A Causal Lens for Controllable Text Generation

Anonymous authors

Abstract

Controllable text generation is one of the most important yet challenging tasks in natural language processing (NLP). Prior methods relying on disentangled representations or language models suffer from limited controllability due to biases or spurious correlations in the data. To remedy this, we present a causal framework for controllable text generation. Our method allows control over one or more attributes such as sentiment and cuisine type and perform counterfactual predictions. Our experiments demonstrate that our models can enable effective control despite data biases.

1 Introduction

The task of controllable text generation is to generate natural sentences whose attributes can be controlled. The attributes to control can be very diverse including sentiments, personal attributes such as gender, age and content. Controllable text generation has many applications. For example, in dialog response, one may want to control persona, sentiment, politeness, and topic. To help non-expert writers, a writing assistant can be used to make the text more formal or meet other advanced writing requirements. Controllable text generation can be a general tool to (post-)process the generated text of many tasks with the desired attributes.

Previous work on controllable text generation proposes methods to solve two separate tasks, attribute-conditional text generation and text style transfer. Previous methods, viewed from the lens of Pearl causal hierarchy (Pearl, 2009), is either associational which learns the probabilistic patterns between input and generated text or is limited to the setting where there is no hidden confounders. We argue that controllable text generation is naturally a causal inference problem. We define a causal graph consisting of attributes or styles to control (treatment) and a latent variable representing the rest of the content (confounders), and the generated text (outcome). Attribute-conditional generation corresponds to intervention on the attributes which is the second ladder of causal inference, and text style transfer corresponds to counterfactual prediction as the third ladder.

Prior controllable generation modeling tends to inherit the biases in the data. For example, it is more likely to generate text with male doctors. The proposed causal framework in contrast mitigates the problem and generates balanced samples through intervention and counterfactual inference, which can be used in downstream tasks such as training unbiased classifiers for language applications.

2 Related Work

With the success of deep learning, a variety of neural methods have been proposed for controllable text generation. If parallel data are provided, standard sequence-to-sequence models are often directly applied for the task (Rao and Tetreault, 2018). However, most use cases do not have parallel data, so controllable text generation on non-parallel corpora becomes a prolific research area. The first line of approaches disentangle text into its content and attribute in the latent space, and apply generative modeling (Hu et al., 2017; Shen et al., 2017). This trend was then joined by another distinctive line of approach, prototype-based text editing (Li et al., 2018) which extracts a sentence template and attribute markers to generate the text. Another paradigm soon followed, i.e., back-translation to generate pseudo-parallel data (Zhang et al., 2018; Jin et al., 2019; Lample et al., 2019), inspired by unsupervised machine translation (UMT). These three directions, (1) disentanglement, (2) prototype-based text editing, and (3) back-translation, are further advanced with the emergence of

Transformer-based models (Sudhakar et al., 2019; Malmi et al., 2020). Different from prior work, we present the first causal framework for controllable text generation. The causal framework enables effective control in the presences of data biases.

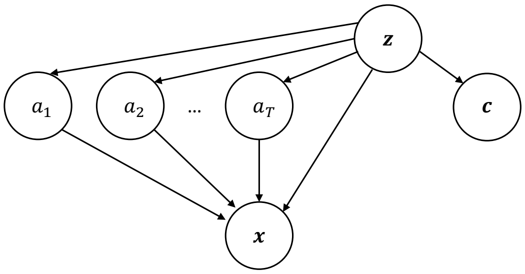

Figure 1: The structural causal model of controllable text generation, where $\boldsymbol{a} = \{a_1, a_2, \ldots, a_T\}$ are attributes to control, $\boldsymbol{z}$ is unobserved confounders, $\boldsymbol{c}$ is partially observed information related to the hidden confounders $\boldsymbol{z}$, and $\boldsymbol{x}$ is the outcome, namely the generated text.

