Dialogue Generation: From Imitation Learning to Inverse Reinforcement Learning

Ziming Li[1] and Julia Kiseleva[1, 2] and Maarten de Rijke[1]
[1]University of Amsterdam
[2]Microsoft Research AI
{z.li, derijke}@uva.nl, julia.kiseleva@microsoft.com

Abstract

The performance of adversarial dialogue generation models relies on the quality of the reward signal produced by the discriminator. The reward signal from a poor discriminator can be very sparse and unstable, which may lead the generator to fall into a local optimum or to produce nonsense replies. To alleviate the first problem, we first extend a recently proposed adversarial dialogue generation method to an adversarial imitation learning solution. Then, in the framework of adversarial inverse reinforcement learning, we propose a new reward model for dialogue generation that can provide a more accurate and precise reward signal for generator training. We evaluate the performance of the resulting model with automatic metrics and human evaluations in two annotation settings. Our experimental results demonstrate that our model can generate more high-quality responses and achieve higher overall performance than the state-of-the-art.

1 Introduction

The task of an open-domain dialogue system is to generate sensible dialogue responses given a dialogue context (Ritter, Cherry, and Dolan, 2010; Shang, Lu, and Li, 2015; Li et al., 2016b; Xing et al., 2017). There are two broad directions for training a dialogue generating system: the first employs defined rules or templates to construct possible responses and the second builds a chatbot to learn the response generation model with a machine translation framework from social dialogue collections (Shang, Lu, and Li, 2015; Sordoni et al., 2015a; Serban et al., 2016, 2017). Sequence-to-sequence (Seq2Seq) models enjoy the advantages of scalability and language independence and the maximum likelihood estimation objectives make it simple to train. However, in dialogue generation, the trained model suffers from generating dull and generic responses such as "I don't know" (Sordoni et al., 2015a; Serban et al., 2016; Li et al., 2016a, 2017), which are meaningless. Li et al. (2016a) suggest that "by optimizing for the likelihood of outputs given inputs, neural models assign a high probability to 'safe' responses." To alleviate this problem, Li et al. (2016b) introduced a neural Reinforcement Learning (RL) generation method to generate coherent and interesting dialogues by optimizing the manually defined reward function covering ideal dialogue properties. However, a handcrafted reward function is expensive to maintain and does not generalize over different domains (Finn, Levine, and Abbeel, 2016; Fu, Luo, and Levine, 2017). Especially for open-domain dialogue systems, it is hard to decide what knowledge is essential to design a proper reward function (Li et al., 2017). Additionally, the accuracy of defined reward functions can degrade when the dialogue context becomes more complex. Li et al. (2017) use adversarial training for dialogue generation where they jointly train two systems, a generative model to produce response sequences and a discriminator to distinguish between the human-generated dialogues and the machine-generated ones. Feedback from the discriminator is used as rewards to push the generator to produce more realistic replies. The discriminator takes a dialogue made of a context-reply pair as input and outputs the probability that this dialogue is from real human dialogues.

In Li et al. (2017), during generator training, the reward of each generated word during decoding should be supplied and Monte Carlo search is applied to estimate the reward for each word position. A potential problem is that the returned reward from the discriminator could be very sparse and unstable, which may lead the generator to produce unintended and nonsense replies. Moreover, Li et al. (2017) put no constraints on the generator policy which can result in two problems. First, the learned policy may prefer to generate general responses. Second, the training step can easily get stuck in a local optimum, which leads the generator to produce identical responses regardless of the input context or even worse – the outputs from the generator are always the same ungrammatical sentence.

In this paper, we first extend the adversarial dialogue generation method introduced by Li et al. (2017) to a new model, DG-AIL, which incorporates an entropy regularization term to the generation objective function. This addition can alleviate the problem of model collapse. Then we adopt adversarial inverse reinforcement learning to train a dialogue generation model, DG-AIRL. This method enables us to both make use of an efficient adversarial formulation and recover a more precise reward function for open-domain dialogue training. Unlike Shi et al. (2018), we design a specific reward function structure to measure the reward of each word in generated sentences while taking account of the dialogue context. We also consider two human evaluation settings to assess the overall performance of our model.

To summarize, we make the following contributions:

- A novel reward model architecture to evaluate the reward of each word in a dialog, which enables us to have more accurate signal for adversarial dialogue training;
- A novel Seq2Seq model, DG-AIRL, for addressing the task of dialogue generation built on adversarial inverse reinforcement learning;
- An improvement of the training stability of adversarial training by employing causal entropy regularization;

2 Background

Preliminaries. We build our dialogue system as a Markov Decision Process (MDP), which is defined by a tuple (S, A, τ, r, γ), where S and A are the state space and action space, respectively, τ is the transition probability, and $\tau(s, a, s')$ is the probability of transitioning from state s to state s' under action a at time t:

$$\tau(s' \mid s, a) = P(s_{t+1} = s' \mid s_t = s, a_t = a). \quad (1)$$

Here, $r(s, a)$ is the immediate reward after taking action a in state s; $\gamma \in [0, 1]$ is a discount factor.

