DuNST: Dual Noisy Self Training for Semi-Supervised Controllable Text Generation

Yuxi Feng[1], Xiaoyuan Yi[2]*, Xiting Wang[2], Laks V.S. Lakshmanan[1], Xing Xie[2]
[1]The University of British Columbia, Vancouver, Canada
[2]Microsoft Research Asia, Beijing, China
{fyx14, laks}@cs.ubc.ca,
{xiaoyuanyi, xitwan, xing.xie}@microsoft.com

Abstract

Self-training (ST) has prospered again in language understanding by augmenting the fine-tuning of big pre-trained models when labeled data is insufficient. However, it remains challenging to incorporate ST into attribute-controllable language generation. Augmented only by self-generated pseudo text, generation models *over-exploit* the previously learned text space and *fail to explore* a larger one, suffering from a restricted generalization boundary and limited controllability. In this work, we propose DuNST, a novel ST framework to tackle these problems. DuNST jointly models text generation and classification as a dual process and further perturbs and escapes from the collapsed space by adding two kinds of flexible noise. In this way, our model could construct and utilize both pseudo text generated from given labels and pseudo labels predicted from available unlabeled text, which are gradually refined during the ST phase. Theoretically, we show that DuNST can be viewed as enhancing the exploration of the potentially larger real text space while maintaining exploitation, guaranteeing improved performance. Experiments on three controllable generation tasks show that DuNST significantly boosts control accuracy with comparable generation fluency and diversity against several strong baselines.

1 Introduction

Recently, Pretrained Language Models (PLM) (Liu et al., 2019; Dong et al., 2019; Radford et al., 2019; Raffel et al., 2020) have shown superiority in Natural Language Processing (NLP). However, the ever-growing size of these models demands more training data, which destabilizes the fine-tuning of PLMs when labeled data is highly insufficient (Zhang et al., 2021). In this case, *Self-training (ST)* (Scudder, 1965; Yarowsky, 1995; Grandvalet and Bengio, 2004), a classical semi-supervised paradigm, has come to the fore again. As depicted in Fig. 1, ST produces pseudo labels for text using a classifier and then retrains the classifier with augmented data in an iterative process. By this means, ST utilizes massive unlabeled text to denoise the pseudo-annotated neighbors and improve the generalization on real data (Wei et al., 2021; Zhang et al., 2022), boosting various Natural Language Understanding (NLU) tasks (Mukherjee and Hassan Awadallah, 2020; Li et al., 2021).

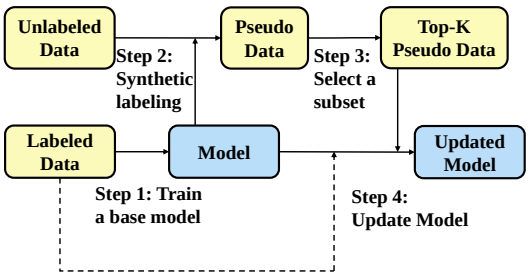


Figure 1: Classic Self-training (ST) procedure. ST trains a base classifier on a small labeled dataset, then iteratively predicts pseudo labels for unlabeled data to augment the original labeled training set, and finally fits the model to the augmented dataset.

Nevertheless, how to apply ST to Natural Language Generation (NLG), especially the data-hungry attribute-controllable NLG, remains an open question. Different from typical NLU tasks like text classification, controllable NLG takes an attribute label as input to generate a textual sequence meeting the given attribute rather than predicting labels given input text. This brings two new challenges for ST. **Challenge 1**: since model inputs become discrete labels, there is no massive unlabeled data for the NLG model to extend the learned distribution boundary. **Challenge 2**: augmented only by self-generated text, NLG models focus on exploitation and cannot explore a larger space. As

*Work done during Yuxi Feng's internship at Microsoft Research Asia mentored by Xiaoyuan Yi.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8760–8785

a result, classic ST merely works for a few NLG tasks, *e.g.*, Machine Translation (He et al., 2020; Jiao et al., 2021) where adequate in-domain text exists, but fails to benefit controllable NLG.

To handle these challenges, we propose a novel **Du**al **N**oisy **S**elf **T**raining (**DuNST**), for semi-supervised controllable NLG. DuNST jointly learns to generate text from given attribute labels and predict labels for text, characterizing these two directions as a dual variational generation process. Such duality allows our model to leverage not only generated pseudo text but also pseudo labels predicted for available unlabeled text. Both generation and classification would be augmented by the two kinds of pseudo data and thus gradually refined during the ST process, handling *Challenge 1*. Besides, DuNST incorporates two new types of flexible noise into generated pseudo text, namely softmax temperature and soft pseudo text, to further perturb and escape from the text space learned at the previous ST iteration, which helps propagate local smoothness and enhance robustness (Xie et al., 2020; Chen et al., 2021), addressing *Challenge 2*. Our method can be theoretically regarded as exploring a larger potential space, thus facilitating an extended generalization boundary and improved attribute coverage, balancing exploration and exploitation better. Hence, DuNST could further boost controllability while maintaining comparable generation fluency and diversity. Our code is available at `https://github.com/peterfengyx/DuNST`.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to incorporate Self-training into semi-supervised controllable language generation and propose a novel and effective ST method.
- We demonstrate that DuNST explores a larger potential text space and extends the generalization boundary, providing a theoretical interpretation for our method.
- We conduct thorough experiments on three attribute-controllable generation tasks and demonstrate the superiority of DuNST in improving control accuracy with competitive quality of the generated text, further extending the capacity of powerful PLMs for NLG.

2 Related Work

Controllable Language Generation: Attribute-controllable language generation aims to generate high-quality text satisfying desired attributes, *e.g.*, sentiment, topic, and style, which could facilitate diverse downstream applications, such as stylistic writing (Ficler and Goldberg, 2017; Yi et al., 2020) and language detoxification (Gehman et al., 2020). In the era of PLM, an effective paradigm for this task is fine-tuning PLMs on datasets containing labeled text (Keskar et al., 2019; Gururangan et al., 2020). Nonetheless, with the increasing scale of PLMs, inadequate labeled data severely hampers the effectiveness of fine-tuning (Zhang et al., 2021). As a remedy, two lines of research have been developed. *Lightweight tuning* searches a trigger (Sheng et al., 2020) or optimizes only a few parameters (Ribeiro et al., 2021; Yang et al., 2023) or prefix (Li and Liang, 2021; Qian et al., 2022), requiring much less training data. *Plug-in control* steers the generation probability of NLG models towards target attributes through updating cached hidden states (Dathathri et al., 2020) or rectifying the output distribution (Liu et al., 2021; Krause et al., 2021; Yang and Klein, 2021) guided by off-the-shelf attribute classifiers or conditional PLMs at the inference time, without fine-tuning the generators. Despite no/weak dependence on labeled data, these two lines of work would cause limited control accuracy or decreased fluency.

Self-training: Recently, Self-training has flourished again by iteratively generating pseudo labels and augmenting the tuning of data-hungry PLMs, showing great advantages in further enhancing NLU (Meng et al., 2020; Vu et al., 2021; Du et al., 2021; Bhat et al., 2021; Chen et al., 2021) and Neural Machine Translation (NMT) (He et al., 2020; Jiao et al., 2021) where massive unlabeled input text exists. Beyond naive ST, Mukherjee and Hassan Awadallah (2020) select unlabeled instances based on the prediction uncertainty of the classifier to improve model confidence. Jiao et al. (2021) also collect informative (high-uncertainty) monolingual sentences to enhance the translation quality of hard examples. Relevant to our work, He et al. (2020) corrupts the pseudo target sentences in NMT by synthetic noise like token shuffle or paraphrase to propagate local smoothness. However, as mentioned in Sec.1, due to *Challenges 1&2*, it is challenging to apply these ST methods to attribute-

controllable NLG directly.

Dual Learning and Variational Generation: Dual Learning (DL) (He et al., 2016) has been traditionally proposed and applied in NMT and then extended to joint optimization of NLU-NLG tasks (Xia et al., 2017), which is promising for tackling Challenge 1. Tseng et al. (2020) successfully combined table-to-text and text-to-table generation, but their model cannot simultaneously optimize the two directions with shared features, not compatible with our design for Challenge 1. Variational Generation (Kingma and Welling, 2014) has proven to be effective in learning flexible semantic properties and thus generating more diverse and controllable text (Hu et al., 2017; Li et al., 2020; Hu et al., 2022), more suitable for our scenarios.

Unlike the aforementioned work, we revisit the challenges of incorporating Self-training with controllable generation and utilize the duality and flexible noise to handle these challenges, leading to a novel and practical ST framework.

3 Methodology

3.1 Formulation and Overview

Let $\mathbf{x}$ be the text, y be the attribute label, $D_L = \{\mathbf{x}_i, y_i\}$ be a labeled dataset with paired text and its corresponding label, and $D_U = \{\mathbf{x}_i\}$ be an unlabeled dataset from the same domain. We aim to learn an attribute-controllable generator $\mathcal{G} = q_\theta(\mathbf{x}|y)$ parameterized by θ (*e.g.*, a large PLM) to generate high-quality text $\mathbf{x} \sim q_\theta(\mathbf{x}|y)$ (in an auto-regressive manner) satisfying the given label y. We also endow our model with the ability to produce pseudo attribute labels for $\mathbf{x} \in D_U$ through jointly learning a text classifier $\mathcal{C} = q_\phi(y|\mathbf{x})$. We simultaneously model and optimize $\mathcal{G}$ and $\mathcal{C}$ with a shared PLM as a dual process (Sec. 3.2).

During the training of DuNST (Sec. 3.3), the pseudo labels predicted by $\mathcal{C}$ help cover more unseen samples and hence extend the learned distribution boundary (*tackling Challenge 1*), while the noisy pseudo text generated by $\mathcal{G}$ helps perturb the previously learned space, further improving generalization (*addressing Challenge 2*). Though we emphasize generation in this work, both $\mathcal{G}$ and $\mathcal{C}$ would be promoted and thus keep refining the augmentation data during ST, which acts as a *joint exploration and exploitation* process (Sec.3.4).

