本周任务：

从acl cl eacl findings 中

选取「QA QG」相关的论文

总结：任务背景；动机、现有问题与新的问题（场景）；文章贡献

至10月14日提交报告。

另外的工作：

方法类似与动机类似论文的归类

QG

23ACL

1 Modeling What-to-ask and How-to-ask for Answer-unaware Conversational Question Generation

对话问题任务主要分成2种任务设置。两者（answer-aware and answer-unaware）的区别在于是否在生成问题时可以参考问题的答案。显然，预先不能知道答案的情况更符合现实的状态。解决answer-unaware情形有2个挑战：问什么和怎么问。现有的启发式方法生成的对话不够自然。因为谈话者们未必是连续地谈论生成问题需要的内容。之前的方法生成的问题不能区分好回答yes or no和一段话的问题。

本文贡献在于给出一个两步架构（问什么、怎么问）SG-CQG；SOTA on answer-unaware CQG on CoQA；首个提出让模型理解估计生成对话的一套标准，以及衡量一个上下文生成对话的多样性的方法；透彻分析和评估生成对话的问题和答案。

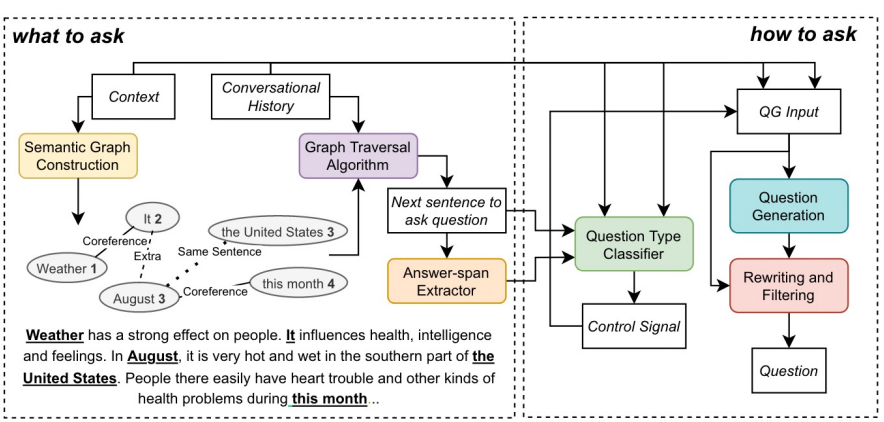


Figure 1: An overview of our proposed *SG-CQG* framework. It consists of two modules: the *what-to-ask* module aims to select a sentence as the rationale from the context and extracts the target answer span from it, and the *how-to-ask* module then predicts the type of the question to be generated and generates the question guided by that type.

2 Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model

是对话问题生成的论文，但是这篇让答案参与问题的生成。这两篇对于共指代的情况都用了聚类的方法解决。

这篇文章引入了Zero-Shot对话生成的办法来缓解标注的对话文本的不足的情况。使用了知识迁移的方法来利用单轮的生成知识于对话式的生成。

现有的问题生成大多关注于一问一答式的问题生成。但是不同于单轮的生成，对话式的生成需要生成上下文和对话历史相关的问题。这篇文章给出的方法可以不依靠人类标注的数据集来做训练；给出了一个多步的知识迁移架构，有效地把单轮问题生成的知识应用到Zero-shot的对话问题生成上；实验、优秀效果。

