Combining Curriculum Learning and Knowledge Distillation for Dialogue Generation

Qingqing Zhu[1], **Xiuying Chen**[2], **Pengfei Wu**[1] , **JunFei Liu**[1*], **Dongyan Zhao** [3*]
[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]Center for Data Science, AAIS, Peking University, Beijing, China
[3]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{Zhuqingqing,xy-chen,wpf9808,liujunfei,zhaody}@pku.edu.cn

Abstract

Curriculum learning, a machine training strategy that feeds training instances to the model from easy to hard, has been proven to facilitate the dialogue generation task. Meanwhile, knowledge distillation, a knowledge transformation methodology among teachers and students networks can yield significant performance boost for student models. Hence, in this paper, we introduce a combination of curriculum learning and knowledge distillation for efficient dialogue generation models, where curriculum learning can help knowledge distillation from data and model aspects. To start with, from the data aspect, we cluster the training cases according to their complexity, which is calculated by various types of features such as sentence length and coherence between dialog pairs. Furthermore, we employ an adversarial training strategy to identify the complexity of cases from model level. The intuition is that, if a discriminator can tell the generated response is from the teacher or the student, then the case is difficult that the student model has not adapted to yet. Finally, we use self-paced learning, which is an extension to curriculum learning to assign weights for distillation. In conclusion, we arrange a hierarchical curriculum based on the above two aspects for the student model under the guidance from the teacher model. Experimental results demonstrate that our methods achieve improvements compared with competitive baselines.

1 Introduction

Along with the enormous prosperity of social media on the Internet, there is a resurgent interest in developing open domain dialogue systems. However, the complexity of conversations crawled from the Internet may vary significantly. Sachan and Xing (2016); Cai et al. (2020); Lison and Bibauw (2017). To adapt to this phenomenon, some prior works (Cai et al., 2020; Sachan and Xing, 2016; Feng et al., 2019) employ curriculum learning (Bengio et al., 2009), in which a model is taught by using easy samples firstly and gradually adding more difficult ones. For example, (Cai et al., 2020) proposes an adaptive multi-curricula learning framework to train the dialogue model with easy-to-complex dataset based on various concepts of difficulty including the specificity and repetitiveness of the response, the relevance between the query and the response, etc. Also, Wan et al. (2020) resolves this problem by introducing self-paced learning (Kumar et al., 2010), which is a special kind of curriculum learning (Eppe et al., 2019). Wan et al. (2020) measures the level of confidence on each training example, where an easy sample is actually one of high confidence by the current trained model. Both curriculum learning and self-paced learning suggest that samples should be selected in a meaningful order for training. The difference is that curriculum learning uses pre-training or human intuitions while the emphasis of self-paced learning can be dynamically determined by model itself.

On the other hand, knowledge distillation (Hinton et al., 2015) is one of the most popular techniques to train efficient models, which aims to transfer the knowledge encoded in a pretrained teacher network into a student model. (Ba and Caruana, 2014) points that the teacher's probability predictions can capture the logarithm relationships between labels that are not obvious in the one-hot ground-truth label. Moreover, the teacher model can spread uncertainty over multiple outputs when it is not confident of its prediction. As a consequence, student models can yield significant performance boost under the guidance of a teacher. Since the knowledge from the teacher to student also has different difficulty degrees, it is intuitive to apply curriculum learning during this distillation process.

To our best knowledge, very little is known about how curriculum learning and knowledge distilla-

[*] Corresponding authors: JunFei Liu and Dongyan Zhao.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1284–1295
November 7–11, 2021. ©2021 Association for Computational Linguistics

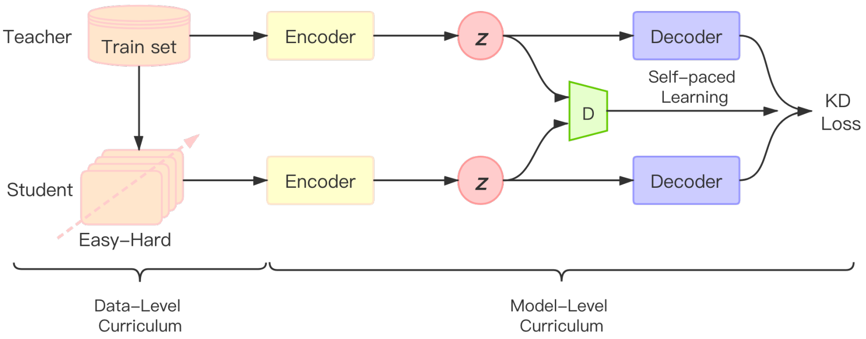

Figure 1: Architecture of our model.

