Answer-Agnostic Question Generation for Educational Questions

Shreyas Chandrashekaran[1], Varun Goyal[2], Rachel Himmel[3], Aakash Patel[1], Narayanan Venugopal[3]

[1]Department of Electrical Engineering and Computer Science, [2]Department of Mechanical Engineering, [3]Department of Statistics
University of Michigan, Ann Arbor, U.S.A.

Motivation

Automatic question generation (QG) is a popular language inference task. However, most state-of-the-art QG models are “answer-aware”, requiring an answer to be directly present in the context sentence and provided in the training data. This is a barrier to generating more sophisticated, open-ended questions, that reflect a deeper understanding of the source material, such as those often asked in school exams. To address this problem, we investigate the use of answer-agnostic question generation methods, including the use of masked language models (MLM), graph-based models (Graph2Seq), and ChatGPT, with a focus on educational questions.

LearningQ TED-Ed Dataset

The LearningQ dataset[3] consists of instructor-designed questions collected from TED-Ed talks along with learner-generated questions and audio transcripts collected from Khan Academy videos.

TED-Ed Data v/s Khan Academy Data

- **Khan Academy** questions were viewer commented, not curated, often grammatically incorrect, and sometimes not even relevant to the content of the video.
- With the intent of generating higher-level questions, often without predetermined answers in mind, the **TED-Ed** data is more suitable as the questions are less contrived.
- Data is split into **train (80%)**, **validation (10%)**, and **test (10%)** sets.

We aim to produce answer-agnostic question generation models, intended to enable QG that involves a greater synthesis of different aspects of the material and requires deeper thought.

TED-Ed Example Contexts

Even though we often say that something tastes spicy, it's not actually a taste, like sweet or salty or sour. Instead, what's really happening is that certain compounds in spicy foods activate the type of sensory neurons called polymodal nociceptors. You have these all over your body, including your mouth and nose, and they're the same receptors that are activated by extreme heat.

It can also rotate around any of these three axes. that's three more, unless it's a linear molecule, like carbon dioxide. There, one of the rotations just spins the molecule around its own axis, which doesn't count because it doesn't change the position of the atoms.

TED-Ed Example Questions

What are the receptors in your mouth that sense spicy compounds called?

How many rotational degrees of freedom does a non-linear molecule have?

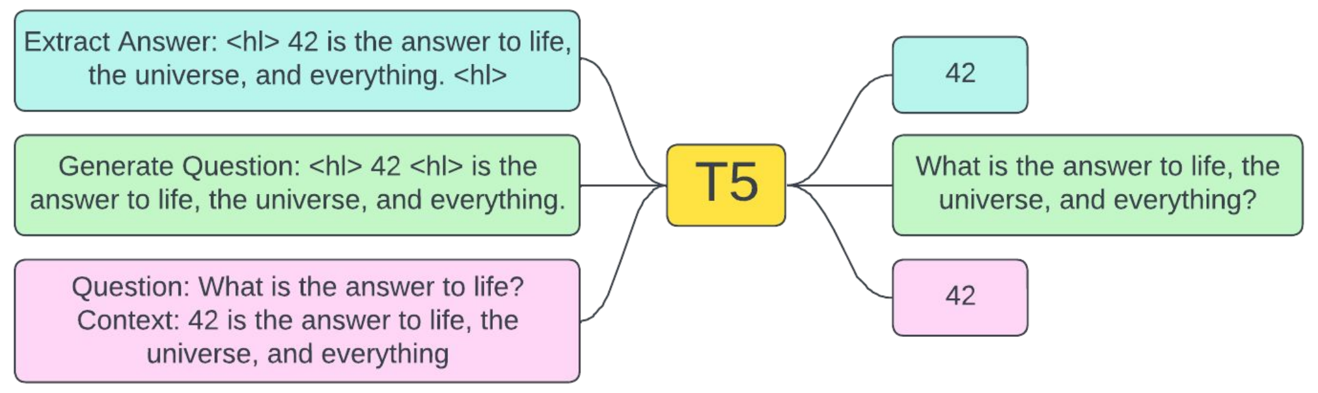
What do you think would happen if you held your nose while eating something really spicy?

