计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(4): 905−914, 2025

自适应的流水线式无监督问题生成方法

李昆泽　张　宇
（哈尔滨工业大学社会计算与交互机器人研究中心　哈尔滨　150001）
（kzli@ir.hit.edu.cn）

Adaptive Pipeline Unsupervised Question Generation Method

Li Kunze and Zhang Yu
（*Research Center for Social Computing and Interactive Robotics*, *Harbin Institute of Technology*, *Harbin* 150001）

Abstract　In traditional question-answering tasks, models generally require extensive data for training, which entails considerable time and manpower costs for data annotation. Unsupervised question generation represents an effective solution to address the scarcity of training data in question-answering tasks. However, the questions generated using this approach currently suffer from issues such as being difficult to answer, lacking variety, and having unclear semantics. To address these issues, we propose an adaptive multi-module pipeline model named ADVICE, with modules improving existing methods in answerability, question diversity and grammatical correctness. Within the question answerability module, we employ coreference resolution and named entity recognition techniques to improve the answerability of questions. For question diversity, we design specific rules for various question types to enhance the diversity of question and answer types. In the grammatical correctness module, a grammar error correction model targeted at questions is trained based on T5 model, and a filtering module is designed to refine the generated question-answer data. Finally, a classifier is trained to automatically select the necessary modules. Experiments demonstrate that the improved question generation method enhances the performance of downstream question-answering models on the SQuAD dataset, with the *EM* (exact match) score increasing by an average of 2.9% and the *F*1 score by an average of 4.4%.

Key words　unsupervised learning; question generation; pre-trained model; deep learning; natural language processing

摘　要　在传统的问答任务中，模型一般需要大量的数据进行训练，而标注这些数据需要较多的时间和人力成本. 无监督问题生成是解决问答任务训练数据匮乏的一种有效方法，但是目前使用该方法生成的问题存在难以回答、种类单一、语义不明等问题. 针对这些问题，提出了一个自适应的多模块流水线式模型ADVICE，多个模块分别从问题可回答性、问题多样性和语法规范性对现有方法进行改进. 在问题可回答性模块中，使用了共指消解和命名实体识别技术来提升问题的可回答性. 在问题多样性模块中，针对不同提问方式的问题设计了不同的规则来提升问题类型多样性与答案类型多样性. 在语法规范性模块中，基于T5训练了一个针对问句的语法错误纠正模型，并设计了一个筛选模块对纠正后的问答数据进行过滤. 最后，训练了一个分类器自动选择所需要的模块. 实验表明，使用改进后的问题生成方法，下游的问答模型在SQuAD数据集上的*EM*值平均提升了2.9个百分点，*F*1值平均提升了4.4个百分点.

收稿日期：2023-10-31；**修回日期**：2024-07-08
基金项目：国家自然科学基金项目（62476066, 62277002）
This work was supported by the National Natural Science Foundation of China (62476066, 62277002).
通信作者：张宇（zhangyu@ir.hit.edu.cn）

关键词 无监督学习;问题生成;预训练模型;深度学习;自然语言处理

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202330857 **CSTR:** 32373.14.issn1000-1239.202330857

问答(question answering, QA)任务要求机器回答以自然语言形式提出的问题. 但是, 现阶段效果比较好的问答模型一般都需要大量的数据进行训练, 而标注这些数据需要大量的时间和人力成本. 在这样的前提下, 问题生成(question generation, QG)任务应运而生. 问题生成任务定义为: 给定一个自然语言的文本, 自动生成内容相关、语法通顺的问题[1], 图 1 给出了一个问题生成任务的示例:

上下文: Oxygen is used in cellular respiration and released by photosynthesis, which uses the energy of sunlight to produce oxygen from water.
问题 1: What life process produces oxygen in the presence of light? 答案: photosynthesis 问题 2: From what does photosynthesis get oxygen? 答案: water

Fig. 1 Example of question generation

图 1 问题生成示例

图 1 中“上下文”部分是一段英文文本, 作为任务的输入; “问题”和“答案”部分是问题生成任务的输出, 由若干个问答对组成. 在本例中, 输出由 2 个问题及对应的答案组成, 每个问题均可以根据上下文中的文本信息作答.

问题生成可以在一定程度上缓解问答领域的数据缺乏问题. 在缺乏人工标注的问答数据时, 可以利用问题生成技术, 在短时间内生成大量数据. 与此同时, 问题生成在教育领域同样有很强的应用价值. 问题生成技术可以缓解教育工作者们在出题上的压力, 从而使他们可以花费更多时间在其他更重要的教学活动上. 并且, 问题生成技术生成的问题可以作为额外的教学资源供学生练习使用[2], 并利用人工智能辅助教育[3].

问题生成通常可以分为有监督和无监督 2 种. 有监督问题生成通常将该任务建模成 Seq2Seq 问题, 并利用现有的问答数据集来对深度学习模型进行训练, 代表工作来自 Dong 等人[4] 提出的 UniLM. 但是, 有监督问题生成模型仍然需要人工标注的数据进行训练, 从而不能完全解决数据匮乏的问题. 无监督问题生成常常直接利用规则等无监督方法来从文本中生成问题, 代表工作为 Lewis 等人[5] 提出的 ClozeQA 方法, 先生成一个完形填空形式的问题, 再转化为最终问题, 或者使用深度学习模型对生成的问题进行优化, 代表工作为 Li 等人[6] 提出的 RefQA, 使用 BERT[7] 模型对生成的问题进行优化. 这些方法不再需要人工标注问答数据, 从而可以更好地解决问答模型训练数据短缺的问题. 因此, 本文主要研究无监督问题生成.