tion work together. Hence, in this work, we propose a dialogue generation model that combines curriculum learning and knowledge distillation. Firstly, from the **data level**, we employ different types of features such as sentence length, word and utterance entropy, and coherence between dialog pairs to estimate the complexity of data for each training example. We preliminary cluster the training cases according to their data level complexity. Then based on these difficulty scores, we construct a learning curriculum for the student model. Secondly, from the **model level,** we employ an adversarial training strategy to evaluate the model-aware complexity. Concretely, the student model is adversarially trained to fool a discriminator, while the discriminator aims to distinguish the outputs from student and teacher networks. We measure the hardness of each sample by taking both the value and history of the discriminator into account, based on the following two intuitions. (1) The discriminator defines an objective of progressive difficulty (Doan et al., 2019), if the discriminator can successfully distinguish the output, then it is a hard case, and vice versa (Doan et al., 2019). (2) The model evolves during training and therefore additional evaluation pass to measure the change in a performance is needed (Matiisen et al., 2020). In this paper we consider the change in the discriminator. If the change is negative, this must mean that the sample is difficult to train. Then based on these model-level difficulty scores, we further transfer the knowledge from teacher to student network gradually by incorporating self-paced learning methodology.

Our contributions are summarized as follows: (1) We make an empirical study on the combination of curriculum learning methods and knowledge distillation for efficient dialogue generation models. (2) We arrange a hierarchical curriculum based on the above two aspects (data and model) for the distillation model. (3) We apply the proposed framework on two real-life open-domain conversation datasets, automatic and manual evaluation shows that our approach can be used to enhance the performance of dialogue models.

2 Related Work

Our work is at the intersection of curriculum learning (Bengio et al., 2009) and knowledge distillation (Hinton et al., 2015) for training dialogue generation models.

2.1 Neural Dialogue Generation

Since (Ritter et al., 2011) propose a data-driven approach that adopt phrase-based statistical machine translation for dialog system. more and more researchers have focused on generation-based conversation system. A popular framework for dialogue generation is using extra information such as conversation topics(Xing et al., 2017) , persona profile (Song et al., 2019), user emotions (Zhou et al., 2018), or out-sourcing knowledge (Liu et al., 2019) is introduced to benefit the dialogue model with more diverse response generation (Serban et al., 2017; Zhao et al., 2017; Gu et al., 2019). Latent variables also benefit the model with more diverse response generations (Zhao et al., 2017). This paper improve the dialogue model from a different angle that we make an empirical study on the combination of curriculum learning methods and knowledge distillation.

2.2 Knowledge Distillation

Our study is also related to the knowledge distillation method (Hinton et al., 2015), which employs a teacher model and tries to minimize the KL divergence between teacher distribution and student distribution. In (Romero et al., 2015), the student network is trained not only using the soft targets, but also using hints from the intermediate layers. Knowledge distillation was first introduced for classification tasks as a way to compress large networks into smaller models. Kim and Rush (2016) extend this to neural machine translation, and then (Zhang et al., 2020) has proposed further applications of dialogue generation task. However, these papers do not consider the order of the learning schedule. In a sense, our method is different from theirs because we borrow the idea of curriculum learning for knowledge distillation.

2.3 Curriculum Learning in NLP

Inspired by the human learning process, curriculum learning (Bengio et al., 2009) is proposed as a machine learning strategy by feeding training instances to the model from easy to hard. It has been applied to many NLP tasks. To name a few, (Sachan and Xing, 2016) propose and study other heuristics that define a measure of easiness and learn the curriculum by selecting samples using this measure. More recently, (Wang et al., 2019) learns a multi-Domain curriculum for neural machine translation. Xu et al. (2020) uses curriculum Learning to distinguish easy examples from difficult ones for natural language understanding by reviewing the trainset in a crossed way. Our paper is quite different from theirs because we arrange a hierarchical curriculum based on the above two aspects (data and model) for the distillation model.

3 Problem Formulation

The overall network architecture is shown in figure 1. The teacher and student model use the same basic architecture that is related to an encoder-decoder (Cho et al., 2014) generative dialogue model based on Variational Autoencoders (VAEs) (Kingma and Welling, 2014).

