Generating Multiple Diverse Responses with Multi-Mapping and Posterior Mapping Selection

Chaotao Chen[1] , **Jinhua Peng**[1] , **Fan Wang**[1] , **Jun Xu**[2] and **Hua Wu**[1]

[1]Baidu Inc., China
[2]Harbin Institute of Technology, China

{chenchaotao, pengjinhua, wangfan04, wu_hua}@baidu.com, jxu@ir.hit.edu.cn

Abstract

In human conversation an input post is open to multiple potential responses, which is typically regarded as a one-to-many problem. Promising approaches mainly incorporate multiple latent mechanisms to build the one-to-many relationship. However, without accurate selection of the latent mechanism corresponding to the target response during training, these methods suffer from a rough optimization of latent mechanisms. In this paper, we propose a *multi-mapping* mechanism to better capture the one-to-many relationship, where multiple mapping modules are employed as latent mechanisms to model the semantic mappings from an input post to its diverse responses. For accurate optimization of latent mechanisms, a *posterior mapping selection* module is designed to select the corresponding mapping module according to the target response for further optimization. We also introduce an auxiliary *matching loss* to facilitate the optimization of posterior mapping selection. Empirical results demonstrate the superiority of our model in generating multiple diverse and informative responses over the state-of-the-art methods.

1 Introduction

Recently, generative models built upon sequence-to-sequence (Seq2Seq) framework [Sutskever *et al.*, 2014; Shang *et al.*, 2015] have achieved encouraging performance in open-domain conversation, with their simplicity in learning the mapping from input post to its response directly. However, an input post in human conversation is open to multiple potential responses, which is typically regarded as a one-to-many problem. Modeling diverse responding regularities as a one-to-one mapping, the Seq2Seq models inevitably favor general and trivial responses [Li *et al.*, 2016a]. Thus the rich and diverse content in human conversation can not be captured.

To address this problem, work from [Zhao *et al.*, 2017; Serban *et al.*, 2017] combines Seq2Seq with Conditional Variational Auto-Encoder (CVAE) and introduces a Gaussian latent distribution to build the one-to-many relationship. By drawing samples from the Gaussian latent distribution, multiple responses can be generated. However, the Gaussian latent
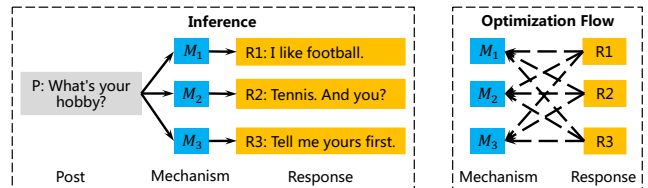

Figure 1: Overview of models with multiple latent mechanisms.

distribution is not compatible to the multi-modal[1] nature of diverse responses and lack of interpretability.

For these issues, recent approaches [Zhou *et al.*, 2017; Zhou *et al.*, 2018a; Tao *et al.*, 2018; Gu *et al.*, 2019] resort to incorporation of multiple latent mechanisms, each of which could model various responding regularities for an input post, as shown in Figure 1. For example, Zhou *et al.* [2017; 2018a] introduce multiple latent embeddings as language responding mechanisms into Seq2Seq framework, and output responses of different language styles by choosing different embeddings; Tao *et al.* [2018] augment Seq2Seq model with multi-head attention mechanism, and generate responses that focus on specific semantic parts of the input post with different heads of attention. Although these methods have shown potential to capture the multi-modal nature of diverse responses, they still fail to fulfill the one-to-many relationship, due to their inaccurate optimization of latent mechanisms. As shown in Figure 1, given a target response, the optimization is distributed to each latent mechanism. However, for more accurate modeling, we assume only the latent mechanism corresponding to the target response should be selected for optimization. For example, given a questioning response, we should only optimize the latent mechanism that models interrogative responding regularities rather than the other irrelevant ones. Although in some methods the optimization to each latent mechanism is guided by a weight from the input post, the weight is inaccurate to represent the selection of the corresponding latent mechanism, considering the semantic gap between the input post and the target response. With such a rough optimization, the latent mechanisms are not guaranteed to capture the diverse responding regularities.

