A Hierarchical Latent Structure for Variational Conversation Modeling

Yookoon Park **Jaemin Cho** **Gunhee Kim**
Department of Computer Science and Engineering & Center for Superintelligence
Seoul National University, Korea
yookoonpark@vision.snu.ac.kr, {jaemin895,gunhee}@snu.ac.kr
http://vision.snu.ac.kr/projects/vhcr

Abstract

Variational autoencoders (VAE) combined with hierarchical RNNs have emerged as a powerful framework for conversation modeling. However, they suffer from the notorious degeneration problem, where the decoders learn to ignore latent variables and reduce to vanilla RNNs. We empirically show that this degeneracy occurs mostly due to two reasons. First, the expressive power of hierarchical RNN decoders is often high enough to model the data using only its decoding distributions without relying on the latent variables. Second, the conditional VAE structure whose generation process is conditioned on a context, makes the range of training targets very sparse; that is, the RNN decoders can easily overfit to the training data ignoring the latent variables. To solve the degeneration problem, we propose a novel model named *Variational Hierarchical Conversation RNNs* (VHCR), involving two key ideas of (1) using a hierarchical structure of latent variables, and (2) exploiting an *utterance drop* regularization. With evaluations on two datasets of Cornell Movie Dialog and Ubuntu Dialog Corpus, we show that our VHCR successfully utilizes latent variables and outperforms state-of-the-art models for conversation generation. Moreover, it can perform several new utterance control tasks, thanks to its hierarchical latent structure.

1 Introduction

Conversation modeling has been a long interest of natural language research. Recent approaches for data-driven conversation modeling mostly build upon recurrent neural networks (RNNs) (Vinyals and Le, 2015; Sordoni et al., 2015b; Shang et al., 2015; Li et al., 2017; Serban et al., 2016). Serban et al. (2016) use a hierarchical RNN structure to model the context of conversation. Serban et al. (2017) further exploit an utterance latent variable in the hierarchical RNNs by incorporating the variational autoencoder (VAE) framework (Kingma and Welling, 2014; Rezende et al., 2014).

VAEs enable us to train a latent variable model for natural language modeling, which grants us several advantages. First, latent variables can learn an interpretable holistic representation, such as topics, tones, or high-level syntactic properties. Second, latent variables can model inherently abundant variability of natural language by encoding its global and long-term structure, which is hard to be captured by shallow generative processes (*e.g.* vanilla RNNs) where the only source of stochasticity comes from the sampling of output words.

In spite of such appealing properties of latent variable models for natural language modeling, VAEs suffer from the notorious *degeneration* problem (Bowman et al., 2016; Chen et al., 2017) that occurs when a VAE is combined with a powerful decoder such as autoregressive RNNs. This issue makes VAEs ignore latent variables, and eventually behave as vanilla RNNs. Chen et al. (2017) also note this degeneration issue by showing that a VAE with a RNN decoder prefers to model the data using its decoding distribution rather than using latent variables, from bits-back coding perspective. To resolve this issue, several heuristics have been proposed to weaken the decoder, enforcing the models to use latent variables. For example, Bowman et al. (2016) propose some heuristics, including *KL annealing* and *word drop* regularization. However, these heuristics cannot be a complete solution; for example, we observe that they fail to prevent the degeneracy in VHRED (Serban et al., 2017), a conditional VAE model equipped with hierarchical RNNs for conversation modeling.

The objective of this work is to propose a novel VAE model that significantly alleviates the degen-

Proceedings of NAACL-HLT 2018, pages 1792–1801
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

eration problem. Our analysis reveals that the causes of the degeneracy are two-fold. First, the hierarchical structure of autoregressive RNNs is powerful enough to predict a sequence of utterances without the need of latent variables, even with the word drop regularization. Second, we newly discover that the conditional VAE structure where an utterance is generated conditioned on context, *i.e.* a previous sequence of utterances, induces severe data sparsity. Even with a large-scale training corpus, there only exist very few target utterances when conditioned on the context. Hence, the hierarchical RNNs can easily memorize the context-to-utterance relations without relying on latent variables.

We propose a novel model named *Variational Hierarchical Conversation RNN* (VHCR), which involves two novel features to alleviate this problem. First, we introduce a global conversational latent variable along with local utterance latent variables to build a hierarchical latent structure. Second, we propose a new regularization technique called *utterance drop*. We show that our hierarchical latent structure is not only crucial for facilitating the use of latent variables in conversation modeling, but also delivers several additional advantages, including gaining control over the global context in which the conversation takes place.

Our major contributions are as follows:

(1) We reveal that the existing conditional VAE model with hierarchical RNNs for conversation modeling (*e.g.* (Serban et al., 2017)) still suffers from the degeneration problem, and this problem is caused by data sparsity per context that arises from the conditional VAE structure, as well as the use of powerful hierarchical RNN decoders.

(2) We propose a novel variational hierarchical conversation RNN (VHCR), which has two distinctive features: a hierarchical latent structure and a new regularization of utterance drop. To the best of our knowledge, our VHCR is the first VAE conversation model that exploits the hierarchical latent structure.

(3) With evaluations on two benchmark datasets of Cornell Movie Dialog (Danescu-Niculescu-Mizil and Lee, 2011) and Ubuntu Dialog Corpus (Lowe et al., 2015), we show that our model improves the conversation performance in multiple metrics over state-of-the-art methods, including HRED (Serban et al., 2016), and VHRED (Serban et al., 2017) with existing degeneracy solutions such as the word drop (Bowman et al., 2016), and the bag-of-words loss (Zhao et al., 2017).

