Select, Extend, and Generate: Generative Knowledge Selection for Open-Domain Dialogue Response Generation

Sixing Wu[1], Ping Xue[2], Ye Tao[2], Ying Li[2,3](✉), and Zhonghai Wu[2,3]

[1] National Pilot School of Software, Yunnan University, Kunming, China
wusixing@ynu.edu.cn
[2] School of Software and Microelectronics, Peking University, Beijing, China
[3] National Research Center of Software Engineering, Peking University, Beijing, China
li.ying@pku.edu.cn

Abstract. Incorporating external commonsense knowledge can enhance machines' cognition and facilitate informative dialogues. However, current commonsense knowledge-grounded dialogue generation works can only select knowledge from a finite set of candidates retrieved by information retrieval (IR) tools. This paradigm suffers from: 1) The knowledge candidate space is limited because IR tools can only retrieve existing knowledge from the given knowledge base, and the model can only use the retrieved knowledge; 2) The knowledge selection procedure lacks enough interpretability to explain the selected result. Moreover, with the increasing popularity of pre-trained language models (PLMs), many knowledge selection methods of non-PLM models have become incapable because of the input/structure restrictions of PLMs. To this end, we propose a simple but elegant *SEG-CKRG*, and introduce a novel PLM-friendly *Generative Knowledge Selection (GenSel)* to select knowledge via a generative procedure. Besides selecting the knowledge facts from the retrieved candidate set, *GenSel* can also generate newly extended knowledge. *GenSel* also improves interpretability because the output of the knowledge selection is a natural language text. Finally, *SEG-CKRG* uses *GPT-2* as the backbone language model. Extensive experiments and analyses on a Chinese dataset have verified the superior performance of *SEG-CKRG*.

Keywords: dialogue generation · knowledge-grounded

1 Introduction

Open-domain dialogue response generation (RG) models enable machines to converse with humans using natural language and play an important role in human-computer interaction [43]. However, machines lack enough real-world knowledge cognition because they can only access the parametric knowledge of a model besides the dialogue history [45]. Thus, machines struggle to thoroughly understand the semantics of dialogue histories and generate informative responses.

X. Wang et al. (Eds.): DASFAA 2023, LNCS 13945, pp. 649–664, 2023.
https://doi.org/10.1007/978-3-031-30675-4_48

Seeking information from external knowledge sources is an effective solution [50], i.e., knowledge-grounded dialogue response generation (KRG) [4,17].

Compared to RG models, the superiority of KRG models derives from the ability to use external knowledge [42]. The general paradigm of KRG can be summarized as three stages [7,39]: 1) *Knowledge-Retrieval stage:* it first employs an efficient Information Retrieval (IR) tool to retrieve a set of knowledge candidates in a coarse-grained way. The retrieved knowledge candidates contain much irrelevant information because IR tools only consider the literal feature; 2) *Knowledge-Selection stage:* To filter out irrelevant information and select contextually-relevant knowledge, KRG also has a knowledge selection stage using more fine-grained methods; 3) *Response Generation stage:* it finally generates the target response by accessing the dialogue history and selected knowledge. Among such three stages, the second knowledge selection stage plays the most crucial role in the research of KRG and has received much attention [10,23,28].

This paper focuses on commonsense knowledge-grounded dialogue response generation (CKRG). Despite many successes [42,46], CKRG still suffers from several challenges, especially in the era of pre-trained language models (PLMs) [14,19]. First, the knowledge candidate space (i.e., the knowledge can be selected and used when generating the response) is fixed and limited. On the one hand, IR tools can only retrieve knowledge candidates already existing in the knowledge base. On the other hand, the model can only use the knowledge candidate already retrieved by IR tools. This may lead to insufficient knowledge coverage [42]. Second, in the knowledge selection stage, previous CKRG works [46,50] often use deep but complex networks, which lack enough interpretability to explain the knowledge selection procedure. For example, it is hard to determine which knowledge facts have been selected. Finally, although PLMs are powerful, they also bring many thorny restrictions to the downstream applications [15], such as the length (most PLMs can only operate at most 512/1024 tokens), the input format (must be plain text), the network structure, and so on. Consequently, many knowledge selection methods originally proposed for non-PLM-based models have become incapable in the era of PLMs; then, knowledge selection can only rely on the external network or the implicit self-attention mechanism [23,49].

Considering these challenges, we propose *SEG-CKRG*, a simple but elegant CKRG model. As shown in Fig. 1, *SEG-CKRG* introduces a novel *Generative Knowledge Selection (GenSel)* mechanism, which regards knowledge selection as a generative problem. *GenSel* uses a PLM to explicitly generate contextually-relevant knowledge based on the dialogue history and the knowledge candidate set retrieved by IR tools. By regarding this task as a generative problem, *GenSel* can not only select knowledge from the candidate set retrieved by IR tools, but can also extend the knowledge by externalizing the inherent knowledge of PLMs. Then, *SEG-CKRG* generates the target response conditioned on both the generated knowledge and the retrieved knowledge. Considering both the generative knowledge selection procedure and the dialogue generation procedure are generative problems, we can train/infer *SEG-CKRG* in an end-to-end fashion. We pre-

