CoLV: A Collaborative Latent Variable Model for Knowledge-Grounded Dialogue Generation

Haolan Zhan[1]*, **Lei Shen**[1,2]*, **Hongshen Chen**[3]†, and **Hainan Zhang**[3]
[1] University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[3]Data Science Lab, JD.com, Beijing, China
zhanhaolan316@gmail.com, shenlei17z@ict.ac.cn, ac@chenhongshen.com

Abstract

Knowledge-grounded dialogue generation has achieved promising performance with the engagement of external knowledge sources. Typical approaches towards this task usually perform relatively independent two sub-tasks, i.e., knowledge selection and knowledge-aware response generation. In this paper, in order to improve the diversity of both knowledge selection and knowledge-aware response generation, we propose a collaborative latent variable (CoLV) model to integrate these two aspects simultaneously in separate yet collaborative latent spaces, so as to capture the inherent correlation between knowledge selection and response generation. During generation, our proposed model firstly draws knowledge candidate from the latent space conditioned on the dialogue context, and then samples a response from another collaborative latent space conditioned on both the context and the selected knowledge. Experimental results on two widely-used knowledge-grounded dialogue datasets show that our model outperforms previous methods on both knowledge selection and response generation.

1 Introduction

Knowledge-grounded dialogue generation (Liu et al., 2018; Zhou et al., 2018a; Lian et al., 2019; Tian et al., 2020), which utilizes external knowledge to enhance conversation backgrounds, has achieved promising performance. To exploit external knowledge efficiently for conversations, typical approaches (Dinan et al., 2019; Kim et al., 2020; Xu et al., 2020; Sun et al., 2020; Chen et al., 2020; Meng et al., 2020; Chen et al., 2021) tend to decompose this task into two streamlined sub-tasks: knowledge selection and knowledge-aware response generation. Besides, some other work (Qin et al., 2019; Tian et al., 2020) also tries to integrate these two sub-tasks in a unified memory-augmented training framework. In both paradigms, knowledge selection plays an important role in the knowledge-grounded dialogue systems.

Dialogue context	What is your favorite number? → I love the number 7. What do you think about that?
Knowledge candidates	**1**. Anyone who dares to kill Cain "will suffer vengeance seven times over". **2**. Seven is the natural number following six and preceding eight. **3**. Islam first came to the western coast when Arab traders as early as the 7th century CE. **4**. The number 7 has been associated with a great deal of symbolism in religion. In western culture, it is often considered lucky. ..... **N**. This genre has been popular throughout the history of culture.
Response a	Yeah. I know that it is before 8 and after 6!
Response b	Yes, it is known as a lucky number in western countries!
Response c	I think 7 is lucky certain cultures. It also depicts some religious importance.

Table 1: An example of knowledge-grounded conversations. Given the dialogue context, knowledge selection and response generation are inherently coupled. Besides, while knowledge selection is diverse, the knowledge-aware response generation could also be diverse based on the same knowledge content. Knowledge No.2 and No.4 are appropriate to the dialogue. Besides, given the same knowledge No.4, both Response *b* and *c* are appropriate.

Observing that the diversity of knowledge selection (given a dialogue context, several pieces of knowledge are appropriate) can be dramatically raised from prior and posterior distributions over knowledge, recent studies (Lian et al., 2019; Kim et al., 2020; Chen et al., 2020) utilize posterior mechanism to select knowledge during training phase. KL loss (Kullback and Leibler, 1951) is employed as one of the training objectives to minimize the gap between training and inference procedure, since posterior information is absent at inference. Kim et al. (2020) enhances this framework with sequential latent variables and Chen et al. (2020) proposes a knowledge distillation training strategy to further bridge the gap between prior and posterior information.

*First two authors contribute equally.
†Corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2250–2261
November 7–11, 2021.

While the success of variational knowledge selection is indisputable, there still exists some challenges that impede the conversational models from selecting appropriate knowledge. **Firstly**, knowledge selection is inherently coupled with knowledge-aware response generation. However, previous methods mostly emphasize the importance of knowledge selection without explicitly modeling the correspondence between the selected knowledge and the generated response. In Table 1, knowledge No.2 (in blue) corresponds to response a (in blue), while knowledge No.4 (in red) is related to response b and c (in red). **Secondly**, the diversity of knowledge selection is effectively improved with variational inference, while the diversity of knowledge-aware response generation (given the selected knowledge, several suitable responses can be generated) is still neglected. As shown in Table 1, response b and c are two different responses that share the same piece of knowledge, i.e., No.4.

In this paper, in order to simultaneously improve the diversity of both knowledge selection and knowledge-aware response generation, we propose a **Co**llaborative **L**atent **V**ariable (CoLV) model to integrate both aspects in separate yet collaborative latent spaces, so as to capture the inherent correlation between knowledge selection and response generation. During generation, our proposed model firstly draws knowledge candidate from the latent space conditioned on the dialogue context, and then samples a response from another collaborative latent space conditioned on both the context and the selected knowledge. Experimental results on two widely-used datasets of knowledge-grounded dialogue generation show that our model outperforms previous methods on both knowledge selection and response generation. Further analysis on collaborative latent variables demonstrates CoLV model's ability to not only improve the diversity of knowledge selection but also generate coherent and diverse responses.

2 Related Work

Our work is mainly related to two research branches: knowledge-grounded dialogue generation and variational auto-encoder learning.

