Hierarchy Response Learning for Neural Conversation Generation

Bo Zhang[1] and **Xiaoming Zhang**[1,2]
[1] School of Cyber Science and Technology
[2] Hefei Innovation Research Institute
Beihang University, China
{zhangnet,yolixs}@buaa.edu.cn

Abstract

The neural encoder-decoder models have shown great promise in neural conversation generation. However, they cannot perceive and express the intention effectively, and hence often generate dull and generic responses. Unlike past work that has focused on diversifying the output at word-level or discourse-level with a flat model to alleviate this problem, we propose a hierarchical generation model to capture the different levels of diversity using the conditional variational autoencoders. Specifically, a hierarchical response generation (HRG) framework is proposed to capture the conversation intention in a natural and coherent way. It has two modules, namely, an expression reconstruction model to capture the hierarchical correlation between expression and intention, and an expression attention model to effectively combine the expressions with contents. Finally, the training procedure of HRG is improved by introducing reconstruction loss. Experiment results show that our model can generate the responses with more appropriate content and expression.

1 Introduction

Neural conversation generation (Xu et al., 2017; Zhang et al., 2018), focusing on responding to humans intelligently on a variety of topics, has drawn great attention from both academic and industry. The sequence-to-sequence model (Seq2Seq) (Sutskever et al., 2014) is one type of neural generation model that maximizes the probability of generating a response given the dialogue context. It enables the incorporation of rich context to generate coherent responses in an unsupervised manner. However, it was found that Seq2Seq models suffer from so-called safe response problem (Xu et al., 2017) , i.e., they tend to generate some dull and generic repetitive responses (e.g., *"I think so"*, *"I don't know'*, etc.), rather than meaningful and conscious expression. Xu et al. (2017) ascribed this to the fundamental nature of statistical models since the distribution of most pieces of information are relatively sparser when compared to the safe response patterns in the open domain conversations. Some works attempted to improve the architecture of Seq2Seq models, including introducing reinforcement learning (Zhang et al., 2018), encouraging responses that have long-term payoff, etc. The other important reason is that the response generation model cannot express the intention and emotion internally. Thus, one line of research has focused on forcing the model to simulate some human's skills by augmenting the input with rich meta information. For example, some recent works biased the responses to some specific personas (Li et al., 2016b) or emotions (Huber et al., 2018).

Usually, in the process of human conversation, a speaker participates in the dialogue including the following steps. The speaker is firstly required to decide what the intention is to reflect the inner feelings or opinions. In the speaker's knowledge base, there may be varieties of appropriate expressions that can be found to represent his current intention. Therefore, a meaningful response can be produced by choosing one of the expressions and filling it with relevant content. For example, if a man wants to ask the way to the Park, he first needs to select an appropriate expression from a cluster of the expressions, e.g., *"Where is the ... ?"*, *"How do I get to the ... ?"* and *"Is the ... far from here?"*, and then replaces "..." as the destination, i.e., the Park. As a crucial feature in natural conversation, **dialog acts** (Poesio and Traum, 1998) have been widely used in the dialogue managers to represent the intentions. Existing works introduce dialog acts to label a cluster of responses and a latent variable is learned to select a dialog act for response generation (Zhao et al., 2017; Serban

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1772–1781, Hong Kong, China, November 3–7, 2019.

et al., 2017a). However, it is not effective to capture the output diversity since the natural correlation between the expression patterns and dialog acts is not learnt. Intuitively, another latent level can be introduced to generate different expressions from the same dialog act, and a hierarchical structure can be used to model the response generation process. That is, the knowledge base is first constructed over the pairs of expressions and dialog acts to capture the latent correlation between them. Then, varieties of expressions can be selected from the knowledge base to be filled with the response content based on the latent correlation.

To learn the hierarchical model, it is quite challenging in large-scale conversation generation due to the following reasons. **First**, the semantic world is populated with a vast number of expressions, each of which corresponds to a specified label that reflects a kind of dialog act. Obtaining high-quality expression-act data is impractical particularly in open domain conversations. **Second**, it is difficult to incorporate expression and content into the generation model in a nature and coherent way because they have different semantic representation patterns. **Last**, this process cannot be efficiently optimized using stochastic gradient descent (SGD) akin to backpropagation on feedforward neural networks.