The *dialogue response strategy* is represented by a policy, which is a mapping $\pi \in \Pi$ from states $s \in S$ and actions $a \in A$ to $\pi(a|s)$, which is the probability of performing action $a_t = a$ by the user when in state $s_t = s$:

$$\pi(a|s) = P(a_t = a \mid s_t = s). \quad (2)$$

Maximum causal entropy. Motivated by the task of decision prediction in sequential interactions, Ziebart, Bagnell, and Dey (2010) propose to use maximum causal entropy to model the availability and influence of sequentially revealed side information. The causal entropy of policy π is defined as:

$$H(\pi) \triangleq E_\pi[-\log \pi(a|s)]) \quad (3)$$

which measures the uncertainty presented in policy π (Ziebart, Bagnell, and Dey, 2010).

Maximum entropy inverse reinforcement learning (MaxEnt-IRL). Given a set of demonstrated (expert) behavior, which can be seen as the resulting trajectories by executing expert policy π_E, Inverse Reinforcement Learning (IRL) aims to find a reward function that can rationalize the given behavior. In Maximum Entropy Inverse Reinforcement Learning (MaxEnt-IRL) (Ziebart et al., 2008), the demonstrated behavior $D_{demo} = \{\zeta_1, \ldots, \zeta_N\}$ is assumed to be the result of an expert acting stochastically and near-optimally with respect to an unknown reward function. Trajectories with equivalent rewards have equal probability to be selected and trajectories are sampled from the distribution:

$$p(\zeta_i \mid \theta) = \frac{1}{Z(\theta)} \exp^{r_\theta(\zeta_i)} = \frac{1}{Z(\theta)} \exp^{\sum_{t=0}^{|\zeta_i|-1} r_\theta(s_t, a_t)}, \quad (4)$$

where $Z(\theta) = \int \exp(r_\theta(\zeta)) d\zeta$ is the partition function and r_θ is the reward function, which takes a state-action pair as input. MaxEnt-IRL maximizes the likelihood of the demonstrated data D_{demo} under the maximum entropy (exponential family) distribution and the objective is given as:

$$L(\theta) = -\mathbb{E}_{\zeta \sim D_{demo}} r_\theta(\zeta) + \log Z \quad (5)$$

This task can be seen as a classification problem where each trajectory represents one class. However, it is difficult to apply vanilla MaxEnt-IRL to complex and high-dimensional settings since computing the partition function $Z(\theta)$ is intractable in the original method. To overcome this drawback, Finn, Levine, and Abbeel (2016) combine sample-based maximum entropy IRL with forward reinforcement learning to estimate the partition function Z, where:

$$L(\theta) = -\mathbb{E}_{\zeta_i \sim p} r_\theta(\zeta_i) + \log \left(\mathbb{E}_{\zeta_j \sim q} \left[\frac{\exp(r_\theta(\zeta_j))}{q(\zeta_j)} \right] \right). \quad (6)$$

Here, p represents the distribution of demonstrated samples, while q is the background distribution for estimating the partition function $\int \exp(r_\theta(\zeta)) d\zeta$. This work alternates between updating the reward function r_θ to maximize the likelihood of the demonstrated data and optimizing the background distribution q to minimize the variance of the importance sampling estimation.

Generative adversarial imitation learning. Recovering the true reward function is intractable in real scenarios (Ziebart et al., 2008; Ho and Ermon, 2016; Fu, Luo, and Levine, 2017). In previous research, if only the optimal policy is pursued, imitation learning is used to rebuild the policy network directly by skipping recovering reward functions. Ho and Ermon (2016) cast the problem of IRL as an optimization problem in the paradigm of Generative Adversarial Networks (GANs), where the discriminator corresponds to the reward function and the generator corresponds to the policy used to sample trajectories. The optimization problem is given as:

$$\max_{r \in R} \left(\min_{\pi \in \Pi} -\lambda H(\pi) - \mathbb{E}_\pi[r(s, a)] \right) + \mathbb{E}_{\pi_E}[r(s, a)]. \quad (7)$$

The optimization of Eq. 7 is converted to an imitation learning algorithm:

$$\min_{\pi \in \Pi} -\lambda H(\pi) + D_{JS}(\rho_\pi, \rho_{\pi_E}), \quad (8)$$

which finds a policy π whose occupancy measure ρ_π minimizes the Jensen-Shannon divergence to the expert's policy π_E (the policy of demonstrated data). The occupancy measure ρ_π can be interpreted as the unnormalized distribution of state-action pairs that an agent encounters when navigating the environment with policy π. Eq. 8 can be solved by finding a saddle point (π, D) of the expression

$$\mathbb{E}_\pi[-\log(D(s, a))] + \mathbb{E}_{\pi_E}[-\log(1 - D(s, a))] - \lambda H(\pi), \quad (9)$$

where D is a binary classifier to distinguish state-action pairs of π and π_E.

