Enhancing Knowledge Retrieval with Topic Modeling for Knowledge-Grounded Dialogue

Presenter: Nhat Tran

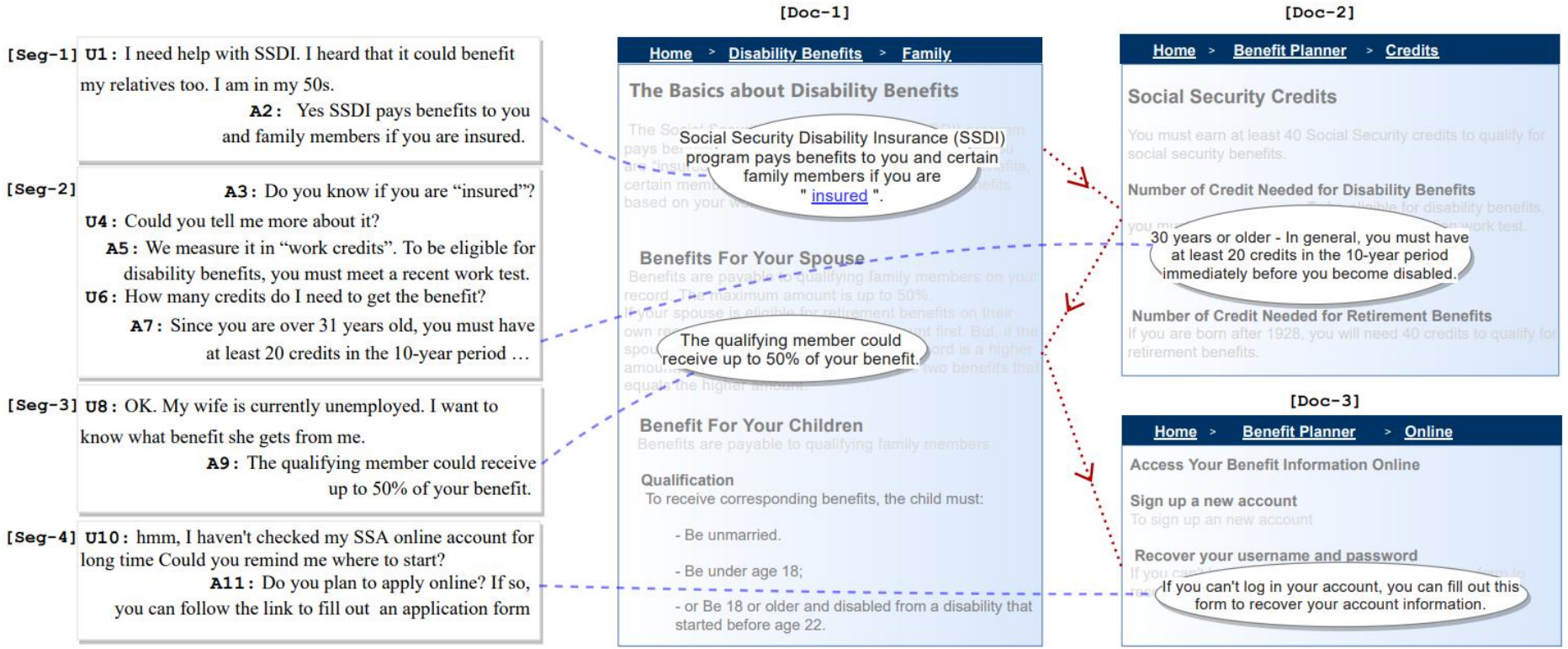
University of Pittsburgh

nlt26@pitt.edu

Background and Motivation

Knowledge-Grounded Dialogue

- Knowledge source is given (e.g., Wiki, websites, ...)
- Responses need to be grounded in the knowledge source
- Large knowledge base => need to find the relevant information



Example from Multidoc2Dial dataset (Feng et al., 2021)

Contributions

- Improve retrieval with topic modeling in the knowledge source
- Suggest a way to find the optimal number of topics T
- Experiment with ChatGPT as a response generator

Why Topic Modeling?