In this paper, in order to capture the one-to-many relationship, we propose to augment the Seq2Seq framework with

---

[1]Multi-modal means the property with multiple modes.

a *multi-mapping* mechanism, employing multiple mapping modules as latent mechanisms to model the distinct semantic mappings between an input post and its diverse responses. More importantly, to avoid the rough optimization of latent mechanisms in previous methods, in training time we incorporate a *posterior mapping selection* module to select the corresponding mapping module according to the target response. By explicitly leveraging the information in the target response (i.e. posterior information), it is easier to select the accurate mapping module. Then only the selected mapping module is updated given the target response. Moreover, to facilitate the optimization of posterior mapping selection, we further propose an auxiliary *matching loss* that evaluates the relevance of post-response pair. Compared with the simple embedding mechanism and the multi-head attention mechanism whose diversity is limited to the semantic attentions in the input post, the proposed multi-mapping mechanism is more flexible to model different responding regularities. And it also introduces multi-modal capacity and interpretability over the Gaussian latent distribution. With the posterior mapping selection to ensure the accurate optimization of mapping modules, our model is more effective to capture diverse and reasonable responding regularities.

Our contributions can be summarized as follow:

- We propose a multi-mapping mechanism to capture the one-to-many relationship with multiple mapping modules as latent mechanisms, which is more flexible and interpretable over previous methods.

- We propose a novel posterior mapping selection module to select the corresponding mapping module according to the target response during training, so that more accurate optimization of latent mechanisms is ensured. An auxiliary matching loss is also introduced to facilitate the optimization of posterior mapping selection.

- We empirically demonstrate that the proposed multi-mapping mechanism indeed captures distinct responding regularities in conversation. We also show that the proposed model can generate multiple diverse, fluent and informative responses, which obviously surpasses the other existing methods.

2 Model

2.1 Model Overview

Following the conventional setting for generative conversation models [Shang *et al.*, 2015; Li *et al.*, 2016b], we focus on the single-round open-domain conversation. Formally, given an input post $X = (x_1, x_2, ..., x_T)$, the model should generate a natural and meaningful response $Y = (y_1, y_2, ..., y_{T'})$.

To address the one-to-many problem in conversation, we propose a novel generative model with multi-mapping mechanism and posterior mapping selection module. The multi-mapping mechanism employs multiple mapping modules to capture the various underlying responding regularities between an input post and its diverse responses. The posterior mapping selection module leverages the posterior information in target response to identify which mapping module should be updated, so as to avoid the rough optimization in
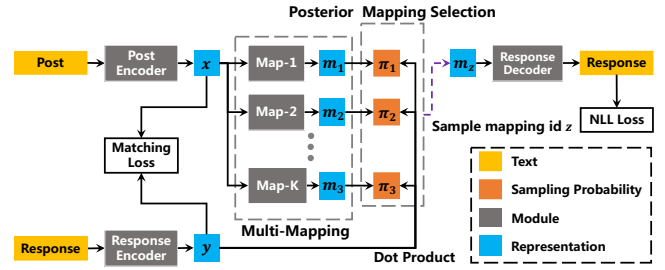

Figure 2: An overview of the proposed Seq2Seq model with multi-mapping and posterior mapping selection.

previous methods. The architecture of our model is illustrated in Figure 2 and it consists of following major components:

Post Encoder encodes the input post X into a semantic representation x and feeds it into different mapping modules.

Response Encoder encodes the target response Y into a semantic representation y for posterior mapping selection.

Multi-Mapping consists of K mapping modules and maps post representation x to different candidate response representation $\{m_k\}_{k=1}^K$ through different mapping module M_k, respectively.

Posterior Mapping Selection selects the z-th mapping module that is corresponding to the target response in the training time.

Response Decoder generates the response based on the candidate response representation m_k.

2.2 Encoder

The post encoder employs a one-layer bidirectional Gated Recurrent Unit (GRU) [Cho *et al.*, 2014] to transform the input post X into a sequence of hidden state h_t as follows:

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{1}$$

$$\overrightarrow{h}_t = \mathbf{GRU}(\overrightarrow{h}_{t-1}, e(x_t)) \tag{2}$$

$$\overleftarrow{h}_t = \mathbf{GRU}(\overleftarrow{h}_{t+1}, e(x_t)) \tag{3}$$

where $[\cdot; \cdot]$ denotes the concatenation of states, $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the forward and backward hidden states at time t, $e(x_t)$ is the embedding of word x_t. The semantic representation of input post is summarized as $x = [\overrightarrow{h}_T; \overleftarrow{h}_1]$.

