Enhancing Knowledge Retrieval for Knowledge-Grounded Dialogue with Topic Modeling

Anonymous ACL submission

Abstract

Knowledge selection is one of the major challenges in building a knowledge-grounded dialogue system. A common method is to use a neural retriever with 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. Experimental results on two datasets show that our model can increase retrieval and generation performance with the correct number of topics chosen. The results also indicate that selecting the right number of topics to segment the knowledge base should be data-dependent and a higher topic coherence of topic modeling does not necessarily lead to better knowledge retrieval performance.

1 Introduction

In knowledge-grounded dialogues, one of the major challenges is to quickly find relevant knowledge passages from a large knowledge base. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020b) has been widely used as a baseline for these knowledge-grounded generation tasks. It uses Dense Passage Retrieval (DPR, Karpukhin et al. (2020)), which utilizes two encoders to encode both dialogue history and the knowledge base into the same vector space, to quickly find the most relevant knowledge passages for the given dialogue history before response generation. Improvement in any of these two encoders can potentially lead to increased performance of knowledge retrieval.

While prior work focused on improving the dialogue history encoder (Tran and Litman, 2022), this paper focuses on the knowledge base encoder. Specifically, we use topic modeling to cluster the knowledge base and train a separate encoder for each cluster. Since the topic distribution of the input query can provide a good signal to find relevant knowledge, we then incorporate it into the similarity score to find the top-K relevant passages. Figure 1 shows our focus within the DPR framework.

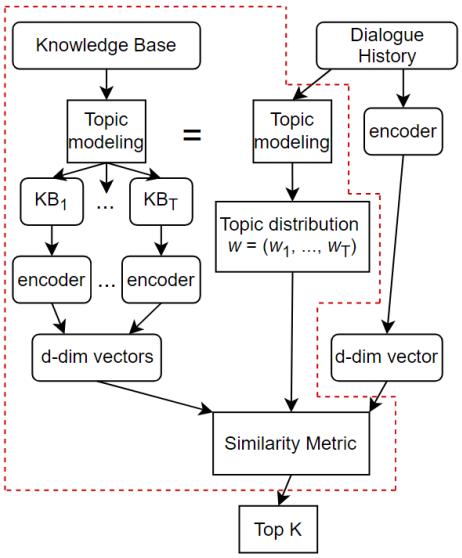

Figure 1: The modified DPR framework with our contribution highlighted. The two topic modeling modules are indeed the same one trained on the knowledge base.

Our contribution is threefold. First, we propose a modification utilizing topic modeling to the RAG model which has been widely used in knowledge-grounded generation that shows improved performance. Second, we investigate the vital parameter, the number of topics T, on two different knowledge-grounded datasets to show that (i) the right choice of T can improve the retrieval and generation performance, (ii) the optimal T is data-dependent and (iii) topic coherence is not a good indicator to find the best T. Finally, we show that combining our approach which manipulates the knowledge base with approaches that focus on building a better input query can further improve performance.

2 Related Work

For knowledge-grounded NLP, knowledge retrieval is a crucial step. Several works have shown that

the retrieval does not strictly have to be performed with a model which contains an explicit memory by embedding the concept of knowledge retrieval into LMs (Petroni et al., 2019; Heinzerling and Inui, 2021; Shin et al., 2020; Roberts et al., 2020). However, these retrieval-free models lack interpretability, and retrieve-then-generate models still yield higher performances in knowledge-intensive tasks (Petroni et al., 2021; Dinan et al., 2019). Our work follows this line of research, in which the response generation is based on the knowledge from a dedicated knowledge retrieval component.

Retrieval methods such as TF-IDF and BM25 (Robertson and Zaragoza, 2009) rely on keyword matching and can be considered as sparse retrieval. In contrast, dense retrieval methods encode text as a latent vector of much smaller dimensionality, in which the relevance of a knowledge passage to an input query is determined by the distance of their vectors. Recent dense retrieval approaches have outperformed the sparse methods (Karpukhin et al., 2020; Lewis et al., 2020b; Xiong et al., 2021). Our work is closely related to recent work on large-scale dense retrieval for dialogue. Specifically, we modify the retriever module and the way to calculate the similarity scores of the popular RAG model (Lewis et al., 2020b) by utilizing 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.

3 Method

We first perform **topic modeling** on the knowledge base. The topic model is used 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 between a knowledge passage and the query is calculated as dot product between their two vector representations.

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:

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

where $\cdot$ is dot product and $\times$ is multiplication[1]. Then, we choose the top-K from these K $\times$ T retrieved passages. We call this version **DPR-topic**.

