# Focus-Constrained Attention Mechanism for CVAE-based Response Generation

**Zhi Cui**[1], **Yanran Li**[2], **Jiayi Zhang**[1], **Jianwei Cui**[1], **Chen Wei**[1], **Bin Wang**[1]

[1]Xiaomi AI Lab
[2]The Hong Kong Polytechnic University
{cuizhi,zhangjiayi3,cuijianwei,weichen,wangbin11}@xiaomi.com
csyli@comp.polyu.edu.hk

## Abstract

To model diverse responses for a given post, one promising way is to introduce a latent variable into Seq2Seq models. The latent variable is supposed to capture the discourse-level information and encourage the informativeness of target responses. However, such discourse-level information is often too coarse for the decoder to be utilized. To tackle it, our idea is to transform the coarse-grained discourse-level information into fine-grained word-level information. Specifically, we firstly measure the semantic concentration of corresponding target response on the post words by introducing a fine-grained *focus* signal. Then, we propose a focus-constrained attention mechanism to take full advantage of *focus* in well aligning the input to the target response. The experimental results demonstrate that by exploiting the fine-grained signal, our model can generate more diverse and informative responses compared with several state-of-the-art models.[1]

## 1 Introduction

Nowadays, developing intelligent open-domain conversational systems has become an active research field (Perez-Marin and Pascual-Nieto, 2011; Shum et al., 2018). Compared with rule-based and retrieval-based methods, neural generative models have attracted increasing attention because they do not need extensive feature engineering and have achieved promising results recently with large-scale conversational data (Vinyals and Le, 2015; Sordoni et al., 2015; Shang et al., 2015).

Typically, neural generative models are trained to learn the post-response mappings based on the Seq2Seq architecture using maximum likelihood (MLE) training objective. This kind of objective induces the model to treat the post-response relationship as one-to-one mappings. However, the conversations in the real world often embodies one-to-many relationships, where a post is often associated with multiple valid responses (Zhou et al., 2017). Due to this discrepancy, standard Seq2Seq models tend to generate high-frequency but trivial responses such as "*I don't know*" or "*I'm ok*" (Li et al., 2016).

To address this issue, one promising research line resorts to Conditional Variational Autoencoder (CVAE), which introduces a latent variable to Seq2Seq models through variational learning. The latent variable is supposed to capture the discourse-level semantics of target response and in turn encourage the response informativeness. Recent literature along this line attempted to improve the model performance by putting extra control on the latent variable (Zhao et al., 2017; Gu et al., 2018; Gao et al., 2019). Despite the control, these methods still relied on the discourse-level latent variable, which is too coarse for the decoders to mine sufficient guiding signals at each generation step. As a result, these variational models are observed to ignore the latent variable (Zhao et al., 2017; Gu et al., 2018; Gao et al., 2019) and to generate semantically irrelevant or grammatically disfluent responses (Qiu et al., 2019).

In this paper, we propose a novel CVAE-based model, which exploits fine-grained word-level information for diverse response generation. Firstly, we transform the discourse-level information into word-level signals, i.e., *focus*. By attending the latent variable to the post words, the *focus* weight measures the response's correlation with the post

[1] https://github.com/cuizhi555/Focus-Constrained-Attention-Mechanism-for-CVAE-based-Response-Generation.

*Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2021–2030
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

words. The higher the weight, the semantics is more likely to concentrate on the corresponding word. To utilize the *focus*, we develop a focus-constrained attention mechanism which better aligns the post words with the response according to the fine-grained signals. In this way, the model is able to produce a semantically different response directed by a different *focus*.

Our contributions can be summarized as three folds: 1) We propose a novel CVAE-based model for diverse response generation, by directing the decoder with fine-grained information. 2) We introduce *focus* to represent the fine-grained information, and propose a focus-constrained attention mechanism to make full use of it. 3). Experimental results demonstrate our model outperforms several state-of-the-art models in terms of response's diversity as well as appropriateness.

## 2 Related Work

The attention mechanism (Bahdanau et al., 2014; Luong et al., 2015) has become a widely-used component for Seq2Seq (Sutskever et al., 2014; Cho et al., 2014) to model Short-Text Conversation (Shang et al., 2015; Vinyals and Le, 2015; Sordoni et al., 2015). Although promising results have been achieved, attention-based Seq2Seq models still tend to generate generic and trivial responses (Li et al., 2016).

There have been many approaches attempted to address this problem. Li et al. (2016) reranked the n-best generated responses based on Maximum Mutual Information (MMI). Shao et al. (2017) adopted segement-level reranking to encourage diversity during early decoding steps. However, these reranking-based methods only introduce a few variants of decoded words. Another group of researches attempted to encourage diversity by incorporating extra information. Xing et al. (2017) injected topic words and Yao et al. (2017) introduced a cue word based on Point-wise Mutual Information (PMI) into generation models. Ghazvininejad et al. (2018) grounded on knowledge bases to provide factual information for the decoder. However, it is difficult to ensure these external information are always appropriate to the conversation context.

