Enhancing Knowledge Retrieval with Topic Modeling for Knowledge-Grounded Dialogue

Nhat Tran, Diane Litman
University of Pittsburgh
Pittsburgh, PA 15260 USA
nlt26@pitt.edu, dlitman@pitt.edu

Abstract

Knowledge retrieval is one of the major challenges in building a knowledge-grounded dialogue system. A common method is to use a neural retriever with a distributed approximate nearest-neighbor database to quickly find the relevant knowledge sentences. In this work, we propose an approach that utilizes topic modeling on the knowledge base to further improve retrieval accuracy and as a result, improve response generation. Additionally, we experiment with a large language model, ChatGPT, to take advantage of the improved retrieval performance to further improve the generation results. Experimental results on two datasets show that our approach can increase retrieval and generation performance. The results also indicate that ChatGPT is a better response generator for knowledge-grounded dialogue when relevant knowledge is provided.

Keywords: knowledge-grounded dialogue, topic modeling, dense retrieval

1. Introduction

In knowledge-grounded dialogues, to find relevant knowledge passages from a large knowledge base, retrieval-based approaches use two encoders to encode both dialogue history and the knowledge base into the same vector space. The encoded dialogue history is treated as an input query to quickly find the relevant knowledge by retrieving the top-K passages in the encoded knowledge base based on a similarity score (e.g., dot product). Improvement in any of these two encoders can potentially lead to increased performance of knowledge retrieval.

While some prior work focused on improving the dialogue history encoder (Tran and Litman, 2022), ours focuses on the knowledge base encoder. Specifically, we use topic modeling to cluster the knowledge base and train a separate encoder for each cluster. We then incorporate the topic distribution of the input query into the similarity score to find the top-K passages. Due to impressive performance across various natural language processing (NLP) tasks of large language models (LLMs) such as ChatGPT, we also experiment with using ChatGPT as the response generator, with and without the retrieved knowledge. Figure 1 shows our focus within the retrieve-then-generate framework.

Our contribution is threefold. First, we propose a modification utilizing topic modeling to the widely used RAG (retrieval-augmented generation) model and achieve improved performance and verify that using validation sets is a reliable way to pick the optimal number of topics. Second, we show that combining our approach which manipulates the knowledge base with an approach focusing on building a better input query can further improve performance. Finally, we find that the relevant knowledge is essential for ChatGPT to achieve the best performances. We also make our source code available at https://github.com/nhattlm95/tm_kg_dialogue.

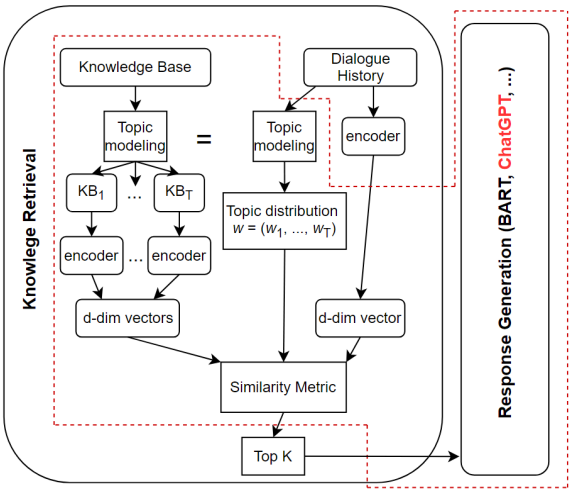

Figure 1: The modified retrieve-then-generate framework (based on RAG) with our contribution highlighted. The two topic modeling modules are the same one trained on the knowledge base.

2. Related Work

For knowledge-grounded NLP, knowledge retrieval is a crucial step (Eremeev et al., 2023). Although LMs can be embedded with knowledge (Petroni et al., 2019; Heinzerling and Inui, 2021; Shin et al., 2020; Roberts et al., 2020), retrieve-then-generate models still yield higher performances in knowledge-intensive tasks (Petroni et al., 2021; Di-

LREC-COLING 2024, pages 5986–5995

nan et al., 2019; Li et al., 2022). Our work follows this line of research, in which a dedicated knowledge retrieval component is used.

