祝胜泉, 邵清. 问题生成应用与前沿研究[J]. 智能计算机与应用,2025,15(3):170-180. DOI:10.20169/j.issn.2095-2163.250325

问题生成应用与前沿研究

祝胜泉, 邵 清
(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: 问题生成是自然语言生成领域的子任务,该任务的目标是通过向机器提供一段文本,让机器生成与该文本相关的问题。问题生成有广泛的应用领域,如阅读理解、对话式问题生成、数据增强等,也因此引起了相关研究人员的关注和重视。本文首先对问题生成任务进行了简单的概述,包括其基本概念、研究意义,其次介绍了相关的数据集,进一步阐述了问题生成任务的技术发展路线,并对问题生成技术的应用方向进行了总结和分类,然后讨论了问题生成的一些前沿研究成果,最后介绍了问题生成的自动化评价指标,以及未来的研究方向和挑战。

关键词: 神经网络; 问题生成; 自然语言生成; 机器阅读理解; 评价指标
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0170-11

Application and frontier research of question generation

ZHU Shengquan, SHAO Qing
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Question generation is one of natural language generation tasks, which aims to generate questions by a given text. Question generation has a wide range of applications, such as reading comprehension, conversational question generation, data enhancement, etc., which has attracted the attention of relevant researchers. This paper firstly gives a brief overview of the question generation task, including its basic concept and research significance, introduces relevant datasets and expounds the technical route of the question generation task, nextly summarizes and classifies the application direction of the question generation technology, and then introduces some frontier researches of question generation. Finally, the paper introduces the evaluation metric of question generation, the future research direction and challenge.

Key words: neural network; question generation; natural language generation; machine reading comprehension; evaluation metric

0 引 言

问题生成[1]是自然语言生成领域下的一个任务,和机器阅读理解密切相关。其定义为:提供一段自然语言文本,生成相关问题,且生成问题的答案包含在提供的文本中。问题生成是检验计算机是否真正理解自然语言的途径之一,是自然语言处理中的重要研究方向。

在学术研究领域,问题生成可以为问答相关任务提供数据,和其他任务模型结合提高其性能,检验模型理解自然语言文本的能力等;在现实应用领域,问题生成可以为搜索引擎提供备选问题和联想问题,让聊天机器人拥有主动提问能力,在教育领域自动生成相关问题,检验相关人员的学习效果等。因其广泛的应用领域,问题生成的相关研究已逐渐引起了学术界的普遍关注。

问题生成和其他的自然语言处理任务类似,从一开始基于规则和模板的方法,发展到使用神经网络的方法,而后随着深度学习的流行,近年来的相关研究大多使用基于深度学习的方法。当前已有一些研究者对问题生成相关的工作进行了梳理,如文献[2]总结了基于深度神经网络的问题生成研究,文献[3]从生成问题类型和技术等方向进行了归纳。本文的贡献在于,以问题生成的应用方向作为分类依据,阐述了问题生成的应用方向的技术发展,并且介绍了部分前沿研究模型。

本文第 1 节介绍问题生成相关的常用数据集;第 2 节概述了问题生成技术的发展路线,从基于规则的方法到基于深度学习的方法;第 3 节对问题生

作者简介: 祝胜泉(1998—),男,硕士研究生,主要研究方向:自然语言处理,自然语言生成。Email:1515394408@qq.com。

收稿日期: 2023-09-03

成的应用领域进行分类，并概述相关成果；第 4 节阐释了问题生成的前沿技术和自动化评价指标；最后总结全文并对未来进行展望。

1　数据集

问题生成模型性能的不断提高，一方面是由于硬件的发展和算法的迭代，另一方面是因为相关数据集的不断完善。因此，本节首先介绍一些问题生成相关的数据集。按照数据的形式，可以将问题生成数据集分为有答案和无答案两种。前者的数据输入形式为文本和答案或答案标记，后者的数据输入形式为纯文本。

在有答案的问题生成研究领域中，备受学术界瞩目的是由斯坦福大学发布的 SQuAD1.1 数据集[4]，该数据集来源为维基百科，最初是用于问答任务，数据集中的每一条数据都是一个三元组（P，Q，A），其中 P 为文本，Q 为针对该文本提出的问题，A 为 Q 对应的答案，且答案在所给文本中可以找到，三者均为文本形式，如图 1 所示。 由于 SQuAD1.1 高质量的数据，现已应用到问题生成任务中。

P：The role of teacher is often formal and ongoing，carried out at a school or other place of formal education. In many countries，a person who wishes to become a teacher must first obtain specified professional qualifications or credentials from a university or college. These professional qualifications may include the study of pedagogy …

Q：What can a teacher use to help student learn?

A：lesson plan.

图 1　SQuAD1.1 数据集的一个数据例子

Fig. 1　Data example of SQuAD1.1

除此之外，有答案问题生成常用的数据集还有：

（1）MS MARCO[5]，来源于必应网络搜索日志，其数据量较大，包含 8 841 823 条文章，由人工进行提问和回答，其答案需要通过推理获得而不是直接从文本中抽取。

（2）RACE[6]，数据来源于中国学生的英语考试试题，包含 28 000 条文本和 10 000 条以上的问题，其特点是一篇文章对应多个答案。

（3）News QA[7]，微软提出的数据来源为新闻语料的数据集，包含 120 万个问答对，问题的难度较高。

（4）NarrativeQA[8]，答案不在单个的段落或者句子中，必须完整阅读上下文才能得到答案。

相对于有答案的问题生成，无答案问题生成任务更加困难，因此其数据集也相应较少。其中，较为常用的数据集有 LearningQ[9]，主要来源于 TED-ed，包含 23 万个文本问题对，分为困难问题和简单问题，但简单问题数目却只有 7 000，剩余的都是需要深入理解文本才能回答的高难度问题，这也是该数据集最大的特点。LearningQ 数据形式如图 2 所示。

Doc：… gases have energy that is proportional to the temperature. The higher the temperature，the higher the energy the gases have. The crazy thing is that at the same temperature，all gases have the same energy …

Q：If you were given oxygen（molecular mass = 18 AMU）and hydrogen（molecular mass = 1 AMU）at the same temperature and pressure，which has more energy?