本文发现现有的无监督问题生成方法虽然能在一定程度上缓解问答领域的数据匮乏问题, 但使用这些方法得到的数据训练问答模型的效果仍不够理想, 与使用人为标注数据训练模型的效果还有较大差距. 除此之外, 生成问题本身的质量也不够高, 这会导致问题很难应用在真实的教育场景中, 从而影响问题的教育价值. 经过分析, 本文发现现有方法存在着难以回答、种类单一、语义不明等问题, 并针对这些问题设计了问题可回答性(answerability)、问题多样性(diversity)和语法规范性(grammatical correctness)的改进模块. 在此基础上, 本文还训练了一个分类器为具体样本自动选择所需要的模块, 最终构建了一个完整的流水线式模型 ADVICE.

在问题可回答性的改进模块中, 本文借助共指消解与命名实体识别技术从不同角度对问题的可回答性进行了改进; 在问题多样性的改进模块中, 以提问方式对问题进行区分, 针对不同的提问方式设计了基于规则和句法分析的方法来提升问题的多样性; 在语法规范性的改进模块中, 基于自建数据集训练了一个语法错误纠正模型来改进问题的语法规范性, 并保证对问题的语义不做过度修改. 这 3 个模块形成了一种流水线式的结构, 具体样本会根据分类器的判断结果通过其中的若干模块.

实验表明, 现有无监督问题生成方法得到的问答数据在经过本文提出的流水线式模型修改后作用在下游的问答模型上, 模型在 SQuAD[8] 数据集上的 *EM* 值和 *F*1 值相较于之前分别平均提升 2.9 个百分点和 4.4 个百分点. 除此之外, 本文还验证了改进方法可以在一定程度上提高生成问题的质量, 从而提升问题的教育价值.

本文的主要贡献有 3 点:

1)提出了一种自适应的多模块流水线式模型, 缓解了现有问题生成方法在可回答性、多样性、语法规范性等方面存在的问题, 在改善生成问题质量的同时有助于提升下游问答模型的性能.

2)提出了一种由分类任务指导下游问题生成方法的新思路，能够根据具体情况选择合适的生成策略，进行更加精细化的问题生成.

3)通过在公共数据集 SQuAD 上进行实验和人工评价，验证了本文方法在提升下游问答模型的性能和改善生成问题的质量方面的有效性.

1 相关工作

近年来，随着问答领域对数据的需求量越来越大，问题生成技术逐渐得到越来越多的关注，而这其中最能有效缓解数据缺乏问题的便是无监督问题生成. 抽取式问答是问答领域的主流方向之一，与此同时抽取式的问题生成也是问题生成领域较早开展的工作. 现有的抽取式无监督问题生成方法大致可以分为两大类：基于规则的方法和规则与深度学习相结合的方法.

1)基于规则的方法. 这类方法使用人为设计的规则进行问题生成，生成问题的效率较高. Lewis 等人[5]选择名词短语或命名实体作为答案，进而对文本中的答案部分进行遮盖(mask)操作，从而生成一个完形填空形式的问题. 进而分别提出 4 种方法将这些问题转化成更自然的问题. 这 4 种方法分别是：①Identity Mapping，直接使用疑问词来替换被遮盖的部分来生成问题；②Noisy Clozes，对完形填空形式的问题施加随机的干扰(如词序变换等)，然后随机或利用启发式算法确定疑问词；③Rule-Based，利用现有的 Heilman 等人[9]提出的方法将陈述句转化成问题；④Seq2Seq，利用一个无监督机器翻译模型进行转化. 经过实验发现，利用无监督机器翻译模型生成问题的效果最好. Fabbri 等人[10]提出了一种两步式的方法. 针对待提问的句子，首先在语料库中寻找与之相似的句子，然后利用这个相似的句子结合模板生成问题. 假设返回的句子为 [*A*][Answer][*B*] 的结构，该方法可以生成形如“*A*+Wh+*B*+?”和“Wh+*B*+*A*+?”两种形式的问题. 基于规则的方法虽然效率很高，但由于规则往往很难考虑到所有可能出现的情况，导致生成问题的质量参差不齐.

2)规则与深度学习相结合的方法. 这类方法将规则与深度学习相结合，比完全基于规则的方法要更加灵活、更加有效. Li 等人[6]提出了 RefQA 方法. 这一方法很好地利用了维基百科的特点，把被引用的文档作为原文，使用当前引用的句子生成问题. 这很好地解决了生成出的问题和原文有大量单词重合的问题. 在生成问题时，该方法利用了依存解析树来将完形填空形式的问题转化成更自然的问题，并使用 BERT 模型进行筛选. Lyu 等人[11]将新闻类的数据作为原文，将人为标注的摘要作为待提问的句子来生成问题，生成模型选用的是 BART[12]. 由于摘要是人为标注的，质量很高，并且和原文之间的重复率低，不容易产生冗余信息，所以最终生成问题的效果也比较好. Nie 等人[13]从答案的多样性角度入手，提出了 DiverseQA 方法，发现 SQuAD, NewsQA[14]等数据集中以命名实体作为问题答案的大约占总问题答案数的 50%，而现有无监督问题生成方法生成出的问题几乎都以命名实体作为答案，过于单一. 文献 [13] 提出了一种改进答案多样性的方法，步骤为：①提取句子中的命名实体；②迭代地对包含命名实体的短语进行拓展，直到该短语的长度超过一定阈值停止迭代；③利用 BERT 模型对生成的问题进行筛选. 规则与深度学习相结合的方法中，深度学习模型常常只起到一个筛选、过滤的作用，导致最终生成问题的质量虽略有提升但仍然不够理想，且生成速度较慢.

现有的无监督问题生成方法基本秉持着“先确定答案，再生成问题”的思路. 在生成问题时，为了避免生成的问题与原文重叠率较高，部分方法选择利用一些外部语料生成问题，导致生成出的问题难以回答，可回答性较差；在确定答案时，现有方法通常选择命名实体作为答案，提问方式也是预先定义好的，导致问题单一、多样性较差；除此之外，部分方法选择基于规则的方法来生成，导致生成出的问题存在语法错误、语义不明、语法规范性较差.

