智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

语义图支持的阅读理解型问题的自动生成

徐坚

引用本文:
徐坚. 语义图支持的阅读理解型问题的自动生成[J]. 智能系统学报, 2024, 19(2): 420–428.
XU Jian. Generating reading comprehension questions automatically based on semantic graphs[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 420–428.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202207001

DOI: 10.11992/tis.202207001
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20231115.1518.006

语义图支持的阅读理解型问题的自动生成

徐坚[1,2]

（1. 云南师范大学 民族教育信息化教育部重点实验室, 云南 昆明 650500; 2. 曲靖师范学院 信息工程学院, 云南 曲靖 655011）

摘　要: 问题自动生成是人工智能领域的一项技术, 其目标是根据输入的文本模拟人类的能力, 自动生成相关问题。目前的问题自动生成研究主要基于通用数据集生成问题, 缺乏专门针对教育领域的问题生成研究。为此, 专注于面向中学生的问题自动生成进行研究。构建一个专门为问题生成模型训练需求而设计的数据集 RACE4QG, 以满足中学生教育领域的独特需求; 开发一个端到端的问题自动生成模型, 该模型训练于数据集 RACE4Q, 并采用改进型“编码器−解码器”方案, 编码器主要采用两层双向门控循环单元, 其输入为单词和答案标记的嵌入表示, 编码器的隐藏层采用门控自注意力机制获得“文章和答案”的联合表示后, 再输入到解码器生成问题。试验结果显示, 该模型优于最优基线模型, 3 个评价指标 BLEU-4、ROUGE-L 和 METEOR 分别提高了 3.61%、1.66% 和 1.44%。

关键词: 语义图; 数据集; 自动问题生成模型; 编码器; 解码器; 答案标记; 图注意力网络; 门控循环单元
中图分类号: TP311　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)02−0420−09

中文引用格式: 徐坚. 语义图支持的阅读理解型问题的自动生成 [J]. 智能系统学报, 2024, 19(2): 420–428.
英文引用格式: XU Jian. Generating reading comprehension questions automatically based on semantic graphs[J]. CAAI transactions on intelligent systems, 2024, 19(2): 420–428.

Generating reading comprehension questions automatically based on semantic graphs

XU Jian[1,2]

(1. Key Laboratory of Educational Informatization for Nationalities, Yunnan Normal University, Kunming 650500, China; 2. School of Information Engineering, Qujing Normal University, Qujing 655011, China)

Abstract: Automatic question generation is a technology in the field of artificial intelligence. Its goal is to simulate human capabilities and automatically generate relevant questions based on input text. Current research on automatic question generation is mainly based on generating questions from general datasets, and there is a lack of research on question generation specifically targeting the field of education. To this end, this article focuses on the automatic generation of questions for middle school students.First, this article constructs a dataset RACE4QG specifically designed for the training needs of question generation models to meet the unique needs of the field of middle school student education. Secondly, we developed an end-to-end automatic problem generation model, which was trained on the RACE4Q dataset.In the improved "encoder-decoder" scheme, the encoder mainly adopts a two-layer bidirectional gated recurrent unit, whose input is the word embedding and answer-tagging embedding, and the hidden layer of the encoder adopts the gated self-attention mechanism to obtain the passage-answer representation, which is then fed to the decoder to generate questions. The experimental results show that the model in this paper is better than the optimal baseline model, and the three evaluation indicators BLEU-4, ROUGE-L, and METEOR are improved by 3.61, 1.66, and 1.44 points, respectively.

Keywords: semantic graph; dataset; automatic question generation model; encoder; decoder; answer tagging; graph attention network; gated recurrent units

阅读是数字时代人类交流的一项重要技能。非英语国家最重要的教学目标之一是提高学习者的阅读理解能力[1]。提问教学作为一种重要的教

收稿日期: 2022−07−01.　**网络出版日期:** 2023−11−16.
基金项目: 国家自然科学基金项目 (62166050); 云南师范大学 2020 年研究生科研创新基金项目 (YSDBS178).
通信作者: 徐坚. E-mail: qjncxj@126.com.

学手段，要求教师能根据指定的教学材料快速有效地生成高质量的阅读理解问题，然而，教师要完成这任务颇具难度。

幸运的是，人工智能的飞速发展[2]为自动问题生成提供了机遇。近年来，自动问题生成 (automatic question generation, QG) 作为自然语言处理领域的一个重要分支，吸引了大批神经语言处理社区研究者的兴趣[3-4]。自动问题生成模型的实质是人工神经网络，其基于海量数据来训练问题生成模型。因此，自动问题生成的主要任务有 2 个：一是构建可用于训练模型的大规模数据集；二是设计优良的模型。传统上，问题生成数据集的构建通常采用“改造法”，即改造传统的问答数据集如 SQuAD[5]、Narrative QA[6]、HotpotQA[7] 和 RACE[8] 等，问题生成模型的成熟架构一般基于循环神经网络。然而，就面向教育领域的自动问题生成任务而言，仍面临 2 个挑战：1) 缺乏特定于教育领域的问题生成数据集；2) 现有的自动问题生成模型主要使用基于 RNN 的一些变体 (如 LSTM[9])，而 RNN 模型具有固有的序列性质并且难以处理长输入序列等缺陷，迫使传统的基于 RNN 的问题生成模型仅能生成句子级而非文章级的问题，例如，Du 等[10] 利用句子级信息来提高自动问题生成模型的性能，鉴于其研究工作未能很好地处理文章级输入 (即多个句子)，Zhao 等[4] 引入最大指针机制 (maxout-pointer) 和门控制自注意力机制来解决此问题，并取得更好的性能。然而，即使是最好的自动问题生成模型也难以在教育领域生成令人满意的问题，这种功能障碍可以归因于双重挑战：教育型自动问题生成数据集的质量和自动问题生成模型的性能。在教育中，解决这些挑战具有突出的现实意义，教师可以利用自动生成的问题来提高学习者的课堂参与度和阅读能力。

