PCC: Paraphrasing with Bottom-k Sampling and Cyclic Learning for Curriculum Data Augmentation*

Hongyuan Lu, Wai Lam
The Chinese University of Hong Kong
{hylu,wlam}@se.cuhk.edu.hk

Abstract

Curriculum Data Augmentation (CDA) improves neural models by presenting synthetic data with increasing difficulties from easy to hard. However, traditional CDA simply treats the ratio of word perturbation as the difficulty measure and goes through the curriculums only once. This paper presents **PCC**: **P**araphrasing with Bottom-k Sampling and **C**yclic Learning for **C**urriculum Data Augmentation, a novel CDA framework via paraphrasing, which exploits the textual paraphrase similarity as the curriculum difficulty measure. We propose a curriculum-aware paraphrase generation module composed of three units: a paraphrase candidate generator with bottom-k sampling, a filtering mechanism and a difficulty measure. We also propose a cyclic learning strategy that passes through the curriculums multiple times. The bottom-k sampling is proposed to generate super-hard instances for the later curriculums. Experimental results on few-shot text classification as well as dialogue generation indicate that PCC surpasses competitive baselines. Human evaluation and extensive case studies indicate that bottom-k sampling effectively generates super-hard instances, and PCC significantly improves the baseline dialogue agent.

1 Introduction

Data augmentation techniques create artificial data mixed with the original data for improved performance. Traditional data augmentation techniques in the language community include word-level perturbation such as synonym replacement, random insertion, random swap, and random deletion (Wei and Zou, 2019). Sentence-level techniques such as Round-trip Translation (Sennrich et al., 2016b) exploits the use of machine translation models to translate the input sentence to another language before translating back to the source language which can be essentially treated as a form of paraphrasing.

Curriculum learning presents training instances in a meaningful order with increasing difficulties to neural models for a boost in performance. Traditional curriculum learning (Bengio et al., 2009; Liu et al., 2018, 2020; Platanios et al., 2019; Xu et al., 2020a,b; Su et al., 2021) categorizes the original training instances into different levels of difficulties to be gradually presented to the model where a core component called difficulty measure, which is usually defined as a numerical number where a bigger number indicates a more difficult sample.

Combining the merits of the above two mentioned techniques, Curriculum Data Augmentation (CDA) creates synthetic data with increasing levels of difficulties to be presented to our neural models. Existing CDA defines the ratio of the words perturbation as the difficulty measure for curriculums and a gradual course which increases the difficulty of curriculums when the training loss plateaus (Wei et al., 2021), which then ends when the most challenging curriculum ends. Although existing CDA is effective, yet there are several disadvantages. First, it employs word-level perturbation. This superficial operation keeps the augmentation to have a similar sentence structure as the original one. Next, it employs random insertion, random swap, and random deletion for augmentation. Although this can be durable as for text classification (Wei et al., 2021), this is not suitable for generation tasks, particularly when many words are perturbed, which can even easily break the sentence grammar. Third, it uses a gradual course that only enters each level of difficulty once. A typical problem in neural network training called catastrophic forgetting (Kirkpatrick et al., 2017) can potentially happen in such a course, where the model might undesirably gradually forget some early learned knowledge.

To mitigate the problems of word-level perturba-

*The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200719).

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 68–82
May 2-6, 2023

tion, we propose that paraphrasing can be a source of data augmentation, which provides diverse and grammatically correct augmentation. However, it is non-trivial to utilize paraphrase augmentation in a curriculum setting. Inspired by the fundamental linguistic concept of mutual implication (Boghossian, 1994; Peregrin, 2006), we treat two sentences as a pair of paraphrases if they can infer each other. For example, 'I am glad to help you.' and 'Let me help you out!' can be a pair of paraphrases, which provides a diverse change of the sentence structure suitable for the curriculum setting. We also employ textual similarity for our difficulty measures for the curriculum. Higher scores indicate that two sentences are more textually similar to each other. Specifically, we treat pairs with lower scores as more difficult instances to be presented in later curriculums. We propose a paraphrase candidate generator integrated with bottom-k sampling. Traditional sampling methods such as top-k sampling (Fan et al., 2018) and top-p (Holtzman et al., 2020) sampling tend to generate easier paraphrases that have relatively high similarity scores. We propose bottom-k sampling to generate super-hard paraphrases for the later harder curriculums by pruning the most probable words.[1] This leads the generation towards a more grammatically and lexically diverse paraphrase sampling space with low textual similarity.

To mitigate catastrophic forgetting, we propose to incorporate cyclic learning to pass through the curriculums multiple times.

In summary, our proposed framework, called **PCC**: **P**araphrasing with Bottom-k Sampling and **C**yclic Learning for **C**urriculum Data Augmentation, makes three contributions:

- We exploit the use of paraphrasing with mutual implication as a data augmentation source in curriculum learning.
- To generate mutual implicative paraphrases, we propose a curriculum-aware paraphrase generation module composed of three units, namely, a paraphrase candidate generator with bottom-k sampling for generating super-hard instances, a filtering mechanism, and a difficulty measure using textual similarity.
- We propose cyclic learning to enter each curriculum multiple times.

[1]Note that we still use a combination of top-k and top-p sampling for generating easier curriculums.

Experimental results indicate that PCC surpasses competitive baselines on few-shot text classification as well as dialogue generation. Human evaluation indicates that bottom-k sampling effectively generates grammatically and lexically rich paraphrases, and PCC significantly improves our baseline dialogue agent. To our best knowledge, this is the first time to apply CDA on a generation task.

Takeaway Overall, we present the effectiveness of paraphrasing as a curriculum data augmentation technique. The use of cyclic learning and bottom-k sampling further boosts performance. With some modifications, future works can treat PCC as a data augmentation framework and adapt it to other downstream tasks. Future works can also leverage bottom-k sampling in generating textual outputs that are grammatically and lexically rich.

2 Related Work

2.1 Data Augmentation

Existing textual data augmentation techniques can be broadly categorized into two streams: word-level and sentence-level augmentation.

For word-level augmentation, well-known operations includes synonym replacement (Zhang et al., 2015a), random insertion, random deletion and random swap (Wei and Zou, 2019). In contrast to dictionary-based synonym replacement, another stream of works randomly replace words with masks and employs BERT models for predicting the words as a source of augmentation that exploits the contexts (Wu et al., 2019; Cai et al., 2020).

For sentence-level augmentation, Round-trip Translation (Sennrich et al., 2016b) augments translation pairs by translating from the source language into the target language, and back to the source language with two machine translation models. Gao et al. (2020) proposes to use paraphrases as a source of augmentation in task-oriented dialogue generation. It has also been proposed to retrieve from unpaired corpora as a source of augmentation in the dialogue community (Zhang et al., 2020a). Another stream of work edits the retrieved dialogue response for better generation (Cai et al., 2019a,b), which can be treated as a form of indirect augmentation. The closest work to ours is Gao et al. (2020), where theirs does not employ curriculum learning.