To tackle the challenges, we propose to take advantage of the hierarchical nature of response generation. In particular, we investigate: (1) how to automatically learn a hierarchical model to naturally capture the response generation process; (2) how to adaptively learn and adjust the influence ratio between expression and content. Our solutions to these questions result in a new architecture for neural response generation. In particular, a novel hierarchical response generation (HRG) framework is proposed to effectively capture the process of response generation. An expression reconstruction model with a two-level probability structure is introduced to randomly generate the expressions, and an expression attention model is proposed to effectively fill the expressions with content. Finally, an efficient training method is proposed to learn the model within the framework of conditional variational autoencoders (CVAE) (Doersch, 2016). The main contributions are outlined as:

- We propose to investigate the problem of generating variety and meaningful responses by imitating the human response process with a hierarchical response generation model.
- We propose an end-to-end framework to incorporate the expression and response content into the dialog generation. Our model is interpretable and even controllable compared to traditional generation model.
- We empirically demonstrate that our approach can generate responses with better expressions and content than traditional generation model.

2 Problem Statement

Our problem is formulated as follows: Given a dialog context C and a dialog act a of the response to be generated, the goal is to generate a response $y = (y_1, y_2, ..., y_n)$ that is coherent with the dialog act a. Essentially, the model estimates the probability: $P(y|C, a) = \prod_t P(y_t|y_{<t}, C, a)$. A simple implementation is to directly embed the act information into the Seq2Seq model. However, as shown in our experiments, it still suffers from safe response problem.

In this paper, we propose a novel hierarchical model to imitate the human thinking process in the conversation generation. The hierarchical generation process is: (1) for each dialog act a, a set $\Omega(a)$ is constructed to contain all the corresponding expressions of a; (2) to generate an expression, a dialog act a is first selected, and then an appropriate expression e is also selected from $\Omega(a)$; (3) a response is obtained by filling the expression e with relevant content according to the dialog context C. This hierarchical model allows us to express the responses with diverse expression templates of the same dialog act by drawing different samples from $\Omega(a)$. However, in addition to the difficulty of constructing high-quality set $\Omega(a)$, it is also needed to maximize the probability of each y in the training set with the objective: $P(y|C, a) = \int_{e \in \Omega(a)} P(y|C, e)\, de$, which is also difficult to compute by the numeric methods.

In our approach, the expression e is modeled as a conditional distribution over the dialog act a, i.e., $p_\theta(e|a)$. The response is then generated by feeding the expression e obtained based on $p_\theta(e|a)$ into the model, i.e., $P(y|C, e), e \sim p_\theta(e|a)$. Now, the training objective is simplified as follows:

$$P(y|C, a) = \int P(y|C, e)\, p_\theta(e|a)\, de. \quad (1)$$

This objective can be transformed as the variational lower bound of CVAE (Doersch, 2016), and thus can be optimized efficiently. Specifically, the variational approximation $q_\phi(e|y)$ is constructed to approximate the intractable posterior $p_\theta(e|a)$. Assuming that the meaning of expression e is independent of C, we train the model by maximizing the variational lower bound,

$$\begin{aligned} L(\theta, \phi; y, C, a) &= -KL(q_\phi(e|y)||p_\theta(e|a)) \\ &+ \mathbb{E}_{e\sim q_\phi(e|y)}[\log p_\theta(y|C, e, a)] \\ &\leq \log p(y|C, a) \end{aligned} \quad (2)$$

where $KL(\cdot||\cdot)$ is the Kullback-Leibler divergence to measure the distance between two distributions.

Note that in our problem statement, we assume that the dialog act of the to-be-generated response is given in advance, rather than predicted depending on the context. Many existing researches (Sacks et al., 1978; Young et al., 2013; Daniel Jurafsky, 2017) have explored dialog act interactions with dialog system and proposed some methods to decide the most appropriate dialog act for the response. In this paper, we only focus on response generation. During the testing process, we simply specify a dialog act to the model. We leave this study to our future work.

3 Hierarchical Response Generation

Building upon the encoder-decoder models (e.g., Seq2Seq, HRED(Serban et al., 2016)), a Hierarchical Response Generation (HRG) framework is proposed to effectively generate more diverse expressions for conversation generation. As shown in Figure 1, HRG contains two main modules: Expression Reconstruction model and Expression Attention model. A training method is proposed to learn the hierarchical HRG model in Section 3.3.

