General Paper Peer-Reviewed

Focused Prefix Tuning for Controllable Text Generation

Congda Ma[†], Tianyu Zhao[††], Makoto Shing[†††], Kei Sawada[††] and Manabu Okumura[†]

In a controllable text generation dataset, unannotated attributes may provide irrelevant learning signals to models that use them for training, thereby degrading their performance. We propose *focused prefix tuning*(FPT) to mitigate this problem and enable control to focus on the desired attribute. Experimental results show that FPT can achieve better control accuracy and text fluency than baseline models in single-attribute control tasks. In multi-attribute control tasks, FPT achieves control accuracy comparable to that of the state-of-the-art approach while maintaining the flexibility to control new attributes without retraining existing models.

Key Words: *Controllable Text Generation, Parameter-Efficient Fine-Tuning*

1 Introduction

Controllable text generation aims to generate text associated with specific attributes. For example, given an attribute TOPIC = *sports* and a prompt "*There is*," a model is supposed to generate a continuation whose TOPIC is *sports*, such as "*There is a tennis match ...*".

In datasets for the controllable text generation task, there exists an annotated attribute, which we call the *explicit attribute* (e.g., the TOPIC attribute in the AGNews dataset). In addition to the *explicit attributes*, the datasets tend to have their own tendencies. For example, up to 98% of training data pieces in the IMDb dataset exhibit "TOPIC = *sci/tech*", while up to 94% of training data pieces in the AGNews exhibit "SENTIMENT = *negative*".[1] This tendency is called an *implicit attribute* (e.g., the TOPIC attribute in the IMDb dataset).

The existence of *implicit attributes* could degrade the performance in controlling an *explicit attribute* when the models are trained on the datasets. Since implicit attributes are at the dataset level and are related to undesired explicit attributes, the probability of generating content with implicit attributes is likely to increase initially. When the text with implicit attributes is generated, the probability of generating content with other undesired explicit attributes increases, and the text with these attributes might be generated next. Consequently, as shown in Table 1, the

[†] Tokyo Institute of Technology
[††] rinna Co. Ltd.
[†††] Stability AI Ltd.

[1] The models used for classification are from (Gu et al. 2022).

Model	Desired Attribute Relevance	Implicit Attribute Relevance
DExperts	81.95	76.54
Vanilla Prefix Tuning	71.94	90.64

Table 1 Relevance of texts generated by different models (e.g., DExperts and Vanilla Prefix Tuning) trained on IMDb dataset. We found a lower desired explicit attribute (e.g., SENTIMENT) relevance is related to a higher implicit attribute (e.g., TOPIC = *sci/tech*) relevance. The relevance is calculated by the classifier models in Sec. 4.1.2.

model generates content with high implicit attribute relevance but low desired explicit attribute relevance (e.g., Vanilla Prefix Tuning (Li and Liang 2021)). In contrast, if the model generates content with low implicit attribute relevance, it will have highly desired explicit attribute relevance (e.g., DExperts (Liu et al. 2021)). We call this phenomenon *attribute transfer*.

To mitigate the effects of attribute transfer, we propose *focused prefix tuning*(FPT), which focuses the generation on the desired explicit attributes. FPT uses *specific* and *general* prefixes to encode the explicit and implicit attributes, respectively. FPT combines the control power of the two prefixes via *logits manipulation* at inference time. Experimental results show that FPT achieves better control accuracy and fluency in single-attribute control tasks. In multi-attribute control tasks, FPT can achieve performance comparable to that of the state-of-the-art approach. Moreover, as we show, FPT enables the training of each attribute prefix individually, allowing for the incremental addition of new attributes without retraining all the prefixes.

2 Related Work

2.1 Controllable Generation

Methods for controlling text generation have rapidly developed (Ficler and Goldberg 2017; Dathathri et al. 2020; Madotto et al. 2020; Alvin et al. 2021). Keskar et al. (2019) trained a large transformer model to generate contents conditioned on up to 55 attributes. However, the cost of training such models is extremely high.

