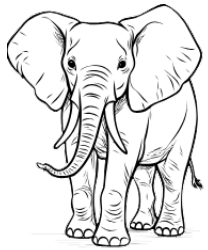


ACL 2024

Bangkok, Thailand

Towards Better Question Generation in QA-based Event Extraction

Zijin Hong, Jian Liu

Jinan University, Beijing Jiaotong University

暨南大學 JINAN UNIVERSITY

北京交通大學 BEIJING JIAOTONG UNIVERSITY

Introduction, Challenges, and Motivations

Context:
Marines$_{attacker}$ were involved in a firefight in the center of Baghdad. They used rocket-propelled grenades and semi-automatic weapons, causing some injuries.

Template Q1: Who is the attacker in firefight?
A1. None ✗

Template Q2: Who was the attacking agent?
A2. weapons ✗

Human-Written Q3: Who was involved in the firefight in the center of Baghdad using grenades and weapons?
A3. Marines ✓

➢ Introduction

- ❑ Given an event context and a target role, the method first generates a question to extract the corresponding arguments from the context, then uses a QA model to answer the question, thus performing the event extraction (EE) task.

➢ Challenges

- ❑ The quality of the generated question significantly affects the accuracy of QA-based EE.
- ❑ Previous methods rely on well-designed templates, which are rigid and less context-dependent.

➢ Motivations

- ❑ To improve question quality for EE, it is essential to define clear criteria for a "good" question.
- ❑ Designing a framework that follows these criteria can enhance question generation, leading to more accurate QA-based EE with high-quality questions.

