Knowledge-aware Attentive Wasserstein Adversarial Dialogue Response Generation

YINGYING ZHANG, National Lab of Pattern Recognition, Institute of Automation, CAS, China, and School of Artificial Intelligence, University of Chinese Academy of Sciences, China
QUAN FANG and SHENGSHENG QIAN, National Lab of Pattern Recognition, Institute of Automation, CAS, China, and University of Chinese Academy of Sciences, China
CHANGSHENG XU, National Lab of Pattern Recognition, Institute of Automation, CAS, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, China, and Peng Cheng Laboratory, China

Natural language generation has become a fundamental task in dialogue systems. RNN-based natural response generation methods encode the dialogue context and decode it into a response. However, they tend to generate dull and simple responses. In this article, we propose a novel framework, called **KAWA-DRG** (Knowledge-aware Attentive Wasserstein Adversarial Dialogue Response Generation) to model conversation-specific external knowledge and the importance variances of dialogue context in a unified adversarial encoder-decoder learning framework. In KAWA-DRG, a co-attention mechanism attends to important parts within and among context utterances with word-utterance-level attention. Prior knowledge is integrated into the conditional Wasserstein auto-encoder for learning the latent variable space. The posterior and prior distribution of latent variables are generated and trained through adversarial learning. We evaluate our model on Switchboard, DailyDialog, In-Car Assistant, and Ubuntu Dialogue Corpus. Experimental results show that KAWA-DRG outperforms the existing methods.

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics**;

Additional Key Words and Phrases: Dialogue system, co-attention, adversarial learning, external knowledge

ACM Reference format:
Yingying Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2020. Knowledge-aware Attentive Wasserstein Adversarial Dialogue Response Generation. *ACM Trans. Intell. Syst. Technol.* 11, 4, Article 37 (May 2020), 20 pages.
https://doi.org/10.1145/3384675

This work was supported in part by the National Key Research and Development Program of China (No. 2017YFB1002804), the National Natural Science Foundation of China under Grant Nos. 61720106006, 61572503, 61802405, 61872424, 61702509, 61832002, 61936005, and U1705262, the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC039, and the K.C. Wong Education Foundation.
Authors' addresses: Y. Zhang, National Lab of Pattern Recognition, Institute of Automation, CAS, China, and School of Artificial Intelligence, University of Chinese Academy of Sciences, 95, ZhongGuanChun East Road, Beijing, China, 100190; email: zhangyingying2017@ia.ac.cn; Q. Fang and S. Qian, National Lab of Pattern Recognition, Institute of Automation, CAS, China, and University of Chinese Academy of Sciences, 95, ZhongGuanChun East Road, Beijing, China, 100190; emails: {qfang, shengsheng.qian}@nlpr.ia.ac.cn; C. Xu, National Lab of Pattern Recognition, Institute of Automation, CAS, China, School of Artificial Intelligence, University of Chinese Academy of Sciences, China, and Peng Cheng Laboratory, 95, ZhongGuanChun East Road, Beijing, China, 100190; email: csxu@nlpr.ia.ac.cn.

2157-6904/2020/05-ART37 $15.00
https://doi.org/10.1145/3384675

1 INTRODUCTION

Dialogue response generation systems aim to produce reasonable utterances given input of conversational history. There has been a growing research interest in applying encoder-decoder models to generate next utterance [31, 48], where dialogue context is considered as the source sequence and the response is the target. They mainly use the recurrent neural network [22] to model utterance in a dialogue, which uses continuous representation to estimate a probability function and generates final response word-by-word. However, recent research pointed out that these models tend to generate generic safe responses, rather than meaningful, context-specific, and diverse replies [15, 32].

It is a matter of interest to researchers how to generate diverse responses in the dialogue system. Basically, there are two major research topics: (1) One popular approach for response generation is to enrich the model with richer conditionals [16, 43, 48]. For example, Li et al. [16] proposed a persona-based neural conversation model that endows data-driven systems with the coherent "persona" needed to model human-like behavior. However, as not all datasets contain the domain-specific information, these approaches are hard to be generalized to other datasets. (2) Some approaches attempted to improve the architecture of the encoder-decoder model itself by optimizing the task metric directly [9, 41]. For instance, in Reference [41], instead of predicting the probability of the next word, it learned to produce non-probabilistic scores for ranking sequences and used a training procedure that defined a loss function in terms of beam search errors.

In addition, recent studies adopted conditional variational auto-encoders (CVAE) [34, 46, 48] to introduce a recognition network to output an approximate posterior distribution over a latent variable, which can capture the discourse-level diversity to obtain diverse responses. Furthermore, unlike CVAE conversation models trained by minimizing the Kullback-Leibler divergence, DialogWAE [10] was proposed to generate multimodal responses by employing a conditional Wasserstein Auto-Encoder, which models the distribution of data by training a GAN within the latent variable space. Although the above approaches have made great progress on generating proper responses, there still exist certain critical issues when it is applied to open-domain conversational generation. (1) *Lack conversation-specific common sense.* Previous methods fail to utilize grounding in the real world and do not have access to any external knowledge, which makes it hard to produce substantive and informative responses. while the daily conversations are generally based on an individual's knowledge [49]. As shown in Table 1, the dialogue is talking about a girl's height, so at least one participant knows about the information of the girl about whom they are talking. (2) *Not all parts of context are equally important to response generation.* Words are differentially informative and important, and so are the responses [44]. In the above example, since utterances containing the word "tall" involve main semantics of the context, they are more important than the first utterance. However, the state-of-the-art models such as DialogWAE encode the context into a single vector without considering different importance, therefore missing rich information in the context, which leads to the responses generated by DialogWAE being really random, varying from "A lot of fun," "Thanks for your help," to "Get a new job."

In this article, we aim to model the common-sense knowledge and the important parts of contexts in a unified framework and thus generate reasonable and diverse responses in dialogue systems. Inspired by the success of the attention mechanism [45] and adversarial learning [17] in NLP tasks, we propose a novel framework, called KAWA-DRG (Knowledge-aware Attentive Wasserstein Adversarial Dialogue Response Generation), in which we introduce a co-attention mechanism to dynamically highlight important parts of word sequences and the utterance sequence as well as exploit the external knowledge into Wasserstein adversarial auto-encoder learning when generating a response. Specifically, for modeling context importance, we propose a novel

Table 1. Examples of Generated Response A, B Indicate Different Actors in the Dialogue

Context
A: Describe her to me.
B: She's not too tall
A: Well, how tall is she?
Expected: She's about 5 feet even.
Responses Generated by DialogWAE
Sample 1 A lot of fun.
Sample 2 Thank you for help.
Sample 3 That's too bad. I thought I'm going to be able to get a new job.

co-attention module between dialogue context and the reference response in the learning phase, to capture the most relative information of them, and add this additive information in the process of generating latent variable, thus, force it to better learn the posterior distribution of the response in the latent space. Furthermore, we incorporate grounding knowledge to prior distribution of response in the latent space. KAWA-DRG introduces latent variables that encode multimodal dialogue semantic structures by transforming context-dependent and knowledge-aware random noise with GAN-based model and minimizes the Wasserstein distance [2] between the prior and the approximate posterior distributions. To provide supervision about knowledge, we recover it from the latent variable and treat it as "fake" while the original as "real," then learn it with adversarial networks. Finally, KAWA-DRG generates diverse responses by drawing samples from the learned distribution and reconstructing them via a decoder neural network.

We evaluate our method on four real benchmark datasets, including Switchboard, DailyDialog, In-Car Assistant, and Ubuntu Dialogue Corpus. Experimental results on the above datasets show that our proposed approach outperforms the existing approaches for natural response generation.