Knowledge-grounded Dialogue Generation has raised broad interest and also has been greatly advanced by many new datasets (Zhou et al., 2020; Wu et al., 2019; Dinan et al., 2019; Moghe et al., 2018; Zhou et al., 2018b). Existing methods on this task mainly focus on resolving two research problems: knowledge selection (KS) and knowledge-aware response generation. Dinan et al. (2019) proposed a memory network to retrieve knowledge and combined it with a Transformer-based model to generate response. External knowledge base was also utilized to facilitate the utterance understanding and knowledge selection (Wang et al., 2020; Zheng et al., 2020). Lin et al. (2020) used memory network and copy mechanism to keep deep interaction between knowledge and utterances. Meng et al. (2020) employed a dual learning paradigm to enhance knowledge interaction. Su et al. (2020) proposed to augment the dialogue generation by utilizing external non-conversational text, which is effective but also introduce noise. Li et al. (2020) and Zhan et al. (2021) proposed to employ pre-training methods on the structured/unstructured knowledge representation and fine-tune the model using the limited knowledge-grounded training examples. Other work took efforts on utilizing future information. Lian et al. (2019) firstly employed posterior network, while Kim et al. (2020) further utilized sequential characteristics of knowledge. Besides, to further bridge the gap between prior and posterior network, Chen et al. (2020) and Chen et al. (2021) devised specific posterior information prediction modules. Hereby, our proposed CoLV model differs from previous work by utilizing collaborative latent variables to model the distributions of knowledge and response simultaneously.

VAE Learning (Kingma and Welling, 2014) is widely used in a variety of natural language processing tasks, including machine translation (Zhang et al., 2016), question answering (Lee et al., 2020), and conversations (Serban et al., 2017; Shen et al., 2019; Li et al., 2020; Shen et al., 2021). The core idea of variational auto-encoder is to utilize the advantage of posterior information or external information during training phase, and optimize the objectives by minimizing the KL divergence (Kullback and Leibler, 1951). Unlike previous work that applied VAEs on dialogue generation (Wu et al., 2020; Serban et al., 2017; Qiu et al., 2019), we aim at using collaborative latent variables to connect the external knowledge, dialogue context, and response, which will further enhance the correlation between knowledge selection and response generation. To the best of our knowledge, our method takes the first attempt to collaboratively model these two different distribu-

tions for knowledge-grounded conversations.

3 Proposed Model

3.1 Task Formulation

Our goal is to simultaneously improve the diversity of knowledge selection and generate diverse knowledge-aware responses. Formally, given a dialogue context $\mathbf{c}$ which contains $|c|$ tokens, $\mathbf{c} = \{c_1, ..., c_{|c|}\}$, and its corresponding knowledge pool KP, which contains $|k|$ knowledge candidate sentences, $KP = \{k_1, ..., k_{|k|}\}$. Each knowledge sentence $k_i \in KP$ contains M tokens, $k_i = \{k_i^1, ..., k_i^M\}$. Our goal has two main steps: (1) selecting the most relevant knowledge sentence $\mathbf{k}$ from knowledge pool KP based on dialogue context. (2) Then, generating a response $\mathbf{r} = \{r_1, ..., r_{|r|}\}$ with $|r|$ tokens, based on the dialogue context $\mathbf{c}$ and the selected knowledge $\mathbf{k}$. We aim at tackling this task by learning the conditional collaborative latent distributions of the knowledge selection and response generation given the dialogue context, which can be formulated as follows: $(\mathbf{k}, \mathbf{r}) \sim p(\mathbf{k}, \mathbf{r}|\mathbf{c})$, We estimate the collaborative distribution $p(\mathbf{k}, \mathbf{r}|\mathbf{c})$ by employing a collaborative latent variable model, named as CoLV model.

3.2 CoLV Framework

Our proposed CoLV model tends to model the conditional collaborative distribution $p(\mathbf{k}, \mathbf{r}|\mathbf{c})$ in relatively separate ye collaborative latent spaces for knowledge and response, which is defined as follows:

$$p_\theta(\mathbf{k}, \mathbf{r}|\mathbf{c}) = \int_{\mathbf{z_k}} \sum_{\mathbf{z_r}} p_\theta(\mathbf{k}|\mathbf{z_k}, \mathbf{c}) p_\theta(\mathbf{r}|\mathbf{z_r}, \mathbf{k}, \mathbf{c}) \cdot p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c}) p_\phi(\mathbf{z_k}|\mathbf{c}) d\mathbf{z_k},$$

where $\mathbf{z_k}$ and $\mathbf{z_r}$ are latent variables for knowledge and response respectively, and the $p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c})$ and $p_\phi(\mathbf{z_k}|\mathbf{c})$ are their conditional prior distributions. Specifically, a Gaussian distribution (Kingma and Welling, 2014) and a categorical distribution (Jang et al., 2017), are employed for $\mathbf{z_r}$ and $\mathbf{z_k}$ respectively, as shown in Figure 1. Knowledge selection is a discriminative task, which is suitable to be modeled by a Categorical distribution. Besides, Gaussian distribution is continuous, which is appropriate to model the response latent variable. As shown in Figure 1, we devise a mutual interaction of these collaborative latent variables for knowledge and response separately.

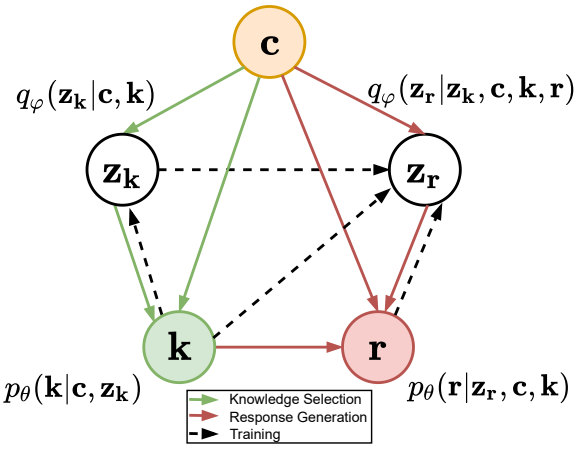


Figure 1: The graphical framework for CoLV model. $\mathbf{c}$: dialogue context, $\mathbf{k}$: knowledge, $\mathbf{r}$: response. The dotted line denotes training procedure solely, while the solid line denotes both training and inference process.

