IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 2 November 2022, accepted 8 December 2022, date of publication 12 December 2022,
date of current version 21 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228964

RESEARCH ARTICLE

Retrieval-Augmented Response Generation for Knowledge-Grounded Conversation in the Wild

YEONCHAN AHN[1], SANG-GOO LEE[2], (Member, IEEE), JUNHO SHIM[3], (Senior Member, IEEE), AND JAEHUI PARK[2,4]

[1]Posicube Inc., Seoul 06234, South Korea
[2]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, South Korea
[3]Department of Computer Science, Sookmyung Women's University, Seoul 04310, Republic of Korea
[4]Department of Statistics, University of Seoul, Seoul 02504, South Korea

Corresponding author: Jaehui Park (jaehui@uos.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) by the Ministry of Science, ICT and Future Planning through the Basic Science Research Program under Grant 2020R1F1A1075952.

ABSTRACT Users on the internet usually have conversations on interesting facts or topics along with diverse knowledge from the web. However, most existing knowledge-grounded conversation models consider only a single document regarding the topic of a conversation. The recently proposed retrieval-augmented models generate a response based on multiple documents; however, they ignore the given topic and use only the local context of the conversation. To this end, we introduce a novel retrieval-augmented response generation model that retrieves an appropriate range of documents relevant to both the topic and local context of a conversation and uses them for generating a knowledge-grounded response. Our model first accepts both topic words extracted from the whole conversation and the tokens before the response to yield multiple representations. It then chooses representations of the first N token and ones of keywords from the conversation and document encoders and compares the two groups of representation from the conversation with those groups of the document, respectively. For training, we introduce a new data-weighting scheme to encourage the model to produce knowledge-grounded responses without ground truth knowledge. Both automatic and human evaluation results with a large-scale dataset show that our models can generate more knowledgeable, diverse, and relevant responses compared to the state-of-the-art models.

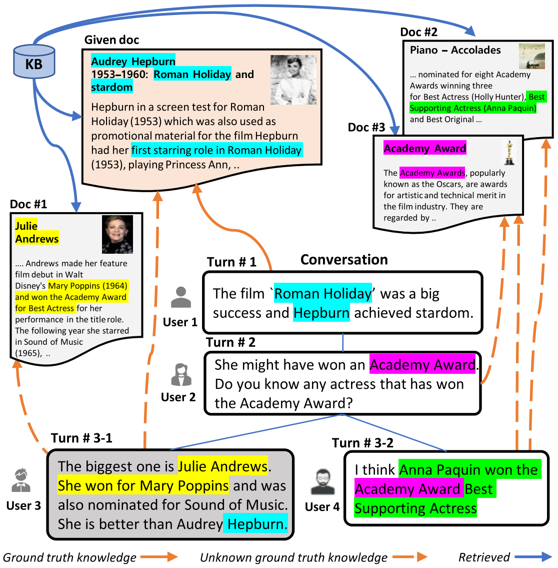
INDEX TERMS Conversation, knowledge-grounded conversation, knowledge retrieval.

I. INTRODUCTION

A knowledge-grounded conversation (KGC) [1] is a task of generating informative responses based on external knowledge and conversation context. Many datasets based on the knowledge of various sources, such as Wikipedia [2] and movie reviews [3], have been proposed for training KGC models by using conversations between crowdsourced workers.

Most conversation datasets contain a significant proportion of knowledge-grounded utterances that are grounded in one or more facts. In specific, an utterance is considered to be knowledge-grounded when there is an association (as defined by the model) between the utterance and one or more documents in the knowledge base (KB). Fig. 1 shows an example of KGC on the internet where multiple users naturally exchange information by referring to various documents without providing ground truth (GT) knowledge. In this study, we define GT knowledge as the documents that each user of the responses refers to. Generating knowledge-grounded responses under such a situation is challenging because each user can refer to multiple documents on various topics simultaneously as shown in Turn # 3-1 or # 3-2 without GT knowledge sharing. We focus on a natural conversation scenario on the internet where a document which we call *given document* is provided to all the users as main topic of the conversation and they can retrieve more documents from a KB while conversing. For example, the user of Turn # 3-1 responds by referring to both the given document on "Audrey Hepburn" and another retrieved document on "Julie Andrews."

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.



Audrey Hepburn
1953–1960: Roman Holiday and stardom

Hepburn in a screen test for Roman Holiday (1953) which was also used as promotional material for the film Hepburn had her first starring role in Roman Holiday (1953), playing Princess Ann, ..

Julie Andrews

.... Andrews made her feature film debut in Walt Disney's Mary Poppins (1964) and won the Academy Award for Best Actress for her performance in the title role. The following year she starred in Sound of Music (1965), ..

Piano – Accolades

... nominated for eight Academy Awards winning three for Best Actress (Holly Hunter), Best Supporting Actress (Anna Paquin) and Best Original ...

Academy Award

The Academy Awards, popularly known as the Oscars, are awards for artistic and technical merit in the film industry. They are regarded by ..

User 1: The film `Roman Holiday' was a big success and Hepburn achieved stardom.

User 2: She might have won an Academy Award. Do you know any actress that has won the Academy Award?

User 3: The biggest one is Julie Andrews. She won for Mary Poppins and was also nominated for Sound of Music. She is better than Audrey Hepburn.

User 4: I think Anna Paquin won the Academy Award Best Supporting Actress

FIGURE 1. Knowledge-grounded conversation between users on the internet, where User 1 begins the conversation by sharing a document that is the main topic of the conversation, and others share their opinions while omitting the documents in which they referred to the so-called ground truth knowledge.

Most existing methods, such as [4], consider only one document when generating a response (as shown in Turn # 2). Such methods may limit the content of the responses, making them difficult to deliver diverse and interesting information. By contrast, the retrieval-augmented KGC models such as [5] can utilize multiple documents simultaneously (as shown in Turn # 3-2). These models are usually based on neural networks and are composed of three modules responsible for understanding a conversation, knowledge selection/retrieval, and response generation.

Retrieval-based models can generate responses composed of multiple keywords from diverse documents. However, they may lack the capability to generate responses using the given documents that people are interested in because they consider only a limited number of tokens that come immediately before the response, which we call the *local context*. Thus, the topic coverage of the model responses is limited to the local context, and thus it incurs the risk of deviating from the main topic of the conversation if the conversation becomes longer.

To alleviate this limitation, we aim at generating interesting responses grounded on both the given document and other relevant documents (as shown in Turn # 3-1). We propose a novel hybrid KGC framework that combines a retrieval-based neural KGC model and several external modules that inject topic information of a conversation into the neural model. We utilize the external module, TextRank [6], in the proposed framework to extract keywords from the conversation and inject them into the neural network to handle cases in which the model deviates from the main topic of the conversation. When the length of the conversation increases, utterances in the distance can affect the response. Therefore, it is important to find the knowledge relevant to the response by identifying the main topic of the conversation, as in the previous studies [7], [8], [9] attempt to solve this issue by considering the flow of the conversation. Inspired by this intuition, we develop an end-to-end (E2E) retrieval-augmented KGC model, an implementation of our proposed framework, that considers both the topic of the whole conversation and the local context to compose an appropriate range of relevant documents. Our conversation encoder accepts a fixed number of tokens before the response (local context), and the external module extract topic keywords from the entire conversation. It then chooses representations of the first N token and ones of keywords from the conversation and document encoders and the following two matching modules compare the two groups of representation from the conversation with those groups of the document, respectively.

Furthermore, we propose a new data-weighting scheme that encourages our model to generate grounded and informative responses to alleviate the data skew of KGCs in the wild. Here, the data skew of KGCs in the wild indicates that the ratio of non-knowledge-grounded responses is relatively high. Our data-weighting scheme uses the estimated groundness of the responses as an instance weight. Using this method, we expect that the model can be trained to be biased toward the knowledge-grounded responses of the training set. We first introduce an estimate of the relevance of the response to GT knowledge using word matching-based similarity. We also estimate the uniqueness of the response and use it as an auxiliary estimate to deal with cases in which our similarity-based groundness cannot be captured.