图 2　LearningQ 数据集的一个例子

Fig. 2　Data example of LearningQ

SQuAD2.0 数据集[10]，同样是由斯坦福大学提出，在 SQuAD1.1 的基础上增加了 53 775 个无法回答的新问题，其中增加的新问题与原文本相关，并且在文本中包含貌似合理的答案，如图 3 所示。也有一些研究者基于此数据集，展开了对无法回答的问题生成的研究，这类研究相较于原本的生成任务更加具有挑战性。

Paragraph：“Other legislation followed，including the Migratory Bird Conservation Act of 1929，a 1937 treaty prohibiting the hunting of right and gray whales，and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society-the species were relatively rare-and little opposition was raised.”

Question 1：“Which laws faced significant opposition ?
Plausible Answer：later laws

Question 2：“What was the name of the 1937 treaty ?”
Plausible Answer：Bald Eagle Protection Act

图 3　SQuAD2.0 数据集的一个例子

Fig. 3　Data example of SQuAD2.0

上述的数据集在数据量、数据形式、答案形式等方面都有所区别。本次研究中，对比其差异，将相关信息进行统计，统计结果见表 1。

表 1　问题生成数据集的信息统计

Table 1　Statistics of question generation dataset

数据集	答案类型	数据源	问题数	难度
SQuAD1.1	文本跨度	维基百科	97 888	简单
MSMARCO	文本跨度	必应搜索	3 563 535	较难
RACE	答案可选	教育领域	72 547	简单
News QA	文本跨度	新闻语料	119 633	较难
NarrativeQA	人工生成	维基百科	46 765	简单
LearningQ	无答案	教育领域	231 470	困难
SQuAD2.0	无答案	维基百科	141 663	困难

2 问题生成的研究发展

2.1 基于规则的问题生成研究

传统的基于规则的方法依赖于手动设计的转换规则,按照设计好的转换规则将文本转换为相应的问题。这些规则大致上可以分为 3 类。

(1)基于模板的方法:使用从训练集中提取的模板生成问题。文献[11]使用模板,从特定类型文章中生成问题来引导阅读。文献[12]在模板中融入语义来进行生成。文献[13]提出了 Ruminator,将总结后的文本作为输入,使用规则和模板,短时间内生成大量简单问题。

(2)基于语法的方法:根据文本语法结构创建语法转换规则生成问题。文献[14]使用 predicate-argument 模式从文本生成问题。文献[15]用一个解析器和简单的转换规则实现多项选择题生成。文献[16]提出了一种过度生成然后排序的方法,首先通过语法转换生成多个问题,然后人工对其排序,选择质量最高的问题作为生成结果,这种方式相对更耗时,但生成的问题质量更高。

(3)基于语义的方法:对文本语义进行分析,从而生成问题。文献[17]提出完全基于语义分析的方法,未使用模板或者语法信息。文献[18]提出利用句子的语义和谓语结构来生成问题,进一步提高对文本信息的利用度。文献[19]在生成问题之前先对文本进行分类,从而确定问题类型,提高生成问题的可用度。

传统的基于规则的方法证明,利用计算机进行问题生成任务可以提高人类的工作效率,但是自然语言非常复杂,人工设计的规则不够完善,基于规则生成的问题较为相似,且规则泛用性很低,因此基于规则的方法慢慢淡出了研究者的视野。然而近几年,也有研究将基于规则的方法应用于特定领域或与神经网络相结合,取得了比主流的神经网络模型更好的效果。文献[20]将基于模板的方法和 Seq2Seq(Sequence to Sequence)模型[21]结合,生成了流畅度高且多样化的问题。Dhole 和文献[22]提出了 Syn-QG,这是一组语法规则,调用了 WordNet[23]、VerbNet[24]中的语义信息,该实验取得了很好的效果。

2.2 基于循环神经网络的问题生成研究

基于循环神经网络的问题生成模型结构相似,如图 4 所示。模型分为编码器部分和解码器部分,这 2 个部分都由循环神经网络组成,例如 LSTM 网络[25](Long-Short Term Memory),因此该类模型也被称为 RNNsearch[21]。在 RNN 模型中,编码器将输入的文本进行编码,解码器将编码后的文本进行解码从而生成问题。

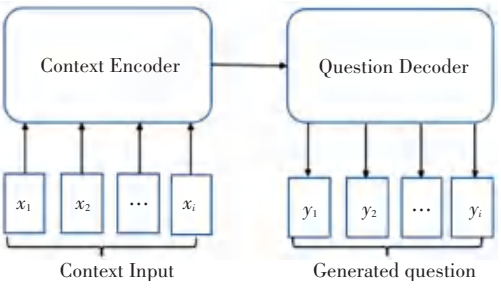

图 4 基于循环神经网络的问题生成模型

Fig. 4 Neural network-based question generation model

文献[26]向神经网络中输入答案的二进制位置信息以及多个上下文事实,生成了多样化的高质量问题。文献[27]设计了一种判别器,将答案和文本分别编码并关注其相关的部分,提高了生成问题与文本的相关度。文献[28]通过 2 个独立的 LSTM 对上下文输入和答案进行分析,最终通过将答案和文章进行匹配生成问题。文献[29]使用答案屏蔽令牌(mask)替换原始段落中的答案,将屏蔽的段落和答案分开编码,更好地利用文本和答案信息。

2.3 基于 Transformer 的问题生成研究

基于 Transformer 的问题生成模型结构与 RNN 模型结构相似。不同的是,前者的编码器和解码器皆由注意力机制实现。注意力机制[30]在 2014 年提出,解决了 RNN 模型编码器输入长度固定的问题。2017 年,Vaswani 等学者[31]提出了 Transformer 模型,该模型基于多头自注意力机制,在机器翻译领域模型表现非常出众,甚至全面超越了当时的 RNN 模型。因此也有不少研究者尝试将该模型应用在问题生成领域。一个基于 Transformer 的问题生成模型结构如图 5 所示。

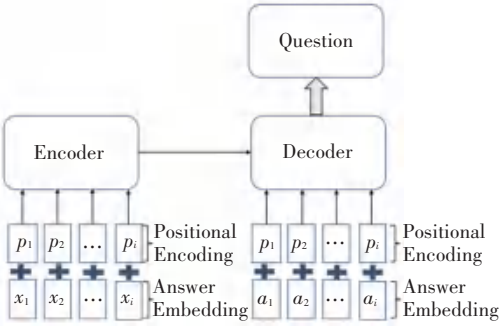

图 5 基于 Transformer 的问题生成模型

Fig. 5 Transformer-based question generation model

由于自注意力机制不关注文本中词语的位置信息，因此在将文本进行编码后，还要对其进行位置编码，再与文本向量相加作为编码器和解码器的输入。由于原始的 Transformer 模型在问题生成领域性能需进一步完善，因此一些研究者尝试对其进行改进。文献[32]在原始的 Transformer 模型中分别加入占位符、拷贝机制和语境嵌入，实验结果表明同时加入 3 种机制时模型达到了最高的性能，其中对模型性能提升帮助最大的是占位符机制。文献[33]使用经过预训练的 Transformer 模型建立了高鲁棒性的问题生成系统。文献[34]将输入文本通过神经网络转化为语义图，通过基于注意力的网络模型对语义图进行编码，向模型输入从而提升生成效果。

