doi：10.13413/j.cnki.jdxblxb.2022422

基于对比学习思想的多跳问题生成

王红斌[1,2,3]，杨何祯旻[1,2,3]，王灿宇[4]

(1. 昆明理工大学 信息工程与自动化学院，昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，昆明 650500；
3. 昆明理工大学 云南省计算机技术应用重点实验室，昆明 650500；
4. 云南农业大学 大数据学院，昆明 650201)

摘要：针对获取大规模的多跳问答训练数据集耗时耗力的问题，提出一个基于对比学习思想的多跳问题生成模型. 模型分为生成阶段和对比学习打分阶段，生成阶段通过执行推理图生成候选多跳问题，对比学习打分阶段通过一个基于对比学习思想的无参考问题的候选问题打分模型对候选问题进行打分排序，并选择最优的候选问题. 该模型在一定程度上缩小了无监督方法与人工标注方法的差距，有效缓解了缺少多跳问答数据集的问题. 在数据集HotpotQA上的实验结果表明，基于对比学习的多跳问题生成模型能有效扩充训练数据，极大减少了人工标注数据的成本.

关键词：多跳问题生成；机器阅读理解；对比学习

中图分类号：TP391 **文献标志码：**A **文章编号：**1671-5489(2023)05-1103-09

Multi-hop Question Generation Based on Contrastive Learning Ideas

WANG Hongbin[1,2,3], YANG Hezhenmin[1,2,3], WANG Canyu[4]

(1. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500, *China*; 2. *Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming* 650500, *China*; 3. *Yunnan Key Laboratory of Computer Technology Application, Kunming University of Science and Technology, Kunming* 650500, *China*; 4. *Faculty of Big Data, Yunnan Agricultural University, Kunming* 650201, *China*)

Abstract: Aiming at the time-consuming and labor-intensive problem of obtaining large-scale multi-hop question and answer training dataset, we proposed a multi-hop question generation model based on the contrastive learning idea. The model was divided into the generation phase and the contrastive learning scoring phase. In the generation phase, candidate multi-hop questions were generated by executing the inference graph. In the contrastive learning scoring phase, candidate questions were scored and sorted through a candidate question scoring model without reference question based on the contrastive learning idea, and the best candidate question was selected. This model had to some extent narrowed the gap between unsupervised methods and manual annotation methods, effectively alleviating the problem of lacking a multi-hop question and answer dataset. The experimental results

收稿日期：2022-10-24.

第一作者简介：王红斌(1983—)，男，汉族，博士，教授，从事自然语言处理、信息检索和机器学习的研究，E-mail：whbin2007@126.com. **通信作者简介**：王灿宇(1979—)，男，白族，硕士，副教授，从事数据分析的研究，E-mail：736559039@qq.com.

基金项目：国家自然科学基金(批准号：61966020)和云南省基础研究计划面上项目(批准号：CB22052C143A).

on HotpotQA dataset show that the multi-hop question generation model based on contrastive learning can effectively expand the training data and greatly reduce the cost of manually labeling data.
Keywords: multi-hop question generation; machine reading comprehension; contrastive learning

问题生成(question generation, QG)是与机器阅读理解(machine reading comprehension, MRC)相关的一个自然语言处理任务[1-2]. 问题生成任务定义为:给定一个文本描述和答案,通过文本描述和答案生成内容相关、语序通顺的问题. 问题生成可分为单跳问题生成和多跳问题生成. 单跳问题的答案出现在单一文档中,且这类问题多数可通过对问题和单一文档使用关键词匹配的方式进行回答,目前主流的阅读理解数据集 SQuAD 所提供的问题都是单跳问题. 使用单跳问答数据集的训练模型虽然在单文本阅读理解任务上表现较好,但其无法评估需要多步推理能力的多文本阅读理解任务. 因此,文献[3]构建了面向自然语言和多步推理问题的数据集 HotpotQA,该数据集是一个多文档、多跳推理、标注支持事实的同构数据集. 面向自然语言和多步推理问题的数据集 HotpotQA 提供了大量的多跳问答数据,多跳问题的回答则需要定位多个信息来源并建模信息间的关系,基于这些信息链进行多步推理和理解,更接近于人类的思维. 因此,本文主要考虑多跳问题生成研究.

现有的神经网络模型在单跳和多跳问答数据集上都表现较好[4-24],但这些方法严重依赖于大规模的人工标注. 尽管当前有监督学习技术表现较好,但人工标注大规模的多跳问答训练数据集成本过高. Pan 等[4]提出的无监督多跳问题生成器 MQA-QG 作为第一个研究无监督多跳问题生成的模型,初步实现了采用自动生成的多跳问题取代人工标注的多跳问题,但其效果仍与人工标注的数据集有一定差距,使用多跳问题生成器 MQA-QG 自动生成的数据训练问答 QA 模型的结果与使用人工标注的数据相比,EM 值和 F_1 值分别相差 14.5 和 14.2,因此还有较大的提升空间. 同时,多跳问题生成器 MQA-QG 将两篇关联的文本分别生成单跳问题,再将单跳问题融合生成多个候选多跳问题,该方法虽然无需标注支持句,但其生成的候选多跳问题不灵活且质量参差不齐没有进行有效筛选.