To construct the collaborative latent variables, we enforce the response latent space to be dependent on the knowledge latent space in $p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c})$, while the knowledge latent space is conditioned on the dialogue context $\mathbf{c}$ in $p_\phi(\mathbf{z_k}|\mathbf{c})$. During the training phase, we use a variational posterior $q_\varphi(\cdot)$ to maximize the Evidence Lower Bound (ELBO) as follows:

$$\begin{aligned}
&\mathcal{L}_{\text{CoLV}} = \\
&- KL(q_\varphi(\mathbf{z_k}|\mathbf{c}, \mathbf{k})||p_\phi(\mathbf{z_k}|\mathbf{c})) \\
&- KL(q_\varphi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c}, \mathbf{k}, \mathbf{r})||p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c})) \\
&+ \mathbb{E}_{\mathbf{z_k} \sim q_\varphi}[\log p_\theta(\mathbf{k}|\mathbf{z_k}, \mathbf{c})] \\
&+ \mathbb{E}_{\mathbf{z_r} \sim q_\varphi}[\log p_\theta(\mathbf{r}|\mathbf{z_r}, \mathbf{k}, \mathbf{c})],
\end{aligned}$$

where θ, ϕ and φ are the parameters of the generation, prior and posterior networks. The graphical framework for our proposed CoLV model is shown in Figure 1.

During training phase, our proposed CoLV model consists of two independent latent variables: $\mathbf{z_k}$ and $\mathbf{z_r}$, which represent the latent variables of knowledge and response respectively. Meanwhile, the variational lower bound includes the reconstruction terms and KL divergence terms (Kullback and Leibler, 1951) based on these two latent variables, which will be optimized in a unified process.

In the generative process, latent variables obtained via prior networks and selected knowledge are fed to the decoder phase, which corresponds to red solid arrows in Figure 1. The generative process is as follows:

Step 1: Sample knowledge latent variable: $\mathbf{z_k} \sim p_\phi(\mathbf{z_k}|\mathbf{c})$.

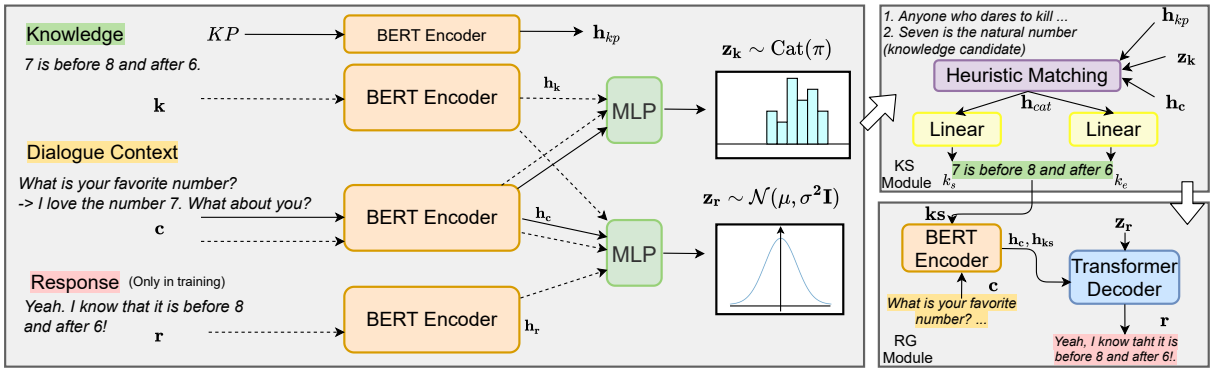


Figure 2: The illustration of our proposed CoLV framework. KP: knowledge pool, c: dialogue context, k: selected knowledge, r: response. The dotted line denotes the training procedure, while the solid line denotes the inference process. "KS" and "RG" denote knowledge selection and response generation, respectively.

Step 2: Sample response latent variable: $\mathbf{z_r} \sim p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c})$.

Step 3: Select a knowledge: $\mathbf{k} \sim p_\theta(\mathbf{k}|\mathbf{z_k}, \mathbf{c})$.

Step 4: Generate a response: $\mathbf{r} \sim p_\theta(\mathbf{r}|\mathbf{z_r}, \mathbf{k}, \mathbf{c})$.