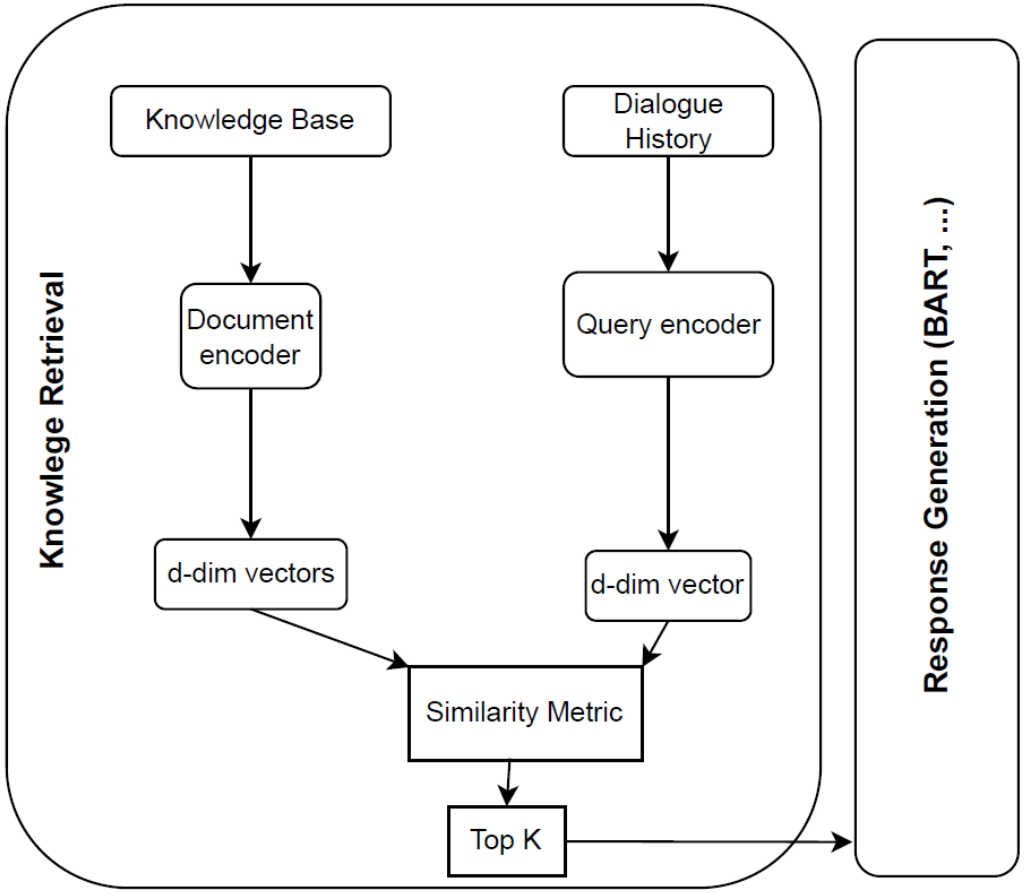
- Structural information of the knowledge base: topics
- Topics of the dialogue history can also be identified
- Retrieval should be guided toward related topics

Method

Retrieval Augmented Generation (RAG)