2.2 Prefix Tuning

Parameter-efficient fine-tuning (PEFT) methods, such as prompt tuning (Lester et al. 2021), have become particularly significant in driving various natural language processing tasks to reduce the high training cost. Prefix tuning (Li and Liang 2021) is a PEFT method that steers pre-

trained models (Radford et al. 2019; Lewis et al. 2020) by applying an additional continuous vector embedding before every activation layer. Qian et al. (2022) proposed a contrastive prefix-tuning method that improves its performance by utilizing the relations between attributes. However, they focused only on explicitly annotated attributes and ignored the effects of implicit attributes.

2.3 Inference-time Methods

Inference time methods (Mireshghallah et al. 2022; Yang and Klein 2021; Dathathri et al. 2020; Madotto et al. 2020), which are lightweight approaches that do not update the parameters, have been used for controllable text generation. To enhance controllability, Krause et al. (2021) proposed a method that combines the computed classification probability distributions. Liu et al. (2021) found that directly applying probability distributions from language models is a simple but effective approach to control generated texts. Inspired by their work, we propose a method that uses probability distributions from language models to eliminate the effects of implicit attributes.

3 Methodology

The task of controllable generation is, given a sequence of prompt tokens $x_{<t}$ and an attribute ATTR = *val* (e.g., TOPIC = *sports*), to generate a sequence of tokens as a continuation x that conforms to both the prompt and specified attribute.

3.1 Vanilla Prefix Tuning

In controllable text generation, a prefix can steer a pre-trained model parameterized by θ to generate texts under a specific attribute value ATTR = *val*. Particularly, vanilla prefix tuning (Li and Liang 2021) prepends a set of continuous vectors before each activation layer of the pre-trained transformer. The continuous vectors are referred to as the prefix $H_\phi^{\text{attr=val}}$, which is parameterized by ϕ.

During training, we freeze the pre-trained model's parameters θ and update only the prefix parameters ϕ to optimize the following objective:

$$-\sum_{x \in \mathcal{D}^{\text{attr=val}}} logP(x_t|x_{<t}, H_\phi^{\text{attr=val}}, \theta), \tag{1}$$

where $\mathcal{D}^{\text{attr=val}}$ is the subset of the entire dataset $\mathcal{D}$ whose attribute ATTR is *val*.

Following Li and Liang (2021), we initialize the prefix H_ϕ with the activation of actual tokens from the pre-trained model's vocabulary.

3.2 Specific and General Prefixes

The prefix in vanilla prefix tuning captures an explicit attribute in a dataset by training it on the subset dataset $\mathcal{D}^{\text{attr=val}}$. In contrast, to capture only implicit attributes while ignoring any explicit attributes, we propose training another prefix on the entire dataset $\mathcal{D}$. To distinguish between the two prefixes, we refer to the prefix trained on $\mathcal{D}^{\text{attr=val}}$ as a *specific prefix* and the one trained on $\mathcal{D}$ as a *general prefix*.

The specific prefix is the same as the one used in vanilla prefix tuning; therefore, we still use Equation 1 to update its parameters. To update the general prefix's parameters, we optimize the following objective:

$$-\sum_{x \in \mathcal{D}} logP(x_t|x_{<t}, H_{\phi'}^{\text{genl}}, \theta), \tag{2}$$

where $H_{\phi'}^{\text{genl}}$ represents the general prefix parameterized by ϕ'.

3.3 Inference-time Logits Manipulation

As shown in Figure 1, FPT suppresses the probability of words with implicit attributes in the generated text by combining logits $z^{\text{attr=val}}$ steered by the specific prefix and logits z^{genl} steered by the general prefix via logits manipulation at inference time. For example, when generating text with the attribute TOPIC = *sports*, the probability of words with implicit attributes (e.g., "*impossible*" with SENTIMENT = *negative*) would be suppressed. During inference, at each step t, we first make two forward runs respectively with the specific and general prefixes to obtain their logits, $z_t^{\text{attr=val}}$ and z_t^{genl}. Since $z_t^{\text{attr=val}}$ encodes both the explicit and implicit attributes while z_t^{genl} encodes mostly the implicit attributes, we use a subtraction operation at the logits