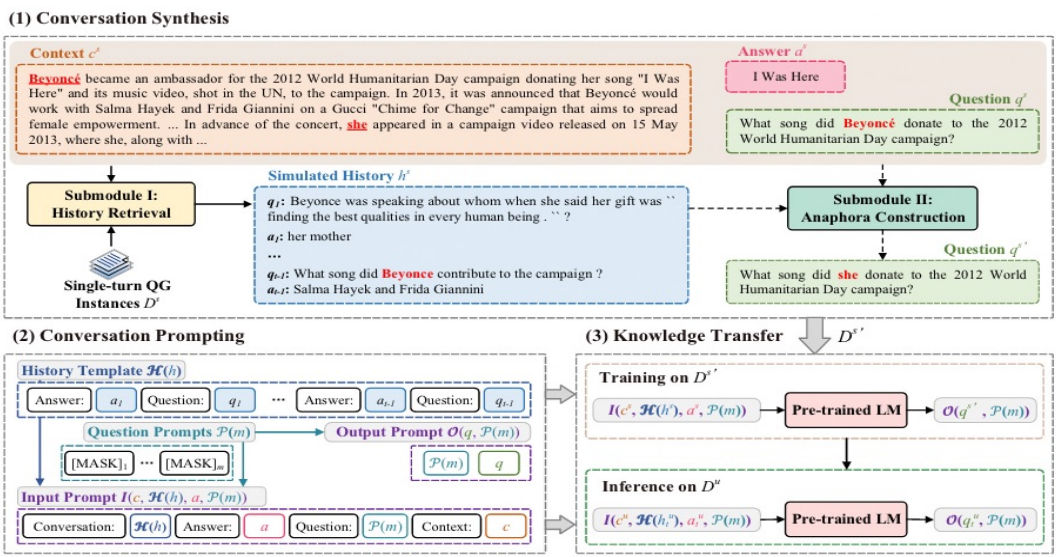


Figure 1: Illustration of the proposed SPARTA for ZeroCQG. **(1) Synthesize.** For each single-turn QG instance $(c^s, a^s, q^s) \in D^s$, the conversation synthesis module will retrieve $t-1$ question-answer pairs from D^s as simulated history h^s, and transform the q^s into $q^{s'}$ if there exist co-reference with pronoun, e.g. *Beyoncé* and *she*. We term the dataset with synthesized conversation as $D^{s'} = \{(c^s, h^s, a^s, q^{s'})\}$. **(2) Prompt.** We propose conversation prompting to reformulate the input and output of conversational QG. **(3) Transfer.** We fine-tune pre-trained LM on the prompted $D^{s'}$. Then, the fine-tuned pre-trained LM with the same conversation prompting is directly applied for inference on D^u to generate conversational questions.

23EACL

3 Socratic Question Generation: A Novel Dataset, Models, and Evaluation

这篇文章介绍了一个专用于生成苏格拉底式问题的数据集。苏格拉底问题是指用特定种类的探究型问题来引出人们对某个话题认识上的偏差，从而引出其他的视角或者更深层的思考。之前的问题生成数据集关注于找到和解决特定问题相关的对话内容、或者是为了训练开放领域的聊天机器人。这些数据集没有进行SoQG所需的上下文信息。

苏格拉底式上下文的三个特点：

·问题的上下文：上下文或者篇章表达了对于某个话题的想法或者意见

·问题的目标：问题不是为了一个正确的答案，而是为了引起反思

·问题的种类：问题强调上下文中表述的思考的完备性和准确性

这篇文章的贡献：提出了一个SoQG数据集、训练了预测问题种类的模型（效果好）、使用QG模型有效地生成近似人提出的问题。

Thought/Passage: I believe that eating meat is ethically wrong. Since we can easily substitute meat with vegan food without much nutritional complications, we have no logical reasons to continue eating living animals. We should stop killing these animals.

Possible Socratic Questions:

1. But what about eating animals that pass on from natural causes?

2. Where is your source that says that vegan food and meat have the same nutrition value?

3. What will happen if we continue killing animals for our consumption?

Figure 1: Example Socratic Questions

Question type	Description	Exemplar
Clarification	Question probing the ambiguities of a thought.	What do mean by ...?
Probing assumptions	Question probing the assumptions behind a thought.	Why do you assume ...?
Probing reasons and evidences	Question probing the justifications or concrete evidences that could have supported a thought.	How did you know that ...?
Probing implications and consequences	Question probing the impacts or implications of a thought.	If ..., what is likely to happen as a result?
Probing alternative viewpoints and perspectives	Question probing other possible viewpoints.	What else should we consider about ...?
Others*	Question unrelated to the question types above (e.g. rhetorical, irrelevant, and/or illogical questions, etc.)	Who wouldn't want to be rich?

Table 1: Description and exemplar for each Socratic Question-Type from Paul and Elder (1990; 2019). *We add the catch-all type Others to refer to questions that do not conform to Socratic categories.