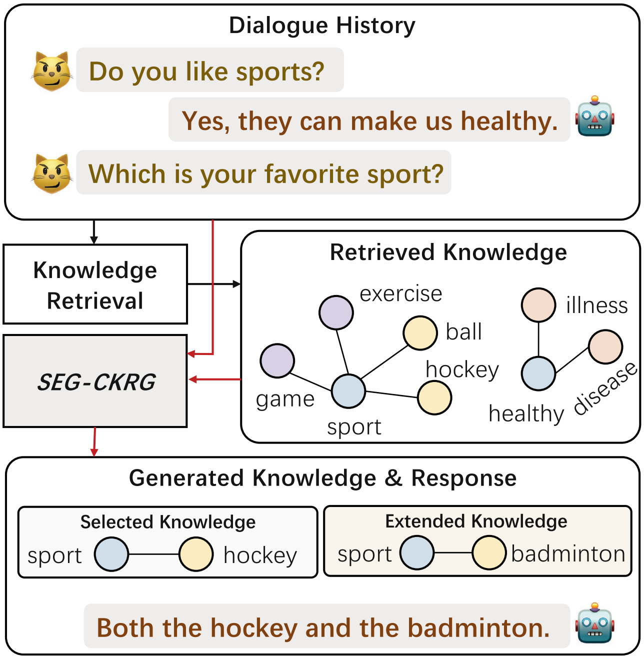

Fig. 1. An example. *SEG-CKRG* can use *Generative Knowledge Selection* to select the existing knowledge and extend the new knowledge, then generates the response.

train two GPT-2 models [27] as the backbone PLMs. To boost the knowledge representation density and the infusing of two generative procedures, we propose an *Efficient Input Representation* technique and a *Dual-Head Generator* technique, respectively.

We conduct extensive experiments on a Chinese conversational dataset *Weibo-ConceptNet* [41], whose dialogues have been aligned to a commonsense knowledge base, ConceptNet [32]. Experimental results have verified that *SEG-CKRG* has significantly outperformed previous state-of-the-art models, and *GenSel* can not only accurately select the knowledge but also generate new contextually-relevant knowledge. We also bring extensive analyses to investigate our approach further.

2 Methodology

2.1 Preliminary

Response Generation (RG). Suppose $\mathcal{D} = \{(H_i, R_i)\}^N$ is a conversational corpus, where $H_i = (h_1, \cdots, h_{|H_i|})$ is the dialogue history, $R_i = (r_1, \cdots, r_{|R_i|})$ is the response. Then, RG learns a conditional language model $P_{RG}(R_i|H_i)$ to generate R_i conditioned on H_i: $P_{RG}(R_i|H_i) = \prod P_{RG}(r_t|r_{<t}, H_i)$.

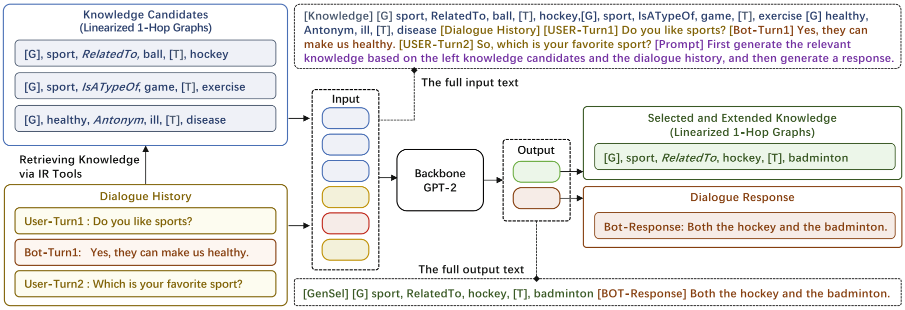
Knowledge Candidates
(Linearized 1-Hop Graphs)
[G], sport, *RelatedTo*, ball, [T], hockey
[G], sport, *IsATypeOf*, game, [T], exercise
[G], healthy, *Antonym*, ill, [T], disease
Retrieving Knowledge via IR Tools
Dialogue History
User-Turn1 : Do you like sports?
Bot-Turn1: Yes, they can make us healthy.
User-Turn2 : Which is your favorite sport?
[Knowledge] [G] sport, RelatedTo, ball, [T], hockey,[G], sport, IsATypeOf, game, [T], exercise [G] healthy, Antonym, ill, [T], disease [Dialogue History] [USER-Turn1] Do you like sports? [Bot-Turn1] Yes, they can make us healthy. [USER-Turn2] So, which is your favorite sport? [Prompt] First generate the relevant knowledge based on the left knowledge candidates and the dialogue history, and then generate a response.
The full input text
Input
Backbone GPT-2
Output
Selected and Extended Knowledge
(Linearized 1-Hop Graphs)
[G], sport, *RelatedTo*, hockey, [T], badminton
Dialogue Response
Bot-Response: Both the hockey and the badminton.
The full output text
[GenSel] [G] sport, RelatedTo, hockey, [T], badminton [BOT-Response] Both the hockey and the badminton.

Fig. 2. An overview of *SEG-CKRG*. We show the input/output examples. In this example, *SEG-CKRG* has selected a knowledge fact '(sport, RelatedTo, Hockey)' and extended a '(sport, RelatedTo, badminton)' in the example.

Knowledge-Grounded Response Generation (KRG). RG models tend to generate generic responses such as 'I don't know.' [13] because P_{RG} can only use the insufficient knowledge hidden in the parameters θ_{RG} and the dialogue history. To address this issue, KRG methods try to seek more knowledge from the external knowledge base, such as encyclopedic knowledge [7], commonsense knowledge [32], and so on [26].

More specifically, in the commonsense knowledge-grounded dialogue response generation (CKRG) scenario, there is a knowledge base $\mathcal{K} = \{k_i = (e_i^h, e_i^r, e_i^t)\}^M$, where k_i is a commonsense fact triplet, e_i^h, e_i^r, and e_i^t are the corresponding head entity, relation, and tail entity, respectively. Then, for each dialogue history H_i, we need to employ an IR tool to retrieve a set of commonsense facts $K_i = \{k_{i,j}\}^L, L << M$ form $\mathcal{K}$. Finally, the problem of CKRG is given by:

$$P_{CKRG}(R_i|H_i) = \prod P_{CKRG}(r_t|r_{<t}, H_i, K_i) \tag{1}$$

where P_{CKRG} is a conditional language model with the ability to access the knowledge K_i. Although K_i is the filtered results via IR tools, IR tools can only consider the token-level literal feature. Thus, a more fine-grained context-aware knowledge selection procedure is needed in P_{CKRG}. In non-PLM CKRG works, this procedure can be explicitly modeled and then integrated into P_{CKRG}. For example, [50] employs graph attention network [34]. In the era of PLM, limited by the input format and network structure, this procedure can only be implicitly performed by the integrated self-attention mechanism or external tools, bringing less interpretability but more limitations to the knowledge selection procedure.

