The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Exploring Implicit Feedback for Open Domain Conversation Generation

Wei-Nan Zhang, Lingzhi Li, Dongyan Cao, Ting Liu[*]
Research Center for Social Computing and Information Retrieval
School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
{wnzhang, lzli, dycao, tliu}@ir.hit.edu.cn

Abstract

User feedback can be an effective indicator to the success of the human-robot conversation. However, to avoid to interrupt the online real-time conversation process, explicit feedback is usually gained at the end of a conversation. Alternatively, users' responses usually contain their implicit feedback, such as stance, sentiment, emotion, etc., towards the conversation content or the interlocutors. Therefore, exploring the implicit feedback is a natural way to optimize the conversation generation process. In this paper, we propose a novel reward function which explores the implicit feedback to optimize the future reward of a reinforcement learning based neural conversation model. A simulation strategy is applied to explore the state-action space in training and test. Experimental results show that the proposed approach outperforms the Seq2Seq model and the state-of-the-art reinforcement learning model for conversation generation on automatic and human evaluations on the OpenSubtitles and Twitter datasets.

Introduction

Conversational robot is one of the most interesting and challenging topics in artificial intelligence research. It is usually developed to imitate the human-human chatting and applied to many scenarios, such as chitchat, interactive question answering, task-oriented dialogue, interactive recommendation, etc. With the blooming of deep neural network, neural conversation models show amazing promise for conversation generation on two major categories. **One** is single turn response generation which is a context insensitive responding process that generates a response by only considering an input message. Many research focuses on the single turn response generation (Shang, Lu, and Li 2015; Vinyals and Le 2015; Li et al. 2016a; Dai and Le 2015; Li et al. 2016b; Yao, Zweig, and Peng 2015; Mou et al. 2016; Xing et al. 2016b; Vougiouklis, Hare, and Simperl 2016; Xing et al. 2016a). It is also worth mentioning that the single turn response generation task is also promoted by the short text conversation (STC) task of NTCIR-12 (http://ntcir12.noahlab.com.hk/stc.htm) and 13 (http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm). However, the main drawback of most single turn response generation models is that they tend to frequently generate generic responses, such as "*Me, too*", "*I don't know*", due to the maximum likelihood training (Li et al. 2016a; Serban et al. 2016c). Moreover, these generic responses are not ease to respond and thus lead to an early close of conversations (Li et al. 2016c).

Table 1: Human-human conversation segments that are sampled from the OpenSubtitles dataset and contain the implicit feedback.

	Conversation Segments	**Implicit Feedback**
1	A: All the girls ogle him, but he doesn't turn me on. B: I don't think he's your type.	*Stance*
2	A: I love you so much. Do you have any feelings for me? B: Would I see you if I didn't?	*Emotion*
3	A: You look pretty in anything. Are you that happy? B: I am.	*Sentiment*
4	A: What's my least favorite word? B: Phlegm. A: You're just guessing. It's "panties". B: Ohh.	*Stalemate*

The other is multi-turn conversation generation which utilizes the recurrent neural network (RNN) based encoder-decoder (Serban et al. 2016a; 2016c; 2016b) and the end-to-end memory network (Bordes and Weston 2016) to model the context information. Recently, the reinforcement learning based conversation generation approaches are proposed to modeling long-term reward (Li et al. 2016c; Asghar et al. 2016; Dhingra et al. 2016). The goal of these models is to generate coherent responses to make the conversations easy to sustain (Li et al. 2016c).

Despite the success of previous work on multi-turn neural conversation generation, a serious issue still remains: The existing approaches for neural conversation generation have not considered the **implicit feedback** on both of offline and online learning of the human-robot conversation process.

[*] Corresponding Author

Table 2: Sampled simulated conversations that are generated by the baseline reinforcement learning (RL) model (Li et al. 2016c)(**Left column**) and our proposed implicit feedback model(**Right column**). Both of the models are trained on the OpenSubtitles dataset. The utterance [1] is input by the authors and the two models then start to chat by taking an output of a model as the input of the other model. The simulation is stopped when the conversations are repetitive.