4 Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels

这篇文章研究的是给定一个输入（答案），输出多个问题的情形。不同于之前的几篇文章，这篇文章不包括上下文的考虑。rule-based方法的表现不及Seq2Seq方法，而且设计规则工作量很大。过去的Seq2Seq方法是高数据导向的，对于专业性强的内容效果不好；而且Seq2Seq的方法大多对于一个输入只有一个输出（要输出多个问题只能指定不同的关键词或者答案范围，然而在这个任务中不提供这些条件），但是这篇文章希望能一次输出多个问题。文章把两者结合，给出了这个任务方向的SOTA表现。

这篇文章提出了一种结合rule-based和Seq2Seq的方法。训练时分别把问答对部分和全部地转换成含有语义角色标签的模式，再加入Seq2Seq模块中做训练；预测时根据输入答案转换后的内容，经过Seq2Seq模块后给出生成的问题。

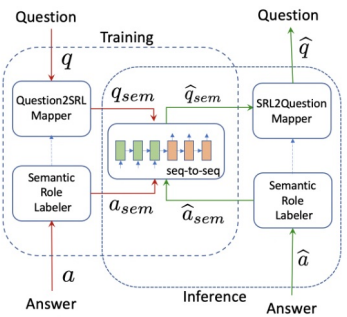


Figure 1: Overview of Proposed Framework

Table 1: Semantic representation of answers and questions (ARG0: agent , ARG1: patient, ARG2: attribute, ARGM-NEG: negation, ARGM-PRP: purpose)

Semantic representation for sample input sentences (answers):
S1. [ARG1: the fuel filler funnel] is [ARG2: under the luggage compartment floor covering] .
S2. [ARG0: this vehicle] has a capless refueling system and does [ARGM-NEG: not] have [ARG1: a fuel cap] .
S3. distribute [ARG1: the trailer load] [ARGM-PRP: so 10 - 15 % of the total trailer weight is on the tongue] .
S4. distribute the trailer load so [ARG1: 10 - 15 % of the total trailer weight] is [ARG2: on the tongue] .
Semantic representation for the questions corresponding to the above sentence representations in the training data:
Q1: where is [ARG1: the fuel filler funnel] ?
Q2: does [ARG0: this vehicle] have [ARG1: a fuel cap] ?
Q3: how much of [ARG1: the trailer load] should be on the tongue ?
Q4: how much of the trailer load should be [ARG2: on the tongue] ?

Table 4: Examples of sentence $\hat{a}$, its semantic representation $\hat{a}_{sem}$, the outcome $\hat{q}_{sem}$ generated by Seq2Seq, the question $\hat{q}$ converted from $\hat{q}_{sem}$, and ground-truth question Q_t from the Car Manuals dataset.

$\hat{a}$:	before placing a child in the child restraint , make sure it is securely held in place .
$\hat{a}_{sem}$:	before placing [ARG1] [ARG2] , make sure it is securely held in place .
$\hat{q}_{sem}$:	what should i do before placing [ARG1] [ARG2] ?
$\hat{q}$:	what should i do before placing a child in the child restraint ?
Q_t:	what should i do before placing a child in the child restraint ?
$\hat{a}$:	adjust the temperature setting using the + and - temperature buttons on the right-hand side of the climate controls .
$\hat{a}_{sem1}$:	adjust the [ARG1] setting using the + and - temperature buttons on the right-hand side of the climate controls .
$\hat{a}_{sem2}$:	adjust the temperature setting using [ARG1] [ARGM-LOC] .
$\hat{q}_{sem1}$:	how do i adjust the [ARG1] setting ?
$\hat{q}_{sem2}$:	how do i adjust the temperature [ARGM-LOC] ?
$\hat{q}_1$:	how do i adjust the temperature setting ?
$\hat{q}_2$:	how do i adjust the temperature on the right-hand side of the climate controls ?
Q_t:	how do i adjust the temperature on the passenger 's side ?

5 Closed-book Question Generation via Contrastive Learning

之前读的文章的任务很多都是“open-book”的（可以参考外界知识，比如抽取的文档、文章），这篇文章讨论的是“closed-book”的情况（没有外界知识）。动机在于：怎样能让一个QG模型更好地理解抽象的长句回答、并且让模型在它的参数中保存更多信息呢？这个任务根据答案来生成问题，但是不依靠其他的知识。

这篇文章的模型有三个部分。其一是问题生成模块；其二是一个对比学习模块，目的是让模型把正确的问答对放到一起，把错误的分开（因为编码的答案应该和问题有一定的近似度，而和其他的问题不近似）；其三是一个答案重构的模块，这里用一个seq2seq模型根据生成的问题给出一个答案，生成时用了gumbel-softmax来增加多样性，生成的句子和问题计算损失函数。把三个模块的损失函数结合起来，来更新问题生成模块的参数。