Another line of research introduced a set of latent responding mechanisms and generated responses based on a selected mechanism. Zhou et al. (2017) learned the post-response mappings as a mixture of the mechanisms, but it is questionable that they only relied on one single mechanism when generating responses given a new post. Chen et al. (2019) adopted posterior selection to build one-to-one mapping relationship between the mechanisms and target responses. Since the target response is missing during testing, it is hard to ensure a satisfactory generated response by a randomly picked mechanism.

Our work centers in the research line of conditional response generation through variational learning (Serban et al., 2017; Zhao et al., 2017). However, the variational methods inevitably suffer from bypassing the latent variable and generating disfluent responses. Zhao et al. (2017) combined CVAE with dialog acts to learn meaningful latent variable, however the discourse-level dialog act is hard to be captured from short conversation. Gu et al. (2018) introduced Gaussian mixture prior network, but it is hard to determine the number of mixtures and the optimization is complicated. Gao et al. (2019) assumed the response generation is driven by a single word, and connected each latent variable with words in the vocabulary. Nevertheless, the difficulty is how to target the driving word for a specific post-response pair. More importantly, all of these methods rely on the coarse-grained discourse-level information, which might be insufficient in leading to a satisfactory response.

Notably, our work induces the response generation with *focus*, a fine-grained feature extracted from the discourse-level latent variable. Compared with the variational attention that models the alignment as latent variable (Bahuleyan et al., 2018; Deng et al., 2018), we are mainly inspired by the idea of coverage vector (Tu et al., 2016) to dynamically adjust the attention based on the attention history and the proposed *focus*. The difference is that Tu et al. (2016) addressed the under/over translation problem and the decoder in their work pays equal attention to the source words. In contrast, our work constrains the decoder to align the decoding attention with the fine-grained *focus* to generate diverse responses.

## 3 Model

### 3.1 Preliminaries and Model Overview

A neural generative model is trained on a collection of post-response pairs $\{(\mathbf{x}, \mathbf{y})\}$, and aimed to generate a response $\mathbf{y}$ word-by-word given an input $\mathbf{x}$. At the basis of our approach is CVAE where a latent variable $z$ is considered to capture

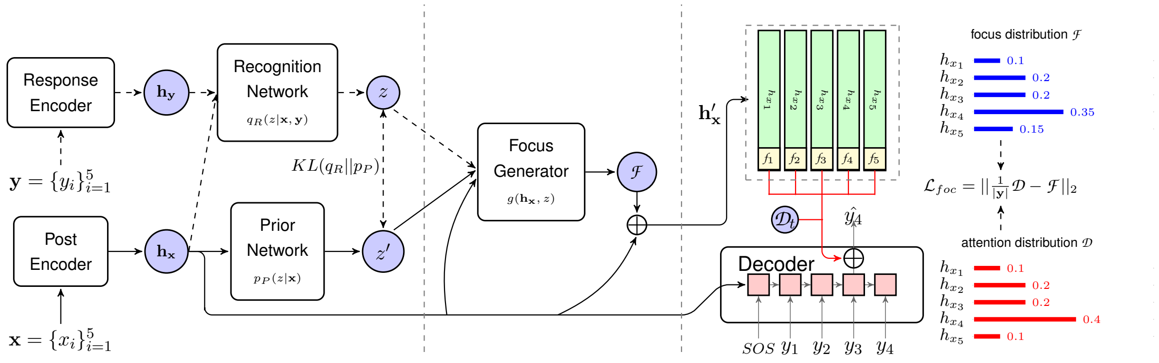

Figure 1: A framework of our proposed model, where the operation $\oplus$ denotes concatenation, the dashed arrow lines are absent during testing, and the proposed focus-constrained mechanism is represented by the red lines.

discourse-level diversity. To extract fine-grained information, we design *focus* $\mathcal{F} = \{f_i\}_{i=1}^{|\mathbf{x}|}$ over the post words, where $f_i$ measures to what extent the latent variable $z$ is concentrating on the post word $x_i$, and $|\mathbf{x}|$ is the length of input $\mathbf{x}$. Besides, we introduce a coverage vector $\mathcal{D}_t = \{d_{i,t}\}_{i=1}^{|\mathbf{x}|}$, where $d_{i,t}$ accumulates the attention weights over the post word $x_i$ up until $t$-th decoding step.