Automatic and human evaluation results show that our model achieves a state-of-the-art performance in terms of the utilization of external knowledge (i.e., grounding) and the quality of the response. The automatic evaluation results regarding the grounding show that our models produce responses that are more reflective of the relevant documents in the KB than the baseline models. In addition, the results indicate that our model yields responses that are more relevant to the context of the conversation and contain more diverse words than the other models. The human evaluation results show that our model outperforms other models in terms of knowledgeableness, interestingness, and relevance.

Our contributions are summarized as follows:

1) We propose a novel retrieval-augmented KGC model that considers both the conversation topic and local context.
2) We propose a novel data-weighting scheme dedicated to an E2E retrieval-augmented model.
3) We empirically show that our models achieve a state-of-the-art performance on a large-scale benchmark dataset.

II. TASK

We focus on a knowledge-grounded response generation task in which the KGC model uses KB as a knowledge source. A model should generate responses grounded on

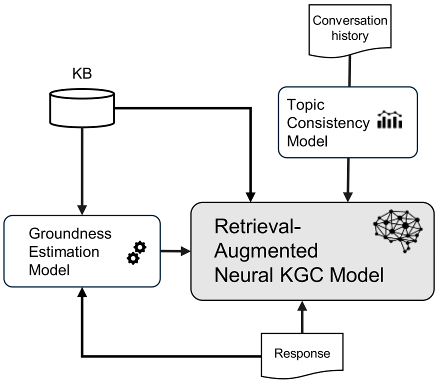

Conversation history
KB
Topic Consistency Model
Groundness Estimation Model
Retrieval-Augmented Neural KGC Model
Response

FIGURE 2. Proposed KGC framework. Retrieval-augmented neural networks with external models for considering the topic consistency and groundness of the responses in the KGCs.

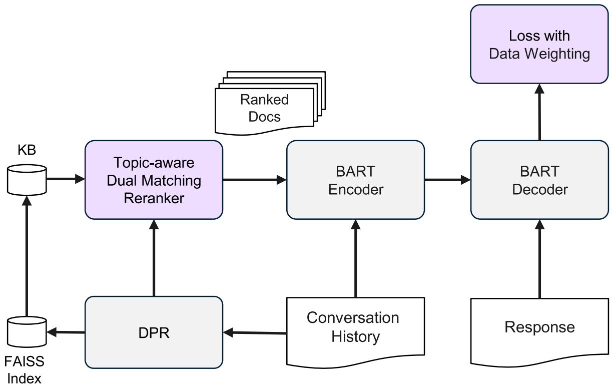


FIGURE 3. Our backbone architecture. Our model is based on RAG token [10], i.e., a topic-aware dual matching reranker and data-weighting scheme, with two significant changes.

factoid documents relevant to the conversation history. More specifically, we are given a KB containing documents without any links between a knowledge-grounded response and corresponding documents, that is, GT knowledge. Formally, the model is given a conversation history of turns $X = (x_0, \ldots, x_M)$ and KB of document KB. With conversation history X, the model needs to generate a natural language response y that is relevant to the documents in the KB and conversation history. The KB consists of all of the pages in Wikipedia and given documents provided in the conversations.

III. METHOD

We propose a novel hybrid framework for KGC, which is composed of a retrieval-augmented neural KGC model, topic consistency model, and groundness estimation model as shown in Fig. 2. The goal of our framework is to train the neural KGC model in the direction of the model designer's guide embedded in the external modules. We introduce two principles in designing the external modules as follows:

1) The probability of the response being generated for a conversation history is proportional to the relevance to the given document rather than any document referenced within the local context.
2) The probability of the response being grounded in documents is proportional to the number of common words between those documents and the response.

The first principle is associated with the situation where people tend to make responses relevant to the main topic, i.e. the given document, over the whole conversation, although they are provided information from other documents. By adopting this principle in the design of our knowledge retriever/reranker, we can endow the ability to present knowledge relevant to the main topic of the conversation in addition to the local context if the time when the given document is provided is given to the model. The second one is based on the observation that responses conveying knowledge often contain keywords from relevant documents in the KB. Based on the second principle, we design a groundness estimation model to calculate the degree of groundness of the response of each training example to guide the model to assign more weights to grounded responses.

In the following, we describe our KGC model in detail. The proposed model comprises a backbone model, knowledge retriever/reranker, and loss with a data-weighting scheme. The backbone model generates responses based on the conversation inputs and retrieved documents. The knowledge retriever/reranker finds multiple documents from KB. The loss with data-weighting scheme provides the training method for the model.

A. BACKBONE MODEL

Fig. 3 shows our backbone model architecture, where our proposed modules in this paper are highlighted in violet. We choose the retrieval-augmented generation (RAG)-token model [10] as our backbone because the intuition of the model is the same as our assumption, which is that each token can be generated based on multiple documents. RAG-token is an answer generation model equipped with a neural retriever trained E2E for open-domain QA tasks. While training, it conducts a K-NN document search from large document KB by utilizing the FAISS index [11]. The training objective of the RAG token is shown as follows:

$$P(y|x) = \prod_{i}^{L} \sum_{z \in topk(p(\cdot|x))} p(z|x)p(y_i|x, z_i, y_{i:i-1}) \tag{1}$$

where x is a question, z is a document, y is a answer of the question, L is the number of tokens in the answer, and i is a index of answer token. The dense passage retriever (DPR) [12] calculates scores of the $|topk|$ documents retrieved by the FAISS index, and the second term $p(y_i|\cdot)$ is a probability of generating each token y_i based on different documents z_i, the input x, and $y_{i:i-1}$.

We replace the input and output of the RAG model with the ones of the KGC setup and utilize them as the basis of our method. x will be the conversation history; z will be the external document, and y will be the response to x. The $p(z|x)$

of our model is computed by using a weighted sum of the score from DPR and the score from the reranker described in the followings.

B. TOPIC-AWARE DUAL MATCHING FOR KNOWLEDGE RERANKER

Fig. 4 shows our topic-aware dual matching (TADM) reranker integrated with the DPR, which comprises a conversation encoder, a document encoder, dual matching layers, a topic keyword extractor, a keyword checker, and a scorer layer. The conversation encoder encodes local context and topic keywords from a conversation history to a sequence of vectors. Then retriever queries the FAISS index to find c candidate documents using the representation corresponding to the [CLS] token. The document encoder encodes retrieved documents. The dual matching layers, like the shallow Transformer head as in [13], conduct fine-grained matching between conversation and documents to yield a score of each using linear layers. Then, the scorer layer outputs the final score of the reranker, which is the weighted sum of the scores from the two matching layers followed by the linear layers.

1) CONVERSATION ENCODER

Our conversation encoder, i.e. the same module as the context encoder of DPR, encodes both tokens before the response and the topic keyword of the whole conversation to retain top-k documents that consider an appropriate range of the relevant topics of the conversation. Specifically, we use the turns right after the given document is provided for topic keyword extraction, because they tend to discuss the content of the given document in conversation. To implement this strategy, we use an external topic keyword extraction module TextRank [6] to recognize important words from the turns and use them as additional input with the tokens before the response. TextRank algorithm is advantageous in that it is trained in an unsupervised way because it is based on the PageRank algorithm [14] on the word graph whose edge weight is calculated using co-occurrences of the words. By adopting TextRank, we can use the keywords without a manual collection.

Then, we assign different segment embeddings to each of these two input types, turns before the response, and topic keywords. Finally, we add them with token embeddings and position embeddings.

2) DUAL MATCHING LAYERS

Prior works [5] and [15] pointed out that the method using only a vector at [CLS] lacks the interaction between the two inputs, which they considered as one of the main causes of the poor performance. Following this line of work, we propose a method, called dual matching, to match the conversation with the documents from each encoder along with multiple representations. Our method first separates the representations from the conversation/document encoder and matches the two groups of representation from the conversation with those groups of the document. Each matching layer is composed of two layers of the Transformer encoder, which consists of the Matching Layer for Local Context and Matching Layer for Keywords. The former matches representations of the first N tokens in the conversation history, and the latter matches representations of the keywords selected by our keyword checker. Our keyword checker determines the word as a keyword if the word is tagged as one among the named entity, non-stopwords, or keywords of TextRank [6], using an external natural language processing tool.