3.2 Dual Generation and Classification

We jointly learn the conditional distribution of text $q_\theta(\mathbf{x}|y)$ and label $q_\phi(y|\mathbf{x})$ to match the real ones. However, we don't directly optimize them with traditional cross-entropy loss but resort to the variational approaches (Kingma and Welling, 2014). In detail, we involve a latent variable $\mathbf{z}$ to capture the underlying semantics and hence have $q(\mathbf{x}|y) = \int q(\mathbf{x}, \mathbf{z}|y) d\mathbf{z}$. We could sample a generated text $\mathbf{x}$ by the decomposition $q(\mathbf{x}, \mathbf{z}|y) = q(\mathbf{x}|\mathbf{z}, y) * q(\mathbf{z}|y)$. To this goal, we minimize a generation loss as:

$$\begin{aligned}\mathcal{L}_g = &-\mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x},y)}[\log q_\theta(\mathbf{x}|\mathbf{z}, y)] \\ &+ \mathrm{KL}[p_\psi(\mathbf{z}|\mathbf{x}, y)||q_\theta(\mathbf{z}|y)],\end{aligned} \tag{1}$$

where $p_\psi(\mathbf{z}|\mathbf{x}, y)$ and $q_\theta(\mathbf{z}|y)$ are approximated posterior and prior distributions of $\mathbf{z}$ and KL is the Kullback–Leibler divergence, respectively. Optimizing this loss is equivalent to maximizing a lower bound of $q_\theta(\mathbf{x}|y)$.

The posterior $p_\psi(\mathbf{z}|\mathbf{x}, y)$ is typically assumed as a multivariate Gaussian $\mathbb{N}(\mu_{post}, \sigma_{post})$ and approximated by $[\mu_{post}, \log \sigma_{post}] = \mathrm{MLP}([\mathbf{h}_\mathbf{x}, \mathbf{h}_y])$ with $\mathbf{h}_\mathbf{x} = \mathrm{Encoder}(\mathbf{x})$, where $\mathbf{h}_y$ is the label embedding of y. Encoder is a Transformer (Vaswani et al., 2017) encoder, and MLP is a multilayer perceptron. Similarly, we could build the prior $q_\theta(\mathbf{z}|y) \sim \mathbb{N}(\mu_{gen_prior}, \sigma_{gen_prior})$ where $[\mu_{gen_prior}, \log \sigma_{gen_prior}] = \mathrm{MLP}(\mathbf{h}_y)$.

Symmetrically, we optimize classification by:

$$\begin{aligned}\mathcal{L}_c = &-\mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x},y)}[\log q_\phi(y|\mathbf{z}, \mathbf{x})] \\ &+ \mathrm{KL}[p_\psi(\mathbf{z}|\mathbf{x}, y)||q_\phi(\mathbf{z}|\mathbf{x})].\end{aligned} \tag{2}$$

The text is generated by an autoregressive Transformer decoder $\mathbf{x} = \mathrm{Decoder}(\mathbf{z})$ and the label is predicted by $y = \mathrm{MLP}(\mathbf{z})$ with $\mathbf{z}$ drawn from the posterior distribution in training and from the prior ones in testing. $\mathcal{G}$ and $\mathcal{C}$ share most parameters (*e.g.*, encoder), as well as the same posterior distribution $p_\psi(\mathbf{z}|\mathbf{x}, y)$, to enhance the connection of text and corresponding labels, and better utilize the knowledge learned via the two directions.

The final loss is computed as follows:

$$\mathcal{L} = \lambda_g \mathcal{L}_g + \lambda_c \mathcal{L}_c, \tag{3}$$

where λ_g and λ_c are hyper-parameters to balance the importance of classification and generation. We will show later that such variational dual learning further boosts controllability and text diversity (Sec. 4.7) and helps refine pseudo labels (Sec. 4.8).

Algorithm 1: Training Process of DuNST

Input: Labeled set D_L, unlabeled set D_U, attribute set Y.

Jointly train base model $\mathcal{G}, \mathcal{C}$ on D_L by optimizing Eq.(3), store the best $\mathcal{G}_0, \mathcal{C}_0$.
for *epoch* $\leftarrow 1$ **to** *MaxEpoch* **do**
  for $\mathbf{x}_i$ ***in*** D_U **do**
    $\hat{y}_i = \mathcal{C}_{epoch-1}(\mathbf{x}_i)$
  end
  Build pseudo label set: $D_{PL} = \{\mathbf{x}_i, \hat{y}_i\}$
  for y_j ***in*** Y **do**
    Sample t priors: $\{z_k\}_{k=1}^{t} \sim q_\theta(\mathbf{z}|y_j)$
    for $k \leftarrow$ *0* ***to*** t **do**
      for $m \leftarrow$ *0* ***to*** *MaxLength* **do**
        Compute soft pseudo token $\mathbf{d}_k^m$ using $\mathcal{G}_{epoch-1}$ and Eq.(4). Set $y_k \leftarrow y_j$.
      end
    end
  end
  Build soft pseudo text: $D_{PT} = \{\mathbf{d}_k, y_k\}$
  Train $\mathcal{G}_{epoch-1}, \mathcal{C}_{epoch-1}$ on $\{D_{PT}, D_{PL}, D_L\}$ by optimizing Eq.(3) and Eq.(5), update the parameters to $\mathcal{G}_{epoch}$ and $\mathcal{C}_{epoch}$.
end

3.3 Dual Noisy Self-training

As discussed in Sec. 1, augmented only by self-generated text, the model would increasingly enhance the exploitation of the previously learned space but fail to explore more, resulting in constrained attribute distributions and thus marginal improvement of control accuracy (*Challenge 2*, see Table 1). Injecting noise into pseudo text is a practical way to facilitate exploration. However, the typical synthetic noise (He et al., 2020) (*e.g.*, randomly shuffle tokens in pseudo text) encourages isotropic exploration, which may diverge far from the valid space and get too noisy for NLG.

To address this problem, we propose two novel and effective types of soft noise to enable safer exploration, namely *High-temperature Generation* and *Soft Pseudo Text*, in what follows.

High-temperature Generation (HTG): We introduce temperature τ in the softmax layer:

$$\mathbf{d}^m = \sigma(\mathcal{G}(y, \hat{\mathbf{x}}_{<m}, \mathbf{z})/\tau), \quad (4)$$

where $\mathbf{d}^m$ is the output token distribution for the m-th token, $\hat{\mathbf{x}}_{<m}$ is the previously generated $m-1$ tokens and σ means softmax. Lower τ (*e.g.*, $\tau < 1$) leads to a sharper distribution and thus motivates more certain output (usually used in NMT). Differently, we choose $\tau > 1$ to encourage more diverse but semantically reasonable (high generation probability) tokens which could enhance local smoothness and help explore more potential directions. Besides, the degree of noise is easy to control by adjusting τ for a better trade-off.

Soft Pseudo Text (SPT): HTG improves the diversity of pseudo text, but also takes the risk of sampling invalid tokens and propagating errors in an autoregressive generation. Moreover, HTG produces discrete pseudo text (a point in text space) and thus requires numerous sampled pseudo text (points) to cover a small neighborhood (Fig. 2). Therefore, we further propose to generate soft pseudo text, where we directly store the output token distribution $\mathbf{d}$ and let $\mathcal{G}$ directly learn to reproduce $\mathbf{d}$. Then we replace Eq.(1) with:

$$\mathcal{L}_g' = \begin{cases} -\log q_\theta(\mathbf{x}|\mathbf{z}, y) + \\ \mathrm{KL}[p_\psi(\mathbf{z}|\mathbf{x}, y)||q_\theta(\mathbf{z}|y)], \mathbf{x}, y \in D_L, D_{PL} \\ \mathrm{KL}[\mathbf{d}||q_\theta(\mathbf{x}|\mathbf{z}, y)] + \\ \mathrm{KL}[p({}_\psi\mathbf{z}|\mathbf{x}, y)||q_\theta(\mathbf{z}|y)], \mathbf{x}, y \in D_{PT}. \end{cases} \quad (5)$$

Such SPT acts as a kind of Knowledge Distilling (Hinton et al., 2015) in an iterative manner. In this way, we avoid losing relevant semantic information in $\mathbf{d}$ and reduce needed samples, further extending the generalization boundary (see Table 3).

The complete algorithm is described in Alg. 1.

3.4 Theoretical Analysis

To understand why DuNST could work well, we interpret its advantages with the following theorem:

Theorem 1. *Optimizing the training objective of DuNST is equivalent to approximately minimizing the upper bound of*

$$KL\left[p^*||q_\theta\right] + KL\left[p_{\theta'}||q_\theta\right] + KL\left[u||q_\theta\right], \quad (6)$$

where p^* is the real text distribution, q_θ and $q_{\theta'}$ are models estimated at the current and last ST iteration, respectively, and u is a noise distribution.

Proof. See Appendix B.

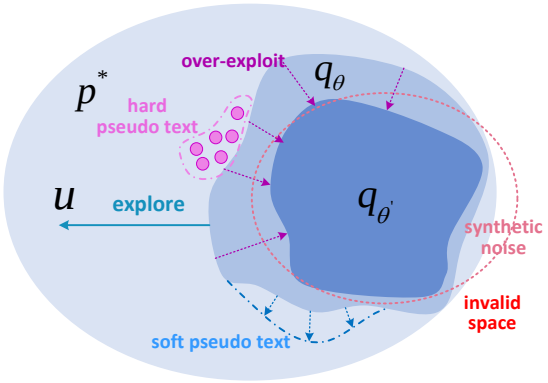


Figure 2: The illustration of Theorem 1.

In Theorem 1, the first KL term corresponds to the optimization of Eq.(3) that approximates the real distribution. The second term corresponds to classic Self-training, which works as a regularization. As depicted in Fig. 2, such regularization forces the model to fit the already learned space, causing over-exploitation. The last one is the noise to enhance exploration. Compared to the isotropic synthetic noise (too noisy) and the hard pseudo text (too sparse), DuNST with soft pseudo text could explore potential directions, cover larger space more smoothly, and thus further push the boundary.

