# IMPROVING DIALOGUE GENERATION VIA PROACTIVELY QUERYING GROUNDED KNOWLEDGE

*Xiangyu Zhao*[1] *Longbiao Wang*[1*] *Jianwu Dang*[2]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2] Japan Advanced Institute of Science and Technology, Japan

## ABSTRACT

Recent advances in pre-trained language models have significantly improved neural response generation. Further, an intelligent dialogue system should be able to give accurate responses that meet the needs of users. However, using appropriate knowledge has so far been proved difficult, because faced with a mass of relevant knowledge, the model not only needs to accurately retrieve the target information, but also integrate the information into dialogue response. In this paper, we propose a novel knowledge-based dialogue system which integrates the strength of a transformer-based generator and a knowledge retriever capable of proactively constructing queries for accurate information. Specifically, generator are used to give a response skeleton and knowledge retriever constructs a query based on response skeleton to obtain specific knowledge, which avoids excessive noise interference when generating responses. Experiments show that conversational systems that leverage knowledge integrator could generate more informative and human-like responses than strong baseline systems.

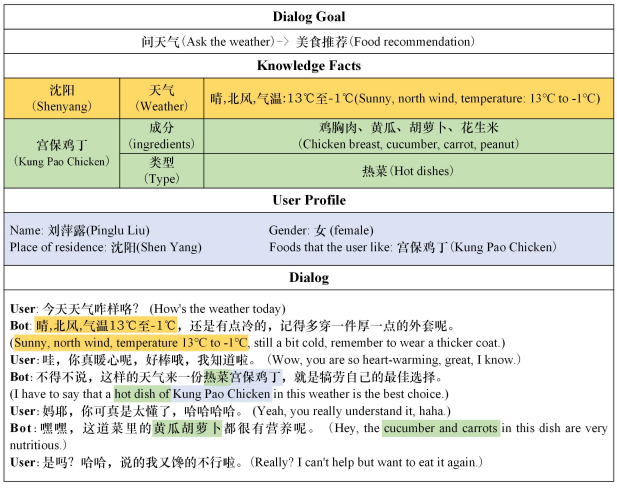
***Index Terms***— natural language generation, dialogue system, knowledge retrieval

## 1. INTRODUCTION

Large-scale pre-training language models using Transformer-based architectures have recently achieved remarkable successes in dialogue generation tasks [1, 2, 3]. While the pre-trained language models are knowledgeable to open domain dialogues [4], integrating external knowledge, e.g. real-time information, user profiles and personalized recommendation also effectively improves the accuracy of generated response and stimulate the interest of users as shown in Fig. 1. However, facing a large amount of external knowledge, models usually have no clear goal and digress to produce less coherent dialogues without accurate information especially for a long conversation. It may be explained as there are lots of noisy information in external knowledge, as a result, models do not have enough effective mechanisms to retrieve the required information accurately.

*Corresponding author, E-mail: longbiao_wang@tju.edu.cn

In order to select the appropriate information, many works have made great progress. While much of researchers focus on selecting knowledge based on the semantic similarity between input utterances and knowledge [5, 6, 7], Lian et al. [8] and Kim et al. [9] consider both dialogue contexts and real responses by a novel knowledge selection mechanism which uses both prior and posterior distributions over knowledge. They usually select specific knowledge related to the real response based on semantic similarity without explicit reasoning. Although this mechanized way of defining the scope of "truth knowledge" can avoid introducing too much noise, it also loses the opportunity to use diversified knowledge, especially in the condition that machine proactively guides the conversation [10]. Moreover, in previous works, we can find that models introduced with external knowledge usually have larger perplexity. Therefore, intelligent dialogue system needs a more flexible reasoning strategy to use external knowledge to balance the contradiction between the introduced noise and the diversification of information when generating replies.



**Dialog Goal**

问天气(Ask the weather)-> 美食推荐(Food recommendation)

**Knowledge Facts**

| | | |
|---|---|---|
| 沈阳 (Shenyang) | 天气 (Weather) | 晴,北风,气温:13℃至-1℃(Sunny, north wind, temperature: 13℃ to -1℃) |
| 宫保鸡丁 (Kung Pao Chicken) | 成分 (ingredients) | 鸡胸肉、黄瓜、胡萝卜、花生米 (Chicken breast, cucumber, carrot, peanut) |
| | 类型 (Type) | 热菜(Hot dishes) |

**User Profile**

Name: 刘萍露(Pinglu Liu) Gender: 女 (female)
Place of residence: 沈阳(Shen Yang) Foods that the user like: 宫保鸡丁(Kung Pao Chicken)