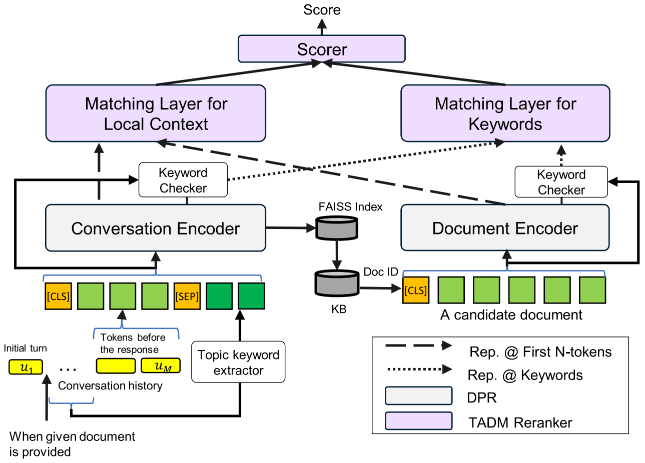


FIGURE 4. Retrieving documents with topic-aware dual matching reranker: The conversation encoder encodes tokens before the response with topic keywords, and then, it retrieves Top-*k* candidate docs from KB. The Matching Layer for Local Context accepts representations on the top of the first N tokens of the inputs. The Matching Layer for Keywords takes representations selected by the keyword checker.

Our intuitions behind the dual matching layers are as follows. First, we expect the Matching Layer for Local Context to match representations having global features related to the current context and the document. Second, we expect that representations of keywords having information relevant to informative or crucial words lead us to more accurate retrieval rather than using representations of other words (e.g., pronoun and article). The analysis on self-attention trained by BERT [16] supports our intuitions. In that work, they visualized the attention scores on some meaningful words, for example, scores on noun phrases are higher than stopwords'.

C. DATA-WEIGHTING SCHEME FOR RETRIEVAL-AUGMENTED GENERATION MODELS

In the KGCs on the internet, users often do not provide GT knowledge that they refer to (shown as unknown GT knowledge in Fig. 1) and make casual responses that are not grounded. To handle this issue, we consider two metrics, the groundness [17] and informativeness [18] to evaluate the quality of the responses. The groundness corresponds to the aforementioned second principle because it is defined as the relevance between a response and a document. Also, we assume that a response is called informative if it has words that rarely occur in the conversation. For example, the term 8,849 in an utterance grounded in some knowledge "The height of Mt. Everest is 8,849 m" rarely occurs in our conversation.

Algorithm 1 Training Algorithm

```
procedure Train($D, M$) ▷ Train model $M$ with dataset $D$
    Calculate IDF of words in turns ▷ Before training
    for $(x, y)_j \in D$ do
        $Z \leftarrow$ Top-K docs relevant to $(x, y)_j$ from KB
        $\hat{z} \leftarrow argmax_{z \in Z}\mathrm{BLEU}(z, y)$
        $w_j \leftarrow gmean(\mathrm{BLEU}(\hat{z}, y), \text{IDF-W}(y))$
    end for
    while $M$ Converges do ▷ Training
        $B \subset \mathrm{shuffle}(D)$
        $w'_b \leftarrow w_b / \sum_{b \in B} w_b$
        Multiply loss of $b$-th example with $w'_b$
        Update Model $M$ with the loss
    end while
end procedure
```

As a result of the above discussion, we propose a novel instance weighting scheme for retriever-augmented KGC models. Algorithm 1 shows the training method using our proposed weighting scheme where the weight of j-th example w_j is the geometric mean of $\mathrm{BLEU}(\hat{z}, y)$ and IDF-W(y). Differently from machine reading comprehension style KGC models such as [4], we do not assume that responses are solely grounded on the given document. Therefore, we retrieve the documents from KB by using the DPR and a query $(x, y)_j$ and use the similarity value as a surrogate of the weight for groundness. Specifically, we use the sentence-level BLEU score for estimating groundness, which we confirmed that it shows better results than the similarity measures based on distributional representations. For informativeness, we use the length-penalized summation of the term's inverse document frequency (IDF) shown in Eq. 2, where we define a document of the IDF is a turn in a conversation. We calculate the weighting score by dividing the sum of IDFs by the length as in [19] to avoid generating only long responses.

$$\text{IDF-W}(y) = \sum_{w_i \in y} IDF(w_i) \frac{(4+1)^{\alpha}}{(4+length)^{\alpha}} \tag{2}$$

Here, α controls the strength of the length normalization. The numbers 4 and 1 are constant values responsible for dealing with weight imbalance between short responses and long ones. When $\alpha = 0$, the weight falls back to the pure sum of IDFs of words.

IV. EXPERIMENT SETTINGS

A. DATASET

We used the Reddit KGC dataset, CbR [17]. The dataset consists of conversation threads extracted from Reddit.com. Each conversation is linked to a given document shared at the beginning. We provided the given document with KGC models except for the retrieval-based models. The dataset has 2.8M, 5.9k, and 13k instances for the training, validation, and testing, respectively. For the evaluation, we used a test set of 2208 instances built by [20], for which 6 responses are available. Additionally, we built another test split that we call internet-grounded split by manually choosing responses that contain knowledge of webpages on the internet such as Wikipedia and news stories. The resulting split includes 204 instances having only one response for each conversation history. We used this test split to evaluate the model's capability of utilizing knowledge of the documents in KB.

B. COMPETING MODELS

- **MemNet** [1]: A memory network designed for KGC. The model uses a memory network to store knowledge.
- **CMR** [17]: A KGC model based on the state-of-the-art machine reading comprehension model [21]. It is trained with a data-weighting scheme to encourage response grounding on the given document.
- **CMR-F** [17]: A model that omits the document reading component of the original CMR model.
- **RAM-T** [4]: A state-of-the-art model in the CbR dataset. The model is trained to generate the memory to resemble the memory induced by the teacher network which accepts response, conversation context, and the given document.
- **BART** [22]: A Transformer-based Seq2seq model pre-trained by reconstructing the original input text which is corrupted with noising functions.
- **RagToken** [10]: A state-of-the-art model on open-domain QA tasks, which is our base model. A detailed description of this model is presented in section III-A.
- **DPRThenPoly** [5]: It shares the same backbone of RagToken, but it reranks documents retrieved by the DPR using Polyencoder [15].
- Our models: **TADM** denotes our KGC model using the TADM reranker. **TADM+IDF-Bw** denotes our model trained with the data-weighting scheme based on BLEU and IDF. **TADM+Bw** and **TADM+IDFw** denote our models with the data-weighting scheme based on only BLEU and IDF, respectively.

C. IMPLEMENTATION DETAILS

For MemNet, CMR, and CMR-F, we used the model provided by [17]. We implemented RAM-T by ourselves because the author did not provide their codes.[1] The above models were trained with hyperparameters that the authors recommended. For Transformer-based models including BART, RagToken, DPRThenPoly, and our models, we used a common codebase[2] provided by [5]. We used the common architectures and

[1] Though [4] said that the authors would make the code available to the public, they have not yet uploaded their codes in the repository https://github.com/tianzhiliang/RAM4CbR. (Accessed: June 24th, 2022)

[2] https://github.com/facebookresearch/ParlAI/tree/main/parlai/agents/rag (Accessed: June 24th, 2022)

parameters of `BART-large`[3] and `DPR-multiset`[4] for the modules. For the Transformer-based models, we used the Adam optimizer for training, with an initial learning rate of 5e-4 and early-stopped when the perplexity with a validation set was not improved with patience 5 with a 13 batch size. During training, all responses were truncated to have a maximum length of 30, and the maximum number of tokens before the response and document were set to 60 and 100, respectively. The number of layers for the Matching Layer for Local Context and Matching Layer for Keywords were set to 2. We utilized $N = 8$ representations for the Matching Layer for Local Context and 16 representations for the Matching Layer for Keywords from each encoder. We used a turn after the given document for topic keyword extraction. For decoding, we used the top-k ($k = 10$) random sampling strategy [23]. We set the number of retrieved documents c and $topk$ to 3 and 5 in the training and testing, respectively. For our model, we implemented our topic keyword extractor and keyword checker with Spacy NLP tools. Also, we set α as 0.5.

