electronics

MDPI

Article

A Multi-Hop Reasoning Knowledge Selection Module for Dialogue Generation

Zhiqiang Ma [1,2,*], Jia Liu [1], Biqi Xu [1], Kai Lv [1] and Siyuan Guo [1]

[1] College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China
[2] Inner Mongolia Autonomous Region Engineering & Technology Research Centre of Big Data Based Software Service, Inner Mongolia University of Technology, Hohhot 010080, China
* Correspondence: mzq_bim@imut.edu.cn

Abstract: Knowledge selection plays a crucial role in knowledge-driven dialogue generation methods, directly influencing the accuracy, relevance, and coherence of generated responses. Existing research often overlooks the handling of disparities between dialogue statements and external knowledge, leading to inappropriate knowledge representation in dialogue generation. To overcome this limitation, this paper proposes an innovative Multi-hop Reasoning Knowledge Selection Module (KMRKSM). Initially, multi-relational graphs containing rich composite operations are encoded to capture graph-aware representations of concepts and relationships. Subsequently, the multi-hop reasoning module dynamically infers along multiple relational paths, aggregating triple evidence to generate knowledge subgraphs closely related to dialogue history. Finally, these generated knowledge subgraphs are combined with dialogue history features and synthesized into comprehensive knowledge features by a decoder. Through automated and manual evaluations, the exceptional performance of KMRKSM in selecting appropriate knowledge is validated. This module efficiently selects knowledge matching the dialogue context through multi-hop reasoning, significantly enhancing the appropriateness of knowledge representation and providing technical support for achieving more natural and human-like dialogue systems.

Keywords: knowledge selection; dialogue generation; knowledge selection; human–computer interaction

check for updates

Citation: Ma, Z.; Liu, J.; Xu, B.; Lv, K.; Guo, S. A Multi-Hop Reasoning Knowledge Selection Module for Dialogue Generation. *Electronics* **2024**, *13*, 3275. https://doi.org/10.3390/electronics13163275

Academic Editors: Ruifeng Xu, Tong Xu and Yanyan Zhao

Received: 26 July 2024
Revised: 14 August 2024
Accepted: 16 August 2024
Published: 19 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Knowledge selection plays a vital role in knowledge-driven dialogue generation methods, significantly influencing the accuracy, relevance, and coherence of the produced responses. Faced with a vast and diverse array of knowledge triples in external knowledge graphs, dialogue generation models need to precisely filter and integrate information. However, when external knowledge becomes overly intricate, models often struggle to make appropriate knowledge selections, leading to responses that contradict dialogue history or external knowledge, or are difficult to verify for accuracy. This study employs existing dialogue generation models to verify whether the model produces inaccurately expressed responses. As illustrated in Figure 1 [1], the model confuses the experiences and attributes of two distinct individuals. Similar issues have been observed in ChatGPT, prompting OpenAI to rely heavily on user feedback to rectify such problems, a time-consuming and labor-intensive process. Thus, enhancing the quality of dialogue generation through improved knowledge selection methods is imperative.



Figure 1. The problem of inappropriate knowledge representation [1].

Previous research indicates that inappropriate knowledge representation in dialogue generation models can be addressed through useful knowledge retrieval [2], control algorithms [3], and post-generation processing [4]. However, these studies have not focused on resolving the gaps between dialogue statements and external knowledge information. In this paper, we introduce a Multi-hop Reasoning Knowledge Selection Module (KMRKSM), which uses multi-hop reasoning on external knowledge graphs as a knowledge selection module to address issues of inappropriate knowledge representation in knowledge-driven dialogue generation models. This module extends concepts from dialogue history into subgraphs, using them as foundations for commonsense knowledge. It begins by encoding multi-relational graphs with composite operations to obtain graph-aware representations of concepts and relationships. Then, the multi-hop reasoning module dynamically infers by aggregating triple evidence across multiple relational paths, generating knowledge subgraphs closely related to the dialogue history. Finally, these knowledge subgraphs are combined with dialogue history features to produce knowledge features through a decoder. Our main contributions are as follows:

1. We introduce the Multi-hop Reasoning Knowledge Selection Module (KMRKSM) to address inappropriate knowledge selection in dialogue generation methods.
2. By combining multi-relational graph coding units with dynamic multi-hop reasoning units, we differentiate between dialogue history and external knowledge during the selection process, enabling effective knowledge fusion.
3. Comparative experiments with a range of baselines on the OpenDialKG dataset, along with ablation studies, show that KMRKSM's knowledge selection performance matches or exceeds that of baseline methods.

