RT-KGD: Relation Transition Aware Knowledge-Grounded Dialogue Generation

Kexin Wang[1], Zhixu Li[2(✉)], Jiaan Wang[1], Jianfeng Qu[1(✉)], Ying He[3], An Liu[1], and Lei Zhao[1]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
{kxwang1,jawang1}@stu.suda.edu.cn, {jfqu,anliu,zhaol}@suda.edu.cn

[2] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
zhixuli@fudan.edu.cn

[3] IFLYTEK Research, Suzhou, China
yinghe@iflytek.com

Abstract. Grounding dialogue system with external knowledge is a promising way to improve the quality of responses. Most existing works adopt knowledge graphs (KGs) as the external resources, paying attention to the contribution of entities in the last utterance of the dialogue for context understanding and response generation. Nevertheless, the correlations between knowledge implied in the multi-turn context and the transition regularities between relations in KGs are under-explored. To this end, we propose a Relation Transition aware Knowledge-Grounded Dialogue Generation model (RT-KGD). Specifically, inspired by the latent logic of human conversation, our model integrates dialogue-level relation transition regularities with turn-level entity semantic information. In this manner, the interaction between knowledge is considered to produce abundant clues for predicting the appropriate knowledge and generating coherent responses. The experimental results on both automatic evaluation and manual evaluation indicate that our model outperforms state-of-the-art baselines.

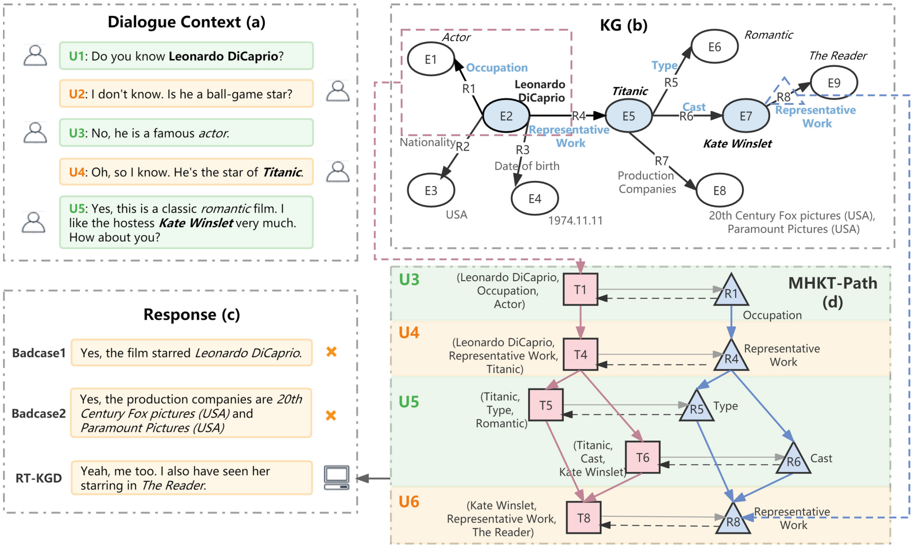
Keywords: Knowledge-Grounded Dialogue · Response generation · Relation transition regularity

1 Introduction

Knowledge-Grounded Dialogue Generation (KGD) aims at generating an informative response based on both dialogue context and external knowledge [6,9]. Current works typically utilize structured knowledge graphs (KGs) [16,32,38] or unstructured texts [9,36] as knowledge resources. Incorporating external knowledge related to the dialogue context has proven to alleviate generating meaningless and bland responses caused by traditional generative models, such as "*I don't know*" and "*You are right*" [11].

K. Wang and Z. Li—The first two authors made equal contributions to this work.

U. Sattler et al. (Eds.): ISWC 2022, LNCS 13489, pp. 319–335, 2022.
https://doi.org/10.1007/978-3-031-19433-7_19



U1: Do you know **Leonardo DiCaprio**?

U2: I don't know. Is he a ball-game star?

U3: No, he is a famous *actor*.

U4: Oh, so I know. He's the star of ***Titanic***.

U5: Yes, this is a classic *romantic* film. I like the hostess ***Kate Winslet*** very much. How about you?

Badcase1: Yes, the film starred *Leonardo DiCaprio*. ✗

Badcase2: Yes, the production companies are *20th Century Fox pictures (USA)* and *Paramount Pictures (USA)* ✗

RT-KGD: Yeah, me too. I also have seen her starring in *The Reader*.

Fig. 1. An illustrative example from KdConv [39]. Based on the dialogue context (a) and the related KG (b), KGD is required to generate a response (c) guided by the MHKT-Path (d). The **bold** denotes the core entities in the dialogue, and the *Italic* denotes related knowledge values involved in the dialogue.

The existing works mainly focus on two aspects in KGD task: knowledge-enhanced context understanding [2,29] and knowledge-fused response generation [13,14]. Traditional efforts [2,6,39] simply treat the relevant external knowledge as the textual complementary to the dialogue context for both context understanding and response generation, neglecting considerable structural information in KGs. Some recent works [8,16,32] realize that the correlation between entities plays an important role in continuing dialogue, thus propose to excavate the valuable structural information between entities in the knowledge graph to predict the entities that might appear in the next response. The predicted entities are further used to guide the response generation. For example, DialKG Walker [16] treats the entities mentioned in the last utterance as the starting nodes and further retrieves relevant entities from KG within two hops. DuConv [30] pre-defines a topic goal including two entities for each dialogue, which guides the model to start with the first entity and gradually transition to the second one.

Despite their great contributions, there are two main drawbacks: on the one hand, the **entity-guided KGD** methods [16,32] consider the entities in the dialogue as the only guidance knowledge for context understanding and response generation, which neglects the importance of **relations** between entities in the KG. However, the regularity behind human conversation can be summarized as a sequence of topics, where each topic may correspond to a relation between entities rather than a single entity in the KG. On the other hand, the existing

KGD methods [8,16] only care about the information in the last dialogue turn for predicting the subsequent knowledge, which is insufficient to learn how human transfer topics across a multi-turn dialogue. Taking Fig. 1 as an example, both badcase 1 and badcase 2 are flawed generated results based on the dialogue context. Badcase 1 demonstrates that the generated response might be redundant and incoherent without modeling multiple turns of knowledge, while badcase 2 reveals an abrupt transition in the topic since the latent relation transition path throughout the dialogue is ignored.