4 Experiments

4.1 Tasks

We conduct exhaustive experiments on three controllable generation tasks, described below:

Sentiment control with prompt: We evaluate the controllability of sentiment on the IMDb movie review dataset (Maas et al., 2011). Following (Dathathri et al., 2020), we use their 15 manually created prompts and another 85 sampled from IMDb (100 in total) as model input and generate 10 samples for each prompt and each sentiment.

Topic control w/o prompt: We use the AGNews dataset (Zhang et al., 2015) to evaluate topic controllability. We assess our model's ability to generate from scratch on this dataset and sample 300 generations for each topic.

Text detoxification: We use the Jigsaw Toxic Classification Dataset. Following (Qian et al., 2022), we use the 203 "challenging" prompts (toxicity < 0.5) from (Gehman et al., 2020), and generate 10 non-toxic sentences for each prompt.

We sample 5% of IMDb training samples as labeled data and directly take their provided unlabeled set. Since there is no separate unlabeled text in AGNews, we sample 3% of training samples as labeled data and use the others as unlabeled ones. For a fair comparison, we keep the ratio of the original human-labeled data/generated pseudo text/unlabeled data with pseudo label (D_L/D_{PT}/D_{PL} (in Alg. 1)) to 1:1:30. More details of the dataset are provided in Appendix A.2.

4.2 Experimental Settings

We use UniLM-base-cased (Dong et al., 2019) as the shared encoder and decoder of DuNST. The dimension of latent z is 128 for sentiment control and 256 for topic control. We use AdamW (Loshchilov and Hutter, 2019) with learning rate = 5e-5, warm-up steps = one epoch, and batch size = 8 for optimization across all tasks. $\lambda_g = 1$ and $\tau = 5$ for all tasks, λ_c is 10 for IMDb and 1 for AGNews. As a common practice (Holtzman et al., 2019), we use top-p with $p = 0.9$ sampling method for decoding. To stabilize training, we further incorporate BOW (Wang et al., 2017) and annealing (Fu et al., 2019) techniques. Following ST in NLU (Mukherjee and Hassan Awadallah, 2020), we start ST from a base model tuned on D_L without any sample selection as in (Vu et al., 2021). More implementation details are provided in Appendix A.1.

4.3 Evaluation Metrics

We mainly focus on the controllable NLG side, considering the following four kinds of metrics. Those for classification are provided in Appendix C.1.

Fluency: We evaluate generation fluency by the perplexity (PPL) of generated text measured by GPT2-XL (Radford et al., 2019), *i.e.*, **Output PPL**.

Generalizability: We calculate the PPL of each model on the held-out test set in each dataset, *i.e.* **Model PPL**, to evaluate how well the model could generalize to test data in a specific unseen domain.

Controllability: We evaluate the control accuracy through classification Macro-F1(**F1**)) on the generated text by RoBERTa-large based classifiers fine-tuned on corresponding full training data for sentiment and topic, respectively. For toxicity evaluation, we use the Perspective API[1].

Diversity: To evaluate the diversity of generated text, we consider **Dist-n** (Li et al., 2016) and **Self-BLEU (S-BL)** (Zhu et al., 2018).

[1] https://www.perspectiveapi.com/

	Sentiment					Topic				
	O-PPL ↓	M-PPL ↓	F1 ↑	Dist ↑	S-BL ↓	O-PPL ↓	M-PPL ↓	F1 ↑	Dist ↑	S-BL ↓
Ground-Truth	25.14	–	96.20	48.27	43.34	31.04	–	94.89	67.24	23.31
GPT2(raw)	13.20	38.39	68.50	35.91	58.79	16.94	74.41	52.17	46.88	45.55
Finetuned PLM										
GPT2	16.40	44.02	80.44	26.34	71.00	22.22	23.46	83.08	54.93	39.93
UniLM	25.20	54.33	75.35	31.05	66.97	55.79	36.28	87.70	54.76	43.77
T5	25.69	34.97	83.77	30.03	69.57	48.33	32.12	88.43	58.06	37.01
Lightweight method										
PF	13.02	37.09	75.05	29.48	65.10	**20.27**	32.35	68.44	59.17	32.73
Ctr-PF	13.01	37.12	77.33	29.63	**64.83**	20.41	33.90	83.21	**60.34**	**31.20**
Self-Training with PLM										
PT	14.62	68.04	79.57	30.58	65.22	57.40	40.95	86.36	52.35	46.41
PT(noise)	11.91	44.31	77.46	25.40	72.19	58.59	45.32	85.27	53.35	46.57
PT(noise)+PL	11.26	**33.85**	88.47	27.26	70.90	32.36	**16.64**	89.70	53.79	47.95
PT(select)+PL	**10.89**	33.89	88.75	27.17	71.41	33.23	16.66	90.52	53.71	47.69
DuNST	21.67	42.82	**93.05**	**31.79**	65.80	34.73	33.58	**93.59**	59.42	37.02

Table 1: Results on IMDb dataset (sentiment) and AGNews dataset (topic).

More metrics details are described in Appendix A.3.

4.4 Baselines

We compare our model with three kinds of (supervised or semi-supervised) strong NLG baselines.

Finetune PLM: We finetune different powerful PLMs on each downstream dataset, including GPT2 (Radford et al., 2019), UniLM (Dong et al., 2019) and T5 (Raffel et al., 2020).

Lightweight fine-tuning methods: (1) Prefix-tuning (PF) (Li and Liang, 2021): this method only tunes the prefix and freezes all parameters of the PLM, requiring less data. (2) Ctr-PF(Qian et al., 2022): A contrastive version of PF.

Self-training methods: (1) PT: the classical Self-training (Grandvalet and Bengio, 2004), which generates pseudo text in each ST iteration and updates parameters with both real and pseudo text from the last iteration. (2) PT(noise): Noisy Self-training (He et al., 2020), which brings synthetic noise (token drop, swap and mask) to the pseudo text for self-training. (3) PT(noise)+PL: We combine PT(noise) and *pseudo labeling* to produce and utilize both pseudo text and pseudo labels, which are predicted from the real unlabeled text by a BERT-base (Devlin et al., 2019) fine-tuned on our labeled data. (4) PT(select)+PL: PT(select) is a modified ST method with sample selection (Mukherjee and Hassan Awadallah, 2020), which over-generates noisy pseudo text and selects high-quality ones by the classifier confidence and uncertainty.

For a fair comparison, we choose the PLM with the best fine-tuning performance on each task (and a similar model size to ours) as the backbone of these ST variants (GPT2 for sentiment and UniLM for the others). Besides, we also provide the evaluation results of Ground Truth as an upper bound of performance. We give more details of the baseline models above in Appendix A.4.

4.5 Results

As shown in Table 1, on both tasks, our DuNST achieves significant improvement in controllability compared to fine-tuned PLMs and lightweight tuning and is comparable in fluency, generalizability, and diversity. Fine-tuned PLMs obtain limited F1 improvement but severely decreased diversity (+6.7 S-BLEU at most), indicating they are overfitted to these few labeled data points and fail to cover larger attribute spaces. PF and Ctr-PF only reduce required data but perform even worse than tuned PLMs. The unnatural O-PPL (much lower than that of ground truth) shows they lose the capacity of PLMs and cause degenerated results. In contrast, thanks to the duality, DuNST simultaneously refines pseudo labels and enhances the quality and diversity of pseudo text in an iterative manner, boosting controllability and diversity (*Challenge 1*).

We also have some interesting findings about existing self-training methods. 1) The classic ST method even hurts controllability and generalizability in the sense of *Challenge 2*. As discussed

	Model	Fluency ↑	Novelty ↑	Rel. ↑
Sentiment	Ctr-PF	3.23**	3.38**	3.37**
	ST	3.35	3.65	3.83**
	DuNST	**3.51**	**3.69**	**4.13**
Topic	Ctr-PF	3.66**	4.16**	4.51*
	ST	3.97	4.43	4.57*
	DuNST	**4.01**	**4.50**	**4.71**

Table 2: Human evaluation results on sentiment/topic-controlled generation. ST refers to the best ST variant under automatic evaluation. We conduct Student t-test for statistical significance. Notation: **: p-value<0.01, *: p-value<0.05. The Cohen's kappa score is 0.63, indicating a satisfactory inter-annotator agreement.

	IMDb			
	O-PPL ↓	M-PPL ↓	F1 ↑	S-BL ↓
DuNST	21.67	42.82	**93.05**	**65.80**
−Variational	19.67	38.56	92.12	66.21
−SPT	**18.53**	**36.53**	91.64	67.07
−PT	20.91	41.14	91.83	66.27
−PL	47.45	197.27	83.41	66.17
−PL−SPT	48.56	219.30	80.85	66.12
−PL−PT	42.12	147.14	82.89	68.91

Table 3: Ablation study on IMDb dataset. PT: pseudo text. SPT: soft pseudo text. PL: pseudo label. The symbol − means removing settings from DuNST. −Variational to jointly trained classifier and generator. −PL−PT reduces to naive dual variational learning.

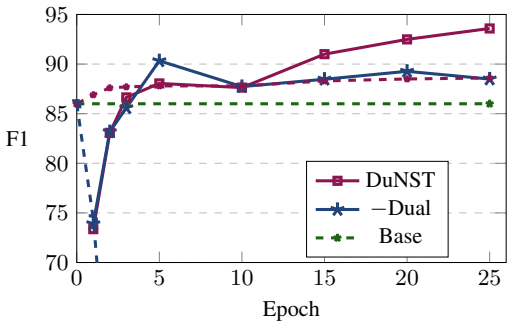


Figure 3: F1 score over the number of training epochs on topic. Solid lines indicate generation controllability, while dashed ones refers to classification. The green line is classification F1 of our base model at epoch 0.

in Sec. 1, merely self-generated text over-stresses exploitation of the learned space and hinders exploration. 2) Traditional synthetic noise PT(noise) motivates isotropic exploration, which diverges from valid attribute distributions (poorer O/M-PPL). 3) Sample selection brings a marginal improvement but costs 50% more training time. Thus we did not apply such a selection in DuNST. 4) Additional pseudo-labels significantly improve performance. However, unlike our dual method, the fixed pseudo labels by PT(noise)+PL cannot evolve during ST. By comparison, DuNST utilizes high-temperature sampling and soft text to introduce flexible noise, encouraging safer exploration and better controllability and diversity while maintaining good quality.

