Adaptively Multi-Objective Adversarial Training for Dialogue Generation

Xuemiao Zhang[1*], **Zhouxing Tan**[1*], **Xiaoning Zhang**[2], **Yang Cao**[4] and **Rui Yan**[3†]
[1]School of Software & Microelectronics, Peking University
[2]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences
[3]Wangxuan Institute of Computer Technology, Peking University
[4]SenseTime Research
{zhangxuemiao, tzhx, ruiyan}@pku.edu.cn, xiaoningzhang42@gmail.com, caoyang@sensetime.com

Abstract

Naive neural dialogue generation models tend to produce repetitive and dull utterances. The promising adversarial models train the generator against a well-designed discriminator to push it to improve towards the expected direction. However, assessing dialogues requires consideration of many aspects of linguistics, which are difficult to be fully covered by a single discriminator. To address it, we reframe the dialogue generation task as a multi-objective optimization problem and propose a novel adversarial dialogue generation framework with multiple discriminators that excel in different objectives for multiple linguistic aspects, called AMPGAN, whose feasibility is proved by theoretical derivations. Moreover, we design an adaptively adjusted sampling distribution to balance the discriminators and promote the overall improvement of the generator by continuing to focus on these objectives that the generator is not performing well relatively. Experimental results on two real-world datasets show a significant improvement over the baselines.

1 Introduction

The end-to-end neural systems [Serban *et al.*, 2016; Luan *et al.*, 2016] based on SEQ2SEQ framework [Sutskever *et al.*, 2014] and using the maximum likelihood estimation (MLE) objective have been extensively studied because of its ability to generate unseen conversations and high scalability. But the dialogue responses generated by these models tend to be generic, dull, repetitive [Tao *et al.*, 2018; Li *et al.*, 2016; Zhang *et al.*, 2018] in practice. Recently, sequence generative adversarial nets (SeqGAN) [Yu *et al.*, 2017] apply the promising adversarial training method of GANs [Goodfellow *et al.*, 2014] on sequence generation task by employing policy gradient and Monte Carlo (MC) search. AdverREGS [Li *et al.*, 2017] employs adversarial training to actual dialogue generation task and trains the generator against a real-fake discriminator. DP-GAN [Xu *et al.*, 2018] and DAL [Cui *et al.*, 2019] try to increase the diversity of the generated responses by designing a fine-grained diversity discriminator and utilizing the duality between query and response generation, respectively.

But in fact, assessing dialogue responses should consider many aspects of linguistics such as diversity, fluency, syntax habits of human, and so on. However, although a well-designed discriminator can make the generator perform well in the corresponding expected aspect, other aspects are often ignored. We notice that the multi-adversarial framework, denoted MDGAN, that extends GANs to multiple discriminators has achieved promising results on image generation tasks. Many models in MDGAN [Durugkar *et al.*, 2017; Neyshabur *et al.*, 2017; Albuquerque *et al.*, 2019] use their specially designed methods to weight the sum of the losses of image discriminators to optimize the image generator overall.

We hope the generator can cover multiple linguistic aspects to achieve overall improvement and perform relatively well in all aspects. Intuitively, we can view dialogue generation as a multi-objective optimization problem [Deb, 2001; Albuquerque *et al.*, 2019]. Each objective is designed for a linguistic aspect, and a corresponding discriminator provide guidances towards this objective, thereby optimizing the generator towards multiple objectives. However, this idea faces challenges. The main challenge is to innovatively propose a theoretically proven and practically effective framework that extends the original adversarial dialogue generator optimization problem to a multi-objective optimization problem, where each objective is defined by policy gradient. The second is to explore how to design and combine various linguistic objectives, and to deal with the problem of asynchronous optimization of different objectives during training.

In this paper, we propose the **a**daptively **m**ulti-objective adversarial dialogue generation framework using **p**olicy gradient, called AMPGAN, to push the generator G_θ to achieve overall improvement. AMPGAN models dialogue generation as a stochastic policy in reinforcement learning (RL) and directly performs gradient policy update to generate discrete dialogue utterances, and is formalized as the multi-objective optimization problem by framing the simultaneous maximization of rewards received from multiple discriminators. We design multiple objectives of linguistics for guiding G_θ of AMPGAN: indistinguishable from the human utterances, high syntactic score and diversity. We propose mul-

[*]Equal contribution.
[†]Corresponding author: Rui Yan (ruiyan@pku.edu.cn)

tiple corresponding discriminators and an adaptively adjusted sampling distribution to dynamically organize them to participate in adversarial training. Through the adaptive distribution, AMPGAN can focus on the aspects that G_θ not perform well relatively. When G_θ is still weak and have much room to improve in aspect a_i, AMPGAN will adaptively increase the probability that the corresponding D_i is selected to participate in adversarial training, to motivate G_θ to optimize towards a_i; Conversely, when G_θ can easily deceive the discriminator D_j of aspect a_j, AMPGAN will adaptively reduce its probability, and pay more attention to other relatively weak aspects. Thereby, G_θ can cover multiple aspects and solve the asynchronous optimization problem.