这篇文章的贡献：一个对比式QG模型，首个close-book任务中采用contrastive learning的工作；在三个数据集表现良好，人工评估良好；把QG模型作为一个扩充数据的策略，用来生成大量的QA对。

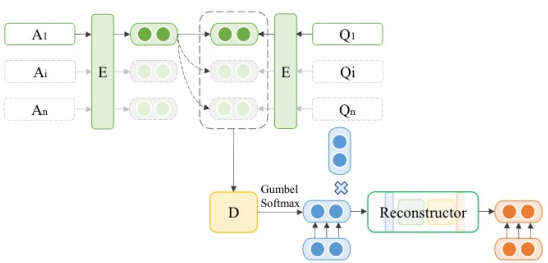


Figure 1: An overview of the proposed closed-book QG framework, which consists of three parts: contrastive learning, question generation, and answer construction. A_i: represents answer i; Q_i represents question i.

Conclusion

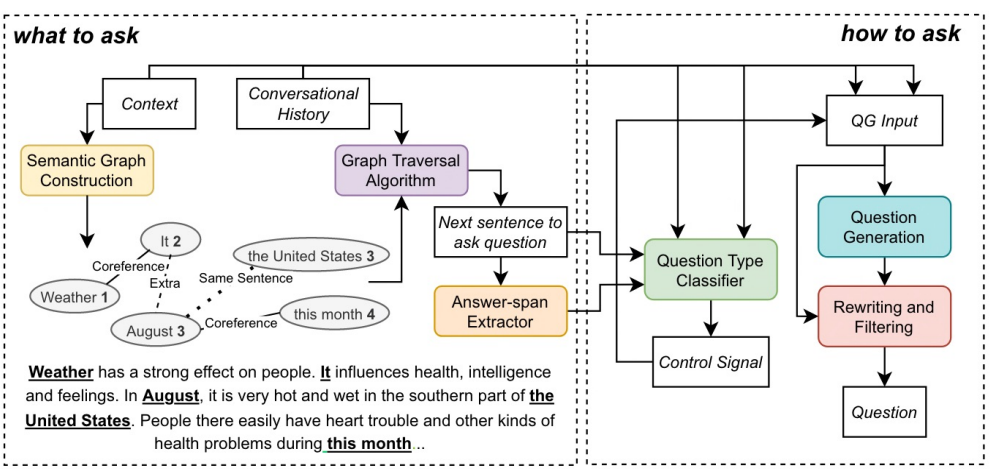
· **对话式问题生成一般需要考虑是否有答案可以辅助。**

1 Modeling What-to-ask and How-to-ask for Answer-unaware Conversational Question Generation

围绕**没有给出答案**的情况展开。这篇文章提到在对话中提出问题是帮助agent和人类更有效交流的重要行为。为了解决之前方法的几个限制：生成重复问题、损失信息的问题以及不能区分回答是否还是一段话，这篇文章提出了一个两步模型：问什么模块鼓励模型生成相关的对话；怎么问模块促进生成自然且多样的问题。

WTA模块构建了**语义图**，首先聚类得到语句的共指代，把每个簇里的节点连接到下文最接近的节点上（共指代）。再抽取所有的命名实体（为什么不先抽取），如果这些节点不在簇里，就把他们额外连接起来。再把同一句话中的所有节点连接起来（同一句话），最后用额外的边把所有子图连接起来（每两句话也这样连接），因为这种边没有任何语义信息，目标是减少这类边的数量。建立图之后，结合对话历史，建立一个存储遍历节点的队列。如果能在目前节点生成知识，就把它相关的未被访问的节点放在队列最前面（这样对话相关性好），否则就把它相关的未被访问的节点放在队列最后，再pop这个节点。根据生成的问题，使用一个t5模型来抽取出它的答案。

HTA模块首先需要根据输入的问题和答案给出问题所需的答案类型（是否/一段话）。用一个t5模型回答生成的问题，再检验问题和答案的相似度，过高的相似度问题会被忽略。另外还有对应错误答案、不相关、信息谬误和赘余的纠错方法。最后根据上下文、对话历史、答案和rationale以及问题类型输出生成的问题。