In our model, there are three elements: dialogue context $X = x_1, x_2 ... x_i$, response $Y = y_1, y_2 ... y_i$ and a latent variable z. The dialogue context X is composed of several history utterances. The response Y is the responses towards the given context. The latent variable z is used to capture the latent distribution over the replies with a standard Gaussian prior, which is defined as follows:

$$P(z) \sim \mathcal{N}(0, I). \quad (1)$$

Our task is to model the true probability of a response Y given an input X, which can be estimated as:

$$P(Y \mid X) = \int_z P(Y \mid z, X) P(z) dz. \quad (2)$$

The hierarchical curriculum strategy for distillation model consists of two parts: one is for the data-level and the other is for the model-level. In the data-level, easier context-response pairs are presented to the student model before harder ones. As for the model level, we design curriculum schedules to gradually transfer knowledge from the the teacher to student, which controls the difficulty of soften labels that are distilled from teacher to students. The samples that discriminator cannot differentiate between the output provided by the student and the teacher are assumed to be easier ones. Starting from easier samples, the model progressively strengthens its relation between the teacher and student models. In the rest of this paper, we give detailed descriptions of the proposed approach.

4 Model

4.1 Data-Level Curriculum

Following existing studies (Platanios et al., 2019; Kocmi and Bojar, 2017) that the model should be trained from easy samples to hard ones, we schedule the curriculum based on three intuitive notions of difficulty: response length (Serban et al., 2017; S. et al., 2017; Baheti et al., 2018), word and utterance entropy (Serban et al., 2017), coherence (Xu et al., 2018). These features compensate each other by capturing the information in a sentence pairs from different aspects. All these features are from previous research and here we integrate them together: we first use the method from these papers to compute the scores for individual sentences; then normalize the scores; finally add all these scores together as a total score. We rank all sentence pairs according to their scores, and we break down the dataset D_o into N subsets, in which those examples with similar complexity are categorized into the same subset.

4.2 Output Knowledge Distillation

Knowledge distillation describes a class of methods for the knowledge transfer from teacher network to student network. In our model, the student network S_θ is trained over the same architecture but different parameters as teacher model T_θ. The teacher has previously been trained, and we freeze its parameters when training the student network.

We transfer the knowledge from teacher to student by minimizing the similarity distance between the output of student network and the soft label generated by the teacher network. We use cross-entropy loss to measure the two logits as (Romero et al., 2015). To further improve the sequence-to-sequence student model, hard-assigned labels are also utilized. The final student network is trained to optimize the following compound objective:

$$\mathcal{L}_{KD}\left(\mathbf{S}_{\theta}\right) = \mathcal{H}\left(S_{\theta}(X), Y\right) + V\mathcal{H}\left(T_{\theta}(X), S_{\theta}(X)\right), \tag{3}$$

where $\mathcal{H}$ refers to the cross-entropy and V is a parameter to indicate the temperature of distillation. Later, we will use the method of the model level curriculum learning to process λ in section 2.5. Note that the first term in Equation (3) corresponds to the traditional cross-entropy between the softmax layer's output of a (student) network and word distribution in response Y, whereas the second term is to learn from the softened output of the teacher network to strengthen its supervision for the student.

In the teacher model, we train it by using all the dataset with original order, while in the student model, the training starts from the step that consists of examples with the lowest difficulty. After that, data in the next step is aggregated to the current training dataset.

4.3 Latent Space Knowledge Distillation

In order to guide the student's learning process of the output layer, we introduce hints (Romero et al., 2015), which are representations in the intermediate layer from the teacher network. Instead of adopting the classic student-teacher strategy of forcing the output of a student network to exactly mimic the soft targets produced by a teacher network, we introduce adversarial networks to transfer the knowledge from teacher to student. Due to the discrete nature of natural language tokens (Shen et al., 2017; Xu et al., 2017), it is difficult to pass the gradient update from the discriminator to the generator (Yu et al., 2017). So we choose to discriminate variable z in high level latent space rather than direct tokens (Gu et al., 2019).

During the process of latent space knowledge distill, we generate student' latent variable representation by training the student network S_{θ} and freezing the teacher parts adversarially against discriminators D. A discriminator D attempts to classify its input as teacher or student by maximizing the following discriminator loss (Goodfellow et al., 2014):

$$\mathcal{L}_{GAN} = W\left(q_{\phi}(z_t, x_t) \| p_{\theta}(z_s, x_s)\right), \tag{4}$$

where $W(\cdot\|\cdot)$ represents the Wasserstein distance between these two distributions (Arjovsky et al., 2017). We choose the Wasserstein distance as the divergence since the WGAN has been shown to produce good results in text generation (Zhao et al., 2018). z_t and x_t denote the latent variable and query representation in T_{θ}. z_s and x_s denote the latent variable and query representation in S_{θ}. Student network attempts to generate similar outputs which fools the discriminator D. D is implemented as a feed-forward neural network which takes as input the concatenation of z and x and outputs a real value.