2.2 Problem Definition and Overview

As shown in Fig. 2, unlike previous CKRG works, *SEG-CKRG* introduces a novel *Generative Knowledge Selection (GenSel)* mechanism, which regards knowledge selection as a generative problem. The objective of *SEG-CKRG* is:

$$P_{GenSel}(K_i^G|H_i, K_i) \cdot P_{ResGen}(R_i|H_i, K_i, K_i^G) \tag{2}$$

where $P_{GenSel}(K_i^G|H_i, K_i)$ first generates the contextually-relevant knowledge K_i^G conditioned both the dialogue history H_i and the retrieved knowledge K_i; subsequently, $P_{ResGen}(R_i|H_i, K_i, K_i^G)$ generates the target response R_i.

2.3 Generative Knowledge Selection

SEG-CKRG uses a generative method to explicitly select and extend knowledge. Similar to other generation tasks, it is a conditional language modeling problem:

$$P_{GenSel}(K_i^G|H_i, K_i) = \prod P_{GenSel}(k_t^G|k_{<t}^G, H_i, K_i) \tag{3}$$

Efficient Input Representation. Most PLMs can only accept plain texts as input, which means the structural commonsense knowledge must be linearized to plain text. Thus, the input of P_{GenSel} is given by:

$$S_i = [\omega_K(K_i), \omega_H(H_i), \mathit{Prompting}] \tag{4}$$

where $\omega_H(H_i)$ linearizes the dialogue history with role (human/bot) labels and turn identifiers; *Prompting* is a prompting text[1] [52] to hint the PLM about the following generation action; $\omega_K(K_i)$ linearizes the structural $K_i = \{k_{i,j} = (e_{i,j}^h, e_{i,i}^r, e_{i,i}^t)\}^M$ to a sequence. To reduce the loss of structural information and improve the representation density, $\omega(K_i)$ uses a graph-level pattern:

$$\begin{aligned} \omega(K_i) &= (\omega_G(g_{i,1}); \omega_G(g_{i,2}); \cdots; \omega_G(g_{i,j}); \cdots) \\ \omega_G(g_{i,j}) &= ([G], e_{i,j}^{hg}, e_{i,j}^{rg}, e_{i,j,1}^{tg}, [T], e_{i,j,2}^{tg}, \cdots) \end{aligned} \tag{5}$$

where K_i is first compressed as a set of 1-hop graphs $G_i = \{g_{i,j} = e_{i,j}^{hg}, e_{i,j}^{rg}, \{e_{i,j}^{tg}\}\}$; namely, $\forall k \in K_i$ that have the same head entity $e_{i,j}^{hg}$ and the same relation $e_{i,j}^{rg}$ are placed to the corresponding 1-hop graph $g_{i,j}$; then, $g_{i,j}$ is sequentially linearized and concatenated with a graph separator $[G]$ and a tail entity separator $[T]$. Compared to previous triplet-level patterns [42,52], our graph-level pattern can reduce the length of the linearized knowledge and achieve higher representation density. Higher representation density means more knowledge facts can be included under the same length limitation.

Generation. The goal is to generate the linearized contextually-relevant knowledge sequence $\omega_K(K_i^G)$. We adopt a widely-used auto-regressive GPT-2 [27] to implement $P_{GenSel}(K_i^G|H_i, K_i)$ and generate the $\omega_K(K_i^G)$:

$$\omega_K(K_i^G) = GPT2(S_i) = GPT2([\omega_K(K_i), \omega_H(H_i), \mathit{Prompting}]) \tag{6}$$

In the training stage, we use a weakly-supervised way [51,52] to construct the generation goal K_i^G. Given a knowledge candidate set K_i retrieved by IR

[1] The translated text is 'First generate the relevant knowledge based on the left knowledge candidates and the dialogue history, and then generate a response.'.

tools, if there is a knowledge candidate $k \in K_i$ whose head entity and tail entity appear in the dialogue history H_i and the dialogue response R_i, respectively; then this k is added to the target K_i^G.

During the generation, K_i^G is fully generated based on the K_i and H_i. Intuitively, the generated K_i^G can select the relevant knowledge from K_i. Besides, as a generative model, GPT2 can also extend to generate the relevant knowledge that is not included in the K_i^G, which is an inherent feature of generative language models [9]. Meanwhile, the generated $\omega_K(K_i^G)$ is a natural language text, which can explicitly explain the results of knowledge selection and extension.

2.4 Dialogue Response Generation

Finally, we use the same GPT2 to generate the dialogue response R_i based on the dialogue history H_i, the retrieved knowledge K_i, and the generated contextually-relevant knowledge K_i^G. We feed the input S_i^{DG} to the GPT2, estimate $P_{ResGen}(r_t|R_{<t}, H_i, K_i, K_i^G)$, and then generate the R_i:

$$\begin{aligned} S_i^{DG} &= [\omega_K(K_i), \omega_H(H_i), Prompting, \omega_K(K_i^G)] \\ R_i &= GPT2(S_i^{DG}) = GPT2([\omega_K(K_i), \omega_H(H_i), Prompting, \omega_K(K_i^G)]) \end{aligned} \tag{7}$$

where the generation head $\mathbf{W^R}$ is newly introduced compared to Eq. 6. This is because two generative procedures have different generation spaces, two separate generation heads help avoid confusion. Such a two-head generation mechanism is called *Dual-Head Generator*.