Figure 1 illustrates the whole framework of our model consisting of three components: CVAE basis, focus generator and response generator. Based on the CVAE framework, we firstly introduce a probabilistic distribution over the latent variable $z$ to model potential responses for a given $\mathbf{x}$. Then, focus generator produce the *focus* $\mathcal{F}$ by attending the latent variable $z$ to hidden representation $\mathbf{h_x}$ of the input. The obtained $\mathcal{F}$ is then concatenated with $\mathbf{h_x}$ to obtain $\mathbf{h}'_\mathbf{x}$ for word prediction. Specifically, the decoder attentively refers to $\mathbf{h}'_\mathbf{x}$ and accumulates decoding attention weights through the coverage vector $\mathcal{D}_t$. To direct response generation using the *focus* $\mathcal{F}$, we not only optimize the variational lower bound on response generation, but also optimize a regularization term named as focus constraint by minimizing the divergence $\mathcal{D}$ and $\mathcal{F}$.

### 3.2 Background of CVAE

Typically, the conditional variational autoencoder (CVAE) introduces a probabilistic distribution over the latent variable to model response diversity. Following CVAE, we firstly encode $\mathbf{x}$ and $\mathbf{y}$ by the post and response encoder, respectively. The two encoders are constructed by the shared bidirectional GRUs (Cho et al., 2014) which generate a series of hidden states $\{h_{x_i}\}_{i=1}^{|\mathbf{x}|}$ for $\mathbf{x}$ and $\{h_{y_i}\}_{i=1}^{|\mathbf{y}|}$ for $\mathbf{y}$. Then, we obtain the sentence representation $\overline{h_x}$ for the post $\mathbf{x}$ by averaging $\{h_{x_i}\}_{i=1}^{|\mathbf{x}|}$. The sentence representation $\overline{h_y}$ for the response $\mathbf{y}$ is calculated from $\{h_{y_i}\}_{i=1}^{|\mathbf{y}|}$ in the same way.

In training phase, we sample a latent variable $z$ from the posterior distribution $q_R(z|\mathbf{x}, \mathbf{y})$. The distribution is modeled as a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, where $\Sigma$ is a diagonal covariance. We parameterize $\mu$ and $\Sigma$ by the recognition network through a fully connected layer conditioned on the concatenation $[\overline{h_x}; \overline{h_y}]$:

$$\begin{bmatrix} \mu \\ \log(\Sigma) \end{bmatrix} = W_q \begin{bmatrix} \overline{h_x} \\ \overline{h_y} \end{bmatrix} + b_q \qquad (1)$$

where $W_q$ and $b_q$ are learnable parameters. To mitigate the gap in encoding of latent variables between train and testing (Sohn et al., 2015; Yan et al., 2015), CVAE requires the posterior distribution $q_R(z|\mathbf{x}, \mathbf{y})$ to be close to the prior distribution $p_P(z|\mathbf{x})$. Notably, $p_P(z|\mathbf{x})$ is parameterized by the prior network and also follows a multivariate Gaussian distribution $\mathcal{N}(\mu', \Sigma')$ in a similar way but only conditioned on $\overline{h_x}$. As usual, we minimize the discrepancy between the two distributions by the Kullback-Leibler divergence:

$$\mathcal{L}_{kl} = KL(q_R(z|\mathbf{x}, \mathbf{y})||p_P(z|\mathbf{x})) \qquad (2)$$

By sampling different $z$, the model is supposed to output semantically different responses. However, such latent variable is too coarse to guide a satisfactory response generation, as discussed previously.

### 3.3 Focus Generator

The core is how to better exploit indicative information from the discourse-level variable for diverse response generation. In this work, we transform the discourse-level latent variable $z$ into fine-grained signal using a focus generator $g(\mathbf{h_x}, z)$ as shown in the middle of Figure 1. To be specific, the focus generator attends the latent variable $z$ to the post representation $\mathbf{h_x}$, and produces the *focus* distribution $\mathcal{F} = \{\mathscr{f}\}_{i=1}^{|\mathbf{x}|}$. Similar to the standard attention (Bahdanau et al., 2014; Luong et al., 2015), the generated *focus* $\mathcal{F}$ measures the response concentration a specific post word, which is calculated by:

$$g(\mathbf{h_x}, z) = \mathcal{F} = \{\frac{\exp(f(h_{x_i}, z))}{\sum_{k=1}^{|\mathbf{x}|} \exp(f(h_{x_k}, z))}\}_{i=1}^{|\mathbf{x}|} \tag{3}$$

where $f(h_{x_i}, z) = v_f^\top \tanh(W_f h_{x_i} + U_f z)$ and $W_f$ and $U_f$ are learnable parameters. This *focus* captures to what extent the response semantics is related to the post words, which will serve as fine-grained signals for the decoder. Notably, the higher the focus, the response is more likely to pay attention to the corresponding word. Compared with the coarse-grained $z$, the word-level *focus* is of great guiding significance when generating responses.

### 3.4 Focus-Guided Generation

The remaining is to properly incorporate the fine-grained *focus* into response generation. Since the *focus* weights imply the semantics of the target response, they are beneficial signals indicating whether a word is attended properly during decoding.