2 自适应的多模块流水线式模型

本节提出的自适应流水线式模型 ADVICE 可以看作是对现有无监督问题生成方法的“后处理”，即不对原方法进行任何修改，而是对其生成的问答数据进行进一步的处理. ADVICE 包含 3 个模块和 1 个分类器. 图 2 是模型的整体结构图.

现有无监督问题生成模型生成出的问答数据会首先输入分类器. 分类器分别从可回答性、多样性和语法 3 个角度对问题进行判断与分类，并根据分类结果选择相应的模块. 这 3 个模块是流水线模型的主体，问答数据会按照编号从小到大的顺序依次通过被分类器选中的模块，并最终输出修改后的数据. 下文将依次介绍图中的 3 个改进模块和分类器.

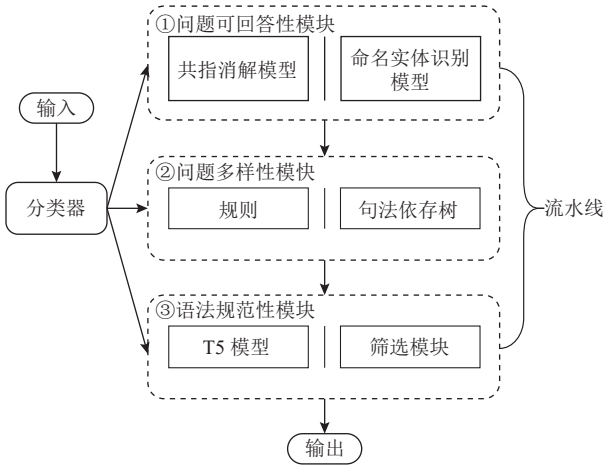


Fig. 2 Overall structure diagram of ADVICE model

图 2 ADVICE 模型整体结构图

2.1 问题可回答性模块

经过分析，本文发现主要有三大类情况导致问题难以回答，它们分别是代词指代模糊、含有冗余信息和疑问词有误. 图 3 中是 3 种情况的实例展示与解释说明.

情况 1： 代词指代模糊
实例：How many his very speech in the post he mistakenly referred to himself as the secretary of state for Ireland In? 说明：his, he和 himself 可能指代的是同一个人物，但是这个人物在这句话中没有出现，因而本文认为它们是指代模糊的.
情况 2：含有冗余信息
实例：Who announced that Iger had agreed to a one -year extension of his term as CEO through July 2? 说明：2 个与日期相关的实体在原文中并没有具体的说明，因而本文认为它们是冗余信息.
情况 3：疑问词有误
实例：Where was the president of the jury for the 17th Shanghai International Film Festival? （答案: Gong Li） 说明：从答案来看，“Gong Li”应该是人名，疑问词应该是“Who”，而不是“Where”.

Fig. 3 Examples of unanswerable questions

图 3 不可回答问题示例

本文将采用共指消解(coreference resolution, CR)技术和命名实体识别(named entity recognition, NER)技术来分别针对上述 3 种情况提出相应的改进方法.

在模型的选择方面，本文选择的共指消解模型是 fastcoref 模型，该模型与当前的 SOTA 模型表现相当，但是推理速度较快. NER 模型本文选择了 huggingface 上的 flair/ner-english-ontonotes 预训练模型，该模型的准确率较高且使用方便.

2.1.1 模糊代词替换

设上下文为 c，基于该上下文利用现有的问答模型生成了 N 个问题，记为 $\{q_i\}_{i=1}^{N}$. 每次选取一个问题进行模糊代词替换. 设第 k 次选择的问题为 q_k，将其与上下文连接后输入至共指消解模型中. 模型会输出若干个聚类 $S=\{s_l\}$，每个聚类中出现的所有单词指代同一个命名实体. 对于 q_k 中的某个代词 w，若存在 l，使得 $w\in s_l$ 且 s_l 中的其他命名实体均不在 q_k 中，就认为 w 是指代模糊的，应该被替换掉. 可以随机选择 s_l 中的任何一个命名实体对 w 进行替换，得到修改后的问题 q_k'.

2.1.2 冗余信息删除

设上下文为 c，基于该上下文利用现有的问答模型生成了 N 个问题，记为 $\{q_i\}_{i=1}^{N}$. 每次选取一个问题进行冗余信息删除. 设第 k 次选择的问题为 q_k，将其与上下文连接后输入至共指消解模型中. 模型会输出若干个聚类 $S=\{s_l\}$，每个聚类中出现的所有单词指代同一个命名实体. 同时利用命名实体识别技术找到 q_k 中的所有实体 $E=\{e_t\}$，对于其中的某个实体 e，若不存在 l，使得 $e\in s_l$，就认为实体 e 是冗余的，应该删去，得到修改后的问题 q_k'.

这里对 $e\in s_l$ 的判断实际上是一个模糊匹配问题. e 是由 NER 模型识别得到的，而 s_l 是由 CR 模型得到的，即使指代的是同一实体，e 和 s_l 中单词的表述也未必完全相同，因此需要进行匹配.