2.5 Training

Two generative procedures can be jointly trained in an end-to-end fashion by sharing the same GPT-2. We have pre-trained two different GPT-2 models and our *SEG-CKRG* on two Nvidia RTX-3090 GPUs:

General GPT2: The general-purpose or dialogue-oriented *base size*[2] Chinese GPT-2 resources are not very abundant [31]. Consequently, we first pre-train a Chinese GPT2 for our experiments. We implement a *base size* GPT2 language model network using the Huggingface transformer library[3] and PyTorch. There are 12 layers of 768-dimensional (for both the hidden states and embeddings) and 12-head Transformer layers. The vocabulary includes 30,000 subwords and 200 special symbols (placeholders). For efficiency, the maximum input length is limited to 512 tokens. This GPT-2 is first pre-trained on massive Chinese unsupervised data, including massive open-released news, movie/product comments, and Wikipedia data. In total, there are 18.4M sessions and 5.22B tokens. During the training, the batch size is 512, the number of total training steps is 80,000, and the optimizer is AdamW. After 4,000 warm-up steps, the learning rate will reach 2e−4; then, the learning rate will linearly decay to 0.

[2] a *base size* PLM models always has about 100M parameters.
[3] https://huggingface.co/.

Dialogue-Oriented GPT2: We also fine-tune a dialogue-oriented GPT-2. We use the Chinese conversational pre-training corpus *LCCC-large* released by [37], which includes 7.2M/4.7M sessions of single/multi-turn dialogues and 380M tokens in total. This GPT-2 is initialized from our general GPT-2, the batch size is 512, the number of total training steps is 180,000, and the optimizer is AdamW. It has the same learning rate strategy as GPT-2, except for the highest learning rate is decreased to 1.5e−4.

SEG-CKRG: Finally, *SEG-CKRG* is fine-tuned on the general GPT2 (by default) or the dialogue-oriented GPT2 (in ablation study). The batch size is set to 32, the maximum training epoch is set to 15. The best epoch on the validation set is adopted in the following test stage.

3 Experiment

3.1 Settings

Dataset. We test models on a Chinese dataset *Weibo-ConceptNet* [41], which has been aligned to a well-known commonsense knowledge base ConceptNet [32]. The training/validation/test set includes 102K/5.6K/5.6K single-turn dialogues. Each utterance has 10.3 words on average. The commonsense graph has 696,466 facts, 27,189 entities, and 26 relations. On average, each dialogue has 77.7 candidate facts that are retrieved from ConceptNet.

Comparison Models. We first selected several non-PLM baselines: *1) Seq2Seq*: an attentive Seq2Seq RG Model [3,24]; *2) PGN*: Seq2Seq + Pointer-Genetor copy network [29]; *3) ConKADI*: a KRG model with the felicitous knowledge selection mechanism [41]; *4) GOKC*: a KRG model with a novel knowledge copy mechanism [1]. We also selected several fine-tuned *base*-size PLM methods: *5) BERT2Seq, 6) BERT-PGN*: We changed the encoder of *Seq2Seq* and *PGN* to the *'hfl/chinese-bert-wwm-ext'* [5] BERT encoder [6]. *7) CDial-GPT2*: An open-released conversational GPT-2 RG models [37]. We select the GPT-2 configuration *'GPT2LCCC-base'*. *8) MHKD-GPT2*: A PLM-based KRG models [42], which is based on *CDial-GPT2*.

Implementation. We use the official codes for *ConKADI*, *GOKC*, *CDial-GPT2*, and *MHKD-GPT2*, and we re-implement the remaining models using PyTorch. For non-PLM models, we use a 2-layer 768d bi-GRU/LSTM[4] encoder, 2-layer 768d GRU/LSTM decoder, Adam optimizer, 1e-4 learning rate. For all baselines, we use 32 batch size, up to 20 epochs, and finally select the best model on the validation set. Due to the different requirements, *BERT2Seq, BERT-PGN* use the corresponding BERT tokenizer and vocab, *Seq2Seq, PGN, GOKC*, and

[4] our codes use GRU, the others keep the original setting.

ConKADI use the original tokenizer and vocab, *CDial-GPT2* and *MHKD-GPT2* use *CDial-GPT2*'s tokenizer and vocab. The implementation of our SEG-CKRG will be released at https://github.com/pku-sixing/DASFAA23_GenSel.

Automatic Evaluation Metrics. Different models use different tokenizers; thus, we conduct character-level evaluations to avoid such differences. We use the following automatic metrics: *1) F1*: it is the F-measure of character-overlapping relevance [1]; *2) BLEU-4*: it is the 4-gram BLEU to evaluate the precision-oriented relevance [25]; *3) ROUGE*: we use ROUGE-L to evaluate the recall-oriented relevance [18]; *4) EM-A/G/X*: we use embedding evaluate the semantic relevance, the embedding is computed using Average/Greedy/Extrema [21]; *5) DI-1/2*: we use Distinct-1/2 to evaluate the diversity [12]; *6) Ent*: we use 4-gram entropy to evaluate the informativeness [30]; *7) Mean*: following [42], we compute the geometric mean of all previous scores to evaluate the overall performance.

3.2 Automatic Evaluation

Table 1. Automatic Evaluation Results. **First**/Second denotes the first/second best.