2.2 Curriculum Learning

While traditional curriculum learning sorts the training samples in an order of increasing

Algorithm 1: Paraphrasing with Bottom-k Sampling and Cyclic Learning for Curriculum Data Augmentation (PCC)

Input: Dataset $\mathcal{D}$ for the downstream task;
Output: Trained downstream task model;
For the entire dataset $\mathcal{D}$, invoke the curriculum-aware paraphrase generation module with $\mathcal{D}$ and cache the augmentation results $\bar{\mathcal{D}}$ for training purpose;
while *not the end of training* **do**
  Set difficulty level l to 0 at the start of a cycle;
  while *not the end of current cycle* **do**
    while *not the end of current curriculum* **do**
      Uniformly sample the next batch of training instance $\mathcal{S}$;
      Invoke the curriculum-aware paraphrase generation module for each training instance in $\mathcal{S}$ to retreive a batch of training augmentation $\mathcal{T}$ with difficulty level l.;
      Invoke the task-specific model trainer to train the downstream task model with the training augmentation $\mathcal{T}$;
    end
    Increase l by 1 to the next level at the end of current curriculum;
  end
end

Algorithm 2: Curriculum-aware Paraphrase Generation Module

Input: A single training instance with textual input x; difficulty level l;
Output: Cache the generated paraphrases into $\bar{\mathcal{D}}$ or retrieve an augmented training instance $\bar{x}$;
if *a cached augmentation exists* **then**
  Retrieve $\bar{x}$ that corresponds to x with the difficulty measure $d = l$;
else
  Invoke the paraphrase candidate generator integrated with bottom-k sampling to generate a bag of paraphrase candidates for x;
  Invoke the mutual implication classifier for each paraphrase candidate to obtain corresponding binary indicator against the input sentence;
  Calculate the textual similarity for each paraphrase candidate against the input;
  Filter the generated paraphrase candidates with the mutual implication and the textual similarity using Equation 3;
  Assign a difficulty measure d to the filtered paraphrases with Equation 4;
  Cache the augmentation results into $\bar{\mathcal{D}}$;
end

difficulties (Bengio et al., 2009; Weinshall et al., 2018; Su et al., 2021), our method follows the other stream of works that applies transformation on the original data with dedicated difficulty level (Korbar et al., 2018; Ganesh and Corso, 2020; Wei et al., 2021). The closest work to ours is Wei et al. (2021). Their work does not consider paraphrasing and focuses on text classification only.

3 Our Proposed Framework

3.1 Background of Curriculum Data Augmentation (CDA)

Existing CDA (Wei et al., 2021) varies the word-level perturbation ratio to achieve different levels of difficulties under curriculum learning with simple word perturbation strategies such as synonym replacement, random insertion, swap, and deletion. As illustrated in Figure 1, such simple word perturbation strategies create problematic instances that break the sentence grammar, which can hamper the model performance. There are two common CDA strategies. One is called two-stage curriculum, which uses a fixed perturbation ratio for a single curriculum as the second stage after training with the original data. The other one is called gradual curriculum. It uses different ratios for a number of (typically 5) curriculums with increasing difficulties. However, such a learning strategy ends after passing through all the curriculums only once, and catastrophic forgetting can happen.

3.2 Our Proposed PCC

We propose curriculum data augmentation with paraphrase augmentation known as **P**araphrasing with Bottom-k Sampling and **C**yclic Learning for **C**urriculum Data Augmentation (PCC). Algorithm 1 depicts an overview of the whole PCC framework. At the start of training, we generate cached training augmentation for the entire dataset with our proposed curriculum-aware paraphrase generation module. Thereafter, we begin with the easiest curriculum. For each training instance, we retrieve the cached augmentation that has an equivalent difficulty measure with the current difficulty level. We then invoke the task-specific model trainer to train the downstream task model with the retrieved training augmentation. At the end of each curriculum difficulty level, we increase the difficulty level to advance to the next harder curriculum. In case it hits the end of the most difficult curriculum, we set the difficulty level to the easiest to start a new cycle. We propose such a cyclic learning strategy for mitigating potential catastrophic forgetting. In order to retrieve paraphrasing augmentation with appropriate difficulty measures, we propose a curriculum-aware paraphrase generation module.

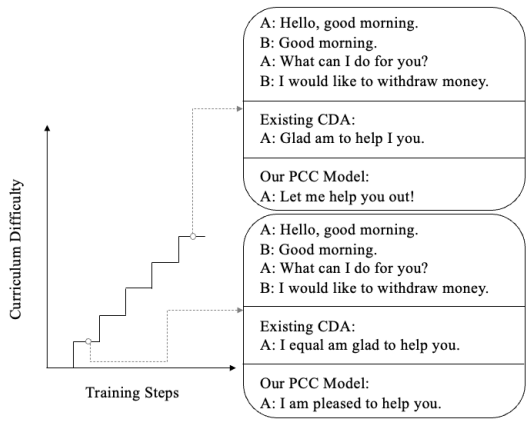

Figure 1: An illustrated example for our PCC model compared to existing CDA for dialogue generation. The original sentence is 'I am glad to help you.'

Sample No.	Sample Text	Sim. Score
1)	I am glad to assist you.	0.888
2)	Let's help you. I am glad to help you.	0.619
3)	Thank you for contacting me. I am glad to help you.	0.371
4)	It is now my pleasure to help you.	−0.038
5)	Let me help you out!	−0.265
6)	Thank you for your question.	−0.506

Table 1: Paraphrases with mutual implication for an input 'I am glad to help you.'

3.2.1 Curriculum-aware Paraphrase Generation Module

Algorithm 2 depicts the curriculum-aware paraphrase generation module. Three components are designed, namely, a paraphrase candidate generator integrated with a bottom-k sampling strategy, a filtering mechanism, and a difficulty measure. The paraphrase candidates are generated and then passed to the filtering mechanism. Finally, the filtered paraphrases are assigned a difficulty measure which represents to which curriculum difficulty level the augmentation belongs.

Paraphrase Candidate Generator with Bottom-k Sampling In order to generate mutual implicative paraphrases for the purpose of curriculum data augmentation, we adopt a Seq2Seq (Sutskever et al., 2014) generator which receives an input sentence x and generates the paraphrases $\bar{x}$ in an autoregressive manner (Nighojkar and Licato, 2021). During training, the paraphrase candidate generator is trained by maximising the following likelihood:

$$P\left(\bar{x} \mid x\right) = \prod_{t=1}^{T} P\left(\bar{x}_t \mid \bar{x}_1, ..., \bar{x}_{t-1}, x\right),$$

where T represents the token length of the paraphrase and x_t represents the word at the position t that has been inferenced.