**Dialog**

**User**: 今天天气咋样咯? (How's the weather today)
**Bot**: 晴,北风,气温13℃至-1℃，还是有点冷的，记得多穿一件厚一点的外套呢。
(Sunny, north wind, temperature 13℃ to -1℃, still a bit cold, remember to wear a thicker coat.)
**User**: 哇，你真暖心呢，好棒哦，我知道啦。(Wow, you are so heart-warming, great, I know.)
**Bot**: 不得不说，这样的天气来一份热菜宫保鸡丁，就是犒劳自己的最佳选择。
(I have to say that a hot dish of Kung Pao Chicken in this weather is the best choice.)
**User**: 妈耶，你可真是太懂了，哈哈哈哈。(Yeah, you really understand it, haha.)
**Bot**: 嘿嘿，这道菜里的黄瓜胡萝卜都很有营养呢。(Hey, the cucumber and carrots in this dish are very nutritious.)
**User**: 是吗? 哈哈，说的我又馋的不行啦。(Really? I can't help but want to eat it again.)

**Fig. 1**. We use different colors to indicate different specific information needed in the conversation.

In this work, we investigate the issue of using accuracy knowledge in the multi-turn knowledge-grounded dialogue.

In knowledge base. Therefore, querying specific knowledge by predicting a fixed relationship can simplify the pressure of generator and query the required information more accurately. While previous works focusing on selecting knowledge first and then generating [8, 11], we propose a knowledge based query generation dialogue system (KBQD) which integrates the strength of knowledge retrieval system and transformer-based generator. Our model generates a response skeleton without specific knowledge but the relationship information with dialogue goal or user first, then it constructs queries based on the relationship and dialogue goal/user to access the knowledge base and get the final response. This method can not only avoid the introduction of noise information in our generative model, but also use the knowledge information in knowledge base.

## 2. KBQD ARCHITECTURE

The whole architecture of our model is shown in Fig. 2. Our model generates responses in two stages:

**Response Skeleton Generator**: On the basis of RoBERTa [12] architecture, we change the model into a generic transformer language model [13] which includes two parts: goal generator and response generator.

**Knowledge Integrator**: Knowledge Reasoners accept the output of Response Skeleton Generator as query and knowledge base as context to get the accuracy knowledge.

### 2.1. Response Skeleton Generator

The generator consists of a paralleled transformer encoder and a loop decoder. Before training process, we replace the specific knowledge information in dialogue with task-specific mask, so the model only needs to judge whether the fragments in response needs to refer relevant knowledge. Specifically, we see the historical dialogue as a long sequence $C = \{s_1, [SEP], s_2, [SEP], ..., s_n\}$, where $n$ is the total number of sentences in the context and we use $[SEP]$ to separate different sentences. Also, we use $[GOAL]$ to separate different goals as $G = \{g_1, [GOAL], g_2, [GOAL], ..., g_n\}$. Before passing the embedding layer, we merge the context information with the goal sequence. We add each goal at the beginning of sentence, so the input is $U = \{g_1, [GOAL], s_1, [SEP], ..., g_n, [GOAL], s_n\}$. For the specific knowledge information that appears in the input and knowledge base at the same time, we use task-specific masks to cover it. It contains the relationship between knowledge item and the goal entity to facilitate the knowledge query generation such as $[Mask - comment]$ is used to replace the song's comments in Fig. 2. For the sake of brevity, we introduce the input as $U = \{w_1, w_2, ..., w_n\}$ in the follow-up content of the paper.

After embedding layer, the encoder employs $N$ layers to make a deep representation of the input context information, in the $i$-th layer of encoder, firstly, the input representation $F_U^{i-1} = \{\boldsymbol{\omega}_1^{i-1}, ..., \boldsymbol{\omega}_n^{i-1}\}$ from the preceding layers are fed into multi-head self-attentions (MH_ATT) [14] layer:

$$\boldsymbol{A}_U^i = \text{MH\_ATT}(\boldsymbol{F}_U^{i-1}, \boldsymbol{F}_U^{i-1}, \boldsymbol{F}_U^{i-1}) \quad (1)$$

where $i \in [1, N]$, $\boldsymbol{F}_U^{(0)} = \text{In}_U$. In the next step, the second sub-layer is a feed-forward network (FFN) which consists of two linear transformations with a GELU activation in between. This process can be stated as $\boldsymbol{F}_U^i = \text{FFN}(\boldsymbol{A}_U^i)$. After $N$ layers, we can get the outputs of encoder:

$$\boldsymbol{u} = \boldsymbol{F}_U^N = \{\boldsymbol{\omega}_1^o, ..., \boldsymbol{\omega}_n^o\} \quad (2)$$

Our decoder generates dialogue goals and response skeleton word by word. The first step is to generate the dialogue goal. We use the Nucleus Sampling method to increase the probability of high frequency words and decrease the probability of low frequency words: $Decoder_{goal} = Nucleus(u)$. The second step is to generate a response skeleton based on the information obtained in the first step. In this process, we found that when the complete response is used as the training target of the model, beam search can generate smoother text than nucleus sampling. So we use beam search as the decoding method:$Decoder_{response} = Beamsearch(u')$.