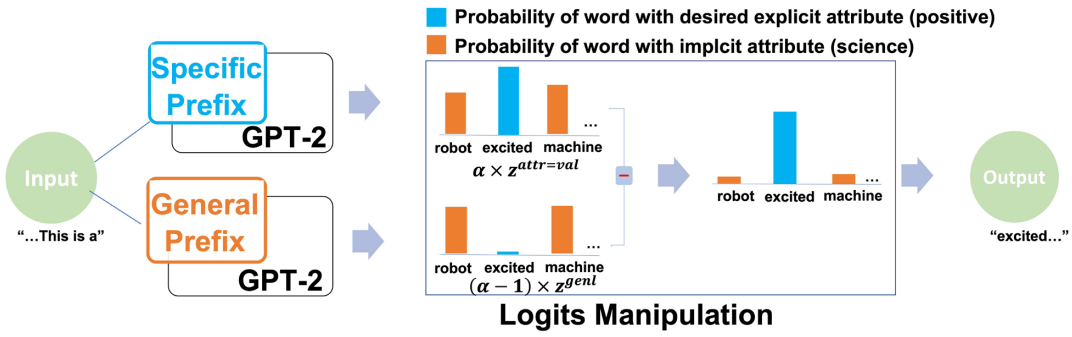

Figure 1 Proposed model framework.

level to suppress the probability of words with implicit attributes:

$$\begin{aligned} &P(x_t|x_{<t}, \text{ATTR} = val) \\ &\quad = P(x_t|x_{<t}, H_{\phi}^{\text{attr=val}}, H_{\phi'}^{\text{genl}}, \theta) \\ &\quad = \text{softmax}(\alpha z_t^{\text{attr=val}} - (\alpha - 1) z_t^{\text{genl}}), \end{aligned} \tag{3}$$

where α is a hyperparameter that can be interpreted as the strength of controlling for implicit attributes. Following Liu et al. (2021), we respectively set α and $\alpha - 1$ as the weight of $z^{\text{attr=val}}$ and z_t^{genl} to make the ratio of logits after the logits manipulation equal to 1.

To ensure the fluency of the generated texts, we follow Liu et al. (2021) and use top-p filtering (Holtzman et al. 2020) to remove the tokens that have low scores in advance before logits manipulation. In particular, we modify the logits produced by the specific prefix by calculating the top-p vocabulary $\widetilde{V}$ and setting all the logits outside $\widetilde{V}$ to $-\infty$:

$$\widetilde{z}[v] = \begin{cases} z[v], & if \quad v \in \widetilde{V} \\ -\infty, & if \quad v \notin \widetilde{V} \end{cases}. \tag{4}$$

Therefore, the logits manipulation in Equation 3 is updated as follows:

$$\begin{aligned} &P'(x_t|x_{<t}, \text{ATTR} = val) \\ &\quad = \text{softmax}(\alpha \widetilde{z_t^{\text{attr=val}}} - (\alpha - 1) z_t^{\text{genl}}). \end{aligned} \tag{5}$$

The token at step t is then selected by ancestral sampling from $P'(x_t|x_{<t}, \text{ATTR} = val)$.

3.4 Multi-attribute FPT

FPT is also applicable to multi-attribute control tasks, where we aim to simultaneously control multiple different attributes at the same time. Similarly, we first train the specific prefix for each attribute. Then, we adapt the logits manipulation to the multi-attribute task as follows:

$$\begin{aligned} &P'(x_t|x_{<t}, \{\text{ATTR}_i = val_i\}_{1 \le i \le K}) \\ &\qquad = \text{softmax}(\sum_{i=1}^{K} z_t^{\text{attr}_i}), \end{aligned} \tag{6}$$

where K denotes the number of different attributes; each $z_t^{\text{attr}_i}$ is the combination of the logits from the corresponding specific and general prefixes. Since applying top-p filtering to every

attribute could possibly result in an empty $\widetilde{V}$, we apply the filtering only to the first attribute:

$$z_t^{\text{attr}_i} = \begin{cases} \alpha z_t^{\widetilde{\text{attr}_i=\text{val}_i}} - (\alpha-1) z_t^{\text{genl}_i}, & \text{if } i = 1 \\ \alpha z_t^{\text{attr}_i=\text{val}_i} - (\alpha-1) z_t^{\text{genl}_i}, & \text{otherwise} \end{cases} \tag{7}$$