3.1 Expression Reconstruction

To maximize the objective Eq.(2), we are first required to model the networks $q_\phi(e|y)$ and $p_\theta(e|a)$. The task of network $q_\phi(e|y)$ is to capture the expression representations from the responses while $p_\theta(e|a)$ to sample an expression representation from the distributions associated with the specified dialog acts. As shown in Figure 2, in the framework of CVAE, the response is first encoded as a latent variable, and then a decoder is introduced to reconstruct the response. But this generation process is not interpretable and controllable. Thus, we

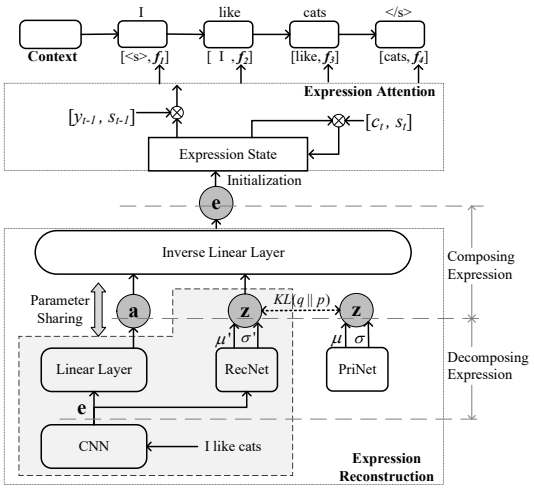

Figure 1: The overview of Hierarchical Response Generation (HRG). The components in the dashed box are removed during testing. The dialog act a is specified manually and the random variable z is generated by PriNet during testing.

propose a novel method to reconstruct the expression with multiple latent variables by establishing links between $q_\phi(e|y)$ and $p_\theta(e|a)$.

Modeling $q_\phi(e|y)$. It was found that the convolutional layer can extract the common patterns within the local regions of the input utterance (Kim, 2014). Therefore, the text convolutional network proposed in (Kim, 2014) is leveraged to mine the relationship between expression patterns and responses. Particularly, the network $q_\phi(e|y)$ consists of several convolutional filtering, local contrast normalization, and max-pooling layers, followed by several connected linear layers. Formally, given a response y, the expression representation can be described as

$$e = \text{CNN}(y). \quad (3)$$

In the experiments, we found that the convolutional layers can effectively extract the expression representation by discarding the content-related information.

Modeling $p_\theta(e|a)$. Given a dialog act a, the network $p_\theta(e|a)$ outputs an expression representation associated with a. To capture output diversity, it is also necessary to generate different expressions each time given the same dialog act. The classical models (e.g., linear layer) are not capable of representing this feature. Instead, the network $p_\theta(a|e)$ is easy to model by a linear layer, where

$$a = g(e) = \boldsymbol{W_f} \cdot e + \boldsymbol{b_f}. \quad (4)$$

Motivated by this observation, we propose an Inverse Linear Layer to model the network $p_\theta(e|a)$ where the dialog act a is mapped inversely into the expression representation e by solving g^{-1}.

Penrose Theorem (Ben-Lsrael and Greville, 1976) gives a general solution of equation $AX = B$ and proofs it. To solve g^{-1}, we simplify Penrose Theorem under an extra constraint and give a Corollary:

Corollary 1 *Let $A \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^m$. The general solution of the equation $Ax = b$ is*

$$x = A^+b + (I - A^+A)z \tag{5}$$

for arbitrary $z \in \mathbb{C}^n$ (if $0 < rank(A) < n$), where A^+ is Moore–Penrose inverse of A and I is an identity matrix.

The transformation of Ax inevitably leads to the information loss of x and thus the random vector z should be supplemented to the solution.

According to Corollary 1, the expression e can be represented by an inverse linear layer as

$$e = g^{-1}(a, z) = \boldsymbol{W_f}^+(a - \boldsymbol{b_f}) + (I - \boldsymbol{W_f}^+\boldsymbol{W_f})z. \tag{6}$$

The expression is uniquely determined by two independent variables, i.e., the dialog act a and the latent variable z. Given the same or similar dialog context, there may exist many valid expressions for the responses with the same dialog act a, each corresponding to a certain configuration of z. This representation allows us to express responses with diverse expression templates of specified dialog act by drawing samples from the learned distribution of z. The network $p_\theta(e|a)$ can be easily computed by setting $p_\theta(a, z) = p_\theta(a)p_\theta(z)$, where a and z are uncorrelated.