Traditional sampling methods such as top-k sampling (Fan et al., 2018) and top-p sampling (Holtzman et al., 2020) sample the next token to be presented in the output from the most probable vocabularies that dominate the probability distribution. For example, at the i-th timestep during inference, top-k sampling samples the next token $\bar{x}_i$ from the most probable k words with the distribution:

$$P_{\bar{x}_i \in \mathcal{V}^{(k)}}(\bar{x}_i \mid \bar{x}_1, ..., \bar{x}_{i-1}, x), \quad (1)$$

where $\mathcal{V}^{(k)}$ represents the most probable k words. However, they are not suitable for generating super-hard instances, i.e., their output paraphrases tend to be textually similar to the original input sentence.[2]

To avoid coping the words and unearth the super-hard paraphrases to be used in later curriculums, we propose bottom-k sampling[3] which excludes a small set of dominating words for the sampling process. Note that we still use the combination of top-k and top-p sampling to generate easier samples for earlier curriculums. Formally, bottom-k modifies the distribution in Equation 1 to:

$$P_{\bar{x}_i \in \mathcal{V} \setminus \mathcal{V}^{(k)}}(\bar{x}_i \mid \bar{x}_1, ..., \bar{x}_{i-1}, x), \quad (2)$$

where $\mathcal{V}$ represents the whole vocabulary. Then, at each time step, we sample the next token with the rescaled distribution in Equation 2. We apply bottom-k for the first $\mathcal{N}$ steps of the generation before fallback to top-k and top-p. Bottom-k tends to generate paraphrases with lower textual similarity. For example, given an input of 'I like to remodel homes', existing sampling methods can generate an output 'Renovations in property I like to remodel homes'. In contrast, bottom-k sampling generates 'Is this what I want to see? Renovating homes are the best choices I have ever had.' where the latter one has a higher difficulty measure. Appendix F presents an extensive analysis.

[2]We found that top-k and top-p sampling tend to copy dominating words from the input into the paraphrases. This is also the reason why we prefer bottom-k over bottom-p, as we would like to effectively prevent from coping dominating words. Appendix F presents a detailed analysis.

[3]We give it such a name to make it catchy. It does not sample from the bottom k words. It samples from the bottom $|\mathcal{V}| - k$ words where $\mathcal{V}$ represents the whole vocabulary.

Paraphrase Filtering The inferential properties or mutual implication (MI) has been argued as a form of equivalent meaning (Boghossian, 1994; Peregrin, 2006), i.e., each sentence should entail each other to be 'paraphrases'. To support curriculum data augmentation, we exploit mutual implicative paraphrases for grammatical and lexical richness. Algorithm 2 (Lines 5, 6, and 7) depicts the filtering mechanism we propose to generate MI paraphrases. In order to determine the MI relationship between a pair of paraphrase $(x, \bar{x})$, we adopt a pre-trained MI classifier $\mathcal{M}(\cdot, \cdot)$ to calculate a binary indicator $\mathcal{M}(x, \bar{x})$. Here, non-MI paraphrases have a score of 0 and MI paraphrases have a score of 1. We also adopt a pre-trained model $\mathcal{G}(\cdot, \cdot)$ to evaluate the textual similarity score of the paraphrases as $\mathcal{G}(x, \bar{x})$. Here, paraphrases with lower similarity scores are treated as grammatically and lexically less similar to the original input sentence. We filter the paraphrase $\bar{x}_i$ based on these two scores:

$$\mathcal{M}(x, \bar{x}_i) + (1 - \mathcal{M}(x, \bar{x}_i))\mathbb{1}(\mathcal{G}(x, \bar{x}_i) \geq \beta). \quad (3)$$

In the formula above, β is a threshold for textual similarity. Here, a paraphrase with a positive mutual implication has a binary output of 1, i.e., it is preserved regardless of its textual similarity score. A paraphrase with a negative mutual implication but high textual similarity also has a binary output of 1, meaning it is preserved as well. In this way, MI paraphrases can be produced. We preserve highly similar paraphrases classified as non-MI, which is a misclassification by the classifier.[4] All paraphrases that are non-MI with low textual similarity have a binary output of 0, meaning we discard those paraphrases. After the filtering, a difficulty measure is computed for each paraphrase.[5]

Difficulty Measure Recall that for a pair of paraphrase $(x, \bar{x})$, we adopt a pre-trained textual similarity model $\mathcal{G}(\cdot, \cdot)$ to calculate its similarity score as $\mathcal{G}(x, \bar{x})$. BLEURT (Sellam et al., 2020) score, a BERT-based pre-trained model, is employed as the textual similarity model $\mathcal{G}(\cdot, \cdot)$. Here, paraphrases with lower similarity scores are treated as more difficult instances with higher difficulty measures. For further illustration, we present 6 samples generated from our model in Table 1 with descending order sorted on the similarity scores. Here, the similarity scores decently represent the grammatical and lexical difference between the paraphrases candidates, and the mutual implicative paraphrase candidates are grammatically (Sample 2, 3, 4, 5, and 6) and lexically (Sample 1, 2, 3, 4, 5, and 6) rich.

As the distribution of the similarity scores for the paraphrases varies for different inputs, we compute the difficulty measure for a paraphrase $\bar{x}_i$ with its rank in a sorted list of similarity scores, denoted as $\text{sort}(\cdot)$, in descending order among a bag of paraphrase candidates $\mathcal{X}$:

$$d_i = \lceil \mathcal{C} \times \frac{\text{sort}_{\bar{x}_i \in \mathcal{X}}(\mathcal{G}(\bar{x}_i, x))}{|\mathcal{X}|} \rceil, \quad (4)$$

where $\mathcal{C}$ represents the total number of curriculum difficulty levels we define, and $|\mathcal{X}|$ represents the total number of paraphrase candidates we have. Here, the paraphrase $\bar{x}_j$ with the highest similarity score, i.e., $\mathcal{G}(x, \bar{x}_j) = \max_{\bar{x}_i \in \mathcal{X}} (\mathcal{G}(\bar{x}_i, x))$, has a rank of 1, therefore, $d_j = 1$. The paraphrase $\bar{x}_k$ with the lowest similarity score, i.e., $\mathcal{G}(x, \bar{x}_k) = \min_{\bar{x}_i \in \mathcal{X}} (\mathcal{G}(\bar{x}_i, x))$, has a rank of $|\mathcal{X}|$, thus $d_k = \mathcal{C}$. Consequently, a larger rank indicates that the paraphrase is more grammatically and lexically different than the original input, and thus belongs to a harder curriculum. We set $d_i = 0$ as the easiest difficulty level for the original data.

3.2.2 Cyclic Curriculum Data Augmentation

Wei et al. (2021) proposed curriculum data augmentation with a gradual course. The training ends after passing the curriculums once. We found that a typical problem called catastrophic forgetting (Kirkpatrick et al., 2017) can hamper the performance during such a gradual course, meaning that the model can gradually forget the knowledge learned in an easier course. The augmentation for later curriculums is a subtask of an easier curriculum and can have lexical overlaps. Formally, the input samples x^{t+1} can have overlapping lexical x_i^t which are the same as x_j^t, where t and $t+1$ represent the curriculum difficulty levels, and i and j represent the word positions in the sentence. Due to catastrophic forgetting, the model can forget what it has learned earlier. Hence, we propose cyclic learning as shown in Algorithm 1 to inform the model which

[4] We postulate it as a flaw introduced by the imbalanced training data with a larger portion of paraphrases that tends to be textually unsimilar against the original sentence. We found in our early experiments that removing these easier examples obviously degrades the results for COVID-Q from 51.7 to 50.0. Furthermore, ignoring non-MI easy examples prevents PCC from collecting enough augmentation for AMZN.