2 Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model

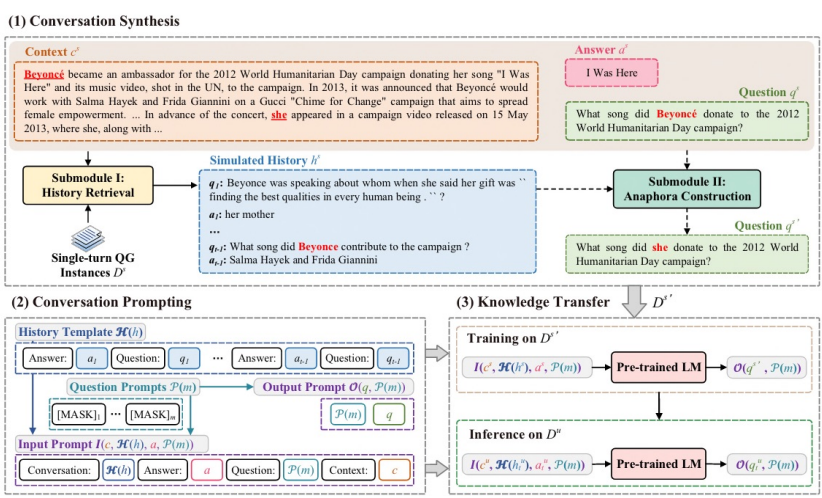
很多问题生成研究主要关注于单轮的问题生成，但是对话式的生成需要综合上下文的内容。此外，这篇文章之前的的对话式问题生成模型虽然在CoQA效果很好，但是严重依赖于大量标记的对话来作为对话历史和接下来的问题的来源。在单轮对话的知识**迁移**到多轮的情景，提出了SPARTA模型，来实现**Zero-shot**的对话式问题生成（Zero-CQG）。这篇文章的问题生成是**answer-awared**的模式。

Zero-CQG不能使用任何人类标注的对话，可以使用一段上下文、一个问答历史、一个给出的答案，目标是生成一个问题。以下分别简单介绍Synthesize, Prompt, Transfer的方法。

· 为了减少单轮和多轮对话生成的领域差别，使用合成对话的方法。首先，对于特定的问题q和上下文c，在整个数据集中检索出有类似的上下文问题和答案对，再计算出问答对和目标上下文c的相似度。由此可以确保可回答性。对得到的QA对按照和q的相似度排序，使用基于BERT的NSP任务，根据每个问题和目标q的概率为前后句的概率做排序，概率越高的和目标q越近，从而对于每组单轮的数据{c,a,q}我们可以得到一个模拟的对话历史序列h {(qi,ai)}。为了解决共指代的问题，这里也用了聚类的方法（聚现实世界的实体），之后用代词将共指代的实体替换掉。

· 为了减少与训练模型和CQG的领域差别，使用Prompt的方法处理数据。把CQG问题转化成一个masked question-filling任务。输入部分对历史序列使用question:和answer:这样的前缀做处理；对于masked question的部分，使用prompt P，用含有[mask]的token遮蔽对话中对应的位置。输出时也通过之前的prompt P来给出模型输出。

· **知识迁移**把单轮的问题生成的知识转移到多轮上。这里使用合成的数据集做训练，用之前的prompt方法来构建数据，损失函数使用了交叉熵。



3 Socratic Question Generation: A Novel Dataset, Models, and Evaluation

这篇文章介绍了一个专用于生成苏格拉底式问题的数据集。苏格拉底问题是指用特定种类的探究型问题来引出人们对某个话题认识上的偏差，从而引出其他的视角或者更深层的思考。之前的问题生成数据集关注于找到和解决特定问题相关的对话内容、或者是为了训练开放领域的聊天机器人。这些数据集没有进行SoQG所需的上下文信息。

Thought/Passage: I believe that eating meat is ethically wrong. Since we can easily substitute meat with vegan food without much nutritional complications, we have no logical reasons to continue eating living animals. We should stop killing these animals.