3.3 Input Representation

We employ a pre-trained BERT_{base} (Devlin et al., 2019) model as encoder to capture the semantic representation of both dialogue context $\mathbf{c}$ and knowledge candidate sentences $\text{KP} = \{k_1, ..., k_{|k|}\}$. Take dialogue context $\mathbf{c} = \{c_1, ..., c_{|c|}\}$ as an example. The initial representation of $\mathbf{c}$ is the sum of word, position and turn-level embeddings:

$$\mathbf{e_c} = \text{WE}(\mathbf{c}) + \text{PE}(\mathbf{c}) + \text{TE}(\mathbf{c}),$$
$$\mathbf{H_c} = \text{BERT}_{base}(\mathbf{e_c}), \mathbf{h_c} = \text{Avgpool}(\mathbf{H_c}),$$

where $\mathbf{e_c}$ and $\mathbf{h_c}$ are the initial representation and hidden representation after BERT of dialogue context. $\text{WE}(\cdot)$, $\text{PE}(\cdot)$ and $\text{TE}(\cdot)$ refer to the word-level, position-level and turn-level embeddings respectively. $\text{Avepool}(\cdot)$ is the average pooling operation (Cer et al., 2018). Similarly, we also employ BERT_{base} to encode the knowledge candidate sentences. The initial representation $\mathbf{e}_{kp}$ and hidden representation $\mathbf{h}_{kp}$ after BERT model and average pooling operation of knowledge candidate sentences are formulated as follows:

$$\mathbf{e}_{kp} = \text{WE}(kp) + \text{PE}(kp) + \text{TE}(kp),$$
$$\mathbf{H}_{kp} = \text{BERT}_{base}(\mathbf{e}_{kp}), \mathbf{h}_{kp} = \text{Avgpool}(\mathbf{H_c}).$$

Similarly, we can also get the posterior information of ground truth response representation $\mathbf{h_r}$ and knowledge representation $\mathbf{h_k}$ for training phase.

3.4 Collaborative Latent Variables

We will use two separate but content-dependent latent variables $\mathbf{z_r}$ and $\mathbf{z_k}$ to represent dialogue response and knowledge respectively. In the following section, we will discuss the prior network and posterior network separately.

3.4.1 Prior Network

We use two different conditional prior networks $p_\phi(\mathbf{z_k}|\mathbf{c})$ and $p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c})$ to model these two tasks. As we know, knowledge selection and response generation belongs to discriminative and generative task respectively. For better collaboratively modelling the relationship between knowledge selection and response generation, we utilize two different distribution models: the standard Categorical distribution $\text{Cat}(\pi)$ for $p_\phi(\mathbf{z_k}|\mathbf{c})$ and Gaussian distribution $\mathcal{N}(\mu, \sigma^2\mathbf{I})$ for $p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c})$. Therefore, the $\mathbf{z_k}$ and $\mathbf{z_r}$ are sampled from:

$$p_\phi(\mathbf{z_k}|\mathbf{c}) = \text{Cat}_\phi(\mathbf{z_k}|\pi),$$
$$p_\phi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c}) = \mathcal{N}_\phi(\mathbf{z_r}|\boldsymbol{\mu}^r, \boldsymbol{\sigma}^r\mathbf{I}),$$

where the parameters σ and μ are estimated by:

$$\boldsymbol{\mu}^r = \text{MLP}^r_\phi(\mathbf{h_c}), \boldsymbol{\sigma}^r = \text{softplus}(\text{MLP}^r_\phi(\mathbf{h_c})),$$

where $\text{MLP}(\cdot)$ denotes the multiple layer perception, and $\text{softplus}(\cdot)$ function is a smooth approximation to ReLU and can be used to ensure positiveness.

3.4.2 Posterior Network

During training phase, we utilize the posterior information to help enforce training. Similar to the prior network, we also use two different conditional posterior networks $q_\varphi(\mathbf{z_k}|\mathbf{c}, \mathbf{k})$ and $q_\varphi(\mathbf{z_r}|\mathbf{z_k}, \mathbf{c}, \mathbf{k}, \mathbf{r})$ to approximate the true posterior distributions of latent variables for both knowledge $\mathbf{k}$ and response $\mathbf{r}$. Therefore, the $\mathbf{z_k}$ and $\mathbf{z_r}$ in the posterior networks

are sampled from:

$$q_\varphi(\mathbf{z_k}|\mathbf{c},\mathbf{k}) = \mathrm{Cat}_\varphi(\mathbf{z_k}|\pi),$$
$$q_\varphi(\mathbf{z_r}|\mathbf{z_k},\mathbf{c},\mathbf{k},\mathbf{r}) = \mathcal{N}_\varphi(\mathbf{z_r}|\boldsymbol{\mu}^r,\boldsymbol{\sigma}^r\mathbf{I}),$$

where the parameters σ and μ in the posterior networks are estimated by:

$$\boldsymbol{\mu}^r = \mathrm{MLP}^r_\varphi([\mathbf{h_c},\mathbf{h_r}]),$$
$$\boldsymbol{\sigma}^r = \mathrm{softplus}(\mathrm{MLP}^r_\varphi([\mathbf{h_c},\mathbf{h_r}])).$$

In the training phase, we adopt the reparameterization trick (Kingma and Welling, 2014) to train our model with back-propagation since the stochastic sampling process of both knowledge selection and response generation is non-differential. Besides, we further employ gumbel-softmax (Maddison et al., 2017) for knowledge selection training procedure, since the latent variables $\mathbf{z_k}$ is discrete.

3.5 Heuristic-based Knowledge Selection

While the efficiency of heuristic matching algorithm (Mou et al., 2016) has been demonstrated in many other tasks, such as question and answering. Following Lee et al. (2020), we also employ a heuristic-based knowledge selection module. Besides, different from previous work, which select out relevant knowledge instance from multiple knowledge sentences, our proposed heuristic-based knowledge selection module regards all candidate knowledge sentences as an integrated paragraph. Then, this module will predict the start and the end word position of an knowledge span. The knowledge span is regarded as the selected knowledge and will be incorporated by the following response generation process.