3 Method

In this section, we will first extend the work (Li et al., 2017) to the framework of adversarial imitation learning, and then introduce our main model which applies adversarial inverse reinforcement learning to train a dialogue system.

3.1 Problem setting

In a dialogue setting, the word sequence $\langle w_1, w_2, \ldots, w_t \rangle$ in an utterance can be regarded as corresponding actions $\langle a_1, a_2, \ldots, a_n \rangle$ taken by the policy network at different time steps. We use a state function f to compress the dialogue context and the words already generated in the current utterance to an intermediate representation, which will be regarded as the current state. For example, $s_0 = f(p)$ represents the state at time step 0 and it takes the dialogue context p as input. State s_t is given as $s_t = f(p, a_1, a_2, \ldots, a_{t-1})$. In this work, we limit the range of the dialogue context to the utterances in the last two conversation turns.

Given an initial state s_0 representing the history of previous dialogues, a well-trained dialogue system should reply with a reasonable sentence $\langle w_0, w_1, \ldots, w_t \rangle$ generated by selecting a specific word at different time steps. The length t is automatically decided by the policy network. We aim to find the optimal policy $\pi(a_t|s_t)$ that selects the most appropriate word at each time step.

3.2 Dialogue generation with adversarial imitation learning (DG-AIL)

In the framework of adversarial imitation learning, we aim to train a dialogue system to imitate the way humans talk by observing real human dialogues. This model DG-AIL can be regarded as an extension of the work of Li et al. (2017). Unlike previous work, we do not only consider the difference between the distributions of real dialogues and generated dialogues but also take into account how the previous state-action pairs affect future words under a specific policy network π, which can be measured by the causal entropy $H(\pi)$.

In adversarial learning, the task of the discriminator D is to distinguish dialogues from the true data distribution and dialogues from the generator. As shown in Fig.1, we adopt a hierarchical structure to represent the discriminator model. The first layer is an input encoder which compresses the utterances from each speaker in the conversation. Then, a context encoder sequentially takes as input the utterance representations and generates a final state to represent the whole dialogue. In the end, the final state is fed to a binary classifier which predicts whether the dialogue is real or fake with a confidence value.

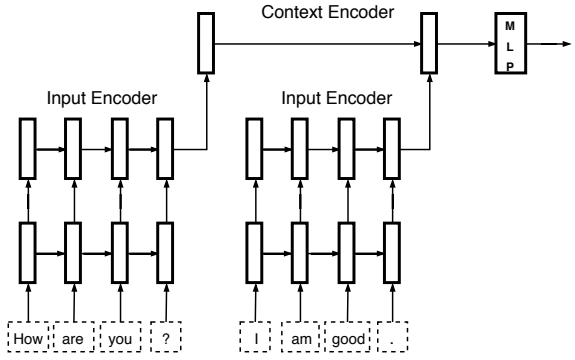

Figure 1: Discriminator architecture in DG-AIL.

According to Eq. 9, the gradient of the discriminator parameters is given as:

$$\begin{aligned}\nabla D_\theta = \mathbb{E}_{\zeta\sim\pi}[\nabla_\theta \log(D_\theta(s,a))] + \\ \mathbb{E}_{\zeta\sim\pi_E}[\nabla_\theta \log(1 - D_\theta(s,a))]\end{aligned} \tag{10}$$

The generative model G attempts to generate high-quality human-like responses to confuse the discriminative classifier D while maintaining high policy entropy. The gradient to update generator parameters can be inferred from Eq. 9 as follows:

$$\begin{aligned}\nabla G_\phi &= \nabla_\phi[-\lambda H(\pi_\phi) - \mathbb{E}_{\zeta\sim\pi_\phi}[D_\theta(s,a)]] \\ &= -\mathbb{E}_{\zeta\sim\pi_\phi}\nabla_\phi[\log(\pi_\phi(a|s))](Q_\theta(s,a) - \\ &\qquad \lambda \log \pi_\phi(a|s)),\end{aligned} \tag{11}$$

where $Q_\theta(s,a) = \mathbb{E}_\zeta[\log(D_\theta(s,a))|s_0 = \bar{s}, a_0 = \bar{a}]$ is estimated with Monte Carlo search.

3.3 Dialogue reward learning with adversarial inverse reinforcement learning (DG-AIRL)

Our main model DG-AIRL adopts inverse reinforcement learning techniques to train a dialogue generation model. We assume that human participants in a dialogue are using a true reward function that guides them to formulate a policy to react with different replies to different dialogue contexts. Unlike the use of a classifier to supply a reward signal in the model DG-AIL, the reward model in DG-AIRL has a more specific architecture to evaluate the reward for each state-action pair, which can provide more accurate and precise reward signal to update the generator.