D. AUTOMATIC EVALUATION

We evaluated the models in terms of four metrics as follows:

1) GROUND-TO-GIVENDOC

This metric is proposed in [17]. It measures the systems' ability to exploit knowledge from the given document by counting the number of words that co-occurred in the given document and the response. If the words also occurred in the conversation history, they are not counted.

2) GROUND-TO-INTERNET

We propose to evaluate the model's ability to utilize knowledge in the KB using the internet-grounded split. For this metric, we use the similarity between one of three responses generated by a model and the human response. We define the metrics for Ground-to-Internet based on several similarity measures as follows:

$$\mathrm{G}_s = \frac{1}{|D_{test}|} \sum_{l=1}^{|D_{test}|} max_{i\in[1,3]} s(y_l, y_{l,i}), \tag{3}$$

where s is chosen from $\{BLEU, NIST, F1\}$; D_{test} is the test split; y_l is lth the human response, and $y_{l,i}$ is one of the generated responses. BLEU and NIST are sentence-level MT metrics, respectively. F1 is the Unigram F1 score.

3) RELEVANCE

We use three metrics measuring the similarity between model response and human responses, i.e., BLEU [24], METEOR [25], and NIST [26], as the surrogates of the relevance between the response and given conversation history. NIST is a variant of BLEU that weighs more informative n-gram.

4) DIVERSITY

We use the system-level diversity metrics, Ent-n [27] and Div-n [18]. Ent-n is the entropy of the n-gram count distribution. Div-n is the number of distinct n-grams in the generated responses divided by the total number of generated tokens.

Note that the metrics of Ground-to-GivenDoc, Relevance, and Diversity are system-level (or corpus-level) scores accumulated for all the responses and their respective references. Thus, we conducted statistical hypothesis tests (Student's paired t-test) for the Ground-to-Internet metric only.

E. HUMAN EVALUATION

We conducted a qualitative evaluation with human annotators from English-speaking countries using Amazon Mechanical Turk. We randomly sampled 200 examples from the test set of 2,208 instances and asked three distinct human annotators to choose a preferred response in terms of **Relevance** and **Interestingness** among two randomly ordered responses of the competing models and ours. We chose the five most competitive models in automatic evaluation metrics, i.e., RAM-T, BART, RagToken, DPRThenPoly, and Human. While comparing the responses, the annotators were additionally asked to agree or disagree with a statement regarding **Knowledgeableness** of each response. The definitions for the measures of response quality are as follows:

- **Relevance**: Is the response relevant to the given conversation history?
- **Interestingness**: Does the response attract your attention because it contains exciting or unusual content?
- **Knowledgeableness**: Does the response provide useful knowledge, not provided in the conversation?

We conducted statistical hypothesis tests for each quality metric. For this, we convert the answer for each question as 1 if judges regard the response as relevant, interesting, or knowledgeable else 0.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. AUTOMATIC EVALUATION RESULTS

Table 1 shows the results of the automatic evaluation of the Relevance, Ground-to-GivenDoc, and Diversity. Our TADM+Bw outperforms the best model RAM-T by 0.13 in terms of Ground-to-GivenDoc. The F1 score of TADM+Bw is increased by 0.13 from the result of RAM-T, which exceeds the human score 0.88. TADM+Bw has improved NIST by 0.01, BLEU by 0.07, and METEOR by 0.19 compared to the best scores in Relevance. In terms of Diversity, our TADM+Bw improved the Ent-*4* by 0.07 and Div-*1* by 0.01.

Notice that our models differ from the backbone model RAG-token in that our model has the reranker and is trained with our data-weighting scheme. We think that the

[3] https://huggingface.co/facebook/bart-large (Accessed: June 24th, 2022)

[4] https://parl.ai/docs/agent_refs/rag.html#rag-options (Accessed: June 24th, 2022)

TABLE 1. Automatic evaluation results in terms of Ground-to-GivenDoc, Relevance, and Diversity. Our best model (TADM+Bw) outperforms the baseline models in terms of grounding considerably and also improves the Diversity and Relevance slightly. Len denotes the length of the generated responses. The best figures among the baselines are underlined.

Model	Ground-to-GivenDoc			Relevance			Diversity			Len
	Precision	Recall	F1	NIST	BLEU	METEOR	Ent-*4*	Div-*1*	Div-2	
Human	3.47%	0.51%	0.88%	2.65	3.13%	8.31%	10.44	0.17	0.67	18.8
MemNet	0.85%	0.11%	0.20%	2.17	0.98%	7.19%	9.79	0.03	0.21	15.4
CMR-F	0.77%	0.10%	0.18%	2.25	0.98%	7.39%	9.79	0.03	0.20	16.0
CMR	1.15%	0.15%	0.27%	2.26	1.30%	7.44%	9.85	0.04	0.25	15.4
RAM-T	2.17%	0.47%	0.77%	2.08	0.96%	8.58%	10.27	0.03	0.17	28.6
BART	2.29%	0.34%	0.60%	2.82	2.04%	8.68%	10.24	0.10	0.43	18.0
RagToken	2.34%	0.34%	0.59%	2.85	2.27%	8.60%	10.21	0.10	0.45	17.3
DPRThenPoly	2.41%	0.35%	0.62%	2.82	2.40%	8.68%	10.24	0.10	0.46	17.5
TADM+Bw	**3.34%**	**0.52%**	**0.90%**	**2.86**	2.47%	8.87%	10.34	**0.11**	0.46	19.1
TADM+IDFw	2.65%	0.44%	0.75%	2.75	2.12%	**9.08%**	**10.39**	0.10	0.46	20.0
TADM+IDF-Bw	3.17%	0.49%	0.85%	2.84	**2.57%**	8.86%	10.32	**0.11**	**0.47**	19.1

TABLE 2. Automatic evaluation results in terms of Ground-to-Internet. Our models outperform the competitive baselines in terms of groundness with retrieved documents in KB based on term-matching similarities. The best figures among the baselines are underlined. Statistically significant improvements ($p < 0.01$, $p < 0.05$, and $p < 0.10$) over the best baseline model are marked with *, **, and ***, respectively.

Models	G_{BLEU}	G_{NIST}	G_{F1}
BART	11.67%	53.40%	15.30%
RagToken	11.64%	51.80%	14.61%
DPRThanPoly	11.32%	50.27%	14.95%
TADM+Bw	**12.62%***	**62.89%***	**16.09%*****
TADM+IDFw	12.29%**	58.31%**	15.67%
TADM+IDF-Bw	12.37%*	57.84%**	15.54%

TABLE 3. Human evaluation results in terms of Relevance, showing preferences (%) for our model (TADM+Bw) vs. the baselines. Statistically significant improvements ($p < 0.01$ and $p < 0.05$) over each baseline are marked with * and **, respectively.

Relevance				
Our best model		Neutral	Baseline	
TADM+Bw	**52.0%***	15.3%	32.7%	RAM-T
TADM+Bw	**42.3%***	26.2%	31.5%	BART
TADM+Bw	**41.2%**	21.5%	37.3%	RagToken
TADM+Bw	**36.3%**	30.7%	33.0%	DPRThenPoly
TADM+Bw	**48.2%****	11.2%	40.7%	Human

reason for the higher performance of our models in terms of Grounding-to-GivenDoc than RagToken's is that our TADM reranker retrieves documents relevant to the given document and the decoder then uses the retrieved results for generating responses that contain keywords relevant to the given document. We presented the evidence of this reasoning in the ablation study V-C. In addition, we reason that better utilization of content from various documents in KB leads to the enhancement in Diversity metrics.