To generate the final response, we use **Retrieval-Augmented Generation (RAG)** (Lewis et al., 2020b). It has a retriever (DPR) and a generator module (BART, Lewis et al. (2020a)). The retriever gets the most relevant passages given the dialogue history as an input query, and the generator takes the query and top-K passages as input 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 query encoder or the generator module in the original RAG model. Our model is called **RAG-topic**. An example comparing RAG-topic with RAG on a given dialogue history from WoW can be found in Appendix C.

[1]We tried different ways to utilize the topic distribution vector w in the formula of relevant scores between two vectors such as all zeros ($w_i = 0$) or one-hot vector for the most probable topic ($w_i = 1$ if $w_i > w_j \; \forall j \neq i$), but Equation 1 gives the best retrieval results.

	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 / 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
RAG-topic / Val.	71.7	72.0	**72.1**	**72.5**	**72.9**	71.1	71.3	71.9	68.0	67.5
RAG-context-topic / Val.	72.0	72.1	**72.2**	**72.6**	**72.7**	71.1	70.1	71.8	71.3	69.8

Table 1: Retrieval Results (R@5) on test and validation data of MultiDoc2Dial. **Bolded** results are significantly better than those in the same row with T=1 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	0.11	**0.15**	**0.21**	**0.32**	**0.34**	**0.35**	**0.29**	**0.33**	**0.38**	**0.35**
RAG-topic / Validation	36.3	36.2	**38.3**	**39.5**	**38.0**	**38.7**	30.5	30.3	25.5	23.4
RAG-topic / Seen Test	37.7	35.9	**38.5**	**40.6**	**40.3**	**40.1**	31.0	30.4	26.7	24.9
RAG-topic / Unseen Test	37.5	34.8	35.3	**39.9**	**39.5**	39.7	31.6	30.8	26.5	24.9

Table 2: Retrieval Results (R@5) of **RAG-topic** on Validation data, Seen and Unseen test data of WoW (average of 3 runs) with the same annotation as Table 1.

4 Experiment Setup

4.1 Datasets

We use two datasets of knowledge-grounded dialogues for this study. In both of them, one speaker in the conversation has to ground their response utterances in a specific knowledge *unit* from the knowledge base. **MultiDoc2Dial** (Feng et al., 2021) consists of around 4800 domain-specific dialogues in the style of information-seeking conversations, grounded in 488 documents from 4 domains. **Wizard of Wikipedia (WoW)** (Dinan et al., 2019) is a large chitchat dataset grounded in knowledge from Wikipedia with two test sets, seen and unseen where the latter has topics never seen before in train or validation. For consistency, we use the term *passage* to refer to the knowledge text spans we want to retrieve for response generation. Examples of the datasets can be found in Appendix B.

4.2 Evaluation and Models

For any RAG-based model, setting **T = 1** is equal to using the original model without our modifications.

For MultiDoc2Dial, to evaluate whether RAG-topic can add value to prior work on this corpus focusing on the dialogue history rather than the knowledge base (Tran and Litman, 2022), we develop **RAG-context-topic**. This approach uses RAG-topic as the model but also 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.

For WoW, we compare the generation performance of RAG-topic with two published baselines. **KnowledGPT** (Zhao et al., 2020) jointly optimizes the knowledge selection and response generation modules with pretrained LMs; **KnowExpert** (Xu et al., 2022) is an end-to-end model that directly injects the knowledge into pretrained language model (e.g., no knowledge extraction step) by using topic modeling to inform the GPT-2 adapters with more relevant "topics" during generation.

The evaluation metric for retrieval is Recall at 5 (R@5) as the generator from RAG uses the top-5 passages to create the response. For generation results, we use unigram-F_1 score between the generated and gold responses. 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).

Due to the randomness of the models (e.g. dropout from CTM training), we run each experiment 3 times and report the average results. Implementation details can be found in Appendix A.

5 Results

Tables 1 and 2 show the **passage retrieval results** with various numbers of topics chosen (T) for Multidoc2Dial and WoW respectively. The topic co-

Model	T = 1	T = 5
RAG-topic	41.1	41.3
RAG-context-topic	41.2	42.1

Table 3: Generation results (F_1) for MultiDoc2Dial, with and without topic modeling (T=5 vs T=1)

Model	WoW Seen	WoW Unseen
KnowledGPT	22.0* / 21.9	20.5* / 20.6
KnowExpert	18.7* / 18.5	16.7* / 16.7
RAG	21.9	20.2
RAG-topic	22.3	20.2

Table 4: Generation results (F_1) for WoW, T = 4. Results with * are numbers reported in the original papers.

herence scores are also reported in the first row of each table. For both tables, T = 1 equals using the original RAG models (no topic modeling).

