Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting

Yi Cheng[1], Siyao Li[2], Bang Liu[3*], Ruihui Zhao[1], Sujian Li[4], Chenghua Lin[5], Yefeng Zheng[1]

ACL 2021

Motivation

- 本文探讨了难度可控问题生成（DCQG）的任务，该任务旨在生成具有所需难度水平的问题。以前对于这项任务的研究缺乏解释性和可控性。

QUESTION REWRITING

Q_1 : Who starred Top Gun?

Q_2 : Who starred the film directed by Tony Scott ? *(BRIDGE)*

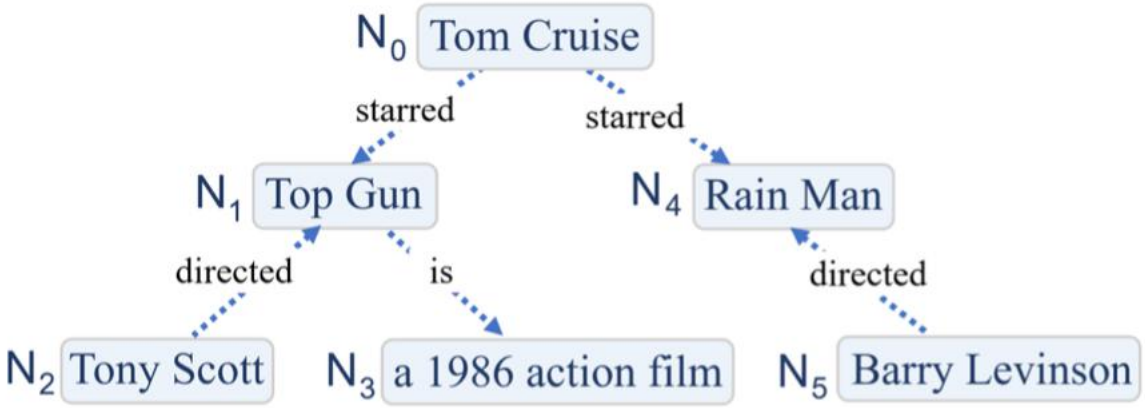
Q_3 : Who starred a 1986 action film directed by Tony Scott? *(INTERSECTION)*

Q_4 : Who starred Rain Man and a 1986 action film directed by Tony Scott? *(INTERSECTION)*

Q_5 : Who starred a film directed by Barry Levinson and a 1986 action film directed by Tony Scott? *(BRIDGE)*