Table 2 shows that our models produce quality responses in terms of Ground-to-Internet metrics based on all the similarity measures. As we argued that in section III-C, TADM-IDFw shows better results compared to the other retrieval-based models', which shows the effectiveness of the data-weighting scheme based on IDF.

TABLE 4. Human evaluation results in terms of Interestingness, showing preferences (%) for our model (TADM+Bw) vs. the baselines. Statistically significant improvements ($p < 0.01$) over each baseline are marked with *.

Interestingness				
Our best model		Neutral	Baseline	
TADM+Bw	**41.3%***	27.7%	31.0%	RAM-T
TADM+Bw	**39.5%***	30.5%	30.0%	BART
TADM+Bw	**40.5%**	20.3%	39.2%	RagToken
TADM+Bw	**35.2%**	32.3%	32.5%	DPRThenPoly
TADM+Bw	41.2%	16.7%	**42.2%**	Human

TABLE 5. Human evaluation results in terms of Knowledgeableness, showing the percentage (%) of responses that human annotators considered knowledgeable. Statistically significant improvements ($p < 0.01$ and $p < 0.05$) over each baseline are marked with * and **, respectively.

Knowledgeableness			
Our best model		Baseline	
TADM+Bw	**48.5%****	42.3%	RAM-T
TADM+Bw	**53.0%***	46.8%	BART
TADM+Bw	**54.7%**	52.0%	RagToken
TADM+Bw	**49.2%**	47.3%	DPRThenPoly
TADM+Bw	48.5%	**49.7%**	Human

B. HUMAN EVALUATION RESULTS

The results in terms of Relevance, Interestingness, and Knowledgeableness are summarized in Tables 3, 4, and 5, respectively. Tables 3 and 4 present comparison results between our best model and the baselines. Table 3 shows that the human annotators judged that our model produces responses relevant to the conversation history more than the other baselines, including the human responses. The reason for these results may be that people on the internet tend not always reply with highly context-relevant responses; instead, they interact with responses that can fit various conversation contexts, such as "I love it." Table 4 shows that our model also outperforms the other models by 10.3 to at least 1.3 in terms of Interestingness. Table 5 shows our model produces knowledgeable responses more frequently than the other models by 6.2 to at least 1.9. Inter-rater agreements

TABLE 6. Results of the ablation study. w/o topic keywords denotes our model that does not use topic keywords as the conversation encoder input; w/o dual matching denotes our model that has a single matching layer; w/o data-weighting denotes our model trained without our data-weighting scheme.

Model	Retrieval	Ground-to-GivenDoc	Relevance			Diversity	
	Recall@5	F1	NIST	BLEU	METEOR	Ent-*4*	Div-*2*
TADM+Bw	**14.07%**	**0.90%**	2.86	**2.47%**	8.87%	10.34	**0.46**
w/o topic keywords	5.23%	0.81%	2.83	2.46%	8.95%	10.35	**0.46**
w/o dual matching	9.98%	0.87%	**2.87**	2.42%	**9.06%**	**10.37**	**0.46**
w/o data-weighting	0.10%	0.63%	2.85	2.15%	8.57%	10.22	**0.46**

measured by Fliess' Kappa of Relevance, Interestingness, and Knowledgeableness were 0.34 ("fair"), 0.31 ("fair"), and 0.31 ("fair") respectively.

C. ABLATION STUDY

Table 6 shows the impact of each module from ours. Recall@5 is the accuracy of the retriever when we assume that the GT knowledge is the given document. Therefore, the results show that Recall@5 is proportional to the Ground-to-GivenDoc, of which the Pearson coefficient is 0.94. We can notice that our data-weighting scheme is the essential factor for training the retriever E2E. Recall@5 gets almost zeros without the weighting scheme. Utilizing the topic keywords and dual matching improves Recall@5 by 8.84 and 4.09, respectively, of which the Ground-to-GivenDoc metric F1 improved by 0.09 and 0.03. In terms of Relevance, the BLEU score decreases by 0.32 without the data-weighting scheme. In addition, the model without dual-matching improved Div-*2*. We presume that it is the result of utilizing the content of other documents rather than the given documents.

D. CASE STUDY

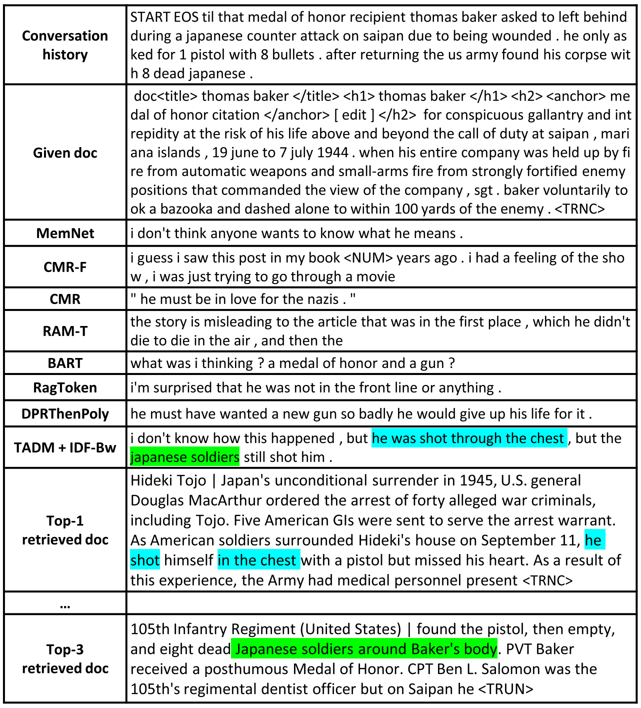
We examine a representative case of the models to learn more about their properties. Fig. 5 shows that our TADM+IDF-Bw can retrieve the document relevant to the given documents and other documents and generates a response that contains the words exactly matched with the words in the retrieved documents. It uses multiple relevant documents (top-1 and top-3 documents) at once. Specifically, the top-3 retrieved document is on the same topic as the given document (an event in which Thomas Backer participated in a battle). Although the top-1 document is not about the event referred to in the conversation history, the model can generate a context-relevant response based on that document.

Compared to our model, other models either do not provide new context-relevant information or generate responses that are less relevant to the conversational history. For example, DPRThenPoly generates a response that is relevant to this context but does not have any useful knowledge. Rather than that, it delivers a thought saying that 'he would give up his life for a new gun' that does not have any ground in the given document.

We provide additional examples to help the readers understand in the appendix.

E. DISCUSSION

As an advantage of our TADM reranker, it maintains topical consistency in a long conversation without processing all tokens in the conversation history, in contrast with recent works on Transformer architecture such as Longformer [28] whose aim is to accept a lengthy input. Our experiment results show that only extracting keywords is sufficient in finding the main reference document. In addition, we observed that our data-weighting scheme is a workhorse for training the E2E retrieval-augmented model in noisy real-world conversations.



