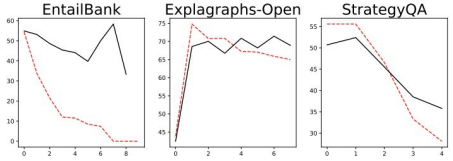
KAIST AI
Kim Jaechul Graduate School

Generative Multi-hop Retrieval

Hyunji Lee, Sohee Yang, Hanseok Oh, Minjoon Seo
KAIST AI

Limitations of Bi-Encoder in Multi-hop

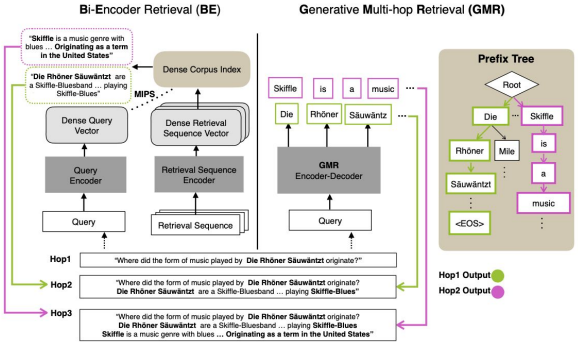


1.BE performance (y-axis) degrades as the input gets longer since the input is encoded into a fixed-sized vector (*bottleneck problem*). It becomes severe as the input length increases with number of hops (x-axis). Red is BE and Black is GMR.

	Minor			Major		
	Str	**Exp**	**Ent**	**Str**	**Exp**	**Ent**
BE	23.6%	46.9%	14.0%	71.2%	91.1%	55.1%
GMR	1.7%	49.3%	11.1%	20.7%	75.8%	39.6%

2. It is *highly vulnerable to error propagation*

Generative Multi-hop Retrieval



GMR is an encoder-decoder model that performs multi-hop retrieval by *iteratively generating the entire target sequences*. It overcomes the bottleneck problem by interacting in the *whole parametric space of the model* rather than the L2 or inner product space as in the bi-encoder approach.

Contributions

1. We show the limitations of bi-encoder retrieval in multi-hop retrieval tasks
2. We show that Generative Multi-hop Retrieval (GMR) which is especially strong in multi-hop retrieval settings close to real-world scenarios and datasets with a low unseen rate.
3. We introduce multi-hop memorization which effectively memorizes the target corpus and improves the performance of GMR.

Two Memorization Methods of GMR

We propose memorization methods to *reduce the unseen rate*.
(1) **LM Memorization**: intermediate task of training on all retrieval candidates with *standard LM objective*
(2) **Multi-hop Memorization**: augments the training data with *pseudo-multi-hop queries*.

Experimental Results

	EntailTree			StrategyQA				EG-Open				RT-Open*		
	ST5	GMR	GMR_L	ST5	GMR	GMR_L	GMR_M	ST5	GMR	GMR_L	GMR_M	ST5	GMR	GMR_L
Fixed R@5	31.5	53.6	**54.3**	37.4	44.9	45.5	**45.6**	27.0	32.9	32.4	**34.6**	-	-	-
Dynamic F1@5	24.9	**48.2**	47.4	38.1	41.9	42.6	**43.1**	25.0	35.5	35.7	**36.2**	-	-	-
Dynamic F1@10	19.4	**52.1**	51.7	36.9	44.3	45.0	**45.2**	24.6	40.0	40.8	**42.1**	-	-	-
Dynamic F1@20	16.9	**52.5**	52.2	36.5	46.6	47.1	**47.9**	25.4	41.5	41.3	**42.6**	17.0	51.0	**65.5**

Method	DPR	MDR-	MDR	fix-GMR	fix-GMR_L
Top-2	25.2	59.9	65.9	57.7	55.0
Top-10	45.4	70.6	77.5	68.8	65.3
Top-20	52.1	73.1	80.2	73.9	71.4

Across all five multi-hop datasets, GMR consistently achieves comparable or higher performance than bi-encoder models while demonstrating *more efficient GPU memory usage (-79.5%) and storage footprint (-69.7%).*
GMR show especially high performance on settings close to real-world scenarios (*Dynamic F1*).

Model	EntailTree	StrategyQA	EG-Open
atomic-DSI*	28.0	-	23.4
naive-DSI*	7.7	-	8.6
fix-GMR	53.6	44.9	32.9

We also show the *importance of explicitly generating all retrieval sequences on multi-hop* retrieval tasks by comparing the performance with DSI, where GMR consistently shows higher performance