3 Methods

3.1 The Structural Causal Graph

We formulate the problem of controllable text generation through the lens of causal inference. Figure 1 shows the proposed structural causal model (SCM). For example, consider the problem of generating a short review text. Then $\boldsymbol{x}$ denotes the generated review (i.e., *outcome*) and $\boldsymbol{a} = \{a_1, a_2, \ldots, a_T\}$ is the attributes (e.g., sentiment) to control, which plays the role of *treatment*. A key component of causal inference is the *confounder* $\boldsymbol{z}$ that affects both the treatment and outcome, such as the food quality of the restaurant, the mood of the customer, and others. Crucially, in practice the confounder is unobserved or it is impossible to measure the confounding factors directly. To model the causal effect of attributes $\boldsymbol{a}$ on the outcome $\boldsymbol{x}$, it is necessary to infer the hidden confounder $\boldsymbol{z}$ from any observed information. In controllable text generation, the text $\boldsymbol{x}$ (e.g., *"The pizza tastes really good."*) as the outcome contains rich information beyond the attributes (e.g., positive sentiment). The additional information is supposed to be from the confounder $\boldsymbol{z}$. Thus the outcome $\boldsymbol{x}$ itself can serve as a crucial source for inferring the confounder. The auxiliary information $\boldsymbol{c}$ offers another clue to infer $\boldsymbol{z}$. For example, in this preliminary study, we consider the category of the review subject as the auxiliary information. That is, $\boldsymbol{c}$ is a binary variable such that $\boldsymbol{c} = 1$ indicates the review text is about a restaurant while $\boldsymbol{c} = 0$ indicates the review text is about any other subjects. In other problems (e.g., to mitigate gender bias), we may consider other $\boldsymbol{c}$ variables (e.g., gender). This is one of the major differences of the causal text model compared to previous causal models studied in other problems, such as medication effect prediction (Louizos et al., 2017), where the outcome is usually a single binary variable (e.g., mortality) and thus the inference of confounder requires $\boldsymbol{c}$ to encode sufficiently rich information.

In most realistic settings, the $\boldsymbol{c}$ information is available only for a small subset of instances (e.g., by human annotation). Our framework allows to incorporate any such available auxiliary labels in the modeling for mitigating biases.

With the above causal model, we are ready to formulate the two common tasks in controllable text generation: **(1) Attribute-conditional generation**: $p(\boldsymbol{x}|do(a_i = a))$ which corresponds to *intervention* through the *do* operation; and **(2) Text style transfer**: $p(\boldsymbol{x}|\boldsymbol{x}', a_i', a_i = a)$ which corresponds to *counterfactual prediction*. We elaborate the two inference tasks later.

3.2 Learning

We want to learn the causal model from a given set of attribute-text pairs $\{(\boldsymbol{a}, \boldsymbol{x})\}$. Optionally, for a small subset of instances, we additionally know the auxiliary information $\boldsymbol{c}$. The small set of $\boldsymbol{c}$ labels enables us to break spurious correlations between the attribute of interest $\boldsymbol{a}$ and the respective confounder. The learned model is then used to perform the above two inference procedures. The model formulates the joint probability distribution $p(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{c}) = p(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{z})p(\boldsymbol{a}|\boldsymbol{z})p(\boldsymbol{c}|\boldsymbol{z})p(\boldsymbol{z})$.

We parameterize the above model distribution using deep neural networks with parameters $\boldsymbol{\theta}$, and introduce the variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{c})$ which is also parameterized as deep neural networks with parameters $\boldsymbol{\phi}$. We then write the variational inference objective as below:

$$\mathcal{L} = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{c})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{z})p_\theta(\boldsymbol{c}|\boldsymbol{z})p_\theta(\boldsymbol{a}|\boldsymbol{z})\right] + \mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{c})\|p(\boldsymbol{z})\right). \tag{1}$$

Note that the reconstruction of $\boldsymbol{c}$ only applies to those instance whose $\boldsymbol{c}$ labels are available. The goal is to minimize the objective and obtain the optimal parameters for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Following (Hu et al., 2017), we additionally use a pretrained attribute classifier $g(\boldsymbol{a}|\boldsymbol{x})$ and encourage the generated counterfactual samples (see below) to entail the desired counterfactual attribute. We omit the details and refer readers to (Hu et al., 2017).