Conversation history	START EOS til that medal of honor recipient thomas baker asked to left behind during a japanese counter attack on saipan due to being wounded . he only asked for 1 pistol with 8 bullets . after returning the us army found his corpse with 8 dead japanese .
Given doc	doc<title> thomas baker </title> <h1> thomas baker </h1> <h2> <anchor> medal of honor citation </anchor> [edit] </h2> for conspicuous gallantry and intrepidity at the risk of his life above and beyond the call of duty at saipan , mariana islands , 19 june to 7 july 1944 . when his entire company was held up by fire from automatic weapons and small-arms fire from strongly fortified enemy positions that commanded the view of the company , sgt . baker voluntarily took a bazooka and dashed alone to within 100 yards of the enemy . <TRNC>
MemNet	i don't think anyone wants to know what he means .
CMR-F	i guess i saw this post in my book <NUM> years ago . i had a feeling of the show , i was just trying to go through a movie
CMR	" he must be in love for the nazis . "
RAM-T	the story is misleading to the article that was in the first place , which he didn't die to die in the air , and then the
BART	what was i thinking ? a medal of honor and a gun ?
RagToken	i'm surprised that he was not in the front line or anything .
DPRThenPoly	he must have wanted a new gun so badly he would give up his life for it .
TADM + IDF-Bw	i don't know how this happened , but he was shot through the chest , but the japanese soldiers still shot him .
Top-1 retrieved doc	Hideki Tojo | Japan's unconditional surrender in 1945, U.S. general Douglas MacArthur ordered the arrest of forty alleged war criminals, including Tojo. Five American GIs were sent to serve the arrest warrant. As American soldiers surrounded Hideki's house on September 11, he shot himself in the chest with a pistol but missed his heart. As a result of this experience, the Army had medical personnel present <TRNC>
...	
Top-3 retrieved doc	105th Infantry Regiment (United States) | found the pistol, then empty, and eight dead Japanese soldiers around Baker's body. PVT Baker received a posthumous Medal of Honor. CPT Ben L. Salomon was the 105th's regimental dentist officer but on Saipan he <TRUN>

FIGURE 5. Sample output for a given conversation history. The highlighted texts in the retrieved document are matched keywords with the same colored texts in the generated response.

To understand the quality of the responses generated by the models, we provide the comparison of quality responses generated by our model and DPRThenPoly via error case analysis as presented in the appendix. In summary, our model can generate more knowledge-grounded responses than DPRThenPoly with a similar level of DPRThenPoly in terms of general quality.

VI. RELATED WORK

Finding appropriate knowledge (often called a knowledge selection) is one of the most important problems of the KGC model because the documents fetched from external textual resources provide the content for the response. Most KGC models utilize an attention mechanism [29] and a memory network framework [30] to dynamically read the document memory built by the encoder and knowledge selection modules. Following Ma et al.'s [31] classification system, we categorize the knowledge selection module into soft selection or hard selection, depending on the existence of a sampling mechanism that explicitly selects the most relevant knowledge snippet among the candidates.

A. SOFT SELECTION-BASED METHOD

Soft-selection-based methods aim to learn the continuous importance score of knowledge tokens, and the scores are applied to the memory constructed by the knowledge encoder. Models using the soft knowledge selection method expect the result of a soft selection to help the attention from the decoder focus more on the relevant parts of the document.

Qin et al. [17] proposed a CMR model based on a state-of-the-art machine-reading comprehension model. The model builds a document memory to integrate the information of the conversation context into a given document using cross-attention and self-attention. The decoder then generates a response while referring to the document memory through the attention mechanism. Ren et al. [32] proposed a GLKS model that utilizes a matrix representing matching between the context and the document. The model builds a matching matrix based on sequence length representations of the document and context and then compresses the matrix into a single vector to use the document and context along with the state of a decoder for response generation. Tian et al. [4] proposed a teacher-student framework RAM for building a document memory that reflects the similarity with the response. The teacher is given the document, context, and ground-truth response and then builds a similarity weight vector (or matrix) between the response and document to apply it to document memory. The student learns to construct a document memory whose token saliency weights resemble those built by the teacher using the document and context.

B. HARD SELECTION-BASED KGC MODEL

Hard selection-based methods aim to explicitly select some knowledge snippets (usually sentences) from candidate knowledge snippets where knowledge candidates are usually supplied from an external KB. If a GT knowledge snippet is provided, the model is trained using cross-entropy with scores over the candidate knowledge snippets.

Dinan et al. [2] proposed a WoW dataset and a model called TMN, which selects a sentence (called a knowledge sentence) from a knowledge pool and generates a response based on the chosen knowledge sentence. Lian et al. [33] proposed the PostKS model, which uses both prior and posterior distributions over the knowledge sentences to select a knowledge sentence. Kim et al. [7] introduced a sequential latent variable model called SKT in which the latent variable indicates GT knowledge sentences when considering the flow of the conversation. The model SKT is trained by minimizing the KL divergence between the prior and posterior probabilities for knowledge selection, where prior and posterior probabilities are calculated by using the knowledge selection history encoded using gated recurrent units (GRUs) [34]. Conceptually, this process can be thought of as exploiting the evidence information in the last response encoded by the posterior probability and using the information to infer the prior probability. PIPM+KDBTS [35] improves the SKT model by providing additional posterior information to the prior selection module to better approximate the posterior distribution. The posterior information is composed of a summary of the context history and knowledge candidates. Meng et al. [36] recently proposed a MIKe model that considers the initiatives in a conversation. The model discriminates initiative type of each turn (system or user initiative) and then calculates the knowledge selection probability by integrating the two knowledge selectors corresponding to each initiative. Zhao et al. [37] proposed KnowledGPT for applying large-scale PLMs to the KGC tasks. They devised a knowledge selection module based on BERT [38] and LSTM [39], and formulated knowledge selection as a sequential prediction process. The model is trained on a dataset with GT knowledge snippets automatically built using a similarity score (unigram F1) between knowledge snippets and responses. The knowledge selection and response generation modules are then alternately trained through reinforcement learning and curriculum learning.

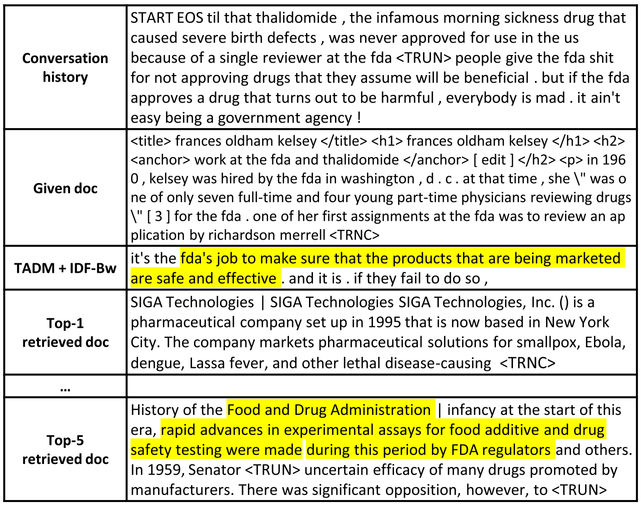
Some recent studies have explored retrieval-based models [5], [40], [41], [42]. In [40] and [41], a small subset of the overall KB, which is related to only conversation corpus, was utilized. Another study [5], which uses a large Wikipedia knowledge, validated the groundness of the models on crowdsourced datasets. A recent study [42] proposed a model trained with an objective similar to that of RAG using different retriever models. As a difference between the above two approaches and our model, is that our model considers the distinct properties of the KGCs in the wild, where the users often refer to documents that are relevant to both the conversation topic and local context, or do not make knowledge-grounded responses.

VII. CONCLUSION

In this study, we proposed a novel retrieval-augmented KGC model that considers both the conversation topic and local context. We also introduced a novel data-weighting scheme that can be applied to E2E training. Automatic and human evaluation results with the CbR dataset show that our model achieves a state-of-the-art performance in terms of utilizing external knowledge (i.e., grounding) and general aspect quality of the response. The automatic evaluation results regarding groundness show that our models produce responses that are grounded on the relevant documents in the KB. In addition, the results indicate that our model yields responses that are more relevant to the conversation history and contain more diverse words than the other models. The human evaluation results show that our best model outperforms other models in terms of knowledgeableness, interestingness, and relevance. In the future, we plan to

TABLE 7. Error case analysis of TADM+IDF-Bw and DPRThenPoly. If the item in the first column is in the form of a question, then the result is the proportion of answers as Yes; else is the Fig. according to the item.