In this paper, we propose a novel KGD model: Relation Transition aware Knowledge-Grounded Dialogue Generation (`RT-KGD`), which models the knowledge transition across multi-turn dialogue by integrating dialogue-level relation transition regularities with turn-level entity semantic information. Specifically, we obtain all the relations and entities contained in the multi-turn dialogue context to construct a so-called Multi-turn Heterogeneous Knowledge Transition Path (`MHKT-Path`), which can be viewed as a subgraph of the external KG integrated with the sequential information of relations and entities in the multi-turn dialogue. Based on the constructed `MHKT-Path`, a knowledge prediction module is proposed to retrieve the triplets that might appear in the subsequent response from the external KG, and they are further fused for triplet prediction. Finally, the subsequent response is generated conditioned on both dialogue context and the predicted triplet. As the example shown in Fig. 1, the `MHKT-Path` grasps the latent conversation regularity of human beings, and the generated response based on the proposed `RT-KGD` is informative and coherent with the dialogue context.

The main contributions of this paper are concluded as follows:

- To the best of our knowledge, we are the first to incorporate the relation transition across multi-turn dialogue into the KGD task. In this manner, the regularity behind human conversation can be portrayed by integrating relation transition paths and entity semantic information.
- We propose to build a Multi-turn Heterogeneous Knowledge Transition Path (MHKT-Path) for each dialogue, which integrates the structure information of external KG and the sequential information of knowledge with the multi-turn dialogue. Based on MHKT-Path, our model then retrieves appropriate knowledge from the KG to guide the next response generation.
- The experimental results on a multi-domain knowledge-driven dialogue dataset (i.e., KdConv [39]) indicate that our model outperforms strong baseline models in both automatic and manual evaluation.

2 Related Work

According to whether to introduce knowledge, we categorize previous dialogue generation works into *Vanilla Dialogue Generation* and *Knowledge-grounded Dialogue Generation*.

Vanilla Dialogue Generation. Early dialogue systems typically employ Sequence-to-Sequence (Seq2Seq) models to generate responses [20,21,31], which is further improved with advanced context encoders [20,31] or more efficient response generation methods [2,33,37]. Recently, pre-trained generative models with the backbone of Transformer [25], such as GPT-2 [18] and BART [10], achieve promising performance in many text generation tasks. There is increasing work focusing on designing Transformer-based pre-trained dialogue models. Among them, Blender [19] enhances Transformer architecture and show their superiority in dialogue generation. DialoGPT [35] extends GPT-2 [18] for response generation. Besides, PLATO [3] pre-trains unified language models for both bi-directional encoding and uni-directional decoding. Nevertheless, they can only implicitly learn dialogue strategies and commonsense knowledge from dialogue corpora, resulting in limited transferability to other dialogue scenes.

Knowledge-Grounded Dialogue Generation. A promising way to generate meaningful and informative responses is to utilize external knowledge to guide the models. Generally, the external knowledge comes from textual corpora [9], commonsense knowledge graphs [29,32,38], and domain knowledge graphs [30,39]. To utilize the knowledge, [6,26] adapt the memory network [23] to store the relevant knowledge and then generate responses conditioned on both dialogue context and stored knowledge. Besides, [12,29] employ the posterior distribution of knowledge to guide its prior distribution, leading to accurate knowledge selection and high-quality generated responses. Furthermore, some work [13,14,29] leverages copy mechanism to copy words from knowledge sources directly and generate more informative responses. Although great progress has been made, the structural information of KG is neglected, which might lead to suboptimal responses.

To effectively excavate the structural information, some researchers attempt to utilize graph neural networks on KG to obtain its structure-aware representation that is further incorporated into dialogue generation [16,32,38]. AttnIO [8] leverages bi-direction attention flows to propagate messages from the entities appearing in the last utterance to their neighbor entities in KG. ConceptFlow [32] applies a graph attention mechanism to attend to appropriate concepts conditioning on dialogue context for responses generation, where the concepts are extracted from ConceptNet [22], a large-scale commonsense knowledge graph. Unlike previous research, our `RT-KGD` (1) refines the dialogue-level knowledge transition from different granularity; (2) incorporates the related knowledge based on the whole dialogue context rather than only the last utterance.

3 Methodology

In this section, we formally define the knowledge-ground dialogue generation task (Sect. 3.1) and then elaborate on four principal components of our `RT-KGD` model. As illustrated in Fig. 2, our model first constructs the multi-turn heterogeneous knowledge transition path (`MHKT-Path`) for the given dialogue context (Sect. 3.2) and then encodes the `MHKT-Path` by a knowledge encoder (Sect. 3.3).

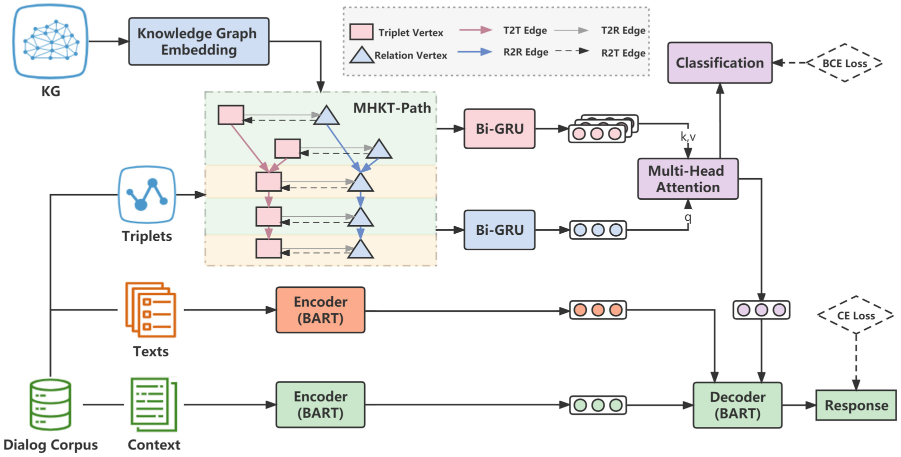


Fig. 2. The architecture of the proposed `RT-KGD` model.