Concretely, we develop a focus-guided mechanism to facilitate the decoder adjust the attention during the generation. To notify the decoder of the *focus*, we concatenate $\mathbf{h_x}$ and $\mathcal{F}$ to obtain a series of combined hiddens of the post $\mathbf{h'_x} = \{h'_{x_i}\}_{i=1}^{|\mathbf{x}|}$ (the green and yellow vectors in Figure 1). After integrating the extra feature $\mathscr{f}_i$, the devised representations $\mathbf{h'_x}$ are then used to calculate the attention weights. Inspired by Tu et al. (2016), we borrow the idea of coverage attention and introduce the coverage vector $\mathcal{D}_t = \{\mathscr{d}_{i,t}\}_{i=1}^{|\mathbf{x}|}$ that records the attention history, where $\mathscr{d}_{i,t} = \sum_{k=1}^{t} \alpha_{i,k}$ accumulates the decoding attention weights on the post word $x_i$. Here, $\alpha_{i,t}$ stands for the attention weight on the post word $x_i$ at $t$-th decoding step ($t \in [1, |\mathbf{y}|]$), which is calculated as:

$$\alpha_{i,t} = \frac{\exp(e_{i,t})}{\sum_{k=1}^{|\mathbf{x}|} \exp(e_{k,t})} \tag{4}$$

$$e_{i,t} = v_a^\top \tanh(W_a h'_{x_i} + U_a s_{t-1} + V_a \mathbf{a}_{t-1}) \tag{5}$$

$$\mathbf{a}_{t-1} = \sum_{i=1}^{|\mathbf{x}|} \mathscr{d}_{i,t-1} h'_{x_i} \tag{6}$$

where $s_{t-1}$ is decoder's hidden state at $(t-1)$-th step, and $\mathbf{a}_{t-1}$ takes into account the attention history before $t$-th step. At the end of each decoding step, a predicted word $\hat{y}_t$ is obtained by:

$$\hat{y}_t = \mathrm{softmax}(W_d[s_t; \sum_{i=1}^{|\mathbf{x}|} \alpha_{i,t} h_{x_i}] + d_d) \tag{7}$$

where $W_d$ and $d_d$ are learnable parameters. Since the *focus* suggests how much attention should be paid to during each decoding step, the devised focus-guided mechanism is able to globally determine a word based on the attention history as well as the current state.

### 3.5 Focus Constraint

Nevertheless, one potential drawback is that the decoder could still ignore the *focus* signals even equipped with the focus-guided mechanism. Considering that *focus* measures the response's significance on a specific post word, it is also essential for the decoder to concentrate on the word with higher *focus* weight, and vice versa.

To prevent the decoder bypassing the *focus* signal, we design a focus constraint to regulate the learning of post-response pairs by taking into account the *focus* weights. As shown in the right side of Figure 1, the focus constraint requires the model to minimize the discrepancy between the focus weight distribution $\mathcal{F}$ and decoding attention distribution $\mathcal{D}$. To implement it, we define the focus constraint $\mathcal{L}_{foc}$ as the Euclidean norm distance between $\mathcal{D}$ and $\mathcal{F}$:

$$\mathcal{L}_{foc} = ||\frac{1}{|\mathbf{y}|}\mathcal{D} - \mathcal{F}||_2 \tag{8}$$

where $\mathcal{D}$ sums up all the decoding attention weights over the post words and $|\mathbf{y}|$ is the total number of decoding steps. Considering $\sum_{i=1}^{|\mathbf{x}|} \mathscr{f}_i = 1$, a division of $|\mathbf{y}|$ from $\mathcal{D}$ makes the two terms being compared at the same magnitude. We name this constrained decoding attention as **Focus-Constrained Attention Mechanism**. Such a constraint will make the decoder draw attention by globally consulting the *focus* $\mathcal{F}$ and distribute the attention dynamically. For example, given a distribution $\mathcal{F}$, if the hidden output $h_{x_i}$ has been attended to a certain degree $d_{i,t-1} \approx f_i$, the model will discourage the decoder to overly emphasize on $h_{x_i}$ after the $t$-th step. In contrast, if the hidden output $h_{x_i}$ has been hardly attended compared with its *focus* weight ($d_{i,t-1} \ll f_i$), the model will encourage the decoder to pay more attention onto $h_{x_i}$ afterwards.

### 3.6 Optimization and Testing

Overall, all the modules described above are jointly trained in an end-to-end way by minimizing the total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{seq} + \mathcal{L}_{foc} + \gamma \mathcal{L}_{kl} + \mathcal{L}_{bow} \qquad (9)$$

Here, $\mathcal{L}_{seq}$ is the sequence cross entropy between the generated response $\hat{\mathbf{y}}$ and the corresponding ground truth $\mathbf{y}$. $\mathcal{L}_{foc}$ is the proposed focus constraint as described above. To address the problem of vanishing latent variable, we follow Bowman et al. (2015) and adopt the annealing weight $\gamma$ for KL divergence $\mathcal{L}_{kl}$, where $\gamma$ is gradually increased during training phase. We also employ the auxiliary bag-of-word loss $\mathcal{L}_{bow}$ to further alleviate the vanishing issue (Zhao et al., 2017).