2. Related Work

To address the issue of inappropriate knowledge representation in dialogue generation models, researchers have been improving knowledge selection methods. Traditional approaches could only extract superficial knowledge from dialogue history. Zhou et al. [5] proposed a dialogue generation model based on commonsense knowledge graphs, using structured knowledge in the form of triples to understand dialogues. Dinan et al. [2] used factual knowledge to guide knowledge selection, showing that effective knowledge improves response quality. Lian et al. [6] applied posterior probability distributions for knowledge selection, enhancing dialogue generation quality. Kim et al. [7] introduced a

sequential latent variable model to improve knowledge selection in multi-turn dialogues. Chen et al. [8] proposed a posterior information prediction module and a distillation-based training strategy to further refine knowledge selection. Dziri et al. [9] developed the NPH model, which uses an illusion evaluation mechanism and query signal propagation during the refinement stage to retrieve entity information relevant to dialogue statements. Shuster et al. [3] explored retrieval loop structures in various neural networks and designed a knowledge retriever using these structures for knowledge selection. Rashkin et al. [10] introduced an illusion control method with resampling techniques during decoding. Wu et al. [4] defined a control mechanism that incorporates lexical control phrases and inductive attention to eliminate non-informative attention links. Ma et al. [11] presented a knowledge-driven dialogue generation model with posterior knowledge selection based on the Siamese network, creating a new mechanism to acquire richer knowledge information.

The aforementioned studies did not emphasize addressing the differences between dialogue statements and external knowledge. Therefore, this paper employs knowledge grounding and multi-hop reasoning techniques to dynamically infer by aggregating triple evidence along multiple relational paths, selecting knowledge subgraphs significantly related to dialogue statements. Throughout the knowledge selection process, grammatical correctness, natural expression, cultural adaptability, and semantic fidelity are ensured, enhancing the fluency and accuracy of the generated responses.

3. Task Formulation

Given the dialogue history $D = \{U_1, R_1, U_2, R_2, \cdots, U_i, R_i, \cdots, U_T\}$ and external knowledge K, knowledge features u_k that align with the dialogue context are selected through the dialogue history D and external knowledge K. Here, U_i represents the word sequence of the user utterance in the i-th dialogue turn,$U_i = \left\{x_{i,1}^U, x_{i,2}^U, \cdots, x_{i,j}^U, \cdots, x_{i,L^U}^U\right\}$, and R_i represents the word sequence of the response sentence in the *i-th* dialogue turn, $R_i = \left\{x_{T,1}^R, x_{T,2}^R, \cdots, x_{T,j}^R, \cdots, x_{T,n}^R\right\}$. $x_{i,j}^U$ and $x_{i,j}^R$ represent the j-th word in the user utterance and response sentence of the i-th dialogue turn, respectively. The external knowledge $KG = (V, E)$ introduced comprises a set of concepts, denoted by V, and the relationships connecting these concept sets, denoted by E. Due to the complexity of direct reasoning on the complete graph, we extract a subgraph, $KG_s = (v, \varepsilon)$ that is composed of h-hop paths that are mutually connected starting from the extracted source entity C_x from the dialogue history, where $v \subset V$ and $\varepsilon \subset E$. Subsequently, we formulate the task to generate the optimal knowledge features u_k, maximizing the following conditional probability, as shown in Equation (1):

$$u_k = \text{argmax} P(KG_s|D) \tag{1}$$

4. Multi-Hop Reasoning Knowledge Selection Module

In order to address the issue of inappropriate knowledge representation in dialogue generation caused by neglecting the handling of differences between dialogue utterances and external knowledge, this paper introduces KMRKSM. The selected knowledge features are used as inputs for generating responses to achieve knowledge-driven dialogue generation, with the module structure depicted in Figure 2.