The response encoder, which encodes the target response Y to the semantic representation y, follows the same structure as post encoder but with different learnable parameters.

2.3 Multi-Mapping

For one-to-many relationship, we introduce a multi-mapping mechanism to capture the different responding regularities, with multiple mapping modules bridging the post encoder and response decoder. Specifically, we employ a linear mapping function as the mapping module for simplicity and leave more advanced mapping structures as future work. Formally, the model maps the post representation x to the K different candidate response representations $\{m_k\}_{k=1}^K$ through different mapping modules as follows:

$$m_k = W_k x + b_k \tag{4}$$

where $\boldsymbol{W}_k$ and $\boldsymbol{b}_k$ are the learnable parameters of the k-th mapping module M_k.

2.4 Posterior Mapping Selection

To ensure accurate optimization of mapping modules in training time, it is necessary to identify which mapping module is responsible for the generation of target response and only update the corresponding mapping module given the target response. Thus, we incorporate a posterior mapping selection module to explicitly select the corresponding mapping module by leveraging the information in target response. With the guidance of target response, we assume it is easier to find the corresponding mapping module for accurate optimization. Specifically, we introduce a categorical distribution π to denote the selection of mapping module conditioned on the target response. And the selection probability of the k-th mapping module is based on its relevance to the target response, which is measured by the dot product between the representations of candidate response and target response as follows:

$$\pi_k = \frac{exp(\boldsymbol{m}_k \cdot \boldsymbol{y})}{\sum_{i=1}^{K} exp(\boldsymbol{m}_i \cdot \boldsymbol{y})} \quad (5)$$

Then for a target response, the corresponding mapping module can be sampled according to their relevance. Given that the z-th mapping module M_z is selected, only the corresponding candidate representation $\boldsymbol{m}_z$ is fed into the response decoder for further decoding optimization. Therefore, optimization of irrelevant mapping module is not conducted and more accurate optimization of latent mechanisms is ensured. In order to back-propagate through the discrete sampling in posterior mapping selection, we leverage the Gumbel-Softmax reparametrization [Jang $et\ al.$, 2017].

2.5 Decoder

The response decoder employs an uni-directional GRU with the selected candidate representation $\boldsymbol{m}_z$ as its initial state and update its hidden state as follows:

$$\boldsymbol{s}_t = \textbf{GRU}(\boldsymbol{s}_{t-1}, \boldsymbol{e}(y_{t-1}), \boldsymbol{c}_t); \quad \boldsymbol{s}_0 = \boldsymbol{m}_z \quad (6)$$

where $\boldsymbol{s}_t$ is the hidden state of decoder at time t, $\boldsymbol{e}(y_{t-1})$ is the embedding of the last generated word y_{t-1}, and $\boldsymbol{c}_t$ is the attended context vector at time t and defined as the weighted sum of hidden states of the post encoder: $\boldsymbol{c}_t = \sum_{i=1}^{T} \alpha_{t,i}\boldsymbol{h}_i$, where $\alpha_{t,i}$ is the attention weight over $\boldsymbol{h}_i$ at time t:

$$\boldsymbol{\alpha}_t = softmax(\boldsymbol{e}_t) \quad (7)$$

$$e_{t,i} = \boldsymbol{v}^T tanh(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t) \quad (8)$$

where $\boldsymbol{v}$, $\boldsymbol{W}_h$ and $\boldsymbol{W}_s$ are learnable parameters. Then at time t, the generation probability conditioned on the input post X and the selected mapping module M_z is calculated as:

$$p(y_t|y_{<t}, X, M_z) = softmax(\boldsymbol{s}_t, \boldsymbol{c}_t) \quad (9)$$

where $y_{<t}$ denotes the previous generated words.

The objective of the response generation is to minimize the negative log-likelihood of the target response Y conditioned on the input post X and the selected mapping module M_z as follows:

$$\mathcal{L}_G = -\ log\ p(Y|X, M_z) \quad (10)$$

where $p(Y|X, M_z) = \prod_{t=1}^{T'} p(y_t|y_{<t}, X, M_z)$ is the conditional probability of target response.

