our approach as CausalST and notions for all baselines same as the above mentioned ones. We test in total four alternatives of our models, considering two aspects: 1) the alternative of base models, denoted as CausalST(BART) and CausalST(T5). 2) the necessity of incorporating $do-calculus$ to solve the non-identifiability

[1] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[2] https://github.com/google-research/bleurt

Model	E&M				F&R			
	ACC↑	BLEU↑	BLEURT↑	GM↑	ACC↑	BLEU↑	BLEURT↑	GM↑
Source	4.5	36.6	0.012	1.41	6.7	32.3	0.006	1.14
NMT-combined	79.7	50.1	-0.100	N/A	79.8	52.7	-0.089	N/A
PBMT-combined	75.3	50.2	-0.088	N/A	78.8	51.7	-0.062	N/A
Bi-direction FT	81.8	55.4	0.023	10.21	83.9	56.8	0.037	13.27
tkBART	85.9	**57.7**	0.044	14.76	88.2	**59.5**	**0.068**	18.89
tkGPT	**92.3**	54.2	-0.007	N/A	91.5	57.2	0.038	14.10
CausalST(BART) w/o bda	86.9	55.6	0.041	14.07	90.5	56.1	0.066	18.31
CausalST(BART)	90.2	57.4	**0.054**	**16.72**	**92.6**	58.1	0.067	**18.98**
CausalST(T5) w/o bda	85.0	55.5	0.033	12.47	N/A	N/A	N/A	N/A
CausalST(T5)	88.3	56.6	0.037	13.60	N/A	N/A	N/A	N/A

Table 1: Overall results of both previous approaches and our model on GYAFC E&M and GYAFC F&R datasets. The best result of each metric is in bold. For all four metrics, a higher score means better model performance.

issue. The alternative approaches without the incorporation of $do-calculus$ are denoted as CausalST(BART) w/o bda and CausalST(T5) w/o bda.

We can see that CausalST(BART) achieve the best GM in both datasets, indicating that it has the highest overall transfer quality. Although tkGPT achieves the best accuracy in E%M dataset, this is probably because auto-regressive approaches aligns better with the nature of text generation. However, such an autoregressive manner also provides the model with excessive freedom, resulting in that it has much lower BLEU and BLEURT scores and consequently lower overall transfer quality. Apart from that, even though tkBART achieves the best BLEU and BLEURT in the domain of F&R, our results are competitive, and outperforms in accuracy by 5%. All above results indicate the success usage of an identifiability estimand.

Apart from that, the T5-based CausalST model has an inferior performance due to the nature of T5, which is pretrained for text summarization for long texts, such that it doesn't have many advantages in handling short texts. Besides, we can see that the success of our approach doesn't come from the design of the base model, as our approach outperforms the alternative approach without the incorporation of causality to a large extent: for every metric on every dataset, our approach with causality incorporated outperforms the one without causality incorporated comprehensively.

4.3 Ablation Study: Graphical Design

The ablation study is conducted using the BART-based version of CausalST using the $E\&M$ domain of GYAFC. Table 2 shows results of our model with and without implementation the decomposition described in Proposition 1 (Decomposition). We can see that our approach with the decomposition setting outperforms the one without decomposition in accuracy for approximately 13%, and in GM for 7%. Although they have competitive BLEU and BLEURT scores, this is possibly resulting from the copying nature of models without decomposition: because X contains factors that confound Y from transferring to the target attribute s, Y tends to copy most concepts from X, including the ones having the original attribute. This conclusion can be supported by the much lower accuracy of CausalST without decomposition. Apart from that, a 1%-2% gap in content preservation can be interpreted attributed to the diversity of outputs generated from pretrained large language models. Overall, the GM is high enough to verify the necessity for having a decomposition process in our model.

	E&M			
	ACC	BLEU	BLEURT	GM
CausalST	**90.2**	57.4	**0.054**	**16.72**
CausalST w/o dcp.	79.1	**58.8**	0.052	15.5

Table 2: Graphical Analysis: necessity for Proposition 1 (Decomposition) CausalST w/o dcp. denotes the alternative of our model, which directly put X through Y without decomposition. Better scores are marked in bold.