设当前问题中识别到实体 e 的开始与结束下标分别是 e_1 和 e_2，聚类 s_l 中对实体的某一次提及 m 的开始与结束下标分别是 m_1 和 m_2. 定义函数 $f(\cdot,\cdot)$ 表示 e 和 m 的匹配情况（$f(\cdot,\cdot)$=1 表示匹配，$f(\cdot,\cdot)$=0 表示不匹配），其形式化表示为：

$$f(e,m)=\begin{cases}1, \text{ if } m_1\leqslant e_1\leqslant m_2 \text{ or } m_1\leqslant e_2\leqslant m_2.\\ 0, \text{ 其他.}\end{cases} \tag{1}$$

2.1.3 疑问词修正

利用 NER 模型对现有答案涉及的上下文中的句子进行命名实体识别，将识别到的实体类型与问题中的疑问词类型进行比对，如果不符合，则按照表 1 中的映射关系进行修正. 如果实体类型对应的疑问

Table 1 Entity Type-Query Word Mapping Table

表 1 实体类型-疑问词映射表

实体类型	疑问词
PERSON，NORP，ORG	who
GPE，LOC，FAC	where
PRODUCT，EVENT，WORK_OF_ART，LAW，LANGUAGE	what
TIME，DATE	when
PERCENT，MONEY，QUANTITY，ORDINAL，CARDINAL	how many/how much

词为“how many”或“how much”，则随机选择其中之一作为修正后的疑问词.

2.2 问题多样性模块

本节针对不同疑问词提问设计了 3 种策略来提升问题的多样性.

1）以 when 或 where 提问的问题

受到 Sun 等人[15]的启发，本文注意到了对时间和地点提问时人类会倾向于用更具体的提问方式（例如图 4 中的例子），而不是单纯地使用“when”或者“where”进行提问.

模型生成的问题	**When** did Beijing host the Olympics?
人类提出的问题	**In what year** did Beijing host the Olympics?
答案	2008

Fig. 4 Example of question diversity

图 4 问题多样性示例

本文注意到现有无监督问题生成的方法在生成时间、地点相关的问题时，在疑问词后常常会出现一个表示时间或地点的介词. 本文利用这一规律进行规则的设计.

设原问题 $q=A+B+C$，其中 A 是“when”或“where”，B 是一个介词，C 是问题中剩余的所有单词；原答案为 a. 本文通过式（2）（3）对原问答对进行修改（设修改后的问题为 q'，答案为 a'）.

$$q' = \begin{cases} B + \text{what} + [type] + C, & \text{if } p < 0.5, \\ A + C, & \text{if } p \geqslant 0.5, \end{cases} \tag{2}$$

$$a' = \begin{cases} a, & \text{if } p < 0.5, \\ B + a, & \text{if } p \geqslant 0.5, \end{cases} \tag{3}$$

其中 [*type*] 表示答案的类型（例如：如果答案是年份，那么其类型就是“year”），随机变量 $p \sim U[0,1]$.

图 5 是一个修改的示例，展示了 p 取不同值时的修改效果.

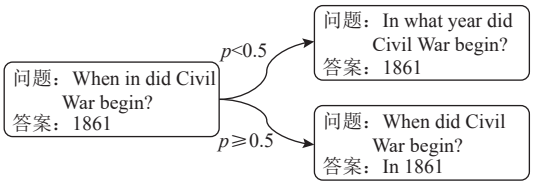


Fig. 5 Modification example

图 5 修改示例

2）以 why 提问的问题

现有的无监督问题生成方法难以生成用“why”来提问的问题. 现有方法通常先选择一个命名实体作为答案，然后生成相应的问题，但“why”往往提问的不是命名实体. 本文调研了 SQuAD 数据集，发现其中有一定比例的用“why”提问的问题. 因此，本文希望利用无监督的方法生成用“why”提问的问题，以增强问题多样性.

本文参考了 Stasaski 等人[16]的方法，设计了 120 个模板，用以提取段落中的因果关系. 本文删去了原有模板中因果关系不强的模板，保留的部分模板如表 2 所示（其中 R 表示“原因”，C 表示“结果”，其他大写字母对应于成分句法分析中的标签）. 最后，本文利用抽取的因果关系生成相应的问题与答案.

Table 2 Examples of Causal Relationship Extraction

表 2 因果关系抽取示例

模板	具体句子
&*C*(,/;/. /--)(&AND) therefore (,)&*R*	He never lies; **therefore**, everyone likes to make friends with him.
&*R* (,)on account of &NP@*C*@	I prefer to call off the meeting **on account of** boss's absence.
&*R*(,)owing to &NP @*C*@	She had a headache, **owing to** the weather here.

对于抽取出的 R 和 C，本文通过一种很简单的方式生成问题 q=*why*+C，答案 a=R. 这样的问题在语法上很有可能是错误的，但考虑到后续可以通过语法规范性模块修正语法错误，并且在这一部分本文主要关注的是问题多样性，因而在这里选择了这种较为简单的问题生成方式.

如果原文中能够抽取因果关系，本文会以大约 10% 的概率替换掉原有的问题（为了保证总的问题数量一致）. 在问题的替换策略方面，优先选择长度较短的问题进行替换.

3）以其他方式提问的问题

在用其他疑问词提问的问题中，本文暂时没有找到一个统一的方法来提升问题类型的多样性. 在这部分中本文主要利用句法分析树来提升答案类型的多样性.

通过句法依存树，可以递归地找到句子中的所有 *NP*（名词短语）和 *ADJP*（形容词短语），具体的算法如算法 1 所示. 得到所需类型的短语后用以下的 3 个条件进行过滤：①包含原答案；②长度不大于 10 个单词；③长度不超过原有答案的 2 倍. ②③这 2 个条件的设计主要是为了保证拓展后的答案尽量不包含太多的冗余信息（例如包含了整个从句），并且尽量和原答案相关. 在使用上述 3 个条件进行过滤后，只要有符合要求的候选答案，本文会以 40% 的概率对原答案进行替换.

算法 1. 提取指定类型的短语算法.

输入：句法分析树中的节点 *node*，短语类型 *type*；

输出: 指定类型的短语列表 *spans*.