Carbon dioxide has ____ rotational and ____ translational modes of motion.

Methodology

Fine-Tuning Transformer LLM

- T5 model[1]: frequently used for transfer learning
- E2E: Only context needed
- Multitask: Extract answer-like spans and generate questions for each span
- Finetune with question and context



Extract Answer: <hl> 42 is the answer to life, the universe, and everything. <hl>

Generate Question: <hl> 42 <hl> is the answer to life, the universe, and everything.

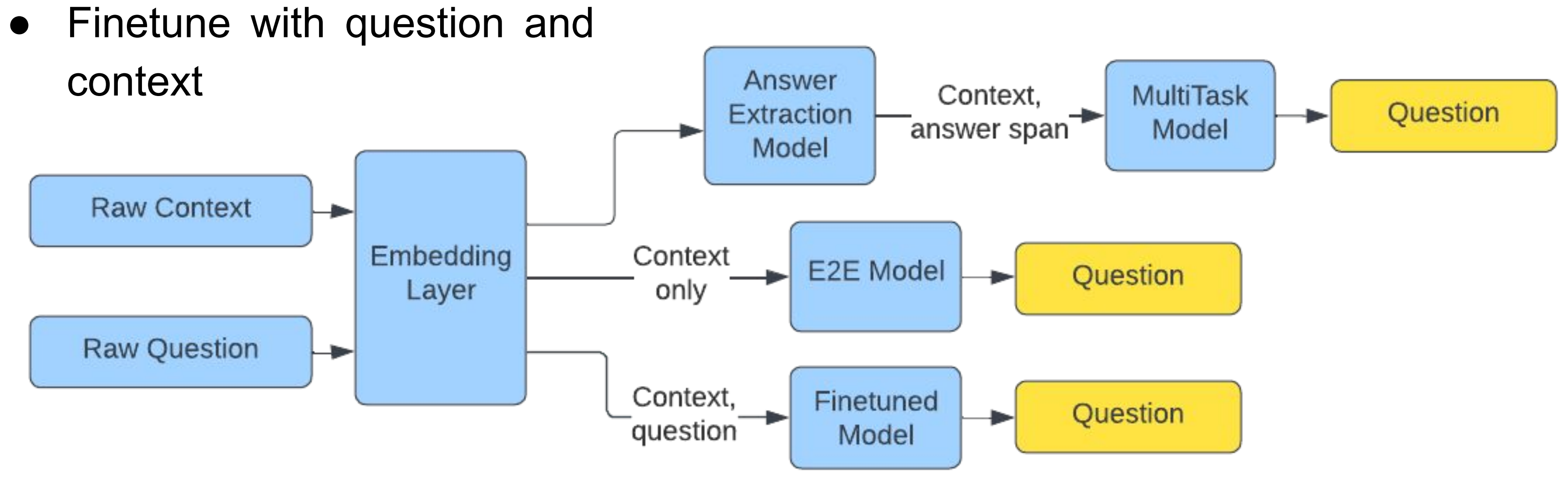
Question: What is the answer to life? Context: 42 is the answer to life, the universe, and everything

T5

42

What is the answer to life, the universe, and everything?