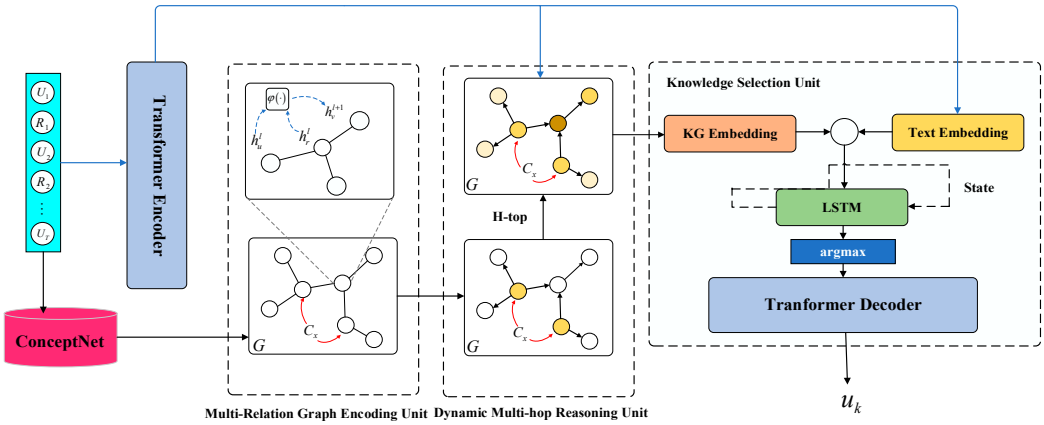


Figure 2. Structure of Multi-hop Reasoning Knowledge Selection Module.

KMRKSM consists of three main components: the multi-relation graph coding unit, the dynamic multi-hop reasoning unit, and the knowledge selection unit.

Utilizing a pre-trained Transformer encoder to capture contextual dependencies within a dialogue's historical text sequence, the model processes the dialogue's historical text sequence $D = \{U_1, R_1, U_2, R_2, \cdots, U_i, R_i, \cdots, U_T\}$ as the input, with the specific procedure denoted by Equations (2) through (4).

$$h_t^o = e_t + p_t \tag{2}$$

$$h_t^l = T_block\left(H_{\leq t}^{l-1}\right), l \in [1, L_D] \tag{3}$$

$$P(s_t|s < t) = softmax\left(W_{LM} h_t^{L_D} + b\right) \tag{4}$$

Here, e_t and p_t represent token-embedding vectors and positional embedding vectors, respectively, T_block is a Transformer module with masked self-attention, and $h_t^{L_D}$ is the hidden state at time step t.

4.1. The Multi-Relation Graph Coding Unit

In reference to Vashishth et al. [12], a multi-relational encoding unit was designed to utilize non-parametric $\varphi(\cdot)$ composition operations on a commonsense knowledge graph aligned with the dialogue history, combining node embeddings and relation embeddings. Specifically, with the input commonsense knowledge graph $KG = (V, E)$ and a GCN with L_G layers, for each node $v \in V$, the information of the $l+1$ layer is updated by aggregating local neighborhood graphs $N(v)$ composed of nodes u and relations. This process is represented through Equations (5) to (6).

$$o_v^l = \frac{1}{|N(v)|} \sum_{(u,r)\in N(v)} W_N^l \varphi\left(h_u^l, h_r^l\right) \tag{5}$$

$$h_v^{l+1} = \mathrm{ReLU}\left(o_v^l + W_S^l h_v^l\right) \tag{6}$$

Here, h_v^0 denotes the initialization for word embeddings, h_r^0 denotes the initialization for relation embeddings, and W_N^l and W_S^l represent two learnable weight matrices in the l-th layer. The non-parametric composition operation $\varphi(h_u, h_r) = h_u - h_r$ is defined through TransE.

The relation embedding is also updated via another linear transformation, as shown in Equation (7).

$$h_r^{l+1} = W_R^l h_r^l \tag{7}$$

Finally, the obtained node embeddings $h_v^{L_D}$ and relation embeddings $h_r^{L_D}$ are used as inputs to the dynamic multi-hop reasoning unit.

4.2. The Dynamic Multi-Hop Reasoning Unit

To enable explicit reasoning on the graph structure during knowledge selection, a dynamic multi-hop reasoning unit was designed. This unit utilizes both the structural knowledge graph and contextual information, passing evidence along relationship paths at each decoding step. The unit updates the scores of external nodes, propagating information through multiple hops until all nodes on G are visited. Initially, nodes corresponding to concepts in C_x are given a score of 1, while the scores for other unvisited nodes are set to 0.