Our contributions are summarized as: (1) We reframe the dialogue generation as a multi-objective optimization problem and propose the novel multiple adversarial generation framework (AMPGAN) to improve the generator towards multiple objectives of linguistics and provide the theoretical derivation and proof; (2) We explore how to choose and combine different discriminators, and design an adaptively adjusting sampling distribution to balance all discriminators; (3) We conduct ample experiments on two datasets, and experimental results show the effectiveness of AMPGAN framework.

2 Methodology

We propose the adaptively multi-adversarial dialogue generation framework using policy gradient, called AMPGAN, to push the dialogue generator to improve towards multiple objectives based on linguistic apects, as shown in Figure 1.

2.1 The AMPGAN Framework

Given a dialogue query utterance $x = \{x_i\}_{i=1}^m$ of m words, the generator G_θ needs to produce a response $y = \{y_j\}_{j=1}^n$ of n words, where $x_i, y_j \in \mathcal{T}$ (the word vocabulary). AMPGAN conducts gradient policy update directly by modeling G_θ defined by a SEQ2SEQ model as a stochastic policy in reinforcement learning (RL). Each discriminator in AMPGAN is designed to judge on the complete generated sequence according to a linguistic objective, and reports the RL reward which is passed back to the intermediate state-action steps using MC search. Note that all discriminators $\{D_i\}_{i=1}^N$ do not share parameters. We extend and divide the types of discriminators into three categories, binary classification task D_C, regression task D_R and direct rule scorer D_I. Among them, D_C and D_R are parameterized models that need training, and D_I is a set of scoring logical rules which can be understood as a fully trained and completely correct discriminator.

More formally, G_θ and $\{D_i\}_{i=1}^N$ play the following minimax optimization game:

$$\min_{G_\theta} \max_{\{D_i\}_{i=1}^N} V(G_\theta, \{D_i\}_{i=1}^N) = \mathbb{E}_{D\sim\pi(Q)}(\mathbb{E}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})} \log(D(\boldsymbol{x})) + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})} \log(1 - D(G_\theta(\boldsymbol{z}))))) \tag{1}$$

where the random variable discriminator D obeys the distribution $\pi(Q)$ on the increment Q of RL rewards. $\boldsymbol{x}$ is the ground truth data and $\boldsymbol{z}$ is the query data drawn from a prior $p_z(\boldsymbol{z})$. G_θ maps $\boldsymbol{z}$ to the fake responses which obeys the distribution p_g. Next we will provide formal theoretical analysis.

Theoretical Analysis

Firstly, we consider the optimization of multiple discirminators for any given G_θ and $\pi(Q)$ under the proposed AMPGAN framework. Note that since the rule scorer D_I is constant, here we only need to consider the variable D_C and D_R.

Proposition 1. For G_θ and $\pi(Q)$ fixed, the optimal discriminator D_i is: ${D_i}^* = p_d(\boldsymbol{x})/(p_d(\boldsymbol{x}) + p_g(\boldsymbol{x}))$.

Proof. According to the induced measure theorem [Somesh Das Gupta, 2000], the two expectations are equal: $\mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}}[f(G(\boldsymbol{z}))] = \mathbb{E}_{\boldsymbol{x}\sim P_G}[f(\boldsymbol{x})]$. The objective function can be rewritten as below:

$$V(G, \{D_i\}_{i=1}^N) = \int_{\boldsymbol{x}} \sum_{i=1}^N p_{D_i}(p_d(\boldsymbol{x}) \log D_i(\boldsymbol{x}) + p_g(\boldsymbol{x}) \log(1 - D_i(\boldsymbol{x})))d\boldsymbol{x} \tag{2}$$

Let $A := \sum_{i=1}^N p_{D_i}(p_d(\boldsymbol{x}) \log D_i(\boldsymbol{x}) + p_g(\boldsymbol{x}) \log(1 - D_i(\boldsymbol{x})))$ then $diag(\{\frac{p_{D_i}(p_d(\boldsymbol{x})(D_i-1)^2 + p_g(\boldsymbol{x}){D_i}^2)}{{D_i}^2(D_i-1)^2}\}_{i=1}^N)$ is the Hessian matrix of $-A$, and all leading principle minors of the matrix are greater than 0, thereby $-A$ is a convex function. So the maximum point D_i^* of A is stationary points on $[0, 1]$:

$$\frac{\partial A}{\partial D_i} = p_{D_i}(\frac{p_d(x)}{D_i} + \frac{p_g(x)}{D_i - 1}) = 0$$
$${D_i}^* = p_d(\boldsymbol{x})/(p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})) \in [0, 1] \tag{3}$$

And D_i^* is also the optimal point of $V(G, \{D_i\}_{i=1}^N)$ Q.E.D.

Secondly, we fix $D_i = D_i^*$ and find the optimal solution G^* for the generator of the proposed AMPGAN framework.

Theorem 1. *Let* $C(G) = \max_{\{D_i\}_{i=1}^N} V(G, \{D_i\}_{i=1}^N)$, *then* $\min C(G) = -\log 4$.