2 Related Work

Conversation Modeling. One popular approach for conversation modeling is to use RNN-based encoders and decoders, such as (Vinyals and Le, 2015; Sordoni et al., 2015b; Shang et al., 2015). Hierarchical recurrent encoder-decoder (HRED) models (Sordoni et al., 2015a; Serban et al., 2016, 2017) consist of utterance encoder and decoder, and a context RNN which runs over utterance representations to model long-term temporal structure of conversation.

Recently, latent variable models such as VAEs have been adopted in language modeling (Bowman et al., 2016; Zhang et al., 2016; Serban et al., 2017). The VHRED model (Serban et al., 2017) integrates the VAE with the HRED to model Twitter and Ubuntu IRC conversations by introducing an utterance latent variable. This makes a conditional VAE where the generation process is conditioned on the context of conversation. Zhao et al. (2017) further make use of discourse act labels to capture the diversity of conversations.

Degeneracy of Variational Autoencoders. For sequence modeling, VAEs are often merged with the RNN encoder-decoder structure (Bowman et al., 2016; Serban et al., 2017; Zhao et al., 2017) where the encoder predicts the posterior distribution of a latent variable $\mathbf{z}$, and the decoder models the output distributions conditioned on $\mathbf{z}$. However, Bowman et al. (2016) report that a VAE with a RNN decoder easily degenerates; that is, it learns to ignore the latent variable $\mathbf{z}$ and falls back to a vanilla RNN. They propose two techniques to alleviate this issue: *KL annealing* and *word drop*. Chen et al. (2017) interpret this degeneracy in the context of bits-back coding and show that a VAE equipped with autoregressive models such as RNNs often ignores the latent variable to minimize the code length needed for describing data. They propose to constrain the decoder to selectively encode the information of interest in the latent variable. However, their empirical results are limited to an image domain. Zhao et al. (2017) use an auxiliary bag-of-words loss on the latent variable to force the model to use $\mathbf{z}$. That is, they train an auxiliary network that predicts bag-of-words representation of the target utterance based on $\mathbf{z}$. Yet this loss works in an opposite direction to the original objective of VAEs that minimizes the minimum description length. Thus, it may be in danger of forcibly moving the information that is better modeled in the decoder to the latent variable.

3 Approach

We assume that the training set consists of N i.i.d samples of conversations $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N\}$ where each $\mathbf{c}_i$ is a sequence of utterances (*i.e.* sentences) $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{in_i}\}$. Our objective is to learn the parameters of a generative network $\boldsymbol{\theta}$ using Maximum Likelihood Estimation (MLE):

$$\arg\max_{\boldsymbol{\theta}} \sum_i \log p_{\boldsymbol{\theta}}(\mathbf{c}_i) \tag{1}$$

We first briefly review the VAE, and explain the degeneracy issue before presenting our model.

3.1 Preliminary: Variational Autoencoder

We follow the notion of Kingma and Welling (2014). A datapoint $\mathbf{x}$ is generated from a latent variable $\mathbf{z}$, which is sampled from some prior distribution $p(\mathbf{z})$, typically a standard Gaussian distribution $\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$. We assume parametric families for conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. Since it is intractable to compute the log-marginal likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$, we approximate the intractable true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ with a recognition model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to maximize the *variational lower-bound*:

$$\begin{aligned}
&\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \qquad (2)\\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})]\\
&= -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]
\end{aligned}$$

Eq. 2 is decomposed into two terms: KL divergence term and reconstruction term. Here, KL divergence measures the amount of information encoded in the latent variable $\mathbf{z}$. In the extreme where KL divergence is zero, the model completely ignores $\mathbf{z}$, *i.e.* it degenerates. The expectation term can be stochastically approximated by sampling $\mathbf{z}$ from the variational posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The gradients to the recognition model can be efficiently estimated using the *reparameterization* trick (Kingma and Welling, 2014).

3.2 VHRED

Serban et al. (2017) propose Variational Hierarchical Recurrent Encoder Decoder (VHRED) model for conversation modeling. It integrates an utterance latent variable $\mathbf{z}_t^{\text{utt}}$ into the HRED structure (Sordoni et al., 2015a) which consists of three RNN components: *encoder RNN*, *context RNN*, and *decoder RNN*. Given a previous sequence of utterances $\mathbf{x}_1, ...\mathbf{x}_{t-1}$ in a conversation, the VHRED generates the next utterance $\mathbf{x}_t$ as:

$$\begin{aligned}
\mathbf{h}_{t-1}^{\text{enc}} &= f_{\boldsymbol{\theta}}^{\text{enc}}(\mathbf{x}_{t-1}) && (3)\\
\mathbf{h}_t^{\text{cxt}} &= f_{\boldsymbol{\theta}}^{\text{cxt}}(\mathbf{h}_{t-1}^{\text{cxt}}, \mathbf{h}_{t-1}^{\text{enc}}) && (4)\\
p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\mathbf{I}) && (5)\\
\text{where } \boldsymbol{\mu}_t &= \text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}) && (6)\\
\boldsymbol{\sigma}_t &= \text{Softplus}(\text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}})) && (7)\\
p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}) &= f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{h}_t^{\text{cxt}}, \mathbf{z}_t^{\text{utt}}) && (8)
\end{aligned}$$