Recent dense retrieval methods (Karpukhin et al., 2020; Lewis et al., 2020b; Xiong et al., 2021), which encode text as a latent vector and use their distances to determine the relevance, have outperformed the sparse methods such as TF-IDF or BM25 (Robertson and Zaragoza, 2009). In this work, we utilize dense retrieval by modifying the retriever module and the way to calculate the similarity scores of the popular RAG model (Lewis et al., 2020b) with the help of topic modeling.

The concept of *topics* has not been explored much in knowledge-grounded dialogue. Xu et al. (2022) proposed an end-to-end framework that uses topic modeling to skip the explicit retrieval process and inject knowledge into the pre-trained language models for knowledge-grounded conversations. Tran and Litman (2022) tries to maintain similar 'topics' (e.g., turns grounded in the same document) in the dialogue history used as input queries in dense retrieval. Those works are different from ours as we focus on improving the knowledge retrieval component with the help of topic modeling on the knowledge base.

Although ChatGPT (OpenAI, 2022) has shown great performance in various NLP tasks (Laskar et al., 2023), recent works in knowledge-grounded dialogue (Li et al., 2022; Zhao et al., 2022; Wu et al., 2022; Gowriraj et al., 2023) have not utilized it as a response generator. Our work tests the potential of ChatGPT to generate responses that need to be grounded in certain knowledge, with the presence/absence of the required knowledge.

3. Method

We first perform **topic modeling** to cluster the training knowledge base into a pre-defined number (***T***) of topic clusters. We use the contextual topic model (CTM) from Bianchi et al. (2021) which has shown better topic coherence compared to traditional methods. The major components of CTM are a neural topic model Neural-ProdLDA (Srivastava and Sutton, 2017) and pre-trained Sentence Transformers embedding (Reimers and Gurevych, 2019). Once trained, the model can output a T-dimension vector $w = (w_1, w_2, ..., w_T)$ given an input sequence, which is the probability distribution of the pre-clustered topics.

To find the top-K relevant knowledge passages from a large knowledge base for a given dialogue history *H*, we modify **Dense Passage Retrieval (DPR)** (Karpukhin et al., 2020). Traditionally, it utilizes two BERT encoders (Devlin et al., 2019), a document encoder ($BERT_d$) and a query encoder ($BERT_q$), to encode the knowledge passages and the dialogue history to the same d-dimensional space. The document encoding is done offline and indexed in a database such as FAISS (Johnson et al., 2021) which can retrieve the top-K at inference time quickly if the relevance score between two vectors is calculated as their dot product.

However, since we have a T-cluster knowledge base, for each cluster t_i, we train a separate document encoder $BERT_d^i$. Given the topic distribution of the dialogue history H calculated using CTM as $w = (w_1, w_2, ..., w_T)$, to find the top-K passages, we first retrieve the top-K passages from each cluster t_i, with the relevant score of a passage p inside the cluster calculated as:

$$BERT_q(H) \cdot BERT_d^i(p) \times w_i \quad (1)$$

where $\cdot$ is dot product and $\times$ is multiplication. Then, we choose the top-K from these K $\times$ T retrieved passages. We call this version **DPR-topic**.

To generate the response, we use **RAG** (Lewis et al., 2020b). It has a retriever (DPR) and a generator module (BART, Lewis et al. (2020a)). Given the dialogue history as an input query, the retriever finds the top-K relevant passages, and the generator takes the dialogue history and retrieved top-K passages to generate the response. The retriever is non-parametric so any pre-trained model can be used. We use **DPR-topic** as the retriever and do not touch the RAG query encoder or the generator module. Our model is called **RAG-topic**.

For MultiDoc2Dial data (Section 5), we also experiment with a RAG-based model (RAG-context) that uses an algorithm to select relevant turns in the dialogue history (Tran and Litman, 2022). Since it only changes the query of RAG, our approach which manipulates the knowledge base can be applied in the same way we modify the RAG model. We call this model **RAG-context-topic**.

Finally, we experiment with **ChatGPT**[1]. Besides asking ChatGPT to generate the response directly given the dialogue history, we feed external knowledge as input to ChatGPT to give it the necessary knowledge. The external knowledge is text retrieved from the retriever (the web page containing the top-1 retrieved passage).

4. Implementation Details

For topic modeling, we use CTM[2] with default parameters and only change the number of topics.

For **RAG-topic**, which is modified from RAG[3] while keeping the default parameters, we have a shared query encoder $BERT_q(H)$, BART-generator and separate document encoders for each cluster *i*, $BERT_d^i(p)$.

