One-to-Many and Many-to-One Dialogue Learning via Sentence Semantic Segmentation Guided Conditional Variational Auto-Encoder

Anonymous ACL submission

Abstract

Due to the complex mapping relations, one-to-many and many-to-one phenomena are huge challenges for open-domain dialogue generation task, which tend to make dialogue models generate irrelevant, incoherent or non-diverse responses. Most existing methods avoid learning such phenomena through introducing the external information, reconstructing the optimization function or manipulating data samples. However, avoiding confronting such challenges ignores valuable information in these responses, and the dialogue models cannot learn the nature of such phenomena. In this paper, we propose a Sentence Semantic Segmentation guided Conditional Variational Auto-Encoder (SegCVAE) to directly learn one-to-many and many-to-one responses. SegCVAE uses prominent semantics to replace the original semantics to learn the distribution of latent variables, which avoids the gap between latent variables and the context, thus ensuring the relevance and coherence of the generated responses. Furthermore, SegCVAE can segment multiple prominent semantics to ensure the diversity of generated responses. To evaluate the model, we first define two new tasks named one-to-many dialogue learning task and many-to-one dialogue learning task. And then provide two new dialogue datasets named One-to-Many and Many-to-One, which are extracted from the well-established dataset. Finally, we also propose the evaluation strategies based on some commonly-used metrics. The experiment results show that our model achieve better performance than the baseline models in addressing these two new tasks.

1 Introduction

One-to-many and many-to-one phenomena, commonly occurring in human dialogue, arise huge challenges for open-domain dialogue generation models (Csaky et al., 2019; Sun et al., 2021): The one-to-many phenomenon could lead the model to generate irrelevant and incoherent responses, while the many-to-one phenomenon could make the model generate non-diverse responses. Facing such phenomenon, most existing methods are trying to avoid directly training models from the one-to-many and many-to-one phenomena to improve their performance. For instance, some methods (Luong et al., 2015; Li et al., 2016b) introduce external information to convert the one-to-many dialogue pairs into one-to-one dialogue pairs, thus reducing the difficulty of training models; Some methods (Li et al., 2016a; Zhang et al., 2018b; Liu et al., 2020) reconstruct the optimization functions, which allows the model to learn to generate qualified responses instead of ground-truth responses, thereby avoiding the directly training on the many-to-one dialogue pairs; Some methods (Xu et al., 2018b; Csaky et al., 2019; Akama et al., 2020) train the model through the filtered datasets, which usually contains little one-to-many and many-to-one dialogue pairs.

context 1	nothing works with my toothache now.
context 2	oh!!! i have a horrible toothache.
response	you should go to the dentist.
context 1	could you tell me how to use it?
context 2	what should i do with the token?
response	you put it in the slot at the turnstile and then push the turnstile to get into the platform.
context 1	how nice these frames are!
context 2	how nice these sunglasses are!
response	yes, they are the latest designs. would you like to try them on?

Table 1: The many-to-one dialogue pairs (multiple contexts with the same response) extracted from DailyDialog dataset.

We do agree that avoiding the one-to-many and many-to-one dialogue pairs is an effective way to improve the performance of dialogue generation models. However, avoiding such dialogue pairs

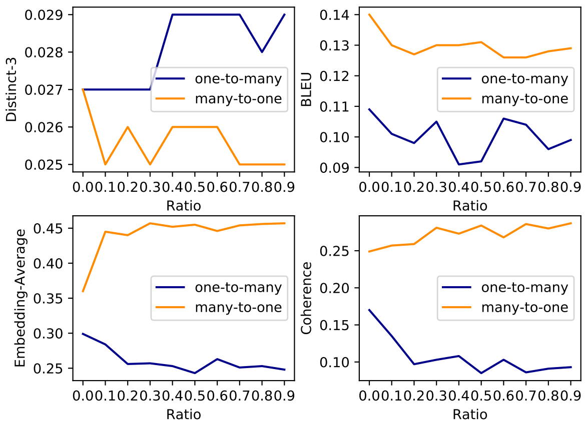

Figure 1: The validation results of responses generated by Seq2Seq model fine-tuned by one-to-many/many-to-one dialogue pairs.

cannot help the model learn the essential knowledge of the one-to-many and many-to-one phenomena in natural human conversation. Furthermore, some one-to-many and many-to-one dialogue pairs are beneficial to help the model in certain aspects. For example, Table 1 shows some many-to-one dialogue pairs extracted from DailyDialog (Li et al., 2017) dataset, which are not only not generic, but can also be used for helping models summarize the response patterns.

In addition, a simple and effective experiment can prove the above point of view. We first pre-trained a Sequence-to-sequence dialogue generation model (Seq2Seq) (Shang et al., 2015) without one-to-many/many-to-one dialogue pairs, and then fine-tuned the model with a certain percentage (0.1-0.9) of the original one-to-many/many-to-one dialogue pairs in the OpenSubtitles dataset (Lison and Tiedemann, 2016). Figure 1 shows the result of the experimental investigation regarding the influence of the ratio of one-to-many and many-to-one dialogue pairs in fine-tuning the Seq2Seq model. In Figure 1, the **Distinct** (Li et al., 2016a) represents the diversity of generated responses; the **BLEU** (Papineni et al., 2002) and the **Embedding-Average** (Liu et al., 2016) represent the difference between generated responses and ground-truth responses in word-overlap level and semantics level, respectively; and the **Coherence** (Xu et al., 2018b) represents the degree of correlation between the generated responses and the context. It can be noticed from this figure that the one-to-many dialogue pairs could increase the distinct of generated responses, but reduce the embedding-average and coherence of the generated responses. On the contrary, the many-to-one dialogue pairs could reduce the distinct, but increase the embedding-average and coherence. Moreover, the BLEU will be reduced while fine-tuning with both one-to-many and many-to-one dialogue pairs, which shows the difficult that training the models with these non-one-to-one dialogue pairs. Table 1 and Figure 1 demonstrate that one-to-many and many-to-one dialogue pairs are both beneficial and harmful to the performance of a dialog generation model. Therefore, except to avoid or filter these dialogue pairs, how to enable the model to effectively learn the essential and useful knowledge from these dialogue pairs while avoiding being affected by the disadvantages is a problem worthy of in-depth study.