2.6 Auxiliary Objective

Although the posterior mapping selection module is designed to select the corresponding mapping module by referring to the target response, we find that its raw implementation quickly converges to selecting the same mapping module and thus the proposed model falls back to the vanilla Seq2Seq. We conjecture that in the early training, the response encoder is inefficient to capture the semantic information in target response. So the posterior mapping selection fails to provide an accurate selection of mapping module and the model falls into a local optima that focuses on single mapping module. To address this issue, we introduce an auxiliary objective from the response retrieval task to improve the semantic extraction of response encoder. The auxiliary objective namely *matching loss* is to evaluate the relevance of post-response pair. Specifically, given an input post X and a target response Y, their relevance probability is estimated by the dot product of their semantic representations:

$$p(r = 1|X, Y) = \sigma(\boldsymbol{x} \cdot \boldsymbol{y}) \quad (11)$$

where r is the label denoting if the response Y is relevant to the post X, and σ is a sigmoid function. Following the previous work [Shang $et\ al.$, 2018], we adopt negative sampling to train this auxiliary task so as to release the burden of human annotation. Particularly, for the input post X and golden response Y, we randomly sample another response Y^- in training set as a negative sample. Formally, the matching loss $\mathcal{L}_M$ is defined as the negative log-likelihood of relevance for the golden response and negative response:

$$\mathcal{L}_M = -log\ p(r = 1|X, Y) + log\ p(r = 1|X, Y^-) \quad (12)$$

With this auxiliary matching loss, the response encoder is more efficient to capture the semantic information from the target response and provide a better relevance measurement for the accurate posterior mapping selection.

2.7 Training and Generation

Overall, the total loss function of our model is a combination of the generation loss $\mathcal{L}_G$ and the matching loss $\mathcal{L}_M$:

$$\mathcal{L} = \mathcal{L}_G + \mathcal{L}_M \quad (13)$$

All the parameters are simultaneously updated with back-propagation.

After optimization with posterior mapping selection, the model is able to capture distinct responding regularities and generate various candidate responses with different mapping modules. For response generation, we assume each mapping module is reasonable and just randomly pick mapping module for responding to avoid selection bias. More advanced response selection such as reranking is left as future work.

3 Experiment

3.1 Datasets

We evaluate the proposed model on two public conversation dataset: **Weibo** [Shang $et\ al.$, 2015] and **Reddit** [Zhou $et\ al.$, 2018b] that maintain a large repository of post-response pairs from popular social websites. After basic data cleaning, we have above 2 million pairs in both datasets. The statistics of datasets are summarized in Table 1.

Dataset	#train	#valid	#test
Weibo	2,630,212	11,811	974
Reddit	2,173,501	6,536	1,298

Table 1: Statistics of datasets.

3.2 Implementation Details

The vocabulary size is limited to 40,000 and 30,000 in Weibo and Reddit dataset, respectively. The hidden size in both encoder and decoder is set to 1024. Word embedding has size 300 and is shared for both encoder and decoder. We initialize the word embedding from pre-trained Weibo embedding [Li et al., 2018] and GloVe embedding [Pennington et al., 2014] for Weibo and Reddit dataset, respectively. The temperature of Gumbel-Softmax trick is set to 0.67. All model are trained end-to-end by the Adam optimizer [Kingma and Ba, 2015] on mini-batches of size 128, with learning rate 0.0002. We train our model in 10 epochs and keep the best model on the validation set for evaluation. The code is available at: https://github.com/LittleChencht/MMPMS.

3.3 Compared Methods

We compare our model with several state-of-the-art generative conversation models in both single and multiple (i.e. 5) response generation:

Seq2Seq [Bahdanau et al., 2015]: The standard Seq2Seq architecture with attention mechanism.

MMI-bidi [Li et al., 2016a]: The Seq2Seq model using Maximum Mutual Information (MMI) between inputs and outputs as the objective function to reorder generated responses. We adopt the default setting: $\lambda = 0.5$ and $\gamma = 1$.

CVAE [Zhao et al., 2017]: The Conditional Variational Auto-Encoder model with auxiliary bag-of-words loss.

MARM [Zhou et al., 2017]: The Seq2Seq model augmented with mechanism embeddings to capture latent responding mechanisms. We use 5 latent mechanisms and generate one response from each mechanism.