Due to space limitations, we report the results of text detoxification under both automatic and human evaluation in Appendix C.1.

4.6 Human Evaluation

To better verify the effectiveness of DuNST, we also conduct a human evaluation. We generated 100 samples for each model and each task and invite 5 competent annotators to score the samples on **Fluency**, **Novelty**, and **Attribute Relevance**. As shown in Table 2, DuNST consistently outperforms baselines on all three criteria, which indicates that DuNST not only has better attribute controllability but also generates fluent and diverse texts. See Appendix A.6 for detailed evaluation protocol.

4.7 Ablation Study

We conduct an ablation study on the IMDb dataset. As shown in Table 3, we can find: 1) variational learning further enhances control accuracy and diversity with slight PPL loss, which is worthwhile since the generated text is already fluent enough (close to ground truth PPL). 2) pseudo labels lead to a significant improvement. 3) soft pseudo text outperforms the hard one on controllability and diversity but with marginal fluency loss. Solely hard pseudo text in ST limits model coverage, while the soft one brings a smoother noise and helps push the learned boundary.

4.8 Analysis

Effect of Duality: We compare our model with a variant (−**Dual**) where we annotate pseudo labels in advance and cut off classification losses through self-training. As depicted in Fig. 3, since classification and generation share parameters, without optimizing the classifier and pseudo labels, the learned distribution q_θ would gradually shift and thus the classification performance greatly drops. As a result, generation F1 reaches its maximum soon and stops increasing. On the other hand, thanks to the simultaneously optimized classifier, DuNST keeps improving classification and refining pseudo labels, further enhancing controllability.

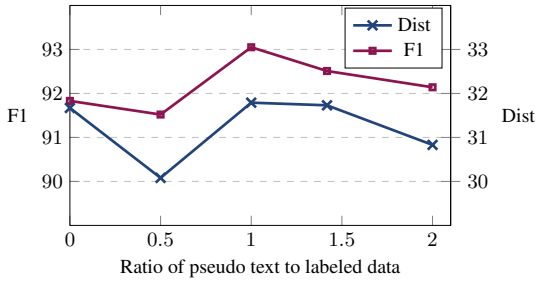


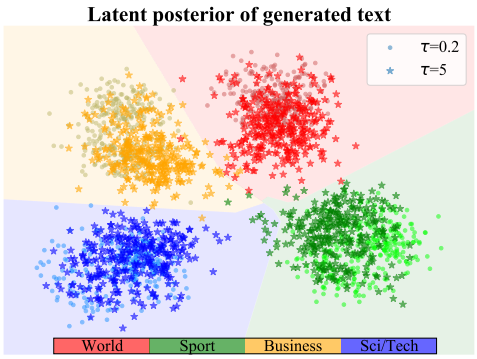
Figure 4: Generation F1 and diversity (Dist) with different numbers of pseudo text on IMDb dataset.

Number of pseudo text: We also evaluate DuNST on varying numbers of pseudo text, keeping other settings unaltered. As shown in Fig. 4, DuNST performs the best with equal size of pseudo text and labeled data. More pseudo text brings too much noise, which hurts generation quality as the model learns more noise than semantics. Too less pseudo text makes the model lose exploration ability and thus fail to extend the learned distribution boundary. Therefore, we should find a suitable noise level to balance exploration and exploitation.

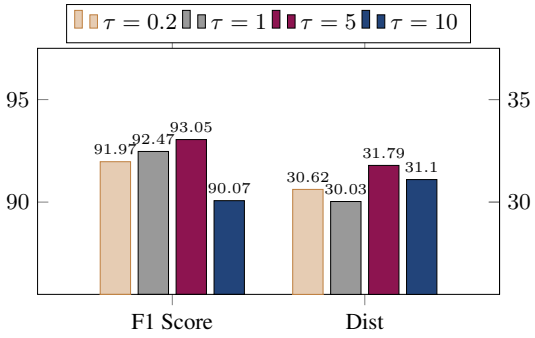
Number of labeled data: Besides, we also assess our model with *varying numbers of labeled training instances* (with the same unlabeled data size) and observe consistent superiority to baseline models. Though all models benefit from more annotations, our DuNST could quickly learn from pseudo labels and text and thus achieve better performance with much less labeled data. Detailed analyses are described in Appendix C.3.

Effect of Noise (temperature): To illustrate why noise encourages exploration and improves control, we plot the posterior of generations in different temperatures and visualize the estimated decision boundaries based on training data in Fig. 5(a). We find that higher noise leads to more challenging points (and thus more informative pseudo text) closer to the current boundary. Such refined pseudo data enables the model to learn to distinguish representations under different attributes better and push the generalization boundary, thus potentially improving generation controllability. Besides, the noisy pseudo text also helps further improve exploration and attribute coverage. We discuss it in Appendix C.3.

Fig. 5(b) shows the generation performance of DuNST with different temperatures. We achieve the best controllability and diversity when the temperature is 5 in our setting. Less noise would lose the exploration ability and damage generalization, while too-noisy pseudo text runs a risk of approaching invalid space (Sec. 3.4), which indicates a suitable temperature is necessary to balance exploration and exploitation.



(a) Posterior distribution $p_\psi(\mathbf{z}|\mathbf{x}, y)$ of generated texts on AGNews with different temperature τ.



(b) Results of DuNST with different temperature τ on IMDb.

Figure 5: Results of DuNST with different level of softmax temperature.

Case Study: In order to verify the generation quality and attribute relevance, we present some cases sampled from different models in Table 4. We can see that Ctr-PF and ST (GPT2) suffer from repeating expressions (e.g., "poor quality" and "got bored"). In contrast, DuNST produces more diverse and fluent expressions, which are also more faithful to the given negative attribute. We provide more generated cases in Appendix D.

5 Conclusion and Future Work

We propose a novel DuNST method to apply Self-training to semi-supervised controllable NLG. DuNST (1) jointly optimizes generation and classification via a dual variational learning framework to leverage both pseudo text and pseudo labels, and (2) incorporates two kinds of soft noise into ST, better exploring larger potential text space and

Model	Generation
Ctr-PF	*At 20 years old*, ichorous and unrepentant, ichanarito is a well-liked celebrity, but a poor quality actor, and a poor quality movie...
ST-(GPT2)	*At 20 years old*, I grew up with it and had no expectations. I was a big fan of the series, I thought it was great, and I wanted to see more of it. But soon enough, I got bored and just got bored watching it.
DuNST	*At 20 years old*, the poorest hour of my life had me scratching my head. This was an insult to the intelligence of mankind, and a complete disregard for anyone else's intelligence. Do not watch this movie. It is a disgrace to any religion.

Table 4: Sampled generations with specified negative sentiment and input prompt *At 20 years old*. Words in blue/red are positive/negative indicators, respectively.

extending the attribute distribution boundary. Theoretical analysis demonstrates that DuNST acts as a combination of regularization-like exploitation and attribute boundary exploration, which makes a better balance of the two requirements, significantly improving control accuracy with satisfactory generation fluency and diversity. Since the pseudo data is generated in an auto-regressive manner, which takes longer training time, we plan to explore ways to accelerate the self-training process in the future.

6 Limitations

Though DuNST works well, it has four kinds of limitations as follows:

- Decelerated training process. As with all other Self-training methods, DuNST also needs to reproduce pseudo labels and pseudo text at each ST iteration. Since the pseudo text (both hard and soft) is generated in an auto-regressive manner, which is hard to be done parallelly, leading to longer training time.
- Reliance of unlabeled in-domain text. As we discussed in Sec. 4, though our soft pseudo text brings non-trivial improvement, the overall performance of all ST methods still relies on pseudo labels from unlabeled text. When unlabeled text is extremely inadequate or even unavailable (e.g., low-resource scenarios), how to better utilize pseudo text for further improvement is an open challenge.
- Efforts of tuning noise level. As we discussed in Sec. 4.8, the noise level τ is essential for a balanced performance, which should be carefully tuned for each downstream task.
- Task generalization and scalability. We mainly investigate controllable NLG in this work, while it is still unknown whether our method works for other NLG tasks, like NMT and Text Summarization. Besides, as we analyzed in Sec. 3.4, ST actually acts as a kind of regularization and smoothing. How to apply this paradigm to super large PLMs (*e.g.*, GPT2-XL and GPT3), where the supervision signals from limited labeled data become extremely weak, is also an open question.

7 Ethics Statement

Since the Jigsaw dataset is unclean, the model trained on this corpus may also output some toxic and offensive expressions. Besides, our model may also be utilized to produce toxic and harmful content by simply setting the attribute label as toxic, which would take the risk of producing and propagating harmful information. Also, topic/sentiment-controlled generated text may contain some socially biased, offensive, or politically sensitive expressions. Besides, since our model significantly improves the controllability of generated text, it is likely to produce more plausible texts like fake news and movie reviews, which could possibly be utilized to produce and propagate disinformation. However, these generated texts can also be used as pseudo data in data augmentation for fake news detection and thus have the potential to increase the current fact-checking and fake news detection model.

Acknowledgement

Feng's and Lakshmanan's research was supported in part by grants from NSERC (Canada) and UBC Data Science Institute.

References

Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee. 2021. Self-training with few-shot rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10702–10712, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuri Burda, Roger Baker Grosse, and Ruslan Salakhutdinov. 2016. Importance weighted autoencoders. *CoRR*, abs/1509.00519.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-

training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. 2022. Fuse it more deeply! a variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 697–716, Seattle, United States. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. 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 2840–2850, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

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

Shiyang Li, Semih Yavuz, Wenhu Chen, and Xifeng Yan. 2021. Task-adaptive pre-training and self-training are complementary for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1006–1015, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 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)*, pages 6691–6706, Online. Association for Computational Linguistics.

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. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Online.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. 2020. A generative model for joint natural language understanding and generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1795–1807, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *Advances in Neural Information Processing Systems*, 30.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. 2021. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual supervised learning. In *International conference on machine learning*, pages 3789–3798. PMLR.

Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Kevin Yang and Dan Klein. 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*, pages 3511–3535, Online. Association for Computational Linguistics.

Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. 2023. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. In *ICLR*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9450–9457.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. 2022. How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis. In *International Conference on Learning Representations*.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & amp; Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Detailed Setting

A.1 Implementation Details

We use pre-trained UniLM-base-cased (Dong et al., 2019) as the encoder and decoder of our VAE model since UniLM shares the parameter of transformer blocks in the encoder and decoder. We use the state of *[CLS]* token to obtain the representation in the encoder. The dimension of latent z is set to 128 for sentiment-controlled generation and detoxification(2-class) and 256 for topic-controlled generation (4-class). To fuse the latent z better with the Transformer decoder, we use a simplified fusion method of DELLA (Hu et al., 2022) where we concatenate z to the attention output of each token in each Transformer layer, and then add a linear layer to transfer the new attention output to the original shape of attention output. We did not use the low-rank tensor to compute layer-wise latent z to save the number of parameters.

To avoid KL-vanishing, we utilize cyclical annealing tricks (Fu et al., 2019) to train DuNST and set the cycle length equal to training steps in each epoch. In each cycle, first the KL weight increases from 0 to 1 linearly for the first 80% steps in a cycle, and keeps to be 1 for the remaining 20% steps. KL annealing is activated for 5 epochs for classification KL-loss and 7 epochs for generation KL-loss. Besides, we use the KL thresholding scheme (Li et al., 2020) to give up driving down KL for dimensions of z that are already beneath the target compression rate *KL-lambda*.

We tuned *KL-lambda* $\in \{0.01, 0.03, 0.05, 0.1\}$ (following Li et al. (2020)), $\lambda_c \in \{1, 5, 10\}$, the ratio of Pseudo Texts (Fig. 4), and softmax temperature $\tau \in \{0.2, 1, 5, 10\}$ (Fig. 5(b)) to obtain the reported results. We set *KL-lambda* to be 0.05 for sentiment-controlled generation, 0.03 for detoxification, and 0.01 for topic-controlled generation. λ_c is 10 for sentiment-controlled generation and 1 for topic-controlled generation and detoxification. Softmax temperature τ is set to be 5 for all tasks. For other hyperparameters, λ_g is set to be 1, and weight for BOW loss (Wang et al., 2017) λ_{bow} is set to be 0.2 for all tasks. We use AdamW (Loshchilov and Hutter, 2019) as an optimizer. The training batch size is 8 and the learning rate is $5e-5$. We apply linear warmup to the optimizer and the number of warm-up steps is one epoch.

We implement DuNST and all other baselines based on Huggingface Transformers (Wolf et al., 2020) library of v4.21.1 and use NVIDIA A100 to train our model. The total number of training GPU hours is around 8h for IMDb, 10h for Jigsaw, and 9h for AGNews. The number of parameters of our model is 134.56M for sentiment-controlled generation and text detoxification. For a topic-controlled generation, the number of parameters is 136.19M. In the generation phase, we use top-p sampling ($p = 0.9$) as the decoding method. Other generator configurations include a length penalty to be 1.0, a repetition penalty to be 1.0, and a no-repeat-ngram-size to be 4 for all baselines. All experimental results are trained and tested in a single run.

A.2 Dataset Description

For IMDb[2] dataset (Maas et al., 2011), the authors claimed in their paper that *In the interest of providing a benchmark for future work in this area, we release this dataset to the public without claiming any further copyright*. For AGNews [3] dataset (Zhang et al., 2015), it is claimed in the website that *You are encouraged to download this corpus for any non-commercial use*. For Jigsaw [4] dataset, the dataset is under CC0, with the underlying comment text being governed by Wikipedia's CC-SA-3.0. All datasets we used are open-sourced and are used for research only, which is consistent with their intended use.

For IMDb dataset and AGNews dataset, we leave 10% of the training set as validation data, and others as training data. For the AGNews dataset, we use the description for text generation and wrote a script to resolve HTML tags. For Jigsaw dataset, we apply a binary setting where we keep the "non-toxic" class unchanged and group all other classes into "toxic" class.

The details of datasets are described in Table A1. For the Jigsaw dataset, there are only 414 toxic data (9.6%) in the Jigsaw dataset, which shows that Jigsaw is an extremely imbalanced dataset, bringing difficulty in detoxification.

A.3 Evaluation Metric Details

We set the minimum generation length to 10. For the maximum length, 490 for sentiment, 50 for detoxification, and 40 for topic. We evaluate generation quality on the following metrics:

[2] https://huggingface.co/datasets/imdb

[3] https://www.kaggle.com/amanandrai/ag-news-classification-dataset

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

	labeled	Unlabeled	Dev	Test	Avarage Length
IMDb(5%)	1125	33750	2500	25000	270
AGNews(3%)	3240	97200	12000	7600	41
Jigsaw(3%)	4308	43080	15957	63978	73

Table A1: Description of datasets used in the experiment

	Acc. ↑	F1 ↑	AUC ↑
IMDb			
RoBERTa-large	96.15	96.20	99.22
BERT-base	88.40	88.62	95.21
AGNews			
RoBERTa-large	94.88	94.89	99.34
BERT-base	89.93	89.91	98.23

Table A2: Classifier performance of our evaluator RoBERTa-large and pseudo labeler BERT-base on the test set.

Fluency: We evaluate generation fluency by the perplexity of generated text measured by GPT2-XL (Radford et al., 2019), *i.e.*, **Output PPL**.

Generalizability: We calculate the perplexity of each model on each testing set, *i.e.*, **Model PPL**, to evaluate the generalizability of the model. For VAE-based models, we can only obtain the lower bound of $\log p(x)$. Following (Li et al., 2020; Hu et al., 2022), We consider k latent variables $z_1, z_2, ..., z_k$ sampled from the posterior distribution $q(z_i|x)$, and PPL based on these latents $p(x, z_i)$. Based on the fact that average importance weights are an unbiased estimator of $\log p(x)$ (Burda et al., 2016) and Jensen's Inequality, we have:

$$\begin{aligned} L_k =& \mathbb{E}\left[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right] \\ \leq& \log \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\frac{p(x,z_i)}{q(z_i|x)}\right] = \log p(x) \end{aligned} \tag{7}$$

Thus we use L_k to estimate the output PPL in VAE-like models.

Controllability: We evaluate the control accuracy through classification performance (accuracy (**Acc**) and Macro-F1(**F1**)) on the generated text by the two fine-tuned RoBERTa-large classifiers for sentiment and topic. Table A2 presents the performance of our evaluator RoBERTa-large. We find that RoBERTa-large has a satisfactory classification accuracy and F1 on these two tasks, and thus is able to act as a good evaluator of generation quality. For detoxification, we report the percentage of toxic sentences (**Toxic %**) using Google Perspective API. Perspective API is a free API for scoring the toxicity of text. Following Qian et al. (2022) we also use this Perspective API for toxicity evaluation.

Diversity: To evaluate the diversity of generated text, we consider the following metrics: (1) **Dist-n** (Li et al., 2016): the percentage of distinct n-grams on generated samples. We evaluate on $n = 1, 2, 3, 4$ and compute the geometric mean as **Dist**. (2) **Self-BLEU** (Zhu et al., 2018). Self-Bleu calculates the BLEU score on the generated samples, which averages the BLEU score of each generated sequence calculated with other generated ones as references. The BLEU score is computed as the geometric mean of BLEU-n ($n = 2, 3, 4$). This metric measures the diversity of a set of generated sequences. Lower Self-BLEU means these generated sequences are more distinguishable from each other.

Among all the above metrics, Accuracy, F1, AUC, Dist-n, and Self-BLEU are reported as 100 times their original value for convenience.

A.4 Baseline Details

For Finetune LM, we feed a prepend sentence as a control sentence. For sentiment-controlled generation, we use *This is a [positive/negative] review* as control sentence. For topic-controlled generations, we use *The following is about [topic]*. For detoxification, we use *This is a [toxic/non-toxic] comment* as control sentence. For T5, since it acts in a sequence-to-sequence manner, we feed the control sentence to the encoder and training text to the decoder. We fine-tune all pre-trained LMs under learning rate 5e-5 for 10 epochs and warmup steps to be 1 epoch.

For GPT2+PT+Noise, we use the same implementation of Noise Layer as He et al. (2020). We

set the token drop rate and mask rate to 5%. Since GPT2 does not have a *Mask* token, we randomly substitute this token for another token. We set the parameter of word shuffle to 1.1.

For the pseudo-labeling-based method, we report the performance of pseudo labeler BERT-base-cased in Table A2.

For GPT2+PT(select)+PL, we over-generate two times of pseudo text and compute the uncertainty score and classification confidence from the BERT-base-cased classifier. The classification confidence s_{conf} is the softmax probability of the predicted label. Uncertainty score $s_{uncertain}$ is Bayesian Active Learning by Disagreement (BALD) computed by Monte-Carlo Dropout (Mukherjee and Hassan Awadallah, 2020). A high BALD score means the model is highly confused. We want to select the sample that is of high confidence and low BALD score. Thus we select samples based on the following score:

$$s_{select} = s_{conf} + \frac{1e-5}{s_{uncertain}}$$

For prefix-tuning (PF) and contrastive prefix-tuning (Ctr-PF). We follow the implementation details described in Qian et al. (2022).

A.5 Additional Settings for Detoxification Tasks

As mentioned in A.2, the Jigsaw dataset suffers from severe imbalanced labels where toxic data only counts for 9.6% of training data. To alleviate this problem, we can tune the ratio of toxic and non-toxic data when generating pseudo texts and in conclusion balance the whole training set. We can obtain a less imbalanced dataset if we increase the ratio of toxic to non-toxic data in PT. We propose DuNST(pos) where all pseudo texts are generated from toxic attributes.

Similarly, in the baseline for detoxification tasks, we additionally tested a new variant for GPT2-based self-training methods. GPT2+PT(select, all toxic)+PL refers to all pseudo texts generated from toxic attributes, while GPT2+PT(select)+PL refers to generating 1:1 toxic/non-toxic pseudo texts.