Next, the predicted triplet from a knowledge prediction (Sect. 3.4) is finally incorporated into the subsequent response, which is generated by a knowledge-enhanced encoder-decoder (Sect. 3.5).

3.1 Task Formulation

Given a dialogue context $C = \{u_1, u_2, \cdots, u_{n-1}\}$, where u_i represents the i-th utterance. Each u_i corresponds to a knowledge triplet set $K_i = \{(h_{i_1}, r_{i_1}, t_{i_1}), (h_{i_2}, r_{i_2}, t_{i_2}), \cdots, (h_{i_{|K_i|}}, r_{i_{|K_i|}}, t_{i_{|K_i|}})\}$ $(|K_i| \geq 0)$, where (h, r, t) means that head entity h and tail entity t have a relation r, and a descriptive text set $S_i = \{s_{i_1}, s_{i_2}, \cdots, s_{i_{|S_i|}}\}$ $(|S_i| \geq 0)$. All knowledge triplets and descriptive texts are from domain knowledge graph $\mathcal{G}$ and corpus $\mathcal{O}$. The goal of knowledge-grounded dialogue systems is to generate a proper response u_n based on the dialogue context C, knowledge graph $\mathcal{G}$, and knowledge corpus $\mathcal{O}$.

3.2 Multi-turn Heterogeneous Knowledge Transition Path

To integrate dialogue-level relation transition regularities with turn-level entity semantic information, we utilize the knowledge triples associated with the given dialogue context, i.e., $K = K_1 \cup K_2 \cup \cdots \cup K_{n-1}$, to construct the multi-turn heterogeneous knowledge transition path, which is called `MHKT-path`. As shown in Fig. 2, `MHKT-path` contains two types of vertices, i.e., triplet vertices and relation vertices. In detail, each triplet vertex represents a knowledge triplet belonging to K, and corresponds with a relation vertex which is extracted from it. There are four types of edges in `MHKT-Path`: (1) the *triplet-to-triplet* edge links the triplet vertices associated in one utterance with others in the neighbor utterances; (2) the paired *triplet-to-relation* and (3) *relation-to-triplet* edges denote the bi-directional interaction between triplet vertices and their corresponding relation vertices; (4) the *relation-to-relation* edge links relation vertices with each other only if their corresponding triplet vertices are connected. In this manner, the knowledge transition of both turn-level triplets and dialogue-level relations is integrated into the `MHKT-Path`.

3.3 Knowledge Encoder

The knowledge encoder learns the representation of the vertices in `MHKT-Path`. Specifically, it contains vertex initializer and graph layers to initialize and update the vertex representations.

Vertex Initializer. Instead of directly using the average word embeddings of the flat texts in entities and relations, we employ a KG embedding algorithm (i.e., TransR [15]) to initialize the representation of vertices in our `MHKT-Path`[1]:

$$h_{e_i}^0 = \text{TransR}(e_i) \tag{1}$$

where $e_i \in \mathcal{K}$ denotes a KG element (e.g., entity or relation), $h_{e_i}^0$ means the initialized representation of e_i. TransR$(\cdot)$ represents the TransR KG embedding function, learned by projecting entities from entity space to different relation spaces and building translations between the projected entities. In this way, the learned representation of KG elements in $\mathcal{K}$ contain the global KG structural information due to their interaction in KG [4,15,34].

For relation vertex in `MHKT-Path`, we directly use $h_{e_i}^0$ as its initial representation. For triplet vertex (h_i, r_i, t_i), we calculate its representation as:

$$\text{TransR}(h_i) \oplus \text{TransR}(r_i) \oplus \text{TransR}(t_i) \tag{2}$$

where $\oplus$ denotes concatenation.

Graph Layers. Graph layers are used to update the vertex representations with the local structural information in the established `MHKT-Path`. Here, we employ the Heterogeneous Graph Transformer (HGT) [7] as the graph layers since it is aware of different types of vertices and edges. Given the `MHKT-Path`, the representation of each vertex v_i is updated by aggregating its neighbor information:

$$HGT(h_{v_i}^{\ell}) = \underset{\forall v_{src} \in N(v_i)}{\mathbf{Aggregate}} \left(\mathbf{Attention}(v_i^{\ell-1}, v_{src}^{\ell-1}) \cdot \mathbf{Message}(v_{src}^{\ell-1}) \right) \tag{3}$$

where $N(v_i)$ is the neighbor vertices set of v_i, the **Aggregate**$(\cdot)$, **Attention**$(\cdot)$, and **Message**$(\cdot)$ are three basic operators in HGT:

- **Attention**$(\cdot)$ calculates the mutual attention of each vertex pair, where each type of vertex and edge has a unique linear projection.

[1] We also attempt to encode entities and relations based on word embedding, as suggested by [27,28], the results underperform that of using TransR.

- **Message**$(\cdot)$ transfers information from different types of neighbor vertices of each vertex v_i.
- **Aggregate**$(\cdot)$ integrates messages from neighbor vertices with attention weights to the core vertex v_i.

Finally, for vertex v_i, we concatenate the final node representation and the corresponding initial node representation with a simple linear projection:

$$h_{v_i}^{HGT} = W([h_{v_i}^0 \oplus h_{v_i}^L]) \tag{4}$$

where W is trainable parameters, L is the number of layers of HGT.

3.4 Knowledge Predictor

After obtaining the final representations of both triplet and relation vertices in `MHKT-Path`, the knowledge predictor is used to predict the knowledge which might be implied in the response. There are three parts to knowledge prediction, i.e., relation prediction, relation-aware triplet prediction, and multi-label triplet classification. Since the knowledge encoder aggregates only local neighborhood information, we further employ the bi-directional gated recurrent unit (Bi-GRU) [5] to enrich the sequential representations of relations and triplets.