① function *Search*(*node*,*type*)
② if *node.label* == ϕ then
③ return ϕ;
④ end if
⑤ *spans*← ϕ;
⑥ for all *child* ∈ *node.children* do
⑦ *spans*←*spans*∪{*Search*(*child*,*type*)}; /*在当前节点的子节点中进行递归查找*/
⑧ end for
⑨ if *node.label* == *type* then
⑩ *spans*←*spans*∪{*str*(*node*)}; /*若当前节点的句法标签与 *type* 相符, 就将其对应的短语加入 *spans**/
⑪ end if
⑫ return *spans*;
⑬ end function

2.3 语法规范性模块

文本语法错误纠正(grammatical error correction, GEC)是自然语言处理领域中的一个重要任务, GEC 任务要求模型能够检测句子中是否有语法错误, 并将检测出的语法错误进行自动纠正. 表 3 是 GEC 任务的一个示例.

Table 3 Example of GEC Task

表 3 语法错误纠正任务样例

输入(有语法错误的)	输出(纠正后的)
She see Tom is catched by policeman in park at last night.	She saw Tom caught by a policeman in the park last night.

现有的 GEC 模型往往针对陈述句, 训练语料也往往以陈述句为主. 考虑到问题生成任务的特殊性, 本文训练了一个针对问句的语法错误纠正模型.

2.3.1 数据集的构建

本文首先构建一个针对问句的 GEC 任务数据集. 构建数据集的思路是: 首先选择一个包含问句的数据集, 然后通过无监督的方法对这些问句进行随机扰动, 生成含有语法错误的问题. 在训练模型时, 将修改后的问句作为模型的输入, 要求模型恢复出原来的问句.

本文选取 Natural Questions[17] 作为数据的主要来源. Natural Questions 来源于 Google 搜索引擎的真实问题, 其数据量充足、类型多样, 非常适合本任务. 在对问句进行随机扰动时, 本文参考了 PIE 模型[18] 中的随机扰动方式(主要有插入、删除、替换、词性变换等). 除此之外, 为了和实际应用场景保持一致, 本文删去了原数据集中不以疑问词开头的问题, 最终保留了 221 209 对问题.

2.3.2 模型的设计

本文提出的 GEC 模型由 2 部分组成: 预训练模型 T5[19] 和筛选模块. T5 模型的结构图如图 6 所示.

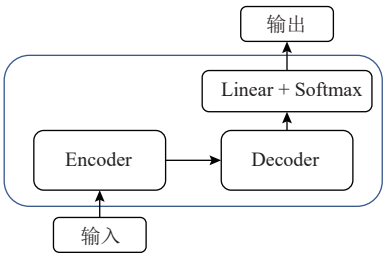


Fig. 6 T5 model structure diagram

图 6 T5 模型结构图

设 GEC 任务的输入为 $x=(x_1, x_2, \cdots, x_n)$, 输出 $y=(y_1, y_2, \cdots, y_m)$. 设在第 t 步解码时, 经过 T5 的 Encoder 和 Decoder 后的隐藏层向量为 $\boldsymbol{z}_t \in \mathbb{R}^d$, d 是隐藏层向量的维数. $\boldsymbol{z}_t$ 可由式(4)计算得到.

$$\boldsymbol{z}_t = Decoder(Encoder(x), y_{<t}). \tag{4}$$

经过线性层得到向量 $\tilde{\boldsymbol{z}}_t \in \mathbb{R}^{|V|}$, $|V|$是单词表的维数. $\tilde{\boldsymbol{z}}_t$可由式(5)计算得到.

$$\tilde{\boldsymbol{z}}_t = Linear(\boldsymbol{z}_t). \tag{5}$$

进而可以通过 Softmax 层将 $\tilde{\boldsymbol{z}}_t$转化成在 $\mathbb{R}^{|V|}$上的概率分布, 如式(6)所示. 式中上标 i 表示该向量第 i 个位置元素的值.

$$p_t^i = \frac{\exp(\tilde{z}_t^i)}{\sum_{k=1}^{|V|}\exp(\tilde{z}_t^k)}. \tag{6}$$

T5 模型优化的损失函数如式(7)所示. 其中 $f(y_t)$ 表示 y_t 在词表中对应的下标.

$$L = -\sum_{t=1}^{m}\log p_t^{f(y_t)}. \tag{7}$$

如果设 T5 模型的参数为 θ, 则式(7)可以写成式(8)的形式.

$$L(\theta) = -\sum_{t=1}^{m}\log P(y_t|y_{<t}, x, \theta). \tag{8}$$

在训练时需要通过调整模型参数 θ 来最小化损失函数 $L(\theta)$.

训练完成后, 在使用 T5 模型进行推理的过程中, 本文发现模型输出的结果存在 2 个问题: 1)修改后语法错误数量并没有减少, 在少数情况下, 语法错误数量甚至有可能增加; 2)修改后问题的语义发生了较大的变化. 于是, 在 T5 模型的基础上, 本文设计

了一个筛选模块来对模型的输出进行筛选与过滤，只保留符合要求的修改结果.

设 T5 模型的输入为 $x=(x_1, x_2, \cdots, x_n)$，输出为 $y=(y_1, y_2, \cdots, y_m)$. 设过滤后的结果为 $\tilde{y}$，$\tilde{y}$ 可由式(9)、式(10)和式(11)计算得到.

$$\tilde{y}=\begin{cases} y, \text{ if } g_1(x,y)\times g_2(x,y)=1, \\ x, \text{ if } g_1(x,y)\times g_2(x,y)=0, \end{cases} \tag{9}$$

$$g_1(x,y)=\begin{cases} 1, \text{ if } err(x)>err(y), \\ 0, \text{ if } err(x)\leqslant err(y), \end{cases} \tag{10}$$

$$g_2(x,y)=\begin{cases} 1, \text{ if } sim(x,y)\geqslant 0.85, \\ 0, \text{ if } sim(x,y)<0.85, \end{cases} \tag{11}$$

其中：函数 $err(\cdot)$ 返回的是当前句子中语法错误的数量，本文使用 language_tool_python 工具进行统计；函数 $sim(\cdot,\cdot)$ 返回的是 2 个句子间的语义相似度，本文首先使用 Sentence Transformers 计算 2 个句子的嵌入表示，进而计算它们的余弦相似度作为相似度度量. 本文分别尝试了 0.8, 0.85, 0.9 这 3 个阈值，实验结果表明 0.85 的实验效果最好，因而本文选择采用 0.85 作为语义相似度的阈值.