[5] As in Appendix A, we use an off-the-shelf paraphrase generator and MI classifier in our experiments.

skills would be useful later before retrospecting to easier curriculums with lower difficulties.

4 Experimental Setup

In our experiments, we define six curriculums ranging from 0 to 5. 0 represents the original data, and 1 and 5 represent the easiest and the most difficult curriculum respectively.[6]

4.1 Few-shot Text Classification Task

For the downstream application task for our experiments, we follow Wei et al. (2021) to conduct the task of few-shot, highly multi-class text classification (Gupta et al., 2014; Kumar et al., 2019), which typically has a large number of classes with only a few samples for each of the class. We use triplet loss, a loss computed with three elements, namely, an anchor a, a positive sample p, and a negative sample n. It origins from the vision community (Schroff et al., 2015), which was later applied to language tasks (Ein Dor et al., 2018; Lauriola and Moschitti, 2020), suitable for the few-shot setting. Precisely, the learning objective is defined as:

$$\mathcal{L} = \mathcal{D}(a, p) - \mathcal{D}(a, n) + \gamma,$$

where $\mathcal{D}$ represents a distance measure that computes the distance between the input encodings. γ represents the margin between the positive and negative samples. We use BERT-based (Devlin et al., 2019) pooled sentence encodings as the input into a two-layer triplet network (Schroff et al., 2015).

Three datasets for the text classification task are used in our experiments, namely, HUFFPOST (Misra, 2018; Misra and Grover, 2021), COVID-Q (Wei et al., 2020), and AMZN (Yury, 2020). For space reasons, we leave their detailed dataset description in Appendix B.

4.2 Dialogue Generation Task

The second downstream task for our experiments is open-domain dialogue generation. We adopt a Seq2Seq neural network (Sutskever et al., 2014) which receives a text concatenation of prepended knowledge k and dialogue context c and generates the dialogue response r in an autoregressive manner (Radford, 2018). We train our dialogue generator by maximising the following likelihood:

$$P\left(r \mid k, c\right) = \prod_{t=1}^{T} P\left(r_t \mid r_1, ..., r_{t-1}, k, c\right),$$

where T represents the length of the generated dialogue response and r_t represents the word at the position t that has been inferenced. Typical prepended knowledge include personal traits (Zhang et al., 2018) and movie description (Zhou et al., 2018). We use DialoGPT (Zhang et al., 2020b) for parameter initialization for PCC.

We use PERSONACHAT (CONVAI2, Zhang et al. 2018) as the dataset for dialogue generation, which is described in Appendix C.

4.3 Baselines for Text Classification

We use the following baselines from existing data augmentation methods for text classification.

Triplet Loss As described in Section 4.1, an anchor, a positive example and a negative example is selected to construct the loss (Schroff et al., 2015).

Token Substitution It substitutes words with their WordNet synonyms (Zhang et al., 2015b; Feinerer and Hornik, 2020).

Pervasive Dropout It uses dropout on words with probability $p = 0.1$ (Sennrich et al., 2016a).

SwitchOut It replaces words with uniformly sampled words (Wang et al., 2018).

Round-trip Translation It translates sentences into another language before translating back into the source language (Sennrich et al., 2016b).

Hard Negative Mining + EDA It combines hard negative mining (Schroff et al., 2015) that chooses hard negative samples and EDA (Wei and Zou, 2019) that employs synonym replacement, word-level random insertion, deletion, and swap.

Hard Negative Mining + EDA + Gradual Curriculum It gradually increases the temperature for EDA augmentation (Wei et al., 2021).

4.4 Baselines for Dialogue Generation

We use the following baselines and data augmentation methods for dialogue generation.

TransferTransfo A Transformer-based model fine-tuned on PERSONACHAT (Wolf et al., 2019).

PerCVAE It uses a memory-augmented architecture with a conditional variational autoencoder to exploit persona information (Song et al., 2019).

DialoGPT It refers to an autoregressive dialogue generator introduced by Zhang et al. (2020b).

[6] We release the code and resource at https://github.com/HongyuanLuke/PCC.

CDA It refers to the curriculum data augmentation technique proposed by Wei et al. (2021) using the augmentation of EDA (Wei and Zou, 2019).

Official & Flatten It refers to the paraphrase augmentation technique that is task-specific to the task-oriented dialogue generation (Gao et al., 2020). To adapt it to our task, we use our generated paraphrase via mutual implication, denoted as Flatten, and the official revised PERSONACHAT paraphrases, denoted as Official.

Round-trip Translation It translates the input into another language before translating back (Sennrich et al., 2016b).

4.5 Evaluation Metrics

For the text classification task, we follow Wei et al. (2021) to use the top-1 accuracy as the metric.

For the dialogue generation task, we use the word-level F1 score, and we adopt the well-known sequence evaluation metric BLEU (Papineni et al., 2002) where we report BLEU-2, BLEU-3 and BLEU-4. We also adopt another well-known sequence evaluation metric, ROUGE, where we report the F-measures for ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004).

To verify our claim that bottom-k sampling generates grammatically and lexically rich paraphrases, we adopt Distinct-N (Li et al., 2016; Gao et al., 2019) with both $N \in \{1, 2, 3\}$ and $N \in \{4, 5, 6\}$ to measure the lexical and grammatical richness respectively using the ratio of distinct N-grams against the total number of N-grams generated.

5 Results and Analysis

5.1 Few-shot Text Classification Results

5.1.1 Main Results

Table 2 presents the results for few-shot text classification. Among the baselines, Triplet Loss + Gradual Curriculum works the best (Wei et al., 2021). PCC improves this baseline significantly. All the models share randomness in data, and our model is the best on all of the random seeds individually. Further, our proposed PCC model surpasses the baselines of Token Substitution, Pervasive Dropout, SwitchOut and Round-trip Translation significantly. Without bottom-k, PCC surpasses all the baselines, and our proposed full model with bottom-k obviously boosts performance. Appendix G additionally presents an analysis of the improvements as a function of the number of data augmentations.

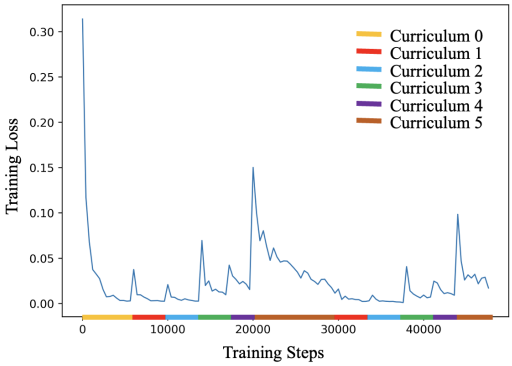

Figure 2: A plot of the training loss for the analysis for cyclic learning. Best viewed in color.