Dialogue response policy. In maximum entropy inverse reinforcement learning, the reward model (discriminator) attempts to assign high rewards to demonstrated trajectories (from the expert policy) and low rewards to sampled trajectories from other policies. In this way, when the reward function is fixed, the expert policy can be found by solving a common reinforcement learning problem:

$$G_\phi(r_\theta) = \arg\min_{\pi\in\Pi} -\lambda H(\pi) - \mathbb{E}_{\zeta\sim\pi}[r_\theta(\zeta)], \tag{12}$$

where ζ represents the sampled dialogues and $H(\pi)$ is the causal entropy regularization term. $r_\theta(\zeta)$ is the reward of dialogue ζ which can be accessed from the reward model. The goal of the generator is to generate dialogues that can achieve higher rewards from the reward model. The found policy maximizes the expected cumulative reward while maintaining high-entropy. The derivative can be inferred as follows:

$$\begin{aligned}\nabla_\phi G(r)_\phi &= \nabla_\phi[-\lambda H(\pi_\phi) - \mathbb{E}_{\zeta\sim\pi_\phi}[r_\theta(\zeta)]] \\ &= -\mathbb{E}_{\zeta\sim\pi_\phi}\nabla_\phi[\log(\pi_\phi(\zeta))](r_\theta(\zeta) - \\ &\qquad \lambda \log \pi_\phi(\zeta)).\end{aligned} \tag{13}$$

If we decompose dialogue ζ into different time steps, the gradient is given as:

$$\begin{aligned}\nabla_\phi G(r)_\phi &= -\mathbb{E}_{\zeta\sim\pi_\phi}\nabla_\phi[\log(\pi_\phi(\zeta))](r_\theta(\zeta) - \\ &\qquad \lambda \log \pi_\phi(\zeta)) \\ &= -\textstyle\sum_t \mathbb{E}_{\pi_\phi(a_t|s_t)}\nabla_\phi[\log(\pi_\phi(a_t \mid s_t))] \\ &\qquad (r_\theta(\zeta_{t:T}) - \lambda \log \pi_\phi(a_t \mid s_t)).\end{aligned} \tag{14}$$

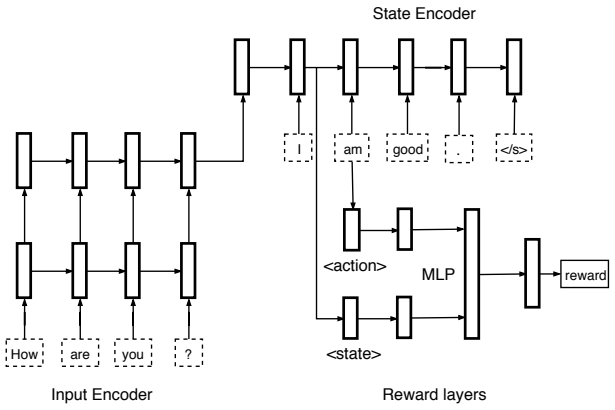

Figure 2: Reward model architecture in DG-AIRL.

The reward $r_\theta(\zeta_{t:T})$ of a partial dialogue from time t to T is estimated with Monte Carlo search.

Reward learning. Following the work of sample-based maximum entropy IRL (Eq. 6), the objective (loss function) of our reward model is given as:

$$L(\theta) = -\mathbb{E}_{\zeta_i \sim \pi_E} r_\theta(\zeta_i) + \log \mathbb{E}_{\zeta_j \sim \pi} \left(\frac{\exp(r_\theta(\zeta_j))}{q(\zeta_j)} \right), \quad (15)$$

where π_E denotes the policy of demonstrated trajectories and π the policy of background samples. The term q denotes the background distribution from which dialogues ζ_j were sampled. In our setting, q is the distribution of dialogues generated with the current dialogue policy π. We use D_{demo} and D_{samp} to represent the set of dialogues generated with policy π_E and π, respectively.

The gradient of the reward function is given by:

$$\begin{aligned} \nabla_\theta L(\theta) = & -\mathbb{E}_{\zeta_i \in D_{demo}} \nabla_\theta r_\theta(\zeta_i) + \\ & \frac{1}{Z} \sum_{\zeta_j \in D_{samp}} w_j \nabla_\theta r_\theta(\zeta_j), \end{aligned} \quad (16)$$

where $w_j = \frac{\exp(r_\theta(\zeta_j))}{q(\zeta_j)}$ and $Z = \sum_j w_j$.

As shown in Fig. 2, our reward model in DG-AIRL consists of two RNN encoders and one MLP network. The input encoder compresses the utterances from the context into a context representation which becomes the initial state in the next step. The state encoder takes as input the dialogue context and generated words before time t and outputs the new state representation s_t for time step t. Then the state and action representations are fed to two separate MLP layers respectively. The outputs of these two models are concatenated and form the input to the third MLP layer to get the final reward value of current state-action pair $\langle s_t, a_t \rangle$.