At testing phase, an intermediate *focus* $\mathcal{F}$ will be obtained with the prior network and focus generator. Notably, this enables us to generate diverse responses by sampling multiple latent variables from the prior network, where each sampled $z$ leads to a semantically distinct response.

## 4 Experiment

### 4.1 Dataset

We conduct experiments on the Weibo benchmark[2] (Shang et al., 2015), a single-round conversational dataset where a post is associated with multiple responses. We follow the default preprocessing step, and obtain 205,164 unique posts and 4,142,299 training post-response pairs in total. After random spilt, we acquire 101,794 post-response pairs for evaluation, and 1,000 distinct posts for testing. Here, each testing post has 5 reference responses for evaluation.

### 4.2 Implementation Details

We implement our model with Tensorflow and run it on NVIDIA Telsa V100. Specifically, the vocabulary size is 50,003 including PAD, UNK and EOS. The word embedding size is 720 as same as the size of latent variable. We build two-layer GRUs for the two parameter-shared encoders as well as for the decoder. In all, our model contains around 130M parameters, which are all randomly initialized with a uniform distribution $[-1, 1]$. We train our model with a batch size of 1,024 by Adam optimizer (Kingma and Ba, 2014). We increase the learning rate from 0 up to 0.0008 within the first 8,000 warmup steps and proportionally decrease it to the inverse square root of step number (Vaswani et al., 2017).

### 4.3 Baseline Models

To demonstrate the necessity and effectiveness of our proposed mechanism alone, we build it on Seq2Seq and exclude as many other interferences as possible when comparing with the following state-of-the-art baseline models:

**S2S** (Bahdanau et al., 2014): It trains a Seq2Seq model with the standard attention and adopts beam search decoding to generate responses.

**MMI** (Li et al., 2016): It is a backward Seq2Seq trained from response to post, and reranks the beam searched candidates under MMI criterion.

**MARM** (Zhou et al., 2017): It is a Seq2Seq model which additionally contains a diverter that consists of 5 latent responding mechanisms. During training, these mechanisms are learned as a mixture by the weighted average.

**CMHAM** (Tao et al., 2018): It is a Seq2Seq model, which is augmented with Constrained-Multi-Head-Attention-Mechanism. The attention heads are constrained by orthogonality and each of them is expected to attend a certain aspect of the post. We set the head number as 5.

**CVAE** (Zhao et al., 2017): It is a vanilla CVAE Seq2Seq trained along with the bag-of-word loss. During testing phase, we take 3 samplings from the prior network to generate each response.

**DCVAE** (Gao et al., 2019): It is a CVAE-based Seq2Seq model trained with discrete latent variables, where the latent variables are connected with words in the vocabulary. To follow their paper, we use their original implementation and pre-train the model with extracted keywords. During testing phase, we adopt their two-stage sampling strategy to generate each response.

**Ours**: In addition, we implement two variants of our proposed model **Ours-FocConstrain**, i.e., 1) **Ours-Foc** introduces the *focus* $\mathcal{F}$, but it does not incorporate the coverage vector $\mathcal{D}_t$, and the de-

[2] https://www.weibo.com/

| Method | Multi-BLEU | | Intra-Dist | | Inter-Dist | | Quality | | Diversity |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-2 | Dist-1 | Dist-2 | Dist-1 | Dist-2 | Acceptable | Good | |
| **S2S** | 26.63 | 9.07 | 45.90 | 60.12 | 9.75 | 34.31 | 61.33 | 2.88 | 1.66 |
| **MMI** | 26.67 | 9.08 | 46.17 | 60.49 | 9.74 | 34.43 | 60.22 | 3.33 | 1.68 |
| **MARM** | 26.70 | 9.30 | 47.00 | 60.90 | 10.92 | 37.88 | 61.77 | 4.22 | 1.72 |
| **CMHAM** | 26.18 | 7.58 | 60.28 | 76.36 | 5.83 | 26.61 | 59.33 | 2.88 | 2.26 |
| **CVAE** | 28.88 | 8.78 | 75.57 | 92.66 | 13.59 | 49.68 | 36.88 | 2.88 | 2.62 |
| **DCVAE** | **30.44** | 8.98 | 73.33 | 90.45 | 14.43 | 53.28 | 58.67 | 4.67 | 2.72 |
| **Ours-Foc** | 27.12 | 9.01 | 44.68 | 59.75 | 10.02 | 35.21 | 60.67 | 4.67 | 1.26 |
| **Ours-FocCoverage** | 29.50 | 9.11 | 66.71 | 85.17 | 15.25 | 54.16 | 62.66 | 3.33 | 2.36 |
| **Ours-FocConstrain** | 30.32 | **9.39** | **80.24** | **95.53** | **16.89** | **59.67** | **65.33** | **9.33** | **2.82** |