5.1.2 Ablation Study

Table 4 presents the results of our ablation study. First, removing the MI paraphrase filtering component described with Equation 3 obviously degrades the results. Replacing bottom-k sampling with pure sampling also decreases the results. Furthermore, paraphrasing in a random or an inverse order of decreasing difficulties, i.e., with neither curriculum learning nor cyclic learning, obviously deteriorates the results. *Therefore, our contribution is the discovery of paraphrasing as an effective CDA method rather than using paraphrasing solely as an augmentation technique.* Moreover, using cyclic learning instead of the gradual curriculum improves the results when trained with and without bottom-k sampling. Training the second cycle in an inversed order of decreasing difficulties degrades the results both with and without bottom-k.

5.2 Analysis on Cyclic Learning

Figure 2 presents the change of the training loss during the progress of the training on the task of text classification on COVID-Q. We observe that catastrophic forgetting exists as the training loss spikes when re-entering the curriculums. For the second time it enters the most difficult curriculum 5, the loss is also further smoothened compared to the first spike. The spike is also desirable as described in Wei et al. (2021), indicating that new instances that are harder to learn are presented and can help to escape the local minima. These support the usefulness of our proposed cyclic learning that can smoothen the gradients, mitigate catastrophic forgetting, and improve generalization by entering curriculums multiple times.

Model	HUFFPOST	COVID-Q	AMZN	Average
Triplet Loss (Schroff et al., 2015)	20.9 ± 1.0	39.7 ± 1.0	11.6 ± 0.6	24.1
Triplet Loss + Token Substitution (Zhang et al., 2015b)	22.7 ± 1.4	43.9 ± 1.3	12.8 ± 0.7	26.5
Triplet Loss + Pervasive Dropout (Sennrich et al., 2016a)	23.1 ± 1.1	43.5 ± 1.8	13.0 ± 0.6	26.5
Triplet Loss + SwitchOut (Wang et al., 2018)	22.9 ± 0.5	41.5 ± 0.6	12.7 ± 0.8	25.7
Triplet Loss + Round-trip Translation (Sennrich et al., 2016b)	24.2 ± 0.7	42.3 ± 1.0	13.0 ± 0.4	26.5
Triplet Loss + Hard Negative + EDA (Wei and Zou, 2019)	22.6 ± 1.8	48.2 ± 0.9	13.7 ± 0.9	28.2
↪ + Gradual Curriculum (Wei et al., 2021)	23.8 ± 0.9	48.9 ± 0.9	14.4 ± 1.5	29.0
PCC with Cyclic Curr. w/o Bottom-k	25.2 ± 1.5	51.4 ± 0.8	17.4 ± 0.7	31.3
PCC with Cyclic Curr. w/ Bottom-k	$\mathbf{25.9 \pm 1.7}$	$\mathbf{51.7 \pm 0.6}$	$\mathbf{18.2 \pm 1.0}$	**31.9**

Table 2: Results in top-1 accuracy for the downstream task of text classification on three datasets. The best results are bolded. We report the results averaged from five random seeds for data selection ranging from 0 to 4, which is the source of the variance here. Our methods report the best performance on all the random data seeds on all the datasets. A combination of top-k and top-p sampling with $k = 120$ and $p = 0.95$ is used for the penultimate row.

Model	F1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
TransferTransfo (Wolf et al., 2019)	16.61 ± 0.09	3.16 ± 0.07	1.04 ± 0.03	0.43 ± 0.02	17.69 ± 0.14	3.96 ± 0.08	16.34 ± 0.13
PerCVAE (Song et al., 2019)	14.33 ± 0.12	1.23 ± 0.06	0.20 ± 0.05	0.04 ± 0.01	13.25 ± 0.10	1.62 ± 0.05	12.02 ± 0.10
DialoGPT (Zhang et al., 2020b)	18.58 ± 0.13	5.25 ± 0.08	1.89 ± 0.07	0.66 ± 0.05	18.42 ± 0.13	4.62 ± 0.09	17.23 ± 0.12
DialoGPT + CDA (Wei and Zou, 2019)	18.38 ± 0.10	5.23 ± 0.10	1.84 ± 0.08	0.63 ± 0.02	18.55 ± 0.31	4.63 ± 0.11	17.40 ± 0.30
DialoGPT + Flatten (Gao et al., 2020)	18.21 ± 0.21	5.03 ± 0.18	1.85 ± 0.11	0.65 ± 0.04	17.97 ± 0.34	4.45 ± 0.16	16.84 ± 0.28
DialoGPT + Official (Gao et al., 2020)	18.12 ± 0.11	4.80 ± 0.27	1.78 ± 0.50	0.59 ± 0.60	17.88 ± 0.24	4.38 ± 0.09	16.84 ± 0.20
DialoGPT + RT (Sennrich et al., 2016b)	18.26 ± 0.49	5.10 ± 0.21	1.80 ± 0.20	0.62 ± 0.08	18.32 ± 0.35	4.47 ± 0.18	17.16 ± 0.31
PCC with Cyclic Curr. w/o Bottom-k	18.76 ± 0.20	5.38 ± 0.14	1.99 ± 0.9	0.71 ± 0.06	18.81 ± 0.18	4.75 ± 0.12	17.53 ± 0.12
PCC with Cyclic Curr. w/ Bottom-k	$\mathbf{18.80 \pm 0.45}$	$\mathbf{5.59 \pm 0.17}$	$\mathbf{2.07 \pm 0.12}$	$\mathbf{0.76 \pm 0.11}$	$\mathbf{19.15 \pm 0.16}$	$\mathbf{4.98 \pm 0.12}$	$\mathbf{17.89 \pm 0.17}$

Table 3: Results for the downstream task of open-domain dialogue generation on PERSONACHAT, averaged from three runs. All the metrics attain better quality with higher scores. We denote Round-trip Translation as RT. A combination of top-k and top-p sampling with $k = 120$ and $p = 0.95$ is used for the penultimate row.

Model	HUFFPOST	COVID-Q	AMZN
PCC w/o MI filtering	25.7 ± 1.4	50.2 ± 1.7	16.7 ± 1.1
PCC w/ Pure Sampling	25.8 ± 1.0	49.7 ± 0.9	16.9 ± 0.8
PCC w/ Inverse Curriculum	23.0 ± 1.7	48.5 ± 1.2	15.0 ± 0.5
PCC w/ Random Curriculum	24.0 ± 1.7	48.9 ± 1.5	15.1 ± 0.8
PCC w/ Gradual Curriculum	24.7 ± 1.3	49.6 ± 1.4	16.5 ± 0.7
PCC w/ Inv. Cyc.	24.9 ± 1.2	50.9 ± 1.0	16.5 ± 0.8
PCC w/ Cyc.	25.2 ± 1.5	51.4 ± 0.8	17.4 ± 0.7
PCC w/ Inv. Cyc., Bottom-k	25.3 ± 1.9	51.3 ± 1.1	17.1 ± 1.2
PCC w/ Cyc., Bottom-k	$\mathbf{25.9 \pm 1.7}$	$\mathbf{51.7 \pm 0.6}$	$\mathbf{18.2 \pm 1.0}$

Table 4: Ablation results in top-1 accuracy for the downstream task of text classification.