4 Experimental Setup

4.1 Dataset

The MovieTriples (Serban et al., 2016) has been developed by expanding and preprocessing the Movie-Dic corpus (Banchs, 2012) of film transcripts and each dialogue consists of 3 turns between two interlocutors. The dialogues are collected from the scripts of more than 600 movies, which span a wide range of topics. We limit the length of the utterances from one speaker in each dialogue turn between 4 and 80. In the final dataset, there are around 157,000 dialogues in the training set, 19,000 in the validation set and 19,000 in the test set. The average length of each dialogue is about 54.

4.2 Experimental settings

We limit the vocabulary table size to the top 20k most frequent words for the MovieTriples dataset. All words that are not in the vocabulary tables are replaced with the token '$\langle unk \rangle$'. Following the preprocessing method from (Serban et al., 2016), all names and numbers are replaced with the '$\langle person \rangle$ and '$\langle number \rangle$' tokens, respectively (Ritter, Cherry, and Dolan, 2010). Since the context input in each dialogue is made up of several utterances from two different speakers, to capture the interactive structure, we insert a special token '$\langle /s \rangle$' between the first turn and the second. The word embedding size is 200.

Next, we list the models we consider. We implement all models based on Tensorflow[1] except VHRED.

DG-AIRL. This is our main model that adopts adversarial inverse reinforcement learning techniques to train a dialogue system. The encoder and decoder in the generator (policy network) are built from a 2-layer GRU with 1024 hidden units and an attention mechanism is incorporated into the decoding step. With respect to reward function structure, we choose a 2-layer GRU with 1024 hidden units as the context encoding layer to compress the input to an intermediate representation. Then a 1-layer GRU with 1024 hidden units is used to take state-action pair as input and output the next state as shown in Fig.2.

Seq2Seq. The encoder and decoder in this baseline are copied from the generator in the DG-AIRL model and built from 2-layer GRU with 1024 hidden units; an attention mechanism is incorporated into the decoding step.

SeqGan. This is the model from (Li et al., 2017). In terms of the generator, SeqGan shares the same architecture as the DG-AIRL model. With respect to the discriminator in this model, both the first encoder layer and the second context layer are built from a 2-layer GRU with 1024 units separately.

VHRED. For the VHRED model, we reuse the original implementation from the authors including their tuning techniques.[2]

DG-AIL. This is the model with the adversarial imitation learning method, which is also an extension of SeqGan. The DG-AIL model shares the same structure as the SeqGan model including generator and discriminator. The only difference is the loss function, as we discussed in Section 3.

We optimize the models using Adam (Kingma and Ba, 2014) and the learning rate is initialized as 0.001 except for VHRED. Dropout with probability 0.3 was applied to the GRUs and we apply gradient clipping for both policy models and reward

[1] https://www.tensorflow.org/

[2] For more details, see `https://github.com/julianser/hed-dlg-truncated`.

models. We set the beam size to 8 for Monte Carlo search during training and beam search during testing. During the training of SeqGan, DG-AIL, and DG-AIRL, we employ the teacher-forcing technique from Li et al. (2017) to increase training efficiency[3].

4.3 Evaluation metrics

To evaluate the response quality in dialogue generation, recent work has adopted word-overlap metrics from machine translation to compare a machine-generated response to a single target response. Since the response to the input context in dialogue could be very diverse and open, a single target response is not able to cover all reasonable answers. Liu et al. (2016) show that word-overlap metrics such as BLEU correlate very weakly with reply quality judgments from human annotators. To assess the performance of our proposed algorithm, we use two evaluation methods, one is to use word embedding based metrics and the other is to employ human annotators to judge the response quality. We also tried to evaluate the response diversity with metric *Distinct* but we found the result is not aligned with the result on human evaluations. In this work we did not attach the result on *Distinct*.

Embedding metrics. With respect to word embedding based methods, we use three metrics that are also used in (Serban et al., 2017):

- *Average embedding*: This method applies cosine-similarity to measure the similarity between the mean word embeddings of the target utterance and the predicted utterance.
- *Greedy embedding*: This metric relies on cosine-similarity but adopts greedy matching to find the closest word in the target response for each word in the generated response (Rus and Lintean, 2012).
- *Extrema embedding*: This method computes the word embedding extrema scores (Forgues et al., 2014) that embed the responses by taking the extrema (maximum of the absolute value) of each dimension, and afterward computes the cosine similarity between them.

A higher score indicates that the generated reply share similar semantic content with the target response. For all three metrics, we use pre-trained Word2Vec word embeddings trained on the Google News Corpus, which is public access.