To address such problems, we present a Sentence Semantic Segmentation guided Conditional Variational Auto-Encoder (SegCVAE). Inspired from the complexity and ambiguity of the language, we found that focusing on different words or word-combinations will highlight different semantic information that we called the prominent semantics from the original semantics. The prominent semantics could explain the one-to-many and many-to-one phenomena naturally: For one-to-many phenomenon, the multiple responses may corresponding to the multiple prominent semantics summarized from different words. In addition, for many-to-one phenomenon, the one response may corresponding to the similar prominent semantics in different contexts. Therefore, we propose the *internal separation* to extract multiple different word-combinations for obtaining such prominent semantics. However, also due to the ambiguity of word, such word-combinations may also have the unclear semantics. Thence, we propose the *external guidance* to obtain multiple instructive words from the vocabulary to constrain the semantic information of the extracted word-combinations. Finally, we use the word-combinations and the instructive words together to summary the prominent semantics, and then generate the response. Furthermore, to build the mapping between the prominent semantics and the response, we propose *semantic alienation norm*, *semantic centralization norm*, and *semantic distillation norm*), which are detailed in Section 4. Our contributions are as follow:

First, we proposed the novel SegCVAE to learn the essential knowledge from the one-to-many and many-to-one dialogue pairs. By using the sentence semantic segmentation, our SegCVAE can construct the mappings between the multiple responses and multiple different possible prominent semantics, thereby naturally explaining one-to-many and many-to-one phenomena. Then, we defined the one-to-many and many-to-one dialogue learning tasks, collected the One-to-Many (O2M) and Many-to-One (M2O) dialogue datasets, and presented some automatic evaluation strategies to assess the ability of the dialogue model on processing one-to-many and many-to-one dialogue pairs. Finally, we conducted extensive experiments to show the superior performance of our SegCVAE in dealing with one-to-many and many-to-one phenomena.

2 Related Work

The open-domain dialogue generation task has received extensive attention since 2014 (Sutskever et al., 2014; Shang et al., 2015; Sordoni et al., 2015). At that time, Sutskever et al. (2014) have identified that the *noisy* dialogue pairs, including one-to-many and many-to-one dialogue pairs, will affect the performance of the dialogue generation models. To address such *noisy* dialogue pairs and improve the performance of the dialogue model, more and more dialogue generation methods have been proposed in recent years. For instance, some methods design a scoring method and filter the *noisy* dialogue pairs (Xu et al., 2018b; Csaky et al., 2019; Akama et al., 2020); some methods introduce the external information to reduce the number of *noisy* dialogue pairs (Luong et al., 2015; Li et al., 2016b; Serban et al., 2016; Zhao et al., 2017; Huber et al., 2018; Ghazvininejad et al., 2018; Tao et al., 2018; Chen et al., 2018; Feng et al., 2020b); and some methods reconstruct the optimization function to avoid training dialogue models directly on such *noisy* dialogue pairs (Li et al., 2016c; Xu et al., 2017; Zhang et al., 2018a; Xu et al., 2018a; Zhang et al., 2018b; Feng et al., 2020a; Liu et al., 2020; He and Glass, 2020).

However, these methods cannot actually learn the essential knowledge of one-to-many and many-to-one dialogue pairs, nor can they make full use of the advantages of such dialogue pairs. For example, Csaky et al. (2019) uses the entropy, calculating based on the conditional probability, to assess the dialogue pairs (high entropy represents low score), which easily filters the one-to-many and many-to-one dialogue pairs before training; Li et al. (2016b) uses personal information to reduce the one-to-many dialogue pairs. They believed that different personal information with the same context will lead to different responses; The Reinforcement Learning based methods only require the generated response could get high reward rather than similar with the ground-truth, which means that some many-to-one dialogue pairs are ignored during training.

In addition to the methods illustrated above, the CVAE-based dialogue generation methods (Shen et al., 2017; Zhao et al., 2017; Chen et al., 2018; Gao et al., 2019; Sun et al., 2021) provide an idea to learn the essential knowledge of the one-to-many and many-to-one phenomena. They try to learn the knowledge into a latent space, a posterior probability distribution, and a prior probability distribution. By sampling latent variables form the latent space based on the probability distributions, the model could easily generate multiple responses for one context. Based on the advantages of the CVAE architecture in solving one-to-many and many-to-one phenomena, we proposed the SegCVAE, which uses the sentence semantic segmentation to regularize and guide the latent variables.

3 Task Definition

One-to-Many Dialogue Learning Let c be a context, and rs=$r_1, r_2, \dots, r_n$ be the responses to c. Follow the general dialogue generation task, we put the c and rs into n dialogue pairs $(c, r_1), (c, r_2), \dots, (c, r_n)$. Let $\mathcal{D}_{1n}$ be the dataset that only contains such one-to-many dialogue pairs. This task requires a dialogue generation model to learn the one-to-many knowledge, and to generate multiple coherent and informative responses for every context sentence.