[1]We use GPT-4 from OpenAI.

[2]https://tinyurl.com/3hb3bkbu

[3]https://tinyurl.com/mstamtct

We use a pre-trained bi-encoder from DPR[4] (Karpukhin et al., 2020) to initialize our encoders and create the index for each cluster in the knowledge base. Then, using the retrieval objective from DPR, we fine-tune the $BERT_d^i(p)$ and $BERT_q(H)$ using the training examples related to the i^{th} cluster. Specifically, in the training data (of either MultiDoc2Dial or KILT-dialogue), if the gold knowledge of a training instance is in cluster *i*, we put it into the training set for $BERT_d^i(p)$. In other words, in the fine-tuning process, each $BERT_d^i(p)$ will have a separate training set that includes only "questions" that require knowledge in cluster *i*. After this step, we use the trained document encoders to create the fixed document index, the trained document encoders are also fixed now (non-parametric).

We continue finetuning the query encoder $BERT_q(H)$ and BART generator using all training data (either KILT-dialogue or MultiDoc2Dial) with the new retrieval results from the non-parametric retrievers. We modify the retriever of RAG to get the top-K passages as described in Section 3.

For **RAG-context-topic**, we modify the code provided by Tran and Litman (2022) in the same way we modify the RAG model to create **RAG-topic**, with the new dialogue history as the query during training and inference.

For **ChatGPT**, we use GPT-4 (8k context) with max_tokens = 100, temperature = 0.5 and other parameters set as default. The following prompt is used, where {Provided Knowledge} is the web page containing the top-1 passage from the retriever.

> Using the given knowledge
> {Provided Knowledge}
> Complete the dialogue with <system> as your role:
> <user>: ...
> <system>: ...
> [...]
> <user>: ...

We choose K=10 in all of our experiments as the baseline RAG model uses the top-10 retrieved passages for generation.

All models were trained on one RTX 3090 card.

5. Experiment Setup

We use two **datasets** of knowledge-grounded dialogues for this study. **MultiDoc2Dial** (Feng et al., 2021) consists of around 4800 information-seeking conversations, grounded in 488 documents from 4 domains. **KILT** (Petroni et al., 2021) is a dataset designed for knowledge-intensive tasks, grounded in Wikipedia. We only use its dialogue subtask (**KILT-dialogue**), in which one speaker must ground their utterances in a specific knowledge sentence, chosen from a Wikipedia page. For consistency, we use *passage* to refer to the knowledge text spans we want to retrieve for response generation. Data examples and statistics are in Appendix A, while an example comparing RAG-topic with RAG given a history from KILT-dialogue is in Appendix B.

Because the number of topics T is a vital hyperparameter, having a way to pick T is crucial. To test the performance sensitivity of T, we experiment with different values of T and report the topic coherence scores and passage retrieval performance. To evaluate the quality of topics from topic model (topic coherence), we follow the authors of our CTM model (Bianchi et al., 2021) and use external word embeddings topic coherence (Ding et al., 2018). The evaluation metric for passage retrieval is Recall at 5 (R@5), which answers the question: out of all relevant passages, how many of them are included in the top-5 retrieved passages.

For **downstream evaluation** to compare to other baselines, following KILT (Petroni et al., 2021), we use page-level Precision at 1 (P@1) to report the final retrieval performances, which is the percentage of correct pages among the top-1 retrieved pages. For generation results, we use unigram-F_1 score between the generated and gold responses. We also use KILT-F_1, which only awards points when the gold-knowledge page is retrieved.

For both datasets, we compare our proposed **RAG-topic** model to a baseline **RAG** model. For MultiDoc2Dial, we also develop a **RAG-context-topic** model to evaluate whether RAG-topic can add value to a prior model developed for this corpus (which we call **RAG-context**), which focused on the dialogue history rather than the knowledge base (Tran and Litman, 2022). The base RAG-context approach has an algorithm and predictive modules to form the dialogue history (input to RAG), based on an assumption that including only turns grounded in the same document as the current turn provides a better input query. Finally, for KILT-dialogue, we use two baselines from KILT (Petroni et al., 2021), RAG and **BART+DPR**, which simply concatenates the retrieved passage from DPR to the dialogue history as input to BART.

