# QA4QG: USING QUESTION ANSWERING TO CONSTRAIN MULTI-HOP QUESTION GENERATION

ICASSP 2022

Dan SU

Peng XU

Pascale Fung

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CAiRE
Center for Artificial Intelligence Research

# Outline

- Introduction
- Related Work
- Methodology
- Experiments
- Conclusion

# Multi-hop Question Generation (QG)

- Question Generation (QG) is a task to automatically generate a question from a given context and, optionally, an answer.
- Multi-hop QG requires aggregating scattered evidence spans from multiple paragraphs, and reasoning over them.
  - Given the answer is Location H, to ask where is T located, the model needs a bridging evidence to know that T is located in C, and C is located in H (T -> C ->H). This is done by multi-hop reasoning.

**Paragraph A:** Marine Tactical Air Command Squadron 28 (*Location T*) is a United States Marine Corps aviation command and control unit based at Marine Corps Air Station Cherry Point (*Location C*) ...

**Paragraph B:** Marine Corps Air Station Cherry Point (*Location C*) ... is a United States Marine Corps airfield located in Havelock, North Carolina (*Location H*), USA ...

**Answer:** Havelock, North Carolina (*Location H*)

**Question:** What city is the Marine Air Control Group 28 (*Location T*) located in?

Fig.1 An example of multi-hop QG in the HotpotQA (Yang et al., 2018) dataset. Figure courtesy: [1]

[1] Dan Su et al, *"Multi-hop Question Generation with Graph Convolutional Network,"* in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

# Related Work on Multi-hop QG

- Extend the existing Seq2Seq framework for single-hop QG with reasoning ability via:
  - Models text as graph structure and incorporates graph neural networks into the traditional Seq2Seq framework [1,5,6].
  - Augment the Seq2Seq framework with extra constraints to guide the generation [7,8,9].

[5] Pan et. al., “Semantic graphs for generating deep questions,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, pp. 1463–1475.
[6] Yu et. al., “Generating multi-hop reasoning questions to improve machine reading comprehension,” in Proceedings of The Web Conference 2020, New York, NY, USA, 2020, WWW ’20, p. 281–291, Association for Computing Machinery.
[7] Gupta et. al., “Reinforced multitask approach for multi-hop question generation,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2760–2775

# Related Work for Multi-hop QG

- The most recent work has shown strong capability of simple architecture design with large pre-trained language models for multi-hop QA [10, 3].
  - Such approaches have outperformed the graph network based methods and achieved comparable performance with state-of-the-art architectures.

[8] Wang et. al., “Answer-driven deep question generation based on reinforcement learning,” in Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), Dec. 2020, pp.5159–5170,
[9] Xie et. al., “Exploring question-specific rewards for generating deep questions,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2534–2546.
[10] Shao et. al, “Is graph structure necessary for multihop question answering?,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7187–7192.

# Issues on Multi-hop QG

- **Framework-wise**: Incorporating graph structure may not be necessary, and can be replaced with **Transformers** or proper use of large pre-trained models for multi-hop QA [2,3].
- **Training objective-wise:** Aim to model ***P(Question|<Context, Answer>)***, but ignored the strong constraint of ***P(Answer|<Question, Context>)***. **QA and QG are dual tasks that can help each other** [4].

[2]Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu, *"Is graph structure necessary for multihop question answering?,"* in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7187–7192.
[3] Groeneveld, Dirk, Tushar Khot, and Ashish Sabharwal. "*A Simple Yet Strong Pipeline for HotpotQA*." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.
[4] Tang, Duyu, et al. "Question answering and question generation as dual tasks." *arXiv preprint arXiv:1706.02027* (2017).

