Learning to balance the coherence and diversity of response generation in generation-based chatbots

Shuliang Wang¹,², Dapeng Li¹, Jing Geng¹,², Longxing Yang³, and Hongyong Leng¹

Abstract
Generating response with both coherence and diversity is a challenging task in generation-based chatbots. It is more difficult to improve the coherence and diversity of dialog generation at the same time in the response generation model. In this article, we propose an improved method that improves the coherence and diversity of dialog generation by changing the model to use gamma sampling and adding attention mechanism to the knowledge-guided conditional variational autoencoder. The experimental results demonstrate that our proposed method can significantly improve the coherence and diversity of knowledge-guided conditional variational autoencoder for response generation in generation-based chatbots at the same time.

Keywords
Variational autoencoder, dialog system, deep learning, response generation, chatbots

Date received: 07 August 2019; accepted: 01 August 2020

Introduction
Together with the rapid growth of Internet social network conversation data and the successful application of deep learning in natural language processing, both academia and industry are paying more and more attention to build open domain nontask-oriented chatbots.¹ Retrieval-based methods and generation-based methods are currently the mainstream methods for building chatbots.² Many well-known chatbots such as MILABOT (from Montreal Institute for Learning Algorithms)³ and XiaoIce (from Microsoft)⁴ use generation-based methods, because they are end-to-end learnable and are good at capturing complicated syntactic and semantic relations between messages and responses.⁵

Sequence-to-sequence (Seq2Seq) model is a recurrent neural network (RNN) model of the encoder–decoder framework, which is commonly used in tasks such as machine translation and text generation in natural language processing. Most of the methods in generation-based chatbots use an improved model of the vanilla Seq2Seq model, which can generate long, diverse, and meaningful responses to meet the needs of response generation. Among them, knowledge-guided conditional variational autoencoder (kgCVAE) is the most distinctive model of these

¹ School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
² Academy of E-Government, Beijing Institute of Technology, Beijing, China
³ Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Corresponding authors:
Dapeng Li and Jing Geng, School of Computer Science and Technology, Beijing Institute of Technology, No. 5 South Street, Zhongguancun, Haidian District, Beijing 100081, China.
Emails: dapengli@bit.edu.cn; janegeng@bit.edu.cn
improved models, which building upon encoder–decoder framework (a.k.a. a Seq2Seq model), adding variational autoencoder to the model for improving the model’s response generation diversity. It assumes that the user’s question-and-answer relationship with chatbot belongs to one-to-many relationship, and this relationship conforms to the normal distribution, as shown in Figure 1. However, kgCVAE using random sampling method based on normal distribution may not conform to the basic law of natural language processing. Bestselling book The black swan: the impact of the highly improbable makes the public realize that the normal distribution is not suitable for all situations, and the abuse of the normal distribution will cause errors. The Zipf law is a basic law of natural language statistics in the study of natural language processing. The distribution of words in corpus should be a power-law distribution rather than a normal distribution. Therefore, we believe that random sampling in the model should be performed in accordance with the Zipf law, as shown in Figure 2.

We used the gamma sampling (GS) that comes from TensorFlow for data sampling and adjusted the parameters to make it conform to the Zipf law. The experimental results demonstrate that GS significantly improves the diversity of model generation dialog. However, it did lead to a decline in the coherence of the model generation dialog. Therefore, we tried to solve this problem by other methods. In fact, attention mechanism has been widely used for the improvement of the Seq2Seq model, as presented in Table 1. Thus, we also tried this method.

After adding attention mechanism to the model, we have greatly improved the coherence of the model’s generated dialog. But this method also led to a decline in the diversity of model generation dialog. Finally, we added the GS and attention mechanism to the model at the same time, which improved the coherence and diversity of response generated by the model. The balance of dialog generation coherence and diversity is achieved, and the best effect of the model is achieved.

Our contributions in this article are 2-folds: (1) we present a novel method that improved the coherence and diversity of response generation of kgCVAE and (2) we verified the effectiveness of our proposed method on a public data set and tested the effectiveness of kgCVAE with GS and kgCVAE with attention mechanism, respectively.

Related work
We briefly review the history of chatbots and introduce the most cutting-edge technology for building chatbots.

Chatbots system
The development of natural language processing is generally considered to be experienced in three eras of rationalism, empiricism, and deep learning. The development of dialog system is basically in sync with the development of natural language processing. The dialog system before the 90s was usually symbolic rule or template-based. Then with the successful application of the probability model in the field of natural language processing, the statistical learning-based dialog system dominates. After 2014, deep learning gradually began to dominate the field of natural language processing, and the dialog system also entered the era of deep learning. The chatbots system is a kind of open domain nontask-oriented dialog system, which is mainly built by two methods: retrieve-based method and generation-based method. The principle of the retrieval-based method is to make trained model select a response from a large corpus; the principle of the generation-based method is to generate a response directly by using trained model. Both retrieval-based method and generation-based method are based on deep learning. Our study focuses on generation-based method.
Vanilla Seq2Seq with attention (Bahuleyan et al.) 30.24 14.33 8.26 4.96
Vanilla Seq2Seq (Du et al.) 31.34 13.79 7.36 4.26

Table 1. BLEU scores of vanilla Seq2Seq and vanilla Seq2Seq with attention of experiments on the SQuAD data set. The results come from Du et al. and Bahuleyan et al.