Reconstructing Expression. As shown in Figure 1, during training, the expression representation e is first captured from the ground-truth response by the network $q_\phi(e|y)$. The expression e is decomposed into multiple independent variables, i.e., a and z, and then these variables are composed to reconstruct the expression e by the network $p_\theta(e|a)$. These variables provide different discourse-level information to force the decoder to focus on multiple global information simultaneously. As shown in Figure 2, different from CVAE, some variables (i.e., a) in HRG are interpretable to make the model controllable while some (i.e., z) are continuous to reflect the latent feature. During testing, the model is controllable by specifying an appropriate dialog act to express the intention.

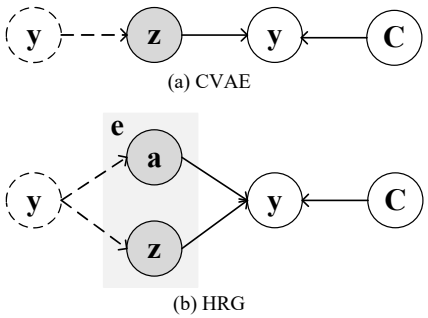

Figure 2: Graphical models for (a) CVAE and (b) HRG. Solid lines denote the generative models (a) $p(z)p(y|z, C)$ and (b) $p(e|a)p(y|e, C) = p(a)p(z)p(y|a, z, C)$; dashed lines denote the variational approximation (a) $q(z|y)$ and (b) $q(e|y) = q(a|y)q(z|y)$.

3.2 Expression Attention

There exist two main methods to incorporate the expression representation into the decoder. First, the concatenation of the context and expression representation e is used to initialize the recurrent of the decoder RNN with a nonlinear transformation. During the decoding, the decoder RNN decodes words based on the current state and previous word embeddings w. The second way is that the concatenation of fixed e and w is fed into decoder to update its state at each step. Formally, the first way updates its state only according to $[w, 0 \cdot e]$, while the second according to $[w, 1 \cdot e]$. However, these methods may cause that the content (expression) is so powerful that the responses are without any effective expression (meaningful content). In this paper, we propose an expression attention (EA) model to attend on different parts of expression representations each step by learning a vector $\alpha = \{\alpha_i \in (0, 1)\}$ adaptively. A balance between content and expression is effectively kept by feeding $[w, \alpha \cdot e]$ into decoder.

In particular, before decoding, an expression state is initialized as $q_0 = e$ to record the current expression representation. At step t, a strength gate β_t is computed based on the input of the previously decoded word y_{t-1} and the previous decoder state s_{t-1}. The expression state is weakened by a certain amount β_t at each step,

$$\beta_t = \text{Sigmoid}(\boldsymbol{W_s}[y_{t-1}, s_{t-1}]), \tag{7}$$

$$f_t = q_{t-1} \otimes \beta_t \tag{8}$$

where $\otimes$ is element-wise multiplication and $\text{Sigmoid}(x) = 1/(1 + exp(-x))$. The decoder updates its state conditioned on the previous token y_{t-1} and the current output f_t as follows:

$$s_t = \text{RNN}^d(s_{t-1}, [y_{t-1}, f_t]). \quad (9)$$

It is a dynamic process that the expression is adjusted adaptively according to the current environment and model behaviors compared to the two existing methods above. After step t, a self-update strategy is designed to update the expression state based on the context vector c_t (computed by Attention Mechanism (Luong et al., 2015)) and the current decoder state s_t. This process is formulated as

$$\gamma_t = \text{Sigmoid}(\boldsymbol{W_u}[c_t, s_t]), \quad (10)$$

$$q_t = q_{t-1} \otimes \gamma_t. \quad (11)$$

The expression representation is integrated into the decoder gradually until the expression state decays to zero through multiple iterations. The expression reconstruction and attention models respectively provide discourse-level and token-level randomness respectively, which can avoid the decoder generates the next token only depending on Neural Probabilistic Language Model instead of the dialog context and the current decoding state.