# Methodologies

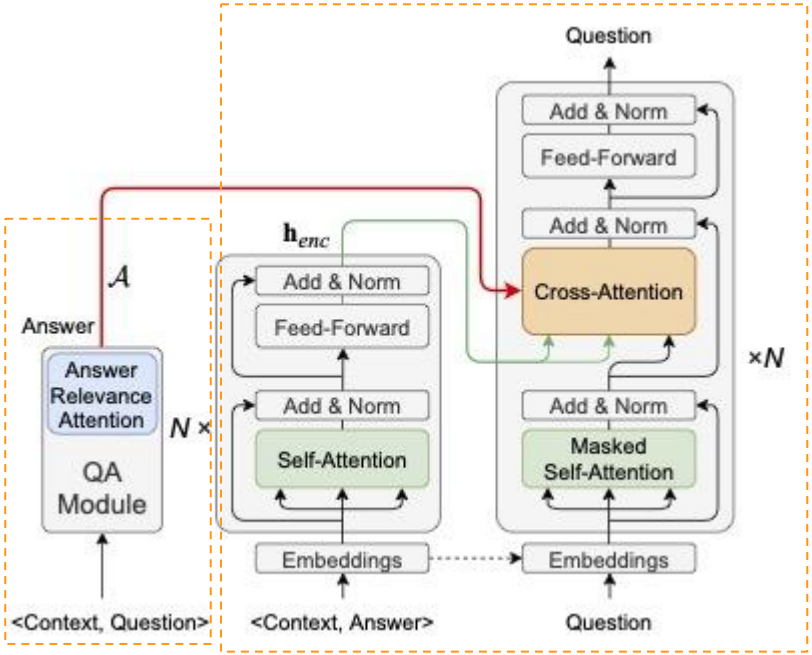


**Fig.2** The architecture of our QA4QG. The output of the QA module is used to bias the cross-attention of Transformer decoder.

- QA Module
  - **Input**: <Context, Question>
  - **Outputs**: the probability of each token being the answer $A$
- Transformer-based Seq2Seq Model
  - **Input**: <Context, Answer>, and $A$
  - **Output**: Question

# Methodologies

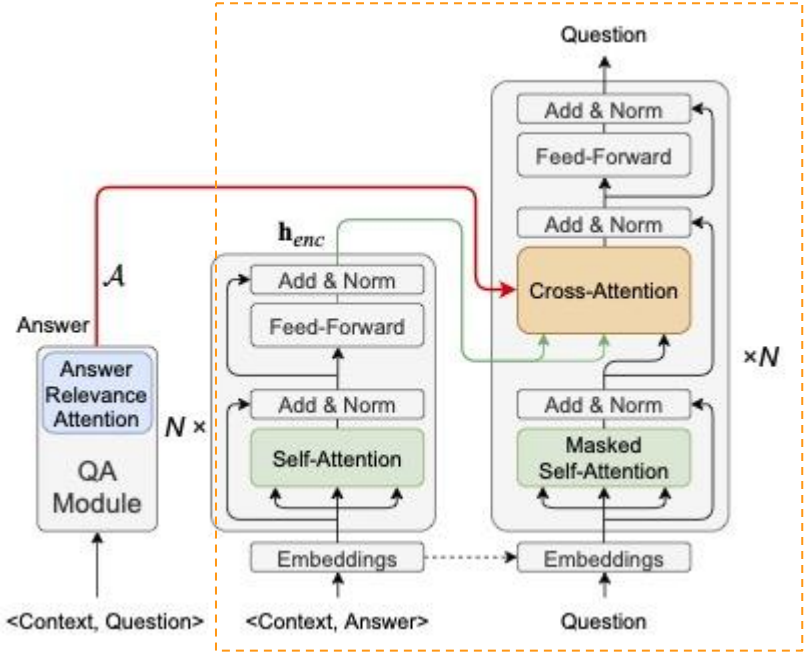


**Fig.2** The architecture of our QA4QG. The output of the QA module is used to bias the cross-attention of Transformer decoder.

- BART as the Transformer-based Seq2Seq model.