In summary, the contributions of this work are threefold:

- We propose a novel KAWA-DRG model for dialogue response generation. The proposed KAWA-DRG can effectively model conversation-specific external knowledge and the importance variances of dialogue context in a unified adversarial encoder-decoder learning framework.
- A novel co-attention module is proposed to calculate the attention matrix and fuse the dialogue context and response to train a better auto-encoder. Moreover, a novel knowledge-aware condition Wasserstein auto-encoder is proposed to condition responses based on words in the context as well as external knowledge related to the conversation history, which can model explicit semantics and implicit common sense in a unified network architecture.
- We evaluate the proposed model over four large, real-world dialogue datasets. Experimental results on benchmark datasets demonstrate that our KAWA-DRG achieves much better performance than the existing approaches.

2 RELATED WORK

In this section, we review recent advances in sequence generative methods, including attention mechanisms in sequence generation and knowledge-guided methods in natural language processing and language understanding, especially in the dialogue systems.

2.1 Generative Model

In recent years, the recurrent neural network has been exploited to address natural language generation problems, e.g., abstractive document summarization [36], visual question answering [1], and image caption [37]. As mentioned above, most recurrent neural network–based generative models fail to generate meaningful and diverse responses. There are several attempts to alleviate this problem. One research line is to change the architecture of the encoder-decoder framework. Li et al. [15] proposed a diversity-promoting objective function, which aims to maximize mutual information between input and output, to encourage diverse responses.

Variation auto-encoders (VAEs) [14] and Generative Adversarial Nets (GANs) [8] are widely used in image generation. Nowadays, these methods are adopted to natural language generation. Park et al. [25] combined VAE with hierarchical RNNs, using a hierarchical structure of the latent variables and exploiting an utterance drop regularization. GANs and reinforcement learning [17] have also been implemented for natural dialogue generation, with a generative model producing a response and a discriminator distinguishing between the reference and machine-generated ones. The output from the discriminator then performs as the reward for the generative model.

2.2 Attention Mechanism in Sequence Generation

Attention mechanism is first proposed for machine translation [3] and is quickly applied to other natural language processing tasks, including response generation in a dialogue system [44], mapping query suggestion [35], identifying the quality of answers in community-based question answering [42].

Xing et al. [44] noticed that there are word-level and sentence-level of attention when modeling a dialogue and proposed a hierarchical recurrent attention mechanism that attends to important parts within and among utterances with word-level attention and utterance level attention during the process of generating every word. Self-attention can mine different aspects of sentences, thus help to learn a better sentence embedding. Lin et al. [19] performed a self-attention mechanism on the top of their sentence embedding model, which extracted different aspects of the sentence into multiple vector representations. Huang et al. [12] applied self-attention in sequential recommendation. Parikh et al. [24] applied an intra-sentence-level attention that encoded compositional relationships between words within each sentence to augment input representations. However, methods with co-attention mechanism are widely used in question answering systems to find the internal relationship between question and the context. Bi-Directional Attention Flow (BiDAF) obtained a similarity matrix to capture relations between the context and question words and used this matrix to obtain context-to-question as well as question-to-context attention vectors in machine comprehension task [30]. The dynamic co-attention network learned co-dependent representations of the question and document [45].

2.3 Dialogue System with External Knowledge

There are many methods that incorporate the external knowledge into natural language processing and language understanding. Zhang et al. [12] learned multimodal taxonomies from the multimedia data on the Web. Many approaches make use of knowledge graphs by sub-graph matching and knowledge graph embedding learning. Hu et al. [11] constructed a semantic query graph to model the query intention and perfectly answered the ambiguous questions by sub-graph matching. Zhang et al. [47] incorporated multi-modal knowledge graph in question answering. Wang et al. [39] discussed how to make use of knowledge graph with knowledge graph embedding and applied it to various natural language processing tasks like relation extraction and

question answering. Wei et al. [40] incorporated the external knowledge base to complete sentences. It jointly learned sentence structures with the internal corpus, relational triples in the external knowledge bases, and semantic information in sentences. However, there are also many other approaches that use unstructured data as external knowledge. Xing et al. [43] and Li et al. [16] incorporated personal and topic information into sequence-to-sequence framework to generate informative responses. The topic information can be captured from social event [27]. Vougiouklis et al. [38] proposed a model based on coupling an RNN that processes each sentence of utterances word-by-word and a Convolution Neural Network (CNN) that extracts features from each set of sentences corresponding to this sequence of utterances. Madotto et al. [21] combined multi-hop attention mechanisms with the idea of pointer networks, which effectively incorporates KB information to choose the best answers. Raghu et al. [28] proposed a novel Hierarchical Pointer Generator Memory Network, in which the next word may be generated from the decoder vocabulary or copied from a hierarchical memory maintaining KB results and previous utterances.

Although the above approaches obtain good performance in knowledge-enhanced natural response generation, they are not well compatible with VAE or GAN frameworks. Different from existing works, we propose a novel knowledge-aware condition Wasserstein auto-encoder to incorporate external knowledge in the latent space of VAE and make use of both well-constructed knowledge graph and unstructured data in a unified framework to generate reasonable and informative responses.

3 METHODOLOGY

In this section, we first define the problem of interest. Then, we introduce the technical details about our proposed model.

3.1 Problem Formulation

Let $\mathcal{D} = [u_1, \ldots, u_{n+1}]$ denote a dialogue of $n+1$ utterances where $u_i = [w_1^i, \ldots, w_{|u_i|}^i]$ represents an utterance and w_j^i denotes the jth word in the utterance u_i. Given n historical utterances as dialogue context, denoted as $c = [u_1, \ldots, u_n]$, and external knowledge e, our goal is to generate the last utterance $x = u_{n+1}$ by estimating the conditional distribution $p_\theta(x|c, e)$. We call the last utterance *response*. We denote the length of *response* as m.

3.2 Overall Framework

Our objective is to learn a better response representation in the latent space. To this end, we present an approach called Knowledge-aware Attentive Wasserstein Adversarial Dialogue Response Generation (KAWA-DRG) to tackle this problem. Figure 1 presents the workflow of KAWA-DRG, which consists of three modules: Utterance and Context Encoder, Co-attention Encoder, and Knowledge-aware Condition Wasserstein Auto-Encoder.

- **Utterance Encoder and Context Encoder.** Utterance encoder is used to encode each utterance in a dialogue into the continuous representation. Context encoder is used to encode all utterances in the context into one representation.
- **Co-attention Encoder.** Co-attention module calculates the attention matrix and fuses the dialogue context and response to train a better auto-encoder.
- **Knowledge-aware Condition Wasserstein Auto-Encoder.** We use a condition Wasserstein auto-encoder to jointly model the distribution of responses with context and external knowledge and learn it with adversarial networks.

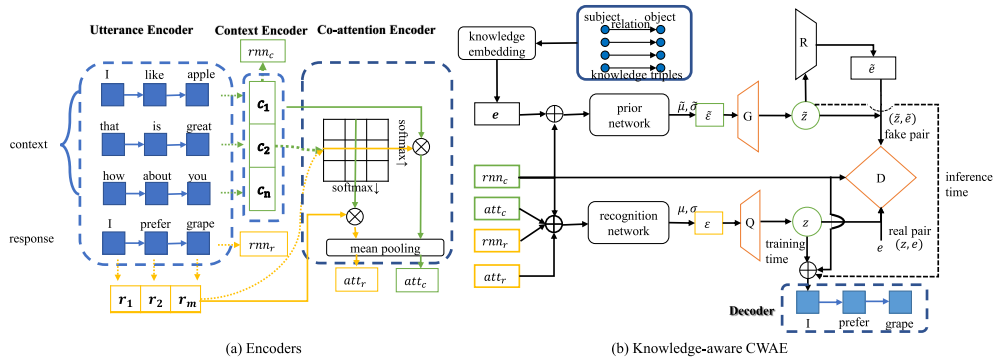


Fig. 1. Architecture of the proposed KAWA-DRG model. ⊕ denotes concatenation of the input vectors. ⊗ denotes matrix multiplication. In the training stage, we use the latent variable from recognition network z to generate responses. In the inference stage, we use $\tilde{z}$ from the prior network instead. The discriminator treats the original knowledge e and z as real pair and the recovered $\tilde{e}$ and $\tilde{z}$ as the fake pair.