本研究基于现有的教育型问答数据集 RACE，将其改造为一个有效的问题生成数据集。此外，考虑到现有自动问题生成模型在教育领域表现欠佳，提出了一个更强大的自动问题生成模型，该模型拥有更先进的编码器和解码器。

本研究的贡献如下：

1) 基于现有的教育数据集 RACE 重建一个问题生成数据集 RACE4QG。

2) 基于 RACE4QG 数据集，提出了一种端到端的自动问题生成模型，该模型主要由编码器和解码器组成。在编码器中，采用门控注意力机制来丰富词嵌入。在解码器中，引入注意力和指针网络机制以动态生成问题。

3) 开展对比试验和消融试验。在试验部分，本研究的自动问题生成模型与 4 个基线模型进行了试验对比。另外，为考察本研究模型各功能组件的作用，开展了消融试验以验证去掉各个组件的模块效果。

1　相关工作

自动问题生成最重要的应用就是为阅读理解教学和评价生成问题[11-13]。问题生成的最初思想主要基于启发式规则，这些规则使用手动构建的模板来生成问题，并对生成的问题进行排名。Heilman 等[14] 首次提出通过启发式策略生成问题，该策略利用手动编写的规则进行句法转换，将陈述句转化为问题，再通过逻辑回归模型对生成的问题进行排名获得最佳问题。Yao 等[15] 提出了一种语义重写方法来构建语义结构和语法规则以生成更好的问题。鉴于创建完整语义表示的无数挑战，Labutov 等[16] 寻求避开语义创建，将自动问题生成任务简化为“本体–众包–相关性”工作流程，先将问题模板众包出去，再选择最佳候选模板来生成问题。由于创建模板需要专家完成，且模型的数据和灵活性有限，抑制了问题生成模型的泛化能力。

鉴于基于规则来生成问题既费时又费力，自 2015 年以来，研究者尝试采用统计方法来生成问题，尤其是基于神经网络的问题生成方法受到青睐，其主要思想是采用具有注意力机制的序列到序列 (sequence-to-sequence) 的问题生成方法，其旨在通过使用 sequence-to-sequence 框架来生成问题[17-18]。Du 等[10] 最早使用 sequence-to-sequence 模型来生成问题，试验结果表明，无论是在问题生成的速度和质量方面，sequence-to-sequence 模型明显优于基于规则的模型。尽管如此，该论文也承认该模型生成的问题不能准确地对应到原始上下文的特定部分。为了应对这一挑战，Zhou 等[3] 建议使用注释向量将答案单词 (答案文本中的词) 的位置信息合并到编码过程中。Song 等[17] 没有采用注释向量来标记答案位置，而是引入统一的框架对文章和答案进行编码。然而，以上问题生成方法在处理长输入上下文时的效果不佳。为此，Zhao 等[4] 引入了一种混合机制，该机制使用门控自注意力和最大指针技术来处理长输入上下文。更进一步，Yuan 等[18] 首先通过在一些自然

语言处理任务上训练大型的预训练神经模型来获得深度语言特征，然后将这些深度语言特征整合到一个 sequence-to-sequence 的问题生成模型中，以引导问题的生成，这个模型将基线模型的 BLEU-4 指标提高了 6.2%。

虽然以上问题生成方法在通用领域已经取得了较好效果，但在教育领域的表现欠佳，原因有二：1）缺乏专门针对教育领域的问题生成数据集；2）现在的问题生成模型也未能充分挖掘文章中的语义结构信息。为此，本研究重构了一个问题生成数据集 RACE4EQG，设计一个答案引导的图注意网络 AO-GAT 来捕获文章的语义图，以生成更好的问题。

2 数据集的构建

RACE[8] 是一个大规模的考试类问答数据集，由卡内基梅隆大学于 2017 年发布，RACE 的语料源自中国初中和高中学生的英语考试真题，该数据集包括 27 933 篇文章，以及 97 687 个阅读理解型问题。在 RACE 中，每个样本是一个四元组“文章，答案，问题，干扰答案”，如果将该四元组的干扰答案删除，形成三元组“文章，答案，问题”，再进行一些预处理工作，就可用于训练本研究的自动问题生成模型。

本研究借鉴 Jia 等[19] 的思想来改造 RACE，重构能用于本研究自动问题生成模型训练的数据集 RACE4QG：

1) 删除干扰答案。鉴于本研究的自动问题生成模型是基于答案引导来生成问题，如果以干扰答案为引导，势必生成不正确的问题，因此，将 RACE 中的干扰答案移除后，数据集的样本格式就由四元组“文章，答案，问题，干扰答案”变成三元组“文章，答案，问题”。

2) 提升数据集的样本质量。RACE 数据集中的问题分为两类：完形填空式的问题和标准的问题。完形填空式的问题也叫填空题，用于传统的问答任务，不能直接用于本研究的自动问题生成任务，因此，需要删除完形填空式的问题。

