Explainable Multi-hop Question Generation: An End-to-End Approach without Intermediate Question Labeling

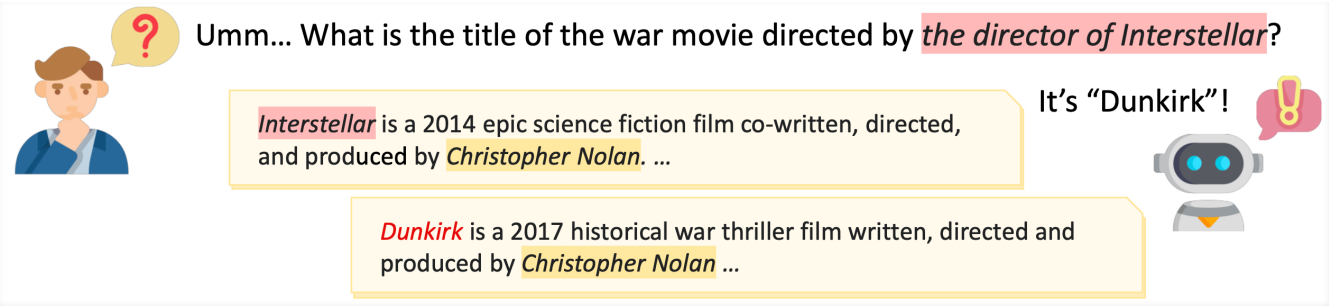
Seonjeong Hwang[1], Yunsu Kim[3], Gary Geunbae Lee[1,2]

[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
[3]aiXplain, Inc. Los Gatos, CA, USA

Task Definition

Multi-hop Question Generation

: Generating questions that require complex reasoning by gathering related information scattered across multiple contexts.



Document A

Dunkirk is a 2017 historical war thriller film written, directed and produced by Christopher Nolan that depicts the Dunkirk evacuation of World War II from the perspectives of the land, sea and air. Midway is a 2019 war film about the Battle of Midway, …

Document B

Interstellar is a 2014 epic science fiction film co-written, directed, and produced by Christopher Nolan. …

Document C

Kip Stephen Thorne (born June 1, 1940) is an American theoretical physicist known for his contributions in gravitational physics and astrophysics. … He continues to do scientific research and scientific consulting, most notably for the science fiction film Interstellar.

Answer Dunkirk

1-hop Question What is the title of the war film directed by Christopher Nolan?

2-hop Question What is the title of the war film directed by the director of Interstellar?

3-hop Question What is the title of the war film directed by the director who received *advice from Kip Thorne in making a science fiction movie*?

Recent Study

1. End-to-end Question Generation

: The input documents and answers are encoded using Transformer-based encoders or Graph neural networks, and then the encoded representation is utilized to decode multi-hop questions[1,2,3].