Human evaluation. A proper quality evaluation of dialogue responses should cover not only topic-similarity but also lexical aspects, informativeness, interestingness and so on. There is currently no reliable metric to assess the overall quality of dialogue responses. For this reason, we create human annotations to evaluate the quality of responses given dialogue context with a crowdsourcing platform.[4] Previous work involving human evaluation usually has two experimental settings: pairwise comparison and pointwise scoring. In pairwise comparisons, annotators are asked to choose the better response from replies generated by two models while in the pointwise method, annotators are asked to rate the overall quality of each response, typically on a scale from 0 (low quality) to 4 (high quality). We employ both pairwise and pointwise assessments. We use a pairwise setting to directly contrast the overall performance of our model against others. Pointwise scoring may be noisier than pairwise judgments since human annotators need to give an exact score. However, pointwise judgments give us a chance to analyze the differences between replies at a fine-grained level of detail.

We randomly sample 1,000 dialogue contexts from the test set of the MovieTriple dataset. Each dialogue context has five replies from five generation models and we have 5,000 context-reply pairs in total; 2,500 are used for pointwise scoring while the remaining 2,500 are grouped into 2,000 comparison pairs for the pairwise setting. Each comparison pair has one dialogue context and two replies, where one is from our DG-AIRL model and the other is from a baseline model.

For the pairwise setting, we ask annotators to judge which of two responses is more appropriate given a dialogue context. We instruct annotators which aspects they should take into account when making a decision. The top priority is that an appropriate response must be relevant; in addition, they should consider:

- If the response is natural;
- If the response is interesting;
- If the response can make the conversation continue which means the response is more proactive;
- If the response is the only possible reply to the given context.

If the annotators think neither of the responses is more appropriate or it is impossible to infer the conversation from the given context, they are asked to choose the third choice – "Neither is more appropriate." We insert test questions to exclude annotators who lack the capacity to finish the tasks, such as limited English skills. We only accept annotators considered "highly trusted" by the crowdsourcing platform and require 90% accuracy on designed "test questions." Each comparison pair is assessed by three annotators.

For pointwise judgments, annotators were asked to judge the overall quality from 0 to 2:

- +2: The response is not only relevant and natural, but also informative and interesting; the response needs not be so interesting, but it is natural and can make the conversation continue (more proactive); the response is the only possible reply to the dialogue.
- +1: The response can be used as a reply to the context, but it is too generic like "I don't know." These replies are usually more reactive.
- 0: The response cannot be used as a reply to the context. It is either semantically irrelevant or disfluent.

At the start of the annotation effort, we instruct the annotators and show them several examples of how to assign grades to a given dialogue. We use the same quality checks and annotator selection criteria as in the pairwise setting. Each context-reply is assigned to three human annotators.

[3]The source code of this work is available at https://bitbucket.org/ZimingLi/dg-irl-aaai2019

[4]FigureEight, `https://www.figure-eight.com/`.

5 Results and Analysis

5.1 Results using embedding metrics

In Table 1, we report scores obtained using the embedding metrics (Section 4.3). All response generation models are fine-tuned to obtain the highest score on the validation dataset. We found 0.01 and 0.1 to be the optimal values of λ for the DG-AIL model and the DG-AIRL model.

The DG-AIRL model achieves the highest scores using the embedding metrics, which means that it can better capture the topic of the target response than other models. The performance of VHRED model is unexpected since this method achieves the lowest value while it is one of the state-of-the-art methods in dialogue response generation. The possible reason is that the other four models adopted an attention mechanism to capture the relation between generated words and input words from context directly. Serban et al. (2017) state that VHRED produces longer responses and its responses are on average more diverse based on unigram entropy. Besides, DG-AIL and DG-AIRL are also supposed to generate more diverse responses. However, a more diverse response does not mean it is an appropriate response. If the response deviates too much from the target topic, the response content will not be relevant to the dialogue context and it deserves a lower quality score. On the other hand, if a diverse response is appropriate to the dialogue context, it is unfair to use embedding-based metrics to assess these kinds of generative models. The same happened to the performance of DG-AIL and SeqGan that the entropy regularization term in the DG-AIL loss function did not improve the score on embedding-based metrics unexpectedly. With this consideration, embedding-based metrics are not able to reflect the overall response quality and it is essential to carry out human evaluations.

Model	Average	Greedy	Extrema	Length
Seq2Seq	0.563 ± 0.003	0.167 ± 0.001	0.352 ± 0.002	8.8
SeqGan	0.564 ± 0.003	0.165 ± 0.001	0.354 ± 0.002	9.7
VHRED	0.507 ± 0.003	0.145 ± 0.001	0.309 ± 0.002	**12.0**
DG-AIL	0.553 ± 0.003	$\mathbf{0.171^* \pm 0.001}$	0.356 ± 0.002	7.7
DG-AIRL	$\mathbf{0.589^* \pm 0.003}$	0.169 ± 0.001	$\mathbf{0.368^* \pm 0.002}$	10

Table 1: Performance in terms of embedding metrics of response generation models, with 95% confidence intervals. * indicates the result is statistically significant ($p < 0.005$) with a paired t-test over DG-AIRL and other baseline models.