At time step t, the encoder RNN $f_{\boldsymbol{\theta}}^{\text{enc}}$ takes the previous utterance $\mathbf{x}_{t-1}$ and produces an encoder vector $\mathbf{h}_{t-1}^{\text{enc}}$ (Eq. 3). The context RNN $f_{\boldsymbol{\theta}}^{\text{cxt}}$ models the context of the conversation by updating its hidden states using the encoder vector (Eq. 4). The context $\mathbf{h}_t^{\text{cxt}}$ defines the conditional prior $p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t})$, which is a factorized Gaussian distribution whose mean $\boldsymbol{\mu}_t$ and diagonal variance $\boldsymbol{\sigma}_t$ are given by feed-forward neural networks (Eq. 5-7). Finally the decoder RNN $f_{\boldsymbol{\theta}}^{\text{dec}}$ generates the utterance $\mathbf{x}_t$, conditioned on the context vector $\mathbf{h}_t^{\text{cxt}}$ and the latent variable $\mathbf{z}_t^{\text{utt}}$ (Eq. 8). We make two important notes: (1) the context RNN can be viewed as a high-level decoder, and together with the decoder RNN, they comprise a hierarchical RNN decoder. (2) VHRED follows a conditional VAE structure where each utterance $\mathbf{x}_t$ is generated conditioned on the context $\mathbf{h}_t^{\text{cxt}}$ (Eq. 5-8).

The variational posterior is a factorized Gaussian distribution where the mean and the diagonal variance are predicted from the target utterance and the context as follows:

$$\begin{aligned}
q_{\boldsymbol{\phi}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{\leq t}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_t', \boldsymbol{\sigma}_t' I) && (9)\\
\text{where } \boldsymbol{\mu}_t' &= \text{MLP}_{\boldsymbol{\phi}}(\mathbf{x}_t, \mathbf{h}_t^{\text{cxt}}) && (10)\\
\boldsymbol{\sigma}_t' &= \text{Softplus}(\text{MLP}_{\boldsymbol{\phi}}(\mathbf{x}_t, \mathbf{h}_t^{\text{cxt}})) && (11)
\end{aligned}$$

3.3 The Degeneration Problem

A known problem of a VAE that incorporates an autoregressive RNN decoder is the degeneracy that ignores the latent variable $\mathbf{z}$. In other words, the KL divergence term in Eq. 2 goes to zero and the decoder fails to learn any dependency between the latent variable and the data. Eventually, the model behaves as a vanilla RNN. This problem is

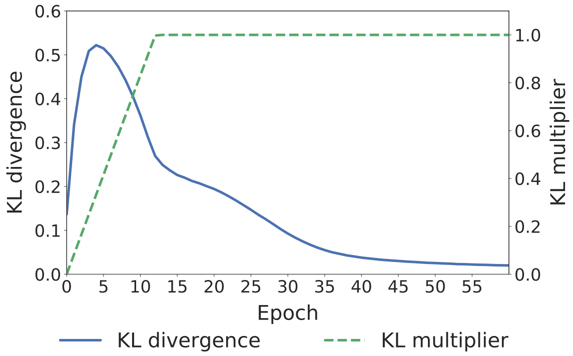

Figure 1: Degeneration of VHRED. The KL divergence term continuously decreases as training proceeds, meaning that the decoder ignores the latent variable $\mathbf{z}^{\text{utt}}$. We train the VHRED on Cornell Movie Dialog Corpus with word drop and KL annealing.

first reported in the sentence VAE (Bowman et al., 2016), in which following two heuristics are proposed to alleviate the problem by weakening the decoder.

First, the *KL annealing* scales the KL divergence term of Eq. 2 using a KL multiplier λ, which gradually increases from 0 to 1 during training:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = -\lambda D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] \quad (12)$$

This helps the optimization process to avoid local optima of zero KL divergence in early training. Second, the *word drop* regularization randomly replaces some conditioned-on word tokens in the RNN decoder with the generic unknown word token (UNK) during training. Normally, the RNN decoder predicts each next word in an autoregressive manner, conditioned on the previous sequence of ground truth (GT) words. By randomly replacing a GT word with an UNK token, the word drop regularization weakens the autoregressive power of the decoder and forces it to rely on the latent variable to predict the next word. The word drop probability is normally set to 0.25, since using a higher probability may degrade the model performance (Bowman et al., 2016).

However, we observe that these tricks do not solve the degeneracy for the VHRED in conversation modeling. An example in Fig. 1 shows that the VHRED learns to ignore the utterance latent variable as the KL divergence term falls to zero.

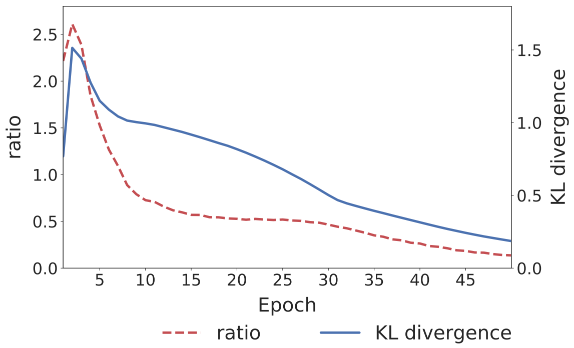

Figure 2: The average ratio $\mathbb{E}[\boldsymbol{\sigma}_t^2]/\text{Var}(\boldsymbol{\mu}_t)$ when the decoder is only conditioned on $\mathbf{z}_t^{\text{utt}}$. The ratio drops to zero as training proceeds, indicating that the conditional priors $p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t})$ degenerate to separate point masses.