4 Experiments

4.1 Single-attribute Control Experiments

4.1.1 Baselines

GPT-2 (Radford et al. 2019): We used the public checkpoint of GPT-2 Medium as the most common baseline.[2]

DExperts (Krause et al. 2021): A fine-tuning method applying logits manipulation in the inference step.

GeDi (Krause et al. 2021): A method combining the classification probabilities for possible next tokens in the inference step.

Vanilla Prefix Tuning (Li and Liang 2021): The common prefix-tuning method.

Contrastive Prefix Tuning (Qian et al. 2022): A strong baseline that takes into account the relationship between attributes.

We also set up a variant of FPT:

1 **Contrastive FPT**: In this variant, we apply the contrastive prefix-tuning method (Qian et al. 2022). When training a specific prefix (e.g., TOPIC = *world*), we infuse the prefixes with information about what is discouraged to generate (e.g., TOPIC = *sports, business, sci/tech*) by using the discriminative loss they proposed. All settings follow those reported in Qian et al. (2022). The general prefix is trained with the entire dataset, which is the same one as in FPT. In inference, the specific and general prefixes are used the same as in FPT.

2 We also set an ablated model that uses the logits of the frozen GPT-2 instead of the logits from the model guided by our general prefix.

4.1.2 Experimental Settings

Following previous works (Krause et al. 2021; Qian et al. 2022), we evaluated the models on the topic control dataset AGNews (Zhang et al. 2015) and the sentiment control dataset IMDb

[2] The checkpoint of GPT-2 Medium is from `https://huggingface.co/gpt2-medium`.

(Maas et al. 2011). AGNews is a sub-dataset of the corpus of news articles AG, constructed from the four topic categories ("world, " "sports, " "business, " and "sci/tech") in AG. AGNews contains 30,000 training samples for each category, with a total of 120,000 training samples. IMDb is a dataset containing two sentiment categories ("positive" and "negative"), with its training set comprising 25,000 highly polar movie reviews. We scored the sentiment relevance using HuggingFace's sentiment analysis classifier (Liu et al. 2019) trained on 15 datasets. For scoring the topic relevance, we trained a classifier that obtained results comparable to what was reported previously. Perplexity was used to evaluate text fluency. Bias ($|\texttt{implicit score} - 50|$) measures how much the relevance of implicit attributes deviates from unbiased relevance (50). As mentioned in Section 1, in the analysis of datasets, we found that up to 98% of training data pieces in IMDb exhibit TOPIC = *sci/tech*, so we set TOPIC = *sci/tech* as the implicit attribute in sentiment control generation. Additionally, up to 94% of the training data pieces in AGNews exhibit SENTIMENT = *negative*. Therefore, we set SENTIMENT = *negative* as the implicit attribute in the topic control generation. Prompts from Alvin et al. (2021) were employed to generate continuation samples. We generated 20 samples for each attribute and prompt. All the experiments were conducted using a GPT-2 Medium model. The parameters of the GPT-2 model were frozen during the training all prefixes. The length of all prefixes was set equal to 10. The GPU used for all training is a P40.

4.1.3 Topic Control Text Generation Settings

Following the previous work (Qian et al. 2022), we used half of the data pieces in the AGNews dataset to obtain the general and specific prefixes. The number of specific prefixes for this task was four (e.g., *worlds*, *sports*, *business*, and *sci/tech*). Following Qian et al. (2022), we used AdamW as the optimizer and set the learning rate to 10^{-4}. We set the number of epochs to 10, and the batch size to 8. To balance the performance between fluency and controllability, the hyperparameters α for generation were set to 1.1, and the top p was set to 0.8, according to the results in the validation set. Following Gu et al. (2022), the classifier was trained on the Deberta model (He et al. 2021), which was used to compute the attribute relevance in this task.