Possible Socratic Questions:

1. But what about eating animals that pass on from natural causes?

2. Where is your source that says that vegan food and meat have the same nutrition value?

3. What will happen if we continue killing animals for our consumption?

Figure 1: Example Socratic Questions

4 Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels

这篇文章是**answer-aware**的。这篇文章结合了**rule-based**和Seq2Seq模式。Seq2Seq方法是高度数据导向的，对于数据有限的领域输入不能很好地理解。此外生成多个问题，Seq2Seq方法需要给出关键词、答案范围等。Beam搜索则需要指定生成多少问题。这篇文章对输入句子做语义角色标识，再训练一个Seq2Seq模型来生成问题。不需要训练集中含有关键词、答案范围标签，也不需要特别指出需要生成多少个问题。语义角色标记增加了训练例子数量，也可以减轻**标记数据有限**的问题。

这篇文章和**Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model**都是希望可以缓解训练数据少的问题。

其中的**语义角色标记**类似之前读到的把实体转换成代词的方法，**训练时**的回答输入需要用SRLer来生成Seq2Seq的输入。如果有多个谓词，有可能会在一个输入生成多个语义表达的输出。而处理问题输入时并不会直接使用SRLer来处理，有两种处理方法。Hard模式用出现在答案中的某个实体的标签I替代掉问题中再次出现的相对应的实体。因为Hard模式可能找不到完全一致的表述（尽管它们的意思一样），为了解决这个问题，使用soft模式来根据问题中单词或者短语和语义角色的相似度给出最佳匹配。Seq2Seq模型使用BART和T5来评估。

推理时先用SRLer来转化输入的答案，再用训练过的Seq2Seq转化成q的语义表示，再用SRL2Question Mapper生成答案。

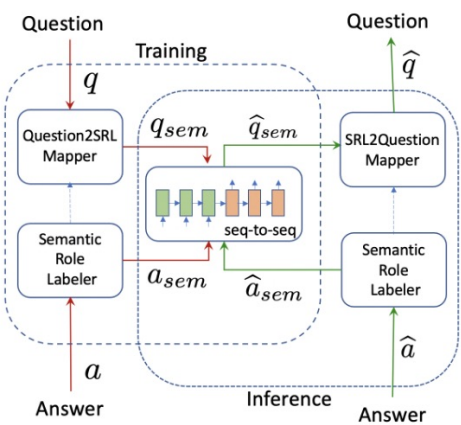
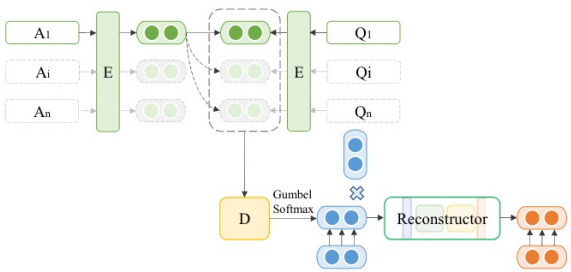


Figure 1: Overview of Proposed Framework

5 Closed-book Question Generation via Contrastive Learning

这篇文章侧重于**无外界知识**的情况。有两个困难：生成问题时不能从外界检索，**只能根据输入的答案生成问题**；多数closed-book的工作使用的数据集忽略了和答案相关的文章。这些答案一般比较短，比如一些实体，相比于长的答案，更容易被语言模型记住并且存储在模型的参数之中。文章使用了对比学习的方法让模型理解答案和问题之间的关系，并且设计了一个重构问题的损失函数来衡量生成的问题的可回答性。这篇文章还使用QG模型作为一种数据增加的策略来生成大规模的QA对，且证明了效果。



首先根据输入的答案生成一个回答，这里和上一篇一样，也是一个Seq2Seq模型。为了让编码的答案应当和它的问题有相关性，但又和其他的问题不那么相关，使用了对比学习的方法。z代表编码后的结果（这里用[cls]），sim是计算余弦相似度的函数。

$$\mathcal{L}_{cl} = -\log \frac{\exp(\text{sim}(\mathbf{z}_{x_i}, \mathbf{z}_{y_i})/\tau_{cl})}{\sum_{i=1}^{2n} \exp(\text{sim}(\mathbf{z}_{x_i}, \mathbf{z}_{y_i})/\tau_{cl})}, \quad (4)$$

为了确保可回答性，这里使用了一个问题重构模块，根据生成的问题再生成答案，再计算它们的损失函数。一个主要的挑战在于生成的问题不够多元化，所以引入了ST Gumbel-Softmax。根据公式计算概率最高的传递到下一层，反向传播的梯度根据Gumbel-Softmax给出。通过ST Gumbel-Softmax得到的one-hot向量和vocabulary嵌入相乘，再送入一个预训练过的Seq2Seq的模型作为生成的问题的表示。

最后把损失函数融合到一起，更新生成问题的参数。

$$\mathcal{L} = \lambda_1 \mathcal{L}_{qg} + \lambda_2 \mathcal{L}_{cl} + \lambda_3 \mathcal{L}_{ar}, \quad (7)$$

The weights λ_1,λ_2, and λ_3 are tuneable hyper-parameters to balance losses and the final objective is to minimize the overall loss.