3.3 Reconstruction Training

The learnt parameters include the embeddings of vocabulary, and those in the encoder-decoder component and HRG. According to Section 3.1, we first identify two key assumptions that are essential: Both z and a are the indigenous properties of the expression e; The meaning of z is independent of the dialog act a. Based on them, we update the objective Eq.2 as

$$\begin{aligned} L(\theta, \phi; y, C, a) = & -KL(q_\phi(z|y)||p_\theta(z)) \\ & - KL(q_\phi(a|y)||p_\theta(a)) \\ & + \mathbb{E}_{z\sim q_\phi(z|y), a\sim q_\phi(a|y)}[\log p_\theta(y|C, a, z)]. \end{aligned} \quad (12)$$

Assuming that z follows isotropic Gaussian distribution, the prior network (PriNet) $p_\theta(z) = \mathcal{N}(\mu, \sigma^2\boldsymbol{I})$ and the recognition network (RecNet) $q_\phi(z|y) = \mathcal{N}(\mu', \sigma'^2\boldsymbol{I})$, where

$$\begin{bmatrix} \mu' \\ \log(\sigma'^2) \end{bmatrix} = \boldsymbol{W_r}\text{CNN}(y) + \boldsymbol{b_r}. \quad (13)$$

Now, the term $KL(q_\phi(z|y)||p_\theta(z))$ is a KL-divergence between two multivariate Gaussian distributions which can be computed in a closed form (Doersch, 2016). Different from z, the dialog act a follows discrete distribution. Minimizing the term $KL(q_\phi(a|y)||p_\theta(a))$ is much simpler than the continuous one, which can be evaluated by

$$\sum_{k=1}^{\mathcal{A}} q_\phi(a = k|y) \log \frac{q_\phi(a = k|y)}{p_\theta(a = k) + \epsilon} \quad (14)$$

where $\mathcal{A}$ is the number of dialog acts and $\epsilon = 10^{-6}$ is used to prevent division by zero. We denote the network $q_\phi(a|y)$ as the act classifier and its probability is evaluated by $softmax(\boldsymbol{W_f}\text{CNN}(y) + \boldsymbol{b_f})$. As shown in Figure 1, note here that the inverse linear layer shares the same parameters $\boldsymbol{W_f}$ and $\boldsymbol{b_f}$ with those in $q_\phi(a|y)$.

In the training process, by introducing the re-parameterization trick (Kingma and Welling, 2014), we obtain the variables z and a from RecNet $q_\phi(z|y)$ and act classifier $q_\phi(a|y)$, and then feed them into the inverse linear layer to capture expression representations. During testing, by specifying a dialog act a, the decoder generates the response according to the sample from the PriNet $p_\theta(z)$.

4 Experiment

4.1 Implementation Details

The dataset DailyDialog Corpus (Li et al., 2017b) is used to evaluate the proposed model. It contains 13,118 multi-turn human-human dialogs annotated with dialog acts and emotions, and covers 10 main topics about daily life. In this Corpus, the dialog act categories are {*Inform, Question, Directive, Commissive*}. In our experiments, HRG is combined into HRED model (Serban et al., 2016) as the expression-aware chatting machine (ECM). PyTorch[1] is used to implement the proposed model. All the RNN modules have 2-layer gated recurrent units (GRU) (Cho et al., 2014) structures with 500 hidden cells for each layer and are set with different parameters. Word embedding has size 300 and is initialized from Glove embedding[2]. The size of the latent variable z is set to be 300. The maximum dialog turn is 5 (10 utterances).

The models are trained end-to-end using Adam optimizer (Kingma and Ba, 2015) with batch size

[1] https://pytorch.org/

[2] https://nlp.stanford.edu/projects/glove/

of 30, learning rate of 0.001 and gradient clipping at 50. To overcome the *latent variable vanishing problem* in CVAE, we use the heuristic method (Bowman et al., 2016) to encode the meaningful information in z. That is, we multiply the first KL term in Eq.12 by a scalar, which starts at 0 and linearly increases to 1 over the first 10,000 batches.

4.2 Baselines

We compare our hierarchical model with two popular baselines: (1) **HAE**: a HRED-based model which we design to embed the dialog act information in the decoder; (2) **kgCVAE**: a knowledge-guided model that introduces dialog acts to guide the learning of the CVAEs (Zhao et al., 2017). A variant of the proposed model is implemented to verify the effectiveness of expression attention (EA) model. We denote the model without EA as **w/o EA**. The hyperparameters of the baselines and variants are the same as ECM.