Proof. Bring D_i^* into the above formula:

$$\begin{aligned} C(G) &= \sum_{i=1}^N p_{D_i}[\mathbb{E}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})} \log(\frac{p_d(\boldsymbol{x})}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}) \\ &+ \mathbb{E}_{\boldsymbol{x}\sim p_g(\boldsymbol{x})} \log(\frac{p_g(\boldsymbol{x})}{p_g(\boldsymbol{x}) + p_d(\boldsymbol{x})})] \\ &= (\sum_{i=1}^N p_{D_i})(-\log 4 + (KL(p_d(x)||\frac{p_d(x) + p_g(x)}{2}) \\ &+ KL(p_g(x)||\frac{p_d(x) + p_g(x)}{2})) \\ &= -\log(4) + JSD(p_d||p_g) \end{aligned} \tag{4}$$

when p_d and p_g are consistent, get the minimum value $-\log(4)$ Q.E.D.

Thirdly, we directly use a policy gradient of RL to optimize G_θ, which can avoid the problem of differentiation difficulty [Yu *et al.*, 2017] for discrete dialogue data in the traditional GAN naturally. Each D_i in AMPGAN guides to improve G_θ by computing a specific reward signal on sequences generated by G_θ using MC rollout strategy. Specifically, starting from the current state $s = Y_{1:t-1}$, action $a = y_t$, using MC search until the complete sequence is generated, the search sequence set $MC^{G_\theta}(Y_{1:t}; N)$ is obtained. And the value function $Q_{D_i}^{G_\theta}$ of single discirminator D_i on this set is set to the average discrimination score as:

$$Q_{D_i}^{G_\theta}(s = Y_{1:t-1}, a = y_t) = \frac{1}{M} \sum_{m=1}^M D_i(Y_{1:T}^m) \tag{5}$$

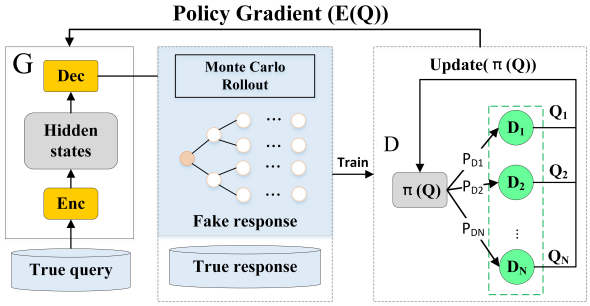

Figure 1: The architecture of AMPGAN. Given queries, the generator generates fake responses. A discriminator will be sampled to compute the reward to optimize the generator according to the adaptively adjusted distribution. All discriminators are supervisedly trained by using true and fake responses.

where $Y_{1:T} \in MC^{G_\theta}(Y_{1:t}; N)$, M is the simulations number of MC. When there is no intermediate reward, we can use the average value $J_{D_i}(\theta)$ assigned by the expectation of value function $Q_{D_i}^{G_\theta}$ to evaluate the policy G_θ [Sutton *et al.*, 2000].

$$J_{D_i}(\theta) = \sum_{y_1 \in Y} G_\theta(y_1|s_0) Q_{D_i}^{G_\theta}(s_0, y_1) \tag{6}$$

In AMPGAN framework, dialog generation is viewed as a multi-objective optimization problem. Formally, the generator solves the following multi-objective problem:

$$\max \mathcal{J}(\theta) = [J_{D_1}(\theta), J_{D_2}(\theta) \dots J_{D_N}(\theta)]^T \tag{7}$$

Specifically, we design an adaptively adjusted distribution $D \sim \pi(Q)$ to organize the multiple objectives as:

$$\max \mathcal{J}(\theta) = \mathbb{E}_{D \sim \pi(Q)} J_D(\theta) \tag{8}$$

So that AMPGAN can continue to pay attention to the objectives that G_θ not performs well relatively and G_θ can be balanced and improved towards many objectives more stably.

Formally, when G_θ is weak in an objective, AMPGAN dynamically strengthens the adversarial training between G_θ and the corresponding D_i by adaptively adjusting the sampling distribution $\pi(Q) = \{p^{(k)}(D = D_i)\}_{i=1}^N$ as:

$$p(D_i)^{(k)} = \frac{1}{Z} p(D_i)^{(k-1)} \exp(\Delta Q_{D_i}^{G_\theta}{}^{(k)}) \tag{9}$$

where $Z = \sum_{i=1}^N p_i^{(k-1)} \exp(\Delta Q_{D_i}^{G_\theta}{}^{(k)})$ and $\Delta Q_{D_i}^{G_\theta}{}^{(k)} = Q_{D_i}^{G_\theta}{}^{(k)} - Q_{D_i}^{G_\theta}{}^{(k-1)} + \mathcal{N}(0, \sigma)$ and $\mathcal{N}(0, \sigma)$ is Gaussian noise, and initial $\pi(Q)$ is $p^{(0)}(D = D_i) = \frac{1}{N}$. When $\Delta Q_{D_i}^{G_\theta}$ is larger, that is, there is much room for G_θ to improve in the corresponding objective, AMPGAN will select D_i with a larger probability.