3) 标注答案单词。使用前文提到的离散型答案标记策略对文章中的单词进行标注，以引导问题的生成。

经过以上 3 个步骤，重构 RACE 得到新的问题生成数据集 RACE4QG，该数据集的样本数达 46 397 个，每个样本的结构为三元组“文章、答案、问题”。这些样本划分到训练集、评估集和测试集的样本数分别为 41 791、2 312 和 2 294。

此外，为了利用数据集 RACE4Q 中文章的答案来引导模型生成好的问题，将在 2.2.1 小节描述如何采用新的答案标记策略来标注该数据集中的答案单词。

3 自动问题生成模型

3.1 模型概述

本研究提出了一种基于重构数据集的端到端自动问题生成模型，其根据给定的输入文章和答案生成多个语法一致且流畅的问题。

形式上，自动问题生成任务被定义为生成问题 $\overline{q}$

$$\overline{q} = \arg\max_{q} P(q|p,a) \tag{1}$$

式中，p、a 和 q 分别表示文章、答案和问题。其中文章 p 有 m 个单词，记为 $p = \{x_t\}_{t=1}^{m}$，$P(q|p,a)$ 是一个条件概率。

自动问题生成的主流模型是基于 sequence-to-sequence 的思想[20]。然而，使用这种策略的自动问题生成模型无法生成满足阅读理解教学需要的问题。为了应对该挑战，本研究首先以选用段落级提问模型[4] 作为基线模型，引入门控循环单元（gated recurrent units, GRU) 和图注意力网络 (graph attention networks, GAT) 后，形成新的端到端自动问题生成模型 (见图 1)，该模型包括 5 个部分：输入层、嵌入表示层、编码器层、解码器层和输出层。

3.2 门控注意力指导文章与答案的联合编码

自动问题生成模型的编码器采用两层双向的 GRU (gated recurrent units), GRU 作为 LSTM 的变体，具有参数少，性能优的特性。GRU 在时间步 t 的隐藏状态 h_t^p 由前向和后向隐藏状态拼接，即 $\boldsymbol{h}_t^p = \left[\overrightarrow{\boldsymbol{h}}_t^p, \overleftarrow{\boldsymbol{h}}_t^p\right]$，如此，所有时间步的隐藏状态可表示为 $\boldsymbol{H}^p = \{\boldsymbol{h}_t^p\}_{t=1}^{m}$。GRU 以文章和答案为输入，然后输出“文章–答案”的向量表示，具体来讲，在每个时间步 t, GRU 的隐藏状态可表示为

$$\boldsymbol{h}_t^p = \mathrm{GRU}(\boldsymbol{h}_{t-1}^p, x_t^p) \tag{2}$$

式中：$\boldsymbol{h}_{t-1}^p$ 为上一时间步的隐藏状态；x_t^p 是时间步 t 的一个输入单词 (即文章的单词)。

3.2.1 离散型答案标记策略

本研究的自动问题生成模型属于答案导引型问题生成模型，以文章和答案为输入，再生成问题。因此，对于数据集中每个样本里面的文章，需要将这些文章中的那些包含在答案中的单词标记出来，以便模型生成质量更高的问题。

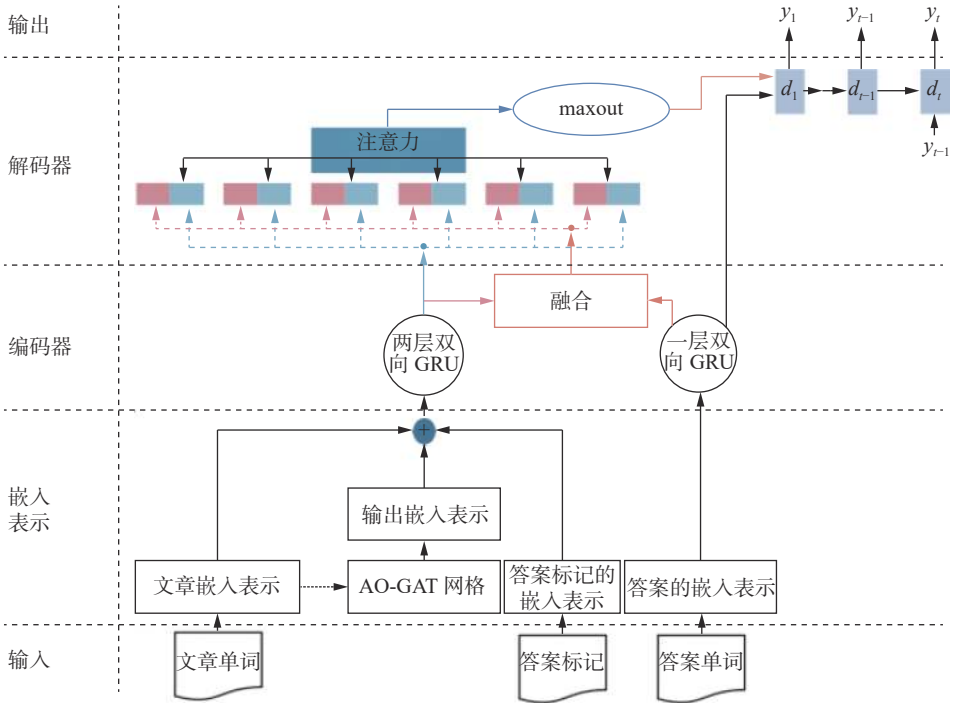

图 1　端到端的自动问题生成模型

Fig. 1　End-to-end automatic question generation model

本研究的 RACE4QG 数据集源自前人的问答数据集 RACE。鉴于 RACE 是一个问答 (question answering, QG) 数据集，其问题和答案均源自实际的英语考试，答案单词分散于文章中，这不同于问答数据集 (例如 SQuAD) 中答案单词是连续出现的序列，因此，传统的连续型答案标记方法[3]不适用于本研究自动问题生成任务。本研究采用离散型答案标注策略，具体来讲，给定一个答案（一般由多个单词构成），本研究首先将答案执行分词操作，获得一个单词集合，再删除停用词，余下的单词放到集合 X。如果文章中的单词属于集合 X，该单词被标记为“A”，其余的文章单词被标记为“O”。