42



Raw Context

Raw Question

Embedding Layer

Answer Extraction Model

Context, answer span

MultiTask Model

Question

Context only

E2E Model

Question

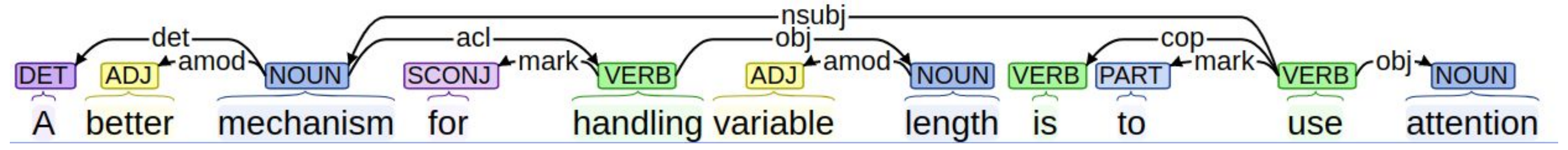
Context, question

Finetuned Model

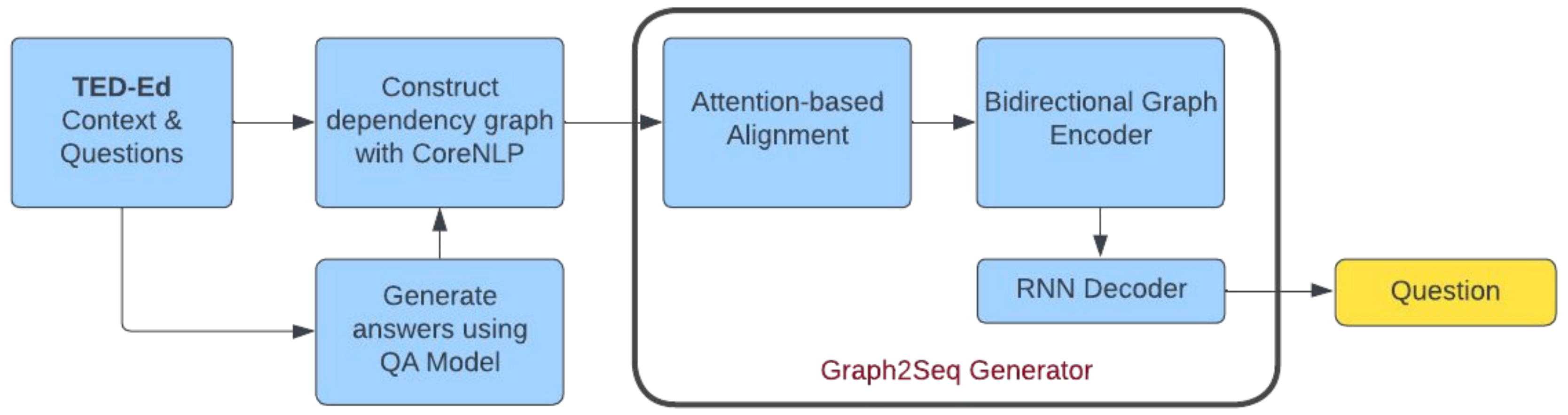
Question

GNN Graph-2-Seq

- Graph2Seq model[2] aims to utilize structural syntactic information to generate more natural text
- Generate dependency graphs using CoreNLP
- QA model generate answers to questions, used as input along with the question and context