For unvisited nodes $v \in V$, their scores $ns(v)$ are calculated by aggregating evidence from $Nin(v)$, which represents the set of visited nodes and their direct connections v through edge r, as shown in Equation (8).

$$ns(v) = \underset{(u,r)\in Nin(v)}{f} (\gamma \cdot ns(u) + R(u,r,v)) \tag{8}$$

Here, γ is a discount factor controlling the strength of information flow from the previous hop. $f(\cdot)$ is an aggregator for pooling scores from connected nodes, primarily utilizing $\max(\cdot)$ as the main aggregator.

The triple correlation, $R(u,r,v)$, reflects the relevance of a triple set (u,r,v) in providing evidence in the current context, computed as shown in Equations (9) and (10).

$$R(u,r,v) = \sigma\left(h_{u,r,v}^{T} W_{sim} h_t^{L_D}\right) \tag{9}$$

$$h_{u,r,v} = \left[h_u^{L_D}; h_r^{L_D}; h_v^{L_D}\right] \tag{10}$$

Finally, after H, the final distribution reaching each node is normalized, as shown in Equations (4)–(11).

$$P(C_t | s < t, G) = soft\max_{u\in V}(ns(v)) \tag{11}$$

Here, C_t represents the concept of the chosen node at time step t.

4.3. The Knowledge Selection Unit

To obtain knowledge more aligned with the conversational context, a knowledge selection unit was devised to acquire semantic embeddings of the dialogue history D through the Sentence-BERT contextual sentence encoder. The model ϕ is a decoder based on LSTM, which, at step t, predicts the probability P_t of an action $\vec{a_t}$ based on a state $\vec{s_t}$. Here, an action denotes the steps of multi-hop reasoning on the graph G, represented by the connections of relations and entity embeddings from the graph G, along with semantic embeddings of context sentences based on Sentence-BERT, as formulated in Equation (12).

$$\vec{a_t} = \left(\vec{e_G} + \vec{e_S}\right) \oplus \left(\vec{r_G} + \vec{r_S}\right) \tag{12}$$

where $\vec{e_S}$ and $\vec{r_S}$ represent the semantic embeddings of the context sentences for entity e and relation r, respectively. The state $\vec{s_t}$ includes a representation of the dialogue history along with the entities and relations the model ϕ has traversed, defined as a tuple $\left(D, \left(\vec{a_1}, \vec{a_2}, \cdots, \vec{a_{t-1}}\right)\right)$. Consequently, the model ϕ simulates the knowledge selection process based on the dialogue history D.

Subsequently, the best knowledge with the highest probability $P_\phi = \prod_t P_t, \phi$ is selected from the knowledge subgraph obtained from the dynamic multi-hop reasoning unit, as shown in Equation (13).

$$u_k = \text{argmax} P_\phi(A = G|D) \tag{13}$$

where A is a set of model ϕ actions $\vec{a_t}$ conditioned on the dialogue history D.

5. Experiment

5.1. Evaluation Indicators

To assess the effectiveness of the knowledge selection module in selecting knowledge, this study conducts experiments using both automatic evaluation metrics and human evaluation metrics.

5.1.1. Automatic Evaluation

To evaluate the quality of selected knowledge, this study assesses the quality of knowledge using automatic evaluation metrics based on word overlap, namely BLUE4 and ROUGE-L [13].

BLEU4: This can measure the fluency of generated sentences, as shown in Equation (14).

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} W_n \times \log P_n\right) \tag{14}$$

where BP is the brevity penalty factor, W_n is the weight for the n-gram typically set to a uniform weight, where for any n, it holds that $W_n = 1/N$, and P_n is the precision for the n-gram.

ROUGE-L [13]: This measures the Longest Common Subsequence (LCS) between the candidate and reference sentences, assessing the fluency of generated sentences, as shown in Equations (15)–(17).

$$R_{LCS} = \frac{LCS(C,S)}{len(S)} \tag{15}$$

$$P_{LCS} = \frac{LCS(C,S)}{len(C)} \tag{16}$$

$$ROUGE - L = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{17}$$

Here, C denotes sentences generated by the model, S denotes reference sentences, R_{LCS} represents the recall rate, and P_{LCS} signifies the precision rate.

To evaluate the appropriateness of knowledge selection, this study employs the source-dependent automatic evaluation metrics FeQA [14], QuestEval [15], and Entity Coverage [16] to assess the appropriateness of knowledge selection.