3.4 Empirical Observation on Degeneracy

The decoder RNN of the VHRED in Eq. 8 conditions on two information sources: deterministic $\mathbf{h}_t^{\text{cxt}}$ and stochastic $\mathbf{z}^{\text{utt}}$. In order to check whether the presence of deterministic source $\mathbf{h}_t^{\text{cxt}}$ causes the degeneration, we drop the deterministic $\mathbf{h}_t^{\text{cxt}}$ and condition the decoder only on the stochastic utterance latent variable $\mathbf{z}^{\text{utt}}$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}) = f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{z}_t^{\text{utt}}) \quad (13)$$

While this model achieves higher values of KL divergence than original VHRED, as training proceeds it again degenerates with the KL divergence term reaching zero (Fig. 2).

To gain an insight of the degeneracy, we examine how the conditional prior $p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t})$ (Eq. 5) of the utterance latent variable changes during training, using the model above (Eq. 13). Fig. 2 plots the ratios of $\mathbb{E}[\boldsymbol{\sigma}_t^2]/\text{Var}(\boldsymbol{\mu}_t)$, where $\mathbb{E}[\boldsymbol{\sigma}_t^2]$ indicates the *within variance* of the priors, and $\text{Var}(\boldsymbol{\mu}_t)$ is the *between variance* of the priors. Note that traditionally this ratio is closely related to *Analysis of Variance (ANOVA)* (Lomax and Hahs-Vaughn, 2013). The ratio gradually falls to zero, implying that the priors degenerate to separate point masses as training proceeds. Moreover, we find that the degeneracy of priors coincide with the degeneracy of KL divergence, as shown in (Fig. 2). This is intuitively natural: if the prior is already narrow enough to specify the target utterance, there is little pressure to encode any more information in the variational posterior for reconstruction of the target utterance.

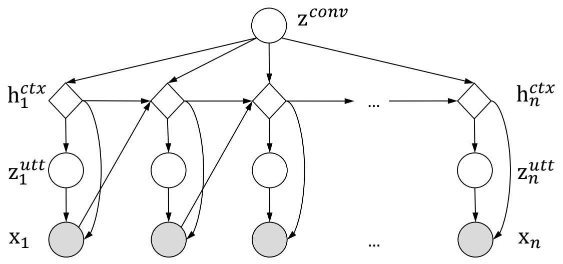

Figure 3: Graphical representation of the Variational Hierarchical Conversation RNN (VHCR). The global latent variable $\mathbf{z}^{\text{conv}}$ provides a global context in which the conversation takes place.

This empirical observation implies that the fundamental reason behind the degeneration may originate from combination of two factors: (1) strong expressive power of the hierarchical RNN decoder and (2) training data sparsity caused by the conditional VAE structure. The VHRED is trained to predict a next target utterance $\mathbf{x}_t$ conditioned on the context $\mathbf{h}_t^{\text{cxt}}$ which encodes information about previous utterances $\{\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}\}$. However, conditioning on the context makes the range of training target $\mathbf{x}_t$ very sparse; even in a large-scale conversation corpus such as Ubuntu Dialog (Lowe et al., 2015), there exist one or very few target utterances per context. Therefore, hierarchical RNNs, given their autoregressive power, can easily overfit to training data without using the latent variable. Consequently, the VHRED will not encode any information in the latent variable, i.e. it degenerates. It explains why the word drop fails to prevent the degeneracy in the VHRED. The word drop only regularizes the decoder RNN; however, the context RNN is also powerful enough to predict a next utterance in a given context even with the weakened decoder RNN. Indeed we observe that using a larger word drop probability such as 0.5 or 0.75 only slows down, but fails to stop the KL divergence from vanishing.

3.5 Variational Hierarchical Conversation RNN (VHCR)

As discussed, we argue that the two main causes of degeneration are i) the expressiveness of the hierarchical RNN decoders, and ii) the conditional VAE structure that induces data sparsity. This finding hints us that in order to train a non-degenerate latent variable model, we need to design a model that provides an appropriate way to regularize the hierarchical RNN decoders and alleviate data sparsity per context. At the same time, the model should be capable of modeling complex structure of conversation. Based on these insights, we propose a novel VAE structure named Variational Hierarchical Conversation RNN (VHCR), whose graphical model is illustrated in Fig. 3. Below we first describe the model, and discuss its unique features.

We introduce a global conversation latent variable $\mathbf{z}^{\text{conv}}$ which is responsible for generating a sequence of utterances of a conversation $\mathbf{c} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$:

$$p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z}^{\text{conv}}) = p_{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{z}^{\text{conv}}) \tag{14}$$

Overall, the VHCR builds upon the hierarchical RNNs, following the VHRED (Serban et al., 2017). One key update is to form a hierarchical latent structure, by using the global latent variable $\mathbf{z}^{\text{conv}}$ per conversation, along with local the latent variable $\mathbf{z}_t^{\text{utt}}$ injected at each utterance (Fig. 3):