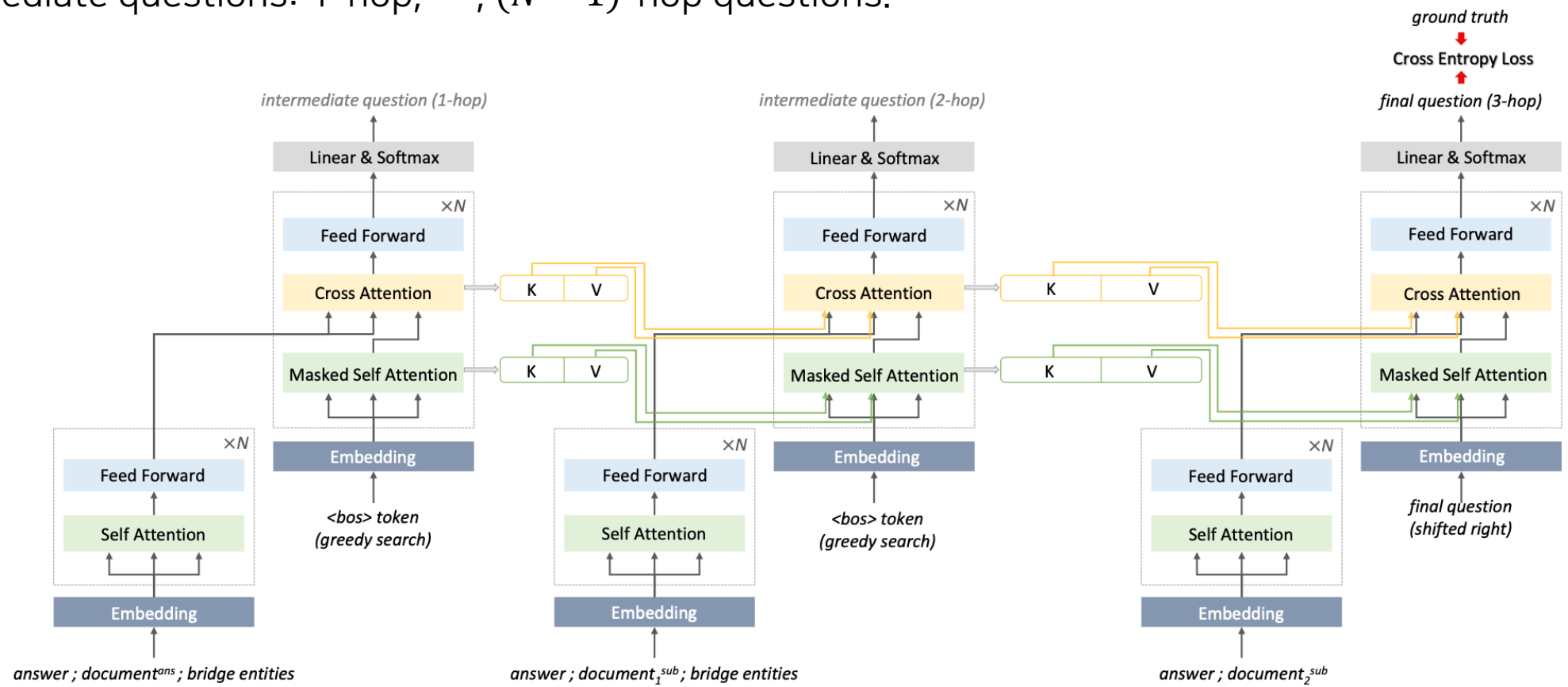
2. Step-by-step Question Rewriting

: A one-hop question is generated by a question generation model, and then question rewriting model progressively increases the question complexity based on new input documents[4].

[1] Pan, Liangming, et al. "Semantic Graphs for Generating Deep Questions." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.
[2] Su, Dan, et al. "Multi-hop Question Generation with Graph Convolutional Network." *Proceedings of Findings of the Association for Computational Linguistics (EMNLP)* (2020).
[3] Fei, Zichu, et al. "CQG: A simple and effective controlled generation framework for multi-hop question generation." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.
[4] Cheng, Yi, et al. "Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

End-to-end Question Rewriting: E2EQR

Given N documents and a target answer, the goal is to generate a N-hop question along with a sequence of intermediate questions: 1-hop, ⋯ , $(N-1)$-hop questions.

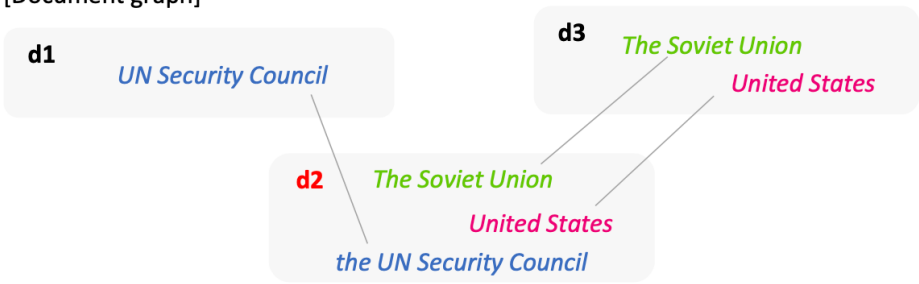


Methodology

Document Arrangement & Bridge Entity Extraction

d1	*UN Security Council* Resolution 731, adopted unanimously on 21 January 1992, after recalling resolutions 286 (1970) and 635 (1989) which condemned acts of terrorism, ...
d2	... On 27 August 67th Fighter Squadron ***aircraft mistakenly attacked facilities in Chinese territory*** and *the Soviet Union* called *the UN Security Council*'s attention to China's complaint about the incident. The *United States* proposed that a commission of India and Sweden determine ...
d3	The "Miracle on Ice" refers to a medal - round game during the men's ice hockey tournament at the 1980 Winter Olympics in Lake Placid, New York, played between the hosting *United States*, and the four - time defending gold medalists, *the Soviet Union*.