FeQA [14]: This is a question-answering (QA)-based evaluation metric designed to assess the reliability of generated text. The metric uses a source (e.g., a document) and its corresponding output (e.g., a summary) as the input. A question generation model creates a question from the source, and an answer a is generated. Then, the model generates another answer b based on the question and the output. The reliability of the generated text is evaluated by calculating the average F1 score between answers a and b. In this study, all triples in the knowledge graph are connected, with the dialogue history as the source and the generated dialogue response as the output. The FeQA score is calculated using models from the official library to evaluate the appropriateness of knowledge selection.

QuestEval [15]: This is another QA-based evaluation metric with two modes: Reference-Dependent (RD) mode, which uses one or more ground truth reference texts, and Reference-Free (RF) mode, which operates without any reference text. The input sources are constructed in the same way as in the FeQA metric. The QuestEval score, used to assess the appropriateness of knowledge selection, is calculated using models from the official library.

Entity Coverage [16]: This uses a named entity recognition model to extract named entities from both the generated dialogue responses and the dialogue history. This study computes the Entity Precision, recall, and F1 scores between the named entities in the selected knowledge and those in the dialogue history to evaluate the appropriateness of knowledge selection.

5.1.2. Manual Evaluation

To evaluate the quality of knowledge selection from different models, human assessments were conducted using Amazon Mechanical Turk. In assessing the appropriateness of knowledge selection, annotators were first asked to determine whether each selected piece of knowledge was Faithful or Hallucinated. Faithful knowledge is defined as being supported by knowledge triplets and the dialogue context, whereas Hallucinated knowledge contradicts or cannot be verified by the dialogue context. Annotators were then asked to further specify whether the Hallucinated knowledge was Extrinsic, Intrinsic, or Both.

5.2. Datasets

This research utilizes the OpenDialKG dataset [17] to perform experiments aimed at achieving interpretable dialogue reasoning through knowledge graph path traversal using attention mechanisms. The dataset features open-ended dialogues between two speakers on specific topics, encompassing approximately 13,000 dialogues and 91,000 dialogue turns. Each dialogue is aligned with relevant knowledge graph paths, connecting the entities and relationships mentioned in the conversation. Consequently, the OpenDialKG dataset is composed of two primary parts. The first part is a parallel corpus of dialogue and knowledge graph paths, where each dialogue turn is associated with a knowledge graph path linking it to the preceding turn, with annotations provided by the participants. The second part consists of the base knowledge graph utilized for dialogue collection and experiments, which is derived from a subset of the Freebase Easy dataset [18]. In the experiments conducted in this study, the dialogues are randomly split into training (70%), validation (15%), and test (15%) sets.

5.3. Experimental Settings

The experimental setup involved an AMD Ryzen 75800H CPU, an NVIDIA GeForce RTX 3060 GPU, a Windows operating system with CUDA 11.0 as the software environment, and a PyTorch 1.7 deep learning framework. Pre-trained BERT (https://huggingface.co/google-bert/bert-base-chinese (accessed on 10 May 2024)) was used to initialize word embeddings in this study, with hidden vectors with a size of 768. The dimensionality of the knowledge graph was set at 300. During training, the SGD optimizer was employed with an initial learning rate of 0.01, a dropout rate of 0.4, a batch size of 64, and 80 iterations.

5.4. Baseline

The following baseline models were employed in the experiment to validate the effectiveness of the proposed method:

EARL [19]: Utilizes external knowledge for knowledge selection without parameterizing specific entity representations.

GPT2 [20]: Fine-tunes GPT-2 using the following settings and hyperparameters: a batch size of 16, a learning rate of 6.25×10^{-5}, and an AdamW optimizer with a linear decay scheduler.

GPT2+NPH [9]: Integrates GPT2 with NPH, where NPH refines generated responses by retrieving entities from the knowledge graph.

BART [21]: Fine-tunes the BART model with a batch size of 16, a learning rate of 3×10^{-5}, and an AdamW optimizer with a linear scheduler.

BART+NPH [16]: Applies a post-processing technique to BART, named BART+NPH, as the baseline for the experiment due to its agnostic nature towards the generative model.

KG-BART [22]: A knowledge-enhanced model based on BART that incorporates relationship information between concepts for commonsense reasoning.