BLEU: bilingual evaluation understudy; Seq2Seq: sequence-to-sequence; SQuAD: stanford question answering dataset.
The bold number of each column means the best experimental result for that column.

**Generation-based methods**

The Seq2Seq model has been successful in the field of machine translation. Due to the similarity of the principle, this model is quickly applied to the field of dialog generation, which enables the dialog generated by a trained model. Sutskever et al. used the vanilla Seq2Seq model to generate dialog. However, the vanilla Seq2Seq model has some problems. The ideal goal is to make the model generate long, diverse, and meaningful responses for users, but the vanilla Seq2Seq model does not do well. Li et al. named this problem “safe reply.” In response to this problem, many researchers have proposed solutions. Most of the proposed methods are to add new modules between the encoder and the decoder of the vanilla Seq2Seq. Serban et al. added latent variables to the vanilla Seq2Seq model that improved the diversity of dialog generation. Similarly, Zhou et al. also added latent responding mechanisms to the vanilla Seq2Seq model that improved the diversity of dialog generation. Xing et al. improved the model by introducing topic information into the vanilla Seq2Seq model. Wu et al. added dynamic vocabulary to the vanilla Seq2Seq model that significantly improved the diversity of dialog generation. On the other hand, Li et al. also applied re-ranking technology, reinforcement learning technology, and adversarial learning technology to dialog generation and achieved better results. In addition, Wu et al. proposed a new paradigm for response generation: prototype-then-edit, which outperforms many existing models on some metrics.

**Variational autoencoder for dialog generation**

In deep learning, Seq2Seq, variational autoencoder (VAE), and generative adversarial networks (GAN) are the most commonly used generation models. Among them, Seq2Seq is mainly used for text generation, and VAE and GAN are mainly used for image generation. Moreover, conditional variational autoencoder (CVAE), as an improved model of VAE, is also commonly used for image generation. Bowman et al. first combined VAE with the encoder and decoder framework and generated more diverse responses through normal distribution sampling in chatbots. However, VAE with the encoder and decoder framework for text generation is not the mainstream of text generation methods, and there are not many related works. Later, Zhao et al. first used CVAE with the encoder and decoder framework to generate dialog, which increased the diversity of dialog generation. In addition, Zhao et al. incorporated the leverage linguistic knowledge into the CVAE and proposed the kgCVAE model, which further improved the diversity of model’s dialog generation. In this article, we propose an improved method for kgCVAE by adding an attention mechanism module between the encoder and the decoder of kgCVAE and replacing normal distribution sampling with gamma distribution sampling.

**The method overview**

CVAE model is the basic of kgCVAE. Based on the traditional encoder–decoder framework, CVAE adds a prior network and a recognition network between the encoder and the decoder to improve the diversity of the trained model’s dialog generation. KgCVAE is an improved model of the CVAE model. On the basis of CVAE, a new language feature variable y (such as dialog act) is added between the encoder and the decoder to further improve the effect of the CVAE. In addition, we add an attention mechanism module to kgCVAE in our proposed method. The process of kgCVAE with attention mechanism is illustrated in Figure 3.

**CVAE model**

Each dyadic conversation of CVAE is represented via three random variables: c, x, and z. c is the dialog context, which is composed of the dialog history: the preceding k−1 utterances; conversational floor (1 means the same speaker and otherwise 0) and meta features m, such as conversational topic. x is the response utterance. z is a latent variable, which is used to capture the latent distribution over the valid responses. CVAE defines p(x, z|c) = p(x|z, c)p(z|c) and sets neural networks (parameterized by θ) to approximate p(z|c) and p(x|z, c). The model generates x through the response decoder pθ(x|z, c), after sampling latent variable z from the prior network pθ(z|c).

CVAE can be efficiently trained with the Stochastic Gradient Variational Bayes framework by maximizing the variational lower bound of the conditional log-likelihood. Specifically, CVAE is trained to maximize the
conditional log-likelihood of $x$ given $c$, which involves an intractable marginalization over the latent variable $z$. In CVAE, the latent variable $z$ follows multivariate normal distribution with a diagonal covariance matrix, and a recognition network $q_z(z|x,c)$ is used to approximate the true posterior distribution $p(x|z,c)$. The variational lower bound for CVAE is defined as follows

$$L(\theta, \phi; x, c) = -KL(q_z(z|x,c)p_z(z)) + E_{q_z(z|x,c)}[\log p_z(x|z,c)] \leq \log p(x|c)$$ (1)

In CVAE, the utterance encoder is a bidirectional gated recurrent unit (Bi-GRU), which encodes each utterance into fixed-size vectors by concatenating the last hidden states of the forward and backward RNN $u_t = \begin{bmatrix} h^t, \overrightarrow{h^t} \end{bmatrix}$. $x$ is simply $u_t$. The context encoder is a 1-layer Bi-GRU network, which encodes the preceding $k-1$ utterances by taking $u_{t:k-1}$ and the corresponding conversational floor as inputs. The last hidden state $h^c$ of the context encoder is concatenated with meta features $m$ and $c = [h^c, m]$. Since $z$ follows normal distribution, the recognition network $q_z(z|x,c) \sim N(\mu, \sigma^2I)$ and the prior network $p_z(z|c) \sim N(\mu', \sigma'^2I)$ are defined as equations (2) and (3), respectively

$$\begin{bmatrix} \mu \\ \log(\sigma^2) \end{bmatrix} = W_r \begin{bmatrix} x \\ c \end{bmatrix} + b_r$$ (2)

$$\begin{bmatrix} \mu' \\ \log(\sigma'^2) \end{bmatrix} = MLP_p(c)$$ (3)