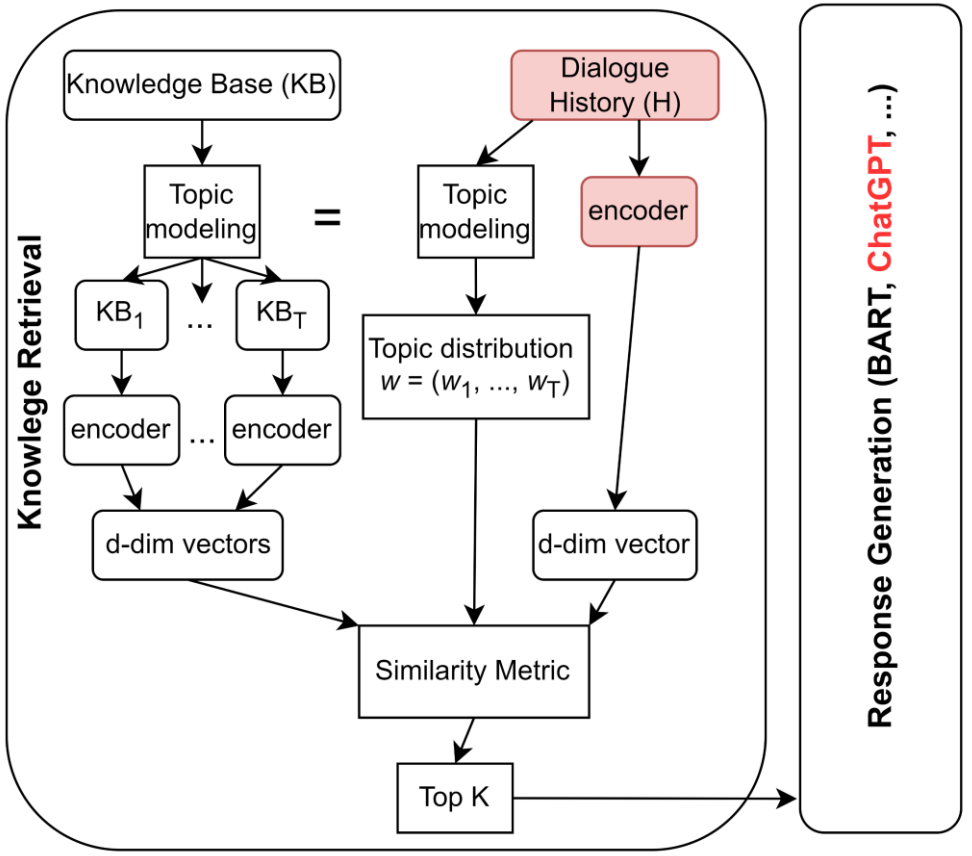
- RAG (Lewis et al., 2022)
- Dense encoders (e.g., BERT): map text to a d-dimensional vector
  - Document encoder encodes the passages: $E_d(p)$
  - Query encoder encodes the dialogue history: $E_q(H)$
- Similarity score to find top-K passages