For both datasets, with the right choices of the number of topics (T), our models can outperform the baseline counterparts with no topic modeling (T = 1). For instance, with T = 4 or T = 5, all models achieve higher R@5 than their non-topic-modeling versions in both tables. On the other hand, certain Ts yield lower results compared to the baselines. For example, with T = 10, all models significantly underperform the baseline T = 1. Also, the best T is consistent among the same dataset but different across datasets. Specifically, the best results are with T = 5 in both tested models for Multidoc2Dial, while T = 4 provides the best RAG-topic results on both WoW Seen and Unseen test data. Additionally, for each dataset, the best T values on validation data and test data are identical. This suggests that the optimal number of topics T is data-dependent and should be tuned on validation data.

In contrast, for both datasets, higher scores in topic coherence do not necessarily lead to higher retrieval results. For MultiDoc2Dial, the best R@5 is at T = 5 while the highest topic coherence is at T = 4. For WoW, T = 9 has the highest topic coherence score, but its models perform worse than the baselines on both seen and unseen data, let alone the best models at T = 4.

Results with RAG-context-topic in Table 1 show that with the right T, our approach which manipulates the knowledge base side compliments the approach that manipulates the input query side from Tran and Litman (2022). RAG-context-topic outperforms the original model in the same row (T = 1) in multiple values of T (e.g., 2, 3, 4, 5) as well as our proposed model when used in isolation (RAG-topic vs RAG-context-topic in the same column).

Table 3 shows the **response generation results** on Multidoc2Dial. For each model, we report the result from T=1 (no topic modeling) and the best T from Table 1 (T = 5). With our proposed approach, all models consistently outperform their baseline versions, even though the gain is very small (less than 1). Similar to Feng et al. (2021); Tran and Litman (2022), the increases in retrieval results do not really transfer to generation performances.

In Table 4, we report the generation results of our best model (T = 4) on the WoW Seen and Unseen test data. Compared to the original RAG with no topic modeling, our RAG-topic approach achieved a higher score on Seen and an equivalent score on Unseen. This implies that utilizing our approach does not decrease generation performances. Our model outperforms KnowledGPT in the Seen data but has a lower F_1 score on the Unseen set. Although KnowExpert uses topic modeling in their approach, the performance is lower than the retrieval-based models, including ours.

6 Conclusion

In this work, we proposed a simple method that utilizes topic modeling on the knowledge base to improve the performance of RAG-based dense retrieval models. Our approach re-uses the same RAG framework but uses topic modeling to cluster the knowledge base. We then build a separate document encoder for each cluster in the knowledge base and incorporate the topic distribution weights into the calculation of similarity scores. The results show that the number of topics T is an important parameter that can affect the retrieval results, either positively or negatively. Additionally, the results suggest that topic coherence is not a good indicator of the optimal T as higher scores do not always lead to better retrieval performances. We also believe that the optimal number of topics is data-dependent since the best Ts are different for the two experimented datasets. Overall, with the right T, we achieve improvement in both retrieval and generation, although the gain in generation performance is small. Future plans include utilizing multi-task training with similar knowledge-intensive tasks and using better generative modules to take advantage of the improved retrieval results.

Limitations

One major limitation of our proposed approach is that the computational requirement is proportional to the number of topics T as we need to retrieve K knowledge passages from each knowledge base cluster in order to get the final top-K. Therefore, this method does not scale well if the optimal T is large. Also, the relation between topic coherence and the best T found in this work is constrained to the metric used to calculate the topic coherence, which is external word embeddings (Ding et al., 2018). Additionally, for generation results, this work lacks human evaluation and analysis of the poor increment of response generation (Tables 3 and 4) despite improvement in knowledge retrieval (Tables 1 and 2).

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.

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.

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.

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.

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.

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, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

A Implementation Details

To train the topic model, we use CTM (Bianchi et al., 2021) and follow the training details from Xu et al. (2022).

For RAG, we use DPR [2] to build the encoder and create the index for each cluster in the knowledge base. This process also initializes the query encoder for RAG. Then, we modify the retriever of RAG [3] to get the top-K passages as described in Section 3. We use the default hyperparameters for these models. For RAG-topic-context, since it only changes the input query to the RAG model, we modify the code provided by Tran and Litman (2022) in the same way we modify the RAG model. KnowledGPT (Zhao et al., 2020) and KnowExpert (Xu et al., 2022) are re-run by using the checkpoint from the source code provided in the original papers without any modification. All models were trained on an RTX 3090 card.