3.2.2　答案引导的图注意网络

文章中各句子内部的语义结构信息和句子之间的关联信息对生成更高质量的问题提供重要信息。本小节将就此问题展开研究。

本研究的自动问题生成任务旨在生成真正的阅读理解问题，这些问题需要考察学习者在句子内和句子之间进行深度推理能力，这个任务对传统的自动问题生成模型而言颇具难度。图神经网络作为一种能够从图结构数据中学习特征的神经网络，是图分析方法与深度神经网络的结合，在拟合单个样本特征之后，进一步提取样本间的关系信息。作为图神经网络的典型代表，图注意力机制 (graph attention networks, GAT) 已经广泛应用于计算机视觉领域[21]。受 Zhang 等[22]的启发，本研究提出了一个答案引导的图注意网络 (answer-oriented graph attention network , AO-GAT) 来捕获文章内部的依赖关系，以生成高质量的问题。AO-GAT 中的 AO 表示构建由答案引导的文章级依存句法图，简称“语义图”，GAT 可将该图转换为嵌入表示。AO-GAT 的构建步骤如下。

1) 构建由答案引导的语义图。为了捕获文章中的内部依赖信息，可对指定文章构建其语义图，如图 2 所示。

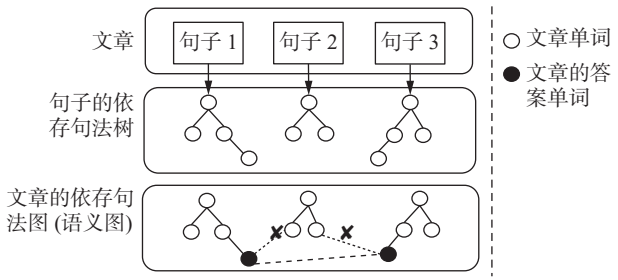

图 2　答案引导的语义图的构建过程

Fig. 2　Building process of answer-oriented graph

具体步骤如下：

①将文章分割为句子。将输入的文章 P 拆分为一个个独立的句子，记为 $p=\{s_i\}_{i=1}^{N}$，其中，s_i 表示文章的第 i 个句子，N 表示文章的句子数量。

②解析句子的句法结构。采用 Stanford-CoreNLP 库[23]生成句子 s_i 的解析依赖树 s_{tree_i}，以此类推，文章的每个句子对应一棵树，整篇文章

就对应一个树的集合。

③生成语义图。为了将相邻句子的依存句法树连接起来，可查找两棵树中对应原始句子首尾单词的边界节点，若这两棵树的边界节点在原文中相邻，则将这 2 个边界节点连接。以此类推，可得到一张由 N 棵树连接而成的语义图。

④修剪语义图。鉴于只有包含答案的句子才助于生成有价值的问题，本研究依次检查语义图中的每棵树是否包含答案单词，若不包含，则将此树从图中删除，最终可获得一个精简后的语义图（简称“精简语义图”），图中的每个节点代表一个文章单词。

2）将精减语义图转换为邻接矩阵。该图存储为一个邻接矩阵 $\boldsymbol{G}$，该矩阵为 0-1 矩阵，取值规则：如果图中的节点 i 和节点 j 相连，则矩阵的第 i 行第 j 列的值 $\boldsymbol{G}_{i,j}$ 为 1，否则为 0。

3）获取精简语义图的 GAT 嵌入表示。为了获得更丰富的图节点（每个节点对应一个单词）特征，可将图注意力机制应用于图邻接矩阵。在编码阶段，句子内和句子间的依赖信息存储在图邻接矩阵 $\boldsymbol{G}$ 中，本研究采用图注意力机制（graph attention networks, GAT）从 $\boldsymbol{G}$ 中提取此信息。图注意力机制以文章单词嵌入和邻接矩阵 $\boldsymbol{G}$ 作为输入，然后输出更高维的词嵌入表示，即 GAT 嵌入。详细步骤如下。

①GAT 的输入为 $\boldsymbol{G}$ 和 $\boldsymbol{e}$，其中 $\boldsymbol{G}$ 为当前文章的邻接矩阵，$\boldsymbol{e}$ 为整个文章的特征表示，$\boldsymbol{e}=\{\boldsymbol{e}_1, \boldsymbol{e}_2,\cdots, \boldsymbol{e}_m\}$，$\boldsymbol{e}_i\in\mathbf{R}^F$，$F$ 表示单个节点（对应文章单词）的特征数，$\boldsymbol{e}_i$ 为当前文章第 i 个单词的一组特征。

②计算节点 i 对其所有邻居的注意力

$$a_{ij}=\frac{\exp\left(P\mathrm{ReLU}\left((\boldsymbol{W}^a)^{\mathrm{T}}[\boldsymbol{e}_i,\boldsymbol{e}_j]\right)\right)}{\sum\limits_{k\in N_i}\exp\left(P\mathrm{ReLU}\left((\boldsymbol{W}^a)^{\mathrm{T}}[\boldsymbol{e}_i,\boldsymbol{e}_k]\right)\right)} \tag{3}$$