研究者们不断改进基于 Transformer 的问题生成方法，期望能够发挥出更好的性能，但研究证明原始的 Transformer 模型在问题生成领域表现一般。文献[35]提出了 GPT 模型，以 Transformer 模型的解码器为基础，进行训练，虽然性能表现没有超过当时表现最好的 RNN 模型，但为后来的预训练大模型提供了技术基础。

3　问题生成技术的应用

问题生成任务有多种应用方向。本节将对难度可控的问题生成、对话式问题生成、多任务问题生成等热门方向的相关工作进行介绍并总结。

3.1　难度可控的问题生成

问题生成模型根据输入的文本生成问题，但生成过程中容易出现提问方式简单、生成问题类型单一等问题。对此，文献[36]提出了难度可控的问题生成任务（Difficulty Controllable Question Generation，DQG），将问题分为简单和困难两种难度，根据问题是否能被 R-Net 模型[37]和 BiDAF 模型[38]同时回答正确作为区分标准，让模型根据需要生成特定难度的问题，如图 6 所示。

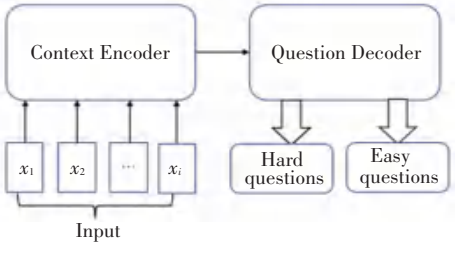

图 6　难度可控的问题生成模型

Fig. 6　Difficulty controllable question generation model

文献[39]结合知识图谱中的信息，将已生成的简单问题重写为复杂问题，为后来的 DQG 任务提供了一种新的方法。文献[40]提出了一个端到端模型完成 DQG 任务，与之前研究不同的是，该研究将问题的难度和知识图谱的跳数关联，问题包含知识图谱的相关信息。节点间跳数越多，问题的难度就越高。文献[41]也将跳数作为问题难度分级的依据，同时采用了文献[39]提出的重写方法，先对单个名词实体进行提问，然后根据需要的难度，结合额外的图节点信息对简单问题重写以此生成复杂问题。该研究发现生成的问题流畅度与节点跳数密切相关，当利用三跳的节点信息时，生成的问题可读性会变差，若使用三跳以上的节点信息会导致生成的问题难以理解。文献[42]利用 DQG 模型构建了一个系统。该系统不仅可以生成特定难度问题，而且可以预测学生在回答这些问题时的正确率，该工作展现了 DGQ 任务的一个实际应用场景。

3.2　对话式问题生成

问题生成的输入通常是一段长文本，可能包含答案。而现实生活中收集到的数据不全是长文本，也可能是对话式文本，对话式问题生成的目标就是从对话式文本中生成问题，如图 7 所示。

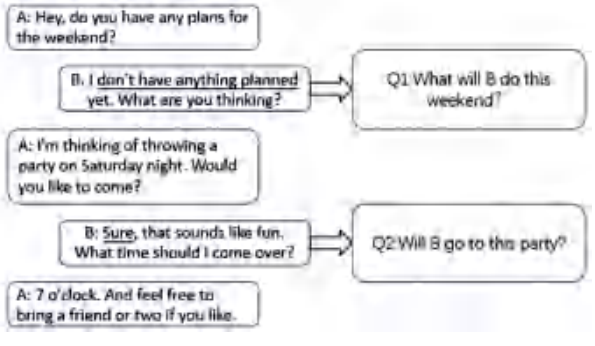

图 7　对话式问题生成

Fig. 7　Conversational question generation

文献[43]提出了 ChainCQG 模型，该模型同时在有答案的对话式问题生成以及无答案的对话式问题生成任务上都有较高的性能表现。文献[44]提出了 VDST 模型，基于 Guesswhat?! 数据集[45]，将图片处理模型与问题生成模型结合，实现了基于图片输入的对话式问题生成。文献[46]针对基于对话式文本的顺序问题生成（Sequential Question Generation，SGQ），提出了首个半自动的顺序问题生成模型，该模型可以把对话中的问题进行分类，再根据分类生成相关的问题。

文献[47]提出了一个分类模型，该分类模型与用户交互获得信息，再基于此进行分类。具体来说，

先向模型中输入需求,然后模型会根据需求生成问题与用户交互,最终根据交互信息确认分类结果。文献[48]从对话式问题简化任务(Conversational Question Simplification,CQS)中获得启发,将对话文本和对话中包含的问题作为输入,再结合上下文信息对输入问题进行重写,生成了高质量的问题。传统的工作会把对话式文本中的三元组,文本、问题、答案分为 2 个部分输入模型进行训练,分别为文本和问题、问题和答案,这种把信息分割的输入方式会使得生成问题与原文一致性不够,对此。文献[49]提出了 GTM 模型。该模型可以将文本,问题和答案作为三元组直接输入模型进行训练,生成连贯性和一致性更好的问题。

3.3 多任务问题生成

多任务问题生成是指将问题生成任务和其他任务结合进行研究,以此提高问题生成模型的性能,完成单个任务模型难以达到的目标。其中,最典型的就是将问题生成模型和问答模型结合,在问答模型训练的过程中,输入问题生成模型生成数据,从而提高问答模型的性能[50],如图 8 所示。