Baseline RL model (Li et al. 2016c)	**Proposed implicit feedback model**
A: What are you like? [1]	A: What are you like? [1]
B: I'm so confused. [2]	B: One of your favs. [2]
A: You know it. [3]	A: That's so cute. [3]
B: Thank you !!! [4]	B: Thank you love. [4]
A: You're welcome. [5]	A: You're welcome xoxo. [5]
B: Lmao I hate this. [6]	B: You're the best! [6]
A: What do you mean? [7]	A: I'll be good to you. [7]
B: Nah I hate it. [8]	B: I like you too. [8]
(Repeat)	A: lol I know. [9]
...	B: This is true. [10]
...	A: That's what I'm saying. [11]
...	(Repeat)

Table 2 shows an example of two simulated conversations[1] which are with (Right column)/without (Left column) modeling the implicit feedback for conversation generation, respectively. To see the entire content in Table 2, we find that the implicit feedback are ubiquitous in the conversations. The utterances of [2], [4], [6], [8] in the left column and [2]-[9] in the right column contain rich sentiments. However, comparing the two simulated conversations, the baseline model fails to capture the implicit feedback that are conveyed by the words of "confused" and "hate" and generate bad actions(the utterances of [3] and [7]).

In this paper, we aim to verify two assumptions:

- The implicit feedback which is involved in an utterance can impact the long term goal of the conversations.
- The reward function which is composed by the implicit feedback can sustain the human-robot conversations in a positive state and not falling into stalemate.

In this paper, we explore the implicit feedback to optimize the generation of open domain conversations for human-robot conversation. More concretely, we integrate the implicit feedback into the reward function to optimize the long-term goal of conversation generation. Two conversational robots are simulating to explore the action-state space for learning to maximize the reward expectation. The sampled results in the right column of Table 2 illustrate that our proposed implicit feedback model can better sustain the conversation than the baseline model.

Implicit Feedback Model

In this section, we will detail the proposed implicit feedback model for conversation generation. We will first analyze the implicit feedback that can be explored during the conversation process. Second, we will briefly introduce the reinforcement learning based conversation model. Third, we will present how to integrate the implicit feedback into the conversation modeling process.

Implicit Feedback Exploration

Implicit feedback, such as stance, sentiment or emotion, etc., are ubiquitous in conversations and have important impact on sustaining conversations. Table 1 shows the conversation segments that involve the implicit feedback. They are sampled from the OpenSubtitles dataset. For the conversation segment 1, the response "*I don't think he's your type.*" shows the opposite stance of speaker B towards the previous utterance of speaker A. The conversation segment 2 and 3 show the emotion of "love" and the sentiment of "happy", respectively. For the conversation segment 4, it is in a stalemate state which makes the conversation hard to carry on.

There may be other types of implicit feedback. As a preliminary attempt, in this paper, we plan to explore the (1)**stance** and (2)**sentiment** as well as the conversation state of (3)**stalemate** for conversational robots to generate open domain conversations.

It is worth noting that the implicit feedback can be used for conversation generation in two ways: First, the implicit feedback can be identified from the candidate utterances generated by a conversational robot and used to estimate the future reward of the generated utterances in human-robot conversations. In this case, the implicit feedback comes from the candidate utterances generated by a conversational robot. Second, the implicit feedback can be detected from the human generated utterances and conditioned as a "guidance" or extra context for a conversational robot to generate responses. In this case, the implicit feedback comes from the human generated utterances. In this paper, we only focus on the first way of using the implicit feedback and leave the use of the second way in future work.

Reinforcement Learning for Conversation Model

The reinforcement learning approach has been widely used in conversation or dialogue systems (Walker 2000; Young

[1] Here, a simulated conversation is generated by chatting of two conversational robots.

et al. 2010; Gasic et al. 2014; Su et al. 2016b; 2016a; Wen et al. 2016b). Generally, a reinforcement learning based conversation model can be represented as a four-tuple $< s, a, r, p >$, where s, a, r and p denote the *state, action, reward* and *policy*, respectively.

In the proposed conversation model, the learning process is through the simulation of two conversational robots. Let A and B denote the two conversational robots, respectively. The conversation process is carried on by the simulated chatting between the two robots. Therefore, a conversation can be represented as an alternate sequence of utterances that generated by the two robots as "$u_{a_1}, u_{b_1}, u_{a_2}, u_{b_2}, ..., u_{a_i}, u_{b_i}, ..., u_{a_n}, u_{b_n}$".

State Taking an utterance u_{b_i} as an example, the state should be denoted by all the previous utterances $u_{a_1}, u_{b_1}, u_{a_2}, u_{b_2}, ..., u_{a_i}$. However, considering the computational complexity on modeling the long-term conversation history, the same to (Li et al. 2016c), we only utilize the previous two utterances $[u_{a_i}, u_{b_i}]$ to denote the current state. The state is then represented by a dense vector by encoding the concatenation of u_{a_i} and u_{b_i}. Similar to (Li et al. 2016a; 2016c; Su et al. 2016b), the concatenation of u_{a_i} and u_{b_i} is fed to an RNN-LSTM model to generate the representation of conversation states.