式中：$\boldsymbol{W}^a$ 是一个可训练的权重向量，$\boldsymbol{W}^a\in\mathbf{R}^{2F'}$；$N_i$ 表示节点 i 的所有邻居节点。值得注意的是，GAT 的原文使用 LeakyReLU 作为激活函数。在本研究试验中，比较了 ELU、ReLU、ReLU6 和 PReLU 等各种激活函数，发现使用 PReLU 的效果更好。

③GAT 的输出被送至编码器。文章的语义图中全部 N 个节点的特征序列 $\boldsymbol{e}=\{\boldsymbol{e}_1, \boldsymbol{e}_2,\cdots, \boldsymbol{e}_N\}$，$\boldsymbol{e}_i\in\mathbf{R}^F$ 表示图中第 i 个节点的全部特征，F 表示特征的个数。GAT 以 $\boldsymbol{e}$ 作为输入，得到对应的输出 $\boldsymbol{e}'=\{\boldsymbol{e}'_1,\boldsymbol{e}'_2,\cdots,\boldsymbol{e}'_N\}$，$\boldsymbol{e}'_i\in\mathbf{R}^F$。$\boldsymbol{e}'$的计算过程可展开为

$$\boldsymbol{e}'_i=\mathrm{sigmoid}\left(\sum_{j\in N_i}a_{ij}\boldsymbol{W}\boldsymbol{e}_j\right) \tag{4}$$

式中：$a_{i,j}$ 表示节点 i 对其邻居节点 j 的注意力；$\boldsymbol{e}_j$ 表示邻居节点 j 的全部特征；$\boldsymbol{W}$ 为可训练的权值矩阵。

为了进一步提高自注意力的性能，本研究利用了多头注意力机制[24]，每个注意力头依次捕获文章中某一方面的特征后，将所有注意力头捕获的特征连接起来作为总的特征。此时，式 (4) 变成

$$\boldsymbol{e}'_i=\mathrm{sigmoid}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}a_{ij}^k\boldsymbol{W}^k\boldsymbol{e}_j\right) \tag{5}$$

式中：K 为注意力头的总个数，本研究的试验结果表明 K=8 时网络的效果最好。将 GAT 的最终输出$\boldsymbol{e}'=\{\boldsymbol{e}'_1,\boldsymbol{e}'_2,\cdots,\boldsymbol{e}'_m\}$ 以及词嵌入和答案标记共同拼接起来作为编码器的输入。

3.2.3 门控自注意力

经过上述步骤，获得了原始的“文章−答案”联合表示，为了聚合来自文章的信息并结合文章内的依赖关系以改进每个时间步的“文章−答案”嵌入，采用门控自注意力机制来获得自匹配表示：

$$\boldsymbol{a}_t^E=\mathrm{softmax}\left((\boldsymbol{H}^p)^{\mathrm{T}}\boldsymbol{W}^s\boldsymbol{h}_t^p\right) \tag{6}$$

$$\boldsymbol{S}_t=\boldsymbol{H}^p\cdot\boldsymbol{a}_t^E \tag{7}$$

式中，$\boldsymbol{S}_t$ 是所有文章单词嵌入的加权求和，表示这些单词在时间步 t 对当前单词的关注度。

将文章答案的原始表示$\boldsymbol{h}_t^p$和自匹配表示 $\boldsymbol{S}_t$ 进行组合，得到一个新的文章答案表示$\hat{\boldsymbol{h}}_t^p$

$$\boldsymbol{f}_t=\tanh\left(\boldsymbol{W}^f[\boldsymbol{h}_t^p,\boldsymbol{S}_t]\right) \tag{8}$$

$$\boldsymbol{g}_t=\mathrm{sigmoid}(\boldsymbol{W}^g[\boldsymbol{h}_t^p,\boldsymbol{S}_t]) \tag{9}$$

$$\hat{\boldsymbol{h}}_t^p=\boldsymbol{g}_t\times\boldsymbol{f}_t+(1-\boldsymbol{g}_t)\times\boldsymbol{h}_t^p \tag{10}$$

式中：$\boldsymbol{f}_t$ 是一个新的自匹配增强型文章答案表示；$\boldsymbol{g}_t$ 是一个门向量，用于在$\boldsymbol{h}_t^p$和 $\boldsymbol{S}_t$ 之间选择信息以获得最终的文章答案表示$\hat{\boldsymbol{h}}_t^p$。

3.2.4 文章与答案的融合

文章单词和答案单词之间的信息可以用于生成好的问题。为此，将原始隐藏状态 $\boldsymbol{H}^p$ 和答案隐藏状态 $\boldsymbol{H}^A$ 统一为

$$\boldsymbol{H}^u=\mathrm{union}\left(\boldsymbol{W}^u\left[\hat{\boldsymbol{H}}^P;\boldsymbol{H}^A;\hat{\boldsymbol{H}}^P\times\boldsymbol{H}^A\right]\right) \tag{11}$$

式中，$\hat{\boldsymbol{H}}^P=\left\{\hat{\boldsymbol{h}}_t^p\right\}_{t=1}^m$，$\boldsymbol{H}^A=\{\boldsymbol{h}_t^A\}_{t=1}^n$。$m$ 和 n 分别是答案的单词个数和文章的单词个数。