RHO [16]: Achieves state-of-the-art performance on OpenDialKG by employing local knowledge, global knowledge, and knowledge reordering for knowledge selection.

6. Results and Analysis

6.1. Comparison Experiment

The knowledge selection automatic evaluation comparative experiments were conducted on the OpenDialKG test set between KMRKSM and multiple baseline models, as shown in Table 1. The automatic evaluation comparative experiments validated and assessed accuracy and appropriateness in terms of the BLEU4, ROUGE-L, FeQA, QuestEval, and Entity Coverage metrics.

Table 1. Automatic evaluation results of comparison experiment.

Model	BLEU4	ROUGE-L	FeQA	QuestEval		Entity Coverage (%)		
				RD	RF	Pre	Recal	F1
EARL	7.97	23.61	39.93	37.88	35.59	86.61	45.17	64.44
GPT2	10.27	29.59	39.40	46.86	42.07	91.61	33.26	52.30
GPT2+NPH	10.41	29.93	40.83	47.45	42.45	95.61	33.39	53.96
BART	14.45	33.33	39.00	46.97	42.75	96.99	44.96	62.87
BART+NPH	15.53	34.99	42.41	47.94	43.56	96.44	44.12	65.98
KG-BART	13.72	33.31	41.87	45.55	42.86	97.68	45.63	64.58
RHO	19.11	38.45	47.99	50.58	46.41	98.53	51.77	72.29
KMRKSM	20.63	39.54	43.04	48.41	43.84	98.70	51.80	73.01

Table 1 presents the results of automatic evaluations on the OpenDialKG test set. KMRKSM outperforms all baseline models in the BLEU4, ROUGE-L, and Entity Coverage metrics, indicating strong performance in both the accuracy and appropriateness of knowledge selection. Specifically, compared to the state-of-the-art model RHO, KMRKSM achieves a 1.53% improvement in BLEU4, a 1.09% increase in ROUGE-L, and gains of 0.17%, 0.03%, and 0.72% in precision, recall, and F1 for Entity Coverage, respectively. KMRKSM also shows competitive results compared to the strong baseline model BART+NPH. However, due to the influence of the text generation model and the lack of local knowledge, KMRKSM performs less well than the RHO model in the FeQA and QuestEval metrics, suggesting that local knowledge in dialogue statements impacts knowledge selection. Overall, the introduction of multi-hop reasoning in knowledge selection enables the model to better align selected knowledge with the dialogue context.

A human evaluation experiment on multi-hop knowledge selection was conducted on the OpenDialKG test set to assess KMRKSM and several baseline models, as shown in Table 2. The evaluation was validated using the Faith and Hallucination metrics. Specifically, 100 samples selected by each model were randomly chosen and evaluated by three different annotators to minimize potential bias. The designated annotators were required to meet the following criteria: an approval rate of 95% or higher on AI tasks and a minimum of 5000 approved AI tasks. These annotators, hailing from Australia, the United Kingdom, and the United States, were employed to evaluate the accuracy and appropriateness of knowledge selection by the models.

Table 2. Manual evaluation results of comparison experiment.

Model	Faith (%)	Hallucination (%)		
		In.	Ex.	Both
GPT2+NPH	72.67	8.67	18.00	0.67
BART+NPH	75.00	9.33	15.33	0.33
RHO	81.67	7.67	10.67	1.00
KMRKSM	82.45	7.72	10.06	1.00

As shown in Table 2, KMRKSM outperforms GPT2+NPH, BART+NPH, and RHO on the Faith metric. Specifically, compared to the state-of-the-art model RHO, KMRKSM shows

a 0.78% improvement in the Faith metric, demonstrating its superiority and effectiveness in the appropriateness of knowledge selection. On a more granular level, the issue of selecting knowledge that contradicts the dialogue history and external knowledge is more common than the issue of selecting knowledge that is difficult to verify against these sources. This trend is observed across all models. Specifically, KMRKSM reduced the selection of contradictory knowledge by 0.61% compared to RHO, indicating its ability to choose knowledge that aligns with the dialogue context. However, it increased the selection of knowledge that is challenging to verify by 0.05% compared to RHO. This slight increase is due to RHO's focus on both global and local knowledge, whereas KMRKSM primarily emphasizes global external knowledge, overlooking the impact of local knowledge on knowledge selection.