Then the gradient of (8) can be computed as:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{D \sim \pi(Q)} \sum_{t=1}^T \mathbb{E}_{Y_{1:t-1} \sim G_\theta} \Big[\sum_{y \in Y} \nabla_\theta G_\theta(y_t|Y_{1:t-1}) Q_D^{G_\theta}(y_t|Y_{1:t-1}) \Big] \tag{10}$$

We can use $\theta \leftarrow \theta + \alpha_k \nabla_\theta \mathcal{J}(\theta)$ to update the parameters θ of the generator, where α_k denotes the learning rate. And in practice, we estimate the expectation in the Eq. (10) by sampling discriminators according to $\pi(Q)$.

Algorithm 1 Training AMPGAN.

Require: generator policy G_θ; N discriminators $\{D_i\}_{i=1}^N$; sampling distribution $\pi(Q)$; dataset $S = \{X_{1:T}\}$

```
Randomly initialize G_θ, {D_i}_{i=1}^N, {π(D_i) = 1/N}_{i=1}^N.
Pre-train G_θ using MLE on S
Generate negative samples using G_θ
Pre-train {D_i}_{i=1}^N separately
while MPGAN do not converges do
    Sample a discirminator D_i using π(Q)
    for g-steps do
        Generate a sequence Y_{1:T} = (y_1, ..., y_T) ~ G_θ
        for t in 1 : T do
            Compute Q(a = y_t; s = Y_{1:t-1}) by Eq. (5)
        end for
        Update generator via policy gradient Eq. (10)
    end for
    for d-steps do
        Combine negative examples generated by current G_θ and positive examples sampled from S to train D_i
    end for
    Update π(Q) by Eq. (9)
end while
```

In summary, Algorithm 1 shows the details of AMPGAN. Before adversarial training, we first pretrain the generator G_θ using MLE on training set $\mathcal{S}$, and then use G_θ to generate negative examples. We sample positive examples from $\mathcal{S}$ and combine them with negative ones to pretrain parameterized discriminators $\{D_i\}_{i=1}^N$. During adversarial training, G_θ and each discriminator are trained alternately. At each training step, a discriminator D_i will be selected from $\{D_i\}_{i=1}^N$ to participate in adversarial training according to distribution $\pi(Q)$. We sample positive examples from $\mathcal{S}$ and use G_θ to synthesize corresponding negative ones and use both to train D_i. G_θ is updated by using the policy gradient and MC search based on the expected final value function received from D_i. Finally, we adaptively adjust the sampling distribution $\pi(Q)$ based on the increments of the generator's scores in various aspects. Following [Li *et al.*, 2017], we employ the *Teacher Forcing* strategy to tell G_θ what sequences are good and guide G_θ push itself to produce these good sequences by exposing the good target sentence to it directly.

2.2 AMPGAN for Dialogue Generation

We train the dialogue generator G_θ against 3 discriminators.

Generator

We use a standard SEQ2SEQ [Sutskever *et al.*, 2014] model implemented by two-layer LSTMs with attention mechanism [Luong *et al.*, 2015] to implement the generator's policy G_θ [Li *et al.*, 2017]. G_θ encodes the input queries into hidden states, and then decodes them into output responses.

Real-Fake Discriminator

We continue the previous works [Li *et al.*, 2017] and use the generic Real-Fake discriminator D_{rf} to judge whether an input dialogue utterances $\{x, y\}$ are generated by humans (denoted as a real label) or machines (fake). D_{rf} is essentially a two-classifier, and uses a hierarchical encoder [Serban *et al.*,

2016] with a two-class softmax layer to report the probability Q_+ that $\{x, y\}$ are real and Q_- that $\{x, y\}$ are fake. The purpose of using D_{rf} is to encourage G_θ to generate dialogue utterances that are difficult to distinguish between real or fake by taking Q_+ as the reward.

Syntactic Discriminator

We notice that the generated sequences are often confusing because the arrangement of words is not in line with human syntax norms. We try to improve the syntax score of the generated sequences by feeding G_θ the syntactic reward reported by the proposed syntactic discriminator D_{syn}. Following [Vashishth *et al.*, 2019], We use graph convolutional networks (GCN) to extract syntactic features embedding $\boldsymbol{e}_{syn}$ of a sequence. We use the Stanford CoreNLP parser [Manning *et al.*, 2014] to parse each word sequence $s = \{x_1, x_2, ..., x_{|s|}\}$ into a dependency parse graph $\mathcal{G} = (\mathcal{V}_s, \mathcal{E}_s)$, whereas $\mathcal{V}_s = \{x_1, x_2, ..., x_{|s|}\}$ and $\mathcal{E}_s$ stands for the labeled directed dependency edges (x_i, x_j, l_{ij}), where l_{ij} is the dependency relaiton label of x_i to x_j. We use the corresponding word embedding $\boldsymbol{v}_i \in \mathbb{R}^d$ to initialize each node embedding $\boldsymbol{h}_i^0 \in \mathbb{R}^d$ in $\mathcal{V}_s$.