Action For a given conversation, an action a is an utterance to be generated. In open domain conversations, the action space is infinite. For example, given an input message "*How is going?*", the responses could be "*Not bad.*", "*It's okay.*", "*Pretty good*", etc. The conversation generation cannot be seen as a supervised learning process as the reward of an action can not be immediately obtained until the end of the conversation.

Reward The reward r indicates the contribution of an action a to the success of a conversation. The reinforcement learning process is to iteratively estimate and maximize the expectation of the future rewards given the current state and action. Li et al. 2016c proposed an approach to approximate the reward function $r_{obj}(a, [u_{a_i}, u_{b_i}])$ (Objective reward) by linearly combining 3 objective factors that avoid to generate dull, repetitive and non-coherent utterances in conversation generation. In this paper, besides the implicit feedback reward, we also consider the objective reward function to our proposed conversation model.

Policy For a reinforcement learning based conversation model, the policy is usually a joint distribution of action and state. The policy can also be seen as a function that input the current state (conversation history) and output an action (response) with its probability. Therefore, the crucial part of learning to conversation is the policy learning. In this paper, we use $p_{RL}(a|u_{a_i}, u_{b_i})$ to denote the conversation policy that is learnt by a reinforcement learning approach. It is worth noting that the object functions of $p_{RL}(a|u_{a_i}, u_{b_i})$ and $p_{seq2seq}(a|u_{a_i}, u_{b_i})$ are different. The former is based on a reinforcement learning function while the latter is based on a cross-entropy function.

Implicit Feedback based Conversation Model

In this section, we will present how the implicit feedback are integrated into the proposed reinforcement learning based conversation model.

A New Reward Function with Implicit Feedback As described in Section , we consider 3 implicit feedback for conversation generation. They are (1)**stance**, (2)**sentiment** and (3)**stalemate** in conversation. We will detail the rewards that are defined by the explored implicit feedback.

Stance Reward We take the stance identification as a binary classification task. For a given utterance u, $f_1(u) \in \{0, 1\}$, where 0 and 1 denote *negative* and *positive* respectively. The function f_1 draws on the definition of binary stance classification function by using (Teng, Vo, and Zhang 2016). We then define the stance reward function as following:

$$r_1 = f_1(a)[\log p_{seq2seq}(a|u_{a_i}, u_{b_i}) + \log p'_{seq2seq}(u_{b_i}|a)] \tag{1}$$

Sentiment Reward The sentiment of a given utterance u is calculated as the following function.

$$f_2(u) = \begin{cases} -2 & \text{if} \quad f_2(u) \leq -1.5 \\ -1 & \text{if} \quad -1.5 < f_2(u) \leq -0.5 \\ 0 & \text{if} \quad -0.5 < f_2(u) \leq 0.5 \\ 1 & \text{if} \quad 0.5 < f_2(u) \leq 1.5 \\ 2 & \text{if} \quad f_2(u) \geq 1.5 \end{cases} \tag{2}$$

Where, each utterance u is assigned a sentiment label from -2 to 2, which denote *very negative*, *negative*, *neutral*, *positive* and *very positive*, respectively. The sentiment labels are generated by the function $f_2(u)$ proposed by (Teng, Vo, and Zhang 2016) and the reward function is defined as:

$$r_2 = f_2(a)[\log p_{seq2seq}(a|u_{a_i}, u_{b_i}) + \log p'_{seq2seq}(u_{b_i}|a)] \tag{3}$$

The sentiment reward is proposed to verify the impact of the sentiments on generating utterances.

Stalemate Reward As the sampled conversations shown in Table 1 and 2, the conversation may come to a stalemate state due to the repetitive turns on simulation process and the perfunctory responses during the conversations.

Given an utterance u, $f_3(u) \in [0, 1]$ denotes the likelihood of u of a conversation to come to a stalemate state. We re-implement the stalemate detection approach (Li et al. 2016d) to calculate the $f_3(u)$. We thus proposed a reward function that considers the stalemate state in conversations.

$$r_3 = -\log f_3(a) - \frac{1}{N_{\mathbb{V}}} \sum_{t \in \mathbb{V}} \frac{\mathbb{1}_t}{N_a} \tag{4}$$

Where, $\mathbb{V}$ represents a stalemate vocabulary[2]. $N_{\mathbb{V}}$ and N_a denote the number of tokens of $\mathbb{V}$ and the action a, respectively. $\mathbb{1}_t$ is an indicator that equals to 1 if t is a token of a, otherwise, equals to 0.

[2] We manually constructed the stalemate vocabulary, which contains 12 words that indicate the conversation may come to a stalemate state.