2.4 分类器

分类器从可回答性、多样性和语法 3 个维度进行判断，并根据结果选择合适的模块.

1)从可回答性的角度. 基于 RoBERTa[20] 模型对问题的可回答性进行判断. 用该模型在 SQuADv2.0[21] 上进行了训练，对“no answer”问题识别的准确率达到了 81.8%. 如果判断问题为“不可回答”，则使用问题可回答性模块，否则不使用该模块.

2)从多样性角度. 可以人为设置使用多样性模块的概率，默认概率为 1.0，即一定使用多样性模块.

3)从语法角度. 基于 DistilBERT[22] 模型对问题是否含有语法错误进行识别. 在语法规范性模块中对自建的数据集进行训练，训练后模型识别语法错误的准确率达到了 98.0%. 如果判断问题“包含语法错误”，则使用语法规范性模块，否则不使用该模块.

3 实验结果与分析

参考现有的关于无监督问题生成的文章，本文首先使用下游问答模型的性能来对问题生成方法进行评价. 除此之外，本文会对生成问题的质量进行评价.

3.1 数据集

RefQA[6] 和 ClozeQA[5] 是无监督问题生成中 2 个比较有代表性的方法，在后续的实验中主要选用这 2 种方法生成的问答数据进行实验. 由于原数据量较大，本文在实验时对其进行了采样，分别保留了 5 万条数据进行实验.

在使用改进方法对问答数据进行修改后，本文会使用修改后的数据对下游的问答模型进行训练. 在对下游的问答模型进行评估时，本文选择的是 SQuAD 的验证集. SQuAD 是问答领域最为权威的数据集之一，数据集里面的文章来自于多个不同领域，问答对质量较高，可以较好地衡量问答模型的性能.

3.2 评价指标

参考现有的无监督问题生成方法，本文采用了精准匹配度(exact match, *EM*)和 *F*1 值这 2 个指标来评价问答模型的性能，从而间接评价无监督问题生成方法的好坏. *EM* 值的计算过程为：

$$EM=\frac{\text{完全预测正确的答案个数}}{\text{全部答案个数}}. \tag{12}$$

*F*1 值的计算过程如式(13)~(15)所示：

$$F1=2\times\frac{P\times R}{P+R}, \tag{13}$$

$$P=\frac{\text{预测答案与真实答案重合的长度}}{\text{预测答案的长度}}, \tag{14}$$

$$R=\frac{\text{预测答案与真实答案重合的长度}}{\text{真实答案的长度}}. \tag{15}$$

3.3 实验设置

在问答模型方面，分别选用 BERT-large[7], XLNet-large[23] 和 ALBERT-large[24] 这 3 个预训练模型进行实验，在实验过程中采用模型的默认参数. 这 3 个模型是典型且常用的预训练模型，具有一定代表性.

在使用不同问答数据进行实验时，上述 3 个问答模型的参数均保持不变，批处理大小固定为 32，BERT 的训练固定为 10 轮，XLNet 和 ALBERT 的训练固定为 5 轮.

在对语法规范性进行改进时，本文使用构建的数据集训练了一个 T5-base 模型，学习率为 2E−5，批处理大小为 32，一共训练了 10 轮.

3.4 实验结果及分析

3.4.1 面向下游问答模型性能的实验与分析

分别记问题可回答性模块、问题多样性模块、语法规范性模块为“ans”“div”“gram”，用“all”表示使用全部 3 个模块，记分类器为“clf”. 实验结果如表 4 所示.

从模型的角度来看，对 BERT 模型，gram 对模型性能的提升最为有效，而 ans 甚至有可能起到负面的效果. XLNet 和 ALBERT 模型的趋势较为相似，div 对模型的性能提升效果明显.

Table 4 Results Comparison of Each Models from Different Data Sources

表 4 不同数据来源下各模型的结果比较 %

模型	RefQA(*EM*/*F*1)	ClozeQA(*EM*/*F*1)
BERT	48.3/60.0	48.3/57.6
BERT + ans	52.0/62.2	46.1/55.0
BERT + div	50.3/63.9	48.3/58.5
BERT + gram	52.9/62.5	49.1/58.6
BERT + all w/o clf	50.7/62.6	46.7/57.3
BERT + all + clf	**53.0/64.9**	**49.5/59.0**
XLNet	46.0/56.4	42.4/50.5
XLNet + ans	49.1/59.8	42.7/51.6
XLNet + div	**49.6/63.0**	44.6/54.1
XLNet + gram	47.7/57.2	43.1/51.9
XLNet + all w/o clf	48.9/61.6	44.6/54.3
XLNet + all + clf	49.2/62.8	**45.5/54.8**
ALBERT	35.8/45.9	38.4/46.7
ALBERT + ans	37.1/47.5	39.2/47.7
ALBERT + div	38.1/51.0	39.6/48.9
ALBERT + gram	38.5/47.7	38.8/47.3
ALBERT + all w/o clf	37.8/50.1	39.8/49.1
ALBERT + all + clf	**39.5/51.8**	**40.3/49.9**

注：黑体数值表示最优结果.

从数据来源的角度来看，在 RefQA 生成的问题上，使用 3 种改进方法后，*EM* 值平均提升 2.8%，*F*1 值平均提升 3.1%；在 ClozeQA 生成的问题上，使用 3 种改进方法后，*EM* 值平均提升 0.5 个百分点，*F*1 值平均提升 1.0 个百分点. 可以看出，本文提出的 3 种改进方法对 RefQA 生成的问题更加有效.