In detail, we first treat the average vertices representation in dialogue order as the input of Bi-GRU. Suppose there are m relation vertices and m triplet vertices in turn i. The relation vertices in turn i are denoted as $\{r_{i,j}\}_{j=1}^{m}$, whose average representation is shown as follows:

$$R_i^0 = Mean(h_{r_{i,1}}^{HGT}, \cdots, h_{r_{i,m}}^{HGT}) \tag{5}$$

Similarly, the triplet vertices in turn i are denoted as $\{t_{i,j}\}_{j=1}^{m} \subset \mathcal{K}$, whose average representation is:

$$T_i^0 = Mean(h_{t_{i,1}}^{HGT}, \cdots, h_{t_{i,m}}^{HGT}) \tag{6}$$

Relation Prediction. The relation prediction part is to obtain the n-th relation hidden state $h_r^{GRU}(n)$ based on the previous $n-1$ turns relation representation. At step t of relation prediction, Bi-GRU generates the t-th relation hidden state as follows:

$$\begin{aligned} h_r^{GRU}(t) &= [h_r^{fw}(t); h_r^{bw}(t)] \\ &= [\overrightarrow{GRU}(R_t^0, h_r^{fw}(t-1)); \overleftarrow{GRU}(R_t^0, h_r^{bw}(t-1))] \end{aligned} \tag{7}$$

Relation Transition Aware Triplet Prediction. Different from the relation, we utilize Bi-GRU to obtain $n-1$ triplet hidden states $h_t^{GRU}(1), \cdots, h_t^{GRU}(n-1)$ based on the input $T^0 = T_1^0, \cdots, T_{n-1}^0$. For the i-th triplet, its hidden state is calculated as follows:

$$\begin{aligned} h_t^{GRU}(i) &= [h_t^{fw}(i); h_t^{bw}(i)] \\ &= [\overrightarrow{GRU}(T_i^0, h_T^{fw}(i-1)); \overleftarrow{GRU}(T_i^0, h_t^{bw}(i-1))] \end{aligned} \tag{8}$$

After obtaining the predicted n-th relation hidden state and $n-1$ triplet hidden states, we employ multi-head attention [25] to jointly attend to the information from both dialogue level and turn level. Thus the predicted triplet representation $h_{t_n}^{ATT}$ is calculated as follows:

$$\begin{aligned} \alpha_i &= softmax_i\left({h_{r_n}^{GRU}}^T h_{t_i}^{GRU}\right) \\ h_{t_n}^{ATT} &= \mathop{\Big\Vert}_{d=1}^{D} \sum_{i=1}^{n-1} \alpha_i^d h_{t_i}^{GRU} \end{aligned} \tag{9}$$

where D denotes the number of attention heads.

Multi-label Triplet Classification. Since there might be multiple knowledge in the next response, the multi-label classification is adapted to map the predicted triplet representation to a label vector, where the number of labels is the total number of triplets in the knowledge graph $\mathcal{G}$.

Formally, let label $l = W_l(h_{t_n}^{ATT}) \in \mathcal{R}^{|\mathcal{K}|}$, where W_l is a trainable parameter and $|\mathcal{K}|$ is the total triplet size. The target label is denoted as $y \in \{0,1\}^{|\mathcal{K}|}$. Then we adapt the binary cross-entropy (BCE) loss to supervise the classification of triplets:

$$L_{BCE} = -\frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} \Big[y_i log(\sigma(l_i)) + (1-y_i) log(1-\sigma(l_i)) \Big] \tag{10}$$

where $\sigma(\cdot)$ is sigmoid function.

3.5 Knowledge-Enhanced Encoder-Decoder

We employ pre-trained BART [10] as the backbone of our KGD model, which aims to generate the final response based on dialogue context C, predicted triplet representation K and corresponding descriptive texts S. The input dialogue context is formed as "`[CLS]` u_1 `[SEP]` u_2 `[SEP]` $\cdots$ `[SEP]` u_{n-1} `[SEP]`", where `[CLS]` and `[SEP]` are two special tokens to indicate the utterance boundaries. Then, the input is automatically tokenized by the BART's tokenizer, followed by a stack of BART encoder layers. Next, the context-aware representation of each token is obtained from the output of the last encoder layer of BART:

$$h_1^C, \cdots, h_{|C_{inp}|}^C = BART_{enc}(C) \tag{11}$$

where $|C_{inp}|$ indicates the number of tokens in the input sequence, $BART_{enc}(\cdot)$ denotes the BART encoder, and h_i^C is the context-aware representation of the i-th token in the sequence.

Similarly, for the descriptive text set $S = \{S_1, S_2, \cdots, S_{n-1}\}$ corresponding to the context C, each S_i is encoded by the BART encoder, where the input is formed as "`[CLS]` S_{i_1} `[SEP]` S_{i_2} `[SEP]` $\cdots$ [SEP]$S_{i_{|S_i|}}$". We take the context-aware final representation of [CLS] as the sentence representation, and the encoded sentence embedding of the i-th turn is obtained as follows:

$$h_i^S = BART_{enc}(S_i) \tag{12}$$

Finally, the response is generated by the BART decoder, conditioning on the BART-encoded dialogue context $h_1^C, \cdots, h_{|C_{inp}|}^C$, descriptive sentences $h_1^S, h_2^S, \cdots, h_{|n-1|}^S$ and predicted triplets $h_{t_n}^{ATT}$:

$$G = BART_{dec}([h_1^C; h_2^C \cdots; h_{|C_{inp}|}^C; h_1^S; h_2^S; \cdots; h_{|n-1|}^S; h_{t_n}^{ATT};]) \tag{13}$$

where G is the representation of generated response, $BART_{dec}(\cdot)$ denotes the BART decoder, ; denotes the token boundaries.

Cross Entropy Loss. We guide the decoder with the ground-truth response $Y = u_n$ by computing the Cross-Entropy Loss:

$$L_{CE} = -\frac{1}{|Y|}\sum_{t=1}^{|Y|} log(P(G_t = Y_t)) \tag{14}$$

where G_t denotes the generated token at the decoding time step t, while Y_t is the t-th token of the ground-truth response. In summary, the final loss is defined by:

$$L_{total} = L_{CE} + \lambda \cdot L_{BCE} \tag{15}$$

where λ denotes the coefficients of the BCE loss.