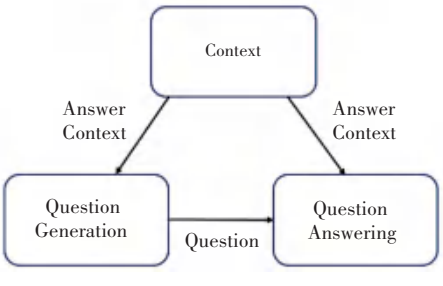

图 8 问题生成模型结合问答模型

Fig. 8 Question generation model combined question answering

文献[51]结合语言模型和问题生成模型,使用语言模型预测前后词汇,然后将语言模型隐藏层作为特征输入到问题生成模型实现生成任务。文献[52]将释义任务和问题生成任务结合,先扩展数据集中的句子和问题得到针对性释义,再结合问题生成模型生成质量更高的问题。此后,文献[53]尝试将预测模型和问题生成模型结合,先预测生成问题的第一个单词,再将预测单词用于后续生成。文献[47]将问题生成模型应用于分类任务,通过生成问题的类型信息提高分类模型准确率。文献[54]结合摘要任务和问题生成任务,通过对数据集做摘要,减少文本和问题之间的差异性,从而达到更好的训练效果。

文献[55]提出了 TMNs(Text Modular Networks)系统,可以将自然语言处理领域的复杂任务分解为多个子任务,输入文本后,TMNs 系统生成问题并让问答模型回答,通过答案将问题分解,再完成分解后的子任务。文献[56]将图像实体识别任务和问题生成任务结合,输入图片后,识别模型会识别图片中的实体,将其问题生成模型的输入,让问题生成模型对输入的图片进行提问。文献[57]提出了 RGX 系统,由实体识别模型、问题生成模型和问答模型组成,该系统识别输入文章的单词实体,通过单词实体生成问题,用于问答模型的训练,从而提高问答模型的性能。文献[58]提出了一个对文本的自动评价方法,首先生成关于该文本的问题,再通过问答模型回答问题的正确率,判断文本的质量,该方法做到了接近人类评价的效果。

3.4 特定领域的问题生成系统

研究者们不仅着眼于如何更好地生成问题,也着眼于如何将问题生成研究落实到特定领域建立可靠的问题生成系统。文献[59]提出的 EQG-RACE 模型用于考试类型问题生成(Examination-type Question Generation,EQG)任务,该模型训练数据来源于处理后的 RACE 数据集[6],生成的问题包含文本中多个句子的信息,答案的跨度较长,可以用于考试。文献[60]试着将问答系统应用于临床医疗领域,发现直接使用已有的模型效果很差,因此提出问题短语预测(Question Phrase Prediction,QPP)机制,将其与问题生成模型结合,生成丰富的问答对用于训练问答模型,取得了很好的效果。文献[61]提出了一个系统,可以模仿医生对病人提出问题。

此外,文献[62]提出反馈训练方法(Back Training)完成问题生成的无监督领域适应任务(Unsupervised Domain Adaptation,UDA),让问题生成模型通过训练后在其他领域生成问题,再将生成的正确问题作为训练数据进行训练,让模型迅速在新的领域完成生成任务。文献[63]通过同时向模型输入区分度大的积极数据和区分度低的消极数据并标记,解决了以多领域文本作为模型输入时生成的问题区分度很低的问题。文献[64]发现传统的问题生成任务生成问题的答案都在输入文本中,这些问题无法让人掌握文本以外的知识,因此提出好奇驱动(Curious Driven)的问题生成任务,目标是生成答案在输入文本之外的问题。文献[65]从小规模的原始数据中输入多跳问题。表 2 罗列了上述领域问题生成模型的性能数据。

表 2　不同领域问题生成模型性能

Table 2　Different domain question generation model

工作	*BLEU* - 1	*BLEU* - 4	*ROUGE* - *L*	*METEOR*
文献[55]	35.10	11.96	34.24	14.94
文献[58]	41.42	17.29	41.58	23.38
文献[59]	28.90	26.20	59.10	32.30
文献[60]	N	19.07	39.41	19.16

3.5　其他领域的问题生成应用

问题生成还有很多其他的应用领域：数据增强，用生成的问题训练问答模型，减少数据标注的成本[66-67]；阐明性问题生成，向用户提出阐明性问题，从而更加了解用户的意图，常在搜索引擎，推荐系统中使用[68-70]；深层问题生成旨在生成多跳问题、数学问题、开放式问题等难度较高的问题，模型经常有较高的复杂度[71-73]；视觉问题生成，利用图像或视频等能被人眼感知的数据进行问题生成任务[74-75,54]；干扰项生成，为生成的多选项问题生成额外的错误选项，更好地检验机器阅读理解能力[76-78]；跨语言问题生成，目标通常是建立跨语言的问题生成模型，利用大语种拥有的大量语料库作为训练数据来源，训练模型生成语料数据相对较少的小语种问题[79-80]。

多样的应用领域，证明了问题生成的研究价值，也期望后来的研究者能提出更多问题生成应用相关的工作。

4　问题生成的前沿技术及评价指标

4.1　问题生成前沿技术及模型

4.1.1　DIFFUSEQ 模型

Gong 等学者[81]于 2022 年提出了 DIFFUSEQ 模型，该模型借鉴了扩散模型[82]，扩散模型在正向过程中，将输入的数据点 $x_0 \sim q(x)$ 逐步污染为标准高斯噪声 $x_T \sim N(0,I)$，反向过程则是逐步从 x_T 采样重建 x_0，用这种方式让训练模型学习，从而提升性能。之前的研究人员大多将扩散模型用于视觉和音频领域，由于文本的离散性，将扩散模型用于文本领域时通常需要大量的分类器，模型的规模会变得很大。而 DIFFSEQ 模型没有分类器结构，减小模型规模的同时提升了模型性能。DIFFSEQ 模型使用的扩散过程如图 9 所示[81]。

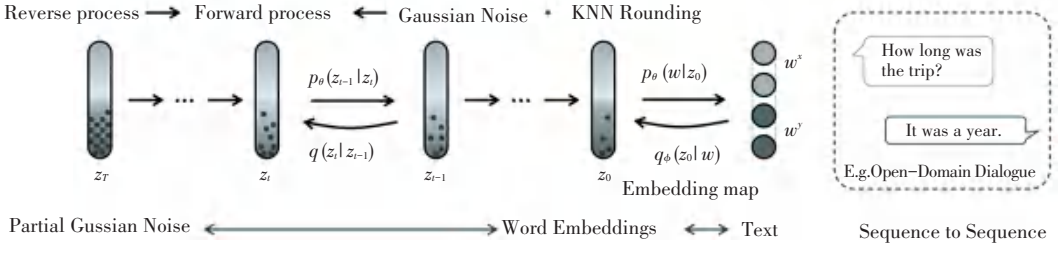

图 9　噪声扩散过程[81]

Fig. 9　Noise diffusion process[81]