Table 1: The results from automatic and human evaluations. The Kappa score is 0.45 and 0.70 for *quality* and *diversity* labeling

coding attention at each step is calculated with only the first two terms in Equation 5. 2) **Ours-FocCoverage** involves both of the *focus* $\mathcal{F}$ and the coverage vector $\mathcal{D}_t$, where the only difference from **Ours-FocConstrain** is that it is optimized without the focus constraint $\mathcal{L}_{foc}$ in Equation 9.

### 4.4 Evaluation Metrics

All models are required to generate 3 responses and are evaluated using both automatic metrics and human judgements:

**Multi-BLEU**: **BLEU** (Papineni et al., 2002)[3] is a common automatic metric to evaluate the response quality. It measures word overlaps between the generated responses and references. We report **Multi-BLEU** scores where each generated response is compared with 5 references.

**Dist-1/2**: **Dist-1/2** measures the diversity of generated responses by counting the distinct uni-grams and bi-grams (Li et al., 2016). In our setting, both **Intra-Dist** and **Inter-Dist** are evaluated on the results to calculate **Dist** of responses for a post and the whole testing set, respectively.

**Human Labeling**: Since there is a gap between automatic metrics and human annotation (Liu et al., 2016), we also consider human labeling to further validate the experiment results. We randomly sample 150 posts and generate 3 responses by each method. Then, we ask 3 professional annotators to label the responses from the aspects of **Quality** and **Diversity**, respectively.

**Quality**: We examine the generated responses from the aspects of informativeness (which measures whether the generated response is informative and interesting), relevance (which measures whether the generated response is relevant to the input post) and fluency (which measures whether the quality of the generated response). Each generated response will be categorized into bad, normal or good (scaled as 0, 1, 2). Note that a generated response will be labled as bad, if it is irrelevant to the post or has grammar mistakes. Besides, a good generated response is more than just fluent but also informative compared with a normal one. We report acceptable ratio for responses that are labeled as 1 or 2, and good ratio only for responses that just are labeled as 2.

**Diversity**: It measures the number of semantically distinct generated responses for a post. The higher the better, the maximum scale is 3.

## 5 Results and Analysis

### 5.1 Comparison Against Baselines

Results of automatic metrics and human labelings are shown in Table 2. The Kappa score is 0.45 and 0.70 for *quality* and *diversity* labeling, indicating that the annotators share a satisfactory agreement in the labeling.

We firstly examine the significance of latent variable. Generally speaking, the compared models without any latent variable (the first 4 rows in Table 1) perform the worst. As shown in Table 1, S2S and MMI achieve the lowest scores. Comparing the 3 generated responses by S2S and MMI shown in Table 2 (the 3 columns), they share similar semantics with only a few word variants. As MMI has to rerank the candidates generated by S2S, their performances are similarly disappointing. This result supports that Seq2Seq is limited in modeling diverse responses for a given post even combined with the reranking strategy. Moreover, MARM per-

[3] http://www.nltk.org/py-modindex.html

| | Post: 炎热的夏天想吃遍所有口味的冰淇淋<br>I want to eat ice-creams of all flavors of in the hot summer | | Gold: 我现在特别想吃薰衣草冰淇淋<br>I really want to eat lavender ice cream right now |
|---|---|---|---|
| **S2S** | 太可爱了，好想吃<br>It's so cute, I want to eat it. | 好想吃。好想吃<br>I want to eat it. I want to eat it | 好想吃啊。好想吃<br>I want to eat it. I want to eat it |
| **MMI** | 好想吃啊，好想吃。<br>I want to eat it. I want to eat it. | 好想吃啊，好想吃<br>I want to eat it. I want to eat it. | 好想吃啊... ...<br>I want to eat it . . . . . . |
| **MARM** | 这是什么口味的冰淇淋<br>What flavor is this ice-cream | 这是什么口味的冰淇淋<br>What flavor is this ice-cream | 这是什么口味的冰淇淋啊<br>What flavor is this ice-cream |
| **CMHAM** | 冰淇淋也可以吃冰淇淋<br>Ice-cream can also eat ice-cream. | 冰淇淋口味太重了吧<br>Ice-creams taste too heavy | 好想吃<br>I want it. |
| **CVAE** | 我好久没吃蛋糕了<br>I haven't eaten cake for a long time | 夏天快来吧，我的夏天。<br>Come on summer, my summer. | 冰淇淋吃冰淇淋，哈哈。<br>Ice-cream eats ice-cream, haha. |
| **DCVAE** | 我也好想吃这个！<br>I really want to eat this too! | 夏天的夏天快到了<br>Summer' summer is comming | 好想吃啊，流口水了。<br>I really want to eat. My mouth is watering. |
| **Ours-Foc** | 哇哦。好想吃哦。<br>Wow. I really want to eat it. | 哇哦。好想吃哦。<br>Wow. I really want to eat it. | 哇哦。好想吃哦。<br>Wow. I really want to eat it. |
| **Ours-FocCoverage** | 真心不喜欢冰淇淋<br>I really don't like ice cream | 夏天都吃冰淇淋了。<br>Always have ice cream in summer. | 夏天吃了这东东...<br>In summer, eat this thing... |
| **Ours-FocConstrain** | 薄荷味真的很好吃！<br>Mint flavor is really delicious! | 爱吃冰激凌的人表示羡慕嫉妒恨。<br>People who love to eat ice cream are green with envy. | 吃货一枚，鉴定完毕。<br>A foodie, the identification is done. |