$$sim(p, H) = E_q(H) \cdot E_d(p)$$



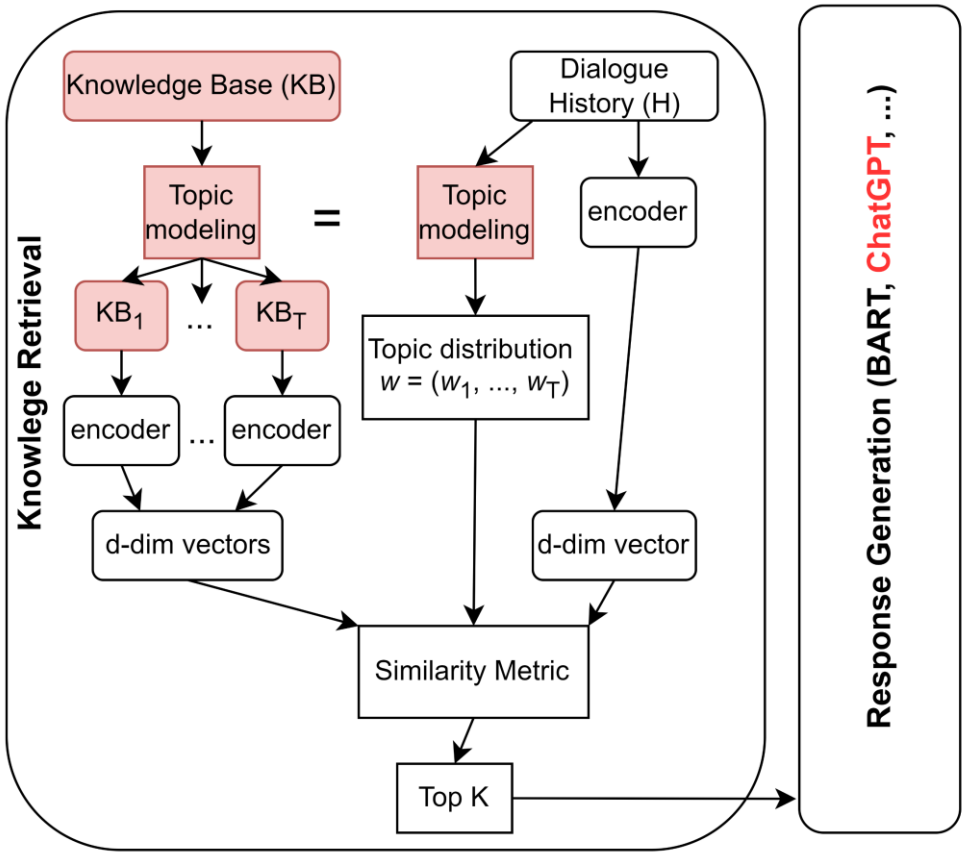
Method

- Shared query encoder $BERT_q(H)$



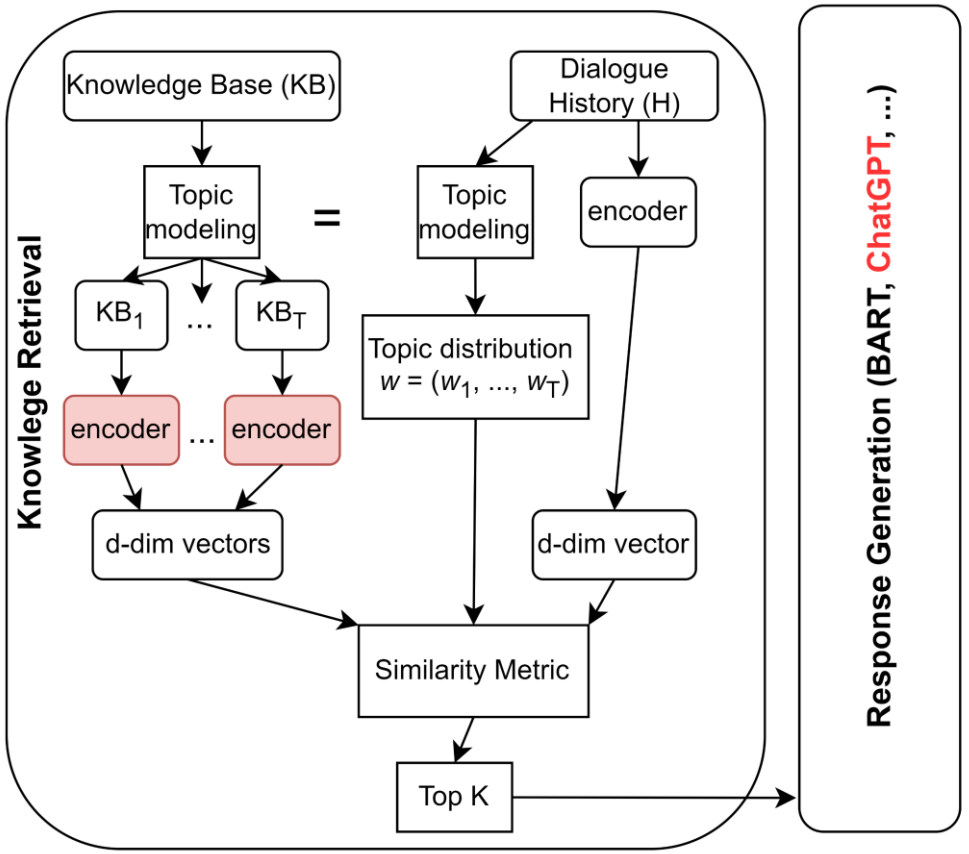
Method

- Shared query encoder $BERT_q(H)$
- Train a topic model and perform topic modeling on the knowledge base (KB)
  - Use CTM (Bianchi et al, 2021)
  - Cluster the KB into **T** topic clusters
  - The topic model can output a T-dimension vector $w = (w_1, w_2, ..., w_T)$ for the dialogue history H