In graph convolutional layers of D_{syn}, we employ the SynGCN [Vashishth *et al.*, 2019] model with Edge Label Gating Mechanism to aggregate each node. Then embedding $\boldsymbol{h}_i^k \in \mathbb{R}^d$ of node i after k GCN layers is computed as $\boldsymbol{h}_i^k = SynGCN(\boldsymbol{h}_i^0)$. We use max-pooling operation to obtain syntactic embedding $\boldsymbol{e}_{syn} = \{e_m\}_{m=1}^d$, where $e_m = \max\{\boldsymbol{h}_{im}^k\}_{i=1}^{|s|}$. Finally, syntactic score s_{syn} is computed as: $s_{syn} = \sigma(BN(\boldsymbol{W}_s \boldsymbol{e}_{syn} + b_s))$, where $\boldsymbol{W}_s \in \mathbb{R}^{1\times d}$ is a parameter vector, b_s is a bias, BN is a batch normalization layer [Ioffe and Szegedy, 2015], and $\sigma(\cdot)$ is the sigmod function. We use s_{syn} as the reward to guide G_θ to produce utterances that conform to human expressions.

Information Discriminator

The generator tends to produce high frequency words. We design the information discirminator D_{info} to directly compute the normalized information amount s_{info} of the generated utterance with n tokens. Formally, s_{info} is computed as $s_{info} = \frac{1}{n}\sum_{i=1}^{n} \frac{I(w_i)-I(w_{min})}{I(w_{max})-I(w_{min})}$, whereas $I(w_i) = -\log p(w_i)$, $p(w_i)$ is the frequency of word w_i, w_{max} and w_{min} are the two words with the largest and smallest word frequencies in the corpus, respectively. We use s_{info} as the reward to push G_θ to explore more novel words.

3 Experiments

3.1 Datasets

Cornell Movie Dataset[*] (denoted as $\mathcal{S}_1$) contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts. It consists of 220579 exchanges between 10292 pairs of characters in the movie, involving 9035 characters in 617 movies, a total of 304713 utterances. **OpenSubtitles Dataset** ($\mathcal{S}_2$) is a well-known human-human scripted dialogue dataset. It is extracted from movie subtitles which are not speaker-aligned [Tiedemann, 2009]. We use the English short message data, which contains about 140M utterances in total, covers 106K movie works.

[*]Cornell Movie–Dialogs Corpus

3.2 Training Details

We sort the words in each corpus by word frequency and use the first 35K words with high word frequency as the vocabulary. We filter out dialogue utterances with less than 5 or more than 40 tokens. We set the training batch size to 128, the size of word embeddings and the graph node embeddings to 512, the hidden size of hidden layers of all encoders in all models to 256, and the number of all LSTM layers to 2, and GCN layer number of D_{syn} to 1. We train all models using Adam optimizer [Kingma and Ba, 2015], and all the dropout rates to 0.7. For the generator G_θ, we use the learning rate decay strategy and set the initial learning rate to 0.1 and the decay factor to 0.99. We set the learning rate to 0.001 fixedly for D_{rf} and D_{syn}. In pre-training, we first pre-train G_θ 5000 iterations, then use G_θ to produce 2500×128 negative examples and sample the same amount of positive ones from the dataset correspondingly. We combine both to pre-train D_{rf} and D_{syn}. In MC search, we employ the experience [Li *et al.*, 2017] that given a partially decoded s_P, G_θ will keep sampling tokens in word distribution until decoding is complete. Repeat this process k times (k is set to 7) and obtain k sequences sharing a common prefix s_P. The average of the corresponding k scores given by the discriminator is used as the reward. During training of each model, if the performance on the validation set has not improved for a long time, stop training and choose the checkpoint with the best performance.

3.3 Experiments Results

We use the following models as the baselines. (1) **MLE** trains a standard SEQ2SEQ generator G_θ using the traditional MLE; (2) **PG-BLEU** [Yu *et al.*, 2017] trains G_θ by policy gradient using the BLEU score of generated sequences as the reward directly, thereby obtaining a higher BLEU score; (3) **AdverREGS** [Li *et al.*, 2017] is a standard adversarial model that G_θ is trained against a real-fake discirminator alone, and can be viewed as an extension of SeqGAN on the SEQ2SEQ framework; (4) **DP-GAN** [Xu *et al.*, 2018] designs a fine-grained diversity discriminator to guide G_θ to produce sequences with high diversity. (5) **DAL** [Cui *et al.*, 2019] tries to increase the response diversity by utilizing the duality between query and response generation.

We regularly combine different discriminators to guide the generator: (1) D_1: only use single D_{rf}, noted as AdverREGS; (2) D_2: only use single D_{syn}; (3) D_3: only use single D_{Info}; (4) $D_1\&D_2$: combine D_{rf} and D_{syn}; (5) $D_1\&D_3$: combine D_{rf} and D_{Info}; (6) $D_2\&D_3$: combine D_{syn} and D_{Info}; (7) $D_1\&D_2\&D_3$: combine D_{rf}, D_{syn} and D_{Info}.

We use two categories of metrics to evaluate each model, namely **Adversarial Evaluation** [Li *et al.*, 2017; Bowman *et al.*, 2016] and widely used **Common Evaluation** metrics.