5.3 Dialogue Generation Results

Table 3 presents the results for dialogue generation on PERSONACHAT. First, we present the results for competitive baselines, namely TransferTransfo and PerCVAE. DialoGPT surpasses these two significantly. Using CDA on DialoGPT has deteriorated BLEU scores, which suggests that using CDA causes grammatical influence, possibly due to the random operations that produce undesirable grammatically incorrect augmentation. We also observe a large variance with the official paraphrase provided by PERSONACHAT, possibly due to the large difference between the manually rephrased sentences. This indicates easier paraphrases seem to be essential for PCC to be effective. Also, the Flatten baseline reported in Table 3 approximates a random curriculum, which degrades the results. It leads to a conclusion about the usefulness of the suggested curriculum. Round-trip Translation (RT) seems not effective, which is somehow reasonable as RT was originally designed for machine translation. PCC achieves the best among all the models, suggesting its usefulness for dialogue generation. Appendix D provides in-depth reasonings on the results. Appendix H presents a human evaluation of the downstream task of dialogue generation.

5.4 Analysis on Bottom-k Sampling

Table 5 presents the automatic results for bottom-k sampling on PERSONACHAT. Here, bottom-k sampling attains the best on Distinct scores with

Model	D1	D2	D3	D4	D5	D6
Pure Sampling	0.187	0.571	0.788	0.881	0.919	0.932
Top-k&p (k=120, p=0.95)	0.145	0.481	0.711	0.826	0.877	0.897
Top-k&p (k=80, p=0.80)	0.125	0.415	0.634	0.762	0.825	0.850
Bot.-k (k=2, $\mathcal{N}$=1)	0.184	0.587	0.824	0.901	0.919	0.925
Bot.-k (k=10, $\mathcal{N}$=1)	0.199	0.630	0.860	0.926	0.940	0.943
Bot.-k (k=2, $\mathcal{N}$=5)	0.223	0.695	0.904	0.945	0.951	0.953
Bot.-k (k=5, $\mathcal{N}$=10)	0.251	0.786	0.950	0.967	0.969	0.970
Bot.-k (k=10, $\mathcal{N}$=15)	**0.262**	**0.851**	**0.971**	**0.978**	**0.979**	**0.979**

Table 5: Automatic results for bottom-k sampling on PERSONACHAT. **D** represents the Distinct-N scores.

Criteria	PCC w/o Bottom-k	PCC w/ Bottom-k
Gramma. Richness	34	66‡
Lexical Richness	33	67‡
Difficulty	34	66‡
Paraphrasing	50	50

Table 6: Human evaluation results for bottom-k in winning percentages. ‡ indicates the results as passing a two-tailed binomial significance test with $p < 0.0001$.

lower grams ($N \in \{1, 2, 3\}$), indicating its lexical richness. It also attains the best on Distinct scores with higher grams ($N \in \{4, 5, 6\}$), indicating its grammatical richness. This helps to generate super-hard instances. Note that the setting of bottom-k sampling employed in PCC with $k = 2$ and $\mathcal{N} = 1$ already gives the best overall diversity against previous sampling methods. Further increasing the value of k and $\mathcal{N}$ leads to higher diversity.

5.5 Human Evaluation on Bottom-k Sampling

We hired three experienced annotators who have degrees relevant to English Linguistics to conduct an evaluation on bottom-k sampling with PERSONACHAT. We present a questionnaire composed of 800 questions with 200 randomly sampled training instances with the paraphrases generated with and without bottom-k sampling to the annotators to compare model outputs under A/B testing:

- **(Grammatical Richness)**: *"Which paraphrase do you think is more grammatically different than the original input sentence?"*
- **(Lexical Richness)**: *"Which paraphrase do you think is more lexically different than the original input sentence?"*
- **(Difficulty)**: *"Which paraphrase is more difficult to read and understood?"*
- **(Paraphrasing)**: *"Which one is more like a mutual implicative paraphrase to the input?"*

Table 6 presents the results of our human evaluation. The paraphrases generated by PCC with bottom-k sampling have a significant advantage in lexical and grammatical richness. Such an advantage correlates well with the difficulty of the paraphrases to be understood by human annotators. Furthermore, bottom-k does not hurt the paraphrasing performance compared to the top-k and top-p sampling. The result of human evaluation verifies our claim that bottom-k generates super-hard paraphrases with grammatical and lexical richness. Appendix F presents how bottom-k sampling is superior over previous methods in our scenario with case studies about the coping mechanism.

6 Conclusions

We propose a novel framework that uses mutual implicative paraphrasing as a curriculum data augmentation technique. Our proposed curriculum-aware paraphrase generation module is composed of three components, a paraphrase candidate generator with a bottom-k sampling strategy for generating superhard paraphrases, a paraphrase filtering mechanism, and a difficulty measure. We propose a bottom-k sampling strategy to effectively generate super-hard instances with grammatical and lexical richness to be used for the later stages in curriculum learning. Moreover, we propose a cyclic learning strategy that mitigates catastrophic forgetting. Experimental results on the task of few-shot text classification as well as dialogue generation support our proposed methodology PCC's usefulness, surpassing several competitive baselines.

Limitations

The proposed PCC cost more computational resources than traditional CDA methods. However, the cost is still affordable. Generating a round-trip augmentation used as one of the baselines costs about 1.5 seconds (1x speed) for PERSONACHAT. In contrast, generating a single paraphrase costs about 0.40 seconds (3x faster) with PCC on our machine with a single GPU.

Ethical Statement

We honour and support the EACL Code of Ethics. The datasets used in this work are well-known and widely used, and the dataset pre-processing does not make use of any external textual resource. In our view, there is no known ethical issue. End-to-end pre-trained dialogue generators are also used,

which are subjected to generating offensive context. But the above-mentioned issues are widely known to commonly exist for these models. Any content generated do not reflect the view of the authors.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*, pages 41–48.

Paul A. Boghossian. 1994. Inferential role semantics and the analytic/synthetic distinction. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 73(2/3):109–122.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019a. Skeleton-to-response: Dialogue generation guided by retrieval memory. 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 1219–1228, Minneapolis, Minnesota. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019b. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, Hong Kong, China. Association for Computational Linguistics.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric from article sections using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Ingo Feinerer and Kurt Hornik. 2020. *wordnet: WordNet Interface*. R package version 0.1-15.

Madan Ganesh and Jason Corso. 2020. Rethinking curriculum learning with incremental labels and adaptive compensation.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6383–6390.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, Online. Association for Computational Linguistics.

Maya R. Gupta, Samy Bengio, and Jason Weston. 2014. Training highly multiclass classifiers. *J. Mach. Learn. Res.*, 15(1):1461–1492.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7774–7785, Red Hook, NY, USA. Curran Associates.

Sawan Kumar, Shweta Garg, Kartik Mehta, and Nikhil Rasiwasia. 2019. Improving answer selection and answer triggering using hard negatives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5911–5917, Hong Kong, China. Association for Computational Linguistics.

Ivano Lauriola and Alessandro Moschitti. 2020. Context-based transformer models for answer sentence selection.