MHAM [Tao et al., 2018]: The Seq2Seq model with multi-head attention mechanism. The number of heads is set to 5. Following the original setting, we combine all heads of attention to generate a response for the single response generation, and generate one response from each head of attention for the multiple response generation. Although the constrained MHAM is reported better performance in the original paper, on our both datasets we see negligible improvement in single response generation and much worse performance in multiple response generation due to the lack of fluency. So we only adopt the unconstrained MHAM as baseline.

MMPMS: Our proposed model with *Multi-Mapping* and *Posterior Mapping Selection* (MMPMS). We set the number of mapping modules to 20.

3.4 Evaluation Metrics

For automatic evaluation, we report: **BLEU** [Chen and Cherry, 2014]: A widely used metric for generative dialogue systems by measuring word overlap between the generated response and the ground truth. **Dist-1/2** [Li et al., 2016a]: Ratio of distinct unigrams/bigrams in the generated responses, which can measure the diversity of generated responses.

Since empirical experiments [Liu et al., 2016] have shown weak correlation between automatic metrics and human annotation, we consider the careful human judgment as major measurement in the experiments. In detail, three annotators are invited to evaluate the quality of the responses for 300 randomly sampled posts. Similar to [Zhou et al., 2017], for each response the annotators are asked to score its quality with the following criteria: (1) **Bad**: The response is ungrammatical and irrelevant. (2) **Normal**: The response is basically grammatical and relevant to the input post but trivial and dull. (3) **Good**: The response is not only grammatical and semantically relevant to the input post, but also meaningful and informative. Responses on normal and good levels are treated as "**Acceptable**". Additionally, to evaluate the diversity of multiple responses generation, for the 5 responses generated for a single post, the annotators also annotate the number of distinct meanings among the acceptable responses, namely **Diversity**. The average Fleiss' kappa [Fleiss and Cohen, 1973] value is 0.55 and 0.63 on Weibo and Reddit, respectively, indicating that the annotators reach moderate agreement.

3.5 Evaluation Results

The evaluation of single response generation are summarized in Table 2. As shown, our model achieves the best performance in human evaluation and *Dist* on both datasets, especially the visible enhancement in *Good* ratio (0.24 vs. 0.15 on Weibo and 0.36 vs. 0.25 on Reddit) compared with the best baseline, indicating our model can generate more informative and diverse responses. Notably, Seq2Seq performs the best in BLEU but poor in human evaluation on Weibo, which further verifies the weak correlation of BLEU to human judgment.

Table 3 shows the evaluation results of multiple response generation. As can be seen, our model outperforms baseline methods by a large margin in human evaluation on both datasets. More importantly, the *Diversity* measure of our model reaches 2.61 on Weibo and 3.16 on Reddit, much higher than other baseline methods, which demonstrates the superiority of our model to generate multiple diverse and high-quality responses. This can also be supported by the examples in Table 4[2], where the multiple candidate responses returned by our model are much more relevant and diverse.

However, CVAE fails to generate appropriate responses (i.e. low *Acceptable* ratio) even though it achieves relatively high *Dist* scores. It seems that its generation diversity comes more from the sampling randomness of the prior distribution rather than from the understanding of responding regularities. And we conjecture that the lack of multi-modal property in Gaussian distribution makes it hard to capture the one-to-many relationship among human conversation.

Instead, MARM performs the worst in response diversity with the lowest score of *Diversity* and *Dist* while it obtains a high *Acceptable* ratio. As shown in Table 4, the responses from MARM are similar and trivial. And the word overlap among the 5 candidate responses is up to 94% and 96% on Weibo and Reddit, respectively, showing that each mecha-

---

[2]Due to space limitation, responses from Seq2Seq and MMI-bidi are omitted, considering their lack of one-to-many mechanism and low diversity in multiple response generation.

Model	Weibo				Reddit			
	Acceptable	Good	BLEU-1/2	Dist-1/2	Acceptable	Good	BLEU-1/2	Dist-1/2
Seq2Seq	0.43	0.08	**0.305/0.246**	0.122/0.326	0.57	0.10	0.205/0.162	0.091/0.254
MMI-bidi	0.46	0.09	0.271/0.218	0.153/0.372	0.54	0.25	**0.345/0.279**	0.107/0.325
CVAE	0.29	0.15	0.252/0.203	0.184/0.542	0.42	0.25	0.287/0.233	0.107/0.428
MARM	0.48	0.11	0.304/0.245	0.132/0.376	0.60	0.09	0.205/0.162	0.100/0.287
MHAM	0.50	0.10	0.304/0.245	0.127/0.347	0.60	0.10	0.192/0.151	0.115/0.331
MMPMS	**0.56**	**0.24**	0.275/0.225	**0.189/0.553**	**0.65**	**0.36**	0.207/0.165	**0.135/0.433**

Table 2: Evaluation results of single response generation on Weibo and Reddit dataset.