Model	F1	ROUGE	BLEU-4	EM-A	EM-G	EM-X	DI-1	DI-2	Ent	Mean
Seq2Seq	16.20	12.40	1.09	0.869	0.677	0.649	0.32	3.22	8.61	2.09
PGN	16.56	12.65	1.23	0.872	0.676	0.651	**0.58**	8.13	9.55	2.54
GOKC	18.13	14.95	1.47	0.881	0.684	**0.695**	0.35	7.95	10.35	2.56
ConKADI	19.20	14.60	1.94	0.885	0.679	0.664	0.38	**11.22**	**12.04**	2.81
BERT2Seq	17.49	13.21	1.93	0.877	0.670	0.658	0.26	2.77	8.83	2.18
BERT-PGN	18.76	13.72	2.52	0.892	0.674	0.664	0.36	6.91	9.69	2.64
CDial-GPT2	14.79	12.31	1.61	0.866	0.675	0.653	0.26	3.69	8.47	2.13
MHKD-GPT2	18.77	16.60	2.45	0.874	0.690	0.667	0.28	4.13	9.43	2.46
SEG-CKRG	**21.02**	**17.15**	**3.11**	**0.896**	**0.708**	0.689	0.48	9.94	11.13	**3.09**

As reported in Table 1, *SEG-CKRG* has achieved tier-1 results (the first and the second best) in all metrics and significantly outperformed previous methods in the *Mean* score, demonstrating the best overall performance and effectiveness. In addition, rather than pursuing the best score on a single-dimensional metric or only using some handpicked metrics, the philosophy of *SEG-CKRG* is multi-dimensional because a single automatic metric is not reliable [21].

Relevance: In the three overlapping-based metrics (i.e., F1, BLEU-4, and ROUGE), *SEG-CKRG* has the best results because our approach can simultaneously seek information from both the pre-trained language model and the external knowledge source to help the dialogue generation. In another three embedding-based relevance metrics (i.e., EMB-A/G/X), *SEG-CKRG* also has the best overall performance, showing the dialogue responses generated by our approach are more semantically relevant to the ground truth. Besides, we can also find that PLM-based models have better relevance performance than non-PLM-based models in the mass. Indicating the necessity of using PLMs in CKRG.

Table 2. Human Annotation Results. **Scores** denotes *SEG-CKRG* is significantly better (sign-test, p-value < 0.005). The 2/3 agreement ratio (at least 2 judges gave the same) is 95.4%, the 3/3 ratio is 54.8.2%.

%	Fluency			Rationality			Informativeness		
Compare to	*Lose*	Tie	*Win*	*Lose*	Tie	*Win*	*Lose*	Tie	*Win*
Seq2Seq	36.7	22.3	**41.0**	31.0	11.7	**57.3**	31.3	7.0	**61.7**
GOKC	8.3	6.0	**85.7**	9.0	7.0	**84.0**	15.0	4.0	**81.0**
ConKADI	11.0	5.3	**83.7**	25.0	3.0	**72.0**	38.6	1.4	**60.0**
BERT-PGN	36.7	6.6	**56.7**	42.3	2.3	**55.4**	45.6	2.4	**52.0**
CDial-GPT	20.6	9.4	**60.0**	38.6	5.7	**55.7**	41.0	3.0	**56.0**
MHKD-GPT	27.6	14.7	**57.7**	38.3	2.7	**59.0**	40.6	3.4	**56.0**
Human	33.6	31.4	**35.0**	46.3	14.0	39.7	62.3	8.7	29.0

Diversity and Informativeness: The situation is different in this part. *ConKADI* and our *SEG-CKRG* notably surpass other models. Between such two models, *SEG-CKRG* is slightly lower than *ConKADI*, and the reason can be summarized as 1) *SEG-CKRG* does not sacrifice the relevance to improving diversity and informativeness; 2) *SEG-CKRG* does not use any copy mechanism. Copy mechanism can copy words from the dialogue history or the external knowledge directly, which can significantly boost diversity and informativeness in the automatic evaluation. For example, compared with *Seq2Seq/BERT2Seq*, the copy variant *PGN/BERT-PGN* has more notable improvements in such metrics. However, we find previous copy works tend to repeat the given query rather than extend the new information, and then we decide not to equip this mechanism.

3.3 Human Evaluation

We employed three well-educated native-speaker to evaluate the practical generation quality of *SEG-CKRG*. The criteria include three dimensions: *1) Fluency*: is this response grammatically correct and fluent? *2) Rationality:* does this response logically conform to the current dialogue context? *3) Informativeness:* can this response provide enough meaningful information?

As reported in Table 2, we sampled 100 comparison cases[5] and compared *SEG-CKRG* with the three best baselines in the automatic evaluation (ConKADI, BERT-PGN, and GOKC) and the naive Seq2Seq. We have several findings: 1) Although Seq2Seq is the naive baseline, the comparison result is not the worst, especially in terms of fluency. This is because the task and the network of Seq2Seq are simple but stable; 2) Compared to GOKC and ConKADI, *SEG-CKRG* has notable advantages, indicating the importance of introducing the PLMs to CKRG; 3) Compared to BERT-PGN, *SEG-CKRG* is still better, demonstrating the effectiveness of using external knowledge. Finally, we also

[5] 5*100 pair-wise comparisons in total.

Table 3. Generated knowledge types. # is the average counting per response.

# Original	#Actual	#Generated	#Selected	#Extend
77.7	52.5	1.282	1.157	0.124

Table 4. Ablation Study.

#	Setting	ROUGE	EMBED-X	DIST2	Mean
0	*Full*	17.15	0.689	9.94	3.09
Different Backbones					
1	*DialogueGPT2*	16.72	0.686	9.69	3.05
2	*FromScratch*	15.64	0.682	5.12	2.56
Different Knowledge Accessing					
3	*w/o GenSelKnow*	16.52	0.690	8.11	2.95
4	*w/o SelKonw*	12.24	0.659	9.70	2.66
5	*w/o All (general GPT2)*	12.44	0.665	8.03	2.56
6	*w/o All (dialogue GPT2)*	12.62	0.665	7.76	2.59

compare *SEG-CKRG* with the human-generated ground-truth response. *SEG-CKRG* is comparable to the human in terms of fluency. However, *SEG-CKRG* is still behind the human in terms of rationality and informativeness. This shows we still have a large room to improve CKRG in future works.