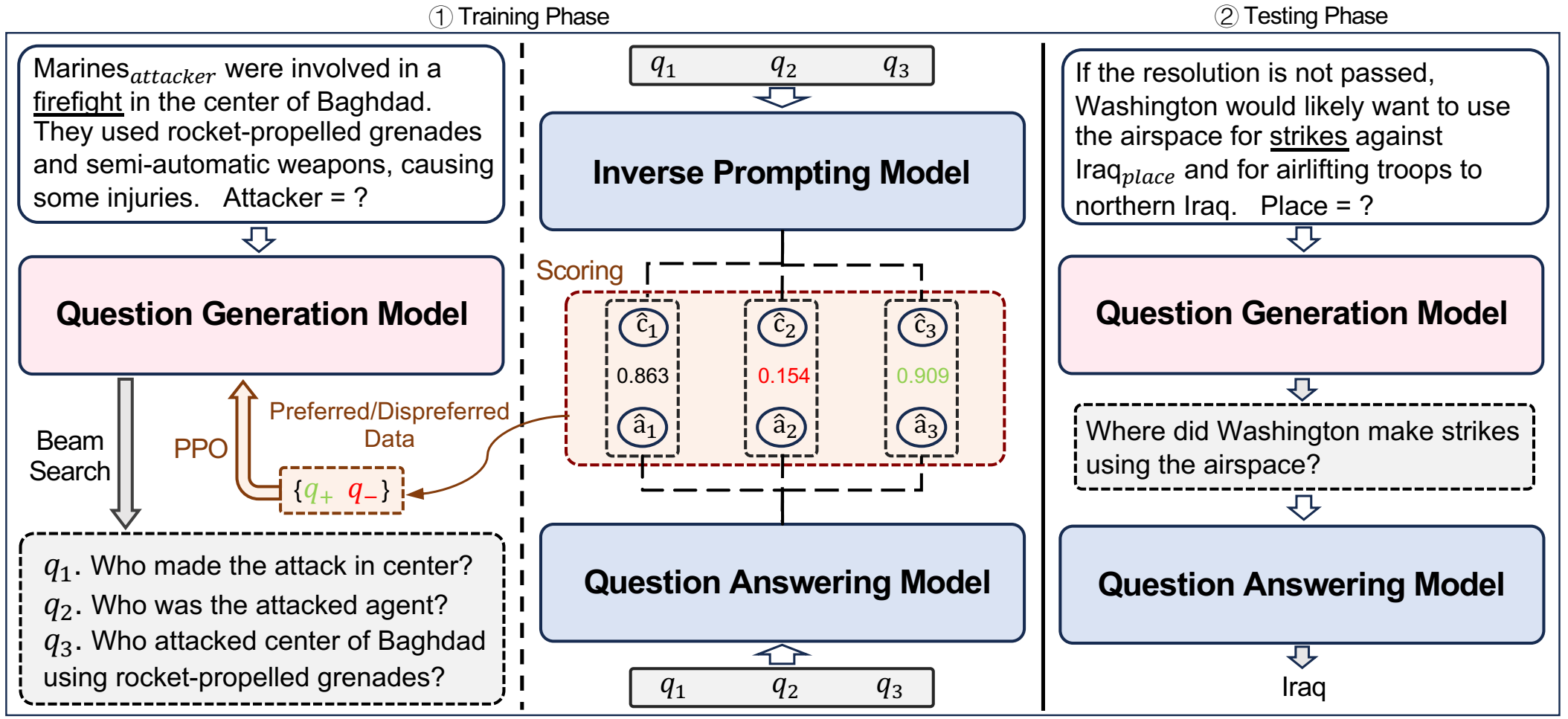
Our Contribution

- We revisit question generation for QA-based EE and suggest four question evaluation criteria **(Fluency, Generalizability, Context-dependence, Indicative guidance)**. We design a model that can generate better questions with these as guidance.

- We introduce a reinforcement learning framework for better question generation for EE, which is considered context-dependent and indicative of question generation.

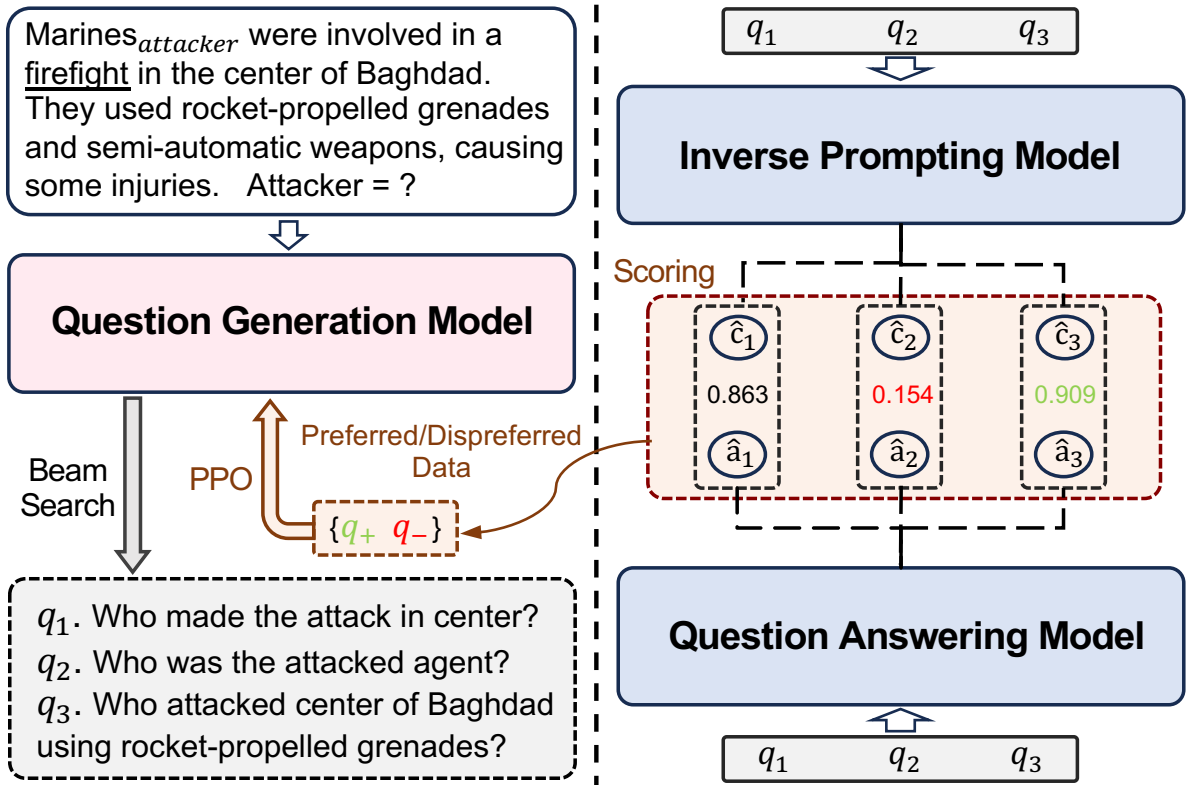
- We have verified the effectiveness of our method on different benchmarks, and show its capability to handle the more challenging data-scarce scenario.