$$\begin{aligned}
\mathbf{h}_t^{\text{enc}} &= f_{\boldsymbol{\theta}}^{\text{enc}}(\mathbf{x}_t) && (15)\\
\mathbf{h}_t^{\text{cxt}} &= \begin{cases} \text{MLP}_{\boldsymbol{\theta}}(\mathbf{z}^{\text{conv}}), & \text{if } t = 0 \\ f_{\boldsymbol{\theta}}^{\text{cxt}}(\mathbf{h}_{t-1}^{\text{cxt}}, \mathbf{h}_{t-1}^{\text{enc}}, \mathbf{z}^{\text{conv}}), & \text{otherwise} \end{cases} \\
p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_t^{\text{utt}}, \mathbf{z}^{\text{conv}}) &= f_{\boldsymbol{\theta}}^{\text{dec}}(\mathbf{x}|\mathbf{h}_t^{\text{cxt}}, \mathbf{z}_t^{\text{utt}}, \mathbf{z}^{\text{conv}}) \\
p_{\boldsymbol{\theta}}(\mathbf{z}^{\text{conv}}) &= \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) && (16)\\
p_{\boldsymbol{\theta}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_{<t}, \mathbf{z}^{\text{conv}}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\mathbf{I}) && (17)\\
\text{where } \boldsymbol{\mu}_t &= \text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}, \mathbf{z}^{\text{conv}}) && (18)\\
\boldsymbol{\sigma}_t &= \text{Softplus}(\text{MLP}_{\boldsymbol{\theta}}(\mathbf{h}_t^{\text{cxt}}, \mathbf{z}^{\text{conv}})). && (19)
\end{aligned}$$

For inference of $\mathbf{z}^{\text{conv}}$, we use a bi-directional RNN denoted by f^{conv}, which runs over the utterance vectors generated by the encoder RNN:

$$\begin{aligned}
q_{\boldsymbol{\phi}}(\mathbf{z}^{\text{conv}}|\mathbf{x}_1, ..., \mathbf{x}_n) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}^{\text{conv}}, \boldsymbol{\sigma}^{\text{conv}} I) && (20)\\
\text{where } \mathbf{h}^{\text{conv}} &= f^{\text{conv}}(\mathbf{h}_1^{\text{enc}}, ..., \mathbf{h}_n^{\text{enc}}) && (21)\\
\boldsymbol{\mu}^{\text{conv}} &= \text{MLP}_{\boldsymbol{\phi}}(\mathbf{h}^{\text{conv}}) && (22)\\
\boldsymbol{\sigma}^{\text{conv}} &= \text{Softplus}(\text{MLP}_{\boldsymbol{\phi}}(\mathbf{h}^{\text{conv}})). && (23)
\end{aligned}$$

The posteriors for local variables $\mathbf{z}_t^{\text{utt}}$ are then conditioned on $\mathbf{z}^{\text{conv}}$:

$$\begin{aligned}
q_{\boldsymbol{\phi}}(\mathbf{z}_t^{\text{utt}}|\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{z}^{\text{conv}}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}'_t, \boldsymbol{\sigma}'_t I) && (24)\\
\text{where } \boldsymbol{\mu}'_t &= \text{MLP}_{\boldsymbol{\phi}}(\mathbf{x}_t, \mathbf{h}_t^{\text{cxt}}, \mathbf{z}^{\text{conv}}) && (25)\\
\boldsymbol{\sigma}'_t &= \text{Softplus}(\text{MLP}_{\boldsymbol{\phi}}(\mathbf{x}_t, \mathbf{h}_t^{\text{cxt}}, \mathbf{z}^{\text{conv}})).
\end{aligned}$$

Our solution of VHCR to the degeneration problem is based on two ideas. The first idea is to build a hierarchical latent structure of $\mathbf{z}^{\text{conv}}$ for

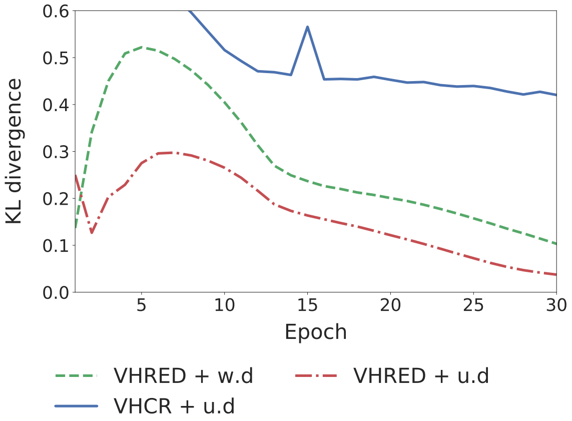

Figure 4: The comparison of KL divergences. The VHCR with the utterance drop shows high and stable KL divergence, indicating the active use of latent variables. w.d and u.d denote the word drop and the utterance drop, respectively.

a conversation and $\mathbf{z}_t^{\text{utt}}$ for each utterance. As $\mathbf{z}^{\text{conv}}$ is independent of the conditional structure, it does not suffer from the data sparsity problem. However, the expressive power of hierarchical RNN decoders makes the model still prone to ignore latent variables $\mathbf{z}^{\text{conv}}$ and $\mathbf{z}_t^{\text{utt}}$. Therefore, our second idea is to apply an *utterance drop* regularization to effectively regularize the hierarchical RNNs, in order to facilitate the use of latent variables. That is, at each time step, the utterance encoder vector $\mathbf{h}_t^{\text{enc}}$ is randomly replaced with a generic unknown vector $\mathbf{h}^{\text{unk}}$ with a probability p. This regularization weakens the autoregressive power of hierarchical RNNs and as well alleviates the data sparsity problem, since it induces noise into the context vector $\mathbf{h}_t^{\text{cxt}}$ which conditions the decoder RNN. The difference with the word drop (Bowman et al., 2016) is that our utterance drop depresses the hierarchical RNN decoders as a whole, while the word drop only weakens the lower-level decoder RNNs. Fig. 4 confirms that with the utterance drop with a probability of 0.25, the VHCR effectively learns to use latent variables, achieving a significant degree of KL divergence.