As shown in Equation (4), the stalemate reward consists of two parts. The left part is estimated by a learning based approach and the right part is calculated by a heuristic approach which depends on a manually collected stalemate vocabulary. Both of the two parts are to penalize the actions that may lead a conversation to a stalemate state.

The final implicit feedback based reward of an action a, given the state $[u_{a_i}, u_{b_i}]$ is as follows:

$$r_{imp}(a, [u_{a_i}, u_{b_i}]) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3 \tag{5}$$

Reward Function In this paper, we take the $r_{obj}(a, [u_{a_i}, u_{b_i}])$ as the base reward. The proposed implicit feedback reward $r_{imp}(a, [u_{a_i}, u_{b_i}])$ is combined with the base reward by linear interpolation. The final reward of an action a, given the state $s_i = [u_{a_i}, u_{b_i}]$ is obtained as:

$$r(a, s_i) = \delta r_{obj}(a, s_i) + (1 - \delta) r_{imp}(a, s_i) \tag{6}$$

Conversation Simulation For the learning of the conversation policy $p_{RL}(a|u_{a_i}, u_{b_i})$, we utilize the strategy of conversation simulation between two conversational robots to explore the state-action space. It is worth noting that the simulation process consists of two steps. The first step is to generate an utterance by a conversational robot. The second step is to learn the generation of simulation turn by turn.

Utterance Generation For utterance generation, we utilize the Seq2Seq model (Vinyals and Le 2015) to initialize the conversation policy $p_{seq2seq}$. Inspired by (Li et al. 2016c; Ranzato et al. 2015), we use policy gradient (Williams 1992; Sutton et al. 2000) for optimization. Given the current state $[u_{a_i}, u_{b_i}]$, the model generates a set of candidate actions $A = \{\hat{a}|\hat{a} \sim p_{RL}\}$, p_{RL} is initialized by $p_{seq2seq}$. The expected reward for a candidate action $\hat{a}$ as $\mathbb{E}[p_{RL}(\hat{a}|u_{a_i}, u_{b_i})]$.

We use the stochastic gradient descent approach to update the parameters of the Seq2Seq model. For training the Seq2Seq model, we use the curriculum learning approach (Bengio et al. 2009) to generate each utterance. Concretely, for a given utterance whose length is U, we use the cross-entropy loss to generate the first R tokens. The rest $U - R$ tokens are then generated by using the reinforcement learning algorithm. R is gradually (batch to batch) annealing to zero during the training process. Following (Zaremba and Sutskever 2016; Li et al. 2016c), we also use an additional neural network which concatenates the representations of the initial source and the generated target of the Seq2Seq model as input and output a score q (Reward baseline), which is used to decrease the learning variance.

Simulation The process of simulation between two conversational robots is run in two steps. 1) given an initial utterance from the training data to robot A as input. 2) robot A generates an output utterance and then feeds as the input to robot B. The simulation is to repeat the above steps until a conversation reaches an end. The same to (Li et al. 2016c), we use a simple rule matching method, with a list of 8 phrases that count as dull responses. Once a dull response is generated, the conversation is ended. During the simulation, we need to learn the conversation policy p_{RL} for each action a given the state $[u_{a_i}, u_{b_i}]$. The policy gradient approach is then used to optimize the expected reward. Please refer to (Zaremba and Sutskever 2016) for more details about the derivation of the objective function and the gradient update.

Similar to (Li et al. 2016c), we also use a curriculum learning for training the conversation generation model. We first generate 2-turn simulations and then gradually increase to generate 5-turn simulations. For each turn of a simulation, 5 candidate utterances are generated. As the proceeding of a simulation, the number of candidate utterances grows exponentially. Considering the computational complexity, each simulation is only carried out for 5 turns at most in the training phase.

Experiments and Results

Experimental Data

In this paper, we empirically compare the performance of the proposed approach and the baselines on two datasets. The first is the OpenSubtitles dataset (Tiedemann 2009), which is also used in (Vinyals and Le 2015; Li et al. 2016c). The second dataset is a Twitter conversation corpus[3] that contains 754,530 messages. 44 million and 376,265 conversation pairs from the two datasets are used for training the Seq2Seq models. 0.8 million and 30,000 of extracted messages that have the lowest likelihood of generating dull responses are respectively used for initializing the conversation simulations for the policy learning.

Parameter Setting

The training epochs are equals to 50 and 124 on the OpenSubtitles and Twitter datasets, respectively. The batch size is set to 128. The maximum length of an input sequence is set to 60 words. The number of hidden state of Seq2Seq model equals to 128. The size of the vocabulary for Seq2Seq model is set to 100,000. The beam size is set to 10 for decoding. For the training and test of the simulation, the numbers of the simulated conversation turns are set to 5 and 8, respectively. The λ_1, λ_2 and λ_3 in Equation (5) equals to 0.4,0.4 and 0.2. δ in Equation (6) equals to 0.5.