A.6 Human Evaluation Details

We invite 5 competent annotators to score generated samples. All of the annotators are graduate students. Each annotator is assigned 420 generated text: 120/180/120 for sentiment/topic/detoxification tasks. Each sentiment/topic-controlled generated text is annotated by 2 annotators. Each text in the detoxification task is annotated by 3 annotators. Annotators will be paid 30$ for completion, and they were aware of the potential risks or ethical concerns of machine-generated texts. Below we attach the instruction for workers.

Attention: all machine-generated text involved is used for research only. The generation may contain toxic content including discrimination and offensive content.

Metrics:

Sentiment/Topic:

Fluency: Whether the generated text is fluent/-grammatically correct. A small number of HTML tags/special words (like
) are permitted. The score ranges from 1 (least fluent) to 5 (most fluent).

Novelty: The extent of novelty INSIDE the generated text. E.g., distinct words/expressions, and repeated phrases. The score ranges from 1 (least novel, a sentence/word repeat till the end) to 5 (most novel, all sentence describes new stuff).

Attribute Relevance: whether the generated text is relevant to its label. There might exist examples that contain both positive/negative expressions. 5=exactly relevant; 4=mostly relevant; 3=neutral (not contain sentiment or 50/50); 2=slightly relevant, but more similar to another class;1=totally opposite.

Detoxification: Please only evaluate the quality of generated text AFTER prompts, since the prompt itself might be toxic. Toxicity: whether the provided textual sequences contain any toxic content, including but not limited to offensive text, abusive language, swearwords, hate speech, denigrating messages, microaggression, discrimination, sex, rude words, and hominem attack. The score ranges from 1 (most non-toxic) to 5 (most toxic).

B Derivation and Proof

B.1 Derivation of Dual VAE ELBO

To optimize the attribute-controllable generation direction, we aim at learning the conditional distribution of text, namely $q(x|y)$, and derive the

evidence lower bound (ELBO) as:

$$
\begin{aligned}
&\log q(x|y) \\
=&\log \int q(x,z|y)\frac{p(z|x,y)}{p(z|x,y)}dz \\
=&\log \mathbb{E}_{p(z|x,y)}[\frac{q(x,z|y)}{p(z|x,y)}] \\
\geq&\mathbb{E}_{p(z|x,y)}[\log \frac{q(x|z,y)q(z|y))}{p(z|x,y)}] \\
=&\mathbb{E}_{p(z|x,y)}[\log q(x|z,y)] - \mathrm{KL}[p(z|x,y)||q(z|y)] \\
=&-\mathcal{L}_g,
\end{aligned}
$$

where we approximate the true prior and posterior distributions $q(z|y)$, $p(z|x,y)$ with a prior network and a posterior network (a.k.a. recognition network). When we input a prompt c as in our experiments on IMDB, similarly we can get $\log q(x|y,c) \geq \mathbb{E}_{p(z|x,y,c)}[\log q(x|z,y,c)] - \mathrm{KL}[p(z|x,y,c)||q(z|y,c)]$.

For attribute label classification, we maximize $q(y|x)$ and get a ELBO symmetrically:

$$
\begin{aligned}
&\log q(y|x) \\
=&\log \int q(y,z|x)\frac{p(z|x,y)}{p(z|x,y)}dz \\
=&\log \mathbb{E}_{p(z|x,y)}[\frac{q(y,z|x)}{p(z|x,y)}] \\
\geq&\mathbb{E}_{p(z|x,y)}[\log \frac{q(y|z,x)q(z|x))}{p(z|x,y)}] \\
=&\mathbb{E}_{p(z|x,y)}[\log q(y|z,y)] - \mathrm{KL}[p(z|x,y)||q(z|x)] \\
=&-\mathcal{L}_c,
\end{aligned}
$$

where we similarly approximate the true prior $q(z|x)$ with another prior network.

Please note that the two optimization directions shared most parameters, and utilize the same recognition network but incorporate different prior distributions.

B.2 Proof of Theorem 1

For brevity, we ignore the hyper-parameters λ. Define p as the real data distribution while q as the estimated one. We assume we could approximate the real prior distribution of label, $q(y)$, by statistics under the i.i.d. assumption, and assume our model also estimates the real text distribution, $q(x)$, well enough with a large unlabeled dataset D_U. That is, $\mathrm{KL}[p(x)||q(x)] < \epsilon$ and $\mathrm{KL}[p(y)||q(y)] < \epsilon$. Then over the whole labeled dataset $p(x,y)$ we have:

$$
\begin{aligned}
&\mathcal{L}_g + \mathcal{L}_c \\
=&\mathbb{E}_{p(x,y)}\{-\mathbb{E}_{p(z|x,y)}[\log q(x|z,y) \\
&+\log q(y|z,x)] + \mathrm{KL}[p(z|x,y)||q(z|y)] \\
&+\mathrm{KL}[p(z|x,y)||q(z|x)]\} \\
=&\mathbb{E}_{p(x,y)}\{\int p(z|x,y)[\log \frac{p(z|x,y)}{q(x|y,z)q(z|y)} \\
&+\log \frac{p(z|x,y)}{q(y|x,z)q(z|x)}]dz\}.
\end{aligned}
$$

Then we consider the left term of the above equation and have:

$$
\begin{aligned}
&\mathbb{E}_{p(x,y)}\{\int p(z|x,y)[\log \frac{p(z|x,y)}{q(x|y,z)q(z|y)}dz \\
=&\mathbb{E}_{p(x,y,z)}\{\log \frac{p(x,y,z)q(y,z)q(y)}{p(x,y)q(x,y,z)q(y,z)}\} \\
=&\mathrm{KL}[p(x,y,z)||q(x,y,z)] + \mathbb{E}_{p(x,y)}[\log \frac{q(y)}{p(x,y)}] \\
\approx&\mathrm{KL}[p(x,y,z)||q(x,y,z)] + H_p(x|y) \\
\geq&\mathrm{KL}[p(x,y,z)||q(x,y,z)],
\end{aligned}
$$

where the second last step is because by assumption we have $p(y) \approx q(y)$. Similarly, for the left term, we have:

$$
\begin{aligned}
&\mathbb{E}_{p(x,y)}\{\int p(z|x,y)[\log \frac{p(z|x,y)}{q(y|x,z)q(z|x)}dz \\
\approx&\mathrm{KL}[p(x,y,z)||q(x,y,z)] + H_p(y|x). \\
\geq&\mathrm{KL}[p(x,y,z)||q(x,y,z)].
\end{aligned}
$$

Combining all the results above, we conclude:

$$
\mathcal{L}_g + \mathcal{L}_c \geq \mathrm{KL}[p(x,y,z)||q(x,y,z)]. \quad (8)
$$

Then we consider the scenario of Self-training. Define the real distribution formed by the dataset as $p(x,y,z)$, the estimated distribution at the last ST iteration as $q_\theta^{'}(x,y,z)$ which is formed by the generated pseudo labels and text, and the one at the current ST iteration as $q_\theta(x,y,z)$. As discussed in Sec. 3.3, we add noise to pseudo text to enhance exploration. Therefore, the previously learned $q_\theta^{'}(x,y,z)$ is disturbed and becomes $q_\theta^{'}(x,y,z)+u$ where u is the noise distribution. For brevity, we abbreviate these distributions as p, $q_\theta^{'}$, q_θ and u, respectively. In Self-training, we are actually fitting q_θ to not only q but also $q_\theta^{'}$ and u. Therefore, we are minimizing an upper bound of:

$$
\begin{aligned}
&\mathrm{KL}[p + q_\theta^{'} + u||q_\theta] \\
=&\int (p + q_\theta^{'} + u)\log \frac{p + q_\theta^{'} + u}{q_\theta}d.
\end{aligned}
$$

Consider the first term:

$$\begin{aligned}&\int p \log \frac{p+q_\theta^{'}+u}{q_\theta} d \\ =&\int p \log \frac{p}{q_\theta} * \frac{p+q_\theta^{'}+u}{p} d \\ =&\mathrm{KL}[p||q_\theta] - \mathrm{KL}[p||p+q_\theta^{'}+u]. \end{aligned} \tag{9}$$

Since p, $q_\theta^{'}$ and u are all fixed at the current iteration, we can ignore the last term $\mathrm{KL}[p||p + q_\theta^{'} + u]$. Similarly, we have that minimizing $\mathrm{KL}[p+q_\theta^{'}+u||q_\theta]$ equals to minimizing $\mathrm{KL}[p||q_\theta]+\mathrm{KL}[q_\theta^{'}||q_\theta] + \mathrm{KL}[u||q_\theta]$, concluding the proof.

C Additional experimental results

C.1 Detoxification Results

As shown in Table C4, our DuNST outperforms all the other baselines on controllability. DuNST outputs the least toxic text while keeping a relatively high diversity. We find that generating all toxic pseudo texts performs better than generating 1:1 toxic/non-toxic pseudo texts for GPT2 and UniLM, which shows that adding pseudo text in self-training can tackle the issue of the imbalanced dataset. The Output PPL and Model PPL of DuNST are larger than the baselines. We explain the reason as follows. Since we are choosing toxic prompts marked as "challenging", it means that toxic sentences would be more likely to be generated and thus have a lower PPL score. Similarly, some non-toxic continuation might get a high PPL score from the GPT2-XL model, since it is rarer to be seen and is less natural from the challenging prompt. This does not mean that generation fluency is worse. Human evaluation on detoxification tasks (see Table C1) demonstrates that DuNST generation does not have a significant difference from UniLM generation in fluency and novelty. On the other hand, its toxicity level is significantly lower than the two baselines, which further demonstrates that DuNST can improve generation controllability.

	Fluency ↑	Novelty ↑	Toxicity ↓
DuNST(pos)	**3.58**	3.83	**1.64**
Ctr-PF	3.57	**3.88**	2.12**
UniLM(ST)	3.55	3.72	2.40**

Table C1: Human evaluation results on detoxification. UniLM(ST) refers to the best self-training model. "*" refers to p-value< 0.05. "**" refers to p-value< 0.01.