3.6 Effectiveness of Hierarchical Latent Structure

Is the hierarchical latent structure of the VHCR crucial for effective utilization of latent variables? We investigate this question by applying the utterance drop on the VHRED which lacks any hierarchical latent structure. We observe that the KL divergence still vanishes (Fig. 4), even though the utterance drop injects considerable noise in the context $\mathbf{h}_t^{\text{cxt}}$. We argue that the utterance drop weakens the context RNN, thus it consequently fail to predict a reasonable prior distribution for $\mathbf{z}^{\text{utt}}$ (Eq. 5-7). If the prior is far away from the region of $\mathbf{z}^{\text{utt}}$ that can generate a correct target utterance, encoding information about the target in the variational posterior will incur a large KL divergence penalty. If the penalty outweighs the gain of the reconstruction term in Eq. 2, then the model would learn to ignore $\mathbf{z}^{\text{utt}}$, in order to maximize the variational lower-bound in Eq. 2.

On the other hand, the global variable $\mathbf{z}^{\text{conv}}$ allows the VHCR to predict a reasonable prior for local variable $\mathbf{z}_t^{\text{utt}}$ even in the presence of the utterance drop regularization. That is, $\mathbf{z}^{\text{conv}}$ can act as a *guide* for $\mathbf{z}^{\text{utt}}$ by encoding the information for local variables. This reduces the KL divergence penalty induced by encoding information in $\mathbf{z}^{\text{utt}}$ to an affordable degree at the cost of KL divergence caused by using $\mathbf{z}^{\text{conv}}$. This trade-off is indeed a fundamental strength of hierarchical models that provide *parsimonious* representation; if there exists any shared information among the local variables, it is coded in the global latent variable reducing the code length by effectively reusing the information. The remaining local variability is handled properly by the decoding distribution and local latent variables.

The global variable $\mathbf{z}^{\text{conv}}$ provides other benefits by representing a latent global structure of a conversation, such as a topic, a length, and a tone of the conversation. Moreover, it allows us to control such global properties, which is impossible for models without hierarchical latent structure.

4 Results

We first describe our experimental setting, such as datasets and baselines (section 4.1). We then report quantitative comparisons using three different metrics (section 4.2–4.4). Finally, we present qualitative analyses, including several utterance control tasks that are enabled by the hierarchal latent structure of our VHCR (section 4.5). We defer implementation details and additional experiment results to the supplementary file.

4.1 Experimental Setting

Datasets. We evaluate the performance of conversation generation using two benchmark datasets: 1) Cornell Movie Dialog Corpus (Danescu-

Model	NLL	Recon.	KL div.
HRED	3.873	-	-
VHRED	≤ 3.912	3.619	0.293
VHRED + w.d	≤ 3.904	3.553	0.351
VHRED + bow	≤ 4.149	2.982	1.167
VHCR + u.d	≤ 4.026	3.523	0.503

(a) Cornell Movie Dialog

Model	NLL	Recon.	KL div.
HRED	3.766	-	-
VHRED	≤ 3.767	3.654	0.113
VHRED + w.d	≤ 3.824	3.363	0.461
VHRED + bow	≤ 4.237	2.215	2.022
VHCR + u.d	≤ 3.951	3.205	0.756

(b) Ubuntu Dialog

Table 1: Results of Negative Log-likelihood. The inequalities denote the variational bounds. w.d and u.d., and bow denote the word drop, the utterance drop, and the auxiliary bag-of-words loss respectively.

	Cornell			Ubuntu		
Model	Total	$\mathbf{z}^{\text{conv}}$	$\mathbf{z}^{\text{utt}}$	Total	$\mathbf{z}^{\text{conv}}$	$\mathbf{z}^{\text{utt}}$
VHRED	0.351	-	0.351	0.461	-	0.461
VHCR	0.503	0.189	0.314	0.756	0.198	0.558

Table 2: KL divergence decomposition. VHRED and VHCR are trained with word drop and utterance drop respectively.

Niculescu-Mizil and Lee, 2011), containing 220,579 conversations from 617 movies. 2) Ubuntu Dialog Corpus (Lowe et al., 2015), containing about 1 million multi-turn conversations from Ubuntu IRC channels. In both datasets, we truncate utterances longer than 30 words.

Baselines. We compare our approach with four baselines. They are combinations of two state-of-the-art models of conversation generation with different solutions to the degeneracy. (i) Hierarchical recurrent encoder-decoder (HRED) (Serban et al., 2016), (ii) Variational HRED (VHRED) (Serban et al., 2017), (iii) VHRED with the word drop (Bowman et al., 2016), and (iv) VHRED with the bag-of-words (bow) loss (Zhao et al., 2017).

Performance Measures. Automatic evaluation of conversational systems is still a challenging problem (Liu et al., 2016). Based on literature, we report three quantitative metrics: i) the negative log-likelihood (the variational bound for variational models), ii) embedding-based metrics (Serban et al., 2017), and iii) human evaluation via Amazon Mechanical Turk (AMT).

4.2 Results of Negative Log-likelihood

Table 1 summarizes the per-word negative log-likelihood (NLL) evaluated on the test sets of two datasets. For variational models, we instead present the variational bound of the negative log-likelihood in Eq. 2, which consists of the reconstruction error term and the KL divergence term. The KL divergence term can measure how much each model utilizes the latent variables.

We observe that the NLL is the lowest by the HRED. Variational models show higher NLLs, because they are regularized methods that are forced to rely more on latent variables. Independent of NLL values, we later show that the latent variable models often show better generalization performance in terms of embedding-based metrics and human evaluation. In the VHRED, the KL divergence term gradually vanishes even with the word drop regularization; thus, early stopping is necessary to obtain a meaningful KL divergence. The VHRED with the bag-of-words loss (bow) achieves the highest KL divergence, however, at the cost of high NLL values. That is, the variational lower-bound minimizes the minimum description length, to which the bow loss works in an opposite direction by forcing latent variables to encode bag-of-words representation of utterances. Our VHCR achieves stable KL divergence without any auxiliary objective, and the NLL is lower than the VHRED + bow model.