Specifically, given the representation of dialogue context $\mathbf{h_c}$ and latent variables $\mathbf{z_k}$, the heuristic-based knowledge selection layer will consider to concatenate the adding and multiplying operation as an new integrated representation $\mathbf{h}_{cat}$, which is formulated as follows:

$$\mathbf{h}_{cat} = [\mathbf{h_c},\mathbf{z_k},|\mathbf{h_c}-\mathbf{z_k}|,\mathbf{h_c}\odot\mathbf{z_k}],$$

where the new representation $\mathbf{h}_{cat}$ will be used to predict the knowledge span in the following steps. Therefore, we will feed the integrated representation $\mathbf{h}_{cat}$ into two separate linear layers (as shown in Figure 2) to predict the start and end position of knowledge span $\mathbf{ks}$. the knowledge span $\mathbf{ks}$ will be extracted and sent into the generation phase.

3.6 Response Generation

In the decoding layer for response generation, we apply a stacked Transformer decoder module equipped with a copying mechanism (See et al., 2017) to generate response. The copy mechanism is used to copy specific knowledge from the selected knowledge span. We feed the dialogue context representation $\mathbf{h_c}$, the selected knowledge span representation $\mathbf{h_{ks}}$ and latent variable $\mathbf{z_r}$ into the decoder phase. Specifically, the probability of generating token y_t at t-th step is modeled as:

$$P(y_t) = \lambda_1 P_{vocab}(y_t|\mathbf{h_c},\mathbf{z_r}) + \lambda_2 P_{cp}(y_t|\mathbf{h_{ks}}).$$

where $P_{cp}(y_t|\mathbf{h_{ks}})$ derives the copying probability from the selected knowledge span $\mathbf{ks}$. The copy mechanism is defined as follows:

$$P_{cp}(y_t|\mathbf{h_{ks}}) = \sum_{i:t_i=y_t} \alpha_{t,i}.$$

$P_{vocab}(y_t|\mathbf{h_c},\mathbf{z_r})$ is the output probability from a stack of Transformer decoder layers (Vaswani et al., 2017). λ_1, and λ_2 are the coordination probability parameters.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct our experiments on two public knowledge-grounded dialogue datasets, *Wizard of Wikipedia* (Dinan et al., 2019) (WoW) and Holl-E (Moghe et al., 2018). In these two benchmarks, both of them contain multiple sessions of dialogues with corresponding knowledge candidate pool. For each dialogue utterance, there is a ground truth knowledge sentence. The statistical details on these two datasets are shown in Table 3.

Baseline Models. We compare our CoLV model with several state-of-the-art models, including:

- **S2SA**: The bidirectional LSTM-based encoder-decoder framework with attention mechanism. This baseline model only consider the dialogue context and do not utilize knowledge information. (Sutskever et al., 2014).
- **Transformer**: an encoder-decoder architecture relying solely on multi-head self-attention mechanisms (Vaswani et al., 2017). It does not consider the knowledge information either.
- **MemNet**: The E2E Transformer with memory mechanism (Dinan et al., 2019), which uses a

Model	WoW Test Seen						WoW Test Unseen					
	ACC	PPL	BLEU-4	RG-1	RG-2	Dist-2	ACC	PPL	BLEU-4	RG-1	RG-2	Dist-2
S2SA	-	93.85	0.46	12.53	0.69	4.81	-	120.81	0.34	9.30	0.76	11.53
Transformer	-	72.42	0.39	14.35	1.36	19.68	-	91.41	0.39	12.87	0.66	12.15
MemNet	21.60	63.52	0.41	16.9	0.64	24.16	13.82	96.47	0.32	14.46	0.82	16.27
PostKS	3.66	79.19	0.57	13.04	1.17	16.70	3.29	152.7	0.36	13.15	1.08	13.38
SKLS	26.83	52.09	1.35	16.87	6.84	23.13	16.59	81.44	1.05	16.16	4.21	16.42
DukeNet	25.96	48.33	2.46	19.02	6.54	25.67	17.49	69.38	1.68	19.36	5.23	17.03
PIPM	27.75	42.71	2.26	19.34	7.36	26.41	**19.43**	65.71	1.56	17.60	5.49	17.74
CoLV	**30.12***	**39.56***	**2.85***	**20.62**	**7.89**	**29.74***	18.91	**54.30***	**2.12***	**19.68***	**6.31**	**20.13***

Table 2: Automatic evaluation results on *WoW Test Seen* and *WoW Test Unseen* (%). The metrics Accuracy, Perplexity, ROUGE-1, ROUGE-2 and Distinct-2 are abbreviated as ACC, PPL, RG-1, RG-2 and Dist-2, respectively. The best results are highlighted with **bold**. "*" denotes that the result is statistically significant with $p < 0.01$.

	WoW	Holl-E
Training size	18,430	7,228
Validation size	1,948	930
Test size	965 (S)/968 (U)	913
Avg. Num of kg	67	53

Table 3: Statistics of two experimental datasets, *Wizard of Wikipedia* (WoW) and Holl-E. "**S**" and "**U**" denotes the test seen and test unseen in WoW dataset respectively.

Model	Holl-E					
	ACC	PPL	BLEU-4	RG-1	RG-2	Dist-2
S2SA	-	150.26	4.84	4.28	2.01	10.38
Transformer	-	120.31	5.09	6.72	2.96	14.29
MemNet	22.75	138.38	5.49	20.19	10.34	23.63
PostKS	1.56	187.20	5.85	15.23	6.08	19.74
SKLS	29.25	48.97	17.81	29.82	23.19	27.43
DukeNet	30.38	42.72	19.15	**32.64**	19.55	28.53
PIPM	30.67	39.22	18.27	30.81	23.96	27.20
CoLV	**32.65***	**34.84***	**20.33***	31.97	**25.84***	**29.86***

Table 4: Automatic evaluation results on *Holl-E* (%). The best results are highlighted with **bold**. "*" denotes that the result is statistically significant with $p < 0.01$.