4.1.2　T5 模型

Raffel 等学者[83]于 2020 年提出了 T5(Transfer Text-to-Text Transformer)模型，该模型以 Transformer 模型为基础，采取了和 Transformer 模型相同的编码器-解码器架构。与 Transformer 模型不同的是，T5 模型由块和子层组成，编码器和解码器层视为块，子层包含自注意力层和前馈网络，通过组装块和子层来建立模型，其次将 Transformer 模型中的 sin/cos 位置编码函数简化为矩阵相加，每层都计算一次位置编码，让模型对位置信息更加敏感，其模型结构如图 10 所示[83]。

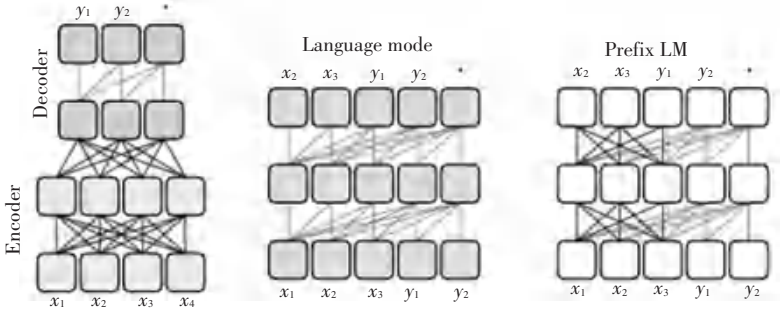

图 10　T5 模型结构图[83]

Fig. 10　Structure of T5 model[83]

4.1.3 Bang 模型

Qi 等学者[84]在 2021 年提出 BANG(Bridge the gap between Autoregressive (AR) and Non-autoregressive (NAR) Generation)模型,该模型降低了自回归生成和无自回归生成之间的性能差距。自回归生成顺序生成令牌,因此执行生成任务会消耗大量的时间,不能用于实时任务;无自回归通常是并行的生成令牌,提高生成任务的效率,但是效果较差。BANG 模型支持使用相同的预训练模型完成自回归、半自回归和非自回归生成,并且显著提高半自回归与非自回归生成的效果,让其达到接近自回归生成的水平。图 11 和图 12 为 BANG 模型的无自回归(NAR)和半自回归生成(semi-NAR)的结构[84]。

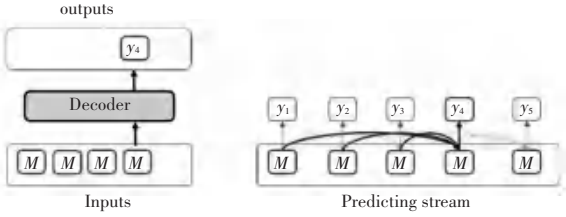

图 11 无回归自动生成[84]

Fig. 11 Non-auto regressive generation[84]

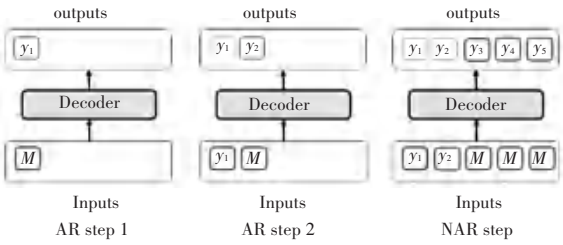

图 12 半自回归生成[84]

Fig. 12 Semi-auto regressive generation[84]

4.1.4 ACS-QG 模型

Liu 等学者[85]提出了 ACS-QG 模型,面向文本中的答案、线索和风格进行提问。ACS-QG 模型拥有一个信息提取器和数据过滤器,放弃对文本进行随机采样,因为随机采样会生成大量无意义的信息组合。信息提取器仅对文本中有意义的输入进行采样,对输入进行 3 次采样,组成(答案,线索,风格)三元组。数据过滤器会对原句和问答对之间的关系进行分类,从而删除低质量的问答对。通过在 SQuAD 数据集上进行训练,ASC-QG 模型自动生成了 2 个大型数据集,分别包含 133 万和 145 个问答对,并且生成的问答对质量都很高。ACS-QG 模型的结构如图 13 所示[85]。

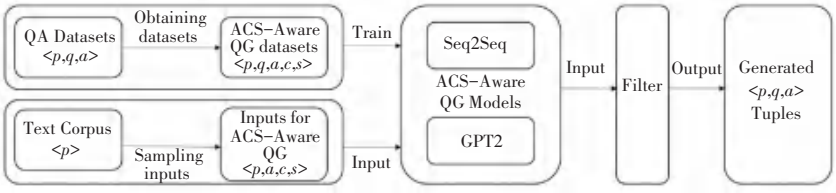

图 13 ACS-QG 模型结构[85]

Fig. 13 Structure of ACS-QG model[85]

4.1.5 VQAG 模型

Shinoda 等学者[86]以变分自动编码器(VAE)模型为基础,提出了 VQAG 模型。该模型采用条件自分编码器框架(Conditional-VAE),在模型中引入了 2 个独立的潜在随机变量,从而提高问题生成的多样性。此外,VQAG 模型还引入了额外超参数来解决后验崩溃问题,防止模型忽略潜在变量生成几乎相同的问题。VQAG 模型的结构如图 14 所示[86]。

4.2 问题生成的评价指标

问题生成任务是自然语言处理领域中的生成任务,在评价指标上也和其他自然语言生成任务类似,如文本生成、机器翻译等,分为人工评价和自动化评价。在早期问题生成任务都是通过人工进行评价,人工评价更注重文本的流利度、相关度、可回答性等属性,一般用作对自动化评价的补充,现在则主要使用自动化的评价指标或人工评价结合自动化评价的方法。

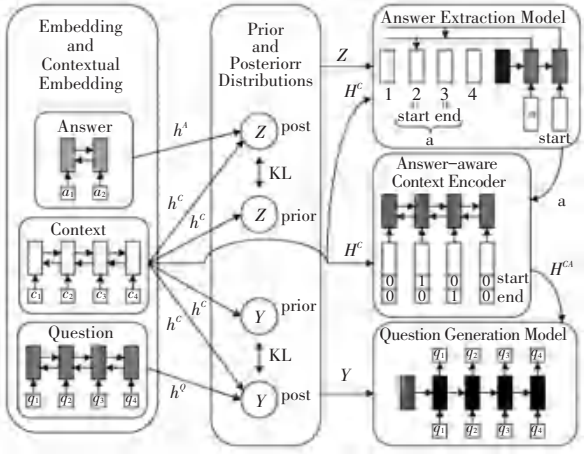

图 14 VQAG 模型结构[86]