Method

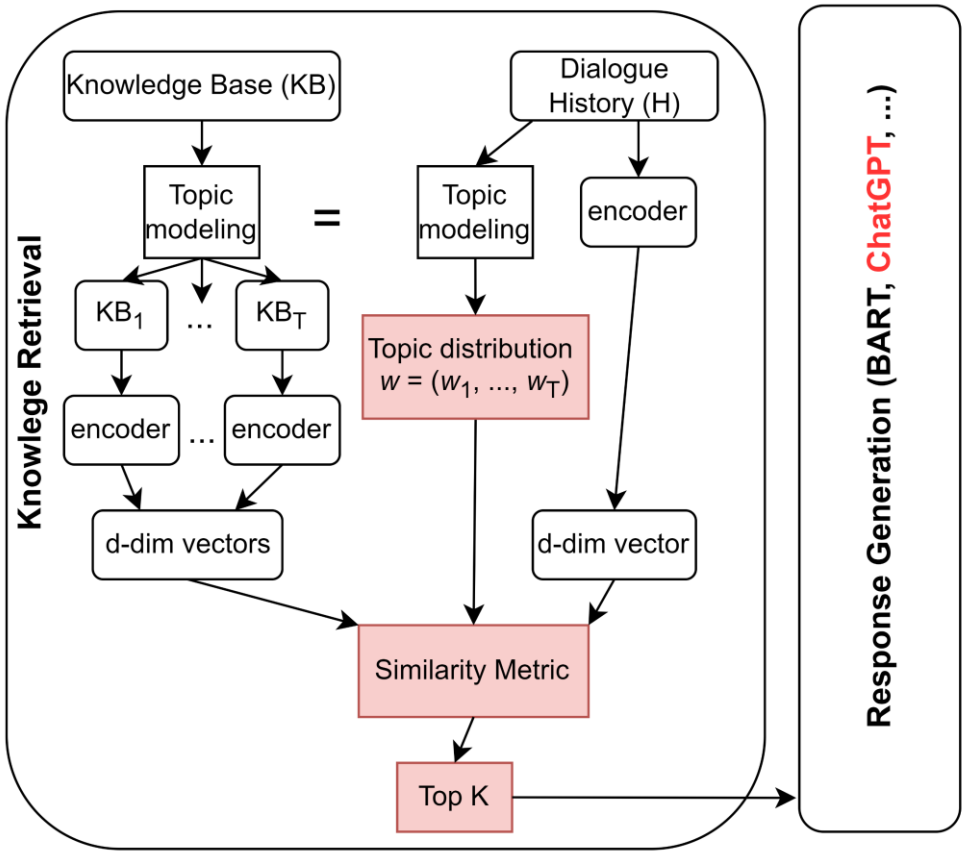
- Shared query encoder $BERT_q(H)$
- Train a topic model and perform topic modeling on the knowledge base (KB)
  - Use CTM (Bianchi et al, 2021)
  - Cluster the KB into **T** topic clusters
  - The topic model can output a T-dimension vector $w = (w_1, w_2, ..., w_T)$ for the dialogue history H
- Train a document encoder $BERT_d^i$ for each cluster *i* (i = 1, 2, ..., T).



Method

- Shared query encoder $BERT_q(H)$
- Train a topic model and perform topic modeling on the knowledge base (KB)
  - Use CTM (Bianchi et al, 2021)
  - Cluster the KB into **T** topic clusters
  - The topic model can output a T-dimension vector $w = (w_1, w_2, ..., w_T)$ for the dialogue history H
- Train a document encoder $BERT_d^i$ for each cluster *i* (i = 1, 2, ..., T).
- To find top-K passages

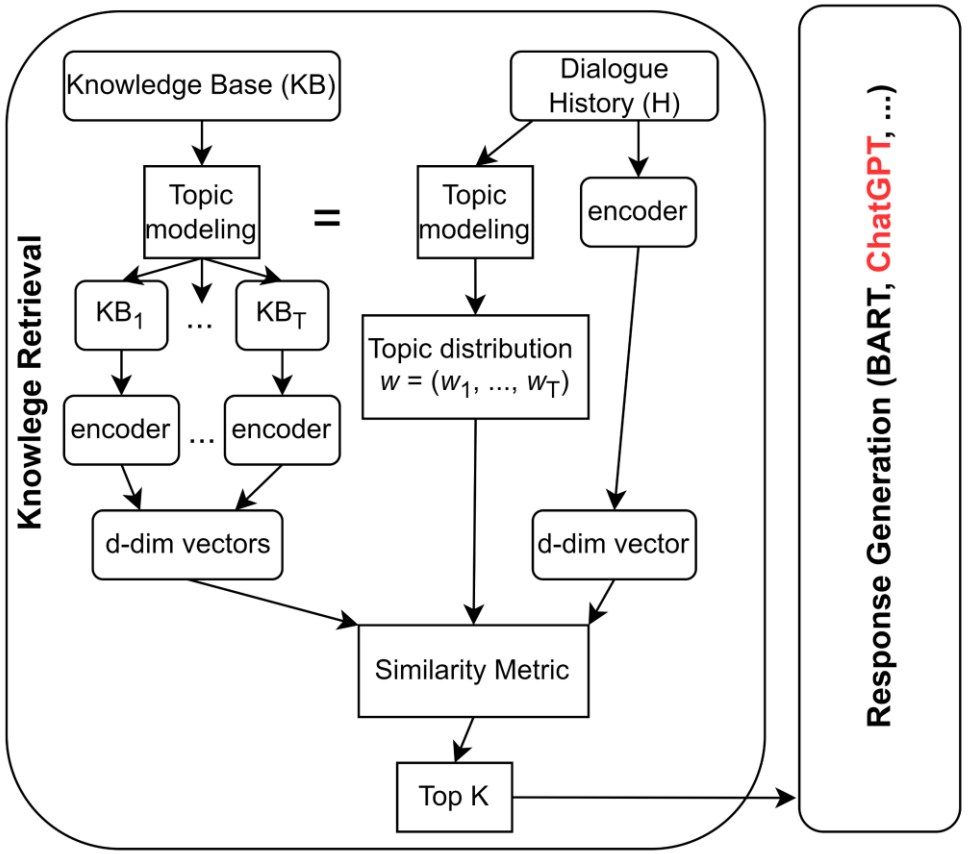
$$BERT_q(H) \cdot BERT_d^i(p) \times w_i$$