Many-to-One Dialogue Learning Relatively speaking, let cs=$c_1, c_2, \dots, c_n$ be the contexts, and r be a response to the cs. Correspondingly, we use $\mathcal{D}_{n1}$ to represent a dataset that only contains many-to-one dialogue pairs $(c_1, r), (c_2, r), \dots, (c_n, r)$. This task requires the dialogue generation model to learn the many-to-one knowledge, and to distinguish which of the contexts can give the same response, and then increase the diversity while keeping the coherence of the generated response.

4 Sentence Semantic Segmentation guided CVAE

Overview This paper proposes the SegCVAE to study the relations of prominent semantics and one-to-many and many-to-one phenomena. SegCVAE uses multiple prominent semantics $(x_1, x_2, x_3, ...)$ to replace the original semantics to learn the probability distribution of latent variables and the generation process. To train our model, We introduce the *Stochastic Gradient Variational Bayes* framework (Kingma and Welling, 2014; Sohn et al., 2015; Yan et al., 2016) and *gradient blocking* trick (Sun et al., 2021):

$$\mathcal{L}(r, x^+) = \max_{i=1,2,3,...} \mathbf{E}_{q_\phi(z|r_e,x_i)}(\log p_{dec}(r|z, x_i)) \\ - KL(q_\phi(z|r_e, x_i)||p_\theta(z|x_i)), \quad (1)$$

The $q_\phi(z|r_e, x_i)$ and $p_\theta(z|x_i)$ are the recognition network and prior network that used for sampling the latent variable z. The $r_e = enc(r)$ is the semantic vector computed by model's encoder enc based on the response r. The dec is the model's decoder, generating the output token based on the conditional probability $p_{dec}(r|z, x_i)$.

To obtain the multiple prominent semantics, the SegCVAE employs the *internal separation* and *external guidance*. To make the prominent semantics meaningful, three novel semantic norms: *semantic alienation norm*, *semantic centralization norm*, and *semantic distillation norm* are proposed.

4.1 Internal Separation

The *internal separation* mainly focuses on extracting the multiple semantics from the context itself, which processes sentences through multiple triggers and extracts multiple sets of different word-combinations, which can be used to get different prominent semantics.

Each trigger consists of a convolution network $Conv$ and a dense network $Dense$. The input of the it is a embedded matrix representation $C_{max\_clen \times N}$ of a context, where max_clen represents the maximum length of a context that can be received and N is the dimension of the word-embedding. The $C_{max\_clen \times N}$ will be processed by the $Conv$ whose kernel K and stride S are $(m, N, 1, chan)$ and $(1, 1, 1, 1)$, respectively.

$$\mathcal{F}_c = Conv(C_{max\_clen \times N}, K, S), \quad (2)$$

where $chan$ is the number of channels of the convolution operation. After that, we can get the semantic features $\mathcal{F}_c$. According to the channel, we squeeze and transpose the $\mathcal{F}_c$ from $(max\_clen - m + 1, 1, chan)$ to $(chan, max\_clen - m + 1)$, and put it into the $Dense$ network. The weight of $Dense$ is $\mathcal{W}_{(max\_clen-m+1, max\_clen)}$, and the activation function of it is **SoftMax**:

$$\mathcal{F}_d = \textbf{SoftMax}(\mathcal{F}_c \times \mathcal{W}), \quad (3)$$

$$\textbf{SoftMax}: \ y_{ij} = \frac{e^{o_{ij}}}{\sum_1^k e^{o_{ik}}}, \quad (4)$$

where $y \in \mathcal{F}_d$ and $o \in (\mathcal{F}_c \times \mathcal{W})$. Hence, the shape of $\mathcal{F}_d$ is $(chan, max\_clen)$, which can represent the probability of words in the context of attention in different channels. Then, we select the word with the highest probability in each channel, which will be processed by the model's encoder to extract a certain semantic information. However, this discrete process will hamper the optimization of the model. In order to ensure the gradient back-propagation, we have introduced the reparameterization tricks (*i.e.* Gumbel SoftMax) to replace the **SoftMax** and selection process, which shown in Eq. 5:

$$\mathcal{F}_d' = \textbf{GumbelSoftMax}(\mathcal{F}_c \times \mathcal{W}), \quad (5)$$

$$\textbf{GumbelSoftMax}: \ y_{ij}' = \frac{e^{o_{ij}/\tau}}{\sum_1^k e^{o_{ik}/\tau}}, \quad (6)$$

where $y' \in \mathcal{F}_d'$ and τ is the temperature parameter. We can control the τ to be as small as possible so that the result of $\mathcal{F}_d'$ is as close as possible to the result of $argmax(\mathcal{F}_d)$.

Finally, we can get the embedded matrix representation of the extracted word-combination $C^{IS}_{chan \times N} = \mathcal{F}_d' \times C_{max\_clen \times N}$. Therefore, the *internal separation* can randomly initializes $\mathcal{M}$ triggers to extract $\mathcal{M}$ embedded matrix representations $(C^{IS,1}_{chan \times N}, C^{IS,2}_{chan \times N}, \dots, C^{IS,\mathcal{M}}_{chan \times N})$ of different word-combinations from a context.