Model	Weibo					Reddit				
	Acceptable	Good	Diversity	BLEU-1/2	Dist-1/2	Acceptable	Good	Diversity	BLEU-1/2	Dist-1/2
Seq2Seq	0.45	0.08	0.79	0.291/0.234	0.037/0.133	0.43	0.10	1.27	0.300/0.242	0.022/0.085
MMI-bidi	0.47	0.09	0.99	0.272/0.219	0.047/0.166	0.52	0.24	1.45	**0.339/0.274**	0.028/0.127
CVAE	0.22	0.13	1.06	0.275/0.224	**0.105/0.404**	0.42	0.27	2.06	0.279/0.226	0.047/0.257
MARM	0.49	0.11	0.62	**0.306/0.246**	0.030/0.091	0.60	0.10	0.67	0.204/0.161	0.021/0.064
MHAM	0.31	0.08	1.41	0.234/0.190	0.074/0.240	0.40	0.12	1.82	0.184/0.149	0.056/0.234
MMPMS	**0.60**	**0.25**	**2.61**	0.270/0.219	0.100/0.389	**0.68**	**0.38**	**3.16**	0.195/0.159	**0.060/0.265**

Table 3: Evaluation results of multiple response generation on Weibo and Reddit dataset.

nism embedding converges to similar and general responding regularities. We attribute this to the lack of accurate selection of latent mechanism for the target response during training. Since each mechanism embedding is roughly optimized with the same target response, they are prone to learn similar and general responding relationships. This result further validates the importance of our proposed posterior mapping selection.

It is also interesting to find the degradation of MHAM in multiple response generation over single response generation. According to Table 4, the responses from different heads of attention are diverse but ungrammatical and irrelevant. The reason may also lies in the absence of accurate selection of latent mechanism during training. Since the response generation is optimized with a combination of all heads rather than the head corresponding to the target response, each head of attention is coupled together and fails to capture independent responding regularity. So the model ends up with inappropriate responses when only single head of attention is utilized in multiple response generation, but ends up with appropriate responses when combining all heads of attention in single response generation. Another potential reason is that there is no enough distinct semantic information in the input post for the model to attend separately.

3.6 Analysis on Mapping Modules
We also conduct analysis to explore the responding regularities that the mapping modules have captured. For the 200 posts sampled from the Weibo test set, we obtain the candidate representation m_k from different mapping modules and apply t-SNE [van der Maaten and Hinton, 2008] for visualization. As shown in Figure 3, the candidate representations are highly clustered by their corresponding mapping modules, indicating the ability of various mapping modules to model various responding regularities by mapping the post representation to significantly different response representations.

For more intuitive understanding, we identify the keywords of different mapping modules from their responses for the Weibo test set. We assume that for a mapping module its
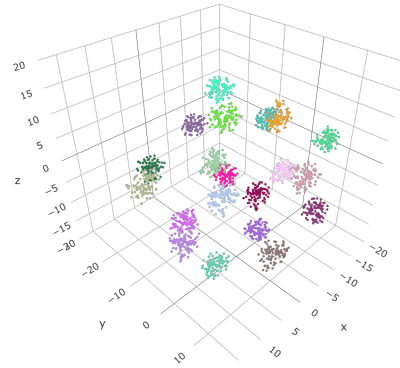

Figure 3: t-SNE visualization of candidate representations. The color represents the mapping module.

keyword should appear frequently in its output responses but rarely occur in other mapping modules. Then, the importance of a word w in the mapping module M_k is measured by $p(w|M_k) = N_k^w / \sum_{i=1}^{K} N_i^w$, where N_k^w is the number of times that word w occurs in the responses from M_k. In addition, only the keywords occurring frequently are considered, namely $N_k^w > 40$. Table 5 illustrates the keywords of several representative mapping modules. As we can see, Map-1, whose keywords are mainly question words (e.g. *where*, *what* and the question mark), probably represents interrogative responding regularity. Map-2 tends to respond with mood words (e.g. *wow*, *lol*, *haha*). The keywords of Map-3 are composed of intensity words (e.g. *too*, *so*) and modifier words (e.g. *horrible*, *cute*, *pretty*), showing that it tends to respond with surprise and emphasis. And Map-4 is more likely to return responses in English. The keywords of Map-5 are mainly subjective words (e.g. *I*, *think*, *believe*, *want*), indicating that it tends to generate responses with respect to personal opinion or intention. These results verify that each mapping module can capture diverse and meaningful responding regularities.