Finally, we train a discriminator using Wasserstein distance to make the prior and posterior distribution closer, and ensure the latent variable learned from the prior distribution contains the information of external knowledge e.

3.3 Utterance Encoder and Context Encoder

We first introduce the utterance encoder and context encoder in our model.

Utterance encoder uses GRU [6] to transform between utterances and the vector representations of sentences. With bidirectional GRU, we encode each word in an utterance as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$, where forward is $\overrightarrow{h_t} = \overrightarrow{GRU}_{enc}(\overrightarrow{h}_{t-1}, w_t^n)$ and backward is $\overleftarrow{h_t} = \overleftarrow{GRU}_{enc}(\overleftarrow{h}_{t+1}, w_t^n)$. The representation of ith utterance in the context is the combination of the last cell's output in each direction, denoted as $c_i = [\overrightarrow{h_m}; \overleftarrow{h_1}]$.

With the utterance encoder, we can encode each word in the response as r_t, the entire *response* as rnn_r, and the response encoding matrix as:

$$R = [r_1, \ldots, r_m] \in \mathbb{R}^{d \times m}. \tag{1}$$

The context encoding matrix can be defined as:

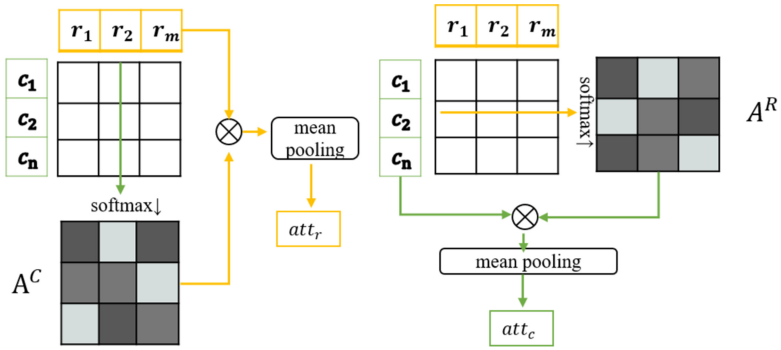
$$C = [c_1, \ldots, c_n] \in \mathbb{R}^{d \times n}. \tag{2}$$

The sentences in the context are computed in a similar manner. The context encoder takes the represention of each utterance c_i and the ith conversation floor (1 if the utterance is from the speaker of response) in the context as the input. Then they are fed into the GRU [6]. We use the last step's output of GRU as the representation of entire context: rnn_c.

3.4 Co-attention Encoder

We propose a co-attention mechanism to link the *response* and the context. The main difference between our method and the previous co-attention approach (DCN) [45] is that DCN concerns the word-level interaction between the document and the question. In contrast, we use co-attention to pay more attention to the important context sentence rather than all contexts. We adopt the following formulation for learning the affinity matrix:

$$A = \mathrm{F}(C)^T \mathrm{F}(R) \in \mathbb{R}^{n \times m}, \tag{3}$$



(a) Computation process of the attention representation of the response

(b) Computation process of the attention representation of the contex

Fig. 2. Co-attention encoder.

where $F(\cdot)$ is a function such as a multilayer perception (MLP) to allow for variation and compute the affinity matrix between the response space $R \in \mathbb{R}^{d\times m}$ and context space $C \in \mathbb{R}^{d\times n}$. The element A_{ij} indicates the ith utterance in the context's similarity with respect to the jth word in the response.

The affinity matrix A is then normalized column-wise with softmax function to produce the attention weights A^C across the response for each utterance in the context. In the same way, A^R is calculated by normalizing A row-wise.

$$A^C_{ij} = \frac{e^{A_{ij}}}{\sum_{j=1}^{m} e^{A_{ij}}}, \quad \text{for} \quad i = 1, \ldots, n; j = 1, \ldots, m, \tag{4}$$

$$A^R_{ij} = \frac{e^{A_{ij}}}{\sum_{i=1}^{n} e^{A_{ij}}}, \quad \text{for} \quad i = 1, \ldots, n; j = 1, \ldots, m. \tag{5}$$

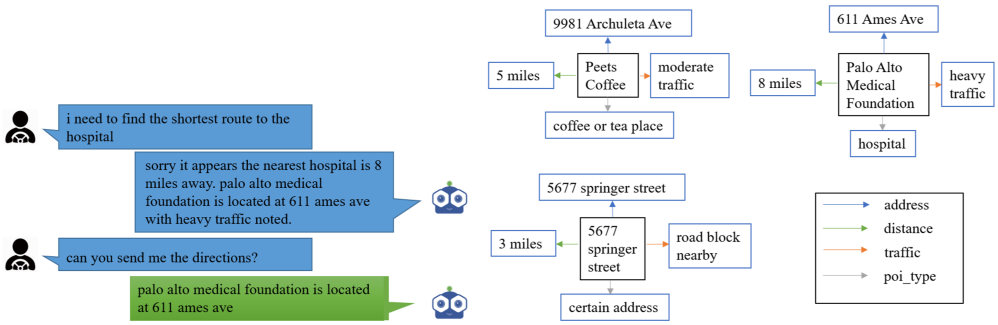
Then, we compute the attention representation of *response* in light of each utterance of the context R^{att} and *context* representation C^{att}:

$$R^{att} = R(A^C)^{\mathrm{T}} \in \mathbb{R}^{d\times n}, C^{att} = CA^R \in \mathbb{R}^{d\times m}. \tag{6}$$

Finally, since the context length and the response length vary in the different dialogue, we add a mean-pooling layer to get fixed size of the context and response representation att_c and att_r by $att_c = mean(C^{att})$ and $att_r = mean(R^{att})$. The procedure of how we calculate the context and response attention representation is shown in Figure 2. Therefore, the final representation of response and context are $\mathbf{r} = [rnn_r; att_r]$ and $\mathbf{c} = [rnn_c; att_c]$.

3.5 Knowledge-aware Wasserstein Adversarial Dialogue Response Generation

As we discussed in Section 1, the traditional Seq2Seq model tends to generate "safe response," i.e., "I don't know," which is due to the fundamental nature of statistical models. Generative Adversarial Nets [8] can jointly train a generative model and a discriminator to generate sharp and realistic images. Thus, this architecture could also be applied to response generation to ease the safe response problem, where the generative part can be a Seq2Seq-based model and the discriminative part can evaluate the quality of the generated responses. However, modeling discrete text tokens is difficult for GAN, since it is non-differentiable. In this manuscript, we follow Reference [10] to model the data distribution by training a GAN within the latent variable space.



(a) Example of utterances in the dialogue (b) Example of knowledge tripes relative to the dialogu

Fig. 3. Example of utterances and knowledge triples pair.

Specifically, we use the conversation-specific external knowledge and the representation of context to model the prior distribution of the response, and use the representation of context and response together to model posterior distribution of response. To force the model to learn from the knowledge, we recover it from the latent space variable. Finally, we train a discriminator using Wasserstein distance to make the prior and posterior distribution closer.