Fig. 14 Structure of VQAG model[86]

目前生成任务主流的自动化评价指标有 BLEU[87]、METEOR[88] 和 ROUGE[89]，其中 BLEU 是机器翻译任务的主要评价指标，根据预测序列和真实序列之间单词重叠的个数分为 BLEU-1，BLEU-2，BLEU-3，BLEU-4。具体而言，其中又以 BLEU-4 为主要的评价指标。METEOR 会考虑基于整个语料库上的准确率和召回率，在计算相似度时会考虑句子流畅性、同义词和词根联系。BLEU 和 METEOR 对于生成语料的长度较为敏感，对散碎的句子会有惩罚。ROUGE 常用于文本摘要任务，主要计算召回率，即生成的文本与原文本之间重叠单元的比率。表 2 为 4.1 节中的模型性能指数。

Nema 等学者[90] 于 2018 年提出了一种专门用于问题生成任务的评价指标，该指标着眼于生成问题的可回答性，从生成问题的内容、单词实体、问题类型和功能四个角度进行评价，再结合原本的 BLEU 值，提出 Q-BLEU 指标，之后的一些工作也采用了 Q-BLEU 指标。也有指标着重评价生成问题的多样性[91]，但相关的工作需进一步验证。

5　问题生成未来研究方向

针对传统的问题生成任务，模型已经在特定场景下可以接近人类的水平。因此未来的问题生成研究任务可能会针对复合任务或者更多样的输入。

5.1　大型复合模型

大型复合模型具有大规模的参数和较长的训练时间，能完成多种任务并且在多个任务上都能取得很好的训练效果，例如文献[92]提出的 UNILMV2 模型，当前颇具盛名的 chatGPT 就是基于大型复合模型。在提出了 GPT 模型后[35]，后来的研究者在 GPT 模型的基础上，陆续提出了 GPT2 模型[93] 和 GPT3 模型[94]。GPT2 模型在前代的基础上增大模型规模，并且使用了大规模未标注的自然语言文本来进行预训练。GPT3 则在 GPT2 的基础更进一步地增大了模型规模，拥有 1 750 亿个参数，是 GPT2 的 100 倍以上，并且 GPT3 作为预训练模型，具有迁移学习的能力，只有少量标注数据的情况下也可以快速适应到新的任务，同时 GPT3 模型也在自然语言处理的各个领域都达到了极高的性能。这些研究也证实了大型模型的潜力。

5.2　多样化输入的问题生成

当前问题生成的研究大多是基于英文阅读理解数据集 SQuAD，对于其他语言或者输入生成问题的研究比较少。从语言上来说，问题生成可以面向更多的语言或者输入类型。生成不同语言的问题，如中文问题的生成[95]。小语种问题生成[80] 等；面向不同类型的输入生成问题，如对输入的图像进行提问[75]、对输入的视频进行提问[96] 等。未来可能还会出现针对音频或其他输入的问题生成研究。通过研究多样化输入的问题生成，可以让问题生成任务拥有更多的应用场景，这也是未来的一个重要研究方向。

6　结束语

问题生成是自然语言生成任务中一个重要的研究方向。本文介绍了问题生成的应用和前沿研究情况。本文首先阐述了问题生成研究的背景和意义，然后回顾了问题生成不同发展阶段使用的方法，接着介绍了该研究方向的一些应用领域、问题生成的前沿研究成果和评价指标，最后对问题生成未来的研究方向进行了展望。

参考文献

[1] PAN Liangming, LEI Wenqiang, CHUA T S, et al. Recent advances in neural question generation[J]. arXiv preprint arXiv, 1905.08949, 2019.

[2] DONG Chenhe, LI Yinghui, GONG Haifan, et al. A survey of natural language generation [J]. arXiv preprint arXiv, 2112.11739,2021.

[3] ZHANG Ruqing, GUO Jiafeng, CHEN Lu, et al. A review on question generation from natural language text [J]. ACM Transactions on Information Systems, 2022,40(1):14.

[4] RAJPURKAR P, ZHANG Jian, LOPYREV K, et al. SQuAD: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv,1606.05250,2016.

[5] BAJAJ P, CAMPOS D, CRASWELL N, et al. MS MARCO: A human generated machine reading comprehension dataset [J]. arXiv preprint arXiv,1611.09268,2016.

[6] LAI Guokun, XIE Qizhe, LIU Hanxiao, et al. RACE: Large-scale reading comprehension dataset from examinations[J]. arXiv preprint arXiv,1704.04683, 2017.

[7] TRISCHLER A, WANG Tong, YUAN Xingdi,et al. NewsQA:A machine comprehension dataset [J]. arXiv preprint arXiv,1611.09830, 2016.

[8] KOČISKÝ T, SCHWARZ J, BLUNSOM P, et al. The narrative QA reading comprehension challenge [J]. arXiv preprint arXiv, 1712.07040, 2017.

[9] AFZAL N, MITKOV R. Automatic generation of multiple choice questions using dependency-based semantic relations [J]. Soft Computing,2014, 18(7):1269-1281.

[10] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: Unanswerable questions for SQuAD [J]. arXiv preprint arXiv,1806.03822, 2018.

[11] MOSTOW J, CHEN Wei. Generating instruction automatically for the reading strategy of self-questioning[C]// Proceedings of the

2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling. Amsterdam, Netherlands: IOS Press, 2009: 465-472.

[12] LINDBERG D, POPOWICH F, NESBIT J, et al. Generating natural language questions to support learning on-line [C]// Proceedings of the 14th European Workshop on Natural Language Generation. ACL, 2013: 105-114.

[13] UREEL I I, FORBUS K, RIESBECK C, et al. Question generation for learning by reading [C]//Proceedings of the AAAI Workshop on Inference for Textual Question Answering. Pittsburgh, USA: AAAI, 2005: 22-26.

[14] HARABAGIU S, HICKL A, LEHMANN J, et al. Experiments with interactive question-answering [C]// Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL' 05). ACL, 2005: 205-214.

[15] MITKOV R, HA L A. Computer-aided generation of multiple-ice tests [C]// Proceedings of the HLT-NAACL'03 Workshop on Building Educational Applications Using Natural Language Processing. ACL, 2003: 17-22.

[16] HEILMAN M, SMITH N A. Good question! Statistical ranking for question generation [C]//Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. New York: ACM, 2010: 609-617.