将问题的难度重新定义为回答问题所需的推理步骤的数量

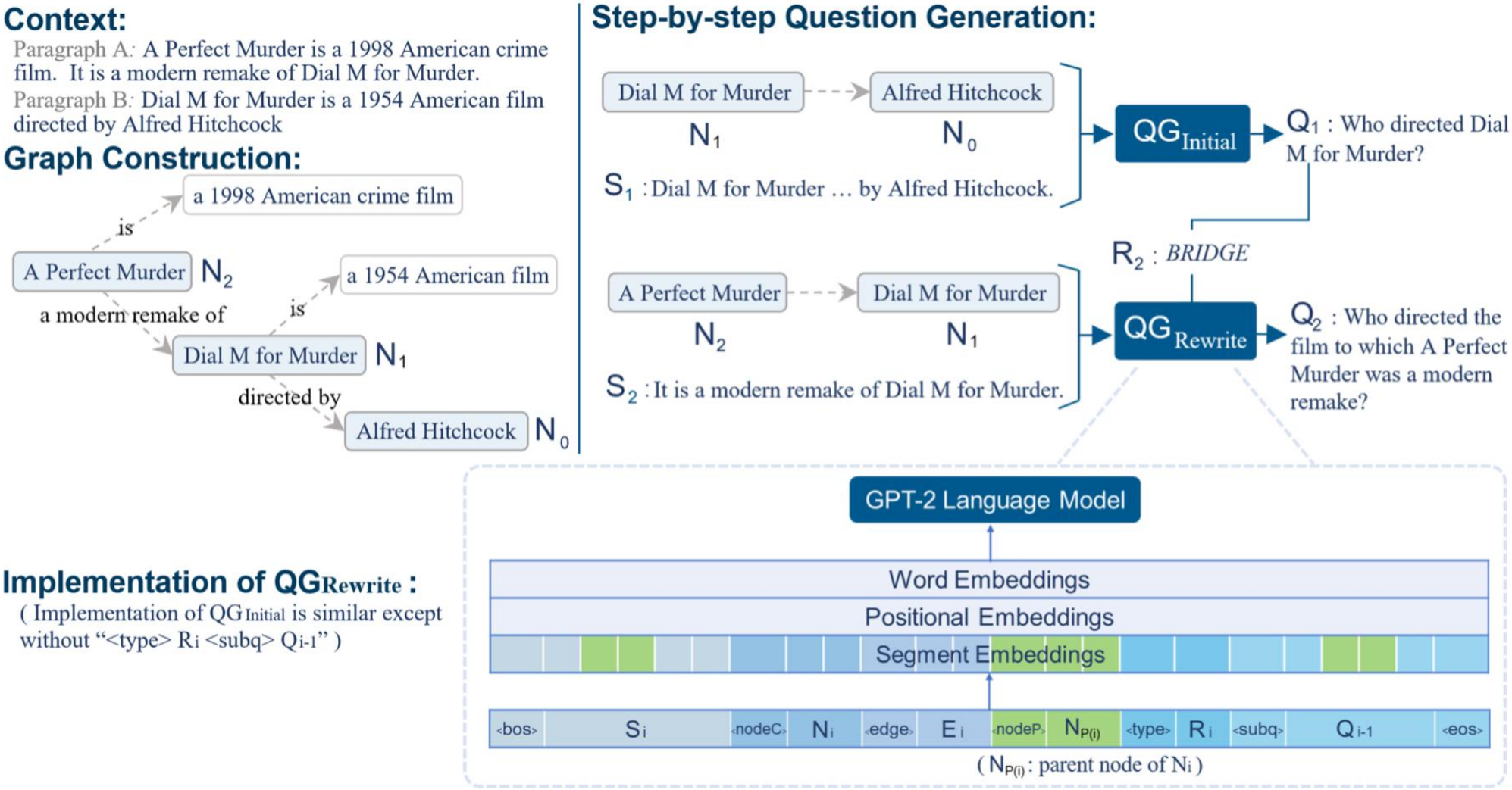
REASONING CHAIN



Contribution

- 对于难度可控问题生成任务，这是首次将问题难度定义为回答问题的推理步骤；
- 提出了一种新的框架，在提取的推理链的指导下，通过逐步重写实现难度可控的问题生成；
- 构建一个数据集，该数据集可以促进将问题改写为更复杂问题的培训，并与构建的上下文图和问题的基本推理链相匹配。