Encoder:

$$h_{enc} = Encoder(< C, Q >)$$

Decoder:

$$H_i^a = \text{MultiHeadAttention}(H_i, H_i, H_i) \quad (1)$$
$$H_i^b = \text{Norm}(H_i + H_i^a) \quad (2)$$
$$H_i^c = \text{MultiHeadAttention}(H_i^b, h_{enc}, h_{enc}), \quad (3)$$

# Methodologies

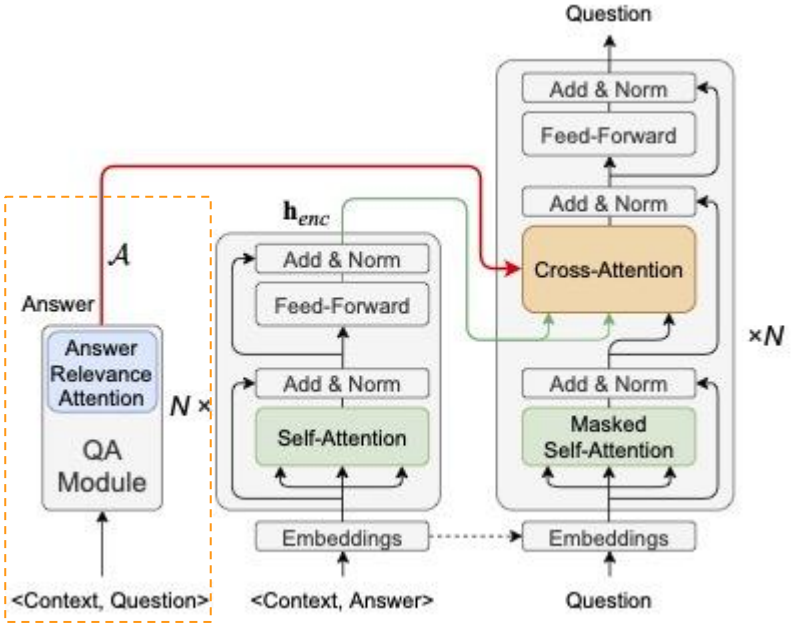


**Fig.2** The architecture of our QA4QG. The output of the QA module is used to bias the cross-attention of Transformer decoder.

**Answer Relevance Attention $A$,** to indicate the answer relevance of each token in context to the target question.

- ***Soft Attention***, when the ground truth question is available (e.g., in the training phase) $A_{soft} = P^s_{ans} + P^e_{ans}$

  $P^s_{ans} / P^e_{ans}$ is the probability that the i-th token is the start/end of the answer span in context C

- ***Hard Attention***, when no question is available (e.g., in the testing phase).
  - $A_{hard}$ is a binary-valued distribution, to indicate the binary relevance of each token in the context to the answer (in our work, py = 1.0, pn = 0.0).

# Methodologies

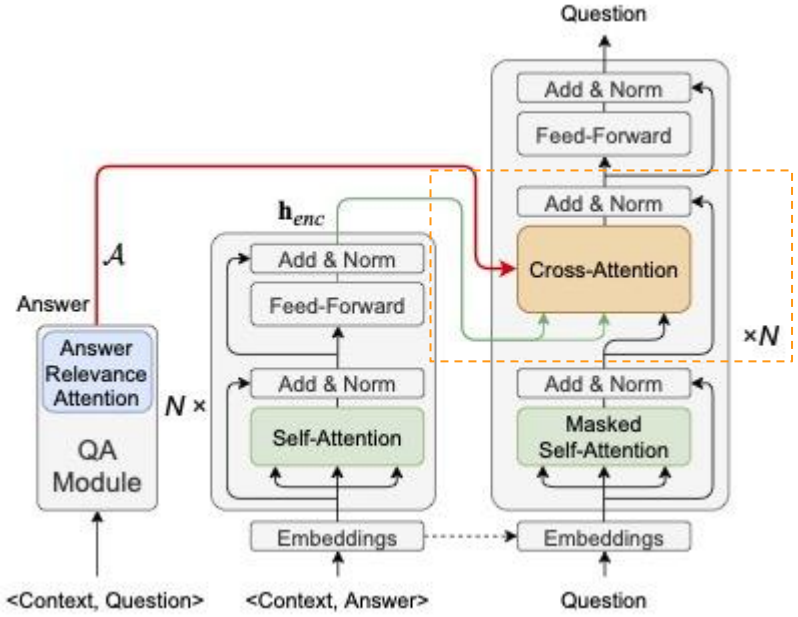


**Fig.2** The architecture of our QA4QG. The output of the QA module is used to bias the cross-attention of Transformer decoder.