4.4 Model-Level Curriculum

4.4.1 Model-Level Difficulty Evaluation

In the first step, we have selected data based on the definition of data difficulty. While in this step, we select the teachers' knowledge by using curriculum learning based on the performance of GAN. GAN can be said to share aspects with curriculum learning: the discriminator defines an objective of progressive difficulty (Doan et al., 2019). We consider two different metrics as scores for measuring generator progress in our curriculum approach, which is defined as follows: (1) **Discriminator evaluation**: $Score_i = D_i$. (2) **Discriminator change**: $Score_i = D_i - D_{i-1}$, where $Score_i$ is the difficulty score of the i th sample.

For comparison, we also use the loss value of distance between the output of teacher and student network to measure the sample difficulty, which is defined as follows: (1) **Loss value**: $Score_i = \mathcal{H}\left(T_{\theta}(x_i), S_{\theta}(x_i)\right)$. (2) **Loss change**: $Score_i = \mathcal{H}\left(T_{\theta}(x_i), S_{\theta}(x_i)\right) - \mathcal{H}\left(T_{\theta}(x_{i-1}), S_{\theta}(x_{i-1})\right)$.

4.4.2 Self-paced Learning

In this section, we aim to decide the order of output distillation. Not that all samples are distilled from teacher to student equally, but to start training from simple samples and gradually select complex samples to join the training process of the model. That is to say, we need to determine the value of V

Algorithm 1 Hierarchical Curriculum Learning Algorithm

```
Input: Dataset $D_o$ ;
Output: Student Model $\mathcal{S}_\theta$;
  Build and Pre-train the Teacher Model $\mathcal{T}_\theta$ and
  then freeze its parameters;
for train step t = 1,...n do
    Uniformly sample one subset of context-
      response pairs $B_t$ from $D_o$ based on the
      data-level curriculum to train $\mathcal{S}_\theta$;
    for $(x_i, y_i)$ in $B_t$ do
        Use a GAN to distill the latent
          variable by using Equation (4)
        Calculate the difficultly score;
        Acquire the self-paced learning
          arrangement and distill the output
          by using Equation (8);
    end for
end for
```

in Equation (3). The conventional self-paced learning selects the samples based on the loss value. While we replace it with our difficulty score described in the last section. Then we use self-paced learning to estimate V by the optimization as:

$$\min_{V} \sum_{i=1}^{n} v_i Score_i + f(\lambda, \mathbf{V}), \tag{5}$$

where $f(\lambda, \mathbf{V})$ determines the way to compute the value of v_i, λ is the self-paced adjustment parameter. lets $\mathbf{V} \in \{0,1\}^n$ and defines $f(\lambda, \mathbf{V})$ as:

$$f(\lambda, \mathbf{V}) = -\lambda \sum_{i=1}^{n} v_i. \tag{6}$$

The optimal V can be calculated by

$$v_i = \begin{cases} 1, \text{ if } Score_i < \lambda \\ 0, \text{ if } Score_i \geq \lambda, \end{cases} \tag{7}$$

where λ is used to control the learning pace of if self-paced learning.

In our paper, suppose T is the total number of training steps and t is the current training step. During training, to select the training instances with desired difficulty, we resort to a pre-defined pacing function $\lambda = f(t)$ to control how fast the output will be distilled from teacher to student. We define three different pacing functions named as linear-scheduler, log-scheduler and exp-scheduler to make a smooth transformation from teacher to student models and verify the effectiveness of the proposed model. Linear-scheduler is increased constantly in the training process. Log-scheduler indicates that the increased speed is from fast to slow, while exp-scheduler is opposite to it. We will compare the effects of these three methods in the next section.

In order to incorporate self-paced learning into the distillation process, we reformulate our objective function 3 as follows:

$$\begin{aligned} \mathcal{L}_{KD}\left(\mathbf{S}_\theta\right) = \sum_{i=1}^{n} (\mathcal{H}\left(\mathcal{S}_\theta(x_i), y_i\right) + \\ v_i \mathcal{H}\left(\mathcal{T}_\theta(x_i), \mathcal{S}_\theta(x_i)\right)). \end{aligned} \tag{8}$$

In conclusion, our hierarchical curriculum learning algorithm framework is described in Algorithm 1.