As for **HAE**, we initialize the embedding of dialog acts using three different methods: (1) **RD** (random): initializing the embedding randomly; (2) **LG** (logic-related): training a Skip-Gram model (Mikolov et al., 2013) to maximize the co-occurrence probability among the acts that appear within a window, w, in the sequence of dialog acts for each dialog (set w to 1); (3) **CT** (content-related): training an act classifier $q_\phi(a|y)$ with the pairs of utterances and dialog acts in training set, and use each row in $\boldsymbol{W_f}$ as the embeddings. The size of act embeddings is set to 300, as the same with the output of EA. The concatenation of dialog act embedding and the previous word embedding is fed into the decoder of HRED to update its state at each step during decoding. HAE is trained to minimize the standard cross entropy loss of the decoder RNN model without any auxiliary loss.

4.3 Quantitative Analysis

Automatically evaluating the quality of the dialog model remains an open question. To evaluate how semantically relevant the response is, we report the results for three word embedding-based similarity metrics proposed by Liu et al. (2016): *Greedy Matching* (GDY), *Embedding Average* (AVG) and *Vector Extrema* (EXT). To evaluate whether the response follows the dialog act, we adopt act accuracy (ACC) as the agreement between the ground-truth dialog act and the dialog act predicted by an act classifier. We trained the act classifier and its precision and average recall in the testing set are 83.4% and 74.3% respectively.

In addition to automatic evaluation metrics, a manual evaluation metric (MUL) is also given to evaluate both the response content and expression, where three workers are employed to score a response in terms of *Content* (rating scale is 0,1,2) and *Act* (rating scale is 0,1). *Content* is evaluated based on whether the response is appropriate and natural to the dialog context, while *Act* based on whether the expression agrees with the ground-truth act. *Content* rating is a widely accepted metric proposed by Shang et al. (2015). And, the workers can easily evaluate *Act* rating based on the context since the number of acts is few in our experiment.

During testing, to efficiently measure output diversity, we generate N responses from HAE models by introducing beam search. For kgCVAE and ECM, we sample N times from the latent variable and only use greedy decoders. Meanwhile, for HAEs and ECM, we specify a as the act of the ground-truth response.

The automatic evaluation metrics focus on comparing the generated responses r_j with the ground truth g_i of the conversation. We compute the scores of models based on all the M test samples as follows:

$$AM = \frac{1}{M}\sum_{i=1}^{M}\max_{j\in[1,N]} d(g_i, r_j) \qquad (15)$$

where $d(\cdot)$ is one of automatic metrics described above, and N is empirically set to 10. Note here that the maximum metric in Eq.15 is more appropriate to measure the output diversity than average one. This is because that taking average metrics may cause that the safety responses get higher scores than meaningful and diverse responses if most of these valid responses are not related to the ground-truth. The maximum metric can greatly reduce the error by increasing the number of samples. The evaluation of MUL metric is unrelated to the ground truth responses. To evaluate MUL metric, we randomly selected 40 dialog contexts from the test set and then generate 400 responses for each model. Each response is evaluated with a rating of *Content-Act* by workers.

The automatic evaluation results are given in Table 1. According to the word embedding-based similarity metrics, responses generated by ECM are substantially more coherent and relevant to

Method (%)	GDY	AVG	EXT	ACC
HAE-RD	11.6	57.7	33.8	64.2
HAE-LG	14.7	66.0	30.0	74.2
HAE-CT	18.2	72.0	35.5	77.0
kgCVAE	23.6	75.3	**39.8**	—
ECM	**28.1**	**84.6**	37.9	80.4
w/o EA	20.6	78.0	34.8	**82.6**

Table 1: Automatic Evaluation Result.

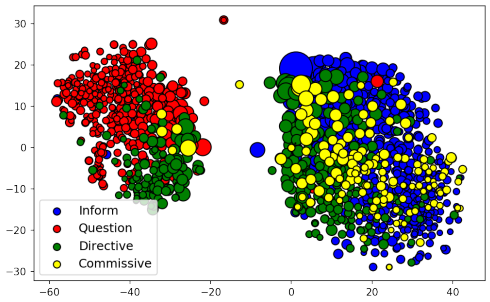

Figure 3: Visualization of expression representations for utterances with dialog acts. The size of circle represents the response length.

the topic compared to HAE models and kgCVAE. ECM obtains higher act accuracy score than HAE as well since the second KL term of Eq.(12) forces the predicted act distribution to approximate the ground-truth. ECM without EA (w/o EA) achieves the best performance in act accuracy but poor performance in embedding-based similarity metrics. It indicates that EA is an efficient model to balance the expression and the content dynamically. On the other hand, HAE-CT gets higher scores both in embedding-based metrics and in accuracy than other HAE models, which suggests that the act classifier can preserve act-related information effectively. Note here that the act accuracy of kgCVAE is not given because the response act is an internal parameter predicted by the dialog context rather than an input during testing. Compared to ECM, kgCVAE may give the decoder a wrong direction to approximate the ground-truth responses with different dialog acts.