3.3 Inference

With the learned model, we then perform inference for attribute-conditional generation and text style transfer, corresponding to intervention and counterfactual prediction in causal inference, respectively.

Attribute-conditional generation Generating text $\boldsymbol{x}$ given an attribute $a_i = a$ is formulated using the do-operation in the causal framework, which effectively removes any links into the intervention node a_i and makes sure the outcome is not biased by confounder. Concretely, the inference problem is written as:

$$p(\boldsymbol{x}|do(a_i = a)) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}|a_i = a, \boldsymbol{z})p(\boldsymbol{z}), \tag{2}$$

where we marginalize out the confounder $\boldsymbol{z}$.

We emphasize the key difference of the above formulation of attribute-conditional generation compared to the conventional formulation studied in previous controllable text generation literature. In particular, previous work without using the causal framework has treated the problem as the simple conditional inference through:

$$p(\boldsymbol{x}|a_i = a) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}|a_i = a, \boldsymbol{z})p(\boldsymbol{z}|a_i = a). \tag{3}$$

Note that here $\boldsymbol{z}$ is marginalized under the conditional distribution $p(\boldsymbol{z}|a_i = a)$, as opposed to the marginal $p(\boldsymbol{z})$ in our framework. The key difference renders the previous methods based on $p(\boldsymbol{x}|a_i = a)$ biased by the confounders.

Text style transfer Text style transfer is another common task in controllable text generation, in which we aim to generate a new piece of text $\boldsymbol{x}$ given the original text $\boldsymbol{x}'$ such that $\boldsymbol{x}$ has the designated attribute $a_i = a$ while preserving all other characteristics of $\boldsymbol{x}'$. Within our causal framework, the problem nicely corresponds to counterfactual prediction, where $\boldsymbol{x}$ is the counterfactual outcome of $\boldsymbol{x}'$ given the new attribute $a_i = a$.

Thus, given the original text $\boldsymbol{x}'$ and the target attribute $a_i = a$, we follow the standard principled counterfactual prediction procedure to generate $\boldsymbol{x}$, by first inferring the latent $\boldsymbol{z}$ from $\boldsymbol{x}'$ through the learned q_ϕ, then setting the attribute a_i to the given value a, and finally generating the counterfactual $\boldsymbol{x}$ through the learned p_θ.

4 Experiments

We evaluate the above causal framework for controllable text generation using two Yelp datasets. For the causal model, we initialize both the encoder q_ϕ and decoder p_θ with GPT2 (Medium), respectively. The encoder-decoder connecting layer follows the architecture of (Li et al., 2020).

4.1 Counterfactual Prediction: Text Style Transfer

We use the commonly used Yelp review dataset (Shen et al., 2017) that aims to manipulate the sentiment of restaurant reviews.

Methods	Accuracy (↑)	self-BLEU (↑)	ref-BLEU (↑)	Perplexity (↓)
(Hu et al., 2017)	86.7	58.4	-	177.7
(Shen et al., 2017)	73.9	20.7	7.8	72.0
(He et al., 2019)	87.9	48.38	18.67	**31.7**
Ours	**90.1**	**59.0**	**24.8**	43.5

Table 1: Results of text style transfer.

Methods	*Controllability*	*Bias*	*Utility*
	Sentiment Accu (↑)	Category Accu (↓)	Downstream Accu (↑)
Conditional LM	80.1	78.7	77.1
Ours	**91.6**	**72.3**	**78.6**

Table 2: Results of control and debiasing.