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.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. *CoRR*, abs/1909.03087.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. Curriculum learning for natural answer generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4223–4229. International Joint Conferences on Artificial Intelligence Organization.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Task-level curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3861–3867. International Joint Conferences on Artificial Intelligence Organization. Main track.

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. *CoRR*, abs/1705.06476.

Rishabh Misra. 2018. News category dataset.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. 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 7106–7116, Online. Association for Computational Linguistics.

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

Jaroslav Peregrin. 2006. Meaning as an inferential role. *Erkenntnis (1975-)*, 64(1):1–35.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021. Dialogue response selection with hierarchical curriculum learning. 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 1740–1751, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *arXiv e-prints*, abs/1610.02424.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods*

in Natural Language Processing, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5493–5500, Online. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What Are People Asking About COVID-19? A Question Classification Dataset. *CoRR*, page CoRR:2005.12522.

Daphna Weinshall, Gad Cohen, and Dan Amir. 2018. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5238–5246. PMLR.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *CoRR*, abs/1901.08149.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *ICCS*.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020a. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. ACL.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2020b. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3977–3989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kashnitsky Yury. 2020. Hierarchical text classification of amazon product reviews.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiaoxi Mao, Yadong Xi, and Minlie Huang. 2020a. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3449–3460, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Australia. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Implementation Details

For text classification, we use the hyper-parameter settings as in Wei et al. (2021) for the gradual course, and we refer to their paper for the detailed settings. For our cyclic learning, we pass through the curriculums twice. We train the same number of steps for each curriculum as we did in the first pass for our second pass, and the remaining hyper-parameters are kept the same. For Token Substituion, Pervasive Dropout, SwitchOut, and Round-trip Translation, we follow Wei et al. (2021) to use the triplet network as the base model and use

a two-stage curriculum for those baselines. Following Wei et al. (2021), we include 20% original data whenever augmentation is used.

For dialogue generation, we use DIALOGPT-SMALL for parameter initialisation. We use a batch size of 4 and a gradient clip of 0.1. We use validation patience of 10 based on the validation loss. We use greedy decoding for all of our experiments. The above settings apply to all our baselines and our proposed model fine-tuned on DIALOGPT. We start to apply the augmentation after 130,000 steps for data augmentation methods. We train the first, second, third, fourth, and fifth curriculums with 60,000 steps. For Official, Flatten, and RT, we perform a two-stage curriculum as described by Wei et al. (2021). We set $\mathcal{N}$ and k as a small value (typically $\mathcal{N} = 1$ and $k = 2$) for bottom-k sampling. We perform a cyclic repetition for our proposed method for the same number of steps for each curriculum until early stopped.

During our experiments, we apply data augmentation methods on the entire textual input for text classification, and we apply data augmentation methods on the personas traits for persona-based dialogue generation. We employ an off-the-shelf pre-trained model for both the paraphrase generator and the MI classifier (Nighojkar and Licato, 2021).

For all of the datasets, we obtain 20 paraphrases after filtering, and we assign 4 paraphrases (Wei et al., 2021) to each of the curriculums we have. We use 2 paraphrases obtained with bottom-k sampling for COVID-Q and we use 4 paraphrases obtained with bottom-k sampling for the remaining datasets.

For our models without bottom-k sampling, we use 20 paraphrases generated with a combination of top-k sampling and top-p sampling with $k = 120$ and $p = 0.95$ for all of the datasets.

We conduct our experiments for dialogue generation on the PARLAI platform (Miller et al., 2017).

B Datasets for Text Classification

- The HUFFPOST dataset is composed of 200k news headlines collected from 2012 to 2018, which is categorized into 41 classes such as politics, entertainment, and travel (Misra, 2018; Misra and Grover, 2021). We use all the classes and a 70% / 30% train / test split by class (Wei et al., 2021).

- The COVID-Q dataset is composed of 87 classes with several questions per cluster which ask about the same thing (Wei et al., 2020). We use the official train / test split with 3 questions per cluster (Wei et al., 2021).

- The AMZN product review dataset (Yury, 2020) categorizes products into given reviews. We consider the use of 318 'level-3' classes with at least 6 samples per product.

For the few-shot scenario, we need to set the number of samples in each class, N_c, to be used to construct the datasets. We use the setting in Wei et al. (2021) where $N_c = 3$ for COVID-Q and $N_c = 10$ for HUFFPOST. We set $N_c = 2$ for AMZN.

C Dataset for Dialogue Generation

CONVAI2 is an official competition built based on PERSONACHAT by adding new training examples as well as a hidden test set. For convenience, we denote the former as PERSONACHAT in the remaining of the paper. Since the test set is not publicly available, we use the official split containing a training / development split with 8,939 / 1,000 multi-turn dialogues conditioned on 1,155 / 100 personas respectively. Each persona is composed of about 4 to 5 persona traits.

D Analysis on Dialogue Generation

Table 3 reports an ablation when we use our PCC to train the dialogue generator without the use of bottom-k sampling. The results suggest that using bottom-k sampling improves all the metrics, especially the ROUGE scores. Table 8 presents the distribution of the textual similarity scores for the paraphrases generated from four methods on PERSONACHAT. The official paraphrase (Zhang et al., 2018) largely differs from the original ones, which we postulate as the reason for the large variance observed in Table 3. This also indicates the neccessity of the easier samples for curriculum learning. The Round-trip Translation generates paraphrases that have higher textual similarity with the input sentence. Our method without bottom-k sampling (we use a combination of top-k and top-p sampling with $k = 120$ and $p = 0.95$ here) generates paraphrases with more evenly distributed scores, with an average of 0.02. In contrast, bottom-k helps to generate harder samples while still capable of generating more easier samples.

E Problematic Cases for EDA

Table 7 presents samples from EDA for a sample input 'I am glad to help you.' with each of the

Sample Number	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$
i)	I equal am glad to help you.	I am glad help to you.	I am gald to happy help you.	To help you.	Glad am to help I you.
ii)	I am glad you help to.	I am gladiola to help you.	I am glad to assistance you.	Help glad am to i you.	I am gladiolus to helper you.
iii)	Am glad you.	I am glad help you.	I am glad you help to.	You I gald to help am.	I am glad help you.
iv)	I am glad to help you.	I am glad equal to help you.	I am glad to help you.	I am glad to happy happy help you.	I am happy to avail you.

Table 7: Randomly selected cases for an input 'I am glad to help you.' using Easy Data Augmentation (Wei and Zou, 2019). We present recommended temperatures τ ranging from 0.1 to 0.5, with four samples for each τ.

Model	$[0.5,$	$[0, 0.5)$	$(-0.5, 0)$	$, -0.5]$	Avg.
Official Paraphrases	1%	14%	33%	52%	−0.46
Round-trip Translation	25%	52%	17%	6%	0.23
PCC w/o Bottom-k	39%	11%	23%	27%	0.02
PCC w/ Bottom-k	16%	8%	18%	58%	−0.43

Table 8: Analysis on the distribution for the textual similarity score with different augmentation methods.