Method (%)	2-1	2-0	1-1	1-0	0-1	0-0
HAE-CT	13.5	8.0	23.5	9.0	31.5	14.5
ECM	26.5	15.5	32.0	7.0	16.0	3.0
w/o EA	21.0	6.0	29.0	10.0	28.5	5.5

Table 2: Manual evaluation result. The percentage of responses with the ratings of *Content-Act*. For instance, 2-1 means *Content* rating is 2 and *Act* rating is 1.

Table 2 shows the manual evaluation result where the content and expression are considered simultaneously. As we can see, responses generated by w/o EA tend to contain obvious act information but a little of content, while HAE generates the responses with lower scores of *Content-Act*. Compared to other methods, responses generated by ECM keep a good balance between *Content* and *Act*. In our experiment, we also find that HAE-CT still faces serious safe response problems. However, EA provides token-level randomness to avoid that the decoder generates the next token only depending on Neural Probabilistic Language Model.

Discussion. ECM performs well than w/o EA not in automatic evaluation but also in manual evaluation obviously. Although ECM includes more trainable parameters in EA than w/o EA, the improvement of performance is mainly due to the effective architecture of EA. The EA model only involves two learnt matrices, i.e, $\boldsymbol{W_s}$ and $\boldsymbol{W_u}$, described in Section 3.2. Compared to the parameters in the multi-layer HRED, expression reconstruction, and the embedding of vocabulary, the number of parameters in EA can be ignored. Therefore, it can be concluded that EA plays a key role in improving output diversity with few parameters due to its efficient architecture.

4.4 Qualitative Analysis

Table 3 shows the responses generated by kgCVAE and ECM. In Example 1, speaker "A" begins with an open domain demand (directive). ECM generated highly diverse answers that cover multiple dialog acts which were fed into the model in advance during decoding. Further, we notice that the generated response with *inform* act (i.e., sample 1) has similar expression with the ground-truth one, implying that the latent z is able to capture the expression-sensitive variations. It verifies the effectiveness of the hierarchical generation process. ECM can obtain effective expression representation, and fill it with appropriate content obtained from the dialog context. Example 2 is a situation where the waiter "A" tells the customer "B" that the order has done. ECM takes the *directive* act as input and generate multiple responses to give "A" some suggestions (or commands). All the responses reflect the similar behaviors with different expression styles. On the contrary, kgCVAE is capable of generating some diverse responses, but cannot accurately understand the intention of

Example 1: **History** (Directive): **A**: Tell me a little bit about yourself , please . **Target** (Inform): **B**: My name is Dunlin and I live in Beijing . I was born in 1980 . I will graduate from Peking University this July .	
kgCVAE + z Samplings	**ECM** + Dialog Acts + z Samplings
1. i've been in china for two weeks .	1. (Inform) i graduated from hebes university .
2. how do you wish i can speak to the teacher's letter ?	2. (Commissive) of course , i ' ve been visiting a new company . i ' m looking for a job .
3. i'd like to speak a little more , but i know him .	3. (Question) yes , what ' s the matter ?
4. excuse me ?	4. (Directive) can you give me a example ?
Example 2: **History** (Inform): **waiter A**: good morning , sir . i ve got breakfast your ordered . **Target** (Directive): **customer B**: just put it on the table please .	
kgCVAE + z Samplings	**ECM** + *Directive* Act + z Samplings
1. oh , no .	1. put it on .
2. yes , i'd like to have a look . it s a good choice .	2. please check out this time .
3. that's ok . it's got wireless only , and we're over .	3. please . i am sorry to have kept you waiting .
4. good morning , thanks .	4. please wait a moment .

Table 3: Generated responses from kgCVAE and ECM in two examples.