The conversation-specific knowledge depends on the dataset. For the SwitchBoard and DailyDialog, we use the topic information of the dialogue and encode the topic information as conversation-specific knowledge e. For these two datasets, the knowledge embedding $\mathbf{e}$ is initialized with one-hot vector. As for In-Car Assistant dataset, we use knowledge triples related to the conversation. Knowledge triples have a form of *(subject, relation, object)*. External knowledge in the In-Car Assistant dataset includes a calendar of event information, a collection of weekly forecasts for nearby cities. There might be several distinct rows of knowledge in each conversation. Take the utterances and knowledge triples in Figure 3; for example, in Figure 3(a), the utterances with the blue background are the context (first three utterances), and the utterance with green background (the last utterance) is the response we suppose our model to generate. This dialogue has 32 distinct knowledge triples related to the dialogue, but due to space limitations, we only display 12 of them in Figure 3(b). In Figure 3(b), the entities in the center with black border are *subjects*, arrows with different colors represent different *relations*, and the entities with the blue border are *objects*. We choose the one whose embedding of *subject* and *relation* are most similar to the words in the context and use the representation of *object* as e. That is, given the embedding of each word in the context $[w_1^1, \ldots, w_{|u_1|}^1, \ldots, w_1^n, \ldots, w_{|u_n|}^n]$ as $\mathbf{H} \in \mathbb{R}^{l \times d}$, the ith knowledge triple's *subject* embedding $\mathbf{s_i} \in \mathbb{R}^{1 \times d}$ and *relation* embedding $\mathbf{r_i} \in \mathbb{R}^{1 \times d}$, where $l = \sum_{i=1}^{n} |u_i|$ is the total number of words in the context, and d is the dimension of the embedding. The embeddings of words in the context and in the knowledge triples are pre-trained by word2vec [23] algorithm with the whole corpus. The similarity score between the knowledge triple and context is calculated as:

$$sim^{\mathbf{s_i}} = \sigma(\mathbf{s_i} \cdot \mathbf{H}^{\mathrm{T}}) \in \mathbb{R}^{1 \times l}, \quad sim^{\mathbf{r_i}} = \sigma(\mathbf{r_i} \cdot \mathbf{H}^{\mathrm{T}}) \in \mathbb{R}^{1 \times l}, score_i = \sum_{j=1}^{l} (sim_j^{\mathbf{s_i}} + sim_j^{\mathbf{r_i}}), \tag{7}$$

and we choose the ith knowledge triple with the highest $score_i$.

In the above example, the knowledge triple <*subject*: Palo Alto Medical Foundation, *relation*: address, *object*: 611 Ames Ave> gets the highest score among all the triples, since "Palo Alto Medical Foundation" directly appears in the context, and "address" is the most relative to the context

among all the relations. Therefore, the representation of words "611 Ames Ave" is used as knowledge representation e.

We model the distribution of the latent variable z by training a GAN within the latent space. We sample from the prior and posterior over the latent variable by transforming random noise ϵ using feed-forward neural networks.

Specifically, the first generator G produces the prior sample $\tilde{z} \sim p_\theta(z|\mathbf{c}, \mathbf{e})$ from the context and knowledge-dependent random noise $\tilde{\epsilon}$, while the approximate posterior sample $z \sim q_\phi(z|\mathbf{c}, \mathbf{r})$ is generated by a generator Q from context-dependent random noise ϵ. Both random noises are drawn from a normal distribution whose mean and covariance matrix are computed by fully-connected networks, *recognition network (RecNet)* for ϵ and *prior network (PriNet)* for $\tilde{\epsilon}$, denoted as:

$$\tilde{z} = G_\theta(\tilde{\epsilon}), \tilde{\epsilon} \sim \mathcal{N}(\tilde{\epsilon}; \tilde{\mu}, \tilde{\sigma}^2 I), \begin{bmatrix} \tilde{\mu} \\ log\tilde{\sigma}^2 \end{bmatrix} = \tilde{W} g_\phi \left(\begin{bmatrix} \mathbf{c} \\ \mathbf{e} \end{bmatrix} \right) + \tilde{b}, \tag{8}$$

and

$$z = Q_\phi(\epsilon), \epsilon \sim \mathcal{N}(\epsilon; \mu, \sigma^2 I), \begin{bmatrix} \mu \\ log\sigma^2 \end{bmatrix} = W f_\theta \left(\begin{bmatrix} \mathbf{c} \\ \mathbf{r} \end{bmatrix} \right) + b, \tag{9}$$

where $f_\theta(\cdot)$ and $g_\phi(\cdot)$ are feed-forward neural networks.

In addition, to provide supervision about knowledge e and ensure the latent variable $\tilde{z}$ to contain information about it, we try to reproduce it with another feed-forward network R. The recovered knowledge is denoted as $\tilde{e}$.

Then, we use the reparametrization trick [14] to obtain samples from *RecNet* (training, z) or *PriNet* (inference, $\tilde{z}$). At last, the decoder, which is a GRU network, predicts the words in *response*.

Our goal is to maximize the log-probability of a reconstructed response from z, while minimizing the divergence between $p_\theta(z|\mathbf{e}, rnn_c)$ and $q_\phi(z|\mathbf{r}, \mathbf{c})$. Thus, our objective function is:

$$\min_{\theta,\phi,\psi} -E_{q_\phi(z|\mathbf{r},\mathbf{c})} log p_\psi(x|z, rnn_c) + W(q_\phi(z|\mathbf{r}, \mathbf{c})||p_\theta(z|\mathbf{e}, rnn_c)), \tag{10}$$

where $p_\theta(z|\mathbf{e}, rnn_c)$ and $q_\phi(z|\mathbf{r}, \mathbf{c})$ are neural networks implementing Equations (4)–(7). $p_\psi(x|z, rnn_c)$ is a decoder. $W(\cdot||\cdot)$ is the Wasserstein distance between these two distributions [2].

During the training phrase, the decoder uses the generated response to optimize *recognition network* Q, encoders and decoder with reconstruction loss:

$$\mathcal{L}_{rec} = -E_{z=Q(\epsilon), \epsilon \sim RecNet(\mathbf{r},\mathbf{c})} log p_\psi(x|rnn_c, z). \tag{11}$$

We match the approximate posterior distribution of the latent variable by introducing an adversarial discriminator. The discriminator D, which treats the latent variable sampled from the prior network $\tilde{z}$ and recovered knowledge $\tilde{e}$ as fake pair and the latent variable sampled from the posterior network z and original knowledge e as the real, is trained by minimizing the discriminator loss:

$$\mathcal{L}_{disc} = E_{\epsilon \sim RecNet(\mathbf{r},\mathbf{c})}[D(Q(\epsilon), e, rnn_c)] - E_{\tilde{\epsilon} \sim PriNet(\mathbf{r},e)}[D(G(\tilde{\epsilon}), \tilde{e}, rnn_c)]. \tag{12}$$

3.6 Algorithm Explanation

All modules are trained together following Algorithm 1. The whole training process is done in three phrases:

(1) Training auto-encoder. Given a dialogue corpus, the steps 5–7 calculate the representation of context, response, and knowledge. Step 8 samples the latent variables from the posterior and prior distribution. The utterance encoder, context encoder, decoder, *RecNet*, and Q are trained simultaneously to minimize the reconstruction in Equation (11).

ALGORITHM 1: Training process of the proposed model (UEnc for utterance encoder, CEnc for context encoder, RecNet: recognition network; PriNet: prior network; Dec: decoder).

Input: Dialogue corpus $\mathcal{D} = (c_i, x_i, E_i)_{i=1}^{|\mathcal{D}|}$, discriminator iterations n_{critic}.
Initialize parameters: $\theta_{UEnc}, \theta_{CEnc}, \theta_{Dec}, \theta_{PriNet}, \theta_{RecNet}, \theta_Q, \theta_G, \theta_R, \theta_D$.