CVAE obtains samples of $z$ either from $N(z; \mu, \sigma^2I)$ predicted by the recognition network or from $N(z; \mu', \sigma'^2I)$ predicted by the prior network via using the reparameterization trick. Lastly, the response decoder is a 1-layer GRU network with initial state $s_0 = W_i[z,c] + b_i$ and can predict the words in $x$ sequentially.

### kgCVAE model and optimization

KgCVAE is the development of CVAE. It has two advantages: (1) by incorporating linguistic feature $y$, such as dialog act, into CVAE, so that CVAE’s latent variable $z$ gets more information to facilitate model training and (2) when the model generates a response, an additional linguistic feature $y'$ is output for each response, which improves the interpretation of the model. In kgCVAE, the generation of $x$ depends on latent variable $z$, context $c$, and linguistic feature $y$, and linguistic feature $y$ relies on latent variable $z$ and context $c$. In the training stage, the initial state of the response decoder is $s_0 = W_i[z,c] + b$ and the input is $[e, y]$ at every step, where $e$ is the word embedding of $r$th word in $x$. Furthermore, kgCVAE has an MLP to predict $y' = MLP_y(z, c)$. The variational lower bound for kgCVAE is defined as follows

$$L(\theta, \phi; x, c, y) = -KL(q_z(z|x,c,y)p_z(z)) + E_{q_z(z|x,c,y)}[\log p(x|z,c,y)] + E_{q_z(z|x,c,y)}[\log p(y|z,c)]$$ (4)

However, there is an optimization problem in both kgCVAE and CVAE. Because they all use a straightforward VAE with RNN decoder, the vanishing latent variable problem will cause. To solve this problem, Zhao et al. proposed that bag-of-word (BOW) loss as an auxiliary loss that requires the decoder network to predict the BOWs in the response $x$. And $x$ is decomposed into two variables: $x_o$ and...
\( x_{bow} \). \( x_o \) has word order \( x_{bow} \) has no word order, and they are conditionally independent given \( z \) and \( c \): \( p(x,z|c) = p(x_o|z,c)p(x_{bow}|z,c)p(z|c) \). Therefore, the latent variable has to capture global information about the target response.

Let \( f = MLP_p(z,x) \in \mathbb{R}^k \), where \( V \) is vocabulary size and has equation (5)

\[
\log p(x_{bow}|z,c) = \log \prod_{i=1}^{|x|} \sum_{j} e^{f_{ji}} 
\]

where \(|x|\) and \( x \) represent the length of \( x \) and the word index of \( ith \) word in \( x \), respectively. The modified variational lower bound for CVAE with BOW loss and the modified variational lower bound for kgCVAE with BOW loss are defined as equations (6) and (7), respectively

\[
L'(\theta, \phi; x, c) = L(\theta, \phi; x, c) + E_q(z|x,c)[log p(x_{bow}|z,c)] \tag{6}
\]

\[
L'(\theta, \phi; x, c,y) = L(\theta, \phi; x, c,y) + E_q(z|x,c,y)[log p(x_{bow}|z,c)] \tag{7}
\]

In the experiment, BOW loss solved the vanishing latent variable problem while complementing the kullback-leibler (KL) annealing technique.

**Gamma sampling**

The gamma distribution is a continuous probability function of statistics. It has two parameters and is subject to an incomplete gamma function. It is usually used to model sums of exponentially distributed random variables.\(^{32}\) It is defined as follows

\[
p(x) = x^{a-1} \frac{e^{-x}}{b^a \Gamma(a)} \tag{8}
\]

where \( \Gamma \) is the gamma function, \( a \) is the shape parameter, and \( b \) is the scale parameter. In our proposed method, we use random gamma function from TensorFlow\(^{10}\) and set \( a = 1 \) to ensure that the function is monotonically decreasing. The graph of data distribution obtained by GS conforms to Zipf’s law.

**Attention mechanism**

In kgCVAE, the decoder generates \( x \) based on latent variable \( z \), context \( c \), and linguistic feature \( y \). Attention mechanism is added to dynamically align \( x \) and \([z, c, y] \) during generation. The attention mechanism computes a probabilistic distribution in the decoder at each step \( j \), it is defined as follows

\[
a_j = \frac{\exp\{\hat{a}_j\}}{\sum_{i=1}^{|z|}\exp\{\hat{a}_i\}} \tag{9}
\]

where \( \hat{a}_j \) is a prenormalized score, computed by \( \hat{a}_j = h_{j}^{(tar)} W^T h_{i}^{(src)} \) in the model. \( h_{j}^{(tar)} \) is the hidden representation of the \( j \)th step in target, \( h_{i}^{(src)} \) is the hidden representation of the \( i \)th in the source, and \( W \) is a learnable weight matrix.