4 Experiments

4.1 Dataset

To verify our model, two requirements should be met in the datasets: (1) each utterance is annotated with related knowledge triples, and (2) containing abundant utterances in each dialogue. Therefore, we conduct our experiments on KdConv [39], a Chinese multi-domain knowledge-driven dialogue dataset, which contains 4.5K dialogues together with 86K utterances from three domains (i.e., film, music, and travel). In KdConv, each dialogue contains 19.0 turns as well as 10.1 triplets on average. For domain-specific knowledge, both structured triplets and unstructured texts are provided. Specifically, the film, music, and travel domain knowledge contain 89K, 56K, and 10K triplets, together with 7.3K, 4.1K, and 1.1K descriptive sentences, respectively.

4.2 Settings

Baselines: We adopt both vanilla and knowledge-grounded (indicating by "+know") dialogue generation models as our baselines:

- **Seq2Seq** [24]: An encoder-decoder model augmented with attention mechanism [1].
- **Seq2Seq+know** [39] fuses the last hidden state of the encoder with the knowledge vector via the attention mechanism and feeds both of them into the Seq2Seq decoder.
- **HRED** [20]: A hierarchical recurrent encoder-decoder model which models utterances and context separately with different RNNs.
- **HRED+know** [39] fuses the context vector with the knowledge vector and treats the fused vector as the initial state of the HRED decoder.
- **BART** [10]: A pre-trained Transformer-based encoder-decoder model which achieves state-of-the-art performance on various text generation tasks.
- **BART+know** incorporates both knowledge entities and relations represented by the average word embeddings of the corresponding flat texts.
- **BART+know(TransR)** incorporates knowledge entities and relations represented by a knowledge graph embedding algorithm (i.e., TransR [15]).

Implementation: We implement the above models with PyTorch and Huggingface Transformers[2] libraries. In Seq2Seq and HRED baselines, we employ GRU architecture [5] as the encoder and the decoder with 200 hidden cells. In terms of word embeddings, we adapt Tencent AI Lab word embeddings of 200d[3]. When encoding context, all models treat the concatenation of the past $n-1$ utterances as the input of the encoder, while the target output of the decoder is the n-th utterance. n is set to 8 in our experiments suggested by KdConv [39]. All models are optimized with ADAM optimizer using an initial learning rate of 5e-5. The mini-batch size is set to 32.

For our `RT-KGD`, the embedding size of entities and relations is set to 200. The implementation of TransR is provided by *OpenKE*[4]. The knowledge encoder is Bi-GRU, the hidden size and the number of layers are set to 300 and 1, respectively. We choose the *Chinese BART*[5] as the baseline pre-training language model with the default hyper-parameter settings. When decoding the response, the beam search size of all models is set to 5. The λ is set to 1 in Eq. 15.

4.3 Evaluation Metrics

Automatic Evaluation: Following [39], we adopt perplexity (PPL), BLEU scores [17], and Distinct scores [11] as automatic metrics. In detail, PPL is used

[2] https://github.com/huggingface/transformers.

[3] https://ai.tencent.com/ailab/nlp/en/embedding.html.

[4] https://github.com/thunlp/OpenKE.

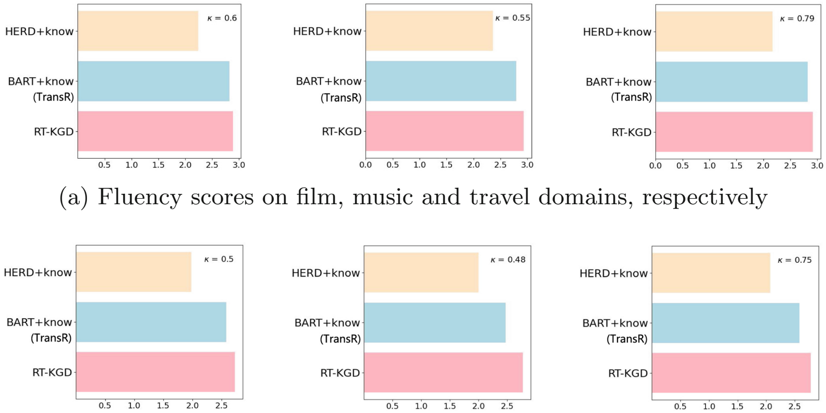
[5] https://huggingface.co/fnlp/bart-base-chinese.

Table 1. Automatic evaluation results on KdConv Corpus. The **bold** indicates the best performance. The "+know" means the models are enhanced by the knowledge base, and the knowledge words are encoded by word embeddings. ↑ indicates higher is better. ↓ indicates lower is better. † denotes the results reported by KdConv [39].