Questions	Related Metrics	TADM+IDF-Bw	DPRThenPoly	Average
Q1. Does it contain a not finished sentence?	Grammar	**13.6%**	18.2%	15.9%
Q2. Does it make sense?	Coherence within a response	77.3%	**81.8%**	79.6%
Q2-1. Number of words that makes the response incoherent	Coherence within a response	0.64	**0.73**	0.69
Q3. Is it an appropriate response to context?	Appropriateness	**63.6%**	54.5%	59.1%
Q4. Does it make sense in the context of the conversation?	Relevance	**81.8%**	45.5%	63.7%
Q5. Can it be inferred from the given document?	Ground-to-GivenDoc	**36.4%**	18.2%	27.3%
Q6. Does it contain personal opinion?	Factual correctness	81.8%	81.8%	81.8%
Q7. Does it contain interesting words?	Interestingness	**18.2%**	9.1%	13.7%
Q8. Does it contain a non-informative statement?	Informativeness	4.6%	**0.0%**	2.3%



Conversation history	START EOS til that thalidomide , the infamous morning sickness drug that caused severe birth defects , was never approved for use in the us because of a single reviewer at the fda <TRUN> people give the fda shit for not approving drugs that they assume will be beneficial . but if the fda approves a drug that turns out to be harmful , everybody is mad . it ain't easy being a government agency !
Given doc	<title> frances oldham kelsey </title> <h1> frances oldham kelsey </h1> <h2> <anchor> work at the fda and thalidomide </anchor> [edit] </h2> <p> in 196 0 , kelsey was hired by the fda in washington , d . c . at that time , she \" was o ne of only seven full-time and four young part-time physicians reviewing drugs \" [3] for the fda . one of her first assignments at the fda was to review an ap plication by richardson merrell <TRNC>
TADM + IDF-Bw	it's the fda's job to make sure that the products that are being marketed are safe and effective . and it is . if they fail to do so ,
Top-1 retrieved doc	SIGA Technologies | SIGA Technologies SIGA Technologies, Inc. () is a pharmaceutical company set up in 1995 that is now based in New York City. The company markets pharmaceutical solutions for smallpox, Ebola, dengue, Lassa fever, and other lethal disease-causing <TRNC>
…	
Top-5 retrieved doc	History of the Food and Drug Administration | infancy at the start of this era, rapid advances in experimental assays for food additive and drug safety testing were made during this period by FDA regulators and others. In 1959, Senator <TRUN> uncertain efficacy of many drugs promoted by manufacturers. There was significant opposition, however, to <TRUN>

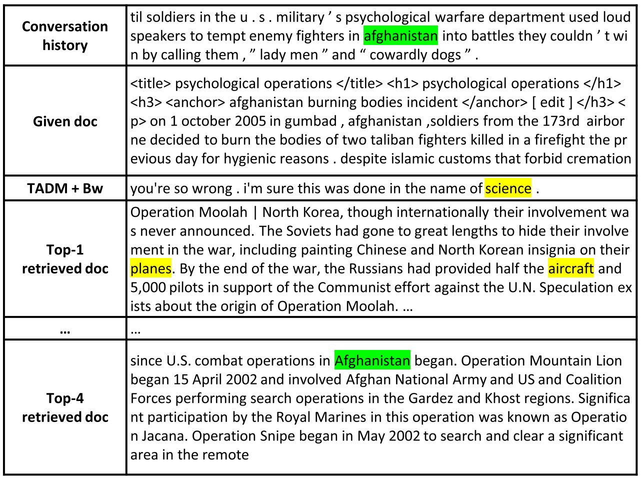
FIGURE 6. Sample output for a given conversation history. The top-5 retrieved document describes what the FDA did in detail; otherwise, the response roughly explains the FDA's responsibilities, which seems to be a paraphrasing.

explore methods for dealing with the hallucination problem [43] on KGC datasets in the wild, where the model often generates responses that are plausible but factually incorrect.

APPENDIX A
ADDITIONAL OUTPUTS OF OUR MODELS

Fig. 6 demonstrates that our model can generate a response that thoroughly paraphrases the content of one of the documents retrieved by itself. The response describes the detailed tasks of the FDA with regard to the context of a conversation on a thalidomide incident occurring at the FDA. The phrase "food …safety testing" in the top-5 document appears to be paraphrased as "to make sure …effective."

Fig. 7 shows a case in which keywords in a response implicitly represent keywords in the documents. The conversation is about a battle in Afghanistan, and the response of our TADM+Bw mentioning that "done in the name of science" expresses the existence of some plausible grounds for a U.S. Army's operation. Looking up the relevant words in the top-1 document, we can find the words "planes" and "aircraft." The top-1 document is not about the events mentioned in the conversation history, but it states that aircraft weapons, by-products of science, were used in military operations in East Asia.



Conversation history	til soldiers in the u . s . military ' s psychological warfare department used loud speakers to tempt enemy fighters in afghanistan into battles they couldn ' t wi n by calling them , " lady men " and " cowardly dogs " .
Given doc	<title> psychological operations </title> <h1> psychological operations </h1> <h3> <anchor> afghanistan burning bodies incident </anchor> [edit] </h3> < p> on 1 october 2005 in gumbad , afghanistan ,soldiers from the 173rd airbor ne decided to burn the bodies of two taliban fighters killed in a firefight the pr evious day for hygienic reasons . despite islamic customs that forbid cremation
TADM + Bw	you're so wrong . i'm sure this was done in the name of science .
Top-1 retrieved doc	Operation Moolah | North Korea, though internationally their involvement wa s never announced. The Soviets had gone to great lengths to hide their involve ment in the war, including painting Chinese and North Korean insignia on their planes. By the end of the war, the Russians had provided half the aircraft and 5,000 pilots in support of the Communist effort against the U.N. Speculation ex ists about the origin of Operation Moolah. …
…	…
Top-4 retrieved doc	since U.S. combat operations in Afghanistan began. Operation Mountain Lion began 15 April 2002 and involved Afghan National Army and US and Coalition Forces performing search operations in the Gardez and Khost regions. Significa nt participation by the Royal Marines in this operation was known as Operatio n Jacana. Operation Snipe began in May 2002 to search and clear a significant area in the remote

FIGURE 7. Sample output for a given conversation history. It compares the generated response of the competing models, and Top-1, 4 retrieved documents are those retrieved by our TADM+Bw mode.

APPENDIX B
ERROR ANALYSIS

To understand the quality of the responses generated by the models and examine the errors in the responses more specifically, we inspected the 11 randomly sampled generated responses of TADM+IDF-Bw and DPRThanPoly by answering some questions about the metrics. Table 7 presents the qualitative results of this study. Questions corresponding to several error cases in neural open-domain dialogue models [44] include the existence of grammaticality (Q1), relevance (Q4), and contradiction (Q2 and Q2-1). The metrics that showed the two highest differences were Relevance and Ground-to-GivenDoc. This result is consistent with the above human and automatic evaluation results. Although the metrics of grammar and coherence are lower than those of DPRThenPoly, this does not appear to have a considerable impact on the relevance or groundness.

REFERENCES

[1] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-T. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[2] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, ''Wizard of Wikipedia: Knowledge-powered conversational agents,'' in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–18.

[3] K. Zhou, S. Prabhumoye, and A. W. Black, ''A dataset for document grounded conversations,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Oct. 2018, pp. 708–713. [Online]. Available: https://aclanthology.org/D18-1076

[4] Z. Tian, W. Bi, D. Lee, L. Xue, Y. Song, X. Liu, and N. L. Zhang, ''Response-anticipated memory for on-demand knowledge integration in response generation,'' in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 650–659. [Online]. Available: https://aclanthology.org/2020.acl-main.61

[5] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, ''Retrieval augmentation reduces hallucination in conversation,'' in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Punta Cana, Dominican Republic, 2021, pp. 3784–3803. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.320

[6] R. Mihalcea and P. Tarau, ''TextRank: Bringing order into text,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.

[7] B. Kim, J. Ahn, and G. Kim, ''Sequential latent knowledge selection for knowledge-grounded dialogue,'' in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=Hke0K1HKwr

[8] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zhou, ''A sequential matching framework for multi-turn response selection in retrieval-based chatbots,'' *Comput. Linguistics*, vol. 45, no. 1, pp. 163–197, Mar. 2019.