Adversarial Evaluation

Adversarial evaluation helps to analyze the impact of different discriminator combinations of AMPGAN. Following [Li *et al.*, 2017], we train automated machine evaluators to distinguish machine-generated responses from human-generated

responses, such as accurately distinguishing real or fake, giving high scores to real responses and low scores to the fake. We design three machine evaluators, namely *RealE*, *SyntaxE* and *InfoE*, corresponding to D_{rf}, D_{syn} and D_{info}, to report the ratio *FoolRate* of the cases that *RealE* are deceived, the syntax score *SynScore*, and the ratio *InfoRate* of the information amount of the generated responses to the ground truth, respectively. Note that each evaluator is in the same architecture as its corresponding discriminator, but trained separately using supervised methods and do not participate in adversarial training. The abilities of parameterized *RealE* and *SyntaxE* are 0.09 and 0.11, indicates how much we can trust the evaluation results given by the evaluators, and the smaller the value, the better. We also report the comprehensive score *NormAve* of *SynScore*, *InfoRate* and *FoolRate* as $\frac{1}{3}\sum_{i=1}^{3}(x_j^i - x_{min}^i)/(x_{max}^i - x_{min}^i)$, where x_j^i, x_{max}^i, and x_{min}^i are the score of model j, maximum and minimum in the table on the i-th indicator respectively.

Table 1 shows the evaluation results. From Table 1 we can find out that AMPGAN can effectively improve the generator G_θ. Specifically, it can be summarized as follows: (1) Using a single discriminator is risky. We can find each model can fool the real-fake evaluator *RealE* to a large extent (greater than 97%) on $\mathcal{S}_1$, which shows that generators can easily perform well in this objective on $\mathcal{S}_1$, even without the guidance reward of D_{rf}. (2) The generators against a single discriminator can go deeper towards expected objectives. We can find the generator against single D_1 or D_2 or D_3 performs better in its corresponding objective than other single adversarial training models. (3) The generator trained against multiple discriminators can find a balance and achieve an overall improvement towards multiple objectives, although it may be weakened unilaterally. We can find the generator against $D_1\&D_2\&D_3$ achieves significantly higher overall performance, although unilateral indicator values decrease a bit. Empirically then, we recommend that designers should design a general objective with the corresponding discirminator to report the basic but universal reward and supplement it with other specific rewards according to actual needs, so that the model can perform well in general and also stand out in several areas.

	Model	SynScore	InfoRate	FoolRate	NormAve
$\mathcal{S}_1$	MLE	0.624	0.817	0.984	0.367
	PG-BLEU	0.540	0.889	0.987	0.498
	AdverREGS(D_1)	0.584	0.864	0.992	0.573
	DP-GAN	0.481	0.921	0.993	0.617
	DAL	0.521	0.917	0.992	0.630
	SynGCN(D_2)	**0.715**	0.832	0.979	0.403
	InfoAmt(D_3)	0.377	**0.950**	0.987	0.474
	$D_1\&D_2$	0.672	0.801	**0.998**	0.624
	$D_1\&D_3$	0.541	0.896	0.993	0.620
	$D_2\&D_3$	0.655	0.880	0.991	0.661
	$D_1\&D_2\&D_3$	0.653	0.871	0.995	**0.710**
$\mathcal{S}_2$	MLE	0.348	0.859	0.221	0.193
	PG-BLEU	0.356	0.875	0.270	0.250
	AdverREGS(D_1)	0.392	0.861	**0.655**	0.600
	DP-GAN	0.428	0.932	0.351	0.536
	DAL	0.401	0.919	0.435	0.595
	SynGCN(D_2)	**0.500**	0.832	0.360	0.466
	InfoAmt(D_3)	0.285	**0.951**	0.373	0.450
	$D_1\&D_2$	0.495	0.822	0.529	0.562
	$D_1\&D_3$	0.355	0.864	0.533	0.457
	$D_2\&D_3$	0.458	0.931	0.401	0.688
	$D_1\&D_2\&D_3$	0.472	0.926	0.477	**0.755**

Table 1: Results of Adversarial Evaluation on both datasets.

Common Evaluation

The metrics of common evaluation are as follows:

(1) **BLEU** score [Papineni *et al.*, 2002] is widely used in sequence generation tasks [Li *et al.*, 2016], measuring the overlaping between the generated word sequences and the ground truth; (2) **Distinctness** metric proposed by [Li *et al.*, 2016] measures the diversity of generated token sequences. (3) **Human Evaluation** (HM): We ask three human annotators with a linguistic background to report on the overall quality of the responses to 200 examples randomly sampled from the test set, which are generated by the full AMPGAN model and all baseline models. (4) **Similarity&Relevance**: We also report the similarity between the ground truth responses and the generated, and the relevance between the queries and the generated responses at semantic level by calculating cosine similarity of the two token sequence embeddings computed by the pretrained BERT [Devlin *et al.*, 2018] model.

Table 2 shows the evaluation results. From Table 2, we can find out that AMPGAN framework do help achieve better results on these common metrics. Specifically, the generator against three discriminators performs very well. It achieves the highest BLEU-1 score and Dist-1, Dist-2, and Dist-3 scores, as well as relatively higher BLEU-2 and Dist-4 scores on both datasets, indicating that the generated token sequences have higher overlap with ground truth responses and higher diversity. It also achieve the highest ranking of Human Evaluation over baselines on both datasets. Moreover, the pairwise adversarial training generators also achieved higher word overlap and diversity than the corresponding single adversarial training ones separately.