We also use **ChatGPT** as a retrieval-free baseline for both datasets, as well as use it as a response generator given the required knowledge. The knowledge source can be knowledge pages retrieved from a model (**+DPR**[5], **+RAG**, **+RAG-topic**, **+RAG-context**, **+RAG-context-topic**) or the **golden knowledge** provided in the datasets. For example, **ChatGPT+RAG** means we feed the retrieved knowledge from the RAG model to ChatGPT's prompt to get the response.

Due to the randomness of the models (e.g.

[4] https://tinyurl.com/mtdeta2w

[5] In KILT-dialogue, the non-ChatGPT counterpart is BART+DPR, but we only need DPR for retrieval.

	Number of Topics (T)									
	1	2	3	4	5	6	7	8	9	10
Topic coherence	0.31	0.25	0.29	**0.38**	**0.35**	0.33	**0.35**	0.29	0.27	0.22
RAG-topic / Validation	71.7	72.0	**72.1**	**72.5**	**72.9**	71.1	71.3	71.9	68.0	67.5
RAG-context-topic / Validation	72.0	72.1	**72.2**	**72.6**	**72.7**	71.1	70.1	71.8	71.3	69.8
RAG-topic / Test	72.5	72.2	72.5	**73.3**	**73.7**	71.5	70.9	72.3	68.3	68.4
RAG-context-topic / Test	72.8	**72.9**	**72.9**	**73.2**	**74.4**	71.5	71.7	72.8	70.5	70.1

Table 1: Retrieval Results (R@5) on Test and Validation data of **MultiDoc2Dial** (average of 3 runs). **Bolded** results are significantly better than those in the same row with T=1 (no topic modeling) in a pairwise t-test ($p < 0.05$). The best result of each row is underlined.

	Number of Topics (T)									
	1	2	3	4	5	6	7	8	9	10
Topic coherence	0.12	**0.16**	**0.22**	**0.34**	**0.35**	**0.37**	**0.27**	**0.33**	**0.38**	**0.36**
RAG-topic / Validation	36.3	36.2	**38.5**	**40.1**	**38.0**	**38.7**	30.6	30.3	25.3	23.5
RAG-topic / Test	37.5	34.8	35.3	**39.9**	**39.4**	39.7	31.6	30.9	26.3	24.7

Table 2: Retrieval Results (R@5) of RAG-topic on Validation and Test data of **KILT-dialogue** (average of 3 runs) with the same annotation as Table 1.

dropout from CTM training), we run each experiment 3 times and report the average results.

6. Results

Tables 1 and 2 show the **passage retrieval results** with various numbers of topics (T) for the 2 tested datasets with topic coherence scores reported in the first row of each table. Although our models can outperform the baseline counterparts with no topic modeling (T = 1) with the right choices of T (e.g., T = 4 or 5), certain Ts can yield lower results compared to the baselines (e.g., T = 10). The results also show that the best T is consistent among the same dataset but different across datasets (i.e., T = 5 for MultiDoc2Dial and T = 4 for KILT-dialogue). In contrast, for both datasets, higher scores in topic coherence do not necessarily lead to higher retrieval results. Therefore, we suggest using the validation set and choosing T that achieves the highest R@5 to find the optimal T for each dataset. We then use the optimal T of each dataset to perform the downstream evaluation (i.e., T = 5 for MultiDoc2Dial and T = 4 for KILT-dialogue). The top keywords of the clusters are in Appendix C.

Tables 3 and 4 show the **page retrieval results** for MultiDoc2Dial and KILT-dialogue, respectively. Our models significantly outperform the baseline counterparts with no topic modeling. For MultiDoc2Dial, we witnessed an increase of 2.71 points from RAG to RAG-topic, and 4.76 points from RAG-context to RAG-context-topic. For KILT-dialogue, performance significantly increases by 5.46 points from RAG to RAG-topic. These results indicate an improvement in retrieval performances when topic modeling is used in our proposed way.

Table 5 shows the **response generation results**

Model	P@1
RAG	64.61
RAG-topic (ours)	67.32*
RAG-context	67.55
RAG-context-topic (ours)	**72.31***

Table 3: Retrieval results on MultiDoc2Dial (T = 5 for our models) with the best result **bolded**. Results with * are statistically significant (pairwise t-test, $p < 0.05$) compared to its non-topic counterpart in the prior row.