Table 2 summarizes how global and latent variable are used in the VHCR. We observe that VHCR encodes a significant amount of information in the global variable $\mathbf{z}^{\text{conv}}$ as well as in the local variable $\mathbf{z}^{\text{utt}}$, indicating that the VHCR successfully exploits its hierarchical latent structure.

4.3 Results of Embedding-Based Metrics

The embedding-based metrics (Serban et al., 2017; Rus and Lintean, 2012) measure the textual similarity between the words in the model response and the ground truth. We represent words using Word2Vec embeddings trained on the Google News Corpus[1]. The *average* metric projects each utterance to a vector by taking the mean over word embeddings in the utterance, and computes the cosine similarity between the model response vector and the ground truth vector. The *extrema* metric is similar to the average metric, only except that it takes the extremum of each di-

[1] https://code.google.com/archive/p/word2vec/.

Model	Average	Extrema	Greedy
1-turn			
HRED	0.541	0.370	0.387
VHRED	0.543	0.356	0.393
VHRED + w.d	0.554	0.365	0.404
VHRED + bow	0.555	0.350	0.411
VHCR + u.d	**0.585**	**0.376**	**0.434**
3-turn			
HRED	0.556	0.372	0.395
VHRED	0.554	0.360	0.398
VHRED + w.d	0.566	0.369	0.408
VHRED + bow	0.573	0.360	0.423
VHCR + u.d	**0.588**	**0.378**	**0.429**

(a) Cornell Movie Dialog

Model	Average	Extrema	Greedy
1-turn			
HRED	0.567	**0.337**	0.412
VHRED	0.547	0.322	0.398
VHRED + w.d	0.545	0.314	0.398
VHRED + bow	0.545	0.306	0.398
VHCR + u.d	**0.570**	0.312	**0.425**
3-turn			
HRED	0.559	**0.324**	0.402
VHRED	0.551	0.315	0.397
VHRED + w.d	0.551	0.309	0.399
VHRED + bow	0.552	0.303	0.398
VHCR + u.d	**0.574**	0.311	**0.422**

(b) Ubuntu Dialog

Table 3: Results of embedding-based metrics. 1-turn and 3-turn responses of models per context.

mension, instead of the mean. The *greedy* metric first finds the best non-exclusive word alignment between the model response and the ground truth, and then computes the mean over the cosine similarity between the aligned words.

Table 3 compares the different methods with three embedding-based metrics. Each model generates a single response (1-turn) or consecutive three responses (3-turn) for a given context. For 3-turn cases, we report the average of metrics measured for three turns. We use the greedy decoding for all the models.

Our VHCR achieves the best results in most metrics. The HRED is the worst on the Cornell Movie dataset, but outperforms the VHRED and VHRED + bow on the Ubuntu Dialog dataset. Although the VHRED + bow shows the highest KL divergence, its performance is similar to that of VHRED, and worse than that of the VHCR model. It suggests that a higher KL divergence does not necessarily lead to better performance; it is more important for the models to balance the modeling powers of the decoder and the latent variables. The VHCR uses a more sophisticated hierarchical latent structure, which better reflects the structure of natural language conversations.

4.4 Results of Human Evaluation

Table 4 reports human evaluation results via Amazon Mechanical Turk (AMT). The VHCR outperforms the baselines in both datasets; yet the performance improvement in Cornell Movie Dialog are less significant compared to that of Ubuntu. We empirically find that Cornell Movie dataset is small in size, but very diverse and complex in content and style, and the models often fail to generate sensible responses for the context. The performance gap with the HRED is the smallest, suggesting that the VAE models without hierarchical latent structure have overfitted to Cornell Movie dataset.

4.5 Qualitative Analyses

Comparison of Predicted Responses. Table 5 compares the generated responses of algorithms. Overall, the VHCR creates more consistent responses within the context of a given conversation. This is supposedly due to the global latent variable $\mathbf{z}^{\text{conv}}$ that provides a more direct and effective way to handle the global context of a conversation. The context RNN of the baseline models can handle long-term context to some extent, but not as much as the VHCR.

Interpolation on $\mathbf{z}^{\text{conv}}$. We present examples of one advantage by the hierarchical latent structure of the VHCR, which cannot be done by the other existing models. Table 6 shows how the generated responses vary according to the interpolation on $\mathbf{z}^{\text{conv}}$. We randomly sample two $\mathbf{z}^{\text{conv}}$ from a standard Gaussian prior as references (*i.e.* the top and the bottom row of Table 6), and interpolate points between them. We generate 3-turn conversations conditioned on given $\mathbf{z}^{\text{conv}}$. We see that $\mathbf{z}^{\text{conv}}$ controls the overall tone and content of conversations; for example, the tone of the response is friendly in the first sample, but gradually becomes hostile as $\mathbf{z}^{\text{conv}}$ changes.