Then, the source information \( \{h_{i}^{(src)}\}_{i=1}^{|z|} \) is summed by weights \( a_j \) to obtain the attention vector; it is defined as follows

\[
a_j = \sum_{i=1}^{|z|} a_j h_{i}^{(src)} \tag{10}
\]

which is fed to the decoder at the \( j \)th step.

Although, the addition of the attention mechanism in the Seq2Seq is a relatively common improving method for the Seq2Seq. However, there is a problem when using variational autoencoder in the Seq2Seq. Bahuleyan et al.\(^{13}\) found the Seq2Seq with a traditional attention mechanism; the variational latent space may be bypassed by the attention model and thus becomes ineffective. Therefore, they proposed a variational attention mechanism for the Seq2Seq, where the attention vector is also modeled as normal distributed random variables. At each step \( j \), it adjusts its hidden state \( h_j^{(tar)} \) with an input of a word embedding \( e_{j-1} \) in the decoding process. This is given by \( h_j^{(tar)} = \text{RNN}_\theta(h_{j-1}^{(tar)}, e_{j-1}) \). In the model, GRU is used as RNN’s transition. Enhanced with attention, the RNN is computed by \( h_j^{(tar)} = \text{RNN}_\theta(h_{j-1}^{(tar)}, [e_{j-1}, a_j]) \). The predicted word is given by a softmax layer \( p(e_j) = \text{softmax}(W_{out} h_j^{(tar)}) \), where \( W_{out} \) is a weight matrix.

We used the variational attention mechanism method of Bahuleyan et al.\(^{13}\) and used gamma distributed random variables for attention vector modeling. Therefore, when building attention mechanism, we treat both the latent space \( z \) and the attention vector \( a \) as random variables. Another noteworthy thing is that we did not use the optimization method proposed by Bahuleyan et al. We still use the optimization method of kgCVAE.

**Experiment**

We tested our proposed method on Switchboard (SW) 1 release 2 corpus\(^{6,33}\) and compared the effects of the GS and attention mechanism on the model.

**Experimental datasets**

SW 1 release 2 corpus was released in 1997 by Godfrey and Holliman.\(^{33}\) The data set consists of more than 2400 two-sided telephone conversations data. Each set of data has a topic tag and contains a total of 70 topics. Zhao et al. randomly selected 2316 dialogs for train, 60 dialogs for valid,
and 60 dialogs for test. And they used the natural language toolkit (NLTK) tokenizer to tokenize, kept the top 10,000 frequent word types as the vocabulary, and deleted nonverbal symbols and repeated words to process the data set. Finally, train data set has 207,833 context and response pairs, valid data set has 5225 context and response pairs, and test data set has 5481 context and response pairs. This data set is unique in that there are 42 kinds of dialog act features labeled by hand and machine. We used data published by Zhao et al. The statistics of SW 1 release 2 corpus are presented in Table 2.

### Multiple reference evaluation

Since the model was tested using one reference, the testing results were not reliable. Therefore, Zhao et al. used information retrieval techniques to collect 10 additional references with the same topic for each reference, and manually filtered out the poor quality of the reference. Finally, the average per context has 6.69 extra references. The specific data statistics are presented in Table 3.

Note that the multiple reference test data set is distinct from the file storage format of the one reference test of SW corpus. The multiple reference test data set does not contain topic information, and the dialog act is a different format. Therefore, it is very difficult to evaluate the model directly using the multiple reference test data set on the existing data interface. In fact, after noting that there are 5481 context and response pairs for the one reference test data set, researchers only need to train the model and then align the 5481 multiple references with the generated hypothesis. To measure the generated hypothesis, Zhao et al. designed precision and recall as metrics. Precision is used to measure the coherence of the generated dialog; recall is used to measure the diversity of the generated dialog. For a given dialog context $c$, there exist $M_r$ reference responses $r_j$, $j \in [1, M_r]$, and a trained model can generate $N$ hypothesis responses $h_i$, $i \in [1, N]$. The generalized response-level precision and recall are defined as equations (11) and (12), respectively

\[
\text{precision}(c) = \frac{\sum_{j=1}^N \max_{i \in [1, M_r]} d(r_j, h_i)}{N} \quad (11)
\]

\[
\text{recall}(c) = \frac{\sum_{j=1}^{M_r} \max_{i \in [1, N]} d(r_j, h_i)}{M_r} \quad (12)
\]

where $d(r_j, h_i) \in [0, 1]$ measures the similarities between reference responses $r_j$ and hypothesis responses $h_i$. The final score of evaluating model is averaged over the test data set with smoothed sentence-level bilingual evaluation understudy (BLEU), which measures the geometric mean of modified n-gram precision with a length penalty. In the experiment, from BLEU-1 to BLEU-4 are used as lexical similarity metric and BLEU scores are normalized to $[0, 1]$.

In addition to BLEU, perplexity is used to measure models in the experiment but not with precision and recall metrics. Perplexity is used to measure the ability of generation dialog model of the syntactic structure of each utterance and the syntactic structure of the dialog. Note that unlike the BLEU, the lower value of perplexity, the better a trained model.