Transformer memory network for knowledge selection and a Transformer decoder for utterance prediction.

- **PostKS**: A LSTM-based model with the posterior knowledge selection mechanism (Lian et al., 2019), which uses the posterior knowledge distribution as a pseudo-label for knowledge selection.
- **SLKS**: A sequential latent knowledge selection model (Kim et al., 2020), which keeps track of prior and posterior distribution over knowledge in a sequential process.
- **DukeNet**: A dual knowledge interaction network (Meng et al., 2020), modeling the knowledge shift and tracking processes with a dual learning paradigm.
- **PIPM**: SLKS model with posterior information prediction module and knowledge distillation training strategy (Chen et al., 2020). It aims to bridge the gap between prior and posterior distributions.

Evaluation Metrics. We report accuracy (Acc) to evaluate the knowledge selection[1]. Besides, we use the traditional indicators, i.e., perplexity (PPL), BLEU-4 (Papineni et al., 2002), ROUGE-1, ROUGE-2 (Lin, 2004) and Distinct-2 (Li et al., 2016) to evaluate the quality of response generation. We also conduct human evaluation for our model. We randomly sampled 300 generated response and then we invite six annotators to select out their preferred response (win), or vote a tie, considering the following aspects: diversity, coherence and knowledge engagement. Each comparison is conducted between two responses generated by our CoLV and a baseline models respectively.

Implementation Details. We implement our proposed model with pytorch (Paszke et al., 2019). For fair comparison, we keep the same defualt settings during dataset pre-processing and the model parameter settings as the same as in (Kim et al., 2020). We employ a pre-trained BERT_{base} model to encoder dialogue context and knowledge sentences. The initial word embedding size is set to 300, and we keep the sentence length of dialogue context and knowledge to 64 and 512 respectively. The hidden size is 768 and vocabulary size is set to 30,522. The batch size is set to 64. Models are trained with

[1]Note that lower perplexity (PPL) indicates better performance. For the evaluation on knowledge selection, only knowledge span with both correct start and end position will be counted in the accuracy. Partially correct sample will not be counted in the accuracy calculation, but we will analysis the KS performance in Section 4.5

Model	WoW Test Seen						WoW Test Unseen					
	ACC	PPL	BLEU-4	RG-1	RG-2	Dist-2	ACC	PPL	BLEU-4	RG-1	RG-2	Dist-2
Full model	**30.12**	**39.56**	**2.85**	**20.62**	**7.89**	**29.74**	**18.91**	**54.30**	**2.12**	**19.68**	**6.31**	**20.13**
- *knowledge latent*	23.65	46.37	2.30	18.41	6.93	26.48	14.70	63.84	2.14	17.36	5.93	18.29
- *response latent*	26.48	58.72	1.75	15.81	4.22	17.93	16.56	72.57	1.56	12.60	3.51	14.96
- *heuristic matching*	24.93	48.16	2.06	17.97	7.14	23.06	13.96	68.29	2.04	15.83	5.92	17.25
- *all*	21.16	74.62	1.26	14.39	3.85	15.39	12.74	81.82	1.43	13.45	3.20	12.77

Table 5: Ablation study results on *WoW Test Seen* and *WoW Test Unseen* (%). The metrics Accuracy, Perplexity, ROUGE-1, ROUGE-2 and Distinct-2 ar abbreviated as ACC, PPL, RG-1, RG-2 and Dist-2, respectively.

Dataset	Model	COLA vs.			kappa
		Win	Loss	Tie	
(a)	S2SA	67%	13%	20%	0.618
	Transformer	56%	21%	23%	0.572
	MemNet	58%	19%	23%	0.538
	PostKS	64%	16%	20%	0.596
	SKLS	47%	28%	25%	0.424
	DukeNet	42%	33%	25%	0.474
	PIPM	49%	24%	27%	0.445
(b)	S2SA	57%	16%	27%	0.538
	Transformer	56%	16%	34%	0.407
	MemNet	52%	19%	29%	0.481
	PostKS	48%	11%	41%	0.523
	SKLS	50%	22%	28%	0.509
	DukeNet	51%	29%	20%	0.426
	PIPM	46%	27%	29%	0.473
(c)	S2SA	71%	8%	21%	0.561
	Transformer	65%	11%	24%	0.539
	MemNet	59%	18%	23%	0.472
	PostKS	54%	16%	30%	0.494
	SKLS	48%	23%	29%	0.586
	DukeNet	52%	27%	21%	0.463
	PIPM	47%	25%	28%	0.535

Table 6: Human evaluations on *Holl-E* and *WoW* datasets. (a): WoW test seen. (b) WoW test unseen. (c) Holl-E.

30 epoch to get the best performance. For training details, we use Adam (Kingma and Ba, 2015) for gradient optimization in our experiments, and the correspondign parameters β_1 and β_2 are set to 0.9 and 0.998. The learning rate is set to 0.001. We use gradient clipping with a maximum gradient norm of 0.4. We run all models on the Tesla P40 GPU and select the best models based on performance on the validation set.

4.2 Experimental Results

Automatic Evaluation Results. The quantitative evaluation results on WoW and Holl-E datasets are shown in Table 2 and Table 4 respectively. Generally, CoLV outperforms baselines on most metrics in these two datasets. In terms of the knowledge selection accuracy, CoLV outperforms three strong baseline SKLS, DukeNet and PIPM on WoW Test Seen dataset by 11.0%, 16.2% and 8.7%, which is significant. Even though the accuracy of CoLV on WoW Test Unseen is a little lower than PIPM, it still outperforms other baselines. The reason why CoLV can improve knowledge selection performance is that CoLV takes two collaborative latent variables simultaneously, which resolving the gap between knowledge and response. Besides, in terms of the generation performance, CoLV also has a significant improvement over baseline models. It helps verify the consistency of improvement on both knowledge selection and response generation.