### 2.2. Knowledge Integrator

Knowledge Integrator retrieves the specific knowledge information in response skeleton. For structured knowledge, because knowledge triples are given, we directly use the relationships (predicate) as task-specific mask. So the structured knowledge reasoner leverage the relationship information to retrieve the knowledge base. First, we filter out relevant knowledge triple based on the relationship information. Then, model concatenates the goal sequence (subject) and relationship (predicate) as query and computes the cosine distance between query and knowledge items (object) to choose specific knowledge. Because the value of cosine similarity calculated directly using BERT word vectors is generally large, we use the word2vec pre-trained word vectors [1].

As for unstructured knowledge, we first use named entity recognition tools spaCy [2] to get the entities both mentioned in the dialogue sentences and grounding knowledge, then use the entity type information as task-specific mask. Specifically, spaCy provides 18 types of entities and we will query specific entity in the knowledge base according to the type information and response skeleton. For unstructured knowledge reasoning process, we regard the reasoning task as machine reading comprehension task, we represent the input query and text as a single packed sequence, where the query is response skeleton, the text is grounded knowledge and specific entities is answer.

[1] https://github.com/Embedding/Chinese-Word-Vectors

[2] https://github.com/explosion/spaCy

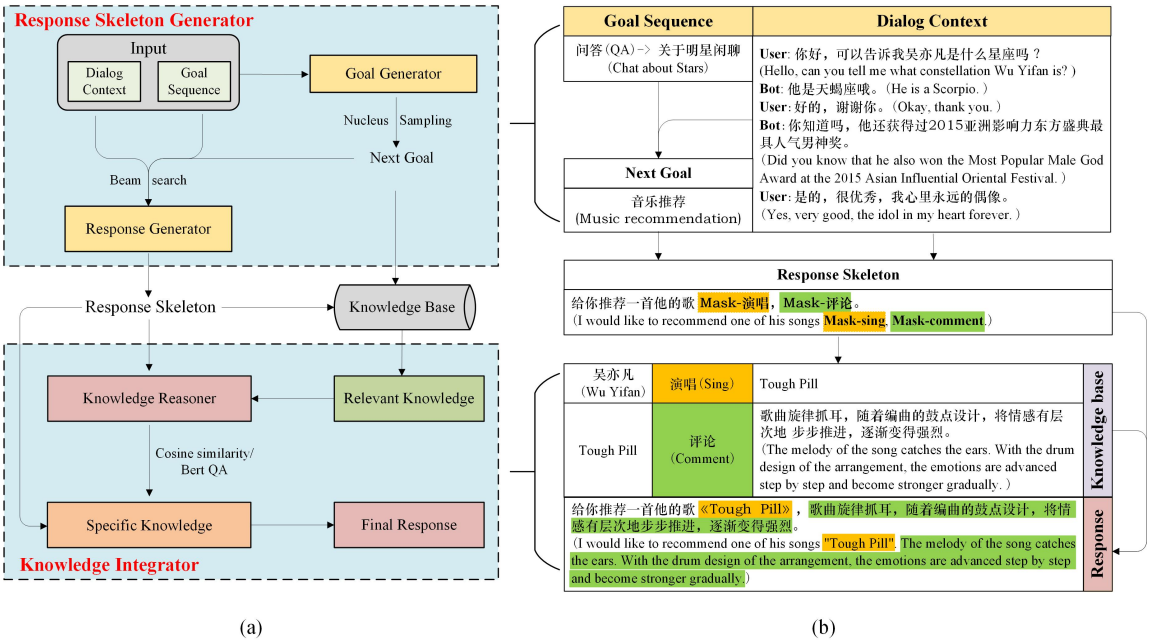



**Fig. 2**. Overview of the dialogue system with knowledge based query generation mechanism.

### 2.3. Loss Function

According to Radford et al. [15], including language modelling as an auxiliary objective to the fine-tuning helps improving generalization of the supervised model and accelerating convergence. We first train the language model, then fine-tune the goal generator and response skeleton generator. The loss function of language model can be written as:

$$\mathcal{L}_{LM} = -\sum_{t} \log P(w_t|w_1, ..., w_{t-1}) \tag{3}$$

In the process of generating, we concatenate goal sequence and dialogue context as input and generate the next goal and response skeleton as output. We leverage the token negative log likelihood with label smoothing algorithm [16]:

$$\mathcal{L}_{GM} = -\sum_{t} \log P(y_t|y_1, ..., y_{t-1}) - D_{KL}(f||P(y_t|y_1, ..., y_{t-1})) \tag{4}$$