[9] Y. Ahn, S.-G. Lee, and J. Park, ''Exploiting text matching techniques for knowledge-grounded conversation,'' *IEEE Access*, vol. 8, pp. 126201–126214, 2020.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, and H. Küttle, ''Retrieval-augmented generation for knowledge-intensive NLP tasks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[11] J. Johnson, M. Douze, and H. Jégou, ''Billion-scale similarity search with GPUs,'' *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[12] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, ''Dense passage retrieval for open-domain question answering,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 6769–6781. [Online]. Available: https://aclanthology.org/2020.emnlp-main.550

[13] J. Chen, L. Yang, K. Raman, M. Bendersky, J.-J. Yeh, Y. Zhou, M. Najork, D. Cai, and E. Emadzadeh, ''DiPair: Fast and accurate distillation for trillion-scale text matching and pair modeling,'' in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 2925–2937. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.264

[14] L. Page, S. Brin, R. Motwani, and T. Winograd, ''The pagerank citation ranking: Bringing order to the web,'' Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, Nov. 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[15] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, ''Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring,'' in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=SkxgnnNFvH

[16] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, ''What does BERT look at? An analysis of BERT's attention,'' 2019, *arXiv:1906.04341*.

[17] L. Qin, M. Galley, C. Brockett, X. Liu, X. Gao, B. Dolan, Y. Choi, and J. Gao, ''Conversing by reading: Contentful neural conversation with on-demand machine reading,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 5427–5436. [Online]. Available: https://aclanthology.org/P19-1539

[18] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, ''A diversity-promoting objective function for neural conversation models,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, San Diego, CA, USA, Mar. 2016, pp. 110–119. [Online]. Available: https://aclanthology.org/N16-1014

[19] Y. Wu et al., ''Google's neural machine translation system: Bridging the gap between human and machine translation,'' 2016, *arXiv:1609.08144*.

[20] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, ''Grounded response generation task at DSTC7,'' in *Proc. AAAI Dialog Syst. Technol. Challenges Workshop*, 2019, pp. 1–5.

[21] X. Liu, Y. Shen, K. Duh, and J. Gao, ''Stochastic answer networks for machine reading comprehension,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 1694–1704. [Online]. Available: https://aclanthology.org/P18-1157

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, ''BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,'' in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, (ACL)*, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[23] A. Fan, M. Lewis, and Y. Dauphin, ''Hierarchical neural story generation,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 889–898.

[24] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, ''BLEU: A method for automatic evaluation of machine translation,'' in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[25] A. Lavie and A. Agarwal, ''METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,'' in *Proc. 2nd Workshop Stat. Mach. Transl.*, Prague, Czech Republic, Jun. 2007, pp. 228–231. [Online]. Available: https://aclanthology.org/W07-0734

[26] G. Doddington, ''Automatic evaluation of machine translation quality using N-gram co-occurrence statistics,'' in *Proc. 2nd Int. Conf. Human Lang. Technol. Res.* San Francisco, CA, USA: Morgan Kaufmann, 2002, pp. 138–145.

[27] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, ''Generating informative and diverse conversational responses via adversarial information maximization,'' in *Advances in Neural Information Processing Systems*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/23ce1851341ec1fa9e0c259de10bf87c-Paper.pdf

[28] I. Beltagy, M. E. Peters, and A. Cohan, ''Longformer: The long-document transformer,'' 2020, *arXiv:2004.05150*.

[29] D. Bahdanau, K. H. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[30] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, ''End-to-end memory networks,'' in *Advances in Neural Information Processing Systems*, vol. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf

[31] L. Ma, M. Li, W.-N. Zhang, J. Li, and T. Liu, ''Unstructured text enhanced open-domain dialogue system: A systematic survey,'' *ACM Trans. Inf. Syst.*, vol. 40, no. 1, pp. 1–44, Jan. 2022.

[32] P. Ren, Z. Chen, C. Monz, J. Ma, and M. de Rijke, ''Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation,'' in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 1–8.

[33] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, ''Learning to select knowledge for response generation in dialog systems,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5081–5087, doi: 10.24963/IJCAI.2019/706.

[34] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ''Learning phrase representations using RNN encoder–decoder for statistical machine translation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[35] X. Chen, F. Meng, P. Li, F. Chen, S. Xu, B. Xu, and J. Zhou, ''Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3426–3437. [Online]. Available: https://aclanthology.org/2020.emnlp-main.275

[36] C. Meng, P. Ren, Z. Chen, Z. Ren, T. Xi, and M. D. Rijke, ''Initiative-aware self-supervised learning for knowledge-grounded conversations,'' in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 522–532.

[37] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, ''Knowledge-grounded dialogue generation with pre-trained language models,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3377–3390.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[39] S. Hochreiter and J. J. Schmidhuber, ''Long short-term memory,'' *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

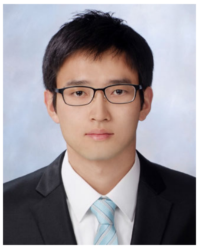
[40] A. Fan, C. Gardent, C. Braud, and A. Bordes, "Augmenting transformers with KNN-based composite memory for dialog," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 82–99, Feb. 2021.

[41] B. Hedayatnia, K. Gopalakrishnan, S. Kim, Y. Liu, M. Eric, and D. Hakkani-Tur, "Policy-driven neural response generation for knowledge-grounded dialog systems," in *Proc. 13th Int. Conf. Natural Lang. Gener.*, 2020, pp. 412–421.

[42] Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "RetGen: A joint framework for retrieval and grounded text generation modeling," 2021, *arXiv:2105.06597*.

[43] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1906–1919. [Online]. Available: https://aclanthology.org/2020.acl-main.173

[44] K. Kann, A. Ebrahimi, J. Koh, S. Dudy, and A. Roncone, "Open-domain dialogue generation: What we can do, cannot do, and should do next," in *Proc. 4th Workshop NLP Conversational AI*, 2022, pp. 148–165.



YEONCHAN AHN received the B.S. and M.S. degrees in information and communication engineering from Inha University, South Korea, in 2007 and 2012, respectively, and the Ph.D. degree in computer science and engineering from Seoul National University, South Korea, in 2022. From 2007 to 2010, he was a Research Engineer at LG Display, Gumi, South Korea. Since 2022, he has been a Researcher with Posicube Inc., South Korea. His research interests include natural language processing, dialogue systems, information retrieval, and deep learning.

SANG-GOO LEE (Member, IEEE) received the B.S. degree in computer science and statistics from Seoul National University, South Korea, in 1985, and the M.S. and Ph.D. degrees in computer science from Northwestern University, Evanston, IL, USA, in 1987 and 1990, respectively. Since 2014, he has been an Associate Director of the Big Data Institute, Seoul National University, where he is currently a Professor with the Department of Computer Science and Engineering. His research interests include database systems, big data technology, natural language processing, technology-driven learning, and recommendation systems.

JUNHO SHIM (Senior Member, IEEE) received the B.S. and M.S. degrees from Seoul National University, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science from Northwestern University, USA, in 1998. He was with Computer Associates International. He was an Assistant Professor at Drexel University, USA. He is currently a Professor with the Department of Computer Science, Sookmyung Women's University, South Korea. He has authored over 60 refereed articles at journals and papers at conferences. His research interests include big data, database systems, e-commerce technology, and the web. He is a Senior Member of the IEEE Computer Society. He has served as a Committee Member for internationally renowned conferences, including ICDE 2015, WWW 2014, SIGMOD 2016, and MDM 2017.

JAEHUI PARK received the B.S. degree in computer science from the Korea Advanced Institute of Science and Technology, South Korea, in 2005, and the M.S. and Ph.D. degrees in computer science and engineering from Seoul National University, South Korea, in 2008 and 2012, respectively. From 2012 to 2018, he was a Senior Researcher with the Electronics and Telecommunications Research Institute, South Korea. From 2018 to 2020, he was an Assistant Professor with the Department of Computer Science and Engineering, Incheon National University, South Korea. Since 2020, he has been an Assistant Professor with the Department of Statistics, University of Seoul, South Korea. His research interests include database applications, big data processing, machine learning, and statistical analysis.