Evaluation

Automatically evaluating a conversation model is still an open problem. The BLEU socre (Papineni et al. 2002) is widely used for evaluating machine translation systems. However, it is not a suitable evaluation metric for conversation generation, as the compatible responses to the same utterance may share less common words. Moreover, it is also hard to construct a reference set with adequate coverage. The perplexity used to evaluate the quality of language modelling, is also not suitable to evaluate the relevance of utterances in conversation (Shang, Lu, and Li 2015; Li et al. 2016c).

To address the above issues, we propose two evaluation measures, namely automatic evaluation and human evaluation. Each evaluation measure includes several metrics. For the empirical comparisons, two **baselines** are chosen.

[3] https://github.com/Marsan-Ma/chat_corpus

Model	# of simulated turns	
	OpenSubtitles	Twitter
Seq2Seq	2.39	5.82
RL-Seq2Seq	4.38	7.32
Ours	**4.64**	**7.79**

Table 3: The comparison of the average number of simulated turns (conversation length) between the baselines and our proposed approach.

- The Seq2Seq model proposed by Vinyals and Le 2015 that used the sequence to sequence learning model for conversation generation.
- The state-of-the-art conversation generation model (RL-Seq2Seq) (Li et al. 2016c) based on deep reinforcement learning and sequence to sequence learning models.

Automatic Evaluation Inspired by the simulation strategy for training the conversation policy. We again adopt the simulation on test. 1,000 and 200 messages that are not used for training, are randomly sampled from OpenSubtitles and Twitter datasets for initializing the test simulation, respectively. The first automatic evaluation metric is the average number of simulated turns generated by the three conversation models. The reason of choosing this metric for evaluation is that resulting in the observations of the training data, we intuitively assume that better conversation models lead to generate more simulated turns without repetitive or dull responses. Table 3 shows the experimental results of the average number of simulated turns of the baselines and our proposed approach. As can be seen, due to the reward combination of the implicit feedback, our proposed approach gains over the baselines on sustaining the simulated conversations.

The second metric for automatic evaluation is the diversity of generated conversations. The diversity score for a conversational robot equals to the number of distinct tokens in its generated utterances divided to the total number of generated tokens. Here, we calculate the diversity score of the distinct unigram and bigram in generated utterances for each model. Experimental results are shown in Table 4. We can see that the proposed approach generates more number of distinct unigrams and bigrams than the baselines in both OpenSubtitles and Twitter datasets. It illustrates that the proposed approach can generate more diverse responses by introducing the stance and sentiment factors as well as avoiding the stalemate in conversations.

To verify the performance of the proposed implicit feedback based reward, we also compare the experimental results of different reward functions. As shown in Table 5, we find that the implicit feedback reward r_{imp} can enhance the performance of the model that only uses the objective reward. Meanwhile, we can also conclude by comparing the performance of r_{obj} and r_{imp} that the r_{obj} based model is adept in increasing simulated turns while the r_{imp} is good at generating longer and more diverse utterances.

Human Evaluation We introduce two settings of human evaluation, namely offline human judgement and online realtime human-robot conversation. The **offline human judgement** includes 2 metrics: "single-turn general quality (S-Q)" and "single-turn ease to respond (E2R)", which are also adopted by Li et al. 2016c. Here, given an input to the two compared models, 3 judges are asked to decide which one of the two outputs is better for the given input(S-Q) and which of the two outputs is easier to respond by human (E2R). We randomly sample 150 messages from each dataset as the inputs of the RL-Seq2Seq and our proposed model. We thus obtain 600 single turn conversations generated by the two compared models for S-Q and E2R evaluations. Ties are permitted. Identical responses to a same input are given the same score. When inconsistencies occur, the final judgements are generated by voting. Specially, when the 3 judges provide totally different judgements, another judge is involved in the judgement. Table 6 shows the human judgement results between our proposed approach and RL-Seq2Seq approach on the above 2 metrics. As can be seen from Table 6, due to the integration of implicit feedback, our proposed approach tends to generate the responses that avoid to make the conversation come to a stalemate and thus easier to respond. Meanwhile, the high tie ratios are also in our expectation and easy to understand as the base model of response generation is Seq2Seq (Vinyals and Le 2015), which is the same to our proposed model and RL-Seq2Seq. It thus leads to that the performance of most single turn responses is similar to the two compared models.