4.2 External Guidance

The *external guidance* is responsible for extracting instructive information from the outside of the sentence (*i.e.* the vocabulary) according to the context semantics. To achieve this goal, we change the hyper-parameter of the dense network in the trigger that defined in the previous section. The new weight matrix of the dense in *external guidance* is $\mathcal{W}'$, whose shape is changed from $(max\_clen - m + 1, max\_clen)$ to $(max\_clen - m + 1, vocab\_size)$. The $vocab\_size$ is the size of the vocabulary. Hence, the results of the dense network represents the probability of words in the vocabulary of attention in different channels. Therefore, the output of *external guidance* is a matrix representation $V^{EG}_{chan \times N}$ of some words in vocabulary related to the semantics of the input sentence:

$$V^{EG}_{chan \times N} = \textbf{GumbelSoftMax}(\mathcal{F}_c \times \mathcal{W}') \times Em$$

where Em is the word-embedding matrix.

Finally, the *external guidance* also randomly initializes $\mathcal{M}$ new triggers to extract $V^{EG,1}_{chan \times N}$, $V^{EG,2}_{chan \times N}$, $\dots$, $V^{EG,\mathcal{M}}_{chan \times N}$.

Therefore, the $C^{IS}_{chan \times N}$ and the $V^{EG}_{chan \times N}$ are used together to calculate multiple different prominent semantics of a context. We concatenate them as $[(C^{IS,1}_{chan \times N}, V^{EG,1}_{chan \times N}), (C^{IS,2}_{chan \times N}, V^{EG,2}_{chan \times N}), \dots, (C^{IS,\mathcal{M}}_{chan \times N}, V^{EG,\mathcal{M}}_{chan \times N})]$, and input them into the enc to get the prominent semantics x_i.

$$x_i = enc((C^{IS,i}_{chan \times N}, V^{EG,i}_{chan \times N})),\ i = 1, \dots, \mathcal{M}$$

4.3 Semantic Norms

We introduce the self-supervise learning ideas, and propose *semantic alienation norm*, *semantic centralization norm*, and *semantic distillation norm*, to constrain the relations between the multiple prominent semantics and the responses.

Semantic Alienation Norm We first propose the *semantic alienation norm* to make each prominent semantics as different as possible from other prominent semantics, which is computed by:

$$\mathcal{L}_{san} = |\mathcal{I}_{\mathcal{M} \times \mathcal{M}} - \textbf{SoftMax}(x_{\mathcal{M} \times N} \times x^T_{\mathcal{M} \times N})|$$

The $\mathcal{I}_{\mathcal{M} \times \mathcal{M}}$ is an identity matrix, and $x_{\mathcal{M} \times N} = concatenate([x_1, x_2, \dots, x_{\mathcal{M}}])$ is the context vectors calculated by the model's encoder enc. The x_i represents one certain semantic vector among $\mathcal{M}$ prominent semantic vectors, so the $x_{\mathcal{M} \times N} \times x^T_{\mathcal{M} \times N}$ can represent the correlation between a certain prominent semantic vector and other prominent semantic vectors.

Semantic Centralization Norm Then we propose the *semantic centralization norm* to ensure the ensemble result of these prominent semantic vectors ($[x_1, x_2, \dots, x_{\mathcal{M}}]$) is similar with the semantics of the original context.

$$\mathcal{L}_{scn} = \mathbf{1} - cosine(enc(C_{max\_len \times N}), \sum_i^{\mathcal{M}} x_i)$$

Semantic Distillation Norm Finally, we propose the *semantic distillation norm*, which uses the relationship knowledge among the ground-truth response to teach the model to learn the semantic relation of these prominent semantic information.

$$\mathcal{L}_{sdn} = KL(\textbf{SoftMax}(C_{B \times N} \times C^T_{B \times N}) || \textbf{SoftMax}(\mathbf{X}^+ \times \mathbf{X}^{+T})),$$

where the $C_{B \times N}$ represents the semantic matrix of batch size B ground-truth responses obtained by the model's encoder enc. And the $\mathbf{X}^+$ is the concatenated result of B positive prominent semantic information x^+ obtained by gradient blocking.

4.4 Likelihood Function

Therefore, the final likelihood function that is used for training our model is:

$$\mathcal{L}_{all} = \mathcal{L}(r, x^+) - \mathcal{L}_{san} - \mathcal{L}_{scn} - \mathcal{L}_{sdn}, \quad (7)$$

where $\mathcal{L}(r, x^+)$ is shown in Eq (1).

5 Experiment[1]

Data Setting We use the processed OpenSubtitles (Lison and Tiedemann, 2016) dataset that proposed by Sun et al. (2021) for general dialogue generation task, which has 5M, 100K, and 50K dialogue pairs in training, validation and, test set, respectively. Meanwhile, we also extract two special datasets from the original OpenSubtitles: One-to-Many and Many-to-One, for our Non-One-to-One dialogue learning tasks. To build these two datasets, we first extract single-turn dialogues from the OpenSubtitles: $T-1$ single-turn dialogues $[(u_1, u_2), (u_2, u_3), ..., (u_{T-1}, u_T)]$ can be extracted from one multi-turn dialogue $(u_1, u_2, ..., u_T)$, where u represents an utterance in each dialogue. Then, we selected and collected a large collection of one-to-many dialogue pairs as the One-to-Many (O2M) dataset, and another large collection of many-to-one dialogue pairs as the Many-to-One (M2O) dataset. Finally, we use the token-list of GloVe (Pennington et al., 2014) to filter the O2M and M2O datasets. For each dialogue pair (context c_i, response r_i), we first obtain its tokens after word segmentation, and then judge whether its tokens are all contained in GloVe's token-list. If the GloVe do not contain any tokens of (c_i, r_i), we drop all dialogue pairs containing the c_i or r_i from the dataset. Table 2 lists key statistics of the dataset after processing.