5 Experiment

5.1 Datasets

We conduct experiments on two English conversation datasets, which have been widely used in open-domain dialogue generation. (1) DailyDialog (Li et al., 2017): it is a collection of real-world daily conversations for an English learner in daily life. It is a multi-turn dataset, and we treat each turn as a single-turn training pair in this work. (2) PersonaChat (Zhang et al., 2018): it is collected by two crowdsourced workers chit-chatting with each other, conditioned on the assigned personas. In our experiments, we only use the conversation text and process it as DailyDialog.

5.2 Evaluation Methods

Automatic Evaluation Method It is challenging to assess the quality of the generated responses. In this paper, we adopt several evaluation methods to measure different aspects of our results: **BLEU** (Papineni et al., 2002): it is used as a reward to evaluate dialog systems by measuring word overlap between the generated reply and the ground truth for the final evaluation. We compute BLEU scores for $n <= 4$ using smoothing techniques [1]. **Entropy-based metrics** : it includes word and sentence entropy as (Serban et al., 2017), which suggests the diversity of responses. **Length**: as proposed by (Mou et al., 2016), the length of an utterance is an objective, surface metric that reflects the substance of a generated reply.

[1] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

Dataset	Method	BLEU	Sentence Entropy			Word Entropy			Length
			1	2	3	1	2	3	
DailyDialog	S2S	0.306	64.924	73.771	73.249	6.709	10.466	11.948	10.079
	CVAE	0.321	61.344	83.954	83.654	6.814	10.688	11.978	8.899
	KD	0.324	65.577	90.873	91.557	6.807	10.653	11.942	9.578
	Curriculum	0.326	65.057	89.625	90.263	**6.817**	10.686	12.004	9.450
	Ours	**0.357**	**96.189**	**134.722**	**145.972**	6.779	**11.568**	**12.904**	**14.336**
PersonaChat	S2S	0.319	65.012	81.221	90.021	6.505	9.959	10.262	9.151
	CVAE	0.329	76.401	81.588	99.398	6.581	10.049	10.207	9.921
	KD	0.334	79.722	84.633	100.79	6.824	10.242	12.197	11.153
	Curriculum	0.333	67.502	90.879	**102.092**	6.623	10.076	**12.381**	10.117
	Ours	**0.345**	**88.839**	**108.539**	96.217	**8.321**	**11.672**	11.724	**11.231**

Table 1: Results of the automatic evaluation on two datasets.

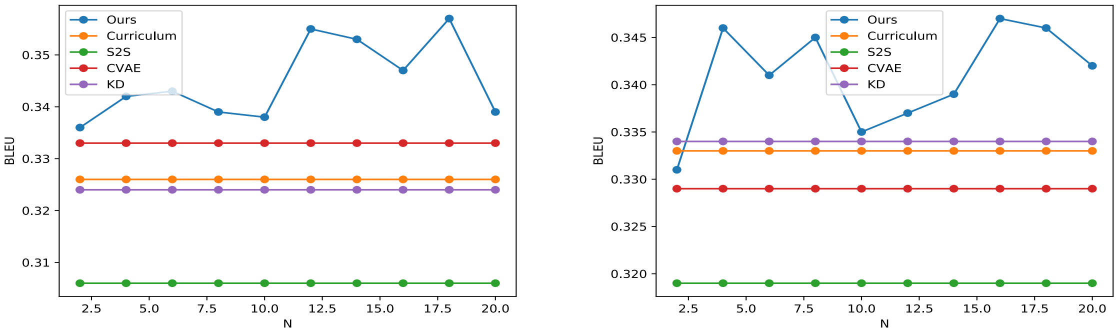

Figure 2: The comparison of BLEU score on two datasets.

Model	**0**	**1**	**2**	**Kappa**
S2S	42.7%	40.1%	17.2%	0.6354
CVAE	26.6%	43.9%	29.5%	0.5982
KD	23.8%	42.5%	33.7%	0.6302
Curriculum	26.7%	43.1%	30.2%	0.6750
Ours	24.1%	36.2%	39.7%	0.6632

Table 2: Results of the human evaluation on DailyDialog dataset.

Human Evaluation Method Considering the limitations of the existing automatic evaluation metrics, we also adopt human judgments. We use DailyDialog as the evaluation corpus since it is more similar to our daily conversations and easier for annotators to make the judgement. We randomly sample 100 cases and three well educated volunteers are recruited to do manual evaluation. For each query-reply pair, volunteers are asked to rate it with three levels: 0, 1, 2. 0 indicates that the selected sentences are either irrelevant or disfluent with grammatical errors; 1 is for the reply that is relevant but not informative enough; 2 means that the queries and replies are extremely related and the replies are natural. We calculate the ratio of each score (0, 1 and 2) for each model. To examine the agreements among all the volunteers, we also calculate the Fleiss kappa (Fleiss and Cohen, 2016) of the human annotations on all models.