CVAE	MARM	MHAM	MMPMS
Post-1: 新京报：雾霾 1 月内第 4 次侵扰京城 (The Beijing News: Smog has hit Beijing for the fourth time in a month)			
强烈抵制环球时报！ Strong boycott of the Global Times! 可怜的帝都啊！ Poor imperial capital! 原来是这样的，有一天。 So it is, one day.	这是一个神奇的国度！ This is a magical country! 这是一个神奇的国度！ This is a magical country! 雾霾的天是晴朗的天。 It's sunny on a foggy day.	京城雾霾！京城雾霾！ Beijing smog! Beijing smog! 这是一个笑话，一个笑话 It's a joke, a joke 这就是所谓的雾霾？ Is this the so-called smog?	北京的空气质量有待提高啊 Beijing's air quality needs to be improved 北京雾霾是不是也该管管了？ Time to tackle the smog in Beijing? 有没有特供的空气净化器？ Are there any special air purifiers?
Post-2: 创业一切的问题都是自己的，搞定了就赚了 (All problems of entrepreneurship are our own. Once settled, we would make a profit.)			
要学创业，可以泡妞 To learn to start a business, you can pick up girls 如何才能成功啊？ How can we succeed? 还真是。刚买了一本书。 It's true. Just bought a book.	创业是一生生活态度。 Entrepreneurship is an attitude towards life. 这是一个神奇的世界！ This is a magical world! 这是一个神奇的世界！ This is a magical world!	我也是自己的，搞定了就赚了 I'm my own guy, too. I'll make money when I'm done 哈哈，我也是，不过我也是 Haha, me too, but me too 这就是传说中的钱荒？ Is this the legendary money scarcity?	创业有风险，投资需谨慎。 Entrepreneurship is risky and investment should be prudent. 那你赚了多少钱呢？ Then how much did you earn? 赚了钱就不是问题了。 That's not a problem if made a profit.
Post-3: Yes! 16gb should be standard for every pc if you ask me.			
I don't think so, but it's still valid. There's nothing to do with it. You should still buy it.	That's what i was thinking. That's what i was thinking. That's what i was thinking.	It should be standard for every game. Didn't know that i hadn't thought of it. 16gb 16gb compatibility compatibility	Have you tried installing the game? Any recommendations for a computer? Even if it's standard for the pc version?
Post-4: I agree! White chocolate isn't chocolate. And low quality white chocolate is a vile substance.			
I don't think it's vanilla custard. Dark chocolate is a poor guy. Don't you worry about the dark chocolate?	I don't know what you're talking about. I don't know what you're talking about. I don't know what you're talking about.	It's not. it's a cake. What color is black ? I agree with this. I like low quality consistency.	Nothing is worse than white chocolate. Have you ever heard of dark chocolate? Yup, same here. It tastes like shit.

Table 4: Examples of multiple response generation. The first two are from Weibo and the last two are from Reddit.

Map-1		Map-2		Map-3		Map-4		Map-5	
哪里	where	嗯	um	可怕	horrible	to	to	以为	think
在	stay	呵呵	lol	太	too	so	so	信	believe
你	you	哇	wow	这么	so	i	i	想	want
什么	what	好吧	well	可爱	cute	have	have	知道	know
意思	mean	哈哈	haha	漂亮	pretty	go	go	我	I
？	？	这个	this	美	beautiful	you	you	只	only
要	want	也	also	请	please	love	love	这样	such

Table 5: Keywords of different mapping modules.