Proposed Method



- Training phase, including supervised fine-tuning and reinforcement learning refining for a question-generation model
- Testing phase, a final (off-the-shelf) question-answering model predicts the final answer based on the given context and question.

Training Details



Context:
Warplanes$_{instrument}$ pounded forward Iraqi positions in the hills overlooking Chamchamal, 35 kilometers ...

Generated Question:
What *instrument* was used in the attack in Iraqi positions?

Recovered Context:
An *instrument* was used to pound the Iraqi positions during the attack.

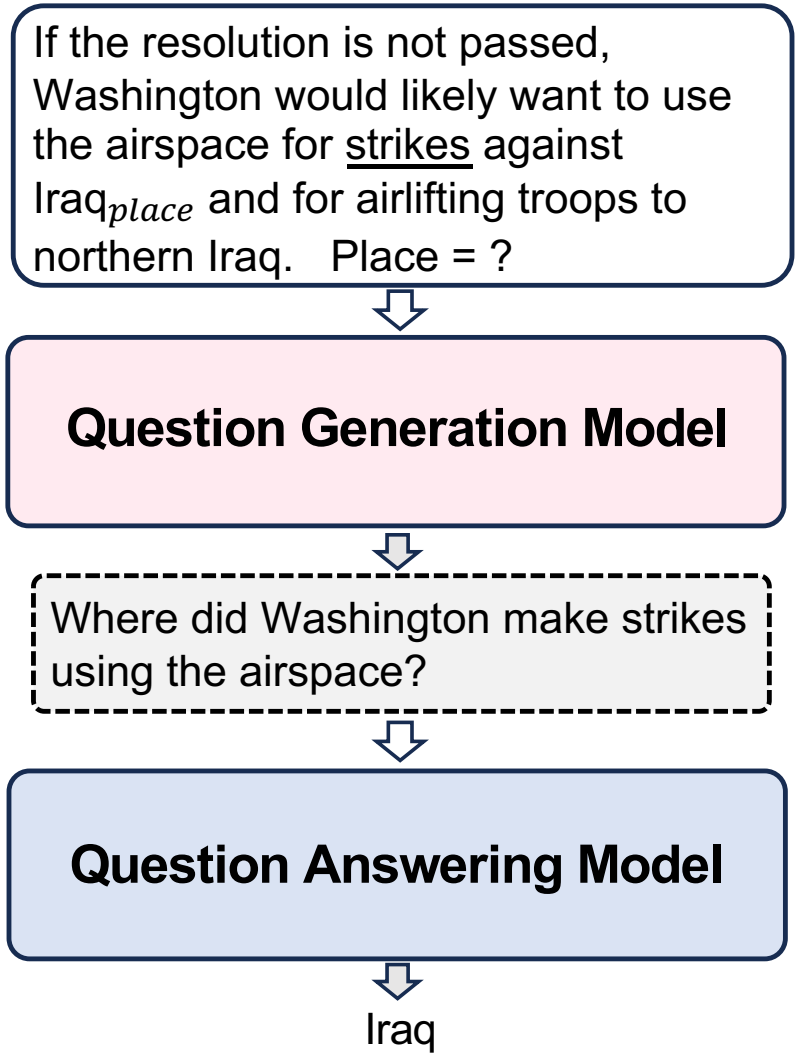
Supervised Fine-tuning

- The backbone model is SFT over the template questions to ensure generalizability and fluency
- Beam search augmentation is utilized to generate question candidates for further refinement