3.3 基于注意力机制和指针网络的解码方案

解码器采用单层单向 GRU 预测下一个单词 $\boldsymbol{y}_t$。在每个时间步 t，对编码器的最终隐藏状态应用注意力机制凸显文章中更重要的单词，以获得原始文本的动态表示，称为上下文向量 $\boldsymbol{C}_t$。再将 $\boldsymbol{C}_t$、解码器先前生成的全部单词 $(y_1, y_2, \cdots, y_{t-1})$ 和当前解码器状态 $\boldsymbol{d}_t$ 拼接起来，拼接结果输入解码

器后使用指针网络生成单词 $\boldsymbol{y}_t$。

3.3.1 注意力机制

注意力机制一直是保证提问模型性能的默认配置。在本研究的解码器中, Luong 注意力机制[25]用于获得原始注意力(见式 (12))。注意力层(见式 (14))作用于解码器状态 $\boldsymbol{d}_t$ 和注意力上下文向量 $\boldsymbol{C}_t$ 的拼接结果, 以获得新的解码器状态 $\hat{\boldsymbol{d}}_t$

$$\boldsymbol{a}_t^D = \mathrm{softmax}\left(\hat{\boldsymbol{H}}^P \boldsymbol{W}_a \boldsymbol{d}_t\right) \tag{12}$$

$$\boldsymbol{C}_t = \hat{\boldsymbol{H}}^p \boldsymbol{a}_t^D \tag{13}$$

$$\hat{\boldsymbol{d}}_t = \tanh\left(W\left[\boldsymbol{d}_t, \boldsymbol{C}_t\right]\right) \tag{14}$$

$$\boldsymbol{d}_{t+1} = \mathrm{GRU}\left(\left[\boldsymbol{y}_t, \hat{\boldsymbol{d}}_t\right]\right) \tag{15}$$

3.3.2 采用指针网络来生成问题

在本研究的自动问题生成任务中, 解码器的输出词汇必须根据输入序列的长度动态变化, 以保证生成更好的问题。然而, 传统的基于 sequence-to-sequence 的自动问题生成模型无法应对这一挑战。为此, 受到 See 等[26]的启发, 本研究采用指针网络 (pointer-generator network) 来解决这个问题。

指针网络是由传统的生成器模型和指针网络混合而成, 可提高问题生成的质量。指针网络的生成器模型和指针网络能分别从词汇表和原文取词。在编码器的每一时间步, 将原文单词动态地添加到词汇表获得“扩展词汇表”, 同时, 动态计算一个生成概率 $p_{\mathrm{gen}} \in [0,1]$ 用作软开关, 可决定模型当前的输出单词是从词汇表生成还是从原文拷贝。由此, 解码器在时间步 t 获得一个基于扩展词汇表的概率分布

$$P(w) = p_{\mathrm{gen}} P_{\mathrm{vocab}}(w) + (1 - p_{\mathrm{gen}}) p_{\mathrm{copy}} \tag{16}$$

式中: w 表示扩展词汇表的任一单词; P_{vocab} 表示词汇表中所有单词的概率形成的概率分布; $P_{\mathrm{vocab}}(w)$ 表示 w 出自词汇表的概率; p_{copy} 表示 w 来自原文的概率; p_{gen} 是一个软开关。

该扩展词汇表中的每个单词都须通过式 (16) 计算出该单词作为新词的概率, 至此, 扩展词汇表中概率值最大的单词即为解码器的当前输出单词。

4 试验配置与模型

4.1 试验配置

本研究的自动问题生成模型的词汇表包含 4.5 万个单词, 其他超纲单词设置为 UNK 符号。输入文章和输出问题的最大长度分别为 400 和 30。使用预训练的 GloVe[27] 来初始化词嵌入表示, 并在训练期间对其进行微调。

该自动问题生成模型主要包括 3 个神经网络: 输入段落的编码器、答案编码器、解码器。输入段落编码器选用一种先进的循环神经网络 GRU, 采用两层双向架构, 隐藏单元大小为 600 维 (每个方向为 300 维); 答案编码器使用隐藏单元大小为 600(每个方向一半) 的单层双向 GRU; 解码器使用单层单向 GRU, 隐藏单元大小为 300 维。

Dropout 值设置为 0.3, 除词嵌入之外的所有可训练参数都设置为均匀分布 (−0.1,0.1)。在训练期间随机梯度下降 (stochastic gradient descent, SGD) 被用作优化器, 其中 batch_size=45, 本研究的自动问题生成模型和基线的初始学习率都设置为 0.01。训练模型期间, 前 8 个 epoch 的学习率固定为 0.01, 然后每隔一个 epoch 学习率就减半, 学习率下降时不能低于阈值 0.001。

解码器采用集束搜索策略来生成问题, 集束搜索的大小设置为 10, 模型的检查点在验证集上选择, 在测试集上报告结果。

4.2 基线模型

为了评估本研究的自动问题生成模型与基线的性能对比, 重写了 4 个可下载到代码的基线模型: 1) sequence-to-sequence (seq-to-seq)[28]。使用注意机制和拷贝策略的 sequence-to-sequence 问题生成模型; 2) Pointer-generator[29]。使用 Pointer-generator 机制来生成问题; 3) Transformer[30]。基于 Transformer 的问题生成模型; 4) ELMo-QG[31]。采用指针网络从输入文章中复制单词来生成问题。

公平起见, 做对比试验时, 本研究的答案标记策略也用在基线模型。进一步, 每个基线的编码器使用两层双向 LSTM, 解码器使用单层单向 LSTM, 其余设置保持不变。