Figure 2 shows the results of similarity&relevance evaluation. We can clearly find the combination of $D_1\&D_2\&D_3$ makes the generated sentences obtain the highest similarity with ground truth responses and the highest relevance with queries on both datasets, indicating the token sequences generated by this generator are closest to the ground truth at the semantic level. Moreover, we can also find out the trend that the generator against multiple combined discriminators can achieve higher similarity and relevance than the one against each of these discriminators alone.

Effectiveness of Adaptive Sampling Distribution

We also investigate the adjustment of the discriminator sample distribution during training, as shown in Figure 3, and conduct a comparative experiment between using adaptive distribution and fixed uniform distribution, as shown in Table 3. From Figure 3, we can find out that different datasets have different characteristics, and AMPGAN can capture these characteristics by adaptively adjusting the sampling distribution of all objectives. And during training, improvement degrees of G_θ towards different objectives are constantly changing, so it is very useful to dynamically adjust the distribution. In addition, from Table 3, we can find that compared with using a fixed uniform distribution, using the adaptive adjustment distribution helps AMPGAN achieve higher scores of adversarial evaluation and diversity on two datasets, which also shows the effectiveness of the adaptive distribution. Tak-

Model	Cornell							Opensubtitles						
	BLEU-1	BLEU-2	Dist-1	Dist-2	Dist-3	Dist-4	HM	BLEU-1	BLEU-2	Dist-1	Dist-2	Dist-3	Dist-4	HM
MLE	12.8	1.54	34.8	62.4	78.6	92.4	5.33	14.4	2.42	28.4	53.1	73.8	91.0	5.67
PG-BLEU	13.2	1.79	37.0	65.2	83.2	97.5	5.00	**15.2**	1.97	30.3	56.7	78.4	95.8	4.33
AdverREGS(D_1)	12.1	1.91	38.3	65.0	82.8	95.6	4.33	14.5	2.30	30.3	56.5	78.1	95.7	4.00
DP-GAN	12.0	1.60	37.8	66.9	84.2	98.3	2.00	14.7	1.92	30.5	56.9	79.1	97.6	2.33
DAL	12.9	1.95	38.7	66.2	83.5	97.9	3.00	14.4	2.35	31.1	57.1	79.5	97.3	2.67
SynGCN(D_2)	12.3	2.17	32.9	65.6	83.2	95.5	-	13.8	2.36	30.1	56.6	79.0	95.2	-
InfoAmt(D_3)	11.8	2.02	36.9	64.7	83.1	95.5	-	14.3	2.42	30.8	57.3	79.4	96.4	-
D_1&D_2	12.6	**2.36**	35.8	66.5	83.5	97.8	-	14.7	2.40	31.7	58.7	80.3	97.2	-
D_1&D_3	12.9	2.14	38.0	66.1	83.7	97.5	-	14.9	2.45	32.4	58.2	80.0	97.4	-
D_2&D_3	13.5	2.22	38.8	67.1	84.7	**98.5**	-	14.8	**2.48**	32.7	59.2	80.8	**97.8**	-
D_1&D_2&D_3	**13.9**	2.21	**42.9**	**70.4**	**87.9**	98.4	**1.33**	**15.2**	2.43	**33.2**	**60.3**	**81.1**	97.6	**2.00**

Table 2: Results of common metrics evaluation on Cornell dataset and Opensubtitles dataset.

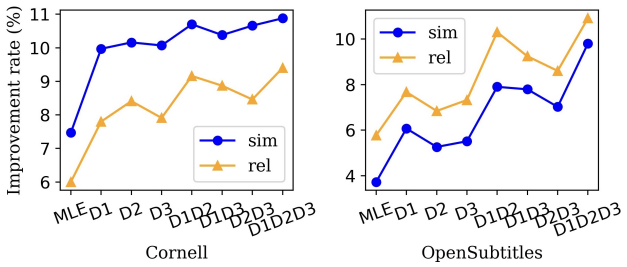

Figure 2: Results of similarity (sim) and relevance (rel) evaluation on the generated sentences by the generators against different discriminator combinations.

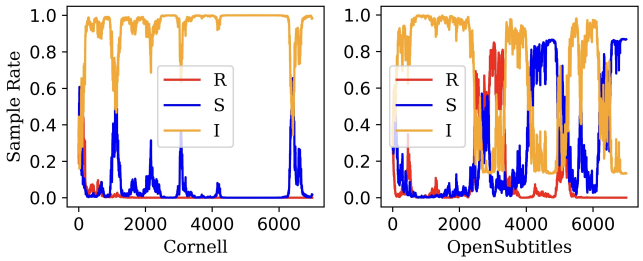

Figure 3: Adaptive adjustment of sampling distribution of Real-Fake(R), SynGCN(S) and InfoAmt(I) as the training progresses.