Generation on a Fixed $\mathbf{z}^{\text{conv}}$. We also study how fixing a global conversation latent variable $\mathbf{z}^{\text{conv}}$ affects the conversation generation. Table 7 shows an example, where we randomly fix a reference $\mathbf{z}^{\text{conv}}$ from the prior, and generate multiple examples of 3-turn conversation using randomly sampled local variables $\mathbf{z}^{\text{utt}}$. We observe that $\mathbf{z}^{\text{conv}}$ heavily affects the form of the first utterance; in the examples, the first utterances all start with a "where" phrase. At the same time, responses show

	Cornell			Ubuntu		
Opponent	Wins	Losses	Ties	Wins	Losses	Ties
VHCR vs HRED	$\mathbf{28.5 \pm 1.9}$	28.2 ± 1.9	43.3 ± 2.1	$\mathbf{52.9 \pm 2.1}$	42.2 ± 2.1	4.9 ± 0.9
VHCR vs VHRED + w.d	$\mathbf{29.9 \pm 1.9}$	28.0 ± 1.9	42.1 ± 2.1	$\mathbf{48.1 \pm 2.1}$	40.1 ± 3.6	11.9 ± 1.4
VHCR vs VHRED + bow	$\mathbf{31.3 \pm 2.0}$	26.9 ± 1.9	41.7 ± 2.1	$\mathbf{46.1 \pm 2.1}$	39.9 ± 2.1	14.0 ± 1.5

Table 4: Results of human evaluation via AMT. Human turkers are asked to choose which response is more appropriate in a given context, without knowing which algorithms generate which responses. For each pair of models, we carry out three evaluation batches, each of which consists of 100 random test samples evaluated by five unique humans. We report mean preferences with ±90% confidence interval.

Context	HRED	VHRED	VHRED+bow	VHCR
i'm going home. → what's wrong? → i don't know. one of these days, i'm gon na see louis. my back's killing me.	i don't know.	i don't want to be here. i'm going to be here.	are you going to go now?	this isn't a joke. you're going to be here.
how's sleeping beauty? → still sleeping. → kiss him, maybe he'll turn into a frog.	he's a good man.	what?	we won't?	maybe he's dead.
when i mount my usb drive i get a permission denied but i can easily see the files if i boot via live cd any ideas on how i → the file permissions of the ext3 file system are likely preventing you → thanks, any idea on options to fix?	no, i don't know how to fix it, but i'm not sure how to fix it	what kind of filesystem is it on?	you did it just first? then you'll need to reboot and then remount it up again, if your problem persists, i'd give you the	the only thing i can think of is to fix the permissions issue, try mounting the drive with the uid and gid options in the forum
hello i need some support, anybody there? → go ahead with your questions. <unk> are not expected on irc. → i have an hp pavilion <unk> here and i can't get it to install ubuntu	what's the problem?	<unk>. what's the problem?	i'm sorry.	they should be supported by canonical. you don't need to use them.

Table 5: Qualitative comparison of generated responses. Top two rows show the samples from Cornell Movie Dialog, while the bottom two rows are from Ubuntu Dialog.

nice to meet you. **→ where's my wife?** **→ she's asleep.**
go on, frank. → i ll tell you what i'm doing. i'm a little tired. → why?
now, you're not going. → not for a long time. → but you're a little <unk>.
get out of here. → all right. → you want to go home?
get out of here!!! **→ it's not your fault, is n't it?** **→ why? what's wrong?**

Table 6: An example of interpolated 3-turn responses over $\mathbf{z}^{\text{conv}}$ on Cornell Movie Dialog.

where is she? → she's the only one who knows where she is, she's going to be all right. → oh, you're the only one who's gon na be. she's a <unk>.
where's my wife? → you've got to get out of here, you know? you're the one who's gon na be here. → oh, that's nice.
where are you? → well, i was just thinking about you and i know what you're doing. i'm going to have to go to the <unk> and i'm → i'm sorry.
where are you going? → to get you to the airport. → you're going to be late?
where are you going? → to the <unk>. i am not going to tell you what i am. i am the only one who has to be. i will be the → you've got to stop!

Table 7: An example of 3-turn responses conditioned on sampled $\mathbf{z}^{\text{utt}}$ for a single fixed $\mathbf{z}^{\text{conv}}$.

variations according to different local variables $\mathbf{z}^{\text{utt}}$. These examples show that the hierarchical latent structure of VHCR allows both global and fine-grained control over generated conversations.

5 Discussion

We introduced the variational hierarchical conversation RNN (VHCR) for conversation modeling. We noted that the degeneration problem in existing VAE models such as the VHRED is persistent, and proposed a hierarchical latent variable model with the utterance drop regularization. Our VHCR obtained higher and more stable KL divergences than various versions of VHRED models without using any auxiliary objective. The empirical results showed that the VHCR better reflected the structure of natural conversations, and outperformed previous models. Moreover, the hierarchical latent structure allowed both global and fine-grained control over the conversation generation.

Acknowledgments

This work was supported by Kakao and Kakao Brain corporations, and Creative-Pioneering Researchers Program through Seoul National University. Gunhee Kim is the corresponding author.

References

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *CoNLL*. https://doi.org/10.18653/v1/K16-1002.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2017. Variational lossy autoencoder. In *ICLR*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *CMCL Workshop*.

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

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547* .

Chia-Wei Liu, Ryan Lowe, Iulian V 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*.

Richard G Lomax and Debbie L Hahs-Vaughn. 2013. *Statistical concepts: A second course*. Routledge.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Building Educational Applications Using NLP Workshop*. ACL. http://www.aclweb.org/anthology/W12-2018.

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

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*. https://doi.org/10.3115/v1/P15-1152.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015a. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015b. A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*. https://doi.org/10.3115/v1/N15-1020.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *EMNLP*. https://doi.org/10.18653/v1/D16-1050.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*. https://doi.org/10.18653/v1/P17-1061.