## Enhanced Cross-Attention

To bias the original cross-attention sub-layer (i.e., Eq. 3) in each BART decoder layer with A:

$$H_i^{c'} = \text{softmax}\left(\frac{H_i^b h_{enc}^T}{\sqrt{d_k}} + \mathcal{A}\right) h_{enc}$$

$$\mathcal{A} = \alpha \mathcal{A}_{hard} + (1-\alpha)\mathcal{A}_{soft},$$

# Experiments

Datasets & Baselines

Results and Analysis

# Datasets

## HotpotQA-datasets

- HotpotQA is a multihop QA dataset, which contains Wikipedia-based question-answer pairs, with each question requiring multi-hop reasoning across multiple paragraphs to infer the answer.
- 90,440 training examples and 6,072 test examples
- Each question is paired with two long documents.

# Baselines

**ASs2s-a:**

**SemQG** :

**F + R + A:**

**SGGDQ:**

**ADDQG:**

**MultiQG:**

**GATENLL+CT:**

**LowResouceQG:**

# Main Results

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L |
|---|---|---|---|---|---|---|
| *Encoder Input: Supporting Facts Sentences* | | | | | | |
| ASs2s-a [14] | 37.67 | 23.79 | 17.21 | 12.59 | 17.45 | 33.21 |
| SemQG [15] | 39.92 | 26.73 | 18.73 | 14.71 | 19.29 | 35.63 |
| F + R + A [12] | 37.97 | - | - | 15.41 | 19.61 | 35.12 |
| SGGDQ (DP) [4] | 40.55 | 27.21 | 20.13 | 15.53 | 20.15 | 36.94 |
| ADDQG [7] | 44.34 | 31.32 | 22.68 | 17.54 | 20.56 | 38.09 |
| QA4QG (*LARGE setting*) | **49.55** | **37.91** | **30.79** | **25.70** | **27.44** | **46.48** |
| *Encoder Input: Full Document Context* | | | | | | |
| MultiQG [5] | 40.15 | 26.71 | 19.73 | 15.2 | 20.51 | 35.3 |
| GATENLL+CT [9] | - | - | - | 20.02(14.5) | 22.40 | 39.49 |
| LowResouceQG [16] | - | - | - | 19.07 | 19.16 | 39.41 |
| QA4QG (*BASE setting*) | 43.72 | 31.54 | 24.47 | 19.68 | 24.55 | 40.44 |
| QA4QG (*LARGE setting*) | **46.45** | **33.83** | **26.35** | **21.21** | **25.53** | **42.44** |

**Table 1.** Comparison between QA4QG and previous MQG methods on the HotpotQA dataset in different encoder input settings. **QA4QG outperforms the best models up to 8 BLEU-4 and 8 ROUGE points.**

# Ablations-1

The effect of QA-module:

- When we remove the QA module, the performance drops in both the large and base settings.
- QA module did not affect the performance in the supporting sentences setting as in the full documents setting.

| Models | BLEU-4 | METEOR | ROUGE-L |
|---|---|---|---|
| QA4QG-large | 21.21 | 25.53 | 42.44 |
| *w/o* QA | 19.32 | 24.65 | 40.74 |
| QA4QG-base | 19.68 | 24.55 | 40.44 |
| *w/o* QA | 17.43 | 23.16 | 38.23 |
| QA4QG-large (sp) | 25.70 | 27.44 | 46.47 |
| *w/o* QA | 25.69 | 27.20 | 46.30 |

**Table 2.** Ablation study on the QA module. The bottom section uses the supporting sentences (sp) as input.

# Ablations-2

The effect of the hyper-parameter:

- In general, the more $A_{soft}$, the greater performance improvement the model can achieve.
- The mixture of both when $\alpha$= 0.3 yields best results, possibly because of the disparity between training and testing, since during testing we only have $A_{hard}$ .