4.4 Ablation Study: Back-door Adjustment Settings

We experiment on the number of back-door adjustment samples to see how the change of λ affects the performance of our model.

Experiments show that a middle value of λ gives the best performance. It is quite interpretable, since with too few back-door samples, models might treat the back-door samples with the same level of importance as the sample we are really trained on. However, excessive back-door samples bring noise to the training, as we can see that the performance suddenly drops when we use $\lambda = 10$.

Overall, we can see that when λ is in an appropriate range, the performance of our model remains quite stable, which indicates the robustness of our approach.

4.5 Sampled Results Analysis

Formal → Informal	
SRC: It does not exist.	
CausalST(BART)	There is no such thing ah !
CausalST(T5)	No where there is no such a thing !
tkBART	There is no such thing as it.
NMT-combined	It does not exist at all.
SRC: I am fairly certain it was a kiss.	
CausalST(BART)	Im pretty sure it was a kiss.
CausalST(T5)	I'm pretty sure it was a kiss.
tkBART	I'm pretty sure it was a kiss.
NMT-combined	I'm pretty sure it was a kiss.
Informal → Formal	
SRC: Whats your favorite movie and why?	
CausalST(BART)	What is your favorite movie and why?
CausalST(T5)	What is your favorite movie and why?
tkBART	What is your favorite movie and why?
NMT-combined	What is your favorite movie and why?

Table 3: Sampled results of both *Formal→Informal* and *Informal→Formal* transfer. SRC is the input sentence.

Table 3 shows sampled results from our approach and selected baselines. All models show their capacities of extracting signal words to perform formal↔informal transfer (e.g. you ↔ u, I am ↔ I'm, Im). However, for sentences lacking in such words, where formal↔informal transfer requires to be performed only if models comprehend the sample on a sentence level instead of merely detecting signal words, baseline models fails to perform satisfying stylization and the results vary from model to model. We can see that results from tkBART and NMT-combined are just paraphrasing the inputs without transferring them into informal ones. This is probably because they are not learning the correct diagram, leading to the occurrence of uninterpretable incorrect outputs. However, CausalST manages to add particles to the transferred outputs, indicating that it really learns the task of performing formal↔informal transfer.

	E&M			
	ACC	BLEU	BLEURT	GM
λ=3	89.9	56.8	0.050	15.97
λ=5	90.2	**57.4**	**0.054**	**16.72**
λ=7	**90.8**	57.1	0.046	15.44
λ=10	88.1	53.2	0.031	12.05

Table 4: Ablation Study: Back-door samples settings. λ has the same meaning as in section 3.3, denoting the number of back-door samples involved in a single iteration.

5 Conclusion

In this paper, we introduce an approach based on causal inference to resolve the non-identifiability issue of current controllable text generation approaches. We formulate a structural causal model and propose an identifiable query. We construct a neural network using pretrained language model to compute this identifiable estimand. Experiments are conducted on a single-attribute controllable text generation task, text style transfer of formality. We empirically show that our model is competitive to other models on two datasets for all four metrics. Besides, outputs of our model outperforms other models in interpretability. In the future, we plan to study problems with multi-attribute transfer, and extend this approach to models with auto-regressive manner.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

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

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. 2020. Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguis-*

tics, ACL 2020, Online, July 5-10, 2020, pages 7530–7541.

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.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

David Galles and Judea Pearl. 1995. Testing identifiability of causal effects. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*, pages 185–195.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70, pages 1587–1596.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 424–434.

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

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1008–1021.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Judea Pearl. 1995. A causal calculus for statistical research. In *Learning from Data - Fifth International Workshop on Artificial Intelligence and Statistics, AISTATS 1995, Key West, Florida, USA, January, 1995. Proceedings*, pages 23–33.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sudha Rao and Joel R. Tetreault. 2018. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *CoRR*, abs/2109.03910.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 837–848.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi S. Jaakkola. 2020. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5186–5198.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3267–3277.

Jin Tian and Judea Pearl. 2002. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada*, pages 567–573.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2236–2249.

Zonghai Yao and Hong Yu. 2021. Improving formality style transfer with context-aware rule injection. 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 1561–1570, Online.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *CoRR*, abs/2201.05337.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *CoRR*, abs/1801.07243.

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