Method

- Shared query encoder $BERT_q(H)$
- Train a topic model and perform topic modeling on the knowledge base (KB)
  - Use CTM (Bianchi et al, 2021)
  - Cluster the KB into **T** topic clusters
  - The topic model can output a T-dimension vector $w = (w_1, w_2, ..., w_T)$ for the dialogue history H
- Train a document encoder $BERT_d^i$ for each cluster *i* (i = 1, 2, ..., T).
- To find top-K passages

  $BERT_q(H) \cdot BERT_d^i(p) \times w_i$
- Optional: ChatGPT for response generation



Experiment

Datasets

- Multidoc2Dial (Feng et al., 2021)
  - Grounded in FAQ webpages
  - 4 domains: Department of Motor Vehicles, Social Security Affairs, Student Aid, Veteran Affairs
  - Information-seeking style
- KILT-dialogue (Petroni et al., 2021)
  - Grounded in Wikipedia
  - Chit-chat style

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

Models

- For both datasets: RAG (Lewis et al., 2022)
  - Our approach: RAG-topic
  - GPT-4 for generation
- Multidoc2Dial: RAG-context (Tran and Litman, 2022)
  - RAG-context: 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
  - Our approach: RAG-context-topic
- KILT-dialogue: BART+DPR as a baseline

Evaluation Metrics

- To find the number of topics T
  - Topic Coherence (Ding et al., 2018)
  - Recall at 5 (R@5)
- Downstream evaluation
  - Retrieval: Page-level Precision at 1 (P@1)
  - Generation
    - unigram F_1
    - KILT-F_1

Results and Discussion

Number of Topics (T)

- T is dependent on the dataset

	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

R@5 on Multidoc2Dial

	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

R@5 on KILT-dialogue

Number of Topics (T)

- T is dependent on the dataset
- Topic coherence is not a good metric

	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

R@5 on Multidoc2Dial

	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

R@5 on KILT-dialogue

Number of Topics (T)

- T is dependent on the dataset
- Topic coherence is not a good metric
- Using a validation set to find T is recommended

	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

R@5 on Multidoc2Dial

	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

R@5 on KILT-dialogue

Retrieval Results

- Our models significantly outperform the baseline counterparts

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

Multidoc2Dial (T = 5)

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

KILT-dialogue (T = 4)

Retrieval Results

- Our models significantly outperform the baseline counterparts
- For Multidoc2Dial, our model complements an approach (RAG-context) that manipulates the input query (i.e., dialogue history)

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

Multidoc2Dial (T = 5)

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

KILT-dialogue (T = 4)

Generation Results

- Our models significantly outperform the baseline counterparts

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

Multidoc2Dial (T = 5)

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

KILT-dialogue (T = 4)

Generation Results

- Our models significantly outperform the baseline counterparts
- GPT-4 without external knowledge performs poorly on Multidoc2Dial

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

Multidoc2Dial (T = 5)

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

KILT-dialogue (T = 4)

Generation Results

- Our models significantly outperform the baseline counterparts
- GPT-4 without external knowledge performs poorly on Multidoc2Dial
- GPT-4 with the better retriever always win

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

Multidoc2Dial (T = 5)

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

KILT-dialogue (T = 4)

Conclusion

- We proposed a method that utilizes topic modeling on the knowledge base to improve the performance of RAG-based models
- Finding T using the validation set is more reliable than metrics such as topic coherence
- ChatGPT might not perform very well without external knowledge, but it is superior when knowledge is provided

Thank you!