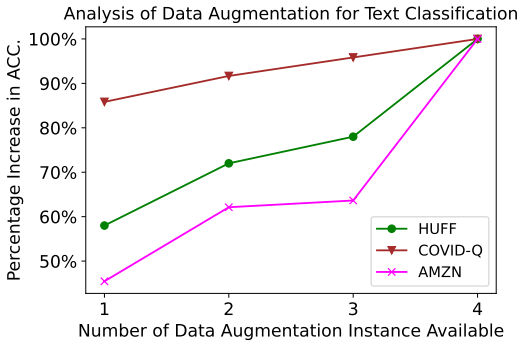

Figure 3: A plot of the percentage performance improvements of the downstream task of text classification against the number of data augmentation instances per curriculum. We use the first row in Table 2 as the baseline and the last row in Table 2 as the full improvements.

temperatures τ ranging from 0.1 to 0.5, which is the recommended setting from Wei et al. (2021). We categorize EDA's problems as the followings:

- Sample i) with $\tau = 0.1$ and sample ii) with $\tau = 0.2$ changes the meaning of the input sentence. 'equal' is possibly produced by random insertion and 'gladiola' is possibly produced by synonym replacement via WordNet (Feinerer and Hornik, 2020).
- Most of the samples produced with $\tau = 0.4$ and $\tau = 0.5$ breaks the grammar, which can be harmful to generation tasks.
- Sample ii) and iv) with $\tau = 0.5$ introduces rare words such as 'avail' and 'gladiolus', which is counterintuitive to see in many tasks.

As illustrated in Figure 1, PCC effectively reduces the above-mentioned issues.

F Analysis on Bottom-k Sampling

Table 9 presents extensive case studies to support that bottom-k sampling generates grammatically rich and lexically rich paraphrases. PCC without bottom-k tends to exploit a coping mechanism at the beginning of generation (Sample 2, 3, 5, 6, 7, 8, 9, 10, 11, 12). By excluding these dominating words to be copied for generation, bottom-k effectively emphasises the content (Sample 5), improves grammatical richness (Sample 1, 2, 3, 4, 5, 6, 7, 10, 12) and lexical richness (Sample 3, 4, 6, 8, 10, 12), does appropriate synonym replacement (Sample 8, 11) and insertion (Sample 4). Without bottom-k sampling, the input that starts with a first-person pronoun 'i' is highly likely to have an output that starts with 'i' (Sample 2, 3, 6, 8, 10). This constrains the model from generating grammatically rich paraphrases. In contrast, bottom-k sampling effectively reduces such cases and biases the generation towards a grammatically rich sampling space. Indeed, out of the 6,126 persona traits from PERSONACHAT, 5,087 of them start with 'i'. PCC without bottom-k generates 2,558 paraphrases that start with 'i', which avoids generating super-hard instances and hampers the PCC performance.[7] In contrast, bottom-k generates 205 paraphrases that start with 'i', indicating its usefulness in improving grammatical richness and generating super-hard instances. Avoiding coping helps to unearth the diverse paraphrases hidden in the tail vocabularies,

[7]Neither a pure top-p sampling with $p = 0.95$, a pure top-k sampling with $k = 120$, nor a greedy sampling helps, which generates 2,313, 2,381 and 3,302 paraphrases that start with 'i' respectively. Compared to these sampling strategies, bottom-k is an effective strategy in preventing copying.

Original Input Sentence	PCC w/o Bottom-k Sampling	PCC w/ Bottom-k Sampling
1): i like to shoot a bow.	When i first started shooting bows, this was the most important method.	Hey, i like to shoot a bow. Just started using a Bow SLR shooter, but a DSLR isn't really necessary.
2): i have four sisters.	i have four sisters	four sisters, and i want four sisters.
3): i believe that mermaids are real.	i believe that mermaids are real " @JesseyHawkins	Marxist philosopher,'mermaids are real," property
4): i work as a stand up comedian.	jesse t	trained comedian, I work as a stand up comedian.
5): my favorite drink is cuba libre.	My favorite beverage is Cuba libre.	Cuba is my favorite drink and I live in Cuba free.
6): i did a few small roles in tv series.	I have done a few small roles in tv series.	over the years i've appeared in a few small roles in television series
7): i love bicycling.	bicycle is my friend. i Love Bicycling.	how wonderful & amp ; inspiring! I love bicycling.
8): i own a hearse.	own a hearse. u could do that?	belongs to a hearse. it's not that expensive.
9): i like to listen to music.	i like to listen to music. How do you make up your mind?	I like to listen to music. by JACK CLINTON
10): i like to party.	I like to party	touts my ambition and passion for parties " by @MargotHillary by @anadulka @KelisStout
11): my favorite band is imagine dragons.	my favorite band is imagine dragons. I am just so happy about that.	i love this band it is awesome
12): i love to sing.	sing, am i love to sing	artist, i love to sing.

Table 9: Extensive case studies on PERSONACHAT support our claim that bottom-k sampling generates grammatically and lexically rich paraphrases that are more different than the input sentence.

which we postulate as the reason for the results observed in human evaluation in Section 5.5.

Note that we use bottom-k sampling to effectively prevent coping to generate instances that are textually more different to the input. There is a stream of work that considers improving the diversity (Vijayakumar et al., 2016). However, these works do not directly consider the similarity between the input paraphrase and the output paraphrase. This is the advantage of bottom-k sampling over this stream of work for our scenario.

G Analysis on Data Augmentation

Figure 3 presents the percentage improvements in accuracy as a function of the number of data augmentation instances available for each curriculum. Here, since we have 5 curriculum difficulty levels in our setting, having 3 instances available for each curriculum means that we have 15 data augmentations in total for each original sample. The improvements are positively correlated with the number of available instances. Furthermore, it seems that the improvements of PCC are not saturated yet. This means that a further increase in the number of data augmentations can lead to even higher performance than reported in our paper.

H More Human Evaluation

- **(Appropriateness)**: *"Who is more appropriate given the previous dialogue context?"*
- **(Informativeness)**: *"Who is more diverse instead of null answers such as I do not know?"*

Criteria	w/o PCC	w/ PCC
Appropriateness	49	**51**
Informativeness	45	**55**†
Engagingness	48	**52**
Human-likeness	49	**51**

Table 10: Human evaluation results for PCC in winning percentages. † indicates the results as passing a two-tailed binomial significance test with $p < 0.05$.

- **(Engagingness)**: *"Who would you prefer to talk with for a long conversation?"*
- **(Human-likeness)**: *"Which speaker do you think sounds more like a real person?"*

We follow Li et al. (2019) and Zou et al. (2021) to conduct a human evaluation of dialogue generation from the four aspects described above. We follow the settings used in Section 5.5 to invite three experienced annotators to mark 200 instances under A/B settings. The results in Table 10 indicate that PCC effectively improves the DIALOGPT baseline in all aspects, especially informativeness.

I Computing Infrastructure

We use an NVIDIA TITAN RTX with 24GB GPU memory for all of the experiments conducted in this paper. Training the text classification model consumes about 1 hour. Fine-tuning the dialogue generator consumes about 15 hours. Generating a single paraphrase to be used in PCC as a CDA method costs about 0.40 seconds on our machine.