### Experiment setup

Except hierarchical recurrent encoder-decoder (HRED) model is implemented by using Texar that is an open-source text generation toolkit based on TensorFlow, most models of experiments are implemented by using TensorFlow, and all models of experiments are run on a single 1080ti GPU. We follow the work of Zhao et al., use Glove Twitter pretrained Word2Vec file, and choose 200 as word embedding size. We choose Bi-GRU for the utterance encoder and set the hidden size of utterance encoder as 300. Furthermore, the hidden size of context encoder is 600, and the embedding size of topic is the same as the embedding size of dialog act is 30. The number of context RNN layers is 1. We set the hidden size of response decoder to 400, keep 10 utterances in the context window, and choose 40 as maximum number of words in an utterance. The dimension of latent variable is 200. We use word drop decoder as the same as Zhao et al. and set 10,000 batches before KL cost weight reaches 1, but change decoder keep probability to 0.95 for adding attention in kgCVAE. Training of models is optimized by Adam algorithm. We change the initial learning rate to 0.005 and change the mini-batch size to 150 to make full use of GPU resources.
Thus maximum number of epoch of training is 20 rather than 60. We set gradient clipping at 5 and set all the initial weights are $\frac{1}{C0}$, 0.08, 0.08 by sampled from a uniform distribution. But we change the dropout rate to 0.95 for adding attention in kgCVAE. Moreover, we adopt early stopping strategy as a regularization strategy and set the improve threshold and patient increase to 0.996 and 2, respectively. It is worth noting that although we change the decoder keep probability, initial learning rate, mini-batch size, and dropout rate, these parameter changes will not improve kgCVAE. We have verified this through experiments. In addition, we used a part of the experimental results from Zhao et al. as baselines.

### Automatic evaluation

We choose three response generation models as baselines, including HRED, CVAE, and kgCVAE. Table 4 demonstrates the automatic evaluation results on SW corpus. It shows that our proposed method has greatly improved kgCVAE and is better than other models of baselines. It is noteworthy that our recurring kgCVAE is a little different from the experimental results of Zhao et al. Our BLEU-2 score for kgCVAE is higher than that of Zhao et al., but BLEU-4 is lower than Zhao et al. BLEU-1 and BLEU-3 have little fluctuations. Hu et al.36 also found the same problem about recurring kgCVAE in their work. In order not to affect our study, we used our recurring kgCVAE as a baseline by referring to the study of Bahuleyan et al. As can be seen from Table 4, the improvement in kgCVAE $+$ GS on perplexity and recall is very obvious, but it causes a significant decline on precision. KgCVAE $+$ attention has a huge improvement on precision, but there is no obvious effect on recall. The effect of kgCVAE $+$ GS $+$ attention is remarkable, with significant improvements on both precision and recall from BLEU-1 to BLEU-4. And perplexity of kgCVAE $+$ GS $+$ attention is lower than kgCVAE. In summary, our proposed method improves the coherence of kgCVAE model generation dialog and improves the diversity of kgCVAE model generation dialog.

Table 4. Experimental results on SW corpus.a

|                | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|----------------|--------|--------|--------|--------|
|                | Perplexity | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| HRED           | 35.4   | 0.405  | 0.336  | 0.300   | 0.281  | 0.272     | 0.254  | 0.226     | 0.215  |
| CVAE           | 20.2   | 0.372  | 0.381  | 0.295   | 0.322  | 0.265     | 0.292  | 0.223     | 0.248  |
| kgCVAE (Zhao et al.) | 16.02 | 0.412  | 0.411  | 0.350   | 0.356  | 0.310     | 0.318  | 0.262     | 0.272  |
| kgCVAE (Ours)  | 16.22  | 0.415  | 0.412  | 0.362   | 0.358  | 0.300     | 0.313  | 0.249     | 0.245  |
| kgCVAE $+$ GS  | 12.97  | 0.515  | 0.454  | 0.317   | 0.392  | 0.266     | 0.342  | 0.206     | 0.259  |
| kgCVAE $+$ ATT | 18.71  | 0.456  | 0.408  | 0.417   | 0.371  | 0.350     | 0.328  | 0.286     | 0.256  |
| kgCVAE $+$ GS $+$ ATT | 13.85 | 0.538  | 0.450  | 0.424   | 0.401  | 0.356     | 0.358  | 0.280     | 0.274  |

BLEU: bilingual evaluation understudy; SW: switchboard; HRED: hierarchical recurrent encoder-decoder; GS: gamma sampling; ATT: attention; CVAE: conditional variational autoencoder; kgCVAE: knowledge-guided conditional variational autoencoder.

*aBold font indicates the method our proposed. The bold number of each column means the best experimental result for that column. BLEU scores are $[0, 1]$ by normalized.