1: **while** not convergence **do**
2: Initialize $\mathcal{D}$
3: **while** $\mathcal{D}$ has unsampled batches **do**
4: Sample a mini-batch of N instances $(x_n, k_n, c_n)_{n=1}^{N}$ from $\mathcal{D}$
5: Get the representations of context, response, and knowledge. $rnn_r = UEnc(x_n), rnn_c = CEnc(c_n), e$
6: Calculate attention representation of context and response att_r, att_c according to Equations (3)–(5).
7: Generate the final representation of context and response $\mathbf{r} = [rnn_r; arr_r], \mathbf{c} = [rnn_c; att_c]$.
8: Generate $z_n = Q(\epsilon_n)$, $\tilde{z} = G(\tilde{\epsilon_n})$, $\tilde{e} = R(\tilde{z})$ according to Equations (8), (9)
9: Update $\{\theta_{UEnc}, \theta_{CEnc}, \theta_{Dec}, \theta_{PriNet}, \theta_{RecNet}\}$ with $\mathcal{L}_{rec}$ according to Equation (11)
10: Update $\{\theta_G, \theta_Q, \theta_R, \theta_{PriNet}, \theta_{RecNet}\}$ with discriminator loss $\mathcal{L}_{disc}$ according to Equation (12)
11: **for** each $i \in [1, \ldots, n_{critic}]$ **do**
12: Repeat 4–8
13: Update θ_D
14: **end for**
15: **end while**
16: **end while**

(2) Training generators. Parameters of $G, Q, R, RecNet, PriNet$ are updated in Step 10 according to Equation (12).
(3) Training discriminator. D is then trained to correctly distinguish the true input signals (z, e) from the false signals $(\tilde{z}, \tilde{e})$ according to Equation (12). Then, we repeat the above training step n_{ctitic} times. We will discuss the parameter n_{ctitic} in the next section.

4 EXPERIMENT

4.1 Experiment Setting

4.1.1 Datasets. We conduct our experiments on four English dialogue datasets: DailyDialog [18], Switchboard [13], In-Car Assistant [7], and Ubuntu Dialogue Corpus [4]. The statistics of four datasets are shown in Table 2. Each dataset is split into training, validation, and testing sets with a ratio of 2,316:60:62 for Switchboards [10, 48], 10:1:1 for DailyDialog [10, 33], and 2,425:302:304 for In-Car Assistant [7]. For Ubuntu Dialogue Corpus [4], we follow the script[1] to create the training, validation, and testing sets with a ratio of 1,000,000:19,560:18,920.

Note that, we remove the last two utterances in In-Car Assistant dataset if the conversation ends with utterances such as "Driver: Thank you!" and "Car: You are welcome," since it is meaningless if we tend to generate the last utterance of the dialogue.

[1] https://github.com/rkadlec/ubuntu-ranking-dataset-creator.

Table 2. Statistics of Four Datasets on Open-domain Dialogue

Dataset	DailyDial	Switchboard	In-Car	Ubuntu
Num. of Talks	13,118	2,438	3,031	1,038,480
Num. of Topics	10	70	-	-
Avg. utterance length	15	15	9	17
Avg. turns	8	90	5.25	5
Train:Valid:Test	10:1:1	2,316:60:62	8:1:1	1,000,000:19,560:18,920

4.1.2 Baselines. To demonstrate the effectiveness of our proposed KAWA-DRG, we employ the following methods as baselines:

(1) CVAE-BOW: CVAE-BOW [48] is a conditional VAE model that introduces an auxiliary loss to require the decoder network to predict the bag-of-word in the response and uses discourse-level knowledge to generate better responses.
(2) CVAE-CO: CVAE-CO [33] is a collaborative variation encoder-decoder that effectively combines VAE and RNN.
(3) DialogWAE: DialogWAE [10] uses a conditional Wasserstein auto-encoder to generate multimodal responses.
(4) HRAN: HRAN [44] uses a two-level hierarchical attention mechanism that focuses on important parts between the utterances in the context and the words in the utterance.
(5) Mem2Seq: Mem2Seq [21] combines the multihop attention over memories with the idea of the pointer network.
(6) BoSsNet: BoSsNet [29] facilitates the disentangled learning of the response's language model and its knowledge incorporation.

We choose these baselines because we want to compare different aspects of our methods with the existing methods. CVAE-BOW and CVAE-CO use conventional measurement Kullback-Leibler distance, DialogWAE uses Wasserstein distance instead, HRAN is a generative method with attention mechanism. Mem2Seq and BoSsNet utilize knowledge information in dialogue generation.

4.1.3 Metrics. We apply three different metrics to evaluate the responses we generate. For fairness of comparison, we sample 10 responses for each context and calculate the following metrics, then normalize the score to 0–1 scale.

(1) BOW Embedding Similarity [20]. Bag-of-word Embedding obtains sentence similarity by calculating the cosine distance between generated responses and the references. We use three different ways to calculate the BOW score: 1. **Average**: cosine similarity between two averaged word embedding in the two utterances. 2. **Extrema**: cosine similarity between the large extreme values among the word embedding in the two utterances. 3. **Greedy**: words are greedily matched in two utterances based on the cosine similarities between their embedding to obtain a score, and then scores are averaged across all words.
(2) Smoothed Sentence-level BLEU [5]. BLEU measures the geometric mean of the modified n-gram precision with a length penalty. We compute BLEU scores for n<4 using smoothing techniques.[2] Then, we define the maximum of BLEU score among all samples as recall and the mean as precision, following DialogWAE [10].

[2]https://www.nltk.org/_modules/nltk/translate/bleu_score.html.

Table 3. The Results on Switchboard (A: Average, E: Extrema, G: Greedy, R: Recall, P: Precision)

Model	BOW			BLEU			DIST	
	A	E	G	R	P	F1	dist-1	dist-2
CVAE-BOW	0.828	0.555	0.840	0.298	0.272	0.284	0.107	0.099
CVAE-CO	0.839	0.557	0.855	0.299	0.269	0.283	0.111	0.110
DialogWAE	0.897	0.627	0.887	0.394	0.254	0.309	0.245	0.413
HRAN	0.891	**0.697**	0.451	0.128	0.037	0.050	0.160	0.135
Mem2Seq	0.673	0.589	0.650	0.063	0.301	0.048	0.054	0.013
KAWA-DRG	**0.927**	0.667	**0.892**	**0.433**	**0.268**	**0.330**	**0.307**	**0.519**

Table 4. The Results on DailyDialog (A: Average, E: Extrema, G: Greedy, R: Recall, P: Precision)

Model	BOW			BLEU			DIST	
	A	E	G	R	P	F1	dist-1	dist-2
CVAE-BOW	0.923	0.540	0.812	0.256	0.224	0.239	0.165	0.206
CVAE-CO	0.914	0.530	0.818	0.259	0.244	0.251	0.106	0.126
DialogWAE	0.948	0.578	0.846	**0.341**	**0.278**	**0.306**	0.327	0.583
HRAN	0.923	**0.661**	0.527	0.337	0.150	0.207	**0.437**	0.625
Mem2Seq	0.853	0.567	**0.861**	0.301	0.276	0.294	0.067	0.084
KAWA-DRG	**0.948**	0.583	0.852	0.339	0.264	0.297	0.357	**0.629**

(3) Distinct [15]. We report the degree of diversity by calculating the number of distinct unigrams (dist-1) and bi-grams (dist-2) among all sampled responses. The value is scaled by total number of generated tokens to avoid favoring long sentences.

4.1.4 Parameter Setting. The size of context window is set to 10 with max utterance length set to 40. The number of hidden units is set to 300 in all GRU cells. The prior and the recognition network are two-layer fully-connected networks of size 300 with tanh non-linearly. The Q, R, G, D are three-layer fully-connected networks of sizes 200, 200, 200, and 600 with ReLU non-linearity. The initial weights of the above fully-connected layers are sampled from a uniform distribution $[-0.02,0.02]$. The dimension of word embedding, latent variable size z, and knowledge embedding are all set to 200. We initialize word embedding with GloVE [26] embedding pretrained on Twitter corpus. For long-turn dialogues in Switchboard dataset, we break one dialogue into several pieces and treat each piece as a data sample.