3.4 More Analyses

Generated Knowledge. As reported in Table 3, we counted the types among the knowledge generated by our *Generative Knowledge Selection (GenSel)*. The original dataset has 77.7 facts per dialogue, but our model could only use 52.5 randomly sampled facts because of the restriction of the backbone GPT-2. We can find *GenSel* generated 1.282 facts per response on average, and about 9.7% of them are extended knowledge. It indicates 1) *GenSel* can indeed extend the knowledge and 2) using IR tools to explicitly retrieve knowledge candidates is still necessary.

Backbone GPT2. We have pre-trained two GPT-2 models; the first general GPT-2 is pre-trained on the general unsupervised data, and the second dialogue-oriented GPT-2 is further fine-tuned on a large-scale conversational dataset [37]. As shown in Table 4, if we train a *SEG-CKRG* based on the second dialogue-oriented GPT-2 (#1), although the perplexity is lower (9.93 vs 11.04), the real performance is even worse than the *SEG-CKRG* trained on the general GPT-2 (#0). We think our GenSel prefers the general backbone. Meanwhile, we also train a *SEG-CKRG* without initializing from a pre-trained GPT-2 (#2). The results are worse in every dimension, indicating the necessity of pre-training.

Table 5. Case Study. For *SEG-CKRG*, we report the generated knowledge (*K:*) and response (*R:*). We use red/blue color to highlight the selected/extended knowledge.

Query #1	水果茶好喝吗? Does fruit tea taste good?
Seq2Seq	还可以，我觉得挺好喝的。Not bad, I think it's good.
ConKADI	还行吧我觉得喝茶水果的最好吃的。 Not bad, I think drinking tea and fruit the best foods.
BERT-PGN	我觉得挺好喝的就是水果茶比较好喝。 I think it's very good, that is the fruit tea is good.
SEG-CKRG	*K:* [G]水果 HasProperty 酸 [G] 樱桃 InvIsA 水果 [G] fruit, HasProperty,sour [G] fruit, InvIsA, cherry *R:* 还不错,酸酸甜甜的,酸樱桃味的很好喝。 Not bad. It's sour and sweet. The sour cherry taste is good.
Query #2	为啥我不喜欢用拍照? Why don't I like taking pictures?
Seq2Seq	因为你好看啊。Because you look beautiful.
ConKADI	因为你不喜欢拍照啊，我喜欢手机。 Because you don't like taking pictures. I like mobile phones.
BERT-PGN	因为手机拍的太好看了。 Because the pictures taken by the phone are so beautiful.
SEG-CKRG	*K:*[G]拍照 InvCapableOf 手机 [G]taking pictures, InvCapableOf, phone *R:* 我也不喜欢,手机拍出来的效果也不好。 I don't like it either. The pictures taken by the phone are bad.

Knowledge Accessing. As shown in Table 4, we first remove the usage of knowledge generated by *Generative Knowledge Selection (GenSel)* (#3), and the performance is worse. It verifies *GenSel* can indeed help *SEG-CKRG* to select the existing knowledge and to extend to generate new relevant knowledge. Subsequently, similar to [52], we remove the usage of retrieved knowledge and try to externalize the implicit knowledge inherent in the backbone PLM (#4). It can be seen that the performance of (#4) is significantly worse than the original model. Thus, we think it is still necessary to explicitly access external knowledge even in the era of PLMs. Finally, we discard all knowledge usages and purely fine-tune two RG models based on our general GPT-2 (#5) and dialogue-oriented GPT-2 (#6). Both (#5) and (#6) are worse than previous model variants, showing the value of modeling dialogue response with external knowledge. We also find if we do not use any external knowledge, the dialogue-oriented GPT-2 is better than the general GPT-2, which differs from the previous (#0 vs #1). It verifies *SEG-CKRG* prefers to general GPT-2 again. In addition, (#2) and (#5) have similar overall performance, showing that explicit commonsense knowledge and implicit PLM knowledge play similar importance in our *SEG-CKRG*.

3.5 Case Study

We report two cases in Table 5 In the first case, *SEG-CKRG* generated two commonsense knowledge facts. We can find one knowledge fact is selected from the retrieved knowledge candidate set, but the other does not exist in the retrieved

Table 6. Error Case. Same notations as the Case Study.

Query #1	一句话证明你是搞游戏的。 One sentence proves that you are a game enthusiast.
Seq2Seq	对啊，我也是这么想的？ Yes, I think so too.
ConKADI	没有啊，电脑的时候电脑剧，你怎么看的。 No, it's a computer drama. What do you think of it.
BERT-PGN	哈哈哈，我就是这么想的，不知道为什么。 Hahaha, that's what I think. I don't know why.
SEG-CKRG	*K:*[G]游戏Synonym玩 [G]game, Synonym, play *R:* 我玩的不是游戏 是情怀。 What I play is not a game, but faith!
Query #2	哈哈哈哈你家的男朋友呢? Ha ha ha, where's your boyfriend?
Seq2Seq	你男朋友是谁？ Who is your boyfriend?
ConKADI	我家的男朋友在家里。 My boyfriend stays at home.
BERT-PGN	我家男朋友都不知道怎么回事了，我还是个孩子。 My boyfriend doesn't know what's going on. I'm still a child.
SEG-CKRG	*K:*[G]男朋友 Causes 拍照 [G]boyfriend, Causes, taking pictures *R:* 哈哈哈哈 拍照技术好。 Haha, haha, he is good at taking pictures

knowledge candidate set. This case verifies that our *Generative Knowledge Selection* mechanism can not only select knowledge but also extend the new knowledge. Subsequently, we can find the response generated by *SEG-CKRG* is highly relevant to the dialogue history and such generated knowledge facts, demonstrating the important role of the generated knowledge. Besides, compared to baselines, *SEG-CKRG* is the only one that can keep fluency, rationality, and informativeness at the same time. In the second case, we can find although ConKADI, BERT-PGN, and *SEG-CKRG* tried to introduce the new knowledge concept *phone*, only *SEG-CKRG* answered a proper response.