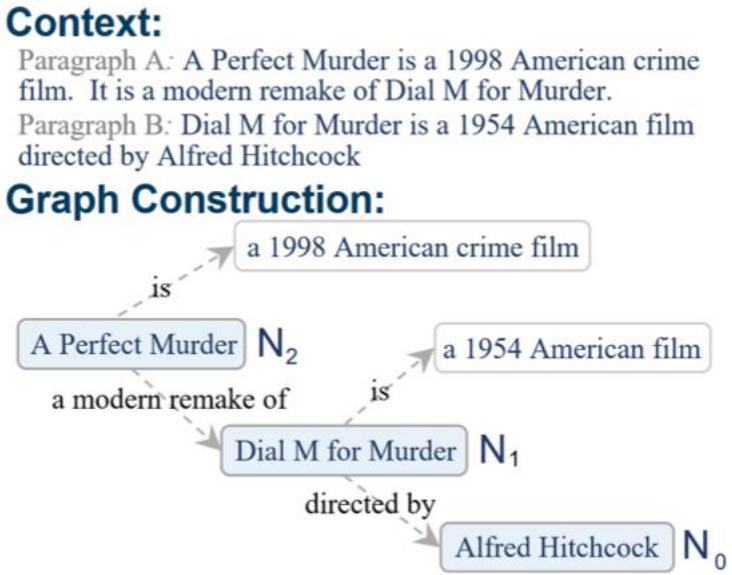
Overview


Context:
Paragraph A: A Perfect Murder is a 1998 American crime film. It is a modern remake of Dial M for Murder.
Paragraph B: Dial M for Murder is a 1954 American film directed by Alfred Hitchcock
Graph Construction:
a 1998 American crime film
is
A Perfect Murder N2
a modern remake of
a 1954 American film
is
Dial M for Murder N1
directed by
Alfred Hitchcock N0
Step-by-step Question Generation:
Dial M for Murder
Alfred Hitchcock
N1
N0
S1 : Dial M for Murder … by Alfred Hitchcock.
QGInitial
Q1 : Who directed Dial M for Murder?
R2 : BRIDGE
A Perfect Murder
Dial M for Murder
N2
N1
S2 : It is a modern remake of Dial M for Murder.
QGRewrite
Q2 : Who directed the film to which A Perfect Murder was a modern remake?
Implementation of QGRewrite :
(Implementation of QGInitial is similar except without "<type> Ri <subq> Qi-1")
GPT-2 Language Model
Word Embeddings
Positional Embeddings
Segment Embeddings
<bos> Si <nodeC> Ni <edge> Ei <nodeP> NP(i) <type> Ri <subq> Qi-1 <eos>
(NP(i) : parent node of Ni)

Method

- Context Graph Construction

首先应用开放信息提取技术从文本句中提取〈主语, 关系, 宾语〉三元组，然后将主语和宾语作为节点，它们之间的边描述关系



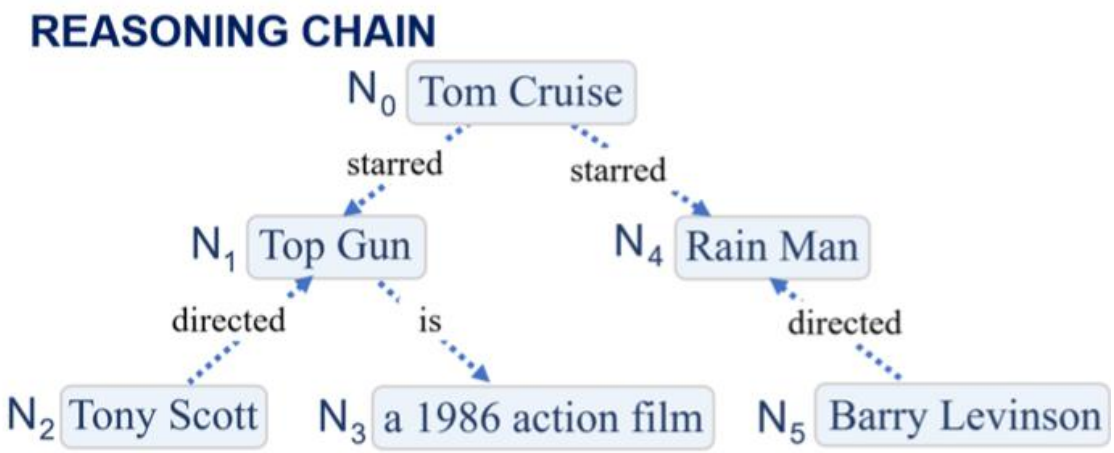
Method

- Reasoning Chain Selection

选取一个由d+1个节点组成的连通子图GT作为生成问题的推理链。

如果节点N0是具有多个节点度的命名实体或与之链接，则首先将其作为问题的答案进行采样。

从子图中提取一个最大生成树，它的根节点是N0，



Method

- Step–by–step Question Generation

Algorithm 1 Procedure of Our DCQG Framework

Input: context $\mathcal{C}$, difficulty level d
Output: $(\mathcal{Q}, \mathcal{A})$