[Document graph]

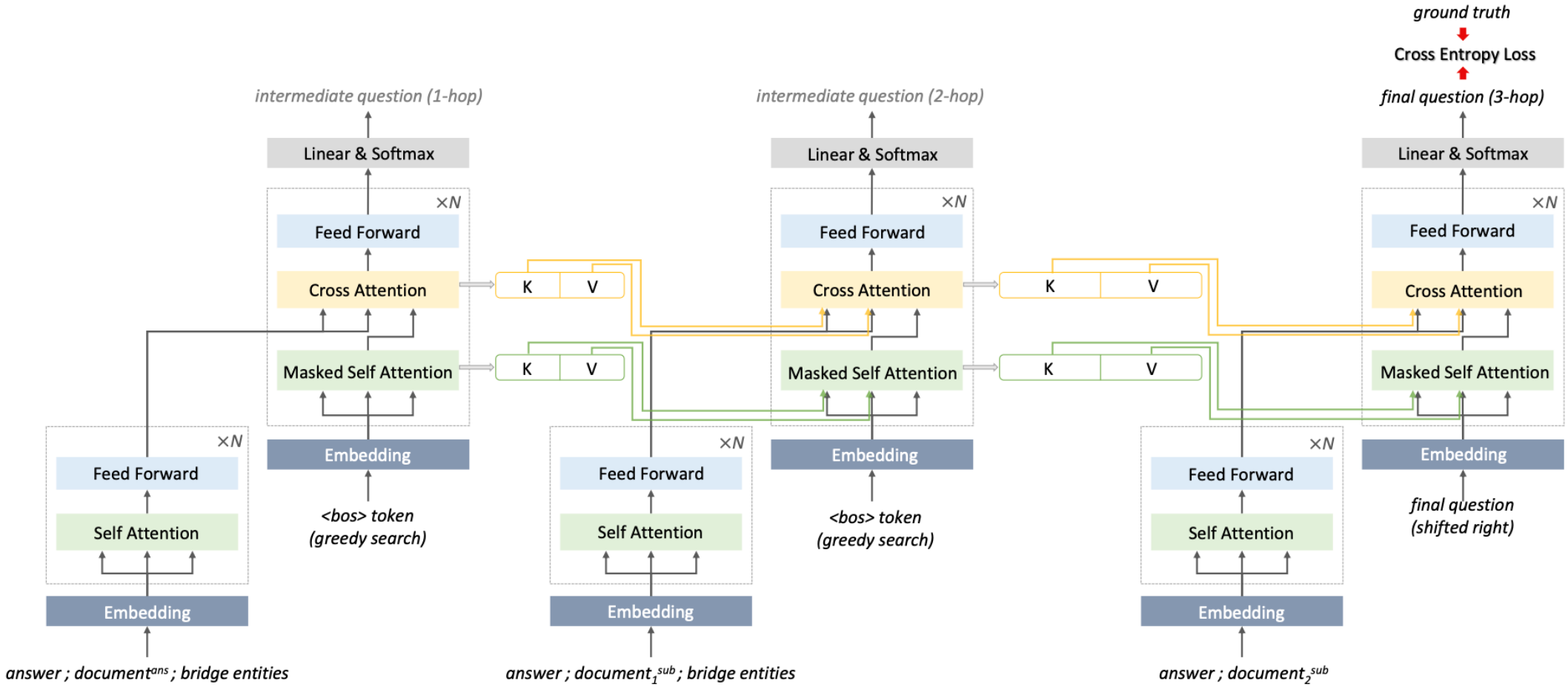


[Graph serialization]

Order	Document	Bridge entities
1st	**d2**	*UN Security Council, The Soviet Union, United States*
2nd	**d1**	*The Soviet Union, United States*
3rd	**d3**	-