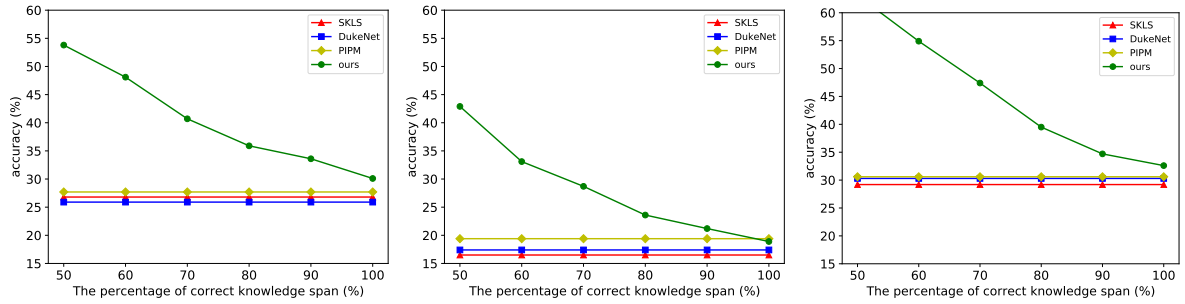
Table 7: Ablation study results on *Holl-E* (%). The knowledge, response, heuristic are abbreviated as kg, res, hr, respectively.

Model	Holl-E					
	ACC	PPL	BLEU-4	RG-1	RG-2	Dist-2
Full model	**32.65**	**34.84**	**20.33**	**31.97**	**25.84**	**29.86**
- *kg latent*	23.94	39.25	18.27	27.86	18.92	26.84
- *res latent*	27.48	56.35	15.36	24.01	13.48	22.05
- *hr matching*	24.12	45.63	16.28	25.86	16.76	24.27
- *all*	14.82	78.29	11.05	19.04	11.67	21.58

Human Evaluation Results. The human-based evaluation results are shown in Table 6. For each case, given a post-knowledge pair, two generated responses are provided, one is from our model and the other is from the compared model. Not surprisingly, CoLV consistently outperforms all the compared models. Meanwhile, we notice that CoLV exhibits significant improvements comparing with vanilla S2SA and Transformer. Besides, CoLV substantially reaches better performances than strong baselines, e.g., SKLS and PIPM. We analyze the bad cases and find that some baselines still suffer from the general or knowledge-irrelevant responses. Augmented with the collaborative latent variables, CoLV introduces a competitive boost in response quality, which is in line with the automatic evaluation, confirming the superior performance of our proposed model. We also employ Fleiss' kappa scores (Fleiss, 1971) to measure the reliability between different annotators, and results show that annotators reach a moderate agreement.

4.3 Ablation Study

To examine the effectiveness of proposed CoLV model we conduct model ablations by removing particular modules from CoLV, including knowl-



(a) WoW Test Seen. (b) Wow Test Unseen. (c) Holl-E.

Figure 3: Analysis on the heuristic-based knowledge selection module. Horizontal axis denotes the percentage of correct knowledge span, ranging from 50% to 100%. Vertical axis denotes the accuracy of knowledge selection. Note that in our experiments, we treat all knowledge candidates as an integrated paragraph, rather than individual sentence.

edge latent, response latent, heuristic matching and all modules. The ablation results on *WoW* and *Holl-E* are shown in Table 5 and Table 7 . We observe that without either knowledge latent and heuristic matching, the performance of knowledge selection drops largely with respect to accuracy metric. The result verifies the effectiveness of integrating these two modules into knowledge selection process. Besides, the values of generative metrics, e.g., PPL, BLEU-4, ROUGE-1/2 and Dist-2, also drop significantly if we remove the response latent variables. It affirms that the collaborative latent variables are helpful to refine the coherence, knowledge engagement and diversity of generated responses. While we remove all these three modules, we can witness a similar performance of our model with the baseline model MemNet, a vanilla Transformer model with knowledge memory network.

4.4 Case Study

To facilitate a better understanding of baselines and our model, we present some examples in Table 8. To better evaluate the performance of response generation, we choose a case from WoW Test Seen that both three baseline models SKLS, DukeNet, PIPM and our model select the same knowledge (marked as yellow in Table 8 from knowledge pool. We observe that even though both baselines and our model can select out the true knowledge sentence, our model still achieve better performances in response generation. For example, SKLS generates a counterfactual response that is not consistent with original knowledge. In orignial knowledge, "*Ireland is the third-large island in Europe*". However, SKLS generates "*Ireland is the largest*". Besides, to show the effectiveness of our model in generating diverse responses, we present several different responses that all generated by our model.

Dialogue Context: Have you been to Europe? → I have! I have been to British. → Great! I've only ever been to Canada.	
Knowledge Candidates: 1. Ireland is the third-largest island in Europe. Thick woodlands covered the island until the Middle Ages . 2. Ireland the second largest island in the British Isles, after Great Britain. 3....	
GT	That's pretty cool, but I'd still love to visit more of Europe. It's cool to explore woodland in Ireland.
SKLS	Yes, Ireland is the largest island in the Europe. I would like to go there again.
DukeNet	Canada is also a beautiful country, like the Ireland.
PIPM	Pretty cool! Canada is fantastic! I wish I lived there instead of the Europe.
Ours-1	That's great! Come to Ireland, the third largest island in Europe.
Ours-2	Pretty great! If you know that Ireland was covered by woodlands before many years.
Ours-3	That's pretty cool. The woodlands in Ireland deserves to be visited.
Ours-4	Yeah, Ireland has a third-largest woodlands. Have you ever heard it?