5.3 Comparison Models

To ascertain the effectiveness and applicability of our approach, we re-implement experiments on these methods: (1) S2S: it is a sequence-to-sequence model with attention mechanism as in (Shang et al., 2015). (2) CVAE: it is a latent variable model using conditional variational autoencoder trained with KL annealing and a BoW loss as in (Zhao et al., 2017). (3) Curriculum (Cai et al., 2020): it employs an adaptive multi-curricula to schedule a committee of organized curricula for dialogue learning. (4) KD (Tahami et al., 2020): it uses two dialogue models as the student and the teacher. The framework uses a teacher-student setting where the student learns from both the ground-truth labels and the soft-labels provided by the teacher.

5.4 Training and Evaluation Details

For the teacher and student model, we use gated recurrent units (GRU) (Cho et al., 2014) for the RNN encoders and decoders. The encoder and decoder are both GRUs with 256 hidden units. The prior and the recognition networks are both 2-layer feed-forward networks of size 200 with tanh non-

DC	OD	MD	MC	BLEU	Sentence Entropy			Word Entropy			Length
					1	2	3	1	2	3	
-	-	-	-	0.315	60.892	77.883	85.351	6.231	9.381	10.711	9.011
-	+	-	-	0.329	77.485	74.433	74.922	6.741	10.232	11.822	10.915
+	+	-	-	0.327	75.873	91.031	99.744	6.509	9.961	12.257	11.043
+	+	+	-	0.339	86.40	91.58	99.398	6.581	10.049	12.207	12.309
+	+	+	+	**0.357**	**96.18**	**134.72**	**145.972**	**6.779**	**11.568**	**12.904**	**14.336**

Table 3: Results of the ablation study on the DailyDialog dataset.

DC	OD	MD	MC	BLEU	Sentence Entropy			Word Entropy			Length
					1	2	3	1	2	3	
-	-	-	-	0.320	67.991	81.887	87.132	6.225	9.251	9.820	9.011
-	+	-	-	0.326	74.924	83.774	90.253	6.709	10.466	11.948	10.079
+	+	-	-	0.334	75.749	90.474	**99.642**	6.562	10.016	**12.363**	11.043
+	+	+	-	0.337	80.539	92.191	81.999	6.755	10.238	11.615	11.109
+	+	+	+	**0.345**	**88.839**	**108.539**	96.217	**8.321**	**11.672**	11.724	**11.231**

Table 4: Results of the ablation study on the PersonChat dataset.

linearity. The dimension of a latent variable z is set to 64. The initial weights for all fully connected layers are sampled from a uniform distribution [-0.02, 0.02]. The generators as well as the discriminator D are 3-layer feed-forward networks with ReLU non-linearity and hidden sizes of 200, 200 and 400, respectively. The gradient penalty is used when training D (Nair and Hinton, 2010) and its hyper-parameter λ is set to 10. We set the vocabulary size to 20,000 and define all the out-of-vocabulary words to a special token $< unk >$. The word embedding size is 200. The longest utterance is set to 40. The baselines are implemented with the same set of hyper-parameters. All the models are implemented with Pytorch [2].

5.5 Evaluation Results

Automatic Evaluation Results The automatic evaluation results of our proposed method and baselines on the two datasets are shown in Table 1. We can see the following observations. (1) Our model outperforms the baselines regarding almost all the evaluation metrics on the two datasets. The overall performance of our model further supports our hypothesis that our model achieves a better trade-off on the whole. (2) Specially, in terms of BLEU scores, compared to the S2S, CVAE, KD and Curriculum, our model obtains impressive 16.7%, 11.2%, 10.2% and 9.5% performance gains on the DailyDialog. As for PersonaChat, our model outperforms the baseline with absolute improvements of about 8.2%, 4.9%, 3.3% and 3.6%. This indicates that our model generates more relevant responses with the highest BLEU scores on both datasets. (3) To show that our model is on average more diverse than other model responses, we compute the average sentence entropy and word entropy, and our model produces responses with higher entropy on both dataset compared to the other baseline models. In particular, we can see that the entropy of the sentences has been considerably enhanced. (4) We also report the average length of responses outputted by each model. Since long responses contain rich content, the results provided quantitative evidence to our claim that we can improve the responses with richer content than other models.