Methodology

End-to-end Question Rewriting



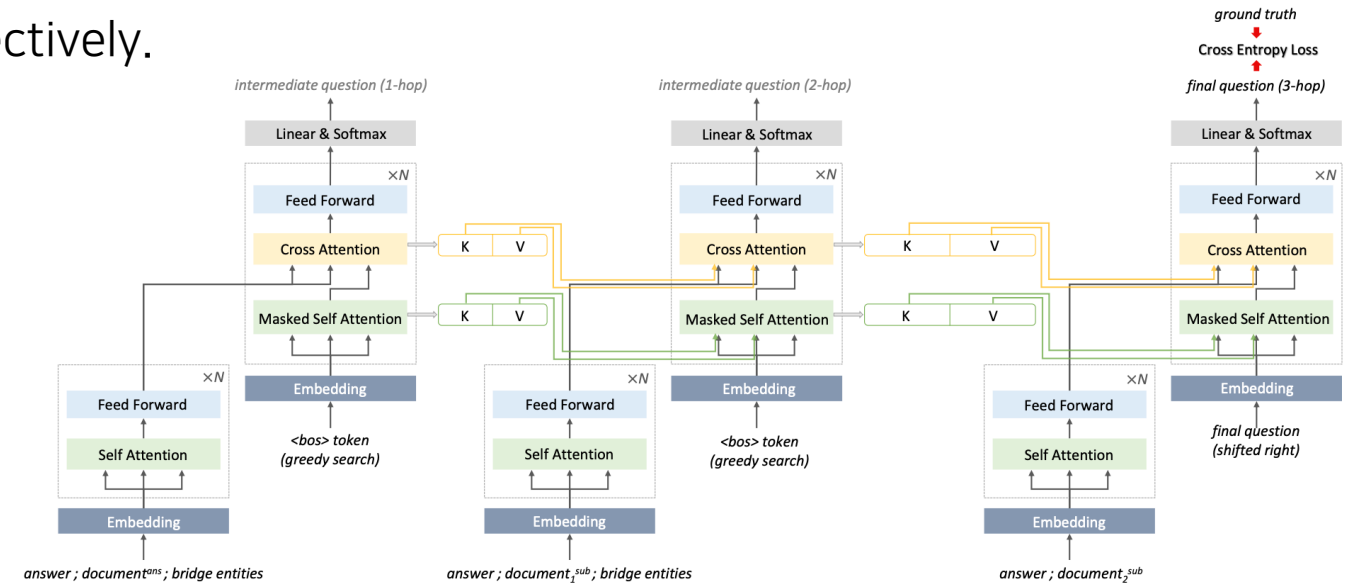
Methodology

End-to-end Question Rewriting

- E2EQR consists of accumulated masked self-attention (SA) and cross-attention (CA) blocks.
- The accumulated key and value matrices ($K_{1:t}$ and $V_{1:t}$) are used in these attention mechanism:

$$\text{Attention}(Q_t, K_{1:t}, V_{1:t}) = \text{Softmax}(\frac{Q_t K_{1:t}^T}{\sqrt{d_k}})V_{1:t}$$

- The accumulated SA and CA allow the use of the information from the intermediate questions and input elements of the prior steps, respectively.



Methodology

Adaptive Curriculum Learning

- The main complexity: $\mathcal{H}$
- Training examples grouped by the question complexity : $\{D_1, D_2, \cdots, D_{\mathcal{N}}\}$
- The suppression ratio ρ for the examples with higher complexity levels than $\mathcal{H}$
- Loss weights for lower and higher complexity levels : γ_{low} and γ_{high}

Algorithm 1 Adaptive Curriculum Learning

Input: $\{D_1, D_2, \cdots, D_{\mathcal{N}}\}$, E2EQR$_\theta$, α, γ_{low}, γ_{high}, ρ