Model	P@1
BART+DPR	25.48*
RAG	57.75
RAG-topic (ours)	**63.21***

Table 4: Retrieval results on KILT-dialogue (T = 4 for our models) with the best result **bolded**. Results with * are statistically significantly different ($p < 0.05$) compared to RAG.

on MultiDoc2Dial. The first 4 rows show that our topic-based RAG models have significantly higher scores compared to their related RAG baselines (row above). Although the F_1 increases are small the increases in KILT-F_1 are larger . The bottom of the table shows that without any knowledge, ChatGPT performs very poorly (35.8). However, when knowledge is provided, ChatGPT generates responses with significantly higher F_1 and KILT-F_1 compared to the original RAG-based versions (same retriever, different generator) in the first four rows. We hypothesize that this dataset focuses on information-seeking conversations, so it is hard to provide the response without relevant information.

In Table 6, we report the generation results on

Model	F_1	KILT-F_1
RAG	41.1	30.71
RAG-topic (ours)	41.3*	34.46*
RAG-context	41.2	32.93
RAG-context-topic (ours)	42.1*	36.21*
ChatGPT	*35.8*	-
+ RAG	*44.5**	36.50*
+ RAG-topic	*47.6**	*38.12**
+ RAG-context	46.9*	38.03*
+ RAG-context-topic	***49.3****	***39.81****
+ golden knowledge	*55.2*	*42.13*

Table 5: Generation results on MultiDoc2Dial (T = 5 for our models). For ChatGPT, the '+' part is only the *knowledge retrieval result* from the mentioned model. Best non-ChatGPT results are underlined and best overall results (not using golden knowledge) are **bolded**. For RAG models (first 4 rows), * indicates statistical significance ($p < 0.05$) compared to equivalent non-topic model (one row above). For ChatGPT-based models, * indicates significance ($p < 0.05$) compared to the non-ChatGPT version (first four rows), *italic* indicates significance compared to one row above.

Model	F_1	KILT-F_1
BART+DPR	15.19*	4.37*
RAG	13.19	9.05
RAG-topic (ours)	15.25*	11.46*
ChatGPT	*16.12*	-
+ DPR	*17.63**	11.97*
+ RAG	*18.21**	*12.07**
+ RAG-topic	***19.46****	***15.41****
+ golden knowledge	*22.39*	*18.72*

Table 6: Generation results on KILT-dialogue (T = 4). For the first 3 rows, * indicates statistical significance ($p < 0.05$) compared to RAG. Annotation for ChatGPT-based models are the same as Table 5.

the KILT-dialogue dataset. The first 3 rows show that RAG-topic achieves the highest scores for both metrics. Although the gain in F_1 is marginal compared to BART+DPR, the KILT-F_1 score is more than double (11.46 versus 4.37). Even without external knowledge, ChatGPT outperforms the three retrieval-based models. There are two reasons we can think of for this behavior. First, Wikipedia's knowledge is already built in ChatGPT internally during training. Second, this dataset is more chitchat-oriented so the response only needs to relate to the latest topic and is less strictly restricted to a specific piece of knowledge. When external knowledge is given to ChatGPT, we observe the same behavior as in MultiDoc2Dial. Specifically, given the same knowledge retrieved from a model (+DPR, +RAG or +RAG-topic), ChatGPT generates responses with higher F_1 and KILT-F_1 scores than its original versions. A brief analysis of the relation between the length of the dialogue history and the generation performance is in Appendix D.

In general, models with higher P@1 have higher F_1 and KILT-F_1 [6]. Models using golden knowledge achieve the highest results. When only retrieved knowledge is used, ChatGPT with the best retriever always wins (+RAG-context-topic for MultiDoc2Dial and +RAG-topic for KILT-dialogue). This suggests that better retrieval leads to better generation.

7. Conclusion

In this work, we proposed a method that utilizes topic modeling on the knowledge base to improve the performance of RAG-based models. Our approach uses topic modeling to cluster the knowledge base, build a separate document encoder for each cluster, and uses the topic distribution weights to calculate similarity scores. Additionally, we experiment with ChatGPT to see its performance with and without external knowledge. We observe that using the validation set to find the optimal number of topics is a reliable approach. Overall, our RAG-based models achieve improvement in both retrieval and generation, and compliment with models focusing on building a better dialogue history representation. We also find that ChatGPT can take advantage of the improved retrieval performance to yield even higher generation results. ChatGPT does not perform very well without external knowledge, but it is superior when knowledge is provided, obtaining higher results given the same knowledge compared to RAG-based models. Future plans include utilizing multi-task training with similar knowledge-intensive tasks and integrating the knowledge-retrieving process into a pipeline of large language models such as ChatGPT.