[17] YAO Xuchen, ZHANG Yi. Question generation with minimal recursion semantics [C]//Proceedings of the 3rd Workshop on Question Generation. Pittsburgh, USA: Citeseer, 2010: 68-75.

[18] MANNEM P, PRASAD R, JOSHI R. Question generation from paragraphs at UPenn: QG-STEC system description [C]// Proceedings of the QG2010: The Third Workshop on Question Generation. Pennsylvania, USA: WikiCFP, 2010: 84-91.

[19] FATTOH I E, ABOUTABL A E, HAGGAG M H. Semantic question generation using artificial immunity [J]. arXiv preprint arXiv, 1809.02393, 2018.

[20] LIU Tianyu, WEI Bingzhen, CHANG Baobao, et al. Large-scale simple question generation by template-based seq2seq learning [C]// 6th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2017). Cham: Springer, 2018: 75-87.

[21] CHO K, MERRIËNBOER V B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv, 1406.1078, 2014.

[22] DHOLE K, MANNIG C D. Syn-QG: Syntactic and shallow semantic rules for question generation [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 752-765.

[23] MILLERG A. WordNet: An electronic lexical database [M]. Massachusetts, USA: MIT Press, 1998.

[24] SCHULERK K. VerbNet: Abroad coverage, comprehensive verb lexicon [D]. Pennsylvania, USA: University of Pennsylvania, 2005.

[25] HOCHREITER S, SCHMIDHUBER J. Long Short-Term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[26] GUPTA D, CHAUHAN H, TEJ A R, et al. Reinforced multi-task approach for multi-hop question generation [J]. arXiv preprint arXiv, 2004.02143, 2020.

[27] WANG Yutong, ZHENG Jiyuan, LIU Qijiong, et al. Weak supervision enhanced generative network for question generation [J]. arXiv preprint arXiv, 1907.00607, 2019.

[28] SONG Linfeng, WANG Zhiguo, HAMZA W, et al. Leveraging context information for natural question generation [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short Papers). ACL, 2018: 569-574.

[29] KIM Y, LEE H, SHIN J, et al. Improving neural question generation using answer separation [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019: 6602-6609.

[30] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv, 1409.0473, 2014.

[31] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[32] SCIALOM T, PIWOWARSKI B, STAIANO J. Self-attention architectures for answer-agnostic neural question generation [C]//Proceedings of the 57th annual meeting of the Association for Computational Linguistics. ACL, 2019: 6027-6032.

[33] LOPEZ L E, CRUZ D K, CRUZ J C B, et al. Transformer-based end-to-end question generation [J]. arXiv preprint arXiv, 2005.01107, 2020.

[34] PAN Liangming, XIE Yuxi, FENG Yansong, et al. Semantic graphs for generating deep questions [J]. arXiv preprint arXiv, 2004.12704, 2020.

[35] RADFORD A, NARSIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [J]. arXiv preprint arXiv, 1806.05266, 2020.

[36] GAO Yifan, BING Lidong, CHEN Wang, et al. Difficulty controllable generation of reading comprehension questions [J]. arXiv preprint arXiv, 1807.03586, 2018.

[37] WANG Wenhui, YANG Nan, WEI Furu, et al. Gated self-matching networks for reading comprehension and question answering [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. ACL, 2017: 189-198.

[38] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension [J]. arXiv preprint arXiv, 1611.01603, 2016.

[39] ZHAO Jie, DENG Xiang, SUN Huan. Easy-to-hard: Leveraging simple questions for complex question generation [J]. arXiv preprint arXiv, 1912.02367, 2019.

[40] KUMAR V, HUA Yuncheng, RAMAKRISHNAN G, et al. Difficulty-controllable multi-hop question generation from knowledge graphs [C]// 18th International Semantic Web Conference on The Semantic Web-ISWC 2019. Cham: Springer, 2019: 382-398.

[41] CHENG Yi, LI Siyao, LIU Bang, et al. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting [J]. arXiv preprint arXiv, 2105.11698, 2021.

[42] SRIVASTAVA M, GOODMAN N. Question generation for adaptive education [J]. arXiv preprint arXiv, 2106.04262, 2021.

[43] GU Jing, MIRSHEKARI M, YU Zhou, et al. Chaincqg: Flow-aware conversational question generation [J]. arXiv preprint arXiv, 2102.02864, 2021.

[44] PANG Wei, WANG Xiaojie. Visual dialogue state tracking for question generation [J]. arXiv preprint arXiv, 1911.07928, 2019.

[45] VRIESD H, STRUB F, CHANDAR S, et al. Guesswhat?! visual object discovery through multi-modal dialogue[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5503-5512.

[46] CHAI Zi, WAN Xiaojun. Learning to ask more: Semi-autoregressive sequential question generation under dual-graph interaction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 225-237.

[47] YU Lili, CHEN H, WANG Sida, et al. Interactive classification by asking informative questions[J]. arXiv preprint arXiv, 1911.03598, 2019.

[48] LIU Zhongkun, REN Pengjie, CHEN Zhumin, et al. Learning to ask conversational questions by optimizing Levenshtein distance [J]. arXiv preprint arXiv, 2106.15903, 2021.

[49] SHEN Lei, MENG Fandong, ZHANG Jinchao, et al. GTM: A generative triple-wise model for conversational question generation[J]. arXiv preprint arXiv, 2106.03635, 2021.

[50] REDDY R G, YADAV V, SULTAN M A, et al. Towards robust neural retrieval models with synthetic pre-training[J]. arXiv preprint arXiv, 2104.07800, 2021.

[51] ZHOU Wenjie, ZHANG Minghua, WU Yufang. Multi-task learning with language modeling for question generation[J]. arXiv preprint arXiv, 1908.11813, 2019.

[52] JIA Xin, ZHOU Wenjie, SUN Xu, et al. How to ask good questions? try to leverage paraphrases[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 6130-6140.

[53] FAN Zhihao, WEI Zhongyu, LI Piji, et al. A question type driven framework to diversify visual question generation[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence(IJCAI'18). Stockholm, Sweden: AAAI, 2018: 4048-4054.

[54] LYU Chenyang, SHANG Lifeng, GRAHAM Y, et al. Improving unsupervised question answering via summarization-informed question generation[J]. arXiv preprint arXiv, 2109.07954, 2021.