5.2 Results using human annotations

Pairwise evaluation. As shown in Table 2, our model DG-AIRL outperforms other response generation models based on the pairwise comparison. Among the first three models, the DG-AIRL model wins them all at the probability 0.46. The difference is that the **Win** rate of VHRED (the lose rate of DG-AIRL) is lower than Seq2Seq and SeqGan. In another word, although VHRED has a higher probability to be tied with DG-AIRL but it loses more compared to Seq2Seq and SeqGan. As we said in the last section, the possible reason is that VHRED does produce longer responses (Table 1) but the contents of these responses deviate too much from the target topic, which could result in lower performance. In our human evaluation setting, fluency is not the only aspect annotators need to consider while judging the preference of two responses. By taking into account different factors, such as relevance, fluency, informativeness, we think the final judgments from human annotators are trustworthy. The Fleiss' kappa which indicates the agreements among labelers (Fleiss and Cohen, 1973) is around 0.23. This value is not high and the possible reason is that judging the response quality is challenging for human annotators when only 1 or 2 utterances of context are provided, especially on the MovieTriples dataset.

Model pair	Win	Tie	Loss
DG-AIRL-Seq2Seq	**0.44**	0.29	0.27
DG-AIRL-VHRED	**0.46**	0.32	0.22
DG-AIRL-SeqGan	**0.47**	0.25	0.28
DG-AIRL-DG-AIL	0.36	**0.37**	0.27

Table 2: Performance in terms of pairwise human annotations of response generation models.

Model	Freq of +2	Freq of +1	Freq of 0	Avg Score
Seq2Seq	0.09	0.22	0.69	0.40
SeqGan	0.09	0.21	0.70	0.39
VHRED	0.12	0.25	0.63	0.49
DG-AIL	0.12	0.29	0.59	0.53
DG-AIRL	0.13	0.28	0.59	**0.54**

Table 3: Performance in terms of pointwise human evaluations of response generation models. "Freq of N" is the frequency of a model's responses with a score of N.

Compared to the first three models, the DG-AIL model achieves better performance. According to the performance of DG-AIL and SeqGan, we can make a statement that causal entropy regularization improves the performance of dialogue models that employ adversarial training because they share the same structure and the only difference is the entropy regularization term. This also echoes the usefulness of entropy regularization in adversarial models. Although the DG-AIRL model still beats DG-AIL, the difference between the **Win** rate and the **Loss** is much smaller compared to other models. DG-AIL model has a higher chance to draw with the DG-AIRL model. In these two models, they both adopt entropy regularization and they have the same generator structure. The performance difference comes from the reward model. DG-AIRL is able to improve the response quality because it has a specific reward model for each state-action pair and adopts importance sampling. The reward signal in DG-AIRL is more concrete and reliable compared to DG-AIL.

Pointwise evaluation. The results of pointwise judgments are shown in Table 3. According to the average score, the DG-AIRL and DG-AIL models outperform other models and the performance between DG-AIRL and DG-AIL are quite close. It does not mean the result of the pointwise evaluation is in conflict with the conclusion we made in the last section, that DG-AIRL model beats DG-AIL model based on the pairwise comparison. Compared to the pairwise comparison, pointwise evaluation needs to assign an exact score to each context-reply pair and this score is independent of the other replies to the same dialogue context. In contrast, the pairwise

Context	Next Reply from Speaker A
Speaker A: we did a story on this guy . he owns half of arizona . **Speaker B:** is he a fraud ?	**Reference:** how would you ever know ? **Seq2Seq:** yeah . no . i do not know who he is . **SeqGan:** sure , i guess not . **VHRED:** he doesn't want to know . **DG-AIL:** no . no , he is not . **DG-AIRL:** no . maybe he is a good man .
Speaker A: and i am honored that you want to be with me because you would never be with anyone who was not truly worthy **Speaker B:** i feel exactly the same way about you	**Reference:** oh, $\langle person \rangle$, do not say that , please , that just makes it worse . **Seq2Seq:** do you love your $\langle unk \rangle$. **SeqGan:** can we i . you . you . you . you . you . you . . you . **VHRED:** $\langle person \rangle$? **DG-AIL:** what is it . **DG-AIRL:** is that what you think .
Speaker A: oh , that was not my father . **Speaker B:** what ?	**Reference:** that was not him . i heard him tell the cops . **Seq2Seq:** i am sorry . i needed to know how i felt about it . **SeqGan:** you said he was a singer . **VHRED:** i can't . i'm sorry . **DG-AIL:** i do not know . i guess . **DG-AIRL:** it is not him . i do not know what happened .