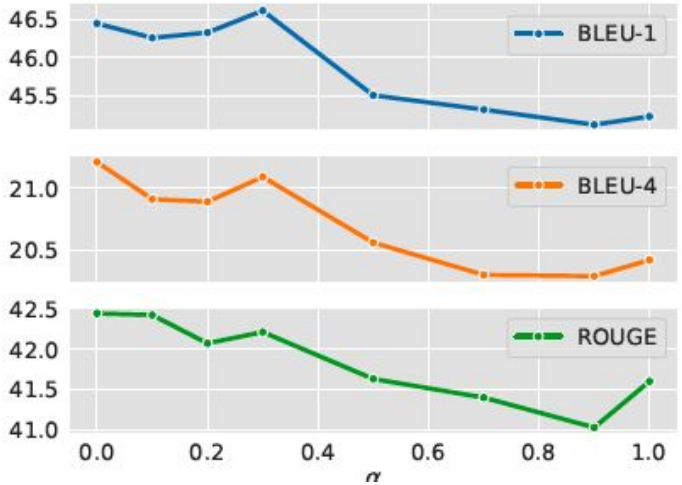


**Table 3.** Ablation study on impact of $\alpha$, with different combinations of the soft attention and hard attention.

# Visualization Example

- As we see, the $A_{soft}$ emphasis on the sentence that contains the multi-hop information ’First for Women’ and ’Jane’ in the context, which then constrains the generation model.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| First | for | Women | is | a | woman's | magazine | published |
| by | Bauer | Media | Group | in | the | USA. | The |
| magazine | was | started | in | 1989. | It | is | based |
| in | Englewood | Cliffs, | New | Jersey. | In | 2011 | the |
| circulation | of | the | magazine | was | 1310696 | copies. | Jane |
| was | an | American | magazine | created | to | appeal | to |
| the | women | who | grew | up | reading | Sassy | Magazine; |
| Jane | Pratt | was | the | founding | editor | of | each. |
| Its | original | target | audience | (pitched | to | advertisers) | was |
| aged | 18–34, | and | was | designed | to | appeal | to |
| women | who | did | not | like | the | typical | women's |
| magazine | format. | Pratt | originally | intended | the | magazine | to |
| be | named | Betty, | but | she | was | voted | down |
| by | everyone | else | involved | in | the | making | of |
| the | magazine. | | | | | | |

**Table 3.** Visualization of soft attention Asoft. Darker color represents higher attention weights. For an answer ’yes’, our Asoft emphasizes the multi-hop information related to ’First for Women’ and ’Jane’ in the context, which then constrains the generation model. The target question is ’Are Jane and
First for Women both women’s magazines?’.

# Conclusion

- Proposed a novel framework, QA4QG, a QA augmented, BART-based framework for MQG.
  - It is the first work to explore large pre-trained language models for MQG.
  - It takes advantage of an additional Multi-hop QA module to further constrain the question generation.
- QA4QG outperforms all state-of-the-art models, with an increase of 8 BLEU-4 and 8 ROUGE points compared to the best results previously reported.
- Our work suggests the advantage of introducing pre-trained language models and QA modules for the MQG task.

# References

[5] Pan et. al., “Semantic graphs for generating deep questions,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, pp. 1463–1475.
[6] Yu et. al., “Generating multi-hop reasoning questions to improve machine reading comprehension,” in Proceedings of The Web Conference 2020, New York, NY, USA, 2020, WWW ’20, p. 281–291, Association for Computing Machinery.
[7] Gupta et. al., “Reinforced multitask approach for multi-hop question generation,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2760–2775
[8] Wang et. al., “Answer-driven deep question generation based on reinforcement learning,” in Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), Dec. 2020, pp.5159–5170,
[9] Xie et. al., “Exploring question-specific rewards for generating deep questions,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2534–2546.
[10] Shao et. al, “Is graph structure necessary for multihop question answering?,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7187–7192.

# Thank you

Any questions are welcome!