Table 5. Generated responses from kgCVAE and our method.

| Topic: Buying a car |
|--------------------|
| **Context A:** Then we have got an 86 Chevy spectrum that we bought new my daughter was a freshman in college when we bought it for her |
| **Target B:** (acknowledge_(backchannel)) uh - huh |
| **kgCVAE** |
| Sample 0 | acknowledge_(backchannel) | oh |
| Sample 1 | acknowledge_(backchannel) | acknowledge_(backchannel) | um - hum |
| Sample 2 | acknowledge_(backchannel) | oh |
| Sample 3 | acknowledge_(backchannel) | acknowledge_(backchannel) | um - hum |
| Sample 4 | acknowledge_(backchannel) | oh |
| Sample 5 | acknowledge_(backchannel) | oh i see |
| Sample 6 | acknowledge_(backchannel) | oh it's |
| Sample 7 | abandoned_or_turn-exit/uninterpretable | um - hum |
| Sample 8 | abandoned_or_turn-exit/uninterpretable | um - hum |
| Sample 9 | acknowledge_(backchannel) | yeah |
| **Our method** |
| acknowledge_(backchannel) | yeah |
| acknowledge_(backchannel) | um - hum |
| acknowledge_(backchannel) | um - hum |
| acknowledge_(backchannel) | um - hum |
| acknowledge_(backchannel) | um - hum |
| acknowledge_(backchannel) | um - hum |

kgCVAE: knowledge-guided conditional variational autoencoder.
We created a case study on generated responses from kgCVAE and our method, as presented in Table 5. Table 5 illustrates that kgCVAE generates five types of responses including oh, oh i see, oh it’s, um - hum, and yeah; and our method generates six types of responses including yeah, um - hum, um - hum yeah, uh - huh, oh, and oh yeah. This means that the diversity of responses generated by our method is better than kgCVAE. In addition, Samples 5 and 9 generated by our method are completely consistent with Target B, and other responses generated by our method are all reasonable responses. However, kgCVAE does not generate a response exactly the same as Target B, and Sample 6 of kgCVAE is not a reasonable response. Therefore, the coherence of response generated by our method is better than kgCVAE. In summary, our method is better than kgCVAE in the coherence and diversity of response generation.

**Analysis**

We visualized values of random sampling and the latent variable $z$ of the prior network and tested the effectiveness of our proposed method on CVAE via model ablation of kgCVAE. In addition, we further explored the different effects by using greedy search decoder and using random sampling decoder in our method.

**Model ablation**

CVAE is the basis of kgCVAE and can be regarded as a model ablation for kgCVAE. We added GS, attention, and
GS + attention to CVAE to verify the effectiveness of our method for CVAE. The experimental results demonstrate that our method can also improve CVAE, as presented in Table 6. When CVAE + GS, perplexity drops sharply, and the precision and recall of BLEU-1 are greatly improved. The recall of BLEU-2 and BLEU-3 is greatly improved, and the precision from BLEU-2 to BLEU-4 is greatly reduced. At CVAE + attention, although perplexity has risen sharply, the precision from BLEU-1 to BLEU-4 has increased significantly. When CVAE + GS + attention, perplexity drops dramatically, and the precision and recall from BLEU-1 to BLEU-4 are greatly improved. Therefore, our improved method is also effective for CVAE.

Greedy search and random sampling

Our improved method uses greedy search decoder because both CVAE and kgCVAE use greedy search decoder. However, one of the baselines, HRED, uses random sampling decoder. Therefore, we also try to use random sampling decoder in our improved method. As presented in Table 7, regardless of kgCVAE + GS and kgCVAE + attention, or kgCVAE + GS + attention, random sampling decoder is not as good as greedy search decoder. Although the perplexity decreased slightly after random sampling decoder was used for kgCVAE + GS, the precision and recall from BLEU-1 to BLEU-4 decreased significantly. The same is true for kgCVAE + attention using random sampling decoder. When kgCVAE + GS + attention adopts random sampling decoder, perplexity drops a little, and the precision and recall from BLEU-1 to BLEU-4 drop sharply. Therefore, our improved method based on kgCVAE uses greedy search decoder better than random sampling decoder.

Conclusions

In this article, we proposed an improved method based on kgCVAE for response generation in generation-based chatbots. This method can help kgCVAE achieve balance and improve the coherence and diversity of dialog generation, which is a key to generate meaningful and diverse response in generation-based chatbots. The experimental results demonstrate that our proposed method has greatly improved the coherence and diversity of kgCVAE for response generation in generation-based chatbots at the same time. In the future, we will study how to apply our proposed method to other applied variational autoencoder tasks and examine if GS with attention mechanism can improve other models.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article was supported by Beijing Municipal Science and Technology Project (Z171100005117002) and Open Fund of Key

**Table 6.** Experimental results of model ablation.

|               | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---------------|--------|--------|--------|--------|
|               | Perplexity | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| CVAE          | 20.2   | 0.372  | 0.381  | 0.295   | 0.322  | 0.265   | 0.292  | 0.223   | 0.248  |
| CVAE + GS     | 12.96  | 0.493  | 0.444  | 0.232   | 0.367  | 0.192   | 0.316  | 0.149   | 0.240  |
| CVAE + ATT    | 20.09  | 0.436  | 0.388  | 0.393   | 0.345  | 0.327   | 0.299  | 0.271   | 0.236  |
| CVAE + GS + ATT | 14.79 | 0.546  | 0.446  | 0.412   | 0.385  | 0.334   | 0.332  | 0.264   | 0.256  |

BLEU: bilingual evaluation understudy; GS: gamma sampling; ATT: attention; CVAE: conditional variational autoencoder. Bold font indicates the method our proposed. The bold number of each column means the best experimental result for that column.