8. Limitations

One major limitation of our approach is that the computational requirement is proportional to the number of topics T as we need to retrieve from each knowledge base cluster to get the final top-K. For each new topic, we need an additional document encoder (BERT - 110M parameters), which results in 110M x (T-1) parameters more than the original RAG model (where T is the number of topics). Therefore, this method does not scale well if the optimal T is large. Additionally, for generation results, this work relies solely on automatic metrics and lacks human evaluation. The lack of diversity of open-source LLMs such as Llama2 or Vicuna makes the findings less generalizable.

[6] Exceptions are BART+DPR vs. RAG (KILT-dialogue) and RAG-topic vs. RAG-context (MultiDoc2Dial).

Ethical Considerations

Although this work focuses on knowledge retrieval performance (e.g. finding the correct knowledge passages as frequently as possible), other aspects of accuracy should be considered, especially in systems that provide information to the user. For example, for a healthcare application, giving the user wrong information is more dangerous than generating an irrelevant response, but both cases are considered equally failed instances when training/testing for most models. Since no NLP/AI model is perfect, depending on the application, further regulation is needed to prevent misinformation.

References

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.

Maksim Eremeev, Ilya Valmianski, Xavier Amatriain, and Anitha Kannan. 2023. Injecting knowledge into language generation: a case study in auto-charting after-visit care instructions from medical dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2373–2390. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Srinivas Gowriraj, Soham Dinesh Tiwari, Mitali Potnis, Srijan Bansal, Teruko Mitamura, and Eric Nyberg. 2023. Language-agnostic transformers and assessing ChatGPT-based query rewriting for multilingual document-grounded QA. In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 101–108, Toronto, Canada. Association for Computational Linguistics.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt blog post.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Nhat Tran and Diane Litman. 2022. Getting better dialogue context for knowledge identification by leveraging document-level topic shift. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 368–375, Edinburgh, UK. Association for Computational Linguistics.

Sixing Wu, Ying Li, Ping Xue, Dawei Zhang, and Zhonghai Wu. 2022. Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 521–531, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107, Dublin, Ireland. Association for Computational Linguistics.

Xueliang Zhao, Tingchen Fu, Chongyang Tao, and Rui Yan. 2022. There is no standard answer:

Knowledge-grounded dialogue generation with adversarial activated multi-reference learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1878–1891, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A. Datasets

Figures 2 and 3 show one example each from our two datasets, MultiDoc2Dial and KILT-dialogue, respectively.

The sizes of training, validation and test set of the two datasets we used can be seen in Table 7. For KILT-dialogue, since the gold answer of the original test set is not released, we use the original validation set as our test set (3054 items). We then use 3054 out of the 63734 instances in the original training set as our validation set to find the optimal T. As a result, our training set consists of 60680 instances.

Dataset	Train	Validation	Test
MultiDoc2Dial	3,474	661	661
KILT-dialogue	60680	3,054	3054

Table 7: Dataset Statistics

B. Examples of Retrieved Passages and Response Generation

In Table 9, we show the top-1 retrieved passage and generated response from RAG and RAG-topic for a given dialogue history in KILT-dialogue. The topic distribution weights from CTM helped guide the search to Cluster 3, which contains knowledge about novels and films, to find a relevant knowledge passage. On the other hand, the original RAG model found an irrelevant knowledge passage and generated an inappropriate response.

C. Themes Among the Clusters

Table 8 shows the list of keywords of each cluster from the knowledge base of KILT-dialogue when the number of topic T for CTM is set as 4 and MultiDoc2Dial when T is set as 5. Generally, there are "themes" among these clusters. For example, in KILT-Dialogue, cluster 1 is related to geography, cluster 2 is about music, cluster 3 is about novels and film. For MultiDoc2Dial, the first 4 clusters are quite representative of the 4 domains in the dataset: Department of Motor Vehicles (dmv), Social Security Affair (ssa), Student Aid (sa) and Veteran Affair (va). The last cluster is a mixture of the information across 4 domains.