A better mechanism for handling variable length is to use attention



TED-Ed Context & Questions

Construct dependency graph with CoreNLP

Generate answers using QA Model

Attention-based Alignment

Bidirectional Graph Encoder

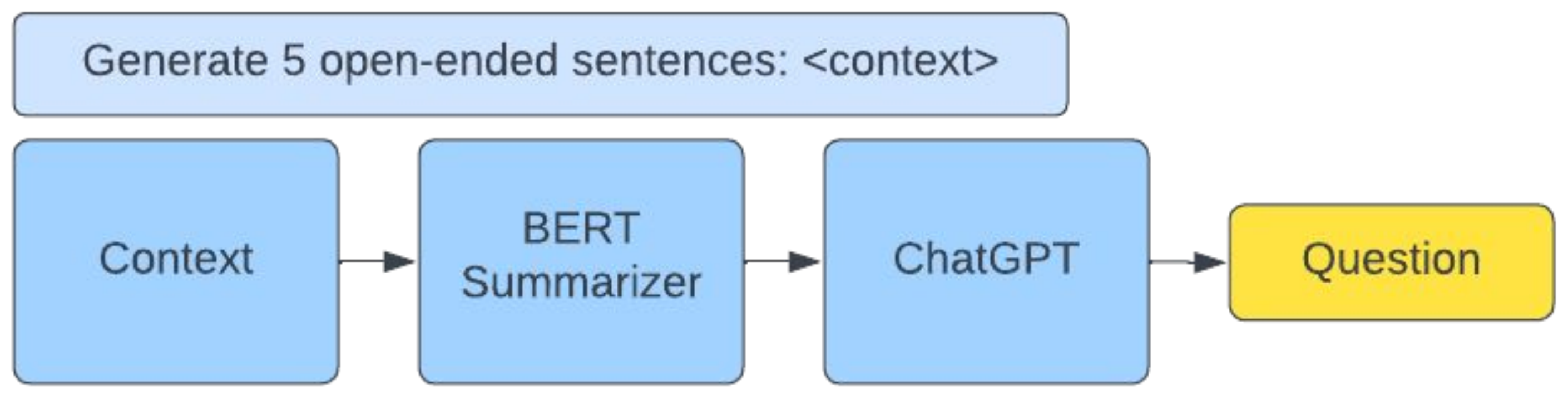
RNN Decoder

Question

Graph2Seq Generator

Zero-Shot Inference with ChatGPT

- Use a pre-trained BERT summarizer[4] to distill context passages into most relevant information, then use ChatGPT to generate questions