$\mathcal{G}_{CG} \leftarrow \mathbf{BuildCG}(\mathcal{C})$
$\mathcal{N}_0 \leftarrow \mathbf{SampleAnswerNode}(\mathcal{G}_{CG})$
$\mathcal{G}_L \leftarrow \mathbf{MaxTree}(\mathcal{G}_{CG}, \mathcal{N}_0)$
$\mathcal{G}_T \leftarrow \mathbf{Prune}(\mathcal{G}_L, d)$
for $\mathcal{N}_i$ **in** $\mathbf{PreorderTraversal}(\mathcal{G}_T)$ **do**
  if $i = 0$ **then** continue
  $\mathcal{N}_{P(i)} = \mathbf{Parent}(\mathcal{N}_i)$
  $\mathcal{S}_i = \mathbf{ContextSentence}(\mathcal{C}, \mathcal{N}_i, \mathcal{N}_{P(i)})$
  $\mathcal{R}_i \leftarrow \begin{cases} \textit{Bridge} & \text{if } \mathcal{N}_i = \mathbf{FirstChild}(\mathcal{N}_{P(i)}) \\ \textit{Intersection} & \text{else} \end{cases}$
  $\mathcal{Q}_i \leftarrow \begin{cases} \mathbf{QG}_{Initial}(\mathcal{N}_i, \mathcal{N}_{P(i)}, \mathcal{S}_i) & \text{if } i = 1 \\ \mathbf{QG}_{Rewrite}(\mathcal{Q}_{i-1}, \mathcal{N}_i, \mathcal{N}_{P(i)}, \mathcal{S}_i, \mathcal{R}_i) & \text{else} \end{cases}$
end for
return $(\mathcal{Q}_d, \mathcal{N}_0)$