5 结果与分析

本研究将自动评估生成问题和真实问题的匹配度。为了全面开展评估任务, 从精度、召回率和语义的视角来选择评价模型的指标。为此, 分别采用 BLEU(表 1 中简称 B)、ROUGE-L 和 METEOR。BLEU 评估生成问题和真实问题之间的 n-gram 精度。ROUGE-L 负责评估召回率。请注意, BLEU 和 ROUGE-L 2 个指标只从浅层面而非语义层面 (深层面) 来评估问题的质量, METEOR 指标可从语义层面来开展评估工作。表 1 列出了本研究模型和基线模型的评估结果。

表 1　BLEU、ROUGE 和 METEOR 对所有系统的自动评估结果

Table 1　Automatic evaluation results on all systems by BLEU, ROUGE, and METEOR

模型	B-1	B-2	B-3	B-4	ROUGE	METEOR
seq-to-seq	23.72	9.95	6.13	5.08	24.32	9.60
Pointer-generator	28.94	13.84	8.44	6.26	30.21	13.11
Transformer	28.91	14.66	9.04	6.50	32.51	14.27
ELMo-QG	34.42	17.76	11.98	**8.82**	**34.01**	**14.87**
本研究模型	**36.24**	**21.11**	**15.12**	**12.43**	**35.67**	**16.31**

通过本研究的自动问题生成模型采用“GRU+GAT”的策略，该模型在所有指标上都优于基线。此外，可以看到 2 个基线模型 (即 sequence-to-sequence 和 Transformer) 之间存在明显的性能差距，主要原因可能是 Transformer 使用了与本研究模型相同的层次结构。此外，该模型比 Transformer 表现更好，这表明 GAT 机制在从句内和句间捕获信息时有重要作用。本研究还比较了集束搜索算法生成的 10 道题的 Jaccard 距离，发现这些题的质量依次下降，原因可能是下一个问题的可能性较低。

为了评估 GAT 机制和 GRU 网络的有效性，开展了 2 种类型的消融试验。首先，本研究模型分别在去掉 GAT 和 GRU 后具有不同程度的性能下降；其次，本研究模型使用“GRU+GAT”的策略，对比论文 AG-GCN[20] 的“LSTM+GCN”的策略，表 2 表明本研究的 GAT 机制优于 GCN，原因是 GAT 可以更好地捕获句子内和句子之间的重要语义和句法信息。同样，如果编码器的 GRU 被传统的 LSTM 替换，本研究模型的性能也会下降，因为 GRU 更适合本研究的数据集。消融试验结果如表 2 所示。

表 2　本研究的智能模型开展消融试验的结果

Table 2　Results of ablation experiments for our model

方法	B-1	B-2	B-3	B-4	ROUGE	METE0R
GRU+GAT	36.24	21.11	15.12	12.43	35.67	16.31
GRU	35.66	18.68	11.61	8.03	35.34	15.90
GRU+GCN	36.11	21.24	15.24	12.36	35.43	15.97
GAT	34.25	17.31	10.34	6.92	34.52	14.75
LSTM+GAT	35.60	21.16	15.21	12.16	34.84	15.45

6　案例研究

为了评估问题生成模型的实际效果，本研究以一段英文为输入，由本研究模型自动生成问题，如下所示。

文章：In Santa Cecilia, Mexico, Imelda Rivera was the wife of a musician. Imelda’s husband left her and her daughter, coco, to pursue a career in music. She banned music in the family and opened a shoemaking family business.

答案 1：Imelda Rivera

问题 1：Who was the wife of a musician in Santa Cecilia, Mexico?

答案 2：coco

问题 2：Who did Imelda’s husband leave her and her daughter to pursue a career in music?

答案 3：shoemaking

问题 3：What business did Imelda’s husband open in Santa Cecilia, Mexico?

第 1 个和第 3 个问题的表达较为流利、准确，缘于生成的问题所在的上下文语义比较明确，句子结构规范，而第 2 个问题的上下文句法结构比较复杂，同时还需要模型具有较好的推理能力，对于此类推理性问题的自动生成研究，仍是本领域的研究的重点和难点。

7　结束语

在阅读理解教学中存在一个长期的痛点，即教师无法根据任意文章自动、及时地产生问题。本研究提出了一个端到端的自动问题生成模型，该模型在重构的 RACE4QG 数据集上训练。为了丰富文章答案嵌入表示，在模型的编码器中引入了一种结合门控策略与自注意力方法的 AO-GAT 机制。此外，模型解码器的性能通过引入注意力机制和指针网络得到增强。试验结果表明，本研究模型在自动评估方面优于基线模型。

然而，试验结果与人类预期仍存在差距。原因有二：一是本研究数据集规模不够大，如前所述，本研究从原始 RACE 数据集中过滤掉 52.5% 的完形填空题来构建的 RACE4QG，因此，未来的工作考虑将完形填空式问题转变为标准问题，这将使数据集 RACE4QG的规模翻倍；二是虽然本研究提出的问题生成模型在数据集 RACE4QG 上取得了较好结果，并为进一步的研究提供了坚实的基础，但是这项具有挑战性的任务仍有很大的改进空间，未来将探索更先进的图结构来编码输入的文章，以进一步生成更好的问题。

参考文献：

[1]　PEARSON P D. Handbook of research on reading comprehension[M]. London: Routledge, 2014: 27−55.

[2] 崔铁军, 李莎莎. 人和人工智能系统的概念形成过程研究 [J]. 智能系统学报, 2022, 17(5): 1012–1020.
CUI Tiejun, LI Shasha. Concept formation process of human and artificial intelligence systems[J]. CAAI transactions on intelligent systems, 2022, 17(5): 1012–1020.