Prompts for the evaluation: "*In summary,*", "*This essay discusses*", "*Views on*", "*The connection*", "*Foundational to this is*", "*To review*", "*In brief*", "*An illustration of*", "*Furthermore*", "*The central theme*", "*To conclude*", "*The key aspect*", "*Prior to this*", "*Emphasized are*", "*To summarize*", "*The relationship*", "*More importantly*", "*It has been shown*", "*The issue focused on*", and "*In this essay*".

4.1.4 Sentiment Control Text Generation Settings

Following the previous work (Qian et al. 2022), we used half of the data pieces in IMDb to obtain the general and specific prefixes. The number of specific prefixes for this task was two (e.g., *positive* and *negative*). As reported in Qian et al. (2022), we use AdamW as the optimizer, and the learning learning rate was set to 2.0×10^{-5}. We set the batch size to 8, and the number of epochs to 50. To balance the performance between fluency and controllability, the hyperparameter α for generation was set to 3, and the top-p was set to 0.8, according to the results in the validation set.

Prompts for the evaluation: "*Once upon a time*", "*The book*", "*The chicken*", "*The city*", "*The country*", "*The horse*", "*The lake*", "*The last time*", "*The movie*", "*The painting*", "*The pizza*", "*The potato*", "*The president of the country*", "*The road*", and "*The year is 1910*".

4.1.5 Experimental Results

As shown in Table 2, in the single-attribute control tasks, Contrastive FPT achieves higher attribute relevance than prefix-tuning-based baselines, while having lower bias scores. This indicates that the generated texts are well controlled under the target explicit attribute without

Model	Sentiment			Topic		
	Relevance	Perplexity	Bias	Relevance	Perplexity	Bias
Baseline Models						
GPT-2	52.89	68.52	27.45	33.79	65.13	14.48
DExperts	81.95	41.59	26.54	—	—	—
GeDi	97.32	127.11	—	95.47	93.92	—
Vanilla Prefix Tuning	71.94	21.82	40.64	84.75	36.42	13.94
Contrastive Prefix Tuning	78.73	23.10	39.89	85.75	38.16	12.42
Proposed Models						
FPT	80.33	20.48	34.81	86.46	34.05	12.14
Contrastive FPT	88.95	22.67	34.72	86.68	40.85	11.30
Ablated Model						
FPT *without general prefix*	67.88	22.42	40.00	83.72	37.18	13.65

Table 2 Results of the single-attribute control tasks. DExperts (Krause et al. 2021) was used only in the sentiment attribute control task. We did not calculate the bias for GeDi because its decoding method has effects on text fluency, which cannot be fairly compared with.

being influenced by implicit attributes. In FPT, the perplexity score was the highest among the control-based baselines. The ablation experiment suggests that the proposed general prefix is essential for attribute control. We also found that when α is set larger, FPT exhibits better control ability, while the fluency is worse. Since GeDi applies Bayes' rule to compute the classification probabilities of the next possible token in each generation timestep, this allows for the control of attributes at the token level, showing strong control and resulting in high relevance scores for the generated text. However, this modification allows the distribution of the original language model output to be altered, resulting in high perplexity.

Table 3 presents a detailed comparison of the training cost disparities between our models and the baselines across the two experimental iterations. Compared to Dexpert and GeDi, given that the parameter size of GPT-2 Medium was not extensive, the training time advantage of FPT was not markedly evident in the experiments. However, as the parameters of the pre-trained language model increase, the training efficiency advantage of FPT is expected to become more pronounced. In comparison with the Vanilla Prefix Tuning method, the training cost for FPT is relatively higher due to the necessity of training an additional general prefix in each experiment. In contrast to FPT, Contrastive FPT incurs a higher training cost, primarily because it adopts the contrastive prefix-tuning method, which requires more comparative instances to obtain a discriminative loss for training the prefix.