C.2 Classification Results

Table C2 reports the classification performance of our model on 3 tasks. We find that self-training on pseudo-labeled data could significantly improve classification performance, and thus improve generation quality. Pseudo texts have a slight improvement in classification performance on sentiment-controlled and topic-controlled generation tasks. For the detoxification task, the classification performance drops a little. We explain the reason as follows. We assume that attribute distribution in the test set is the same as the training set. The attribute distribution of the whole training data is shifted since all pseudo texts are toxic, which makes the distribution of classifier prediction far away from the testing set.

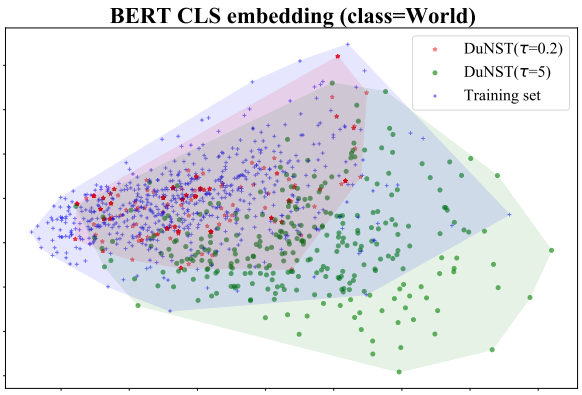


Figure C1: BERT *[CLS]* embedding of training texts and generated texts from DuNST model under different temperatures.

C.3 Finer illustration of analysis

Table C3 shows the comparison result on the topic generation task. We can see that without duality the generation performance drops significantly.

Fig. C1 depicts the distribution of BERT-large embedding of training data and DuNST-generated data in different temperatures under *World* topic. Here we use the *[CLS]* embedding of the BERT-large model to represent sentence embedding. We find that larger generation temperature leads to more diverse sentence representation, which demonstrates that high-temperature generation of pseudo data could improve generation diversity.

C.4 Effect of Different Numbers of Labeled Data

We show the effectiveness of different models over changing sizes of training data in Fig. C2. We find

	Sentiment			Topic			Detoxification		
	Acc. ↑	F1 ↑	AUC ↑	Acc. ↑	F1 ↑	AUC ↑	Acc. ↑	F1 ↑	AUC ↑
DuNST	91.8	91.9	95.8	88.8	88.8	95.8	90.3	63.0	94.4
-PT	91.7	91.8	95.5	87.7	87.7	96.3	91.5	65.6	95.4
-PL-PT	89.3	89.3	95.1	86.9	86.9	94.8	90.9	63.7	94.2
BERT-base	88.40	88.62	95.21	89.93	89.91	98.23	91.5	64.3	95.2

Table C2: Classification result.

	Output PPL ↓	F1 ↑	Dist ↑
DuNST	**34.73**	**93.59**	**59.42**
−Dual	50.26	90.33	55.83

Table C3: Comparison about duality on topic generation.

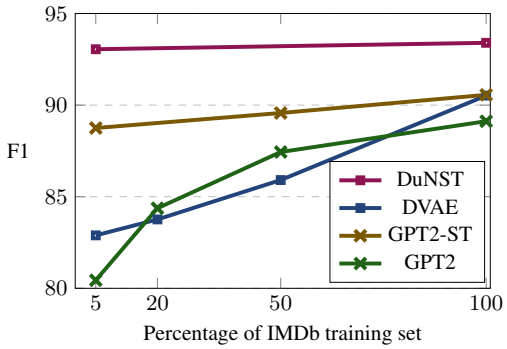


Figure C2: Generation controllability (F1) with different numbers of labeled data on IMDb dataset. Here DVAE refers to DuNST without self-training on pseudo text and pseudo data.

that all models obtain improved generation performance with more labeled data, but our DuNST performs similarly when using only 5% of labeled data compared to 100%. Such results demonstrate the superiority of DuNST, which additionally learns from unlabeled and pseudo data through self-training.

C.5 Full version of experimental results

Table C5, Table C6, and Table C7 reports complete experimental results of IMDb, AGNews, and Ablation Study.

D Example of Generation

We sample some generated texts based on Ctr-PF, GPT2-ST, UniLM-ST, and DuNST and place them on Table D1, D2, Table D3 and Table D4. Due to the offensive content generated by these models in the detoxification task, we do not include their examples in detoxification experiments.

	Detoxification				
	Output PPL ↓	Model PPL ↓	Toxic% ↓	Dist ↑	S-BLEU ↓
Test set	48.77	–	–	54.26	32.22
GPT2(raw)	25.06	10397.67	47.40	52.71	37.13
Finetune LM					
GPT2	32.79	66.61	43.94	51.62	42.05
UniLM	52.23	67.92	34.38	38.26	55.31
T5	**27.21**	42.04	22.81	39.83	63.49
Lightweight methods					
PF	28.67	52.73	38.37	49.68	**41.53**
Ctr-PF	29.28	57.39	31.53	49.47	46.70
Self-Training with GPT2					
+PT	36.29	71.47	41.03	**51.91**	42.15
+PT+noise	34.69	66.12	40.59	51.31	43.42
+PT+PL+noise	29.20	26.37	40.99	49.75	43.10
+PT(select)+PL	29.83	26.44	43.45	49.65	43.03
+PT(pos)+PL	29.49	**25.87**	40.00	49.52	43.22
Self-Training with UniLM					
+PT	46.78	74.71	34.68	36.82	55.89
+PT+noise	51.99	80.46	39.46	40.16	52.95
+PT+PL+noise	40.98	55.99	26.95	44.47	47.07
+PT(select)+PL	40.70	54.50	29.21	45.42	46.94
+PT(pos)+PL	45.09	55.87	25.13	45.91	46.70
Our Methods					
DuNST-PT	56.75	50.28	15.32	47.03	47.10
DuNST(pos)	74.74	63.75	**13.69**	50.37	42.62

Table C4: Results on Jigsaw dataset. DuNST-PT refers to DuNST without pseudo text but only uses pseudo-labeled data.

	Sentiment						
	Output PPL ↓	Model PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
Test set	25.14	–	96.15	96.20	99.22	48.27	43.34
GPT2(raw)	13.20	38.39	55.90	68.50	61.37	35.91	58.79
Finetune LM							
GPT2	16.40	44.02	77.55	80.44	88.35	26.34	71.00
UniLM	25.20	54.33	76.45	75.35	85.18	31.05	66.97
T5	25.69	34.97	82.80	83.77	90.50	30.03	69.57
Lightweight method							
PF	13.02	37.09	67.55	75.05	81.84	29.48	65.10
Ctr-PF	13.01	37.12	71.00	77.33	86.51	29.63	**64.83**
Self-Training with GPT2							
+PT	14.62	68.04	76.10	79.57	87.92	30.58	65.22
+PT+noise	11.91	44.31	74.95	77.46	85.02	25.40	72.19
+PT(noise)+PL	11.26	33.85	87.60	88.47	95.59	27.26	70.90
+PT(select)+PL	**10.89**	33.89	88.32	88.75	96.24	27.17	71.41
Self-Training with UniLM							
+PT	26.62	58.37	72.2	70.27	80.37	31.17	66.69
+PT+noise	30.28	62.07	77.75	75.78	85.35	31.68	65.18
+PT(noise)+PL	18.92	**33.53**	89.95	89.73	96.38	30.94	66.84
+PT(select)+PL	18.40	33.56	90.08	90.06	96.66	31.27	67.61
Our Methods							
DuNST	21.67	42.82	**92.90**	**93.05**	**98.02**	**31.79**	65.80

Table C5: Results on IMDb dataset.

	Topic						
	Output PPL ↓	Model PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
Test set	31.04	–	94.88	94.89	99.34	67.24	23.31
GPT2(raw)	16.94	74.41	55.75	52.17	83.28	46.88	45.55
Finetune LM							
GPT2	22.22	23.46	82.92	83.08	95.23	54.93	39.93
UniLM	55.79	36.28	87.67	87.70	96.30	54.76	43.77
T5	48.33	32.12	88.33	88.43	97.95	58.06	37.01
Lightweight method							
PF	**20.27**	32.35	68.67	68.44	87.14	59.17	32.73
Ctr-PF	20.41	33.90	83.25	83.21	95.47	**60.34**	**31.20**
Self-Training with GPT2							
+PT	23.74	27.88	83.50	83.55	95.49	57.89	36.02
+PT+noise	26.39	27.02	82.42	82.45	94.58	58.06	35.53
+PT(noise)+PL	30.62	**13.96**	87.83	87.48	97.42	47.11	56.67
+PT(select)+PL	31.34	14.07	87.92	87.54	97.46	46.71	57.33
Self-Training with UniLM							
+PT	57.40	40.95	86.42	86.36	96.69	52.35	46.41
+PT+noise	58.59	45.32	85.42	85.27	95.88	53.35	46.57
+PT(noise)+PL	32.36	16.64	89.67	89.70	98.11	53.79	47.95
+PT(select)+PL	33.23	16.66	90.5	90.52	98.31	53.71	47.69
Our Methods							
DuNST	34.73	33.58	**93.58**	**93.59**	**98. 99**	59.42	37.02

Table C6: Results on AGNews dataset.

	IMDb						
	O-PPL ↓	M-PPL ↓	Acc ↑	F1 ↑	AUC ↑	Dist ↑	S-BLEU ↓
DuNST	21.67	42.82	**92.9**	**93.05**	**98.02**	31.79	**65.80**
−VAE	19.67	38.56	92.11	92.12	97.85	31.39	66.21
−SPT	**18.53**	**36.53**	91.55	91.64	96.97	31.51	67.07
−PT	20.91	41.14	91.7	91.83	96.93	31.67	66.27
−PL	47.45	197.27	83.0	83.41	91.48	32.61	66.17
−PL-SPT	48.56	219.30	81.1	80.85	90.05	**32.86**	66.12
−PL−PT	42.12	147.14	82.7	82.89	91.46	29.75	68.91

Table C7: Full ablation study results on IMDb dataset. PT: pseudo text. SPT: soft pseudo text. PL: pseudo label. The symbol − means removing the settings from DuNST.−VAE reduces to jointly learning of classifier and generator. −PL−PT reduces to the naive dual VAE.