Model	OpenSubtitles		Twitter	
	unigram	bigram	unigram	bigram
Seq2Seq	0.0027	0.0013	0.0157	0.0211
RL-Seq2Seq	0.0034	0.0038	0.0212	0.0301
Ours	**0.0042**	**0.0045**	**0.0219**	**0.0316**

Table 4: The diversity of the generated conversations of the baselines and our proposed approach. Diversity score equals to the number of distinct unigrams and bigrams in the generated utterances divided to the total number of generated tokens, respectively.

We developed a human-robot conversation platform for the **online realtime human-robot conversation**. In this setting, 3 judges are talking to two anonymous conversational robots (One is based on our proposed model, the other is based on the RL-Seq2Seq model), respectively. The judges can decide when to end a conversation if they are not willing to continue the current conversation. Each of the 3 judges is asked to finish 25 online realtime conversations for each robot. For each online conversation, the starting sentence for the two conversational robots is the same. The chat content is controlled to be chit-chat. The content of question answering, task-oriented dialogue and information recommendation is not permitted in the online conversations. We thus totally collected 150 conversations for the two conversational robots. After the ending of each conversation, the judges are asked to give a feedback on 2 metrics, namely user satisfaction (S) and conversation fluency (F). The user satisfaction is ranged from $0 \sim 4$, which denote *very dissatisfied*, *dissatisfied*, *neutral*, *satisfied* and *very satisfied*, respectively. The

Reward	OpenSubtitles				Twitter			
	# of turns	unigram	bigram	$Avg(N_t(u))$	# of turns	unigram	bigram	$Avg(N_t(u))$
r_{obj}	4.38	0.0034	0.0038	4.41	7.32	0.0212	0.0301	4.9
r_{imp}	3.71	0.0037	0.004	4.64	6.55	0.0216	0.0308	5.15
$r_{obj} + r_{imp}$	**4.64**	**0.0042**	**0.0045**	**5.38**	**7.79**	**0.0219**	**0.0316**	**6.08**

Table 5: The average number of simulated turns and the diversity of the generated conversations of the reinforcement learning model with different reward functions.$Avg(N_t(u))$ denotes the average number of tokens in the generated utterances.

Dataset	Metric	Our-win	Our-lose	Tie
OpenSubtitles	S-Q	0.153	0.12	0.726
	E2R	0.147	0.1	0.753
Twitter	S-Q	0.16	0.073	0.767
	E2R	0.147	0.073	0.78

Table 6: Human judgement results between our proposed approach and RL-Seq2Seq approach on the 2 metrics.

Model	S	F	# of turns	$\mathbf{Avg(N_t(u))}$
RL-Seq2Seq	1.974	1.26	8.84	2.92
Ours	**2.304**	**1.532**	**11.53**	**3.42**

Table 7: The average scores of the user satisfaction, conversation fluency, the number of conversation turns and the average number of generated tokens $\mathbf{Avg(N_t(u))}$ of our proposed approach and RL-Seq2Seq approach.

conversation fluency measures the coherence of the conversation. It is ranged from 0 to 2 denoting *disfluency*, *neutral* and *fluency*, respectively. Table 7 shows the human evaluation results of our proposed approach and RL-Seq2Seq approach on 4 metrics. We can see that our approach outperforms the RL-Seq2Seq model in both user satisfaction and conversation fluency. To see the average number of turns of conversation simulation, it denotes that the users are more willing to talk with our proposed conversation model due to its ability on modeling the human stance, sentiment, emotion, etc, as well as breaking the stalemate in conversations. The satisfaction and fluency scores also illustrates that our proposed model are more capable of maintaining the conversation topics and generate coherent conversations than that of the RL-Seq2Seq model. It also verifies that the long-term goal of conversation generation can be better optimized by exploring the implicit feedback. We can also see that the average number of tokens in the generated utterances is different between the proposed conversation model and the RL-Seq2Seq model. It indicates that the implicit feedback reward tends to increase the length of the generated utterances in conversation. Meanwhile, longer utterances may provide more information in conversations. That may be helpful to sustain the conversations.