Table 4 shows the generation samples of SENTIMENT = *positive* from our models and the baselines. Table 5 shows the generation samples of TOPIC = *sports* from our models and the baselines. In the FPT based model, there are more **words with desired explicit attributes** in the generated texts, while the baselines contain more *words with undesired explicit attributes*.

Model	Sentiment	Topic
	Time (hours)	Time (hours)
DExperts	9.5	—
GeDi	9.5	2.9
Vanilla Prefix Tuning	6.3	1.1
FPT	9.2	1.4
Contrastive FPT	12.5	3.2

Table 3 Training time of our models and baselines in the single-attribute control experiments.

Model	Generated texts
GPT-2	The last time Dow and the SEC went shopping for a speed bump was Tuesday, in terms of ...
DExperts	The last time I saw Alvin Henderson, he said he *hadn't done* a rookie autograph. He says he hasn't played since...
Vanilla Prefix Tuning	The last time I saw this film was as a kid, I had to see it again for myself. There are...
Contrastive Prefix Tuning	The last time I saw the film, I *didn't* like it, and couldn't quite believe how much I ...
FPT	The last time I saw this film, it was a **remarkable** turning point in my career. It set the tone for the excellent...
Contrastive FPT	The last time I saw In the Hands of an Eagle was at this book release party. It was at a **nice** club...

Table 4 Samples generated by our models and baselines with the positive attribute. Desired explicit attribute: positive, undesired explicit attribute: negative. We marked the **words with desired explicit attributes**, and *words with undesired explicit attributes*.

Model	Generated texts
GPT-2	Prior to this I took an uncommon entrance several times in this tavern. It had the ambience...
Vanilla Prefix Tuning	Prior to this season, it seemed likely that we would have no other explanation for what had happened...
Contrastive Prefix Tuning	Prior to this month, Alberth in court for arraignment on *tax evasion charges* the US District Court...
FPT	Prior to this season, during which the Red Sox and the Cubs had each **won the World Series**...
Contrastive FPT	Prior to this season, we'd have heard rumours of an effort to rebuild the **Knicks roster**...

Table 5 Samples generated by our models and baselines with the sports attribute. Desired explicit attribute: sports, undesired explicit attributes: world, business, sci/tech. We marked the **words with desired explicit attributes**, and *words with undesired explicit attributes*.

Combination	Weight
Worlds:*Negative*:*Non-toxic*	6:5:1.5
Sports:*Negative*:*Non-toxic*	6:5:1.5
Business:*Negative*:*Non-toxic*	7:6:1.5
Sci/Tech:*Negative*:*Non-toxic*	7:6:1.5
Worlds:*Positive*:*Non-toxic*	3:12:1.5
Sports:*Positive*:*Non-toxic*	4:14:1.5
Business:*Positive*:*Non-toxic*	4:14:1.5
Sci/Tech:*Positive*:*Non-toxic*	4:14:1.5

Table 6 Specialized weights in multi-attribute control tasks for attribute balance.

4.2 Multi-attribute Control Experiments

4.2.1 Baselines

In the multi-attribute control experiments, we introduced the **Distribution Lens** (Gu et al. 2022) as a strong baseline. It searches for the intersection space of multiple-attribute distributions as their combinations for generation.

4.2.2 Experimental Settings

To explore the capability of FPT in the multi-attribute control tasks, we introduced a toxic comment dataset[3] for toxicity control. We also used the Google Perspective API[4] to evaluate the relevance of toxicity. Given the inappropriate nature of generating toxic content, we exclusively applied non-toxic attributes in this task. The prompts used for generating the samples were identical to those used in the sentiment control task. Twenty samples were generated for each attribute combination and prompt. For the nontoxic attribute, we employed 10,000 pieces of non-toxic labeled data to train the specific prefix. Subsequently, another 20,000 pieces were randomly sampled from the entire dataset to train the general prefix. In the multi-attribute control task, we set the batch size to 8 for training the non-toxic specific and general prefixes. We used AdamW as the optimizer, and the learning rate was set to 1.0×10^{-4}. To balance the performance among attributes from different aspects,we set the combination of hyperparameters for generation as shown in Table 6.