**Table 7.** Experimental results of different types of decoder in our proposed method.

|               | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---------------|--------|--------|--------|--------|
|               | Types of decoder | Perplexity | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| kgCVAE + GS   | Greedy search   | 12.97   | 0.515   | 0.454  | 0.317   | 0.392  | 0.266   | 0.342  | 0.206   | 0.259  |
| kgCVAE + GS   | Random sampling | 12.80   | 0.458   | 0.436  | 0.280   | 0.360  | 0.226   | 0.304  | 0.181   | 0.235  |
| kgCVAE + ATT  | Greedy search   | 18.71   | 0.456   | 0.408  | 0.417   | 0.371  | 0.350   | 0.328  | 0.286   | 0.256  |
| kgCVAE + ATT  | Random sampling | 18.73   | 0.380   | 0.375  | 0.376   | 0.338  | 0.315   | 0.295  | 0.263   | 0.233  |
| kgCVAE + GS + ATT | Greedy search | 13.85   | 0.538   | 0.450  | 0.424   | 0.401  | 0.356   | 0.358  | 0.280   | 0.274  |
| kgCVAE + GS + ATT | Random sampling | 13.67   | 0.466   | 0.432  | 0.371   | 0.366  | 0.299   | 0.311  | 0.238   | 0.239  |

BLEU: bilingual evaluation understudy; GS: gamma sampling; ATT: attention; kgCVAE: knowledge-guided conditional variational autoencoder. The bold number of each column means the best experimental result for that column.
Laboratory for National Geographic Census and Monitoring, National Administration of Surveying, Mapping and Geoinformation (2017NGCMZD03).

ORCID iD
Dapeng Li https://orcid.org/0000-0002-4735-7651

References
1. Wu Y, Furu W, Shaohan H, et al. Response generation by context-aware prototype editing. In: Proceedings of the 33rd AAAI conference on artificial intelligence (eds Van Hentenryck P and Zhou Z-H), Honolulua, Hawaii, USA, 27 January–1 February 2019, pp. 7281–7288. Palo Alto, CA: AAAI Press.
2. Wang S, Dapeng L, Jing G, et al. Learning bi-utterance for multi-turn response selection in retrieval-based chatbots. Int J Adv Robot Sys 2019; 16(2): 1–10.
3. Serban IV, Chinnadhurai S, Mathieu G, et al. A deep reinforcement learning chatbot. CoRR 2017. abs/1709.02349.
4. Shum H-Y, Xiaodong H, and Di L. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Front Inf Technol Electron Eng 2018; 19(1): 10–26.
5. Wu Y, Wei W, Dejian Y, et al. Neural response generation with dynamic vocabularies. In: Proceedings of the 32nd AAAI conference on artificial intelligence (eds Mellraith SA and Weinberger KQ), New Orleans, Louisiana USA, 2–7 February 2018, pp. 5594–5601. Palo Alto, CA: AAAI Press.
6. Zhao T, Ran Z, and Maxine E. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Proceedings of the 55th annual meeting of the association for computational linguistics (eds Barzilay R and Min-Yen K), Vancouver, Canada, 30 July–4 August 2017, pp. 654–664. Palo Alto, CA: ACL.
7. Taleb NN. The black swan: the impact of the highly improbable. New York: Random House; Bristol: Allen Lane, 2007.
8. Zipf GK. Selected studies of the principle of relative frequency in language. Cambridge, MA: Harvard Univ. Press, 1932.
9. van der Walt S, Chris Colbert S, and Gael V. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng 2011; 13: 22–30.
10. Abadi M, Ashish A, Paul B, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. CoRR 2016. abs/1603.04467.
11. Rajpurkar P, Jian Z, Konstantin L, et al. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing (eds Su J, Xavier C, and Kevin D), Austin, Texas, 1–5 November 2016, pp. 2383–2392. Stroudsburg, PA: ACL.
12. Du X, Junru S, and Claire C. Learning to ask: neural question generation for reading comprehension. In: Proceedings of the 55th annual meeting of the association for computational linguistics (eds Barzilay R and Min-Yen K), Vancouver, Canada, 30 July–4 August 2017, pp. 1342–1352. Stroudsburg, PA: ACL.
13. Bahuleyan H, Lili M, Olga V, et al. Variational attention for sequence-to-sequence models. In: Proceedings of the 27th international conference on computational linguistics (eds Bender EM, Leon D, and Pierre I), Santa Fe, New Mexico, USA, 20–25 August 2018, pp. 1672–1682. Stroudsburg, PA: ACL.
14. Deng L and Yang L. Deep learning in natural language processing. Singapore: Springer, 2018.
15. Hu B, Zhongdong L, Hang L, et al. Convolutional neural network architectures for matching natural language sentences. In: Proceedings of the 27th international conference on neural information processing systems (eds Z Ghahramani, M Welling, C Cortes, et al.) Montreal, Canada, 08–13 December 2014, pp. 2042–2050. Cambridge, MA: MIT Press.
16. Wu Y, Wei W, Chen X, et al. Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th annual meeting of the association for computational linguistics (eds Barzilay R and Min-Yen K), Vancouver, Canada, 30 July–4 August 2017, pp. 496–505. Stroudsburg, PA: ACL.
17. Wu Y, Wei W, Zhiouju L, et al. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers) (eds Gurevych I and Yusuke M), Melbourne, Australia, 15–20 July 2018, pp. 420–425. Stroudsburg, PA: ACL.
18. Zhou X, Li L, Dong D, et al. Multi-turn response selection for chatbots with deep attention matching network. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers) (eds Gurevych I and Miyao Y), Melbourne, Australia, 15–20 July 2018, pp. 1118–1127. Stroudsburg, PA: ACL.
19. Sutskever I, Oriol V, and Le Q.V. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014 (eds Zoubin G, Welling M, Cortes C, et al.), Montreal, Quebec, Canada, 8–13 December 2014, pp. 3104–3112. Cambridge, MA: MIT Press.
20. Vinyals O and Quoc L. A neural conversational model. CoRR 2015. abs/1506.05869.
21. Li J, Michel G, Chris B, et al. 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 (eds Knight K, Ani N, and Owen R), San Diego, California, 12–17 June 2016, pp. 110–119. Stroudsburg, PA: ACL.
22. Serban IV, Alessandro S, Yoshua B, et al. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th AAAI conference on artificial intelligence (eds Schuurmans D and Wellman MP), Phoenix, Arizona, 12–17 February 2016, pp. 3776–3783. Palo Alto, CA: AAAI Press.
23. Serban IV, Tim K, Gerald T, et al. Multiresolution recurrent neural networks: an application to dialogue response generation. In: *Proceedings of the 31st AAAI conference on artificial intelligence* (eds Singh SP and Shaul M), San Francisco, California USA, 4–9 February 2017, pp. 3288–3294. Palo Alto, CA: AAAI Press.