[1]See Appendix A for other experiment settings.

dataset	type	# tokens	# pairs	# contexts(c)	# responses(r)	avg # r	avg # c	max # r	max # c
O2M	training	40,875	778,658	284,516	778,658	2.74	-	1,546	-
	validation	-	222,126	81,057	222,126	2.74	-	689	-
	test	-	110,446	40,710	110,446	2.71	-	497	-
M2O	training	40,331	768,183	768,183	279,978	-	2.74	-	1,588
	validation	-	217,474	217,474	79,552	-	2.73	-	957
	test	-	109,815	109,815	39,795	-	2.76	-	321

Table 2: Statistics for One-to-Many (O2M) and Many-to-One (M2O) datasets. The **# tokens** is the vocabulary size, and the **# pairs/contexts/responses** is the number of the dialogue pairs/contexts/responses in datasets. The **avg/max # r** is the average/maximum number of responses for each context, and the **avg/max # c** is the average/maximum number of contexts for each response. "-" means the cell is not necessary for this **type/dataset**.

model	ppl	Distinct-1	Distinct-2	Length	BLEU-1	BLEU-2	BLEU-3	Average	Coherence
Seq2Seq	45.9±.13	0.002±.00	0.010±.00	11.8±.81	0.236±.04	-	-	0.465±.08	0.281±.05
CVAE+BOW	12.2±.17	0.005±.00	0.095±.00	13.1±.26	0.172±.02	-	-	0.285±.04	0.195±.03
K-CVAE+BOW	12.1±.20	0.006±.00	0.098±.00	13.1±.10	0.203±.02	-	-	0.311±.06	0.200±.05
SepaCVAE	2.0±.06	0.016±.00	0.282±.01	12.6±.11	0.417±.00	-	-	0.836±.01	0.707±.01
SegCVAE	3.0±.09	0.011±.00	0.232±.01	12.4±.10	0.412±.01	0.339±.01	0.287±.00	0.842±.00	0.719±.01
Seq2Seq	-	0.003±.00	0.015±.00	11.8±.82	-	0.193±.03	0.163±.03	0.465±.08	0.281±.05
CVAE+BOW	-	0.009±.00	0.131±.00	13.1±.24	-	0.144±.02	0.123±.02	0.285±.04	0.195±.03
K-CVAE+BOW	-	0.010±.00	0.135±.00	13.1±.10	-	0.169±.02	0.144±.01	0.308±.06	0.198±.05
SepaCVAE	-	0.025±.00	0.330±.03	13.5±.58	-	0.326±.01	0.276±.01	0.807±.02	0.677±.01
SegCVAE	-	0.021±.00	0.323±.01	14.4±.80	0.437±.01	0.364±.01	0.310±.01	0.836±.00	0.707±.01

Table 3: Mterics results on validation data (up) and test data (down) of the OpenSubtitles dataset. The best score in each column is marked with underline. Note that our BLEU-1,2,3 scores are normalized to [0, 1]. - represents the result is not calculated or not published in the reference.

Evaluation Strategy for Non-one-to-one Tasks The non-one-to-one tasks require the new strategies to apply the automatic evaluation metrics.

Diversity: This is mainly used to evaluate whether the model can learn the ability to generate multiple diverse responses. Therefore, we assess the diversity by calculating the distinct-n of multiple generated responses $[\hat{r_1}, \hat{r_2}, \ldots, \hat{r_\mathcal{M}}]$ generated based on one context:

$$\textbf{Diversity} = \frac{\textbf{unique}(Tokens_{[\hat{r_1}, \hat{r_2}, \ldots, \hat{r_\mathcal{M}}]})}{Tokens_{[\hat{r_1}, \hat{r_2}, \ldots, \hat{r_\mathcal{M}}]}}$$

Word consistency: We use the maximum BLEU of each ground-truth response and multiple generated responses to represent the word consistency:

$$\textbf{WordCons} = \frac{1}{\mathcal{R}s} \sum_{i=1}^{\mathcal{R}s} \max_{j=1,\cdots,\mathcal{M}} (Bleu(r_i, \hat{r}_j)),$$

where $\mathcal{R}s$ is the number of the ground-truth responses $(r_1, r_2, \ldots, r_{\mathcal{R}s})$ for the context, and the $\mathcal{R}s = 1$ for Many-to-One task.

Semantics consistency: We use the maximum embedding-average value of each ground-truth response and multiple generated responses to represent the semantics consistency.

Complex coherence: We use the ratio of the average coherence between the context and generated responses and that between the same context and ground-truth responses to evaluate the complex coherence of the model:

$$\textbf{CompCohe} = \frac{\sum_{i=1}^{\mathcal{M}} cohence(c, \hat{r}_i)/\mathcal{M}}{\sum_{j=1}^{\mathcal{R}s} cohence(c, r_j)/\mathcal{R}s}$$

The bset CompCohe should be close to 1.0, which means that the model has learned the semantic relationship between the context and the true response.