Human Evaluation Results The results of human evaluation against all baseline methods are listed in Table 2. The Kappa scores on all models are larger than 0.5, which indicates the correlation of the human evaluation. From the results we can again observe that, similar to the automatic evaluation results, our model consistently achieves the best performance, which further demonstrates the effectiveness of our proposed method.

6 Further Analysis

6.1 Ablation Study

There are four important parts in the proposed framework: Data Level Curriculum (DC), Output Distillation (OD), Middle Layer Distillation (MD), Model Level Curriculum (MC) and we remove them one at a time. Table 3 and 4 present the results of model variants by ablating specific parts of our model. Overall, we observe that all parts of our method lead to improvements, which further demonstrates the neural dialogue generation model not only benefits from curriculum learning

[2] https://pytorch.org/

Dataset	Method	BLEU	Sentence Entropy			Word Entropy			Length
			1	2	3	1	2	3	
DailyDialog	Ours-Lin	0.333	78.598	79.209	79.039	6.9018	10.72	12.153	9.58
	Ours-Log	0.343	85.555	84.701	83.275	6.9076	10.756	12.114	11.28
	Ours-Exp	**0.357**	**96.189**	**134.722**	**145.972**	6.779	**11.568**	**12.904**	**14.336**
PersonaChat	Ours-Lin	0.332	78.346	72.852	81.039	7.238	10.72	10.023	10.92
	Ours-Log	0.342	75.358	79.391	86.385	8.502	10.756	10.381	10.32
	Ours-Exp	**0.345**	**88.839**	**118.539**	96.217	**8.321**	**11.672**	**11.724**	**11.231**

Table 5: Results of different scheduler on the on two datasets.

Method	BLEU	Sentence-Entropy			Word-Entropy			Length
		1	2	3	1	2	3	
None	0.339	86.40	91.58	99.398	6.5809	10.049	12.207	12.309
Loss Value	0.341	61.344	83.954	83.654,	6.814	10.688	11.978	8.899
Loss Change	0.324	65.577	90.873	91.557	6.807	10.653,	11.942	9.578
D	0.326	65.057	89.625	90.263	**6.817**	10.686	12.004	9.450
D change	**0.357**	**96.189**	**134.722**	**145.972**	6.779	**11.568**	**12.904**	**14.336**

Table 6: Performance comparison on the DailyDialog dataset.

but also knowledge distillation. Specially, we find that the MC is slightly more important in overall performance. Meanwhile, without other parts also decreases the performance on most evaluation metrics, which further proves the effectiveness of combining these two techniques together.

6.2 The Effect of Different settings of subsets in Data Level Curriculum

We further explore the effects of different number of subsets for our data-level curriculum strategies, which also decides the granularity of sample selection in one epoch. Experiments are conducted on the proposed two datasets and we report BLEU scores in Figure 2. We select a wide range of choices: from 2 to 20. In general, its performance significantly outperforms the baseline system on the test set with different settings of subsets, which indicates that our approach is robust and effective. We also evaluate extreme situations. For example, when we divide our data set into 100 groups, the result is 0.295 on BLEU score (0.011 below our baseline with the worst effect), which is as expected because an over-small subset leads to the problem of overfitting.

6.3 The Effect of Different Schedular Functions

Since we design three pacing functions in model-level curriculum arrangement, we compare and analyze the proposed functions in experiments. We conduct experiments on the two datasets and the performance of different pacing functions can be found in Table 5. We have the following two observations. (1) The exp-scheduler method consistently outperforms others on two datasets. We suspect that is because in the case of the exp-scheduler function, the student network starts learning less from the teacher model and therefore has more time to learn a better discriminator. (2) Compared with other pacing functions, the linear-scheduler pacing function results in the worst performance, which indicates the effectiveness of changing learning speed.

6.4 The Effect of Different Model-Level Curriculum Strategy

To further glean the insights regarding the different model-level curriculum strategy, we present the results in Table 6. We can see that D change achieves the best results when compared to the baseline and other methods, which indicates that D does reflect the complexity of students' models compared to teachers'. The loss-based complexity performed worse than D and D change. We suspect that because the loss function is not a good signal to judge the model complexity compared to discriminator.

6.5 Case Studies

To empirically analyze the quality of generated responses, we present examples generated from our model and baselines in Table 7 . For each query, we show the best samples of generated responses from each model. On the table, we see that our model generates both long and informative replies compared with others.