Table 2: The gold and generated responses by each method.

forms similarly with S2S and MMI in terms of the automatic scores, human judgments as well as the generated responses shown in Table 2. Despite that MARM introduces a set of latent embeddings, its poor performance is attributed to the lack of extra disentanglement control on the mixture learning of latent mechanisms, as analyzed in the previous section. Things become interesting when we examine the performance of CMHAM. It seems that CMHAM effectively improves the diversity over other Seq2Seq models if we only checked the indicators in Table 1. However, the responses generated by CMHAM from Table 2 are either too short or ungrammatical. Such inconsistency between the results from Table 1 and Table 2 might be resulted from several causes. We conjecture one primary reason is the gap between model training and testing. During training, the semantic representation in CMHAM is learned as a mixture of all attention heads. While during testing, CMHAM is limited to use one single constrained head to focus on a certain post word.

We then examine the variational models equipped with latent variable (the fourth to sixth rows) to investigate which method(s) are more effective in utilizing the latent information. From Table 1, CVAE brings obvious improvements on *Dist* and *Diversity* as compared with the non-variational models (the first four rows). However, the responses generated by CVAE in Table 2 are of low quality. It is because that the vanilla CVAE has no extra control on the latent variable, and the stochasticity injected in the latent variable is too overwhelming for the decoder when generating responses. In turn, hardly the decoder is able to balance the latent semantics with the response fluency. As a result, the latent variable fails to effectively direct a high-quality response generation. When comparing DCVAE with CVAE, we can see noticeable increases especially on *Quality* and *Diversity*. This is not surprising in that DCVAE introduces additional control on each latent variable and connects the variables with the words in the vocabulary. Though it is more meaningful to incorporate the latent variable in this way, DCVAE is still insufficient. Take the 2nd generated response from DCVAE in Table 2 as an example where the driving word is "夏天(summer)". In this case, DCVAE is unable to adjust the attention, and thus directs the flawed response to overly emphasize on "夏天(summer)". This example partially proves that even though DCVAE has taken control over the latent variable, it is still problematic to guide response generation through a coarse-grained signal.

On the contrary, the proposed model and its variants **Ours-FocCoverage Ours-FocConstrain** base on the fine-grained *focus* signal and successfully improve the overall generation quality as well as response diversity. Especially, our full model **Ours-FocConstrain** performs the best in terms of almost every metric except *BLEU-1*. The highest scores of human evaluations in Table 1 and the responses in Table 2 together show that our proposed method **Ours-FocConstrain** is able to generate high-quality and diverse responses. In brief, our model introduces a performance boost as it fully leverages the word-level information for response generation.

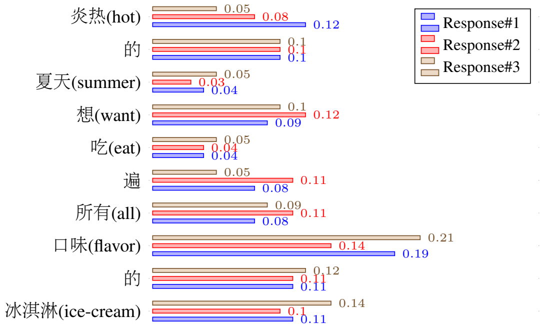

Figure 2: The focus distributions of the 3 test cases by Ours-Foc from Table 2

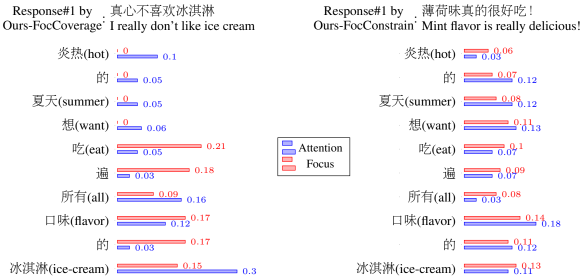

Figure 3: The focus and attention distribution of the the test cases by Ours-FocCoverage and Ours-FocConstrain from Table 2.