6 Results and Analysis

General Dialogue Generation Task Table 3 reports the automatic results of SegCVAE and baseline models on validation and test data of the OpenSubtitles dataset. These results show that our SegCVAE achieves the best performance in terms of **BLEU**, **Average**, and **Coherence**, which demonstrates the superior performance of our model on generating coherent and semantically related responses. In addition, the **Distinct** of our SegCVAE is far superior to Seq2Seq, CVAE+BOW and **K**-CVAE+BOW, and is closer to the state-of-the-art SepaCVAE, which illustrates the ability of our model in generating diverse responses.

model	ppl	Distinct-1	Distinct-2	length	BLEU-1	BLEU-2	Average	Coherence
CVAE+BOW	15.79±.22	0.003±.000	0.050±.007	12.18±.13	0.425±.006	0.346±.005	0.849±.005	0.738±.007
K-CVAE+BOW	15.72±.10	0.003±.001	0.045±.008	12.04±.18	**0.448±.006**	0.360±.005	**0.865±.005**	0.742±.008
SepaCVAE	2.49±.02	**0.006±.000**	0.185±.006	**12.63±.22**	0.432±.002	0.354±.002	0.846±.006	0.712±.014
SegCVAE	3.58±.10	0.005±.000	**0.145±.011**	12.26±.11	0.441±.017	**0.361±.013**	0.848±.002	**0.714±.005**
GroundTruth	0.0	0.0103	0.1315	12.49	1.0	1.0	1.0	0.7078
CVAE+BOW	11.10±.09	0.002±.000	0.032±.004	9.35±.03	0.424±.003	0.338±.003	0.843±.003	0.743±.004
K-CVAE+BOW	11.15±.11	0.002±.000	0.032±.001	9.28±.20	**0.451±.001**	0.357±.002	**0.858±.001**	0.741±.003
SepaCVAE	3.03±.02	**0.005±.000**	0.137±.012	**9.56±.23**	0.449±.009	**0.358±.007**	0.830±.012	0.685±.024
SegCVAE	4.66±.13	0.003±.001	**0.077±.007**	9.75±.44	0.413±.010	0.332±.007	0.839±.002	**0.716±.006**
GroundTruth	0.0	0.0093	0.0792	9.57	1.0	1.0	1.0	0.7077

Table 4: Metrics results on validation data of O2M (up) and M2O (down). The score closest to the GroundTruth in each column is shown in bold. The best score in each column is marked with underline.

model	Diversity-1	Diversity-2	Diversity-3	WordCons	SemaCons	CompCohe	MaxCohe	MinCohe
CVAE+BOW	0.007±.001	0.078±.009	0.280±.023	0.318±.001	**0.901±.001**	**1.017±.011**	0.828±.002	0.593±.023
K-CVAE+BOW	0.007±.001	0.070±.010	0.262±.023	0.313±.002	0.898±.002	1.039±.013	0.837±.003	**0.626±.021**
SepaCVAE	**0.015±.001**	0.261±.022	0.694±.029	**0.318±.004**	0.894±.002	0.953±.043	0.810±.010	0.493±.094
SegCVAE	0.012±.001	**0.193±.009**	**0.626±.011**	0.315±.003	0.895±.000	0.973±.001	**0.798±.000**	0.554±.002
GroundTruth	0.0341	0.2244	0.5073	1.0	1.0	1.0	0.7822	0.6965
CVAE+BOW	0.002±.000	0.032±.004	0.144±.009	0.313±.000	**0.901±.000**	2.669±.033	0.830±.001	0.604±.011
K-CVAE+BOW	0.002±.000	0.032±.001	0.144±.003	0.309±.002	0.896±.000	2.598±.083	0.832±.001	**0.608±.008**
SepaCVAE	**0.005±.000**	**0.130±.011**	0.466±.022	**0.315±.001**	0.893±.003	2.436±.096	0.807±.005	0.470±.072
SegCVAE	0.004±.001	0.072±.007	**0.328±.018**	0.309±.003	0.893±.001	**2.421±.074**	**0.803±.002**	0.564±.012
GroundTruth	0.0250	0.1381	0.2838	1.0	1.0	1.0	0.7352	0.7352

Table 5: Mterics results on test data of O2M (up) and M2O (down). The score closest to the GroundTruth in each column is shown in bold. The best score in each column is marked with underline.

model	Informativeness	Relevance	Erudition
CVAE+BOW	3.19	2.20	2.33
K-CVAE+BOW	3.40	2.11	2.35
SepaCVAE	1.52	2.79	2.21
SegCVAE	1.79	2.28	1.89
CVAE+BOW	2.84	2.00	1.96
K-CVAE+BOW	3.13	1.83	1.89
SepaCVAE	1.79	2.53	1.92
SegCVAE	2.00	2.11	1.92

Table 6: Human evaluation results on test data of O2M (up) and M2O (down). The best score in each column is marked with underline.

One-to-Many and Many-to-One Dialogue Learning Tasks To evaluate whether the model has learned the knowledge of one-to-many and many-to-one phenomena, we not only underlined the best scores, but also bolded the scores that are closest to the ground-truth in Table 4 and 5. As can be seen, our SegCVAE is closer to the information collected in the dataset in terms of coherence and distinct, which proves to a certain extent that our model can learn some specific knowledge from the dataset. In Table 4 and 5, the performance of CVAE+BOW and **K**-CVAE+BOW is greatly improved compared to Table 3, which is due to the presence of noise in the O2M and M2O dataset. When we checked the dataset, we found that there are samples with the same semantics but different performance, such as "is that" and "ls that", "ok" and "okay", etc. These samples make the difference between the maximum and minimum coherence getting smaller, resulting in a concentrated prior distribution. This increases the coherence and relevance performance of CVAE+BOW and **K**-CVAE+BOW but decrease the diversity of them.