Limitations. Although *SEG-CKRG* has surpassed all baselines, we also find a limitation in the current work, i.e., *Error Propagation*. *SEG-CKRG* sequentially generates the selected/extended knowledge and the dialogue response. Thus, if irrelevant knowledge has been generated in the first knowledge generation procedure, the next response generation procedure will be impacted. We report two typical error cases in Table 6. In the first case, *SEG-CKRG* generated a new but incorrect knowledge fact '(game, Synonym, play)' in the knowledge generation procedure. *SEG-CKRG* wrongly predicted the relation between 'game' and 'play', where the correct relation should be 'CapableOf'. But fortunately, this level of error has little impact on the following dialogue response generation. *SEG-CKRG* still generated a better response than other baselines. In the next case, *SEG-CKRG* has generated an existing but contextually-irrelevant knowledge fact. This error has significantly impacted the relevance of the generated response. Without considering the dialogue query, the response generated by *SEG-CKRG* is still fluent.

4 Related Work

Knowledge-Grounded Response Generation (KRG): Due to the inability to access enough knowledge, traditional RG models [33,35] always generate safe but boring responses in spite of the given query [12,13]. Consequently, KRG models try to solve this issue by accessing the external knowledge bases [20,36,48]. Commonsense knowledge is a popular knowledge type in the current research [32,51], which helps a model to understand the dialogue, extend the topic, and then generate informative responses [38,40,41,44,46,50].

Pretrained Language Models (PLMs): PLMs such as BERT [6], RoBERTa [22], GPTs [2,27], and BARTs [11] have shown the dominate advantages in many NLP tasks [16]. PLMs can transfer the knowledge learned from massive unsupervised corpus to the open-domain RG models and bring significant improvements [8,31,37,47]. As for KRG models, previous works have shown PLMs can further prompt the text knowledge-grounded dialogue response generation [4,49] and the commonsense knowledge-grounded dialogue response generation [52].

Knowledge Selection: It is a research focus in KRG [7]. Non-PLM KRG models often adopt specific modules to conduct this job. [50], and [46] adopt graph neural networks, [41] uses the posterior response to help the learning of knowledge selection, [20], and [1] introduces copy networks to select the knowledge, [10] proposes a sequential knowledge selection paradigm, [28] proposes a global-to-local paradigm. In the era of PLMs, most knowledge selection methods that are originally designed for non-PLM KRG models become incompatible due to the restrictions of PLMs. Thus, the knowledge selection can only rely on the self-attention implemented by the Transformers of PLMs [23] or use the external module [49]. Meanwhile, such works can only select knowledge from a fixed and limited knowledge space, and the selection procedure is not very transparent. Different from such works, *SEG-CKRG* proposes a PLM-friendly *Generative Knowledge Selection* mechanism, which regards knowledge selection as a generative problem. Thus, our method can not only select the existing knowledge but also extend the new knowledge. Another difference is our work can explain the selection result using the human understandable natural language. In addition, although *TBS* [52] uses a PLM to generate knowledge, it does not include any knowledge selection procedure. The methodology of *TBS* is similar to our model (#4) in Table 4. Please refer to the corresponding results.

5 Conclusion

We propose an end-to-end CKRG model *SEG-CKRG*. Unlike previous works that can only use the limited and fixed knowledge retrieved by IR tools, *SEG-CKRG* introduces a novel *Generative Knowledge Selection (GenSel)* mechanism to select existing knowledge and extend new knowledge in a generative way.

More importantly, the knowledge selection/extension procedure has higher interpretability than previous works because the output is a natural language text. *SEG-CKRG* is implemented based on two Chinese GPT-2s pre-trained by ourselves. Finally, experimental results have shown the very competitive performance of *SEG-CKRG*.

Our future work includes three directions. First, we will continue to address the mentioned limitation; Second, we will explore and verify the effectiveness of *GenSel* in more different types of knowledge, such as text-based and table-based knowledge; Third, we are considering jointly modeling the CKRG task and the conversational relation extraction simultaneously by extending the potential of *GenSel*.

References

1. Bai, J., Yang, Z., Liang, X., Wang, W., Li, Z.: Learning to copy coherent knowledge for response generation. In: AAAI 2021 (2021)
2. Brown, T.B., et al.: Language models are few-shot learners. CoRR abs/2005.14165 (2020). https://arxiv.org/abs/2005.14165
3. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M. (eds.) SSST@EMNLP 2014 (2014)
4. Cui, L., Wu, Y., Liu, S., Zhang, Y.: Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In: EMNLP 2021, November 2021
5. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for Chinese bert. IEEE/ACM TASLP (2021)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019 (2019)
7. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: Knowledge-powered conversational agents. In: ICLR 2019 (2019)
8. Gu, X., Yoo, K.M., Ha, J.: Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In: AAAI2021 (2021)
9. Ippolito, D., Kriz, R., Sedoc, J., Kustikova, M., Callison-Burch, C.: Comparison of diverse decoding methods from conditional language models. In: ACL 2019, July 2019
10. Kim, B., Ahn, J., Kim, G.: Sequential latent knowledge selection for knowledge-grounded dialogue. In: ICLR 2020 (2020)
11. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL 2020 (2020)
12. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: NAACL 2016, June 2016
13. Li, J., Monroe, W., Jurafsky, D.: A simple, fast diverse decoding algorithm for neural generation. CoRR abs/1611.08562 (2016). http://arxiv.org/abs/1611.08562
14. Li, J., Tang, T., Zhao, W.X., Nie, J., Wen, J.: A survey of pretrained language models based text generation. CoRR abs/2201.05273 (2022). https://arxiv.org/abs/2201.05273
15. Li, J., Tang, T., Zhao, W.X., Wei, Z., Yuan, N.J., Wen, J.R.: Few-shot knowledge graph-to-text generation with pretrained language models. In: Findings of ACL-IJCNLP 2021 (Aug 2021)