Overall, the consistency between human evaluation results and automatic evaluation results validates the feasibility and reliability of this experimental approach. Under the experimental conditions, KMRKSM not only ensures the selection of knowledge aligned with the dialogue context, but also enhances the appropriateness of knowledge expression when combined with the dialogue generation model. This confirms the effectiveness and superiority of KMRKSM in the knowledge selection task.

6.2. Ablation Experiment

To validate the effectiveness of the units constructed within the module presented in this paper, ablation experiments were conducted. The results on the OpenDialKG dataset, as shown in Table 3, demonstrate the performance on the BLEU4, ROUGE-L, FeQA, QuestEval, and Entity Coverage metrics in the absence of the multi-relational graph coding unit (MGCU) and the dynamic multi-hop reasoning unit (DMRU).

Table 3. Results of ablation experiment.

Model	BLEU4	ROUGE-L	FeQA	QuestEval		Entity Coverage		
				RD	RF	Pre	Recal	F1
KMRKSM w/o MGCU	19.23	37.67	41.67	47.31	42.45	96.75	48.56	71.45
KMRKSM w/o DMRU	18.54	35.45	40.34	45.74	40.98	95.67	47.87	70.25
KMRKSM	20.63	39.54	43.04	48.41	43.84	98.70	51.80	73.01

From Table 3, it is evident that the absence of the dynamic multi-hop reasoning unit (DMRU) results in the lowest scores across all metrics. This indicates the significant role that DMRU plays in KMRKSM, suggesting that through multi-hop reasoning with knowledge, the module can engage in more appropriate knowledge selection. On the other hand, the absence of the multi-relational graph coding unit (MGCU) leads to lower scores across all metrics compared to KMRKSM, but higher scores than those without DMRU, highlighting the supportive role of MGCU in knowledge selection. This illustrates the effectiveness of both MGCU and DMRU in KMRKSM, showcasing their crucial roles in knowledge selection.

6.3. Case Study

To further validate the effectiveness of the knowledge selection in KMRKSM, a set of dialogues was selected from the OpenDialKG dataset. The responses generated by both KMRKSM and the RHO model were compared for these dialogues, and the responses generated by these two models were documented for case analysis, as recorded in Table 4.

Table 4. An example response generated by RHO and KMRKSM.

Knowledge Graph	Dialogue Context	Real Response	RHO	KMRKSM
<The Ref, starred actors, Judy Davis>	User: Who else was in that movie? Assistant: Aside from Woody Allen, Judy Davis and Greta Gerwig also have star credits. User: What else do I know Judy Davis from?	Assistant: She also starred in "The Ref" and "My Brilliant Career."	Assistant: She starred in The Referendum and The Golden Compass.	Assistant: Judy Davis also starred in The Ref. Have you seen this movie?

As demonstrated by the example in Table 4, the RHO model produced an incorrect response, while the proposed KMRKSM generated a factually accurate reply. This indicates that KMRKSM can effectively select appropriate knowledge based on dialogue history, enabling the dialogue generation model to produce responses that are more accurate and better aligned with the input compared to the current state-of-the-art model, RHO. This confirms the effectiveness and superiority of the proposed method in knowledge selection.

7. Conclusions

To address the challenges posed by the heterogeneity between external knowledge and dialogue history in knowledge selection for dialogue generation, this paper proposes a knowledge selection module based on knowledge multi-hop reasoning, inspired by multi-hop reasoning techniques. This module is designed to enhance the appropriateness of knowledge selection in dialogue generation. Additionally, through the collaboration of the multi-relational graph coding unit and the dynamic multi-hop reasoning unit, the module differentiates between dialogue history and external knowledge during the selection process, ensuring that the selected knowledge aligns with the dialogue context.

The experimental results on the OpenDialKG dataset demonstrate that KMRKSM exhibits performance in knowledge selection that is comparable to or exceeds baseline models, validating the effectiveness and superiority of the proposed knowledge selection module. In future work, we will further explore the impact of KMRKSM on emotion prediction, aiming to enhance the quality and empathy of generated emotional responses by considering both emotional and textual dimensions.

Author Contributions: Z.M.: methodology, supervision, funding acquisition. J.L.: methodology, conceptualization, software, investigation, writing—review and editing. B.X.: data curation, formal analysis, software. K.L.: data curation, formal analysis, software. S.G.: data curation, formal analysis, software. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62166029) and the Basic Scientific Research Fund for Universities directly under the Inner Mongolia Autonomous Region (JY20220074, ZTY2024062).