4.2 Experimental Result and Analysis

4.2.1 Quantitative Results. The results of our evaluation experiments are presented in Tables 3–6.[3] From these results, we have the following observations:

(1) CVAE-BOW and CVAE-CO perform poorly on four datasets when generating responses with respect to the distinct metrics, especially for SwitchBoard and DailyDialog. This is because CVAE's variants do not capture the underlying semantic-related information and tend to produce generic responses. CVAE-CO has the high score with the diversity measure DIST on In-Car and Ubuntu Dialogue Corpus, because it generates more "random" responses, but these responses may not match the context.

[3]The results on SwitchBoard and DailyDialog datasets (except for HRAN) of baselines are from Reference [10].

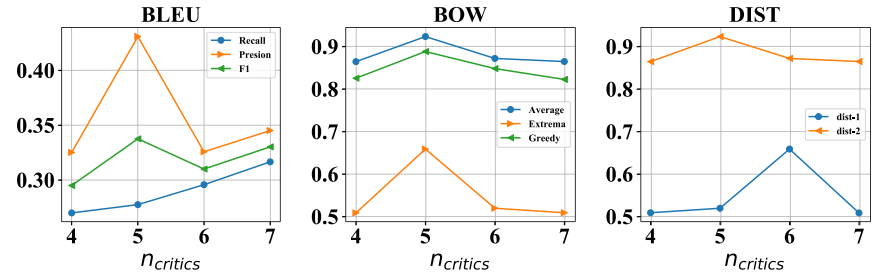
Table 5. The Results on In-Car Assistant (A: Average, E: Extrema, G: Greedy, R: Recall, P: Precision)

Model	BOW			BLEU			DIST	
	A	E	G	R	P	F1	dist-1	dist-2
CVAE-BOW	0.902	0.572	0.760	0.264	0.195	0.224	0.121	0.167
CVAE-CO	0.189	0.157	0.251	0.115	0.092	0.102	**0.530**	**0.748**
DialogWAE	0.896	0.535	0.789	0.294	0.275	0.284	0.141	0.161
HRAN	0.734	0.456	0.277	0.014	0.003	0.005	0.423	0.560
Mem2Seq	0.888	**0.661**	**0.957**	0.301	0.301	0.301	0.098	0.098
BossNet	0.0.898	0.633	0.668	0.313	0.313	0.313	0.088	0.096
KAWA-DRG	**0.905**	0.496	0.790	**0.324**	**0.313**	**0.318**	0.128	0.154

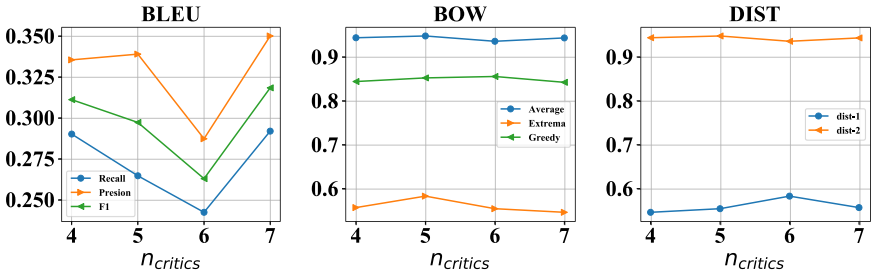
Table 6. The Results on Ubuntu (A: Average, E: Extrema, G: Greedy, R: Recall, P: Precision)

Model	BOW			BLEU			DIST	
	A	E	G	R	P	F1	dist-1	dist-2
CVAE-BOW	0.862	**0.542**	0.686	**0.368**	0.182	0.243	0.467	0.816
CVAE-CO	0.105	0.107	0.103	0.155	0.145	0.150	**0.807**	**0.958**
DialogWAE	0.854	0.522	0.764	0.347	0.236	0.281	0.168	0.240
HRAN	0.861	0.524	0.655	0.184	0.076	0.108	0.133	0.298
Mem2Seq	0.547	0.282	0.618	0.125	0.101	0.112	0.067	0.005
KAWA-DRG	**0.896**	0.519	**0.776**	0.361	**0.293**	**0.323**	0.254	0.469

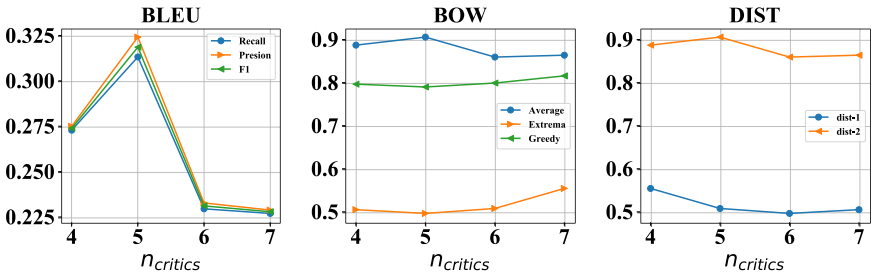
(2) DialogWAE obtains better performance than CVAE-BOW and CVAE-CO, which shows that it is useful to learn the latent space by employing adversarial learning.
(3) HRAN uses a two-level hierarchical attention mechanism. From Tables 3–6, we observe that HRAN gains the low BLEU score on all of the datasets, showing that it fails to produce accurate responses.
(4) Mem2Seq outperforms most baselines with respect to BOW and BLEU metrics, because it incorporates knowledge information. However, we can see that it gets the low scores on the diversity metrics on four datasets, showing that it fails to produce diverse responses.
(5) BoSsNet performs a little bit better than Mem2Seq, but it also fails to generate diverse responses.
(6) The proposed KAWA-DRG has higher BOW and BLEU scores than the baseline models in the majority of experiments, which indicates that it can generate more semantic responses with respect to the reference. Higher DIST scores show that it can generate more diverse responses. Especially on the SwitchBoard dataset, it has over 6% and 10% improvement on dist-1 and dist-2. However, in the other three datasets, our KAWA-DRG fails to gain the highest DIST score. In Table 5, the dist-1 is only slightly lower in HRAN. In Table 5 and Table 6, the DIST measure is much lower than CVAE-CO, because CVAE-CO tends to generate random and irrelevant answers due to its lower BOW and BLEU scores. We can also observe that KAWA-DRG fails to produce higher BLEU scores on the DailyDialog dataset. The possible reason is that there are synonyms of words of reference in the generated sentences, which leads to higher BOW and DIST scores.
(7) The results also show that KAWA-DRG consistently outperforms other baselines on all four datasets with different training settings in major of experiments. It demonstrates that the co-attention network together with the knowledge-aware condition Wasserstein auto-encoder



(a) SwitchBoard dataset



(b) DailyDialog dataset



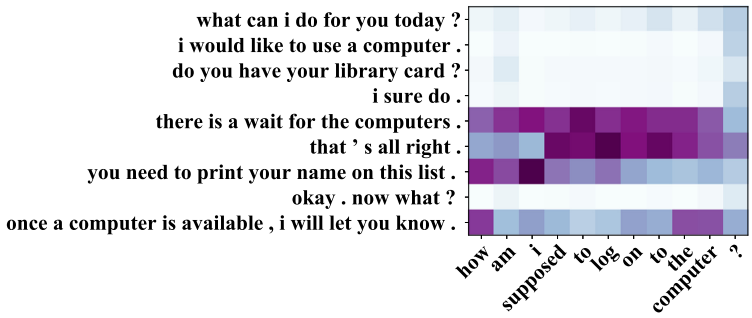
(c) In-Car Assistant dataset

Fig. 4. Performance with different discriminator iterations.