16. Li, J., Tang, T., Zhao, W.X., Wen, J.: Pretrained language models for text generation: A survey. CoRR abs/2105.10311 (2021). https://arxiv.org/abs/2105.10311
17. 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 2021 (2021)
18. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain, July 2004
19. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. CoRR abs/2106.04554 (2021). https://arxiv.org/abs/2106.04554
20. Lin, X., Jian, W., He, J., Wang, T., Chu, W.: Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In: ACL 2020 (2020)
21. Liu, C.W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, J.: How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In: EMNLP 2016, November 2016
22. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692
23. Lotfi, E., Bruyn, M.D., Buhmann, J., Daelemans, W.: Teach me what to say and I will learn what to pick: Unsupervised knowledge selection through response generation with pretrained generative models. CoRR abs/2110.02067 (2021). https://arxiv.org/abs/2110.02067
24. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: EMNLP 2015 (2015)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL, pp. 311–318. ACL (2002)
26. Qin, L., Liu, Y., Che, W., Wen, H., Li, Y., Liu, T.: Entity-consistent end-to-end task-oriented dialogue system with KB retriever. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) EMNLP-IJCNLP 2019 (2019)
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
28. Ren, P., Chen, Z., Monz, C., Ma, J., de Rijke, M.: Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In: AAAI 2020, pp. 8697–8704 (2020)
29. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. In: Barzilay, R., Kan, M. (eds.) ACL 2017 (2017). 10.18653/v1/P17-1099
30. Serban, I.V., et al.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: AAAI 2017 (2017)
31. Shao, Y., et al.: CPT: a pre-trained unbalanced transformer for both Chinese language understanding and generation. CoRR abs/2109.05729 (2021). https://arxiv.org/abs/2109.05729
32. Speer, R., Havasi, C.: Conceptnet 5: a large semantic network for relational knowledge. In: The People's Web Meets NLP, Collaboratively Constructed Language Resources (2013)
33. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27 (2014)
34. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: ICLR 2018 (2018)
35. Vinyals, O., Le, Q.V.: A neural conversational model. CoRR abs/1506.05869 (2015). http://arxiv.org/abs/1506.05869

36. Wang, S., et al.: Modeling text-visual mutual dependency for multi-modal dialog generation. CoRR abs/2105.14445 (2021). https://arxiv.org/abs/2105.14445
37. Wang, Y., et al.: A large-scale chinese short-text conversation dataset. In: Zhu, X., Zhang, M., Hong, Yu., He, R. (eds.) NLPCC 2020. LNCS (LNAI), vol. 12430, pp. 91–103. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60450-9_8
38. Wu, S., Li, Y., Xue, P., Zhang, D., Wu, Z.: Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In: COLING 2022, pp. 521–531. International Committee on Computational Linguistics (2022). https://aclanthology.org/2022.coling-1.43
39. Wu, S., Li, Y., Zhang, D., Wu, Z.: Improving knowledge-aware dialogue response generation by using human-written prototype dialogues. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of EMNLP 2020 (2020)
40. Wu, S., Li, Y., Zhang, D., Wu, Z.: Generating rational commonsense knowledge-aware dialogue responses with channel-aware knowledge fusing network. IEEE ACM Trans. Audio Speech Lang. Process. **30**, 3230–3239 (2022). https://doi.org/10.1109/TASLP.2022.3199649
41. Wu, S., Li, Y., Zhang, D., Zhou, Y., Wu, Z.: Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In: ACL 202 (2020)
42. Wu, S., Wang, M., Li, Y., Zhang, D., Wu, Z.: Improving the applicability of knowledge-enhanced dialogue generation systems by using heterogeneous knowledge from multiple sources. In: WSDM 22 (2022)
43. Yan, R.: “Chitty-chitty-chat bot”: deep learning for conversational AI. In: IJCAI 2018 (2018)
44. Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., Huang, M.: Augmenting end-to-end dialogue systems with commonsense knowledge. In: AAAI 2018 (2018)
45. Yu, W., et al.: A survey of knowledge-enhanced text generation. CoRR abs/2010.04389 (2020). https://arxiv.org/abs/2010.04389
46. Zhang, H., Liu, Z., Xiong, C., Liu, Z.: Grounded conversation generation as guided traverses in commonsense knowledge graphs. In: ACL 2020 (2020)
47. Zhang, Y., et al.: DIALOGPT: large-scale generative pre-training for conversational response generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, July 2020
48. Zhao, X., Wu, W., Tao, C., Xu, C., Zhao, D., Yan, R.: Low-resource knowledge-grounded dialogue generation. In: ICLR 2020 (2020)
49. Zhao, X., Wu, W., Xu, C., Tao, C., Zhao, D., Yan, R.: Knowledge-grounded dialogue generation with pre-trained language models. In: EMNLP 2020 (2020)
50. Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X.: Commonsense knowledge aware conversation generation with graph attention. In: IJCAI 2018 (2018)
51. Zhou, P., et al.: Commonsense-focused dialogues for response generation: an empirical study. In: SIGdial 2021 (2021)
52. Zhou, P., et al.: Think before you speak: explicitly generating implicit commonsense knowledge for response generation. In: ACL 2022, May 2022