4 Related Work

The *safe response* problem [Li *et al.*, 2016a] in Seq2Seq models [Sutskever *et al.*, 2014; Shang *et al.*, 2015] remains an open challenge. In order to generate multiple diverse responses, many approaches resort to enhanced beam search [Li *et al.*, 2016a; Li *et al.*, 2016b]. But these methods are only applied to the decoding process and limited by the semantic similarity in the decoded responses. Another line of research turns to the different factors that determine the generation of diverse responses, such as sentence function [Ke *et al.*, 2018], specificity [Zhang *et al.*, 2018], dialogue act [Xu *et al.*, 2018] and keywords [Gao *et al.*, 2019]. However, such methods require annotations and can capture only one aspect of one-to-many relationship. Work from [Zhao *et al.*, 2017; Serban *et al.*, 2017] combines Seq2Seq with Conditional Variational Auto-Encoder and employs a Gaussian distribution to capture discourse-level variations. But it is observed that these methods suffer from the *posterior collapse* problem [Bowman *et al.*, 2016]. Moreover, the Gaussian distribution is not adaptive to the multi-modal nature of diverse responses.

The most relevant work to ours is those incorporating multiple latent mechanisms for one-to-many relationship. Zhou *et al.* [2017; 2018a] propose a mechanism-aware machine that introduces multiple latent embeddings as language responding mechanisms. Tao *et al.* [2018] propose a multi-head attention mechanism to generate diverse responses by attending to various semantic parts of an input post. Gu *et al.* [2019] incorporate a Gaussian mixture prior network and employ Gaussian component as the latent mechanism to capture the multi-modal nature of diverse responses. However, without an accurate selection of the latent mechanism corresponding to the target response, these methods suffer from a rough optimization of latent mechanisms. Given a target response, instead of optimizing the corresponding latent mechanism, Zhou *et al.* [2017] just roughly optimize each mechanism embedding while Tao *et al.* [2018] optimize a rough combination of all heads of attention according to the input post. Whereas, our model maintains a more accurate selection of the corresponding latent mechanism by referring to the posterior information in the target response. Although posterior information is also utilized in [Gu *et al.*, 2019], it is for the optimization of the prior Gaussian component which is roughly inferred by the input post, rather for the accurate selection of the Gaussian component.

5 Conclusion

In this paper, we augment the Seq2Seq model with a multi-mapping mechanism to learn the one-to-many relationship for multiple diverse response generation. Particularly, our model incorporates a posterior mapping selection module to select the corresponding mapping module according to the target response for accurate optimization. An auxiliary matching loss is also proposed to facilitate the optimization of posterior mapping selection. Thus each mapping module is led to capture distinct responding regularities. Experiments and analysis support that our model works as expected and tends to generate responses of diversity and high quality.

Acknowledgments

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Bowman *et al.*, 2016] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

[Chen and Cherry, 2014] Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Fleiss and Cohen, 1973] Joseph L. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973.

[Gao *et al.*, 2019] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. Generating multiple diverse responses for short-text conversation. In *AAAI*, 2019.

[Gu *et al.*, 2019] Xiaodong Gu, Kyunghyun Cho, Jungwoo Ha, and Sunghun Kim. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. In *ICLR*, 2019.

[Jang *et al.*, 2017] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[Ke *et al.*, 2018] Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. Generating informative responses with controlled sentence function. In *ACL*, pages 1499–1508, 2018.

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

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

[Li *et al.*, 2016b] Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*, 2016.

[Li *et al.*, 2018] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. In *ACL*, pages 138–143, 2018.

[Liu *et al.*, 2016] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132, 2016.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

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

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

[Shang *et al.*, 2018] Mingyue Shang, Zhenxin Fu, Nanyun Peng, Yansong Feng, Dongyan Zhao, and Rui Yan. Learning to converse with noisy data: Generation with calibration. In *IJCAI*, pages 4338–4344, 2018.

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

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

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, Nov(9):2579–2605, 2008.

[Xu *et al.*, 2018] Can Xu, Wei Wu, and Yu Wu. Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255*, 2018.

[Zhang *et al.*, 2018] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. Learning to control the specificity in neural response generation. In *ACL*, pages 1108–1117, 2018.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pages 654–664, 2017.

[Zhou *et al.*, 2017] Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. Mechanism-aware neural machine for dialogue response generation. In *AAAI*, pages 3400–3407, 2017.

[Zhou *et al.*, 2018a] Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. Elastic responding machine for dialog generation with dynamically mechanism selecting. In *AAAI*, pages 5730–5737, 2018.

[Zhou *et al.*, 2018b] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.