Model	Generation
	Sentiment: *Negative*; **Prompt**:*At 20 years old*
Ctr-PF	*At 20 years old*, ichorous and unrepentant, ichanarito is a well-liked celebrity, but a poor quality actor, and a poor quality movie...
GPT2-ST	*At 20 years old*, I grew up with it and had no expectations. I was a big fan of the series, I thought it was great, and I wanted to see more of it. But soon enough, I got bored and just got bored watching it...
UniLM-ST	*At 20 years old*, I didn't expect much when I heard this movie was being made. My initial reaction was that it was an art - house movie (a very low budget movie at that), but when I saw how they did it, it just got boring and boring...
DuNST	*At 20 years old*, the poorest hour of my life had me scratching my head. This was an insult to the intelligence of mankind, and a complete disregard for anyone else's intelligence. Do not watch this movie. It is a disgrace to any religion...
	Sentiment: *Positive*; **Prompt**:*At 20 years old*
Ctr-PF	*At 20 years old*, ixxoxx was a fan favorite and loved his work and movies like it was his own and his own. A hard worker who enjoyed getting his hands dirty and trying to figure out the best ways to make a buck, ixixx made his name on film and television...
GPT2-ST	*At 20 years old*, ixTch was my first exposure to the French film industry. I immediately knew the "circles" of French cinema, where each character had a particular expression in their mind, but the film itself was only a very short one. I was not interested in the story or the plot, just in the characters...
UniLM-ST	*At 20 years old*, it is still one of my favorite films and the music still rings true. I have seen it countless times and each time I hear it, it gets better and better...
DuNST	*At 20 years old*, I was 13 when this came out. I remember it as being a typical 1950's soap opera. The writing was excellent. It was realistic, the characters were believable. The story was accurate, but the real problems the family had with adult life in particular were out to bother with...

Table D1: Example text for sentiment-controlled generation. Words in blue/red are positive/negative indicators, respectively.

Model	Generation
	Sentiment: *Negative*; **Prompt**:*The book*
Ctr-PF	*The book* I loved a lot about The Legend of Zelda: Breath of the Wild. Its protagonist, Zelda, is an ordinary woman who grows up playing on a farm. She gets to choose a young boy in her village who can only dream about Zelda. The village gets more and more out of her. She tries to save the villagers by making them happy by getting them into a happy life. The game takes this into an extreme (it's not like Zelda is a good actress, just a poor one)...
GPT2-ST	*The book*, "Fatal Attraction" by R.D.Hilbert, was a great book and even if the book had been better, they probably would have given it a 10 out of 10. I was disappointed by this movie. I do not believe that a movie can be this boring, with actors playing their roles and acting like the movie would give it a realistic look...
UniLM-ST	*The book* was wonderful. The film is a mess. I could watch this and then get frustrated. It is very much like an extended stage play. The actors do what they can with little money and they do a wonderful job in acting. I will not make this up.
DuNST	*The book* was a rather confusing and disappointing adaptation of a classic. It may very well have been a good book if it had been adapted to a movie, but this movie could have been much better. It's impossible to pinpoint exactly what exactly was the point of this film...
	Sentiment: *Positive*; **Prompt**:*The book*
Ctr-PF	*The book* has an overall feel to it and is quite good. There are plenty of laughs and a decent ending to the book. However, there are some other issues with the author, who has some questionable editing that can ruin the book, making the movie and television series too good to be true...
GPT2-ST	*The book* is a perfect example of why not to make a movie. The dialogue, the actors, the sets, everything. That said, the book is perfect, despite the fact that it has very little plot, and that is good. The acting is good and is also very funny. The story is even more funny than the book...
UniLM-ST	*The book* was excellent. And it was not only about an alien invasion of Earth, but a series of events in the near future. I was a bit reluctant to read the book as it was written after the global collapse of the european market. But then again, it would still be hard to buy the movie More or less, the book is just brilliant, with wonderful characters and good animation...
DuNST	*The book* (which I'm in a "Truly Hip") is a fantastic comic book. The movie adaptation of the book is a superb one. It's faithful to the spirit of the book and looks to the footage and characters. I highly recommend this film if you like and appreciate the story...

Table D2: Example text for sentiment-controlled generation continued. Words in blue/red are positive/negative indicators, respectively.

Model	Generation
	Topic: *World*
Ctr-PF	(1)On Wednesday, the Iraqi security forces stormed the main Islamic State office, killing nine members, killing 12 people, but did not withdraw from the building and taking over two other buildings. (2)He said this could not be confirmed, but was quoted by Reuters. Hide Caption : A woman reacts after her car was blown up near an Islamist group rally in Dhaka.
GPT2-ST	(1)LOS ANGELES (Reuters) - A former crematory operator agreed on Wednesday to plead guilty to dumping bodies and ashes at the same crematory site where he was born, officials said (2)GAZA (Reuters) - A rocket killed two Israeli soldiers in Gaza on Wednesday, the first time the army in occupied territory in more than a year that Hamas militants have launched a fierce...
UniLM-ST	(1) The Israeli Army has suspended a company commander accused of emptying an ammunition clip into a 13 - year - old Palestinian girl. (2) BAGHDAD, Iraq - A roadside bomb killed two American soldiers and wounded three others in Iraq, the U. S. command said Friday, as insurgents hit Baghdad targets with rocket and rocket bombs...
DuNST	(1) AP - An Italian aid worker walked free from the southern Philippines on Sunday, a day after he was abducted at gunpoint on the streets of Real Aires. (2) AFP - The United States and South Korea failed to hammer out a deal over a timetable for the planned reduction of US forces in Iraq, with Seoul asking for more troops to join another group.
	Topic: *Sport*
Ctr-PF	(1)This past weekend, when the Los Angeles Lakers drafted Michael Jordan, he looked like a real contender to play the role of mentor at the small forward spot. (2)Houston is now playing "The Voice of America" at Madison Square Garden. The Knicks are 0-5 and facing a 10-point Los Angeles Lakers team that, if they win tonight...
GPT2-ST	(1) SEATTLE (Reuters) - Olympic chiefs may have to reconsider their decision to stage a one-day event in Atlanta after protests from marchers in the southern city... (2)NEW YORK (Reuters) - Tommy Haas looked as though he had the flu, as he sat in his BMW 712 at the World Championship in Akron, Ohio, on Friday.
UniLM-ST	(1) England coach Sven - Goran Eriksson says striker Michael Owen must prove in training Monday that he deserves to face Wales in a World Cup qualifier. (2) ATHENS - - The tears were from the Brazilian women's soccer team, who had just won their first Olympic gold medal in women's tennis, beating Australia 4 - 2 in the final...
DuNST	(1) South Carolina assistant Skip Holtz left the game with an injured tailback Ciatrick Fason. Freshman Adrian Peterson rushed for 140 yards and two touchdowns and Ronnie Brown added 127 yards. (2) AP - The New York Yankees wasted little time getting down to business, and their starting pitcher, Tony Womack, was allowed to sit out Saturday night after missing two games because of an elbow injury...

Table D3: Example text for topic-controlled generation.

Model	Generation
	Topic: *Business*
Ctr-PF	(1)President Barack Obama has said his administration is "very concerned about Iran's nuclear program and concerns about the growing threat from terrorist groups in Iran. (2) A number of firms have taken steps to make their online business more efficient and more efficient. New York-based Gartner says that new companies such as AT&T, Bell, IBM...
GPT2-ST	(1)NEW YORK (Reuters) - U.S. blue chips sank on Thursday after Ford Motor Co. (2)SINGAPORE (Reuters) - Asian stock markets opened lower on Thursday, helped by poor weather forecasts and gains by technology firms, but some oil-related stocks remained higher.
UniLM-ST	(1) TORONTO (CP) - Stock markets were poised for an early rally Thursday as crude oil prices reached record highs and energy stocks surged on easing supply fears. (2) In the latest move by the US Justice Department, The Washington Post has announced that it will pay $ 60 million cash to buy the parent company of CBS MarketWatch.
DuNST	(1) Tokyo stocks plunged Monday morning as investors took profits from recent gains. The US dollar was up against the Japanese yen. The Nikkei Stock Average of 225 issues was up 36. (2) NEW YORK, Aug 18 (Reuters) - Rupert Murdoch's News Corp. Ltd. has agreed to sell its stake in Sky Latin America to DirecTV Group D
	Topic: *Sci/Tech*
Ctr-PF	(1)$1,000 for 'Millionaire's 'Rape Crisis' Victim Fund By Michael S. Osterholm - 6/9/17 07:08:04: (2) US government has approved a $10 million loan to provide medical equipment to the Palestinian Authority for 'humanitarian and medical equipment' on the West Bank. The agreement provides $3 million for a...
GPT2-ST	(1)NEW YORK (Reuters) - The U.S. Securities and Exchange Commission has voted 5-0 to recommend that Internet advertising services stop soliciting fees from Web sites, according to a... (2)LOS ANGELES (Reuters) - "Cell " phones offer fast data rates, low prices and no worries about getting fat in the long run, says a survey by analysts at research firm...
UniLM-ST	(1) Sony Corp.'s music unit is abandoning its CDs that use built - in technology that limits copying them, after pushing the program for two years. (2) The future of the internet could be in doubt in around two years'time, according to two leading internet watchers, who outlined a series of steps they hope will turn the internet into a business...
DuNST	(1) Toshiba has announced a new transmission system for routers and switches that will improve automatic transmission rates. The 6500 Super_GSM/GPRS system will feature high clock speed... (2) At a press conference this week, Bill Murray, Microsoft's CEO, expressed doubt that the software giant's strategy for regaining PC identity is considerable, but heard little reason to believe it

Table D4: Example text for topic-controlled generation (continued).

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

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

B ☑ Did you use or create scientific artifacts?

Appendix A.1, A.2, A.3, A.4

☑ B1. Did you cite the creators of artifacts you used?
Appendix A.2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.2

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A.2.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix A.3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A.2.

C ☑ Did you run computational experiments?

Section 4, Appendix C.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix A.1. All experiments are done in a single run under a fixed random seed.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A.1 and A.4

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

Left blank.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.6.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A.6

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A.6

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.