Related Work

The conversation generation research mainly focuses on two categories. First is the open domain conversation generation. Ritter, Cherry, and Dolan 2010 proposed an unsupervised approach to model dialogue response by clustering the raw utterances. They then presented an end-to-end dialogue response generator by using a phrase-based statistical machine translation model (Ritter, Cherry, and Dolan 2011). Banchs and Li 2012 introduced a search-based system, named IRIS, to generate dialogues using vector space model and then released the experimental corpus for research and development (Banchs 2012). Recently, benefit from the advantages of the sequence to sequence learning framework with neural networks (Sutskever et al. 2014) and Shang, Lu, and Li 2015 had drawn inspiration from the neural machine translation (Bahdanau, Cho, and Bengio 2014) and proposed a RNN encoder-decoder based approach to generate dialogue by considering the last one sentence and a larger range of context respectively. Serban et al. 2016a preseneted a hierachical recurrent encoder-decoder (HRED) approach to encode each utterance and recurrently model the dialogue context to generate context dependent responses. Serban et al. 2016c further introduced a stochastic latent variable at each dialogue turn to improve the ambiguity and uncertainty of the HRED model for dialogue generation. Serban et al. 2016b proposed a parallel stochastic generation framework which first generates a coarse sequence and then generate an utterance conditioned on the coarse sequence. To address the problems of generating generic and repetitive response of the RNN encoder-decoder framework, Li et al. 2016c proposed a deep reinforcement learning approach to either generate meaningful and diverse response or increase the length of the generated dialogues. Xing et al. 2017 proposed a hierarchical recurrent attention network (HRAN) to jointly model the importance of tokens in utterances and the utterances in context for context-aware response generation. Dhingra et al. 2016 presented an end-to-end dialogue system for information accquisition from knowledge base by using reinforcement learning. Asghar et al. 2016 proposed an active learning approach to learn user explicit feedback online and combine the offline supervised learning for response generation of conversational agents.

Second is task-oriented dialogue generation. Previous research on task-oriented dialogue generation usually employed handcrafted generator to define the generation decision space with the handcrafted features or statistical models (Langkilde and Knight 2002; Walker, Rambow, and Rogati 2002; Paiva and Evans 2005; Isard, Brockmann, and Oberlander 2006; Mairesse and Walker 2008; Rieser and Lemon 2009). These approaches have the limitation on scaling to new domains. Mairesse et al. 2010 proposed a statistical language generator which used a dynamic

Bayesian networks to generate dialogue response. Mairesse and Young 2014 learned to generate paraphrases in dialogue through a factored language model that was training from the data collected by crowdsourcing. Both of them completely learn from the data and thus has no limitation on domain transfer. Recently, as the powerful of deep neural network on learning from large-scale data, Wen et al. 2015a proposed a statistical dialogue generator based on a joint recurrent and convolutional neural network, which can directly learn from the data without any semantic alignment or handcrafted rules. Further, Wen et al. 2015b proposed a semantically conditioned LSTM to generate dialogue response and then compared it with an RNN encoder-decoder generator on multi-domain data to verify the ability of domain adaptation of the two generators (Wen et al. 2016a) . Bordes and Weston 2016 utilized an end-to-end memory network to model the context information and generate task-oriented response for dialogue system.

Conclusion and Future Work

In this paper, we explore the implicit feedback in the conversation to optimized the long-term goal of conversation generation. Based on the reinforcement learning framework, we proposed a new reward function that integrate the implicit feedback for conversation generation. A simulation strategy is utilized to explore the action-state space for training and test the conversation model. Experimental results show that the proposed approach outperforms the state-of-the-art approach in both automatic and human evaluations on the OpenSubtitles and Twitter datasets.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their insightful comments. Meanwhile, the authors would like to thank Prof. Yue Zhang (Singapore University of Technology and Design) and Mr. Yiming Cui (iFLYTEK Research) for their help on revising the manuscript. The paper is supported by the 973 Program (No. 2014CB340503) and NSFC (No. 61502120, 61472105). The paper is also supported by the HIT Tencent Joint Lab. The technical collaborators in Tencent Corp. are Mr. Lianqiang Zhou, Mr. Yue Zhang and Mr. Wenzhi Xu.

References

Asghar, N.; Poupart, P.; Xin, J.; and Li, H. 2016. Online sequence-to-sequence reinforcement learning for open-domain conversational agents. *arXiv:1612.03929v3*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Banchs, R. E., and Li, H. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *ACL*, 37–42.

Banchs, R. E. 2012. Movie-dic: a movie dialogue corpus for research and development. In *ACL*, 203–207.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *International Conference on Machine Learning*, 41–48.

Bordes, A., and Weston, J. 2016. Learning end-to-end goal-oriented dialog. *CoRR* abs/1605.07683.

Dai, A. M., and Le, Q. V. 2015. Semi-supervised sequence learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, 3079–3087.

Dhingra, B.; Li, L.; Li, X.; Gao, J.; Chen, Y.; Ahmed, F.; and Deng, L. 2016. End-to-end reinforcement learning of dialogue agents for information access. *CoRR* abs/1609.00777.