B Examples from Datasets

Figures 2 and 3 show one example each from our two datasets, Multidoc2Dial and WoW, respectively. Notice that for the WoW dataset, we do not take advantage of the topic given in the dataset. We instead assume that no topics are given during the conversation and the relevant knowledge passages need to be found from the entire knowledge base.

C Examples of Retrieved Passages and Response Generation

Table 5 shows the list of keywords of each cluster from WoW when the number of topic T for CTM is set as 4. In Table 6, we show the top-1 retrieved passage and generated response from RAG and RAG-topic for a given dialogue history in WoW. 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

[2] https://github.com/facebookresearch/DPR

[3] https://github.com/huggingface/transformers/tree/main/src/transformers/models/rag

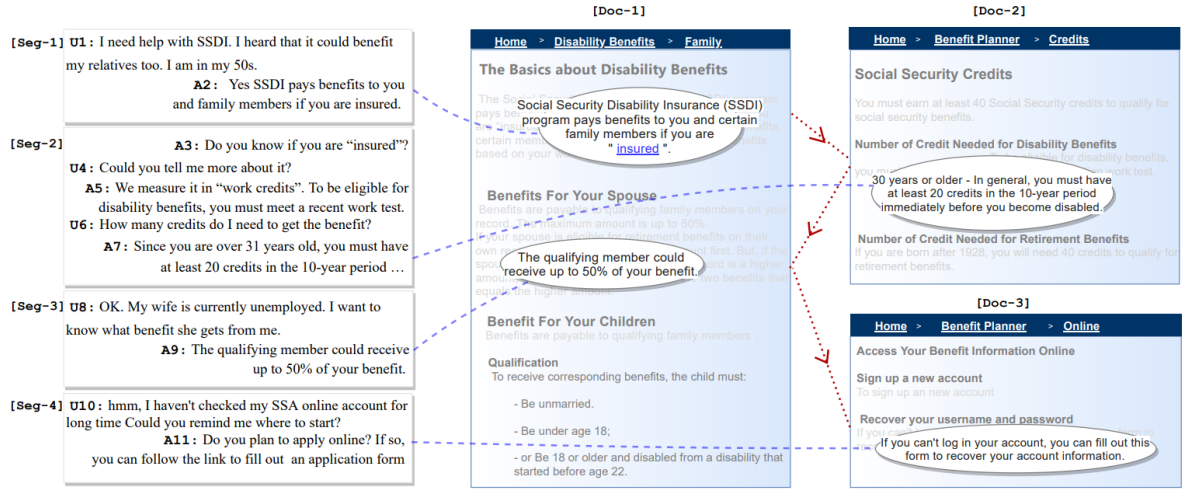

Figure 2: An example dialogue from Multidoc2Dial borrowed from Feng et al. (2021). The conversation (on the left) is grounded in 3 documents Doc-1, Doc-2, and Doc-3. Each dialogue segment indicates that all turns within it are grounded in the same document (e.g., A3 to A7 in Seg-2 are all grounded in 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).

Topic:	Armadillo
Wizard:	I love animals and think armadillos are awesome with their leathery shell.
Apprentice:	I don't think I've ever seen an armadillo in real life!
Wizard:	I've seen them at the zoo. Armadillo means little armored one in Spanish.
Apprentice:	Are they native to a Spanish-speaking part of the world?
Knowledge:	Armadillos are New World placental mammals in the order Cingulata …
	The word "armadillo" means "little armoured one" in Spanish.
	…
	The nine-banded armadillo ("Dasypus novemcinctus"), or the nine-banded, long-nosed armadillo, is a medium-sized mammal found in North, Central, and South America.
Wizard:	Yes, they are most commonly found in North, Central, and South America

Figure 3: An example dialogue from WoW copied from Dinan et al. (2019). Two speakers talk about a given topic (e.g., Armadillo). In the data collection process, only the wizard has access to an information retrieval system over Wikipedia (around 61 knowledge candidates per turn) to make statements relevant to the conversation. The knowledge passage chosen by the wizard is highlighted in blue. However, in this study, we assume the information about the topic is not given to the speakers and perform the retrieval on the entire knowledge base.

original RAG model found an irrelevant knowledge passage and generated an inappropriate response.

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

Table 5: Top 10 words for each cluster of the knowledge base on WoW

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 with T = 1)	**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 6: An example from WoW 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 5 from the dialogue history.