Sample	Cornell			Opensubtitles		
	NormAve	BAve	DAve	NormAve	BAve	DAve
Uniform	0.679	7.89	73.8	0.698	**8.84**	67.5
Adaptive	**0.710**	**8.06**	**74.9**	**0.755**	8.82	**68.1**

Table 3: Comparison of two sampling distributions. *BAve* is the average score of BLEU-1 and BLEU-2, and *DAve* is the average score of Dist-1, Dist-2, Dist-3 and Dist-4.

ing the *OpenSubtitles* curve as an example, when G_θ has been promoted sufficiently towards one objective and its improvement room becomes relatively small, then AMPGAN will reduce its probability and focus on other objectives, so as to achieve the purpose of balancing all objectives and improving overall performance. And because D_{info} is absolutely accurate and can basically not be deceived by G_θ, D_{info} can always provide reward to guide G_θ to improve, which is reflected in its high probabilities. However, the D_{rf} has the risk of saturation [Xu *et al.*, 2018]. From *Cornell* curve, we can find that there is no improvement room towards the objective of D_{rf} soon, indicating that G_θ can quickly fool D_{rf}.

4 Related Works

SeqGAN [Yu *et al.*, 2017] successfully employs the promising GAN framework into the discrete sequence generation task by utilizing policy gradient and MC search. AdverREGS [Li *et al.*, 2017] employs adversarial training to guide the generator to produce dialogue utterances that difficult for evaluators to distinguish between human-generated and machine-generated. Many models focus on increasing the diversity of the generated responses: DP-GAN [Xu *et al.*, 2018] designs a fine-grained discriminator which has a better diversity evaluation method; The Adversarial Information Maximization (AIM) [Zhang *et al.*, 2018] optimizes a variational lower bound on pairwise mutual information between query and response; And DAL [Cui *et al.*, 2019] utilizes the duality between query generation and response generation.

Recent literatures have demonstrated promising results in image generation tasks for extending the basic GAN framework using multiple discriminators to train against the generator. D2GAN [Nguyen *et al.*, 2017] combines KL and inverse KL divergence into a unified objective function and uses the complementary statistical characteristics of these divergences to effectively disperse the estimated density to capture multiple modes. The Generative Multi-Adversarial Nets (GMAN) [Durugkar *et al.*, 2017] trains the generator against a softmax weighted arithmetic average of K different discriminators in the same model architecture. HVM [Albuquerque *et al.*, 2019] proposes hypervolume maximization to efficiently optimize a weighted loss from all discriminators for the case of large neural networks.

5 Conclusions

In this paper, we propose the adaptively multi-objective adversarial dialogue generation framework using policy gradient, called AMPGAN, to push the generator to improve itself towards multiple objectives of linguistics by feeding these reward signals from the discriminators corresponding to these objectives. Comparative experimental results of different combination models and baseline models show that AMPGAN can significantly improve overall performance with only minor sacrifices in each objective.

Acknowledgments

This work was supported by the National Science Foundation of China (NSFC No. 61876196).

References

[Albuquerque *et al.*, 2019] Isabela Albuquerque, João Monteiro, Thang Doan, Breandan Considine, Tiago H. Falk, and Ioannis Mitliagkas. Multi-objective training of generative adversarial networks with multiple discriminators. In *ICML*, 2019.

[Bowman *et al.*, 2016] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *SIGNLL*, 2016.

[Cui *et al.*, 2019] Shaobo Cui, Rongzhong Lian, Di Jiang, Yuanfeng Song, Siqi Bao, and Yong Jiang. DAL: Dual adversarial learning for dialogue generation. In *NAACL Workshop on NeuralGen*, 2019.

[Deb, 2001] Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*, volume 16. John Wiley & Sons, 2001.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Durugkar *et al.*, 2017] Ishan P. Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. In *ICLR*, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, and Jianfeng Gao. A diversity-promoting objective function for neural conversation models. In *NAACL*, 2016.

[Li *et al.*, 2017] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, 2017.

[Luan *et al.*, 2016] Yi Luan, Yangfeng Ji, and Mari Ostendorf. Lstm based conversation models. *arXiv preprint arXiv:1603.09457*, 2016.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[Manning *et al.*, 2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL: System Demonstrations*, 2014.

[Neyshabur *et al.*, 2017] Behnam Neyshabur, Srinadh Bhojanapalli, and Ayan Chakrabarti. Stabilizing gan training with multiple random projections. *arXiv preprint arXiv:1705.07831*, 2017.

[Nguyen *et al.*, 2017] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. In *NIPS*, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Serban *et al.*, 2016] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 2016.

[Somesh Das Gupta, 2000] Jun Shao. Somesh Das Gupta. Mathematical statistics. *Springer*, 2000.

[Sutskever *et al.*, 2014] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.

[Sutton *et al.*, 2000] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 2000.

[Tao *et al.*, 2018] Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, 2018.

[Tiedemann, 2009] Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. 2009.

[Vashishth *et al.*, 2019] Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, and Partha Talukdar. Incorporating syntactic and semantic information in word embeddings using graph convolutional networks. In *ACL*, 2019.

[Xu *et al.*, 2018] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *EMNLP*, 2018.

[Yu *et al.*, 2017] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

[Zhang *et al.*, 2018] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*, 2018.