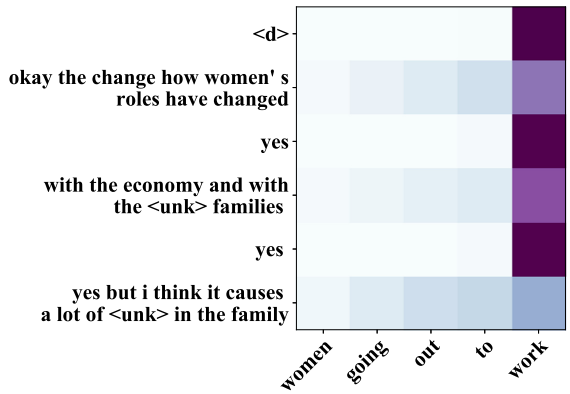
can better capture data distribution and the underlying relationship among data by integrating the interactive formation and grounding knowledge into a unified framework.

4.2.2 Parameter Analysis. **Different number of discriminator iterations$n_{critics}$**: We vary the number of critics from 4 to 7 and report the results on three datasets. In Figure 4(a), the BLEU, BOW, and dist-1 reach highest when the $n_{critics}$ is 5, and 6 for dist-2 on SwitchBoard dataset. In Figure 4(b), the best $n_{critics}$ is 7 for BLEU, 5 for BOW and DIST on the DailyDialog dataset. In Figure 4(c), the best $n_{critics}$ is 5 for BLEU, BOW, and DIST. In general, the performance increases as the number of critics increases in most cases and decreases once it achieves a threshold. The possible reason of this degradation is that the training difficulty increases as the number of critics increases. The optimal $n_{critics}$ is around 5 among all the datasets.

4.2.3 Attention Mechanism. To show the significance of the co-attention mechanism on exploiting important context and response pairs, we visualize the heat maps of A_C and A_R in Figure 5.



(a) Attention across the response for each utterance in the context



(b) Attention across the context for each word in the response

Fig. 5. Co-attention between the response and context.

Each interaction in the figure is marked with various background colors. The stronger the background color is, the more important the interaction is. Figure 5(a) shows the attention across the response for each utterance in the context. We can see that utterances containing "computer" (the fifth and ninth utterances) gain more attention with respect to the words in the response. One might wonder why the cell <"that's all right" (the sixth utterance), "supposed" (the fourth word)> is also highlighted, since the expression "that's all right" is really general. This is because the attention scores heavily rely on the embedding of the word, and the co-occurrence of "that's all right" and "supposed" is also high. The same situations occur between "that's all right" and "log"/"on." Figure 5(b) shows the attention across the context for each word in the response. We can see that the last word gains the most attention. Because we use the GRU as the encoder and the last representation of the word contains the information of previous words, we add the information of previous important words into the latent variable with attention.

4.2.4 Ablation Study. We study the impacts of various architectural decisions on model performance. We design four different variants of our model. Table 7 reports the ablation analysis conducted on In-Car dataset.

(1) In the variant KAWA-DRG-NoKnowledge, we remove the knowledge to observe the impact of knowledge information on latent variable generation. The full model gains

Table 7. Ablation Study on In-Car Assistant (A: Average, E: Extrema, G: Greedy, R: Recall, P: Precision)

Model	BOW			BLEU			DIST	
	A	E	G	R	P	F1	dist-1	dist-2
KAWA-DRG-NoKnowledge	0.897	**0.536**	0.781	0.313	0.287	0.299	**0.145**	**0.174**
KAWA-DRG-NoAttention	0.903	0.507	**0.796**	0.321	0.308	0.314	0.135	0.148
KAWA-DRG-NoDiscriminator	0.884	0..497	0.742	0.277	0.273	0.275	0.106	0.108
KAWA-DRG-MaxPooling	0.898	0.507	0.775	0.316	0.306	0.311	0.115	0.123
KAWA-DRG-Full	**0.905**	0.496	0.790	**0.324**	**0.313**	**0.318**	0.128	0.154

higher BLEU and BOW scores than NoKnowledge, indicating that our model has learned the knowledge information in the latent space, thus benefits the process of generating responses. Table 7 shows the result of the ablation study. KAWA-DRG-NoKnowledge has the higher DIST score than others, showing that it can generate more diverse responses, since knowledge here plays a role as a constraint. For instance, if the model obtains the knowledge "tomorrow will be sunny" when facing the question "what's the weather tomorrow?," then the model is less likely to generate a response such as "tomorrow will be rainy" or "tomorrow will be cloudy."

(2) KAWA-DRG-NoAttention is the variant that removes co-attention and treats all words and sequences as the same. The result slightly decreases after co-attention is removed, showing that co-attention between the context and response does provide additional information to assist in latent variable modeling.

(3) KAWA-DRG-NoDiscriminator, where we remove the discriminator. The BOW, BLEU, and DIST scores drop compared with the KAWA-DRG-Full, which shows the discriminator is necessary.

(4) KAWA-DRG-Maxpooling, where we replace the mean pooling with the max pooling to get the representation of the context and response. The result slightly decreases after the mean pooling layer is replaced by max pooling layer, validating that we make full use of the context information.

4.2.5 Case Study. Table 8 shows generated responses. We can see that our model can generate more coherent and meaningful responses. In the first dialogue, we can judge from the context that speaker A is an interviewer and speaker B is an interviewee, and the last utterance is speaker A asking about speaker B's experience of being a secretary. Under such context, our model tends to generate "job"-related responses, such as "No, it is a very difficult job" (Sample 2) and "I think I'm very good at that" (Sample 3).

The second dialogue most likely happened in a restaurant, and speaker A is a waiter and speaker B is a customer, and they are deciding what to eat. Under such a context, our proposed KAWA-DRG captures "fish" and "food" in the context, and therefore generates "food"-related response "Do you know what kind of food is it." However, the baseline model DialogWAE cannot generate such context-relevant responses.

In the third dialogue, speaker A (a seller) is trying to sell a credit card to speaker B. The baseline model does not generate anything relevant to the main topic "credit card," but our proposed model generates expressions like "I'm afraid I'll have to pay for my credit card" (Sample 3). At the same time, Sample 2 is also reasonable, since it is to respond to "But our credit card saves you 10 percent" in the last utterance of the context.

Table 8. Examples of Generated Responses. A, B Indicate Different Actors in the Dialogue

Index	Context	DialogWAE	KAWA-DRG
1	A: Have you got any experience in advertising? B: Yes, I have been working in the public relations. A: Do you have any practical experience as a secretary?	**1.** What a good idea. **2.** What the price is the most important month. **3.** Yes I've been thinking about it.	**1.** Yes, but this is my first time. **2.** No, it is a very difficult job. **3.** I think I'm very good at that. I think I'll have to be friend to find a job that I can do for the job.
2	A: Could I have some fish? B: Certainly. And what vegetable would you like? A: Oh, spinach, I think.	**1.** What is your market? **2.** Right. Do you have any supposed about your computer? **3.** A per room, please?	**1.** No problem. **2.** Ok, let me see, how much is the appreciate for you? **3.** Do you know what kind of food is it?
3	A: Could I interest you in our store credit card? B: No thanks, I already have credit cards. A: But our credit card saves you 10 percent.	**1.** That's right, but it's not a long time. **2.** That sounds good, how much is it. **3.** Oh, I don't know, but it's not a long time.	**1.** No, thanks. **2.** Not. It is a very expensive. **3.** Oh, yes. that's very kind of you. I'm afraid I'll have to pay for my credit card.

However, we can see there are still some grammar errors in the generated responses, for instance, Sample 2 in dialogue 3 cannot be regarded as a complete sentence. We will pay special attention to this problem and try to solve it in the future.

4.2.6 Human Evaluation. Except for the case study, we also compared KAWA-DRG with one of the baselines DialogWAE using one-vs-one human annotation.