Gasic, M.; Kim, D.; Tsiakoulis, P.; Breslin, C.; Henderson, M.; Szummer, M.; Thomson, B.; and Young, S. 2014. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains.

Isard, A.; Brockmann, C.; and Oberlander, J. 2006. Individuality and alignment in generated dialogues. In *INLG*, 25–32.

Langkilde, I., and Knight, K. 2002. Generation that exploits corpus-based statistical knowledge. *Programmierte Personalpolitik in Kreditinstituten*.

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.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003.

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

Li, X.; Mou, L.; Yan, R.; and Zhang, M. 2016d. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *International Joint Conference on Artificial Intelligence*.

Mairesse, F., and Walker, M. A. 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In *ACL*, 165–173.

Mairesse, F., and Young, S. 2014. Stochastic language generation in dialogue using factored language models. *Computational Linguistics* 40(4):763–799.

Mairesse, F.; , M.; Jurcicek; Ek, F.; Keizer, S.; Thomson, B.; Yu, K.; and Young, S. 2010. Phrase-based statistical language generation using graphical models and active learning. In *ACL*, 1552–1561.

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3349–3358.

Paiva, D. S., and Evans, R. 2005. Empirically-based control of natural language generation. In *ACL*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence level training with recurrent neural networks. *Computer Science*.

Rieser, V., and Lemon, O. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. In *EACL*, 105–120.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *NAACL*, 172–180.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *EMNLP*, 583–593.

Serban, I.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models.

Serban, I. V.; Klinger, T.; Tesauro, G.; Talamadupula, K.; Zhou, B.; Bengio, Y.; and Courville, A. 2016b. Multiresolution recurrent neural networks: An application to dialogue response generation. *arXiv preprint arXiv:1606.00776*.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2016c. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

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 and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1577–1586.

Su, P. H.; Gasic, M.; Mrksic, N.; Rojasbarahona, L.; Ultes, S.; Vandyke, D.; Wen, T. H.; and Young, S. 2016a. Continuously learning neural dialogue management.

Su, P. H.; Gasic, M.; Mrksic, N.; Rojasbarahona, L.; Ultes, S.; Vandyke, D.; Wen, T. H.; and Young, S. 2016b. Online active reward learning for policy optimisation in spoken dialogue systems. In *Meeting of the Association for Computational Linguistics*, 2431–2441.

Sutskever, I.; Vinyals, O.; Le, Q. V.; Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. *NIPS* 4:3104–3112.

Sutton, R. S.; Mcallester, D.; Singh, S.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems* 12:1057–1063.

Teng, Z.; Vo, D. T.; and Zhang, Y. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1629–1638.

Tiedemann, J. 2009. *News from OPUSłA Collection of Multilingual Parallel Corpora with Tools and Interfaces*.

Vinyals, O., and Le, Q. 2015. A neural conversational model. *Computer Science*.

Vougiouklis, P.; Hare, J.; and Simperl, E. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3370–3380.

Walker, M. A.; Rambow, O. C.; and Rogati, M. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech & Language* 16(3C4):409–433.

Walker, M. A. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research* 12(1):387–416.

Wen, T. H.; Gasic, M.; Kim, D.; Mrksic, N.; Su, P. H.; Vandyke, D.; and Young, S. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *Computer Science*.

Wen, T. H.; Gasic, M.; Mrksic, N.; Su, P. H.; Vandyke, D.; and Young, S. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *Computer Science*.

Wen, T.-H.; Gašić, M.; Mrkšić, N.; Rojas-Barahona, L. M.; Su, P.-H.; Vandyke, D.; and Young, S. 2016a. Multi-domain neural network language generation for spoken dialogue systems. In *NAACL*, 120–129.

Wen, T. H.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojasbarahona, L. M.; Su, P. H.; Ultes, S.; and Young, S. 2016b. A network-based end-to-end trainable task-oriented dialogue system.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8(3):229–256.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. 2016a. Topic augmented neural response generation with a joint attention mechanism. *CoRR* abs/1606.08340.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. Y. 2016b. Topic aware neural response generation.

Xing, C.; Wu, W.; Wu, Y.; Zhou, M.; Huang, Y.; and Ma, W. 2017. Hierarchical recurrent attention network for response generation. *CoRR* abs/1701.07149.

Yao, K.; Zweig, G.; and Peng, B. 2015. Attention with intention for a neural network conversation model. *Computer Science*.

Young, S.; Gasic, M.; Keizer, S.; Mairesse, F.; Schatzmann, J.; Thomson, B.; and Yu, K. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language* 24(2):150–174.

Zaremba, W., and Sutskever, I. 2016. Reinforcement learning neural turing machines - revised. *Computer Science* (July).