Q_i表示每一步生成的问题， Q_d表示最终问题， Q_{i+1}表示从Q_i中重写通过添加一跳推理

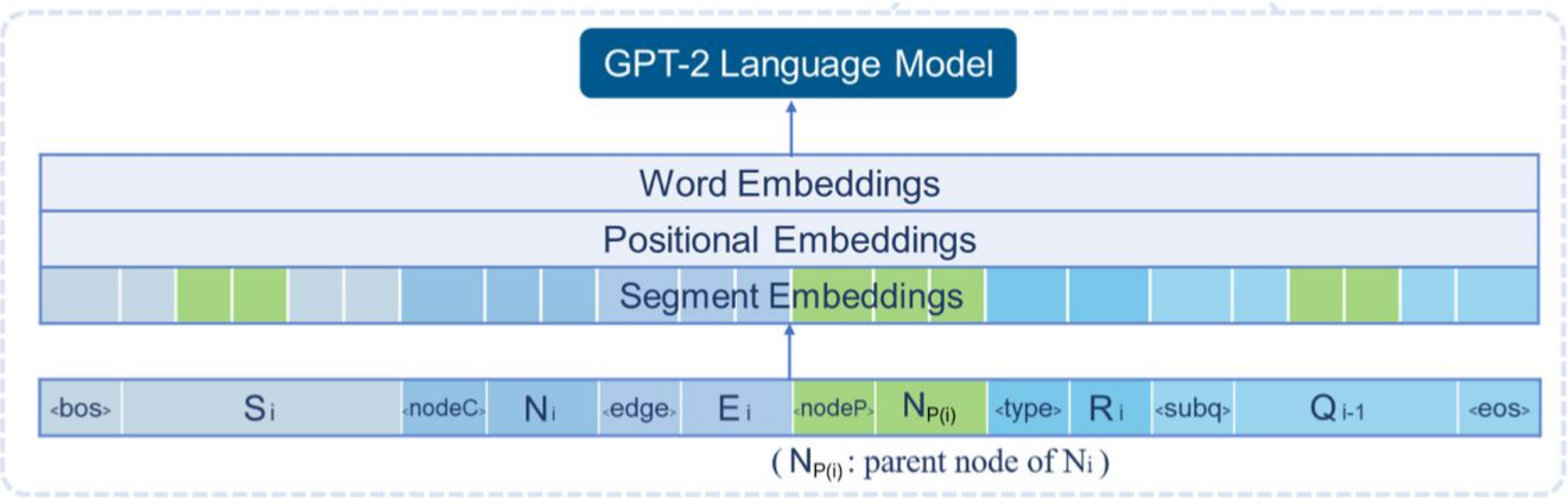
S_i代表上下文句子

R_i表示重写类型，这里分为两种：

Bridge–style：重写使用修改过的子句重新放置实体

Method

- $QG_{initial}$和$QG_{rewrite}$都是用预先训练好的GPT–2模型初始化的



Automatic Dataset Construction

- 现有的数据集HotpotQA 中包含的是多跳问题，需要将其分解

Algorithm 2 Procedure of Data Construction

Input: context $\mathcal{C} = \{\mathcal{P}_1, \mathcal{P}_2\}$, QA pair $(\mathcal{Q}_2, \mathcal{A}_2)$, supporting facts $\mathcal{F}$

Output: $\mathcal{R}_1, (\mathcal{Q}_1, \mathcal{A}_1), \mathcal{S}_1, \mathcal{S}_2, \{\mathcal{N}_0, \mathcal{E}_1, \mathcal{N}_1, \mathcal{E}_2, \mathcal{N}_2\}$

$\mathcal{R}_1 \leftarrow \mathbf{TypeClassify}(\mathcal{Q}_2)$

if $\mathcal{R}_1 \notin \{\textit{Bridge}, \textit{Intersection}\}$ **then** return

$subq_1, subq_2 \leftarrow \mathbf{DecompQ}(\mathcal{Q}_2)$

$suba_1, suba_2 \leftarrow \mathbf{QA}(subq_1), \mathbf{QA}(subq_2)$

$$\mathcal{Q}_1, \mathcal{A}_1 \leftarrow \begin{cases} subq_2, suba_2 & \text{if } \mathcal{A}_2 = suba_2 \\ subq_1, suba_1 & \text{else} \end{cases}$$

$$\mathcal{S}_1, \mathcal{S}_2 \leftarrow \begin{cases} \mathcal{F} \cap \mathcal{P}_1, \mathcal{F} \cap \mathcal{P}_2 & \text{if } \mathcal{Q}_1 \text{ concerns } \mathcal{P}_1 \\ \mathcal{F} \cap \mathcal{P}_2, \mathcal{F} \cap \mathcal{P}_1 & \text{else} \end{cases}$$

$\mathcal{N}_2 \leftarrow \mathbf{FindNode}(\mathcal{A}_2)$

$\mathcal{N}_0, \mathcal{E}_1, \mathcal{N}_1, \mathcal{E}_2 \leftarrow \mathbf{Match}(subq_1, subq_2)$

Experiment

- Evaluation of Question Quality

  BLEU3, BLEU4，METEOR， CIDEr以n-gram为单位衡量生成结果与参考问题之间的相似性。

- Human Evaluation

  - 格式正确：检查问题是否完全正确。注释者被要求将问题标记为“是”、“可接受”或“否”。如果问题在语法上不正确，但其含义仍然可以推断，则选择“可接受”。
  - 简明：它检查QG模型是否过度拟合，产生带有冗余修饰符的问题。如果没有一个单词可以删除，则问题标记为“是”；如果问题有点长，但仍然是自然的，则可以接受；如果问题异常冗长，则标记为“否”。
  - 可回答：根据给定的上下文检查问题是否可回答。注释为是或否。
  - 答案匹配：检查给定答案是否为问题的正确答案。注释为是或否。

Experiment

Model	BLEU3	BLEU4	METEOR	CIDEr
NQG++	15.41	11.50	16.96	-
ASs2s	15.21	11.29	16.78	-
SRL-Graph	19.66	15.03	19.73	-
DP-Graph	19.87	15.23	20.10	1.40
GPT2	20.98	**15.59**	**24.19**	1.46
$\text{Ours}_{\text{2-hop}}$	**21.07**	15.26	19.99	**1.48**

Table 1: Automatic evaluation results of the baseline models and the 2-hop questions generated by our method ($\text{Ours}_{\text{2-hop}}$).

$\text{Ours}_{\text{2-hop}}$	GPT2
When was the first theatre director of African descent born?	When was the first theatre director of African descent to establish a national touring company in the UK born?
What play by Carrie Hamilton was run at the Goodman Theatre in 2002?	What play by Carrie Hamilton and Carol Burnett ran at the Goodman Theatre and on Broadway in 2002?
What was the review score for the album that has been reissued twice?	What was the review of the album that includes previously unreleased tracks by Guetta from its first major international release?