Data Availability Statement: The processed data and code relate to ongoing research and therefore cannot be shared at this time.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Santhanam, S.; Hedayatnia, B.; Gella, S.; Padmakumar, A.; Kim, S.; Liu, Y.; Hakkani-Tur, D. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv* **2021**, arXiv:2110.05456.
2. Emily, D.; Stephen, R.; Kurt, S.; Angela, F.; Michael, A.; Jason, W. Wizard of Wikipedia: Knowledge-powered conversational agents. *arXiv* **2019**, arXiv:1811.01241.
3. Kurt, S.; Spencer, P.; Moya, C.; Douwe, K.; Jason, W. Retrieval augmentation reduces hallucination in conversation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3784–3803.
4. Wu, Z.Q.; Galley, M.; Brockett, C.; Zhang, Y.Z.; Gao, X.; Quirk, C.; Koncel-Kedziorski, R.; Gao, J.F.; Hajishirzi, H.; Ostendorf, M.; et al. A controllable model of grounded response generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Online Event, 2–9 February 2021; Volume 35, pp. 14085–14093.

5. Zhou, H.; Young, T.; Huang, M.; Zhao, H.Z.; Xu, J.F.P.; Zhu, X.Y. Commonsense knowledge aware conversation generation with graph attention. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, 13–19 July 2018; pp. 4623–4629.
6. Lian, R.Z.; Xie, M.; Wang, F.; Peng, J.H.; Wu, H. Learning to select knowledge for response generation in dialog systems. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 5081–5087.
7. Kim, B.; Ahn, J.; Kim, G. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Online Event, 26 April–1 May 2020.
8. Chen, S.H.; Zhang, F.; Sone, K.; Roth, D. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online Event, 2–5 June 2021; pp. 5935–5941.
9. Dziri, N.; Madotto, A.; Zaïane, O.; Bose, A.J. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2197–2214.
10. Hannah, R.; David, R.; Gaurav, S.T.; Dipanjan, D. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online Event, 2–5 June 2021; pp. 704–718.
11. Ma, T.H.; Zhang, Z.; Rong, H.; Al-Nabhan, N. SPK-CG: Siamese Network based Posterior Knowledge Selection Model for Knowledge Driven Conversation Generation. *ACM Trans.* **2023**, *22*, 1–16. [CrossRef]
12. Shikhar, V.; Soumya, S.; Vikram, N.; Partha, P.T. Composition-based multirelational graph convolutional networks. In Proceedings of the 8th International Conference, ICLR 2020, Online Event, 26 April–1 May 2020.
13. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the 42th Proceedings of Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004; pp. 74–81.
14. Durmus, E.; He, H.; Diab, M. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online Event, 5–10 July 2020; pp. 5055–5070.
15. Thomas, S.; Paul-Alexis, D.; Sylvain, L.; Benjamin, P.; Jacopo, S.; Alex, W.; Patrick, G. Questeval: Summarization asks for fact-based evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6594–6604.
16. Ji, Z.; Liu, Z.; Lee, N.; Yu, T.Z.; Bryan, W.; Zeng, M.; Pascale, F. RHO (ρ): Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding. In Proceedings of the 61st Annual Meeting of The Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 4504–4522.
17. Moon, S.; Shah, P.; Kumar, A.; Subba, R. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July 28–2 August 2019; pp. 845–854.
18. Bast, H.; Buchhold, B.; Buchhold, B.; Haussmann, E. Easy access to the Freebase dataset. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014.
19. Zhou, H.; Huang, M.; Liu, Y.; Chen, W.; Zhu, X. EARL: Informative knowledge-grounded conversation generation with entity-agnostic representation learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2383–2395.
20. Alec, R.; Jeffrey, W.; Rewon, C.; David, L.; Dario, A.; Ilya, S. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
21. Lewis, M.; Liu, Y.; Goyal, N.; Marjan, G.; Abdelrahman, M.; Omer, L.; Veselin, S.; Luke, Z. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online Event, 5–10 July 2020; pp. 7871–7880.
22. Liu, Y.; Wan, Y.; He, L.F.; Peng, H. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online Event, 2–9 February 2021; Volume 35, pp. 6418–6425.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.