for $\mathcal{H} = 1$ to $\mathcal{N}$ **do**
  Training examples $\mathfrak{D} \leftarrow D_1 \cup \cdots \cup D_{\mathcal{H}}$
  for $h = \mathcal{H} + 1$ to $\mathcal{N}$ **do**
    $\mathfrak{D} \leftarrow \mathfrak{D} \cup D_h^{part}$, where $n(D_h^{part}) = \rho \cdot n(D_h)$
  end for
  for $i = 1$ to $n(\mathfrak{D})$ **do**
    $x_i, y_i \leftarrow \mathfrak{D}_i$
    $\mathcal{L}_i \leftarrow$ *CrossEntropyLoss*(y_i, E2EQR$_\theta(x_i)$)
    if $i < \mathcal{H}$ **then**
      $\mathcal{L}_i \leftarrow \gamma_{low} \cdot \mathcal{L}_i$
    end if
    if $i > \mathcal{H}$ **then**
      $\mathcal{L}_i \leftarrow \gamma_{high} \cdot \mathcal{L}_i$
    end if
  end for
  $\mathcal{L} \leftarrow \sum_{i=1}^{n(\mathfrak{D})} \mathcal{L}_i / n(\mathfrak{D})$
  $\theta \leftarrow \theta - \alpha \cdot \frac{\partial \mathcal{L}}{\partial \theta}$
end for

Experiments

Datasets

- MuSiQue[1]
- HotpotQA[2]

Metrics

- **Automatic evaluation:** BLEU4, METEOR, ROUGE-L
- **Human evaluation:** Fluency, Complexity, Answer Matching

[1] Trivedi, Harsh, et al. "♫ MuSiQue: Multihop Questions via Single-hop Question Composition." *Transactions of the Association for Computational Linguistics* 10 (2022): 539-554.
[2] Yang, Zhilin, et al. "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

Main Results

Automatic Evaluation & Human Evaluation

Model	2-hop			3-hop			4-hop			Intermediate Question
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L	
DP-Graph (Pan et al., 2020)	5.14	10.80	28.69	4.33	10.37	28.33	4.47	9.87	28.01	
CQG (Fei et al., 2022)	9.64	16.20	33.98	7.79	13.77	31.12	5.14	11.93	26.48	
MulQG (Su et al., 2020)	9.56	15.41	37.18	9.23	14.35	35.66	7.13	12.43	31.88	
BART (Lewis et al., 2020)	20.84	25.91	43.81	17.64	22.93	41.16	16.11	20.63	37.02	
E2EQR	20.33	25.64	44.01	17.02	22.33	40.04	15.34	19.78	36.98	✓

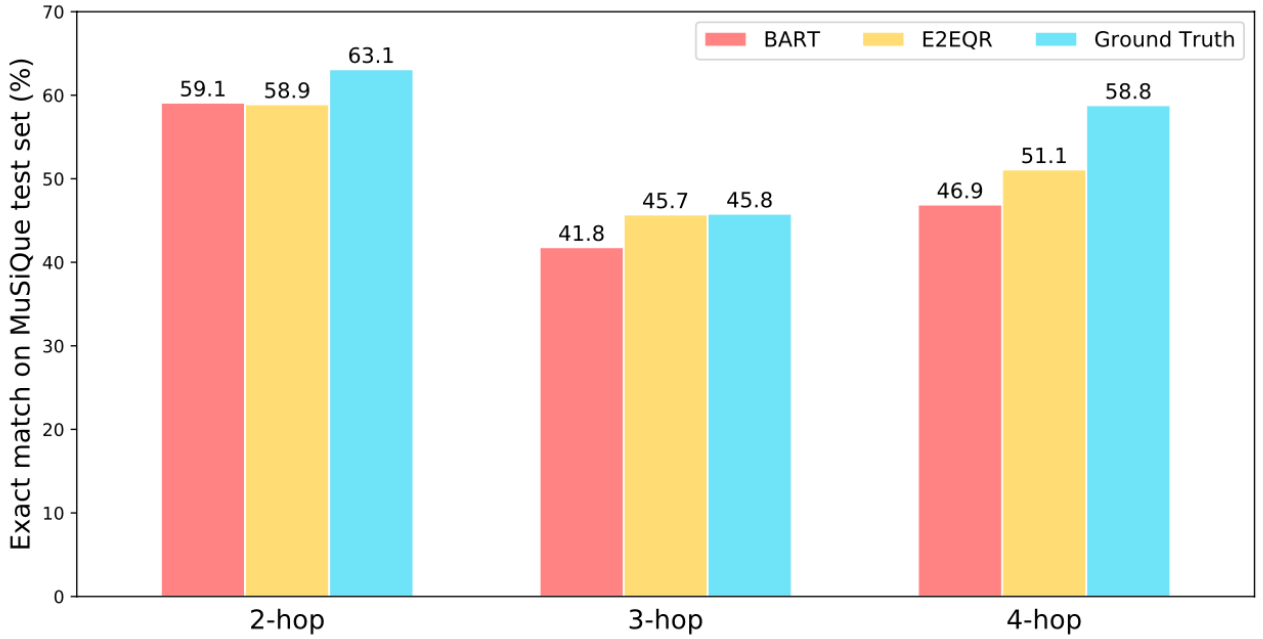
Table 1: Automatic evaluation results on the MusiQue test set.