Query	that is good . are you married ? yes and have two animals what kind of animals ? i love taking my dog for runs to stay in shape .
groud-truth	they are cats . they do not like being outside
S2S	that is cool . i am not sure . are you ?
CVAE	oh . do you have any pets ?
KD	i have a dog and i am not into it .
Curriculum	i am going to have a dog , what happened to your dogs ?
Ours	yeah dogs are cool, what kind of dog do you have ?
Query	i am fluent in three different languages . do you speak more than one language ? i do speak multiple languages . which ones do you speak ? i speak , english , spanish and french .
groud-truth	those happen to all be languages i speak . i want to visit france sometime .
S2S	I am not sure .
CVAE	I am not good for it
KD	yes, i am a teacher and you ?
Curriculum	i have a lot of time to speak .
Ours	that is awesome . where are you from ?

Table 7: Case studies of generated replies.

6.6 Error analysis

To enhance the performance of our model in the future, we take the worse cases in human judgment as an example to analyze our errors. We find that although our model improves the response diversity significantly, the model still has a “safe response" problem. Compared with the response generated by the teacher model, we find that the “safe response" generated by the teacher model can greatly affect students. 80.1% of the “safe response" is from the teacher model. That is, soft labels that are generated by a teacher model largely determine the performance of its student model. Therefore, in the future, we will study methods that can learn the good parts of the teacher model, and filter the bad parts of the teacher model.

7 Conclusion

In this work, we consider open-domain dialogue systems. To induce model learning from effective teachers, we propose a learnable distillation model to dynamically distill knowledge by hierarchical curriculum learning. Experiments conducted on two public conversation datasets show that our proposed framework is able to boost the performance of existing dialogue systems. Besides, our framework is not limited to the neural dialogue generation task. In the future, we would extend our method to deal with other text generation tasks (e.g., abstract summarization) and examine this approach’s adaptability to these tasks.

References

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR*, abs/1701.07875.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2654–2662.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3970–3980. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Hengyi Cai, Hongshen Chen, Cheng Zhang, Yonghao Song, Xiaofang Zhao, Yangxi Li, Dongsheng Duan, and Dawei Yin. 2020. Learning from easy to complex: Adaptive multi-curricula learning for neural dialogue generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7472–7479. AAAI Press.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Thang Doan, João Monteiro, Isabela Albuquerque, Bogdan Mazoure, Audrey Durand, Joelle Pineau, and R. Devon Hjelm. 2019. On-line adaptative curriculum learning for gans. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3470–3477. AAAI Press.

Manfred Eppe, Sven Magg, and Stefan Wermter. 2019. Curriculum goal masking for continuous deep reinforcement learning. In *Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2019, Oslo, Norway, August 19-22, 2019*, pages 183–188. IEEE.

Jiazhan Feng, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3805–3815. Association for Computational Linguistics.

Joseph L. Fleiss and Jacob Cohen. 2016. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational Psychological Measurement*, 33(3):613–619.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2019. Dialogwae: Multimodal response generation with conditional wasserstein autoencoder. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tom Kocmi and Ondrej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 379–386. INCOMA Ltd.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1189–1197. Curran Associates, Inc.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Pierre Lison and Serge Bibauw. 2017. Not all dialogues are created equal: Instance weighting for neural conversational models. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 384–394. Association for Computational Linguistics.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1782–1792. Association for Computational Linguistics.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2020. Teacher-student curriculum learning. *IEEE Trans. Neural Networks Learn. Syst.*, 31(9):3732–3740.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3349–3358. ACL.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1162–1172. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 583–593. ACL.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sharath T. S., Shubhangi Tandon, and Ryan Bauer. 2017. A dual encoder sequence to sequence model for open-domain dialogue modeling. *CoRR*, abs/1710.10520.

Mrinmaya Sachan and Eric P. Xing. 2016. Easy questions first? A case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Haoyu Song, Weinan 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 2019, Macao, China, August 10-16, 2019*, pages 5190–5196. ijcai.org.

Amir Vakili Tahami, Kamyar Ghajar, and Azadeh Shakery. 2020. Distilling knowledge for fast retrieval-based chat-bots. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2081–2084. ACM.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1074–1080. Association for Computational Linguistics.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1282–1292. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6095–6104. Association for Computational Linguistics.

Xinnuo Xu, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3981–3991. Association for Computational Linguistics.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural response generation via GAN with an approximate embedding layer. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 617–626. Association for Computational Linguistics.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2852–2858. AAAI Press.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiaoxi Mao, Yadong Xi, and Minlie Huang. 2020. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3449–3460. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.