Algorithm 1: QG framework.

Input: Pre-trained language model $p(q|a)$, answer reconstruction model $p(a|q)$ and answer-question pairs

Output: Question generator $p(q|a)$

for $i \leftarrow 1$ **to** *Epoch* **do**
  $\hat{q} = p(q|a)$
  Compute $\mathcal{L}_{qg}$ via Eq. 3
  /* contrastive learning */
  $a_i = \text{Encoder}([\text{CLS}] \oplus a)[:, 0:]$
  $a_i^+ = \text{Encoder}([\text{CLS}] \oplus q)[:, 0:]$
  Get $\mathcal{L}_{cl}$ to (a_i, a_i^+) via Eq. 4
  /* answer reconstruction */
  $\hat{a} = p(a|\hat{q})$
  Compute $\mathcal{L}_{ar}$ via Eq. 5
  Calculate total loss $\mathcal{L}$ via Eq. 7
  Update generator $p(q|a)$ with $\mathcal{L}$
end
return question generator $p(q|a)$

任务的总结：

1 Modeling What-to-ask and How-to-ask for Answer-unaware Conversational Question Generation 对话问题生成 给定上下文，给定对话历史，预测当前问题与答案

2 Synthesize, Prompt and Transfer: Zero-shot Conversational Question Generation with Pre-trained Language Model 对话问题生成(Zero-shot) 给定上下文，给定对话历史，给定答案来生成问题（不使用任何人工标注的对话）

4 Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels

给定答案，生成问题

5 Closed-book Question Generation via Contrastive Learning(禁止外界知识) 给定答案，生成问题

共同点的总结

把3刨除在外，因为它是一篇关于特定任务的新的数据集的文章。

对话式：1 2

answer-aware：2 4 5

禁用外界知识（至少不使用对话历史）：4 5

用聚类来处理共指代问题： 1 2

Seq2Seq: 4 5

rule-based:1 4

常用metrics
BLEU ROUGE METEOR BERTSCORE

https://en.wikipedia.org/wiki/BLEU

https://en.wikipedia.org/wiki/ROUGE_(metric)

https://en.wikipedia.org/wiki/METEOR

https://arxiv.org/abs/1904.09675

Findings

Open-World Factually Consistent Question Generation

预训练模型经常遇到事实的前后矛盾和错误实体识别的问题，导致生成无法回答的问题。此外还有领域迁移的问题，训练的数据集领域和测试的数据集领域不同更加剧了幻觉问题。文章提出了一种有效的数据处理方法。这种方法缓解了幻觉问题，增强了领域切换的鲁棒性，而且可以在不显著影响传统的测试指标的情况下生成事件级的事实可信问题。

这篇文章尝试解决问题生成中的事件级别的事实矛盾（首个）。不同于之前总结工作，这篇文章的工作独立于模型和训练过程。不会通过过滤的方法来减少数据集的数量，而是预处理数据集来促使生成关于输入的可信问题。

大概方法：先抽取实体，然后根据某个规则替代掉抽取到的实体，然后在生成后替换回来. 之后有两步筛选：若生成问题包括了输入不存在中的实体，认为是幻觉；选择前五个beam中事实前后一致且困惑值最低的作为输出，没有事实一致的就输出困惑最低的。

这种方法在不处理模型结构和过滤数据集的情况下得到了很好的效果。

Evaluation of Question Generation Needs More References

这篇文章给出了一个衡量QG任务（给出上下文和答案）的新指标。过往的方法经常是根据一个gold问题来计算生成的问题的表现（Single Reference Evaluation），但是一个问题可以有多种正确的问法，而只依靠一个gold问题做计算很可能让好的生成被忽视掉。现有的根据gold生成reference的办法要么限制于其他生成任务，要么难以用于问题生成。这篇文章使用大模型来生成gold的补充来评价QG结构（首个）。这种方法(Multi-Reference Evaluation)可以用于多种旧的metrics，用来更好地评估模型的表现。