24. Serban IV, Alessandro S, Ryan L, et al. A hierarchical latent variable encoder-decoder model for generating dialogues. In: *Proceedings of the 31st AAAI conference on artificial intelligence* (eds Singh SP and Shaul M), San Francisco, California USA, 4–9 February 2017, pp. 3295–3301. Palo Alto, CA: AAAI Press.

25. Zhou G, Ping L, Rongyu C, et al. Mechanism-aware neural machine for dialogue response generation. In: *Proceedings of the 31st AAAI conference on artificial intelligence* (eds Singh SP and Shaul M), San Francisco, California USA, 4–9 February 2017, pp. 3351–3357. Palo Alto, CA: AAAI Press.

26. Xing C, Wei W, Yu W, et al. Topic aware neural response generation. In: *Proceedings of the 31st AAAI conference on artificial intelligence* (eds Singh SP and Shaul M), San Francisco, California USA, 4–9 February 2017, pp. 3351–3357. Palo Alto, CA: AAAI Press.

27. Li J, Will M, Alan R, et al. Deep reinforcement learning for dialogue generation. In: *Proceedings of the 2016 conference on empirical methods in natural language processing* (eds Su J, Xavier C, and Kevin D), Austin, Texas, 1–5 November 2016, pp. 1192–1202. Stroudsburg, PA: ACL.

28. Li J, Will M, Tianlin S, et al. Adversarial learning for neural dialogue generation. In: *Proceedings of the 2017 conference on empirical methods in natural language processing* (eds Palmer M, Rebecca H, and Sebastian R), Copenhagen, Denmark, 9–11 September 2017, pp. 2157–2169. Stroudsburg, PA: ACL.

29. Bowman SR, Luke V, Oriol V, et al. Generating sentences from a continuous space. In: *Proceedings of the 20th SIGNLL conference on computational natural language learning (CoNLL)* (eds Riezler S and Yoav G), Berlin, Germany, 11–12 August 2016, pp. 10–21. Stroudsburg, PA: ACL.

30. Zhao T, Kysoung L, and Maxine E. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In: *Proceedings of the 56th annual meeting of the association for computational linguistics (long papers)* (eds Gurevych I and Yusuke M), Melbourne, Australia, 15–20 July 2018, pp. 1098–1107. Stroudsburg, PA: ACL.

31. Kingma DP and Welling M. Auto-encoding variational bayes. *CoRR* 2013. abs/1312.6114.

32. Hahn GJ and Shapiro SS. *Statistical models in engineering*. Hoboken, NJ: John Wiley & Sons, Inc., 1994, p. 88.

33. Godfrey JJ and Edward H. *Switchboard-1 release 2*. Philadelphia, PA: Linguistic Data Consortium, 1997.

34. Bird S, Ewan K, and Edward L. *Natural language processing with python*. Newton, MA: O’Reilly Media, 2009.

35. Chen B and Colin C. A systematic comparison of smoothing techniques for sentence-Level BLEU. In: *Proceedings of the ninth workshop on statistical machine translation* (eds Bojar O, Christian B, Christian F, et al.), Baltimore, Maryland USA, June 26–27, 2014, pp. 362–367. Stroudsburg, PA: ACL.

36. Hu Z, Haoran S, Bowen T, et al. Texar: a modularized, versatile, and extensible toolkit for text generation. *CoRR* 2018. abs/1809.00794.

37. Pennington J, Richard S, and Manning CD. Glove: global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (eds Moschitti A, Bo P, and Walter D), Doha, Qatar, 25–29 October 2014, pp. 1532–1543. Stroudsburg, PA: ACL.

38. Kingma DP and Ba J. Adam: a method for stochastic optimization. In: *Proceedings of the 3rd international conference for learning representations*, San Diego, USA, 7–9 May 2015.