## 5.2 Ablation Study and Analysis

To verify the effectiveness of each proposal in our work, we conduct ablation studies by comparing with several variants of our model, **Ours-Foc**, **Ours-FocCoverage** and **Ours-FocConstrain**. The ablation results are summarized in the last three rows in Table 1 and Table 2.

Clearly, the performance gap among our variants indicate that all the three modules in the proposed model are of great importance. One thing to note in Table 1 is the unsatisfactory performance achieved by the bare-bone variant **Ours-Foc** that it performs similarly with the vanilla Seq2Seq. **Ours-Foc** solely contains the focus generator which dissembles the discourse-level latent variable into word-level guiding signals—*focus*—for each decoding step. This setting is insufficient because the model is prone to bypass the guiding signals. We observe such unexpected phenomenon in Table 2 where the three responses from Ours-Foc are generated with one single model and thus they are similar to each other. This phenomenon is further validated in Figure 2, where we plot the *focus* distributions that are correlated with the three responses from Table 2. From this experiment, we can see that the generated responses do not attach much attention to the word "口味(flavor)" even though the word is assigned with the highest *focus* weight. This verifies that, despite that **Ours-Foc** incorporates the fine-grained *focus*, it still lacks mechanism(s) and strategy(s) to make full use of it.

Upon the bare-bone model, **Ours-FocCoverage** incorporates the proposed focus-guided mechanism and increases the metric scores a lot especially on the metric *Dist* and *Diversity*. We attribute this increase to the use of coverage vector. In such way, the model is able to adjust attention based on attention history as well as the *focus*, rather than simply considering the current relevant words as in the standard attention mechanism. Therefore, the *focus* tends to show guiding significance for the decoder to generate qualified responses. From Table 2 we can see, the responses generated by **Ours-FocCoverage** differ from each other with respect to both semantic meaning and their expressions.

More importantly, **Ours-FocConstrain** further employs the novel focus constraint to properly align the target response with input post according to the *focus*. To examine in detail, we plot both *focus* and decoding attention distribution of the test cases by **Ours-FocCoverage** and **Ours-FocConstrain**. As shown in Figure 3, the latent variable of **Ours-FocCoverage** addresses the highest *focus* to the word "吃(eat)". However, the decoder does not follow such guidance and pays more attention to the word "冰淇淋(ice-cream)", resulting in an improper response. In contrast, the latent variable in **Ours-FocConstrain** concentrates more on the word "口味(flavor)" than the others. With the help of focus constraint, the decoder of **Ours-FocConstrain** makes it to direct the generated response embody the meaning of "口味(flavor)". In other words, though **Ours-FocCoverage** introduces the coverage vector and potentially encourages the diversity using different sampled latent variables, **Ours-FocConstrain** steps further and kills the chance of generating responses regardless of the *focus* by using the constraint $\mathcal{L}_{foc}$. Drawing on the highest scores achieved by **Ours-FocConstrain**, we conclude that the proposed focus constraint is an indispensable design and is potentially beneficial for CAVE-based response generation models.

Overall speaking, the proposed **Focus-Constrained Attention Mechanism** consists of: (1) focus generator to produce fine-grained signals; and (2) focus-guided mechanism and focus constraint to fully utilize the signal. This ablation study validates the necessity of fine-grained latent information, and demonstrates the effectiveness of each component in the proposed method. By leveraging the proposed Focus-Constrained Attention Mechanism, the decoder is able to tell the importance of each word and start a holistic-planned response generation under the fine-grained focus guidance.

## 6 Conclusion

In this paper, we identify the insufficiency of discourse-level latent variable in response generation. To address this, we develop a novel CVAE-based model, which exploits a fine-grained word-level feature to generate diverse responses. On a real-world benchmarking dataset, we demonstrate that our proposed model is able to fully leverage the fine-grained feature, and generate better responses as compared to several SOTA models. Based on the ablation studies, we verify the contribution of each proposal in our method and highlight the significance of fine-grained signal in response generation.

## Acknowledgments

We would like to thank anonymous reviewers for their helpful comments and suggestions to improve our work.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. 2018. Variational attention for sequence-to-sequence models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1672–1682.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. *arXiv preprint arXiv:1906.01781*.

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 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, pages 9712–9724.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. 2019. A discrete cvae for response generation on short-text conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1898–1908.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. *arXiv preprint arXiv:1805.12352*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Diana Perez-Marin and Ismael Pascual-Nieto. 2011. Conversational agents and natural language interaction: Techniques and effective.

Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are training samples correlated? learning to generate dialogue responses with multiple references. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3826–3835.

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

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

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2015. Attribute2image: Conditional image generation from visual attributes. *arXiv preprint arXiv:1512.00570*.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2199.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. 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 (Volume 1: Long Papers)*, pages 654–664.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.