为了衡量生成的reference的好坏，标记了3164个{上下文，答案，生成的reference}的数据，对于每个样本，人类对流畅、可回答且符合上下文的标记1，否则标记0，取平均值后选择标记为1的作为reference问题，剩下的问题使用Peason correlation coefficient和Speaman's rank correlation coefficient衡量人类分数和自动评价的分数。

	Pearson Correlation					Spearman Correlation				
	SRE	MRE				SRE	MRE			
		HRQ-VAE	GPT-3 (0-shot)	GPT-3 (3-shot)	ChatGPT (0-shot)		HRQ-VAE	GPT-3 (0-shot)	GPT-3 (3-shot)	ChatGPT (0-shot)
BLEU-4	0.2028	0.2443	0.2782	0.3162	**0.3630**	0.2772	0.3224	0.2688	0.3021	**0.3340**
ROUGE-L	0.2908	0.3325	0.3241	0.3447	**0.3799**	0.2787	0.3270	0.3050	0.3330	**0.3637**
RQUGE	0.2932	-	-	-	-	0.2571	-	-	-	-
METEOR	0.3447	0.2968	0.3480	0.3877	**0.4116**	0.3111	0.2822	0.3159	0.3562	**0.3780**
BERTScore	0.3556	0.3634	0.3552	0.3877	**0.4033**	0.3462	0.3568	0.3327	0.3723	**0.3859**
MoverScore	0.4383	0.3835	0.4297	0.4693	**0.4953**	0.3882	0.3643	0.3885	0.4214	**0.4292**
BLEURT	0.4739	0.4287	0.4656	0.4803	**0.5019**	0.4566	0.4193	0.4456	0.4648	**0.4816**

Table 1: Results of the correlation coefficient between measured metrics and human score. The best scores in methodology are in bold, and the best scores in metrics are underlined. These depend on the types of correlation measures. '-' denotes unreported results.

RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question

这篇文章介绍了新的metric。过去的Bleu、ROUGE及BERTScore等方法高度依靠给出的reference问题的多样性，而且对于小的变化不够敏感，可能会误判好的问题。这篇文章提出了RQUGE，通过一个QA模块和一个评估模块来评价问题的生成质量，不需要获得任何reference问题，和人类判断也更加接近，此外鲁棒性更好，还通过RQUGE重排序的QG模型生成的合成数据，微调后提升了QA模型在领域外数据集的表现。

大概的方法：给定答案和上下文生成问题，再对这个上下文和生成问题生成一个新的答案，按照[CLS]生成问题[SEP]新的生成答案[SEP]标准答案[SEP]上下文[SEP]的模式拼接，输入一个给分模块，给出对于生成的问题的得分。

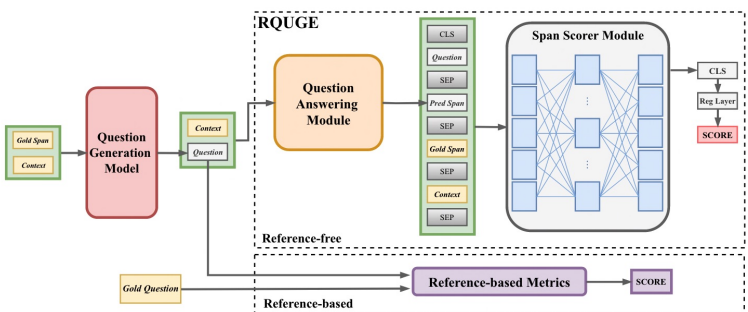


Figure 2: The architecture of RQUGE metric (upper-side) for the question generation task, which consists of a question answering and a span scorer modules to compute the acceptability of the candidate question. Reference-based metrics are also shown at bottom of the figure, where the score is calculated by comparing the gold and predicted questions.

Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation

大模型难于训练与部署，缺乏简单且鲁棒的办法从随机的样本中得到最佳输出。fine-tune方法很贵，prompt-tuning需要数据的ground-truth标签，也需要资源。这篇文章采用prompt方法来增强大模型的zeroshot问题生成（给出context和答案）能力，提出了两种方法：round trip和prompt-based score。metrics：对SQuAD使用了BLEU-4；对FairytaleQA使用了ROUGE-L还引入了人类评估。

Round-trip: q'=QG(c, a); a'=QA(c, q') (GPT-3回答) ; a'=a 计算a'和a对相似度 使用BLEU或者ROUGE

Prompt-based Score: 首先给出上下文和问题，令GPT-3回答。再另GPT-3给出评分。