Our final loss function combines Eq.3 and Eq.4:

$$\mathcal{L}_{all} = \lambda_{LM} \cdot \mathcal{L}_{LM} + \mathcal{L}_{GM} \tag{5}$$

where $\lambda_{LM}$ and $\lambda_{RM}$ are the weights of the language model loss function during fine-tuning. In the language modeling stage, we set $\lambda_{LM} = 1$ and $\lambda_{GM} = 0$, and in the dialogue generation stage we set $\lambda_{LM} = 0.2$ and $\lambda_{GM} = 0.8$.

## 3. EXPERIMENT AND DISCUSSION

### 3.1. Dataset

In order to assess the performance of our methods, we conduct experiments on unstructured knowledge based dataset, Wizard-of-Wikipedia dataset [17] and a structured knowledge based dataset, DuRecDial dataset [10]. Wizard-of-Wikipedia is a large-scaled knowledge-grounded chit-chat dataset. This dataset covers 1365 natural, open-domain dialogue topics and each topic is linked to a Wikipedia article. The final dialogue dataset we leverage consists 166,787 turns for training, 17,715 turns for validation, and 17,497 turns for testing. The testing set is split into two subsets: Test Seen contains 8,715 turns of conversations on some overlapping topics with the training set, while Test Unseen contains 8,782 turns on topics never seen before in train or validation. DuRecDial is a human-to-human recommendation dialog dataset (about 10k dialogs, 164k utterances and 21.93 knowledge facts on average), where there are multiple sequential dialogs for a pair of a recommendation seeker (user) and a recommender (bot). Each seeker has an explicit profile for the modelling of personalized recommendation. In this data set, the specific knowledge information that appears in the dialogue sentence can be found in the ground knowledge and user profiles.

### 3.2. Baselines

Since we want to verify the effectiveness of our method in structured and unstructured knowledge bases, we adopted advanced models on the two datasets respectively. The baselines

**Table 1**. Automatic evaluation on dataset WoW.

| Model | Wizard-of-Wikipedia (Test Seen/Unseen) | | | |
|---|---|---|---|---|
| | BLEU1 | BLEU2 | F1 | PPL |
| PostKS | 16.72/14.81 | 6.88/4.66 | 0.18/0.13 | 59.8/91.8 |
| SKT | 17.68/16.14 | 7.32/5.93 | 0.19/0.16 | 50.2/81.4 |
| DukeNet | 17.99/16.34 | 7.51/5.99 | 0.19/0.17 | 42.3/64.1 |
| **KBQD** | **19.24/18.21** | **10.36/8.74** | **0.21/0.19** | **21.4/33.8** |

**Table 2**. Automatic evaluation on dataset DuRec.

| Model | DuRec Dataset | | | |
|---|---|---|---|---|
| | BLEU1/BLEU2 | F1 | PPL | Distinct-1/2 |
| MGCG | 35.34/25.19 | 0.36 | 17.69 | 0.101/0.339 |
| KnowHRL | 35.26/21.14 | 0.37 | 18.53 | 0.092/0.297 |
| GOKC | 41.30/31.80 | 0.47 | **11.38** | 0.072/0.193 |
| **KBQD** | **51.42/41.15** | **0.53** | 11.86 | **0.125/0.314** |

on unstructured knowledge dataset :

**PostKS**: This model employs a knowledge selection mechanism where both prior and posterior distributions over knowledge are used to facilitate knowledge selection [8].

**SKT**: It first leverage a sequential latent variable model to do knowledge selection. This model can keep track of the prior and posterior distribution over knowledge [9].

**DukeNet**: This model leverages an unsupervised learning scheme to explore knowledge beyond the demonstrated ones in the dataset[18].

The baselines on structured knowledge dataset :

**MGCG**: It includes an automatic goal planning model and a response generation model. The generative-based model is proposed by [19], which is the baseline on DuRecDial.

**KnowHRL**: It is a three-layer knowledge aware hierarchical reinforcement Learning based model. It leverage hierarchical goal planning to facilitate chatting topic management [20].

**GOKC**: It has a goal-oriented knowledge discernment mechanism to discern the knowledge facts that are highly correlated to the dialog goal and the dialog context. [11].