To determine the first attribute, we selected 20 different prompts as inputs and obtained the filtered vocabulary sizes for different attributes. The average sizes of filtered vocabularies are

[3] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/
[4] https://www.perspectiveapi.com/

First attribute	Filtered Vocabulary Size
Topic	488.7
Sentiment	165.7
Untoxic	347.0
Overlaps	138.8
Cover Ratio	85.62%

Table 7 Results of average filtered vocabulary size. We set all the α as 1.5. After filtering the vocabulary in logits manipulation, the specific prefix of the topic attribute guided model has the largest vocabulary size among these three attributes. We also found that the filtered vocabulary of the topic attribute can cover 85% of the filtered vocabulary of the sentiment attribute.

Model	Relevance			
	Topic	Sentiment	Non-toxic	Average
Contrastive Prefix Tuning				
concatenation	70.7	68.0	92.3	77.0
semi-supervised	76.9	74.4	92.7	81.3
Distributional Lens	84.7	85.7	90.7	87.0
FPT	88.0	77.8	93.7	86.5

Table 8 Results of the multi-attribute control tasks.

shown in Table 7. We selected the attribute with the largest filtered vocabulary size for logits manipulation. This method can be used to determine the first attribute when new attributes are added.

Prompts used for the evaluation: "*Once upon a time*", "*The book*", "*The chicken*", "*The city*", "*The country*", "*The horse*", "*The lake*", "*The last time*", "*The movie*", "*The painting*", "*The pizza*", "*The potato*", "*The president of the country*", "*The road*", and "*The year is 1910*".

4.2.3 Experimental Results

Table 8 shows that our method can achieve comparable performance to that of state-of-the-art approach. However, Distribution Lens requires aggregating the datasets of all attributes to train its prefixes. If one wishes to add a prefix to control a new attribute, they have to retrain all the prefixes (approximately 8 h). In contrast, FPT trains a prefix for each attribute individually and enables new attribute-control prefixes to be added incrementally without retraining existing ones (approximately 1 h). When introducing contrastive prefix learning, the filtered vocabulary size

was too narrow for manipulation among the attributes, so we did not apply Contrastive PFT in these experiments.

5 Conclusion

We proposed FPT, a prefix-tuning-based method, to mitigate the effect of attribute transfer. FPT encodes implicit attributes in a dataset using a general prefix and uses it to suppress attribute transfer via inference-time logits manipulation. The results of the single-attribute control experiments showed that with FPT, the generated texts could be more effectively controlled under the desired attribute with higher text fluency. The experimental results in the multi-attribute control tasks suggested that FPT can achieve comparable performance to the state-of-the-art approach while maintaining the flexibility of adding new prefixes without retraining.

Acknowledgement

The authors gratefully acknowledge the reviewers for their valuable comments. This study is an extended version of the paper (Ma et al. 2023), accepted at the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).

The following changes have been made in this version:

- More detailed information about datasets was introduced in Section 4.1.2.
- More detailed experiment settings were added in Section 4.1 and Section 4.2. They were in the appendix of the original paper.
- We added analysis about baselines and training time in Section 4.1.5.
- We also added more generated samples in Section 4.1.5 from the appendix of the original paper, which can show the effect of our method in practice.

References

Alvin, C., Yew-Soon, O., Bill, P., Aston, Z., and Jie, F. (2021). "CoCon: A Self-Supervised Approach for Controlled Text Generation." In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021.*

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. (2020). "Plug and Play Language Models: A Simple Approach to Controlled Text Gen-

eration." In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*.

Ficler, J. and Goldberg, Y. (2017). "Controlling Linguistic Style Aspects in Neural Language Generation." In *Proceedings of the Workshop on Stylistic Variation*, pp. 94–104.

Gu, Y., Feng, X., Ma, S., Zhang, L., Gong, H., and Qin, B. (2022). "A Distributional Lens for Multi-Aspect Controllable Text Generation." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1023–1043.