D. Relations between Dialogue History Length and Performance

Table 10 shows the Pearson correlation values between the length of the dialogue history (number of tokens) and generation performance (F1) on two datasets. Generally, there are negative correlations between the two variables, indicating that the performance decreases when the dialogue history is longer. This is reasonable as not all information in the dialogue history is relevant to the current turn and the redundancy can create noise for the retrieval process. We also observe that our proposed approaches (RAG-topic and RAG-context-topic) help mitigate this negative relation because their absolute Pearson correlation values are smaller compared to RAG. Especially in MultiDoc2Dial, RAG-context-topic can filter out irrelevant turns in the history, and thus there might be less noise in the selected dialogue history used as a query for the retrieval process.

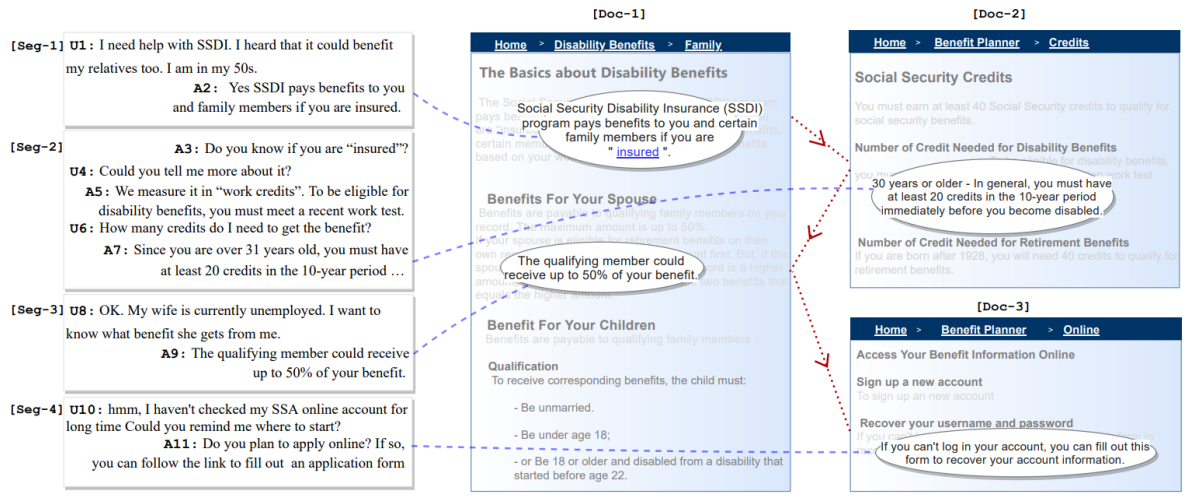

Figure 2: An example dialogue from MultiDoc2Dial borrowed from Feng et al. (2021). The conversation (on the left) is grounded in 3 documents Doc-1, Doc-2, and Doc-3. Each dialogue segment indicates that all turns within it are grounded in the same document (e.g., A3 to A7 in Seg-2 are all grounded in Doc-2). A dialogue turn and its corresponding relevant span in a document are connected by a blue dashed line. The red dotted lines with arrows show the dialogue flow shifts among the grounding documents through the conversation (e.g., Doc-1 → Doc-2 → Doc-1 → Doc-3).