Human Evaluation This result is shown in Table 6. As discussed above, the CVAE+BOW and **K**-CVAE+BOW sample latent variables from a concentrated prior distribution, which leads high relevance but low informativeness. The SepaCVAE using the orthogonal vectors for sampling latent variables, which increases the informativeness but decreases the number of relevant responses. Our SegaCVAE generates multiple responses based on multiple prominent semantics, resulting in a proper result. Moreover, SegaCVAE achieves the best Erudition score, which demonstrates the superior ability of it in handling one-to-many samples. Following the existing work (Xu et al., 2018a; Feng et al., 2020a), the Pearson's correlation coefficient is 0.83 on Informativeness, 0.55 on Relevance, and 0.51 on Erudition, with $p < 0.0001$ and below 0.001, which indicates high correlation and agreement.

model	Diversity-1	Diversity-2	Diversity-3	WordCons	SemaCons	CompCohe	MaxCohe	MinCohe
SegCVAE	0.012±.001	0.193±.009	0.626±.011	0.315±.003	0.895±.000	0.972±.001	0.798±.001	0.554±.002
-wo. IS	0.011±.001	0.156±.031	0.506±.076	0.317±.002	0.892±.001	0.942±.011	0.790±.002	0.508±.019
-wo. EG	0.012±.001	0.183±.009	0.598±.019	0.316±.000	0.896±.000	0.979±.010	0.801±.001	0.547±.048
-wo. $\mathcal{L}_{san}$	0.013±.001	0.218±.022	0.655±.024	0.318±.001	0.895±.002	0.980±.022	0.801±.004	0.568±.044
-wo. $\mathcal{L}_{scn}$	0.011±.002	0.187±.023	0.596±.072	0.315±.002	0.895±.001	0.969±.010	0.801±.001	0.519±.062
-wo. $\mathcal{L}_{sdn}$	0.013±.001	0.200±.020	0.621±.026	0.314±.003	0.892±.003	0.932±.041	0.792±.006	0.485±.077
SegCVAE	0.004±.001	0.072±.007	0.328±.018	0.309±.003	0.893±.001	2.421±.074	0.803±.002	0.564±.012
-wo. IS	0.003±.000	0.058±.010	0.270±.041	0.306±.005	0.892±.003	2.358±.066	0.802±.003	0.564±.026
-wo. EG	0.003±.000	0.064±.001	0.314±.006	0.314±.003	0.895±.000	2.523±.029	0.809±.001	0.589±.010
-wo. $\mathcal{L}_{san}$	0.004±.000	0.071±.000	0.299±.005	0.304±.003	0.892±.000	2.149±.088	0.800±.002	0.471±.022
-wo. $\mathcal{L}_{scn}$	0.003±.001	0.058±.019	0.238±.095	0.301±.006	0.889±.005	2.075±.456	0.799±.006	0.279±.239
-wo. $\mathcal{L}_{sdn}$	0.004±.001	0.078±.009	0.337±.016	0.311±.000	0.894±.001	2.418±.048	0.803±.002	0.542±.039

Table 7: Ablation results on test data of O2M (up) and M2O (down).

Context	I'm sorry, you're mistaken.
EG	Confided Confided <pad>
IS	I Mistaken <pad>
SegCVAE	**So, I'll help** my mate and **you**. listen, one day to tell me to go from the fields together.
Context	Move! What have you done?
EG	Rendezvous Humiliate <pad>
IS	Move ! <pad>
SegCVAE	Hey, **please. relax**.
Context	Not this year, dani. Mom said you have to.
EG	Tying <pad> Tying
IS	Said Not Said
SegCVAE	I'm compounded you **talk about our great <unk> in the other times.**

Table 8: Generated responses and their corresponding keyword-combinations of SegCVAE. EG and IS represent the External Guidance and the Internal Separation. Note that the results of EG and IS are used for extracting prominent semantics.

Context	I'd rather die than live with you! freaking unk!
Responses	Relax! where does it hurt? Stop! ma'am, ma'am!
CVAE+BOW	I'm gonna get you to know! That's a bad idea, mister. I have a hell! It's a joke that you said he's a special agent! why do you want me to believe? You have something to do with this? aah. Hey, you're ready? yeah. The world's in the mood! Here, put your hands in the bowl.
SegCVAE	Yep tonight really... to me. sean? **Calm down.** hurry any, hurry unk. Nothing, they are hot / hey, **No-no,** your unk. i... God? uh... did not fit... Be it then let's abandon it. 9 pigs. 1 50,000. open. Really is going with nothing? all unk came in the past hours. Most way. hell and i are unk

Table 9: Generated responses from the baseline and SegCVAE on O2M dataset. Note that the generated "Calm down." and "No-no," are corresponding to the "Relax!" and "Stop" in true responses.

Ablation Study Table 7 reports the ablation results of the SegCVAE. As can be seen, the Internal Separation (IS) and External Guidance (EG) mainly affect the performance of the model, while the semantic norms (*i.e.* $\mathcal{L}_{san}$, $\mathcal{L}_{scn}$, and $\mathcal{L}_{sdn}$) mainly affects the stability of the model.

Case Study Table 8 reports several generated samples and their related word-combinations. Table 9 and Table 10 (in Appendix B) show two samples of the generated responses of contexts in test set of O2M and M2O datasets This result illustrates that the SegCVAE could effectively build the relations between the multiple prominent semantics and the multiple responses.