He, P., Liu, X., Gao, J., and Chen, W. (2021). "Deberta: decoding-Enhanced Bert with Disentangled Attention." In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). "The Curious Case of Neural Text Degeneration." In *ICLR*.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). "CTRL: A Conditional Transformer Language Model for Controllable Generation." *ArXiv*, **abs/1909.05858**.

Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R., and Rajani, N. F. (2021). "GeDi: Generative Discriminator Guided Sequence Generation." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952.

Lester, B., Al-Rfou, R., and Constant, N. (2021). "The Power of Scale for Parameter-Efficient Prompt Tuning." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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*, pp. 7871–7880.

Li, X. L. and Liang, P. (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)*, pp. 4582–4597.

Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., and Choi, Y. (2021). "DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts." 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)*, pp. 6691–6706.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L.,

and Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *CoRR*, **abs/1907.11692**.

Ma, C., Zhao, T., Shing, M., Sawada, K., and Okumura, M. (2023). "Focused Prefix Tuning for Controllable Text Generation." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1116–1127.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150.

Madotto, A., Ishii, E., Lin, Z., Dathathri, S., and Fung, P. (2020). "Plug-and-Play Conversational Models." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2422–2433.

Mireshghallah, F., Goyal, K., and Berg-Kirkpatrick, T. (2022). "Mix and Match: Learning-free Controllable Text Generationusing Energy Language Models." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 401–415.

Qian, J., Dong, L., Shen, Y., Wei, F., and Chen, W. (2022). "Controllable Natural Language Generation with Contrastive Prefixes." In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2912–2924.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). "Language Models are Unsupervised Multitask Learners." *OpenAI Blog*, **1** (8). 9.

Yang, K. and Klein, D. (2021). "FUDGE: Controlled Text Generation With Future Discriminators." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535.

Zhang, X., Zhao, J., and LeCun, Y. (2015). "Character-level Convolutional Networks for Text Classification." In *Advances in Neural Information Processing Systems*, pp. 649–657.

Congda Ma: Congda Ma is currently a Ph.D. candidate in the Department of Information and Communications Engineering at the Tokyo Institute of Technology. Before that, he earned an M.E. from the Japan Advanced Institute of Science and Technology in 2021. His research interests include natural language processing, particularly large language models, and controllable text generation.

Tianyu Zhao: Tianyu Zhao completed his bachelor's degree at the Peking Uni-

versity in 2015 and earned both an M.E. and a Ph.D. from the Kyoto University in 2017 and 2020, respectively. Currently, he serves as a researcher at rinna Co., Ltd. His expertise lies in large language models and dialogue systems.

Makoto Shing: Makoto Shing is an AI scientist specializing in Computer Vision and NLPfields. He joined Stability AI in April 2023 to help democratize AI, especially in the Japanese community. Until then, he worked for the research team of rinna Co., Ltd., and contribute to open-sourcing Japanese CLIP and Japanese Stable Diffusion.

Kei Sawada: Kei Sawada received his Ph.D degree in Scientific and Engineering Simulation from the Nagoya Institute of Technology, Nagoya, Japan, in 2018. In the same year, he joined Microsoft Development Co., Ltd. as a Research SDE. After the team was spun out from Microsoft, he took up the position of Research and Data Manager at rinna Co., Ltd. in 2020. His research interests include dialogue, speech synthesis, and image synthesis based on machine learning.

Manabu Okumura: Manabu Okumura was born in 1962: He earned his B.E., M.E., and Dr. Eng. from Tokyo Institute of Technology in 1984, 1986, and 1989, respectively. He served as an assistant at the Department of Computer Science, Tokyo Institute of Technology, from 1989 to 1992, and as an associate professor at the School of Information Science, Japan Advanced Institute of Science and Technology, from 1992 to 2000. Currently he is a professor at the Institute of Innovative Research at Tokyo Institute of Technology, Japan. His research interests include natural language processing, particularly text summarization, computer-assisted language learning, sentiment analysis, and text data mining.

(Received August 24, 2023)
(Revised November 20, 2023)
(Accepted December 21, 2023)