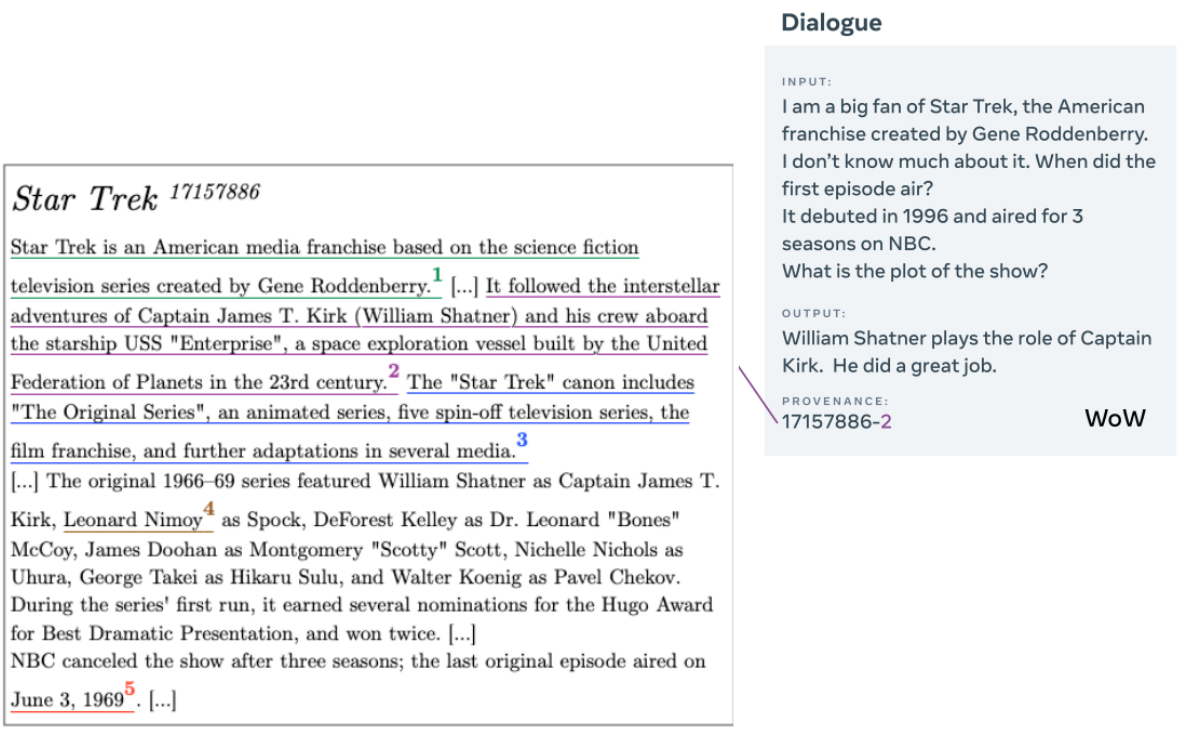

Figure 3: An example dialogue from KILT-dialogue borrowed from Petroni et al. (2021). Two speakers talk about a given topic (e.g., Star Trek) grounded in a Wikipedia page.

KILT-Dialogue	Number of Topics (T) = 4	
	Cluster 1	east, west, south, river, north, state, area, city, district, center
	Cluster 2	rock, band, records, music, song, album, team, record, club, studio
	Cluster 3	story, fiction, characters, book, disney, novel, film, episode, films, comic
	Cluster 4	pain, bon, Canberra, rutgers, blocked, khalil, edmonton, capitals, auckland, auburn
MultiDoc2Dial	Number of Topics (T) = 5	
	Cluster 1	car, dmv, vehicle, plate, license, driver, toll, registration, insurance, hearing
	Cluster 2	benefit, social, disabled, number, income, retirement, document, children, child, security
	Cluster 3	student, aid, school, apply, scholarship, college, aids, program, grant, loans
	Cluster 4	va, status, appeal, account, claim, evidence, review, compensation, deposit, allowance
	Cluster 5	test, benefits, address, registrations, information, website, programs, accounts, online, office

Table 8: Top 10 words for each cluster of the knowledge base of KILT-dialogue and MultiDoc2Dial

Dialogue history		
Speaker 1: the Draco lizard is so cool they can glide from trees Speaker 2: Lizards are just cool in general but i havent heard of that one before Speaker 1: have you heard of Draco Malfoy?		
Model	**RAG**	**RAG-topic** (T = 4)
Topic distribution	w = (1.00)	w = (0.21, 0.09, 0.55, 0.15)
Retrieved passage (Top-1)	Members of Draco are primarily arboreal, inhabiting tropical rainforests, and are almost never found on the forest floor	Draco Lucius Malfoy is a character in J. K. Rowling's "Harry Potter" series.
Generated response	Yes, you can find them in tropical rainforests.	Yes, he is a character in harry potter series.

Table 9: An example from KILT-dialogue in which our proposed RAG-topic successfully retrieved a relevant knowledge passage while the original RAG failed to do so for the same given dialogue history. For RAG-topic, vector w represents the topic distribution of the four clusters in Table 8 from the dialogue history.

	MultiDoc2Dial	KILT-dialogue
RAG	-0.35*	-0.24*
RAG-topic	-0.28*	-0.20*
RAG-context-topic	-0.19*	n/a

Table 10: The Pearson correlation values between the length of the dialogue history (number of tokens) and generation performance (F1). The columns represent the different datasets and the rows represent the different models. * indicates $p < 0.05$ in a two-tailed t-test.