Model	PPL ↓	BLEU-1/2/3/4 ↑				Distinct-1/2/3/4 ↑			
Film									
Seq2Seq	23.88†	26.97†	14.31†	8.53†	5.30†	2.32†	6.13†	10.88†	16.14†
Seq2Seq+know	25.56†	27.45†	14.51†	8.66†	5.32†	2.85†	7.98†	15.09†	23.17†
HRED	24.74†	27.03†	14.07†	8.30†	5.07†	2.55†	7.35†	14.12†	21.86†
HRED+know	26.27†	27.94†	14.69†	8.73†	5.40†	2.86†	8.08†	15.81†	24.93†
BART	**2.66**	28.54	19.28	14.21	11.00	2.46	14.12	25.72	36.12
BART+know	2.85	29.38	20.18	15.02	11.74	2.55	15.26	28.01	39.45
BART+know(TransR)	2.82	29.68	20.43	15.26	11.97	2.50	15.12	27.96	39.56
RT-KGD(ours)	2.86	**32.11**	**22.21**	**16.68**	**13.18**	**3.05**	**16.34**	**31.36**	**44.68**
Music									
Seq2Seq	16.17†	28.89†	16.56†	10.63†	7.13†	2.52†	7.02†	12.69†	18.78†
Seq2Seq+know	17.12†	29.6†	17.26†	11.36†	7.84†	3.93†	12.35†	23.01†	34.23†
HRED	16.82†	29.92†	17.31†	11.17†	7.52†	2.71†	7.71†	14.07†	20.97†
HRED+know	17.69†	29.73†	17.51†	11.59†	8.04†	3.80†	11.70†	22.00†	33.37†
BART	2.46	31.65	23.04	18.22	15.05	2.80	13.69	24.73	34.59
BART+know	**2.40**	32.20	23.24	18.20	14.89	2.74	13.54	24.96	35.41
BART+know(TransR)	2.44	32.27	23.40	18.44	15.22	2.80	13.68	25.19	35.61
RT-KGD(ours)	2.47	**40.75**	**31.26**	**25.56**	**21.64**	**4.18**	**17.38**	**30.05**	**41.05**
Travel									
Seq2Seq	10.44†	29.61†	20.04†	14.91†	11.74†	3.75†	11.15†	19.01†	27.16†
Seq2Seq+know	10.62†	37.04†	27.28†	22.16†	18.94†	**4.25**†	13.64†	24.18†	34.08†
HRED	10.90†	30.92†	20.97†	15.61†	12.30†	4.15†	12.01†	20.52†	28.74†
HRED+know	11.15†	36.87†	26.68†	21.31†	17.96†	3.98†	13.31†	24.06†	34.35†
BART	1.83	34.77	29.11	25.69	23.33	2.70	13.39	21.92	29.53
BART+know	1.67	36.19	29.83	26.04	23.41	2.59	13.31	22.01	29.69
BART+know(TransR)	1.69	36.61	30.29	26.54	23.92	2.56	13.58	22.85	30.87
RT-KGD(ours)	**1.61**	**47.56**	**41.46**	**37.40**	**34.31**	3.58	**15.50**	**26.10**	**35.72**

to evaluate whether the generation result is grammatical and fluent. BLEU-n (n=1, 2, 3, or 4) estimates how many n-grams overlap between generated sentences and ground truth references. Distinct-n (n=1, 2, 3, or 4) evaluates the diversity of generated responses.

Human Evaluation: Considering the complexity of the knowledge-grounded dialogue generation task and the limitation of automatic evaluation, it is necessary to further conduct the human evaluation. Following KdConv [39], The criteria of human evaluation include two aspects: (1) Fluency evaluates whether the generated responses are reasonable and relevant to the given dialogue context. (2) Coherence measures how relevant the knowledge contained in the generated responses and the counterpart in the ground truth responses. We randomly select 100 dialogue contexts from KdConv in three domains, respectively, and then ask five well-educated evaluators to judge the generated responses by different models. The scoring adopts a 3-point scale.



(a) Fluency scores on film, music and travel domains, respectively

(b) Coherence scores on film, music and travel domains, respectively

Fig. 3. Human evaluation in three domains, including means and variances of the Fluency (a) and Coherence (b). $\boldsymbol{\kappa}$ is the Fleiss' kappa value.

4.4 Experimental Results

Table 1 shows the automatic evaluation results. We analyze the results from the following perspectives:

(1) Comparison between models: Compared with all baseline models, `RT-KGD` achieves the best results on most of automatic metrics in three domains, which indicates that our knowledge-guided method is extremely effective in improving the coherence and diversity of generated responses. Specifically, compared with Seq2Seq-based and HRED-based models, our `RT-KGD` obtains not only lower PPL scores but also higher BLEU-n and Distinct-n scores in three domains. This is because we utilize the pre-trained language model to encode contexts and generate responses, which makes use of the implicitly learned knowledge from the pre-trained corpus. On the other hand, compared with BART-based models, our `RT-KGD` works better in terms of BLEU-n and Distinct-n scores, however worse on PPL scores. Based on our manual sampling analysis of the experimental results, the reason might be that our `MHKT-Path` takes the knowledge transition into consideration. At the same time, diverse knowledge information may result in responses that have never appeared in the corpus, thus reducing the PPL scores.

Moreover, it can be seen that all models with knowledge perform better than those without knowledge in terms of BLEU-n and Distinct-n, indicating the benefits of incorporating knowledge. However, the addition of knowledge works worse in PPL. The reason may be that the sentence with knowledge is less common and more difficult to understand for the model. We also observe that all models with "know(TransR)" obtain higher BLEU-n and Distinct-n scores than models with "know", demonstrating that introducing of knowledge graph embedding algorithm has a positive influence on generating high-quality responses. It is worth noting that in the music domain, BART performs better than BART+know in terms of Distinct-1 and Distinct-2 but worse in Distinct-3 and Distinct-4, which is due to that BART prefers to use individual words with low frequency rather than common phrases. Furthermore, it is possible to get a high Distinct-1 for putting together a response with entirely random words. The same analysis comparing *BART* and *BART+know* also applies to the travel domain.

(2) Comparison between domains: As we can see, models in the travel domain perform better than that in film and music domains on PPL and BLEU-k, while models in the film domain obtain higher Distinct-n scores than the same model in music and travel domains. The reason might be that there are more entities and relations in the film domain, which leads to more diverse knowledge tokens but a lower similarity with the ground-truth.

4.5 Human Study

Here, we estimate three knowledge-grounded dialogue generation models which perform better than other baselines. The experiment results are shown in Fig. 3. As can be seen, RT-KGD outperforms other models significantly on both metrics in all three domains, which indicates that our model can generate more human-like responses. Moreover, the performance gap between models behaves differently on different metrics. The fluency scores in the music domain (the middle one in Fig. 3(a)) are increased from 1.36 (HRED+know) to 1.93 (RT-KGD), while the coherence scores in the music domain (the middle one in Fig. 3(b)) are increased from 1.00 (HRED+know) to 1.77 (RT-KGD). We also show Fleiss' Kappa values of our human study. A higher score indicates higher agreements among evaluators. The kappa scores demonstrate a good inter-agreement among our evaluators.