Experiment

Difficulty Level	Model	Well-formed			Concise			Answerable		Answer Matching	
		Yes	Acceptable	No	Yes	Acceptable	No	Yes	No	Yes	No
2-hop	DP-Graph	28%	41%	31%	41%	53%	6%	49%	51%	39%	61%
	GPT2	57%	34%	9%	47%	50%	3%	69%	31%	66%	34%
	$\text{Ours}_{\text{2-hop}}$	**74%**	19%	7%	**67%**	30%	3%	**78%**	22%	**69%**	31%
	$\text{Gold}_{\text{2-hop}}$	72%	22%	6%	56%	40%	4%	92%	8%	87%	13%
1-hop	$\text{Ours}_{\text{1-hop}}$	**46%**	46%	8%	**65%**	25%	10%	**81%**	19%	**72%**	28%
	$\text{Gold}_{\text{1-hop}}$	56%	39%	5%	80%	16%	4%	84%	16%	79%	21%

Experiment

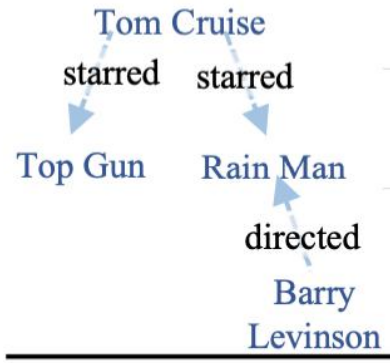
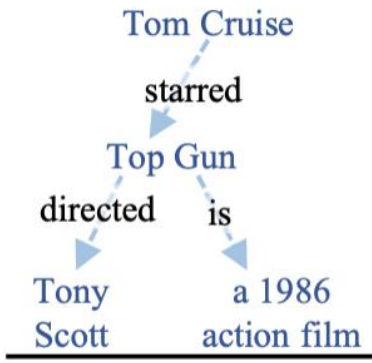
Model	Inference Steps			
	1-hop	2-hop	3-hop	>3-hop
DP-Graph	26.1%	55.1%	8.7%	10.1%
GPT2	23.3%	57.1%	13.2%	6.4%
$\text{Ours}_{\text{2-hop}}$	4.3%	**67.7%**	25.8%	2.2%
$\text{Ours}_{\text{1-hop}}$	**70.7%**	28.2%	1.1%	0.0%

Table 4: Human evaluation results of the number of inference steps required by the generated questions.

Test Set	BERT		RoBERTa	
	EM	F1	EM	F1
DP-Graph	0.436	0.615	0.552	0.678
GPT2	0.419	0.581	0.669	0.772
$\text{Ours}_{\text{2-hop}}$	0.295	0.381	0.506	0.663
$\text{Ours}_{\text{1-hop}}$	0.618	0.737	0.882	0.937

Table 5: Performance of BERT- and RoBERTa-based QA models on different generated QA datasets.

More-hop Question Generation

Reasoning Chain	QG Process	
	Q_1:	Which actor who starred in Top Gun?
	Q_2:	Which star of Top Gun was also in the movie Rain Man?
	Q_3:	Which star of Top Gun was also in the movie directed by Barry Levinson?
	Q_1:	Which actor who starred in Top Gun?
	Q_2:	What actor starred in the film that was directed by Tony Scott?
	Q_3:	What actor starred in the film that was directed by Tony Scott and was released in 1986?

我们可以看到，作为跳板的中间问题被QGRewrite有效地用于生成更复杂的问题。通过只包含1-hop和2-hop问题的训练数据，我们的框架能够生成一些高质量的3-hop问题，证明了我们框架的可扩展性。但是，当生成超过3跳的问题时，我们发现问题质量会急剧下降。
这可能是因为QGRWRITE的输入变得太长，以至于无法通过GPT2小模型进行精确编码，因为问题的长度越来越长。我们未来的工作将是探索如何有效地将我们的方法扩展到更高跳数的问题生成。

Thanks！