Model	2-hop			3-hop			4-hop		
	Fluency	Complexity (≤ 2)	Answer Matching	Fluency	Complexity (≤ 3)	Answer Matching	Fluency	Complexity (≤ 4)	Answer Matching
BART	4.80	1.93	87.5%	4.88	2.54	75.0%	4.78	2.92	73.8%
E2EQR	4.92	1.99	87.5%	4.83	2.50	82.5%	4.60	2.96	80.0%
Ground Truth	4.85	1.99	95.0%	4.85	2.65	78.8%	4.65	3.29	77.5%

Table 2: Human evaluation of questions generated by multi-hop QG models and the ground truth.

Analysis

Multi-hop QA Data Augmentation



Analysis

Ablation Study

Model	2-hop	3-hop	4-hop
E2EQR	44.61	40.31	38.41
E2EQR w/o Accumulated SA	43.32	39.94	35.14
E2EQR w/o Accumulated CA	41.84	36.69	35.35

Analysis

Examples of generated questions

Answer: July 22, 1864

[Document A] Battle of Atlanta

The Battle of Atlanta was a battle of the Atlanta Campaign fought during the American Civil War on July 22, 1864, just southeast of Atlanta, Georgia. Continuing their summer campaign to seize the important rail and supply center of Atlanta, Union forces commanded by William Tecumseh Sherman overwhelmed and defeated Confederate forces defending the city under John Bell Hood. . . .

Intermediate Question (1-hop): When did the battle of Atlanta happen?

[Document B] List of municipalities in Georgia

The largest municipality by population in Georgia is Atlanta with 420,003 residents, and the smallest municipality by population is Edge Hill with 24 residents. . . .

Intermediate Question (2-hop): When did the Battle of the largest municipality in Georgia happen?

[Document C] WEKL

WEKL, known on-air as "K-Love", is a Contemporary Christian radio station in the United States, licensed by the Federal Communications Commission (FCC) to Augusta, Georgia, broadcasting on 102.3 MHz with an ERP of 1.5 kW. . . .

Final Question (3-hop): When did the Battle of the largest municipality in the state WEKL broadcasts in happen?

E2EQR

Answer: *Milledgeville*

[Document 1] Blackberry Smoke is an American rock band from Atlanta, Georgia, United States. ...

[Document 2] WEKL, known on-air as "102.3 K-Love", is a Contemporary Christian radio station in the United States, ...

[Document 3] Georgia has had five different capitals in its history. The first was Savannah, the seat of government during British colonial rule, followed by Augusta, Louisville, *Milledgeville*, and Atlanta, the capital city from 1868 to the present day. ...

Standard Generation

Where is the location of the studios for the radio station in Augusta where the American rock band Blackberry Smoke originates from?

Incremental Generation

1-hop question (using Document3): What city served as the capital of Georgia before Atlanta?

2-hop question (using Document2): What city in Georgia is the 102.3 K-Love radio station located in?

3-hop question (using Document1): What city in Georgia is the band Blackberry Smoke from?

Ground Truth

What was the capital of the state where WEKL operates, before the city where Blackberry Smoke was formed?

GPT-3.5-turbo

Conclusion

- The generation results of our model are explainable in terms of its reasoning process.
- End-to-end training, eliminating the requirement for labeling intermediate questions.
- The ability to generate complex questions aligned with target answers.
- The additional benefit of synthetic questions for training multi-hop QA models.

Email: seonjeongh@postech.ac.kr