从全局的角度来看，使用完整的流水线式模型在不同数据来源和不同模型上进行实验，问答模型在 SQuAD 数据集上的 *EM* 值平均提升 2.9 个百分点，最高提升 4.7 个百分点；*F*1 值平均提升 4.4 个百分点，最高提升 6.4 个百分点. 对比有无 clf 的结果可以发现，clf 对问答模型的性能总是起到正面的效果.

除此之外，本文使用完整的 ADVICE 模型对 SQuAD 训练集进行优化，以测试本文方法对现有高质量问答数据集的有效性. 实验结果如表 5 所示.

表 5 表明，使用经过优化的数据进行训练后，3 种问答模型的 *EM* 值没有显著改善，但 *F*1 值均有一定提升. 这是由于 SQuAD 数据集质量较高，优化改善了原有问答数据多样性等问题，但另一方面也引入了一定的噪声. 不过，对 *F*1 值的提升仍然反映出 ADVICE 模型具有一定的实用价值.

Table 5 Results Comparison of of Each Model Before and After Optimization on SQuAD Dataset

表 5 经过优化前后各模型在 SQuAD 数据集的结果比较 %

模型	*EM*	*F*1
BERT	86.0	92.6
BERT + ADVICE	85.9	92.7
XLNet	85.3	92.4
XLNet + ADVICE	85.4	92.7
ALBERT	83.1	90.2
ALBERT + ADVICE	83.1	90.4

在 ans 模块中，本文分别设计了模糊代词替换、冗余信息删除和疑问词修正 3 个子模块提升问题的可回答性，现探究这 3 个模块在不同数据来源、不同模型的条件下起到的作用. 实验结果如表 6 所示. 由于 ClozeQA 生成的问题中不存在可以用冗余信息删除的情况，因而本文没有对这一情况进行实验.

Table 6 Ablation Experiment of Each Module for Question Answerability

表 6 问题可回答性各模块消融实验 %

模型	RefQA(*EM*/*F*1)	ClozeQA(*EM*/*F*1)
BERT*	52.0/62.2	46.1/55.0
BERT* w/o part1	49.8/60.6	44.7/53.8
BERT* w/o part2	51.8/62.3	
BERT* w/o part3	48.7/59.0	45.3/54.0
XLNet*	49.1/59.8	42.7/51.6
XLNet* w/o part1	46.4/56.9	41.1/49.8
XLNet* w/o part2	47.1/58.1	
XLNet* w/o part3	47.2/58.2	43.5/51.6
ALBERT*	37.1/47.5	39.2/47.7
ALBERT* w/o part1	39.0/49.3	38.4/46.8
ALBERT* w/o part2	35.3/45.0	
ALBERT* w/o part3	35.1/44.3	39.6/47.8

注：模糊代词替换、冗余信息删除和疑问词修正分别记作 part1、part2、part3；模型右上角的“*”表示使用全部 3 个模块.

总的来看，对于不同模型以及不同来源的问答数据，这 3 个模块起到的作用与效果不尽相同. 在大部分情况下这些模块对模型性能的提升都能起到正面的效果，但在少数情况下，部分模块的加入没有起到作用或起到负面作用.

3.4.2 面向问题质量的实验与分析

本节将分别从可回答性和语法规范性 2 个角度衡量问题的质量.

本文参考了 Bartolo 等人[25] 的论文，给出了问题

可回答性的定义. 一个可回答的问题必须满足 3 点要求：1)问题可以用原文中的若干连续的单词来回答；2)问题必须有一个明确的答案；3)只要给予足够的时间，人类应当能够正确回答问题.

本文随机抽取了 200 个 RefQA 方法生成的问题进行了人工评价，参照上文提到的标准比较这些问题的可回答性在经过问题可回答性模块前后是否有所变化. 结果如表 7 所示.

Table 7 Comparative Analysis of Question Answerability Before and After Modification

表 7 修改前后问题可回答性对比分析 %

类别	不可回答				可回答
	代词指代模糊	含有冗余信息	疑问词有误	其他	
修改前	8.5	14	7	16.5	54
修改后	1	3.5	3	20	72.5

表 7 中，对不可回答的问题按照不可回答的原因进行了分类. 经过模糊代词替换、冗余信息删除和疑问词修正这 3 个模块的修改后，问题的可回答性有明显的提升(从 54% 提升至 72.5%). 这一提升主要是因为前 3 种原因出现的比例有了明显的下降，原来因为这 3 种原因造成问题不可回答的比例大约为 30%，修改后降低至 10% 左右. 可以发现，“其他”的比例有略微的上升，这是因为在对前 3 种情况进行修改时，有时并不能完全解决不可回答的问题. 虽然这 3 种情况得到了解决，但问题本身依然是不可回答的，这类问题也会被归类至“其他”类别. 总的来看，问题的可回答性在修改后得到了显著的提高.

除此之外，本文对问题的语法规范性进行了探究. 在语法规范性模块中，本文会对语法错误数量增加和原有语义发生较大变化这 2 种异常情况进行过滤，表 8 详细说明了 2 种异常情况的比例.

Table 8 Abnormal Situation Statistics

表 8 异常情况统计 %

问题生成方法	异常类型	
	语法错误数量增加	语义发生较大变化
RefQA	1.7	13.8
ClozeQA	3.8	5.7

本文使用 language_tool_python 工具对经过语法规范性模块修改前后问题的语法数量进行了统计(修改后的问题排除了上述 2 种异常情况)，结果如表 9 所示.