4.6 Ablation Study

To analyze which components are driving the improvements, we further design three graph variants for detailed comparison and ablation study: (1) "w/o tri" removes the triplet vertices in MHKT-Path; (2) "w/o rel" removes the relation vertices in MHKT-Path; (3) "w/o edge" removes the edges between the triplet and the relation vertices in MHKT-Path.

Table 2 shows the results of ablation studies. First, we observed that models suffer the performance drop when removing any of the components, demonstrating the effectiveness of integrating triplets and relations. Second, the degree

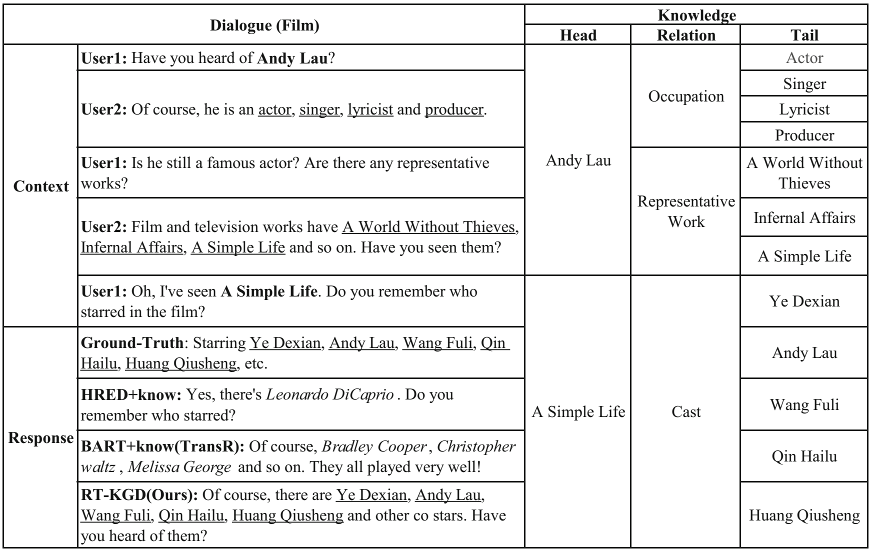
Table 2. Ablation study on KdConv. The **bold** and underline denote the best and the worst performances, respectively.

Model	PPL ↓	BLEU-1/2/3/4 ↑				Distinct-1/2/3/4 ↑			
Film									
RT-KGD(ours)	2.86	**32.11**	**22.21**	**16.68**	**13.18**	**3.05**	**16.34**	**31.36**	**44.68**
- w/o tri	**2.85**	30.17	20.82	15.58	12.22	2.61	15.79	29.28	41.16
- w/o rel	3.37	30.10	20.64	15.42	12.10	2.56	15.76	29.31	41.44
- w/o edge	3.35	30.13	20.76	15.52	12.22	2.53	15.79	29.42	41.68
Music									
RT-KGD(ours)	2.47	**40.75**	**31.26**	**25.56**	**21.64**	**4.18**	**17.38**	**30.05**	**41.05**
- w/o tri	2.43	32.22	23.24	18.22	14.94	2.74	13.17	24.26	34.42
- w/o rel	2.49	32.53	23.66	18.67	15.44	2.85	14.12	26.28	37.22
- w/o edge	**2.42**	32.28	23.44	18.50	15.26	2.83	13.92	25.36	35.55
Travel									
RT-KGD(ours)	**1.61**	**47.56**	**41.46**	**37.40**	**34.31**	**3.58**	**15.50**	**26.10**	**35.72**
- w/o tri	1.70	36.92	30.69	26.95	24.33	2.71	13.89	23.32	31.76
- w/o rel	1.84	36.98	30.59	26.74	24.06	2.64	13.63	23.01	31.17
- w/o edge	1.82	37.39	31.02	27.21	24.55	2.58	13.43	22.14	29.79

of impact increases from the film domain to the travel domain after removing components. For example, the BLEU-n scores decrease by 1.4, 7.4, and 10.4 on average in film, music, and travel, respectively, which shows that our `MHKT-Path` plays a more significant role in the travel domain in improving the quality of generated response. Third, the contribution of each component is not equal in different domains. Specifically, if the triplet vertices are removed, BLEU-n and Distinct-n scores are dramatically dropped in the music domain, indicating that turn-level entity information is capable of enhancing knowledge comprehension. While removing the relation vertices, BLEU-n scores declined most significantly in film and travel domains, demonstrating the advantage of explicitly modeling dialogue-level relation transition regularities. Lastly, without the edges between the triplet and relation vertices, the performance of `RT-KGD` in all three domains is reduced to varying degrees. This is because the edge between triplet vertices and relation vertices effectively propagates the information between these two vertices.

4.7 Case Study

As shown in Fig. 4, we show the responses generated by HRED+know, BART+know(TransR) and `RT-KGD`. We can observe that given the context and corresponding knowledge triplets, HRED+know tends to generate generic or irrelevant responses, and BART+know(TransR) can generate coherent and



Dialogue (Film)		Knowledge		
		Head	Relation	Tail
Context	**User1:** Have you heard of **Andy Lau**?	Andy Lau	Occupation	Actor
	User2: Of course, he is an actor, singer, lyricist and producer.			Singer
				Lyricist
				Producer
	User1: Is he still a famous actor? Are there any representative works?		Representative Work	A World Without Thieves
	User2: Film and television works have A World Without Thieves, Infernal Affairs, A Simple Life and so on. Have you seen them?			Infernal Affairs
				A Simple Life
	User1: Oh, I've seen **A Simple Life**. Do you remember who starred in the film?	A Simple Life	Cast	Ye Dexian
Response	**Ground-Truth**: Starring Ye Dexian, Andy Lau, Wang Fuli, Qin Hailu, Huang Qiusheng, etc.			Andy Lau
	HRED+know: Yes, there's *Leonardo DiCaprio* . Do you remember who starred?			Wang Fuli
	BART+know(TransR): Of course, *Bradley Cooper*, *Christopher waltz* , *Melissa George* and so on. They all played very well!			Qin Hailu
	RT-KGD(Ours): Of course, there are Ye Dexian, Andy Lau, Wang Fuli, Qin Hailu, Huang Qiusheng and other co stars. Have you heard of them?			Huang Qiusheng

Fig. 4. Example dialogue cases. The **bold** is the core entity under discussion. Underline is the appropriate knowledge used in the dialogue. *Italic* is inconsistent with the context.

informative responses but utilizes the inconsistent knowledge. While our RT-KGD is superior to generating high-quality responses with appropriate knowledge.