Generate 5 open-ended sentences: <context>

Context

BERT Summarizer

ChatGPT

Question

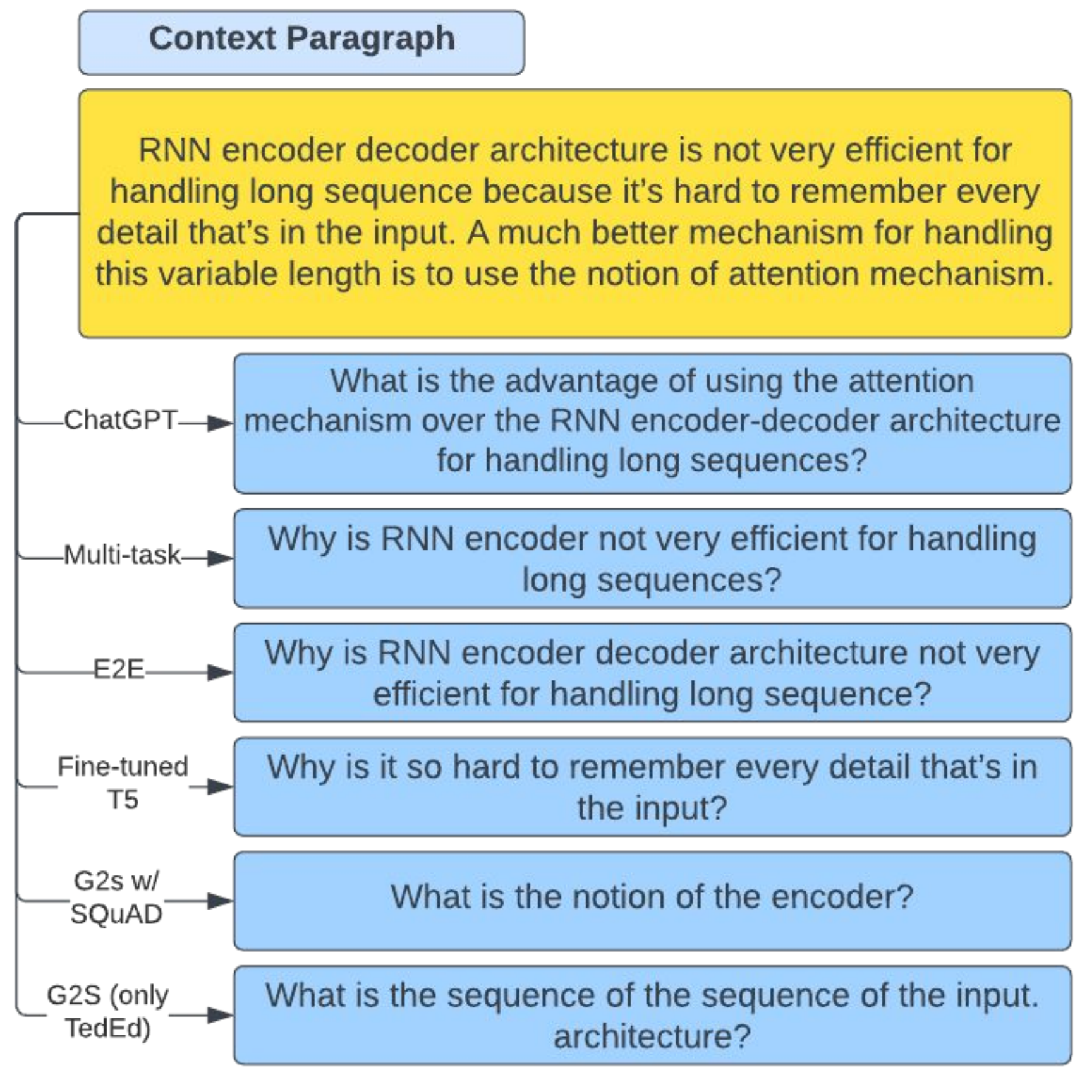
Evaluation

Evaluate using BLEU, ROUGE, and METEOR, as well as human evaluation (**C**oherence 1-5, **R**elevance 1-5, **O**pen-**E**ndedness 0-1)

Model	BLEU-1	BLEU-4	ROUGE-L	METEOR	C	R	OE
T5 End-To-End	**0.4885**	**0.1053**	**0.2624**	0.2358	4.37	3.25	0.18
T5 Multitask	0.4740	0.0659	0.2476	0.2171	4.0	3.63	0.02
T5 Finetuning	0.4702	0.0824	0.2581	**0.2699**	4.4	3.5	0.14
G2S (no answer)	0.1593	0.0270	0.2278	0.0723	2.2	1.8	0.1
G2S (generated)	0.1819	0.0230	0.2377	0.0762	2.67	2.6	0.1
0-shot ChatGPT	0.4585	0.0	0.2287	0.2641	**5**	**5**	**1**

ChatGPT surpasses all other methods on human evaluation

Case Study



Context Paragraph

RNN encoder decoder architecture is not very efficient for handling long sequence because it’s hard to remember every detail that’s in the input. A much better mechanism for handling this variable length is to use the notion of attention mechanism.

ChatGPT: What is the advantage of using the attention mechanism over the RNN encoder-decoder architecture for handling long sequences?

Multi-task: Why is RNN encoder not very efficient for handling long sequences?

E2E: Why is RNN encoder decoder architecture not very efficient for handling long sequence?

Fine-tuned T5: Why is it so hard to remember every detail that’s in the input?

G2s w/ SQuAD: What is the notion of the encoder?

G2S (only TedEd): What is the sequence of the sequence of the input. architecture?

Here are some samples of questions generated by the different methods on a passage taken from Prof. Lee’s lecture on RNNs.

The questions from ChatGPT and T5 are much more coherent than the G2S model. The degenerative behavior (“the sequence of the sequence”) observed in the G2S model is a common issue with sampling via beam search.

Conclusion

- We find that answer-agnostic generation of open-ended questions is a challenging task, but can be improved by leveraging question answering and answer extraction methods.
- The transformer-based models both significantly outperform the graph-based approach, in line with current state-of-the-art methods.
- Future work may involve experimenting with different input embeddings and sampling methods, such as nucleus sampling.

References

[1] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

[2] Chen, Yu, Lingfei Wu, and Mohammed J. Zaki. "Reinforcement learning based graph-to-sequence model for natural question generation." *arXiv preprint arXiv:1908.04942* (2019).

[3] Chen, Guanliang, et al. "LearningQ: a large-scale dataset for educational question generation." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. No. 1. 2018.

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).