Table 4: Example output. People names are replaced with the token '$\langle person \rangle$'. All words are converted to lowercase.

comparison looks at a pair of replies to the same context at the same time and it is more natural and reliable if we want a performance ranking list. The advantage of pointwise setting is that it can provide the quality distributions of different models and help us find out what makes the model performance different.

As shown in Table 3, we can find that VHRED, DG-AIL, and DG-AIRL generate almost the same number of high-quality responses (responses who get score "+2"). The VHRED model loses the competition with DG-AIRL and DG-AIL models because it generates more low-quality replies (responses who get score "0"). In our experiment setting, we ask annotators to grade the reply quality as "+1" (fine quality) if the response can be used as a reply to the message, but it is too generic. In Section 1, we expect to generate more diverse responses and avoid producing too generic responses, such as "I don't know". However, in some dialogue contexts, "I don't know" is still an appropriate and reasonable response. As shown in Table 4, DG-AIRL and DG-AIL improve the proportion of high-quality responses without losing the capacity to generate fine quality replies.

6 Related Work

With the developing of sequential neural networks, Shang, Lu, and Li (2015); Sordoni et al. (2015b) propose to generate high-quality replies in a dialogue system with a recurrent neural network. To formulate the complex dependencies between different utterances in multi-turn dialogs, Serban et al. (2016) propose to adopt hierarchical recurrent encoder-decoder neural network (HRED) to the dialogue domain, where word-level and utterance-level Recurrent Neural Networks (RNN) are used. Built on HRED, the same group extend it to a more powerful generative architecture (Serban et al., 2017)with latent stochastic variables that span a variable number of time steps (VHRED). To train these RNN models, supervise training is commonly used which minimize the cross-entropy between the generated reply and an oracle reply. However, in terms of open-domain dialogue systems, there could be multiple reasonable replies for the same input context. In other words, the entropy of the target replies is high. Li et al. (2017) cast the task of open-domain dialogue generation as a RL problem and train a generator based on the signal from a discriminator to generate response sequences indistinguishable from human-generated dialogs.

7 Conclusion

In this work, we have investigate two different adversarial training methods for open-domain dialogue systems. We firstly adopt adversarial imitation learning to push our model to generate human-like dialogue responses. Besides, we incorporate an entropy regularization term to the generator objective function, which can alleviate the problem of model collapse. Our second and main method DG-AIRL relies on the techniques of adversarial inverse reinforcement learning. We design a specific reward architecture to supply more accurate and precise reward signal for the generator training. To confirm the overall performance of our model, we propose two different human-evaluation settings. We adopt the results from pairwise setting to show our model can beat state-of-the-art methods in open-domain dialogue generation. To analyze the reply difference from different models, we explore the results from the point-wise setting which can provide a general quality distribution for different models.

In the future, we plan to extend the reward learning to multi-turn dialogue generation which can propagate the reward signal between different conversation turns. Other research direction is to explore the more potential usefulness of recovered reward models, such as evaluating the quality of generated responses from other models.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banchs, R. E. 2012. Movie-dic: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 203–207. ACL.

Finn, C.; Christiano, P.; Abbeel, P.; and Levine, S. 2016. A connection between generative adversarial networks,

inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*.

Finn, C.; Levine, S.; and Abbeel, P. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, 49–58.

Fleiss, J. L., and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33(3):613–619.

Forgues, G.; Pineau, J.; Larchevêque, J.-M.; and Tremblay, R. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

Fu, J.; Luo, K.; and Levine, S. 2017. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*.

Ho, J., and Ermon, S. 2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 4565–4573.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kramer, G. 1998. *Directed information for channels with feedback*. Ph.D. Dissertation, Swiss Federal Institute of Technology Zurich.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. 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*, 110–119.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1192–1202.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2157–2169.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132.

Permuter, H. H.; Kim, Y.-H.; and Weissman, T. 2008. On directed information and gambling. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, 1403–1407. IEEE.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180. ACL.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, 583–593. ACL.

Rus, V., and Lintean, M. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 157–162. ACL.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, 3776–3784.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 3295–3301.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, 1577–1586.

Shi, Z.; Chen, X.; Qiu, X.; and Huang, X. 2018. Towards diverse text generation with inverse reinforcement learning. *arXiv preprint arXiv:1804.11258*.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015a. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 196–205.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015b. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Wen, T.-H.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojas-Barahona, L. M.; Su, P.-H.; Ultes, S.; and Young, S. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic aware neural response generation. In *AAAI*, volume 17, 3351–3357.

Yu, Z.; Xu, Z.; Black, A. W.; and Rudnicky, A. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 404–412.

Ziebart, B. D.; Bagnell, J. A.; and Dey, A. K. 2010. Modeling interaction via the principle of maximum causal entropy. In *27th International Conference on Machine Learning, ICML 2010*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *AAAI*, 1433–1438. AAAI Press.