5 Conclusion

In this paper, we proposed a novel KGD model: Relation Transition aware Knowledge-Grounded Dialogue Generation (RT-KGD), which models the knowledge transition across multi-turn dialogue by integrating dialogue-level relation transition regularities with turn-level entity semantic information. Furthermore, our RT-KGD model utilizes the predicted knowledge to generate a response given the dialogue context. According to automatic and manual evaluation, our model generates high-quality responses which utilize more appropriate knowledge and are closer to the responses given by humans.

Supplemental Material Statement: The KdConv dataset and part of the baselines in Sect. 4 are publicly available from Github[6]. Source codes for RT-KGD model are available at https://github.com/tigerwww-git/RT-KGD.

Acknowledgments. Zhixu Li and Jianfeng Qu are the corresponding authors. This research is supported by the National Natural Science Foundation of China (Grant No. 62072323, 62102276), the Natural Science Foundation of Jiangsu Province (Grant No.

[6] https://github.com/thu-coai/KdConv.

BK20191420, BK20210705, BK20211307), the Natural Science Foundation of Educational Commission of Jiangsu Province, China (Grant No. 21KJD520005), the Major Program of the Natural Science Foundation of Jiangsu Higher Education Institutions of China (Grant No. 19KJA610002), NH33714722 Youth Team on Interdisciplinary Research Soochow University - Research on Subjectivity and Reasoning Theory in Artificial Intelligence, the Priority Academic Program Development of Jiangsu Higher Education Institutions, Suda-Toycloud Data Intelligence Joint Laboratory, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
2. Bai, J., Yang, Z., Liang, X., Wang, W., Li, Z.: Learning to copy coherent knowledge for response generation. In: AAAI (2021)
3. Bao, S., He, H., Wang, F., Wu, H.: Plato: pre-trained dialogue generation model with discrete latent variable. In: ACL (2020)
4. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS (2013)
5. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP (2014)
6. Ghazvininejad, M., et al.: A knowledge-grounded neural conversation model. In: AAAI (2018)
7. Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: Proceedings of the Web Conference 2020 (2020)
8. Jung, J., Son, B., Lyu, S.: Attnio: knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue. In: EMNLP (2020)
9. Kim, B., Ahn, J.H., Kim, G.: Sequential latent knowledge selection for knowledge-grounded dialogue. In: ICLR (2020)
10. Lewis, M., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL (2020)
11. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B.: A diversity-promoting objective function for neural conversation models. In: NAACL (2016)
12. Lian, R., Xie, M., Wang, F., Peng, J., Wu, H.: Learning to select knowledge for response generation in dialog systems. In: IJCAI (2019)
13. Liang, Y., Meng, F., Zhang, Y., Chen, Y., Xu, J., Zhou, J.: Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In: AAAI, no. 15 (2021)
14. Lin, X.V., Jian, W., He, J., Wang, T., Chu, W.: Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In: ACL (2020)
15. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: AAAI (2015)
16. Moon, S., Shah, P., Kumar, A., Subba, R.: Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In: ACL (2019)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019)

19. Roller, S., et al.: Recipes for building an open-domain chatbot. In: EACL (2021)
20. Serban, I., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI (2016)
21. Sordoni, A., et al.: A neural network approach to context-sensitive generation of conversational responses. In: NAACL (2015)
22. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: AAAI (2017)
23. Sukhbaatar, S., Szlam, A.D., Weston, J., Fergus, R.: End-to-end memory networks. In: NIPS (2015)
24. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)
25. Vaswani, A., et al.: Attention is all you need. In: NIPS (2017)
26. Vougiouklis, P., Hare, J.S., Simperl, E.P.B.: A neural network approach for knowledge-driven response generation. In: COLING (2016)
27. Wang, J., et al.: Knowledge enhanced sports game summarization. In: WSDM (2022)
28. Wang, J., et al.: Incorporating commonsense knowledge into story ending generation via heterogeneous graph networks. In: Database Systems for Advanced Applications (2022)
29. Wu, S., Li, Y., Zhang, D., Zhou, Y., Wu, Z.: Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In: ACL (2020)
30. Wu, W., et al.: Proactive human-machine conversation with explicit conversation goal. In: ACL (2019)
31. Xing, C., Wu, W.Y., Wu, Y., Zhou, M., Huang, Y., Ma, W.Y.: Hierarchical recurrent attention network for response generation. In: AAAI (2018)
32. Zhang, H., Liu, Z., Xiong, C., Liu, Z.: Grounded conversation generation as guided traverses in commonsense knowledge graphs. In: ACL (2020)
33. Zhang, R., Guo, J., Fan, Y., Lan, Y., Xu, J., Cheng, X.: Learning to control the specificity in neural response generation. In: ACL (2018)
34. Zhang, T., et al.: Aligning internal regularity and external influence of multi-granularity for temporal knowledge graph embedding. In: Database Systems for Advanced Applications (2022)
35. Zhang, Y., et al.: Dialogpt: large-scale generative pre-training for conversational response generation. In: ACL (2020)
36. Zhao, X., Wu, W., Xu, C., Tao, C., Zhao, D., Yan, R.: Knowledge-grounded dialogue generation with pre-trained language models. In: EMNLP (2020)
37. Zheng, D., Xu, Z., Meng, F., Wang, X., Wang, J., Zhou, J.: Enhancing visual dialog questioner with entity-based strategy learning and augmented guesser. In: Findings of EMNLP 2021 (2021)
38. Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X.: Commonsense knowledge aware conversation generation with graph attention. In: IJCAI (2018)
39. Zhou, H., Zheng, C., Huang, K., Huang, M., Zhu, X.: KdConv: a Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In: ACL (2020)