Reinforcement Learning

- Each question in the candidate set is input into the inverse prompting and the QA model
- PPO training is conducted based on reward modeling of the recovered context from inverse prompting and the answer from QA

Better Questions Generation in QA-based EE



- The event context, trigger, and target role are input into the RL-refined question generation model, which generates a better question with context-dependence and clear guidance.
- This improved question will guide the QA model in extracting accurate event arguments.

Main Experiments

Methods	Practical Eval.			Full Eval.		
	EM	COR	SemSim	EM	COR	SemSim
Template						
Simple-Q (Liu et al., 2020)	35.41	40.23	60.93	14.38	16.17	24.55
Standard-Q (Du and Cardie, 2020)	37.42	43.87	63.70	15.60	17.36	25.92
Back-Trans-Q	36.13	41.39	62.41	15.14	16.67	25.28
Guideline-Q (Du and Cardie, 2020)	38.51	45.28	65.54	17.61	19.96	28.14
Dynamic-Q (Lu et al., 2023)	38.70	45.79	65.55	20.45	23.12	30.79
Supervised Fine-tuning						
SFT (*Standard*)	37.63	42.95	62.36	15.31	17.13	25.72
SFT (*Back-Trans*)	38.24	43.56	64.11	17.47	18.90	27.32
SFT (*Guideline*)	38.62	44.69	64.66	17.33	19.61	27.77
SFT (*Dynamic*)	39.31	46.78	66.24	20.35	23.05	30.53
In-context learning						
LLaMA-2-13b-Chat (*0shot*)	1.21	3.50	35.88	0.43	1.25	21.78
LLaMA-2-13b-Chat (*5shot*)	27.97	33.04	53.69	13.01	14.93	23.54
GPT-4 (*0shot*)	28.97	35.83	57.90	11.14	13.54	23.35
GPT-4 (*5shot*)	39.24	47.59	65.92	16.32	19.37	27.46
RLQG (Ours)	**41.39**	**48.58**	**67.94**	**21.71**	**24.19**	**31.80**

Table 2: Event extraction results with Practical Evaluation and Full Evaluation on the ACE test dataset, where EM, COR, and SemSim indicate exact match accuracy, context overlap ratio, and semantic similarity, respectively.

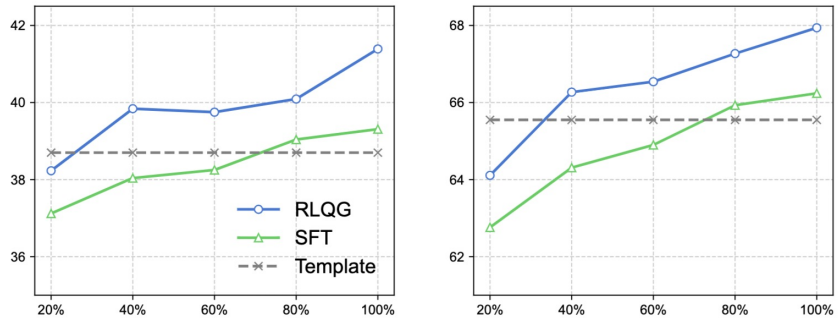
Methods	EM	COR	SemSim
Template			
Standard-Q	17.65	23.02	47.96
Back-Trans-Q	16.45	21.47	46.43
Supervised Fine-tuning			
SFT (*Standard*)	18.10	23.84	48.79
SFT (*Back-Trans*)	18.29	24.11	49.32
RLQG (Ours)	**19.61**	**25.43**	**50.69**

Table 3: Event extraction results on the RAMS test dataset with practical evaluation.

- Our proposed method outperforms all baseline methods in every evaluation metric on both evaluation settings, including the proprietary model GPT-4.
- The RLQG further refines the questions generated by the SFT model, making them more context-dependent and indicative.