Specifically, we find three students who speak English as human annotators. We show each annotator one context and two sets of responses, one from KAWA-DRG and the other from DialogWAE. There are 10 responses in each set, which is the top-10 results in the beam search. Then, we ask the annotator to choose which set is better. The criteria are, set A is better than set B if:

(1) A has more relevant, consistent to the context, and fluent responses than B;
(2) Both A and B have the same amount of relevant, consistent, and fluent responses, but responses in A are more informative and responses in B are general (like "Thanks" or "OK");
(3) Both A and B have the same amount of relevant, consistent, fluent, and informative responses, but responses in A are more diverse; for example, set A only contains two same responses, but all responses in set B are the same.
(4) If, finally, the annotator cannot tell which set is better, then she/he was supposed to label a "tie" to the test example.

Each annotator individually judges all the test samples for each KAWA-DRG/Baseline pair. All examples are randomly shuffled before they are represented to the annotators. The results of human evaluation are listed in Table 9. From the results, we can observe that our proposed

Table 9. Human Evaluation Results (in %)

Model	Win	Loss	Tie	Gains (Win-Loss)
KAWA-DRG vs CVAE-BOW	37.3	21.7	41.0	15.6
KAWA-DRG vs CVAE-CO	47.8	17.4	34.8	30.4
KAWA-DRG vs DialogWAE	42.7	29.3	28.0	13.4
KAWA-DRG vs HRAN	44.7	21.3	34.0	23.4
KAWA-DRG vs Mem2Seq	41.3	30.4	28.3	10.9

KAWA-DRG outperforms all the baselines. Even compared with Mem2Seq, the winning percentage of KAWA-DRG is 41.3%, and the tie rate is 28.3%. In addition, KAWA-DRG achieves performance gains (win-loss) over 10%. The results indicate that our model can generate more readable, relative, fluent, and diverse responses.

5 CONCLUSION

We propose a novel framework, called **KAWA-DRG**, to model conversation-specific external knowledge and the importance variances of dialogue context in a unified adversarial encoder-decoder learning framework. We analyze our model from different aspects including quantitative analysis, parameter analysis, components analysis, ablation study, case study, and human evaluation. All the experimental results show that the proposed KAWA-DRG is reasonable and outperforms the existing methods.

In the future, we plan to incorporate richer knowledge into the framework to generate more grounding responses, and will apply the proposed model for various goal-oriented dialogue systems.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 2425–2433.

[2] Martïn Arjovsky, Soumith Chintala, and Lëon Bottou. 2017. Wasserstein GAN. *arXiv preprint arXiv:* 1701.07875 (2017).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR'15)*.

[4] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference*. The Associationfor Computer Linguistics, 285–294.

[5] Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the 9th Workshop on Statistical Machine Translation*. Association for Computational Linguistics. 362–367.

[6] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1724–1734.

[7] Mihail Eric and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the SIGDIAL Conferences*. Association for Computational Linguistics, 37–49.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'14)*. Curran Associates, Inc., 2672–2680.

[9] Kartik Goyal, Graham Neubig, Chris Dyer, and Taylor Berg-Kirkpatrick. 2018. A continuous relaxation of beam search for end-to-end training of neural sequence models. *arXiv preprint arXiv:* 1708.00111 (2018).

[10] Xiaodong Gu, Kyunghyun Cho, JungWoo Ha, and Sunghun Kim. 2018. DialogWAE: Multimodal response generation with conditional Wasserstein auto-encoder. *arXiv preprint arXiv:* 1805.12352 (2018).

[11] Sen Hu, Lei Zou, Jeffrey Xu Yu, Haixun Wang, and Dongyan Zhao. 2018. Answering natural language questions by subgraph matching over knowledge graphs. *IEEE Trans. Knowl. Data Eng.* 30 (2018), 824–837.

[12] Xiaowen Huang, Shengsheng Qian, Quan Fang, Jitao Sang, and Changsheng Xu. 2018. CSAN: Contextual self-attention network for user sequential recommendation. In *Proceedings of the ACM Multimedia Conference (MM'18)*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 447–455.

[13] John Godfrey and Edward Holliman. 1997. Switchboard-1 release 2. In *Proceedings of the Linguistic Data Consortium*.

[14] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:* 1312.6114 (2013).

[15] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'16)*. Association for Computational Linguistics. 110–119.

[16] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 994–1003. Retrieved from http://www.aclweb.org/anthology/P16-1094.

[17] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Daniel Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2157–2169.

[18] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the the 8th International Joint Conference on Natural Language Processing*. AFNLP, 986–995.

[19] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *CoRR* abs/1703.03130 (2017).

[20] 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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2122–2132.

[21] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. *arXiv preprint arXiv:* 1804.08217 (2018).

[22] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'10)*. ISCA, 1045–1048.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 3111–3119.

[24] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[25] Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1792–1801.

[26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

[27] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao. 2016. Multi-modal event topic model for social event analysis. *IEEE Trans. Multimedia* 18, 2 (2016), 233–246. DOI: https://doi.org/10.1109/TMM.2015.2510329

[28] Dinesh Raghu, Nikhil Gupta, and Mausam. 2018. Hierarchical-pointer generator memory network for task oriented dialog. *arXiv preprint arXiv:* 1805.01216 (2018).

[29] Dinesh Raghu, Nikhil Gupta, and Mausam. 2019. Disentangling language and knowledge in task-oriented dialogs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT'19) Volume 1 (Long and Short Papers)*. 1239–1255.

[30] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*.

[31] Iulian 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 *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI, 3776–3783.

[32] Iulian 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 31st AAAI Conference on Artificial Intelligence*. AAAI, 3295–3301.

[33] Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. *arXiv preprint arXiv:* 1802.02032 (2018).

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'15)*. Curran Associates, Inc., 3483–3491.

[35] Jun Song, Jun Xiao, Fei Wu, Haishan Wu, Tong Zhang, Zhongfei Zhang, and Wenwu Zhu. 2017. Hierarchical contextual attention recurrent neural network for map query suggestion. *IEEE Trans. Knowl. Data Eng.* 29 (2017), 1888–1901.

[36] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'17)*.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 3156–3164.

[38] Pavlos Vougiouklis, Jonathon S. Hare, and Elena Paslaru Bontas Simperl. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of the International Conference on Computational Linguistics (COLING'16)*. The International Committee on Computational Linguistics, 3370–3380.

[39] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.* 29 (2017), 2724–2743.

[40] Xiaochi Wei, Heyan Huang, Liqiang Nie, Hanwang Zhang, Xian-Ling Mao, and Tat-Seng Chua. 2017. I know what you want to express: Sentence element inference by incorporating external knowledge base. *IEEE Trans. Knowl. Data Eng.* 29 (2017), 344–358.

[41] Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1296–1306.

[42] Fei Wu, Xinyu Duan, Jun Xiao, Zhou Zhao, Siliang Tang, Yin Zhang, and Yueting Zhuang. 2017. Temporal interaction and causal influence in community-based question answering. *IEEE Trans. Knowl. Data Eng.* 29 (2017), 2304–2317.

[43] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 3351–3357.

[44] Chen Xing, Wei Wu, Yu Wu, Ming Zhou, Yalou Huang, and Wei-Ying Ma. 2018. Hierarchical recurrent attention network for response generation. *arXiv preprint arXiv:* 1701.07149 (2018).

[45] Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*.

[46] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2Image: Conditional image generation from visual attributes. *arXiv preprint arXiv:* 1512.00570v2 (2016).

[47] Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2019. Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering. In *Proceedings of the 27th ACM International Conference on Multimedia, (MM'19)*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1089–1097.

[48] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics, (Volume 1: Long Papers)*. Association for Computational Linguistics, 654–664.

[49] Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv preprint arXiv:* 1709.04264 (2017).

Received July 2019; revised January 2020; accepted February 2020