[55] KHOT T, KHASHABI D, RICHARDSON K, et al. Text modular networks: Learning to decompose tasks in the language of existing models[J]. arXiv preprint arXiv, 2009.00751, 2020.

[56] SHUKLA P, ELMADJIAN C, SHARAN R, et al. What should I ask? using conversationally informative rewards for goal-oriented visual dialog[J]. arXiv preprint arXiv, 1907.12021, 2019.

[57] LUO Hongyin, LI Shangwen, GAO Mingye, et al. Cooperative Self-training of machine reading comprehension[J]. arXiv preprint arXiv, 2103.07449, 2021.

[58] REBUFFEL C, SCIALOM T, SOULIER L, et al. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation[J]. arXiv preprint arXiv, 2104.07555, 2021.

[59] JIA Xin, ZHOU Wenjie, SUN Xu, et al. Eqg-race: Examination-type question generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2021: 13143-13151.

[60] YUE Xiang, ZHANG Xinliang, YAO Ziyu, et al. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering[C]//2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway, NJ: IEEE, 2021: 580-587.

[61] LEHMAN E, LIALIN V, LEGASPI K Y, et al. Learning to Ask Like a Physician[J]. arXiv preprint arXiv, 2206.02696, 2022.

[62] KULSHRESHTHA D, BELFER R, SERBAN I V, et al. Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval[J]. arXiv preprint arXiv, 2104.08801, 2021.

[63] CHO W S, ZHANG Yizhe, RAO S, et al. Contrastive multi-document question generation[J]. arXiv preprint arXiv, 1911.03047, 2019.

[64] SCIALOM T, STAIANO J. Ask to learn: A study on curiosity-driven question generation[J]. arXiv preprint arXiv, 1911.03350, 2019.

[65] YU Jianxing, LIU Wei, QIU Shuang, et al. Low-resource generation of multi-hop reasoning questions[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 6729-6739.

[66] LIANG D, XU P, SHAKERI S, et al. Embedding-based zero-shot retrieval through query generation[J]. arXiv preprint arXiv, 2009.10270, 2020.

[67] ZHANG Shiyue, BANSAL M. Addressing semantic drift in question generation for semi-supervised question answering[J]. arXiv preprint arXiv, 1909.06356, 2019.

[68] PYATKIN V, HWANG J D, SRIKUMAR V, et al. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations[J]. arXiv preprint arXiv, 2212.10409, 2022.

[69] OUSIDHOUM N, YUAN Zhangdie, VLACHOS A. Varifocal question generation for fact-checking[J]. arXiv preprint arXiv, 2210.12400, 2022.

[70] WU Kun, WANG Lijie, LI Zhenghua, et al. Data augmentation with hierarchical SQL-to-question generation for cross-domain text-to-SQL parsing[J]. arXiv preprint arXiv, 2103.02227, 2021.

[71] CAO S, WANG L. Controllable open-ended question generation with a new question type ontology[J]. arXiv preprint arXiv, 2107.00152, 2021.

[72] LIU Tianqiao, FANG Qian, DING Wenbiao, et al. Mathematical word problem generation from commonsense knowledge graph and equations[J]. arXiv preprint arXiv, 2010.06196, 2020.

[73] SACHAN D S, WU Lingfei, SACHAN M, et al. Stronger transformers for neural multi-hop question generation[J]. arXiv preprint arXiv, 2010.11374, 2020.

[74] YANG A, MIECH A, SIVIC J, et al. Learning to answer visual questions from web videos[J]. arXiv preprint arXiv, 2205.05019, 2022.

[75] YEH M H, CHEN V, KU L W. Multi-VQG: Generating engaging questions for multiple images[J]. arXiv preprint arXiv, 2211.07441, 2022.

[76] GAO Yifan, BING Lidong, LI Piji, et al. Generating distractors for reading comprehension questions from real examinations [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019: 6423-6430.

[77] LIANG Chen, YANG Xiao, DAVE N, et al. Distractor generation for multiple choice questions using learning to rank [C]//Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. New Orleans, USA: dblp, 2018: 284-290.

[78] BITEW S K, HADIFAR A, STERCKX L, et al. Learning to

reuse distractors to support multiple choice question generation in education[J]. IEEE Transactions on Learning Technologies, 2022, 17: 375-390.

[79] CHI Zewen, DONG Li, WEI Furu, et al. Cross-lingual natural language generation via pre-training[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020: 7570-7577.

[80] KALPAKCHI D, BOYE J. Quinductor: A multilingual data-driven method for generating reading-comprehension questions using universal dependencies[J]. arXiv preprint arXiv, 2103.10121, 2021.

[81] GONG Shansan, LI Mukai, FENG Jiangtao, et al. Diffuseq: Sequence to sequence text generation with diffusion models[J]. arXiv preprint arXiv, 2210.08933, 2022.

[82] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[83] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[84] QI Weizhen, GONG Yeyun, JIAO Jian, et al. Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining[J]. arXiv preprint arXiv, 2012.18828, 2021.

[85] LIU Bang, WEI Haojie, NIU Di, et al. Asking questions the human way: Scalable question-answer generation from text corpus[C]//Proceedings of The Web Conference 2020. New York: ACM, 2020: 2032-2043.

[86] SHINODA K, AIZAWA A. Variational question-answer pair generation for machine reading comprehension[J]. arXiv preprint arXiv, 2004.03238, 2020.

[87] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. ACL, 2002: 311-318.

[88] DENKOWSKI M, LAVIE A. Meteor universal: Language specific translation evaluation for any target language [C]// Proceedings of the Ninth Workshop on Statistical Machine Translation. ACL, 2014: 376-380.

[89] LIN C Y. Rouge: A package for automatic evaluation of summaries [C]//Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. ACL, 2004: 74-81.

[90] NEMA P, KHAPRA M M. Towards a better metric for evaluating question generation systems [J]. arXiv preprint arXiv, 1808.10192, 2018.

[91] SULTAN M A, CHANDEL S, ASTUDILLO R F, et al. On the importance of diversity in question generation for QA [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 5651-5656.

[92] BAO Hangbo, DONG Li, WEI Furu, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training [J]. arXiv preprint arXiv, 2002.12804, 2020.

[93] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[94] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[95] LIU Ming, RUS V, LIU Li. Automatic chinese factual question generation[J]. IEEE Transactions on Learning Technologies, 2016, 10(2): 194-204.

[96] WANG Y S, SU H T, CHANG C H, et al. Video question generation via semantic rich cross-modal self-attention networks learning[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 2423-2427.