Further Analysis

Performance on Data Scarcity



(a) Exact Match Accuracy (b) Semantic Similarity

Figure 3: Experimental results in the ACE dataset for the data-scarce scenario. The x-axis represents different ratios of training data, y-axis is the value of the metric.

Ablation Study

Methods	ACE	RAMS
Standard-Q	37.42	17.65
SFT (*Standard*)	37.63	18.10
RLQG (*Standard*)	**39.29**	**19.23**

Table 4: Ablation study of the different template as the starting point on exact match accuracy

Method	EM	COR	SemSim
RLQG	**41.39**	**48.58**	**67.94**
-w/o *IP reward*	40.21	47.49	66.96
-w/o *QA reward*	39.86	47.17	66.88

Table 5: Results of ablation studies for removing different reward.

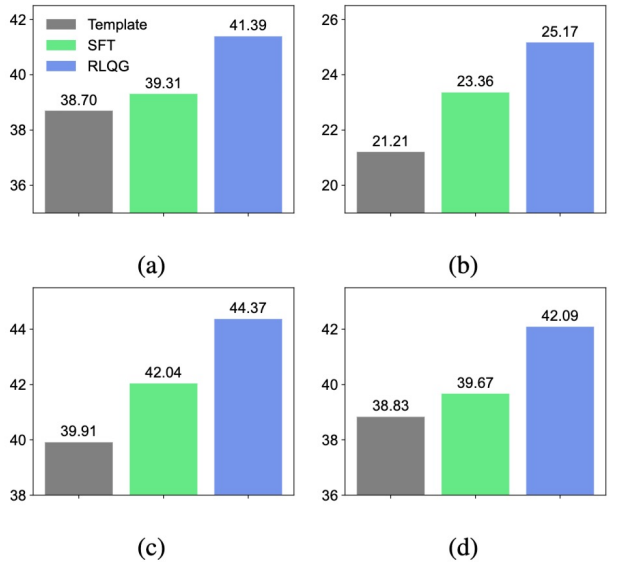


Figure 4: The performance with exact match accuracy on different QA model (a) LLaMA-2-13b-Chat (*5shot*)

- In data-scarce scenarios, RLQG is particularly effective for QA-based EE compared to the template and SFT methods.
- RLQG maintains its effectiveness with different template starting points and various QA models.
- Every component is essential for RLQG's success.

Case Study

Context:
Former senior banker Callum McCarthy$_{person}$ **begins** what is one of the most important jobs in London's financial world in September, when incumbent Howard Davies steps down.

Template Question:
Who is the employee?
Answer: Davies ✗

SFT Question:
Who was hired by banker?
Answer: Howard Davies ✗

RLQG Question:
Who was hired as one of the most important jobs?
Answer: Callum McCarthy ✓

Human-Written Question:
Who was the former senior banker that began an important job in September?
Answer: Callum McCarthy ✓

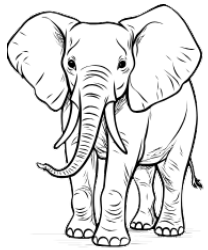
- The questions generated by the template and SFT models are rigid and misleading, lacking event context.
- The context-dependent RLQG question correctly trigger the answer, which is the most similar to the human-written question.

Conclusion

- **Advancement in EE Methods:** Event extraction (EE) has evolved from traditional classification to QA-based methods, emphasizing the design of quality questions to guide QA models for better answers.

- **Reinforcement Learning Framework:** Our proposed reinforcement learning framework generates context-dependent, fluently phrased questions that are generalizable and indicative, addressing the limitations of rigid, template-based questions.

- **Improved Performance:** Our methodology demonstrates improved performance on ACE and RAMS benchmarks, especially in data-scarce scenarios, highlighting its efficacy in generating effective questions for QA-based EE without extensive manual intervention.



ACL 2024

Bangkok, Thailand

Thanks for your listening!

暨南大學

JINAN UNIVERSITY

北京交通大學

BEIJING JIAOTONG UNIVERSITY