Table 8: Case study on the WoW Test Seen dataset. For fair comparison of the generation performance, we choose the case that all model select the same knowledge (marked as yellow). We further present several different responses that all generated by our model. "GT" denotes the ground truth.

As shown in Table 8, our model is able to engage different parts in the knowledge sentence and then generate diverse and coherent response. The reason why our model has the ability to generate different and semantic coherent response is that the collaborative latent variables in CoLV consider these two distributions collaboratively.

4.5 Analysis of Heuristic-based KS

We conduct a further experiments to analyze if our fine-grained knowledge selection perform better than traditional sentence-level match methods. The results of knowledge selection on three datasets are shown in Figure 3. Different from previous sentence-level knowledge selection. Our method firstly treats all knowledge sentences as an integrated paragraph and select knowledge span from

Dialogue Context: Thierry Henry is one of my all time favorite players. What about you?	
Knowledge Candidates: 1. Thierry Daniel Henry is a retired French professional footballer . 2. He played as a forward and is the second assistant manager of the Belgium national team . 3. Henry made his professional debut with Monaco in 1994. 4. A year later he signed for Premier League club Arsenal .	
GT Pair	Kg: a retired French professional footballer. Response: He was good. he is a retired French professional footballer.
Pair-1	Kg: Premier League club Arsenal . Response: I know him. He has served in the premier league club Arsenal.
Pair-2	Kg: a retired French professional footballer . Response: Henry is a retired French footballer, he was so famous.
Pair-3	Kg: the second assistant manager of the Belgium national team . Response: Yes, I love him too. He was also the second assistant manager of the Belgium team.

Table 9: Qualitative analysis of collaborative latent variables. Knowledge-response pairs generated by our model. "GT pair" denotes the ground truth knowledge-response pair in the dataset. "Pair-1", "Pair-2" and "Pair-3" are generated from our model.

this paragraph. Only the start and end position of knowledge span are totally matched (100% matching) with original knowledge sentence, it will be counted by accuracy metric. However, we observe that in our test set, many bad cases also select out partial correct knowledge content. Therefore, we conduct a further statistics on accuracy of different percentage of correct knowledge span, as shown in Figure 3. Take the WoW Test Seen dataset as example ((a) in Figure 3), our model can reach around 0.35 accuracy on the 80% and 90% percentage of correct knowledge span, which is significantly higher than baseline models. Considering that conversational model usually do not engage all the knowledge context into response generation, we claim that the 80% and 90% percentage of correct knowledge span are acceptable in real application scenarios. Therefore, CoLV is more practical and flexible than the existing methods.

4.6 Effects of Collaborative Latent Variables

We conduct a further qualitative analysis on the collaborative latent variables. Firstly, we utilize the knowledge variable in multiple times to get different knowledge. Then, for each selected knowledge, we employ the decoder phase to generate corresponding responses. As shown in Table 9, our CoLV is able to select different knowledge context and then generate corresponding responses. We can notice that all responses in Pair-1 to Pair-3 are coherent and fluency to the dialogue context. Besides, knowledge information is appropriately engaged into the response. Therefore, the two latent variables in our CoLV model are effective to help select diverse knowledge and then generate coherent response.

5 Conclusion

In this paper, we propose a novel collaborative latent variable (CoLV) model to simultaneously learning to select knowledge and generate responses in knowledge-grounded dialogue generation. CoLV model helps improve the diversity not only on knowledge selection but also help generate diverse response while given a specific knowledge. Besides, the CoLV model uses two collaborative latent variables for coupling the knowledge and dialogue. Extensive experiments on two benchmark datasets show that CoLV achieves satisfied performance, indicating that CoLV can select more diverse knowledge and further generate more coherent and diverse responses than baseline models.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their constructive comments and suggestions. We would also like to thank JD.com for their support in computing resource.

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *ArXiv preprint*, abs/1803.11175.

Xiuyi Chen, Feilong Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021. Unsupervised knowledge selection for dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1230–1244, Online. Association for Computational Linguistics.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

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, San Diego, California. Association for Computational Linguistics.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087. ijcai.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52, Online. Association for Computational Linguistics.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1151–1160. ACM.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers), pages 130–136, Berlin, Germany. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 3826–3835, Florence, Italy. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Lei Shen, Yang Feng, and Haolan Zhan. 2019. Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5497–5502, Florence, Italy. Association for Computational Linguistics.

Lei Shen, Fandong Meng, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. GTM: A generative triple-wise model for conversational question generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3495–3506, Online. Association for Computational Linguistics.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097, Online. Association for Computational Linguistics.

Yajing Sun, Yue Hu, Luxi Xing, Jing Yu, and Yuqiang Xie. 2020. History-adaption knowledge incorporation mechanism for multi-turn dialogue system. In *AAAI*. AAAI.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Zhiliang Tian, Wei Bi, Dongkyu Lee, Lanqing Xue, Yiping Song, Xiaojiang Liu, and Nevin L. Zhang. 2020. Response-anticipated memory for on-demand knowledge integration in response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 650–659, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. *Proceedings of the Thirty-Third AAAI Conference onArtificial Intelligen*, pages 9169–9176.

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 53–65, Online. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with

explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9338–9345. AAAI.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630, Online. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.