Table 1 shows the results, where we measure different metrics including accuracy (how accurate the generation encodes the target attribute), self-BLEU and ref-BLEU (how well the generation preserves all other characteristics of the original text, by measuring the overlap with the original text and human-written ground-truth text, respectively), and perplexity (how fluent the generation is). We can see that our method achieves the best balance across the different metrics, and improving the self-BLEU and ref-BLEU scores over the previous approaches by around 10 and 6 absolute points.

4.2 Intervention: Debiasing Attribute Classifier

We evaluate how well the causal model could be used to generate unbiased data which in turn can be used to train unbiased classifiers. Following previous causality work (e.g., Sauer and Geiger, 2021), we created a biased dataset out of the full Yelp dataset [1], such that 90% of text examples in the resulting dataset have the same sentiment (0:negative and 1:positive) and category (0:non-restaurant, 1:restaurant) labels. That is, sentiment $\boldsymbol{a}$ and category $\boldsymbol{c}$ have a strong correlation. We make the dataset of the same size of the dataset used in section 4.1. Different from the sentiment labels, we assume that the category label as an auxiliary information is only accessible on a very small portion of instances (10K, around 2% of all instances). The goal is to learn to disentangle sentiment from the confounding category and to generate text with controlled sentiment that is not correlated with category. With the learned model, we apply the do-operation (Eq.2) to generate samples which are expected to have weaker correlation between sentiment and category. As a preliminary comparison, we train a sentiment-conditional language model (initialized with GPT2) on the biased Yelp dataset, and use the resulting model to generate the same amount of new samples where half is positive and half is negative.

To measure the controllability and sentiment-category correlation of the samples, we use two methods: (1) Given a text sample and its conditioning sentiment code, we respectively use a pretrained sentiment classifier and a pretrained category classifier to measure the "accuracy" of the classifiers predicting the sentiment code given the text sample. The higher the sentiment classifier accuracy, the more accurate the controllable generation. On the other hand, the lower the category classifier accuracy, the weaker the sentiment-category correlation (i.e., less biased); Table 2 shows the results, where we can see that our method substantially improve the controllablility and reduces the bias, compared to the vanilla conditional LM; (2) We use the generated samples to train a sentiment classifier. To achieve a high test accuracy, the generated samples as the training set need to be of high quality in general (i.e., accurate on sentiment, fluent, and not biased). We see from Table 2 that the classifier trained with the samples from our method achieves a higher accuracy. We are excited to conduct more comprehensive study in the future.

5 Conclusion and Future Work

Controllable text generation is a fundamental problem in natural language processing. Prior approaches apply disentanglement, language model or back translation to effect control. However, due to inherent and prevalent biases in datasets, these approaches achieve limited controllability. In this paper, we formulate controllable text generation as a causal inference problem. Through intervention and counterfactual predictions, we show that our models can achieve effective control.

[1] https://www.yelp.com/dataset/challenge

Our preliminary experimental results show that we can achieve improved and promising results. The generated balanced samples could potentially be used to mitigate biases in downstream classifiers.

For future work, we plan to conduct more thorough analysis of the framework, and extend our framework to control text summarization and dialog systems. We would also like to explore more for debiasing language models and knowledge grounding applications.

REFERENCES

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 (ICLR)*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, R. Salakhutdinov, and E. Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning (ICML)*.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. https://doi.org/10.18653/v1/D19-1306 IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

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

Chunyuan Li, Xiang Gao, Yuan Li, Xiujun Li, Baolin Peng, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. *arXiv preprint arXiv:2004.04092*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. https://doi.org/10.18653/v1/N18-1169 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, New Orleans, Louisiana. Association for Computational Linguistics.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems (NeurIPS)*, pages 6446–6456.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680.

Judea Pearl. 2009. *Causality*. Cambridge university press.

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

Axel Sauer and Andreas Geiger. 2021. Counterfactual generative networks. *arXiv preprint arXiv:2101.06046*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems (NeurIPS)*, pages 6830–6841.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP)*.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.