### 3.3. Experimental Details

In our generation model, we reuse the language model RoBERTa[3] and RoBERTa for Chinese[4] which is published by [12]. We fine-tune the attention mask mechanism of the RoBERTa by adding a attention mask of the lower triangle method. Specifically, we train the model 3 epochs for pre-training the language generation model on dialogue datasets we processed and 5 epochs to minimize overall loss. As for our unstructured knowledge reasoner module, we reuse the fine-tuning BERT on the SQuAD 2.0 task, which is based on $\mathrm{BERT}_{\mathrm{LARGE}}$[5]. We first use Named Entity Recognition tool spaCy to extract the entities mentioned both in the sentences and knowledge base. We divide the identified named entities into 18 categories and use the categories to mask specific knowledge information. After processing, Wizard-of-Wikipedia has 485,762 named entities appeared in dialogue sentences and ground knowledge. We fine-tune BERT QA on Wizard-of-Wikipedia dataset to predict the masked knowledge and achieves 41.13% accuracy on our dataset.

### 3.4. Overall Results

Table 1 and 2 lists the automatic evaluation results for each baseline and our method. Our models outperform almost the baselines. Among them, our method reaches an extremely low ground truth perplexity. This is within our expectations, because we did not introduce relevant knowledge in our generative model, and generator doesn't need to give specific knowledge. When the model generates a response, the uncertainty is reduced, and when it encounters a part (specific knowledge) that is difficult to generate, it will be handed over to the next module for processing. Also, the performance of KBQD on F1 is higher than other models which indicates that our model have the ability to understand and determine which pieces of knowledge fit into the current conversation. What's more, KBQD is more stable on both Test Seen and Test Unseen, which indicates that the robustness of our model is better. At the same time, we can find that the performance of our model on the DuRecDial dataset is much higher than WoW dataset. This is because the knowledge in DuRecDial is triples form and appeared in the dialogue completely. This makes it easier for the model's knowledge integrator to determine the knowledge needed in the dialogue. This also illustrates the role of generating knowledge base query methods in promoting dialogue generation from the side.

## 4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new architecture for open-domain knowledge grounded dialog system, which incorporates a generative model to generate response skeleton and knowledge integrator to retrieve specific knowledge. In the future, we intend to incorporate the multi-task learning into the dialog system and knowledge retrieval system to enhance the quality of generated response.

## 5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200 and the National Natural Science Foundation of China under Grant 62176182.

[3]https://github.com/pytorch/fairseq/tree/master/examples/roberta

[4]https://github.com/ymcui/Chinese-BERT-wwm

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

## 6. REFERENCES


[1] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, JJ (Jingjing) Liu, and Bill Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," in *arXiv:1911.00536*, November 2019.

[2] Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha, "Dialogbert: Discourse-aware response generation via learning to recover and rank utterances," *arXiv preprint arXiv:2012.01775*, 2020.

[3] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al., "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[4] Yufan Zhao, Wei Wu, and Can Xu, "Are pre-trained language models knowledgeable to ground open domain dialogues?," *arXiv preprint arXiv:2011.09708*, 2020.

[5] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra, "Towards exploiting background knowledge for building conversation systems," in *Proceedings of the 2018 Conference on EMNLP*, 2018, pp. 2322–2332.

[6] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu, "Commonsense knowledge aware conversation generation with graph attention," in *IJCAI-18*, 7 2018, pp. 4623–4629.

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

[8] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu, "Learning to select knowledge for response generation in dialog systems," *arXiv preprint arXiv:1902.04911*, 2019.

[9] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim, "Sequential latent knowledge selection for knowledge-grounded dialogue," in *International Conference on Learning Representations*, 2020.

[10] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu, "Towards conversational recommendation over multi-type dialogs," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. July 2020, pp. 1036–1049, Association for Computational Linguistics.

[11] Jiaqi Bai, Ze Yang, Xinnian Liang, Wei Wang, and Zhoujun Li, "Learning to copy coherent knowledge for response generation," 2021.

[12] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," 2018.

[16] Sergey Edunov, Myle Ott, Michael Auli, David Grangier, et al., "Classical structured prediction losses for sequence to sequence learning," in *Proceedings of the 2018 Conference of the NAACL*, 2018, pp. 355–364.

[17] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," in *International Conference on Learning Representations*, 2019.

[18] Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke, "Dukenet: A dual knowledge interaction network for knowledge-grounded conversation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1151–1160.

[19] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu, "Towards conversational recommendation over multi-type dialogs," *arXiv preprint arXiv:2005.03954*, 2020.

[20] Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che, "Knowledge graph grounded goal planning for open-domain conversation generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9338–9345.