[3] ZHOU Qingyu, YANG Nan, WEI Furu, et al. Neural question generation from text: a preliminary study[C]// HUANG X, JIANG J, ZHAO D, et al. National CCF Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2018: 662−671.

[4] ZHAO Yao, NI Xiaochuan, DING Yuanyuan, et al. Paragraph-level neural question generation with maxout pointer and gated self-attention networks[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 3901−3910.

[5] RAJPURKAR P, ZHANG Jian, LOPYREV K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 2383−2392.

[6] KOČISKÝ T, SCHWARZ J, BLUNSOM P, et al. The Narrative QA reading comprehension challenge[J]. Transactions of the association for computational linguistics, 2018, 6: 317–328.

[7] YANG Zhilin, QI Peng, ZHANG Saizheng, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 2369−2380.

[8] LAI Guokun, XIE Qizhe, LIU Hanxiao, et al. RACE: large-scale ReAding comprehension dataset from examinations[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 785−794.

[9] 李倩玉, 王蓓, 金晶, 等. 基于双向 LSTM 卷积网络与注意力机制的自动睡眠分期模型 [J]. 智能系统学报, 2022, 17(3): 523–530.
LI Qianyu, WANG Bei, JIN Jing, et al. Automatic sleep staging model based on the bi-directional LSTM convolutional network and attention mechanism[J]. CAAI transactions on intelligent systems, 2022, 17(3): 523–530.

[10] DU Xinya, SHAO Junru, CARDIE C. Learning to ask: neural question generation for reading comprehension[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2017: 1342−1352.

[11] 刘明, 张津旭, 吴忠明. 智能提问技术及其教育应用 [J]. 人工智能, 2022, 9(2): 30–38.
LIU Ming, ZHANG Jinxu, WU Zhongming. Intelligent questioning technology and its educational application[J]. AI-View, 2022, 9(2): 30–38.

[12] RUS V, PIWEK P, STOYANCHEV S, et al. Question generation shared task and evaluation challenge: status report[C]//Proceedings of the 13th European Workshop on Natural Language Generation. New York: ACM, 2011: 318−320.

[13] DHOLE K, MANNING 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. Stroudsburg: Association for Computational Linguistics, 2020: 752−765.

[14] HEILMAN M, SMITH N A. Good question! Statistical ranking for question generation[C]//HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. New York: ACM, 2010: 609−617.

[15] YAO Xuchen, BOUMA G, ZHANG Yi. Semantics-based question generation and implementation[J]. Dialogue & discourse, 2012, 3(2): 11–42.

[16] LABUTOV I, BASU S, VANDERWENDE L. Deep questions without deep understanding[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2015: 889−898.

[17] SONG Linfeng, WANG Zhiguo, HAMZA W. A unified query-based generative model for question generation and question answering[EB/OL]. (2017−09−04)[2020−01−01]. https://arxiv.org/abs/1709.01058.pdf.

[18] YUAN Wei, HE Tieke, DAI Xinyu. Improving neural question generation using deep linguistic representation[C]// Proceedings of the Web Conference 2021. Ljubljana, Slovenia. New York: ACM, 2021: 3489−3500.

[19] JIA Xin, ZHOU Wenjie, SUN Xu, et al. EQG-RACE: examination-type question generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2021: 13143−13151.

[20] 李武波, 张蕾, 舒鑫. 基于 Seq2Seq 的生成式自动问答系统应用与研究 [J]. 现代计算机, 2017(36): 57–60.
LI Wubo, ZHANG Lei, SHU Xin. Application and research on generative automatic question answering system based on Seq2Seq[J]. Modern computer, 2017(36):

57–60.

[21] 李景聪, 潘伟健, 林镇远, 等. 采用多路图注意力网络的情绪脑电信号识别方法 [J]. 智能系统学报, 2022, 17(3): 531–539.
LI Jingcong, PAN Weijian, LIN Zhenyuan, et al. Emotional EEG signal recognition method using multi-path graph attention network[J]. CAAI transactions on intelligent systems, 2022, 17(3): 531–539.

[22] ZHANG Yuhao, QI Peng, MANNING C D. Graph convolution over pruned dependency trees improves relation extraction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 2205−2215.

[23] MANNING C, SURDEANU M, BAUER J, et al. The stanford CoreNLP natural language processing toolkit[C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2014: 55−60.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000−6010.

[25] LUONG T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1412−1421.

[26] SEE A, LIU P J, MANNING C D. Get to the point: summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2017: 1073−1083.

[27] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1532−1543.

[28] LIU Bingran. Neural question generation based on Seq2Seq[C]//Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence. New York: ACM, 2020: 119−123.

[29] SUN Xingwu, LIU Jing, LYU Yajuan, et al. Answer-focused and position-aware neural question generation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 3930−3939.

[30] LOPEZ L E, CRUZ D K, CRUZ J C B, et al. Simplifying paragraph-level question generation via transformer language models[C]//Pham DN, Theeramunkong T, Governatori G, et al. Pacific Rim International Conference on Artificial Intelligence. Cham: Springer, 2021: 323−334.

[31] ZHANG Shiyue, BANSAL M. Addressing semantic drift in question generation for semi-supervised question answering[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2019.

作者简介：

徐坚，教授，主要研究方向为机器学习、自然语言处理、智慧教育。出版著作 6 部，发表学术论文 48 篇。E-mail：qjncxj@126.com。