7 Conclusion and Future Outlook

This paper mainly focuses on the one-to-many and many-to-one phenomena in dialogue generation task. Therefore, we present the one-to-many and many-to-one dialogue learning tasks, collect two datasets, and provide multiple automatic evaluation strategies. Futuremore, we also propose the SegCVAE, which has three novel components: internal separation, external guidance and semantic norms. SegCVAE uses the sentence semantic segmentation to analyze and learn the essential knowledge of one-to-many and many-to-one phenomena. As demonstrated in the experimental results, the SegCVAE could learn the essential knowledge of one-to-many and many-to-one phenomena, and uses such knowledge to handle these two tasks better than the baseline models. In future, we plan to (1) clean the O2M and M2O data sets; (2) study new semantic segmentation approaches; (3) study new Non-One-to-One dialogue learning frameworks.

References

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *EMNLP*, pages 941–958.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *WWW*, pages 1653–1662. ACM.

Richard Csaky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *ACL (1)*, pages 5650–5669.

Shaoxiong Feng, Hongshen Chen, Kan Li, and Dawei Yin. 2020a. Posterior-gan: Towards informative and coherent response generation with posterior generative adversarial network. In *AAAI*, pages 7708–7715.

Shaoxiong Feng, Xuancheng Ren, Hongshen Chen, Bin Sun, Kan Li, and Xu Sun. 2020b. Regularizing dialogue generation by imitating implicit scenarios. In *EMNLP*, pages 6592–6604.

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 *EMNLP-IJCNLP*, pages 1898–1908. Association for Computational Linguistics.

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 *AAAI*, pages 5110–5117.

Tianxing He and James R. Glass. 2020. Negative training for neural dialogue response generation. In *ACL*, pages 2044–2058.

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

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

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

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016b. A persona-based neural conversation model. In *ACL (1)*.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP(1)*, pages 986–995.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *LREC*.

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

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *ACL*, pages 1417–1427.

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

Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *CoRR*, abs/1703.01619.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

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

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*, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL (1)*, pages 1577–1586.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *ACL (2)*, pages 504–509.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *NIPS*, 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 *HLT-NAACL*, pages 196–205.

Bin Sun, Shaoxiong Feng, Yiwei Li, Jiamou Liu, and Kan Li. 2021. Generating relevant and coherent dialogue responses using self-separated conditional variational autoencoders. In *ACL/IJCNLP*, pages 5624–5637. ACL.

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

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018a. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *EMNLP*, pages 3940–3949.

Xinnuo Xu, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018b. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *EMNLP*, pages 3981–3991.

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

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. In *ECCV (4)*, volume 9908 of *Lecture Notes in Computer Science*, pages 776–791.

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

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*, pages 1815–1825.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL (1)*, pages 654–664.

A Experiment Settings

Automatic Evaluation Metrics We use **Distinct-n**, **BLEU**, **Embedding Average (Average)**, and **Coherence** introduced in the Section 1 to assess our model and baseline models. In addition, we also employ the **Perplexity (ppl)** (Neubig, 2017) and **Response length** (Csaky et al., 2019): **ppl** is an indicator commonly used in dialogue generation tasks, is usually used to evaluate the degree of convergence of the model. **Response length** is the average number of words of all generated responses.

Human Evaluation We conduct human evaluation to further evaluate our model and baseline models. First of all, each model received 50 identical contexts randomly extracted from the test sets of the two dialogue datasets respectively, and generated 400 responses. Then, three annotators were invited to rank our SegCVAE and baseline models with respect to three aspects of their generated responses: **Informativeness**, **Relevance** and **Erudition**. Ties are allowed. **Informativeness** indicates how much diverse and informative responses are provided by the generative models. **Relevance** means how many generated responses are relevant to the context. **Erudition** specifies whether multiple generated responses have the same information and semantics as the ground-truth responses.

Baseline Models We compare our model with several state-of-the-art generative dialogue models: A sequence-to-sequence (Seq2Seq) (Shang et al., 2015; Sordoni et al., 2015), a general CVAE (Shen et al., 2017), a knowledge guide CVAE (Zhao et al., 2017), and a self-separated CVAE (Sun et al., 2021) are used as the baselines in our experiment. Due to the lack of the knowledge information, we introduce the cluster method (*i.e.* K-means(**K**)), and use the cluster results as the knowledge.

Training Details For a fair comparison, we used the 300-dimensional GloVe embeddings as the word-embedding matrix. The hidden size of all models are set to 300. The maximum length of context and response are set to 25. We set the batch sizes to 32 for all datasets (OpenSubtitles, O2M, and M2O). Adam is utilized for optimization. The initial learning rate is set to 0.001. We train all models in 50 epochs on a RTX 2080Ti GPU card with Tensorflow, and save the generated responses when the **ppl** reaching minimum. The random seed is set as 123456. Greedy search is used to generate responses for evaluation.

B Several Cases

Context	A sacrifice that the island demanded. excuse me?
CVAE+BOW	Why? because it's only a strange. No, really. what were you talking about? No. this is my job. It's a unk. i was a member of the united states states states. When you've been here, i will get back to your senses. you must have it. I'm not sure. you know why? What are your parents? he's just gonna take his place after your marriage, he lives. Why? because it's just like that.
SegCVAE	Yes, unk. yes. Pretty much, unk. we're looking. Yeah. a kid that call it before you put him off. Then everybody in red. there's tom. That's disgusting. brother! Everyone, that's in a way and that brain's trapped in strength feelings, but all holy unk. I take. he said i was dead. In her is the master. she's the.

Table 10: Generated responses from the baseline and SegCVAE on M2O dataset.