由表 9 可以发现，经过语法规范性模块的修改，

Table 9 Statistics on the Number of Grammatical Errors

表 9 语法错误数量统计

问题生成方法	每句语法错误数量	
	修改前	修改后
RefQA	1.64	1.01
ClozeQA	1.24	0.98

RefQA 和 ClozeQA 这 2 种方法生成的问题包含的语法错误数量都有了一定程度的下降，下降的比例分别达到了 38% 和 21%.

问题的语法规范性和可回答性对问题本身的质量有重要的影响. 通过上述实验与分析可知，现有无监督问题生成方法生成的问题在经过 ADVICE 模型的改进后质量有了一定的提升.

4 结论及展望

本文针对现有无监督问题生成方法存在的问题，从问题可回答性、问题多样性和语法规范性入手设计并提出了一个自适应的多模块流水线式模型 ADVICE. 实验证明，在现有方法的基础上使用本文提出的模型，可以有效提升下游问答模型的性能. 除此之外，本文的方法还可以在一定程度上改善生成问题的质量. 未来的工作将尝试从数据集、模型和参数选择等方面对语法规范性模块的性能进行改进，并尝试提出更加通用的方法提升问题的可回答性和多样性.

作者贡献声明：李昆泽提出了算法思路，设计并完成实验，撰写论文；张宇提出指导意见并修改论文.

参 考 文 献

[1] Wu Yunfang, Zhang Yangsen. A survey of question generation[J]. Journal of Chinese Information Processing, 2021, 35(7): 1−9 (in Chinese)
(吴云芳，张仰森. 问题生成研究综述[J]. 中文信息学报，2021，35(7)：1−9)

[2] Kurdi G, Leo J, Parsia B, et al. A systematic review of automatic question generation for educational purposes[J]. International Journal of Artificial Intelligence in Education, 2020, 30: 121−204

[3] Liu Tieyuan, Chen Wei, Chang Liang, et al. Research advances in the knowledge tracing based on deep learning[J]. Journal of Computer Research and Development, 2022, 59(1): 81−104 (in Chinese)
(刘铁园，陈威，常亮，等. 基于深度学习的知识追踪研究进展[J]. 计算机研究与发展，2022，59(1)：81−104)

[4] Dong Li, Yang Nan, Wang Wenhui, et al. Unified language model pre-training for natural language understanding and generation[J]. Advances in Neural Information Processing Systems, 2019, 32: 13063−13075

[5] Lewis P, Denoyer L, Riedel S. Unsupervised question answering by cloze translation[C]//Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4896−4910

[6] Li Zhongli, Wang Wenhui, Dong Li, et al. Harvesting and refining question-answer pairs for unsupervised QA[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6719−6728

[7] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: ACL, 2019: 4171−4186

[8] Rajpurkar P, Zhang Jian, Lopyrev K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]//Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 2383−2392

[9] Heilman M, Smith N A. Good question! statistical ranking for question generation[C]//Proc of the 2010 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2010: 609−617

[10] Fabbri A R, Ng P, Wang Zhiguo, et al. Template-based question generation from retrieved sentences for improved unsupervised question answering[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4508−4513

[11] Lyu Chenyang, Shang Lifeng, Graham Y, et al. Improving unsupervised question answering via summarization-informed question generation[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 4134−4148

[12] Lewis M, Liu Y, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 7871−7880

[13] Nie Yuxiang, Huang Heyan, Chi Zewen, et al. Unsupervised question answering via answer diversifying[C]//Proc of the 29th Int Conf on Computational Linguistics. New York: ACM, 2022: 1732−1742

[14] Trischler A, Wang Tong, Yuan Xingdi, et al. NewsQA: A machine comprehension dataset[C]//Proc of the 2nd Workshop on Representation Learning for NLP. Stroudsburg, PA: ACL, 2017: 191−200

[15] Sun Yuan, Liu Sisi, Dan Zhengcuo, et al. Question generation based on grammar knowledge and fine-grained classification[C]//Proc of the 29th Int Conf on Computational Linguistics. New York: ACM, 2022: 6457−6467

[16] Stasaski K, Rathod M, Tu T, et al. Automatically generating cause-and-effect questions from passages[C]//Proc of the 16th Workshop on Innovative Use of NLP for Building Educational Applications. Stroudsburg, PA: ACL, 2021: 158−170

[17] Kwiatkowski T, Palomaki J, Redfield O, et al. Natural questions: A benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 452−466

[18] Awasthi A, Sarawagi S, Goyal R, et al. Parallel iterative edit models for local sequence transduction[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 4260−4270

[19] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1−67

[20] Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach[J]. arXiv preprint, arXiv: 1907.11692, 2019

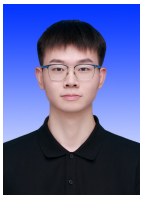
[21] Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: ACL, 2018: 784−789

[22] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter[J]. arXiv preprint, arXiv: 1910.01108, 2019

[23] Yang Zhilin, Dai Zihang, Yang Yiming, et al. XLNet: Generalized autoregressive pretraining for language understanding[C]//Proc of the 33rd Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 5753−5763

[24] Lan Zhenzhong, Chen Mingda, Goodman S, et al. ALBERT: A lite BERT for self-supervised learning of language representations[J]. arXiv preprint, arXiv: 1909.11942, 2019

[25] Bartolo M, Thrush T, Jia R, et al. Improving question answering model robustness with synthetic adversarial data generation[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 8830−8848



Li Kunze, born in 2000. Master candidate. Student member of CCF. His main research interests include natural language processing and question generation.

李昆泽，2000 年生. 硕士研究生. CCF 学生会员. 主要研究方向为自然语言处理、问题生成.



Zhang Yu, born in 1972. PhD, professor, PhD supervisor. Member of CCF. His main research interests include natural language processing, question answering system, and information retrieval.

张　宇，1972 年生. 博士，教授，博士生导师. CCF 会员. 主要研究方向为自然语言处理、问答系统、信息检索.