"A" and thus the responses lack of coherency. The human-human dialogues in the dataset follow some dialog flow patterns, such as *Question-Inform*, *Directive-Commissive* (Li et al., 2017b). kgCVAE predicts the dialog act exactly in example 1 but wrongly in Example 2 since the pattern *Inform-Directive* is not common.

In our work, a CNN module is leveraged to filter the content-related information of utterances and get a discourse-level representation, i.e. expression vector, where meaningful expression information is preserved. CNN models have been shown to be efficient for NLP and have achieved excellent results in sentence modeling and classification. So we conjecture that the expression vectors are highly correlated with the dialog acts, and each one reflects a concrete expression representation of the specified dialog act. Figure 3 visualizes the expression vectors in the test dataset in 2D space using t-SNE (Der Maaten and Hinton, 2008). We find that the expression vectors are clustered into meaningful groups associated with the dialog acts, which confirms that CNN is an efficient tool to extract the expression information.

5 Related Works

Vanilla Seq2Seq model usually ends up with generic and dull responses. To tackle this problem, one line of research has focused on forcing the model to imitate some human's skills by augmenting the input with rich meta information. For example, some works separately gave chatbots the ability of emotions (Zhou et al., 2018), persona (Li et al., 2016b), vision (Huber et al., 2018; Wu et al., 2018) and thinking over the knowledge base (Liu et al., 2018; Zhu et al., 2017). In this work, we consider open domain dialogue generation with dialog acts. But, only a little works (Zhao et al., 2017; Serban et al., 2017a) on open domain end-to-end modeling take dialog acts into account.

On the other hand, many attempts have also been made to improve the architecture of Seq2Seq models by changing the training methods. Li et al. (2016a) attributed safe response problems to the use of MLE objective. Some works separately attempted to replace the MLE method with maximum mutual information (Li et al., 2016a), reinforcement learning (Zhang et al., 2018; Li et al., 2016c) and adversarial learning (Xu et al., 2017; Li et al., 2017a). Serban et al. (2017b) viewed the dialog context as prior knowledge and combined HRED model into the CVAE framework. Zhao et al. (2017) further introduced dialog acts to guide the learning of CVAE. In our paper, we use CVAE to learn the hierarchy generation model.

6 Conclusion and Future Work

In this paper, we investigate the problem of generating meaningful responses by imitating the hierarchical process of human response. Specifically, a hierarchical response generation model is proposed to hierarchically generate the expressions and fill them with appropriate content naturally and coherently. The experiment results show that HRED model equipped with HRG can generate responses appropriate not only in content but also in expression. Different from existing works, our model is interpretable and controllable.

In the future work, we will explore the act interactions with HRG. Instead of specifying a dialog act manually, the most appropriate one can be decided automatically.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No.U1636210, No.U1636211, No.61602025) and Beijing Natural Science Foundation of China (No.4182037).

References

Adi Ben-Lsrael and Thomas N. E. Greville. 1976. Generalized inverse: Theory and applications. *Journal of the Royal Statistical Society*, 139(1).

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *conference on computational natural language learning*, pages 10–21.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

James H. Martin Daniel Jurafsky. 2017. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Laurens Van Der Maaten and Geoffrey E Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Carl Doersch. 2016. Tutorial on variational autoencoders. *CoRR*, abs/1606.05908.

Bernd Huber, Daniel Mcduff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *CHI*, page 277.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

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

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *ACL*, volume 1, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Daniel Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *EMNLP*, pages 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sebastien Jean, Alan Ritter, and Daniel Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *EMNLP*, pages 2157–2169.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, pages 986–995.

Chiawei Liu, Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *ACL*, pages 1489–1498.

Minh Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR, Workshop Track Proceedings*.

Massimo Poesio and David Traum. 1998. Towards an axiomatization of dialogue acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology*. Citeseer.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3783.

Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017a. A deep reinforcement learning chatbot. *CoRR*, abs/1709.02349.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*, pages 1577–1586.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Qi Wu, Peng Wang, Chunhua Shen, Ian D Reid, and Anton Van Den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, pages 6106–6115.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, Xiaolong Wang, Zhuoran Wang, and Chao Qi. 2017. Neural response generation via gan with an approximate embedding layer. In *EMNLP*, pages 617–626.

Steve Young, Milica Gasic, Blaise Thomson, and J. D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pages 4567–4573.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *ACL*, 1:654–664.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*, pages 730–739.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264.