为解决该问题,本文在生成任务中引入对比学习的思想,提出基于对比学习思想的多跳问题生成方法,通过将生成的候选问题与原文本进行对比,并引入与参考问题的对比损失,在一定程度上提高了生成问题的质量且无需标注支持句. 同时对生成的多跳问题集进行数据增强,扩充后的训练集放入 QA 模型训练能得到更好的训练效果. 将本文方法在多文本同构数据集 HotpotQA 上进行实验,实验结果表明,所提出的基于对比学习思想的多跳问题生成方法不需要对支持句进行标注,只需要一定的参考问题和对应的文档数据进行训练,成功生成并有效筛选了质量较高的多跳问题集,扩充了原始训练集,极大减少了人工标注数据的需求,在部分机器阅读理解任务上取得了一定的性能提升.

1 相关工作

1.1 基于规则方法的问题生成

早期的问题生成研究主要采用规则方法,基于规则的问题生成系统[5]需要人工设计将陈述句转换为疑问句的复杂规则,如句法树转换模板,这些方法成功的关键取决于将陈述句转换到疑问句的规则设计是否足够好,而转化规则通常需要设计者具有深厚的语言知识. 为改进纯基于规则的系统,Heilman 等[6]提出可以使系统生成过量的问题,然后采用基于监督学习的排序算法对问题质量进行排序,选出排序最好的;而监督算法又要求人工设计复杂的特征集,即使设计出来,系统所生成的问题也与文档有较多重叠,导致生成的问题容易回答. 早期基于规则的方法初步证明了自然语言处理技术可以帮助生成问题从而减少人工劳动. 但由于语言本身的复杂性,人工发现和归纳出所有的问题规则几乎是不可能的,而且规则方法难以扩展,为某个领域制定的规则通常很难在其他领域快速移植.

1.2 基于深度神经网络的问题生成

随着深度学习技术的发展,基于深度学习的问题生成研究取得了许多成果. 第一种方法是采用序列到序列模型实现问题生成. Bahdanau 等[7]提出了一个采用双向循环神经网络 BiRNN 的序列到序列模型,首次将注意力机制引入到自然语言处理(NLP)领域. Gulcehre 等[8]和 Gu 等[9]引入了复制机制,

将文本中的内容复制到输出问题中，强化了输入与输出之间的关联. 融入了复制机制的序列到序列模型，采用双向长短时记忆网络(BiLSTM)编码器编码段落，采用融入复制机制的注意力双向长短期记忆(LSTM)解码器生成问题. Du 等[10]改进了序列到序列的模型，编码端采用了句子级和篇章级的基于注意力机制的双向 LSTM 编码器，解码端采用了 LSTM. Song 等[11]通过在编码端加入多视角匹配机制，考虑了答案位置信息与所有文本词的匹配信息，丰富的匹配信息可以引导解码器生成更准确的问题. 第二种方法是基于 Transformer 模型的问题生成研究，Transformer 结构的提出使神经网络机器翻译任务取得了重大突破，性能获得很大提升. Scialom 等[12]提出了使用 Transformer 模型进行答案无关的文本问题生成任务，研究人员将 Transformer 结构同样应用到了问题生成任务中，仍可以与复制机制结合，从而提高文本问题生成任务的性能. 第三种方法是基于预训练技术的问题生成研究. 该技术从大量的语料中学习自然语言知识，从而提高下游任务的表现. 其中较常用的是预训练模型 BERT[13]，该模型提出了掩码训练的方式训练语言模型. 在下游任务中，通过对语言模型的微调即可获得较好的表现. 后续的研究工作探索了多种不同的预训练模型，并将其用于问题生成任务中，使该任务的性能得到进一步提升. 这类方法的模式为“预训练-微调”，且一次预训练就能利用大量的语料，微调过程时间更短，其在文本问题生成领域具有广阔的应用前景.

1.3　多跳问题生成

近年来，多跳问题生成研究成为问题生成领域的一个研究热点. Pan 等[4]提出了一个无监督的多跳问答框架 MQA-QG，是一个多跳问题生成器，使用多跳问题生成器 MQA-QG 可以从同构或异构的数据集中生成接近人类思维的多跳问答训练集，再用生成的训练集去训练多跳 QA 模型可得到更好的结果，并且极大减少了人工标注数据的需求. 同样作为问答的数据源，知识图谱与无结构的纯文本数据源相比，其优势主要体现在数据的结构化、精度、关联度等方面，结构化的知识图谱以一种更清晰、更准确的方式表示人类知识，对于生成更契合实际应用中复杂场景的复杂多跳问题效果更佳. Yu 等[14]将知识图谱和知识推理融入多跳问题生成研究中，生成的多跳问题需要多个句子或段落的证据进行推理才能回答，缓解了多跳数据短缺的问题，并提升了多跳机器阅读理解的性能. Fei 等[15]提出了一个简单有效的可控生成框架 CQG，保证了问题的复杂性和质量，同时引入了一种新的基于 Transformer 的可控解码器，以保证关键实体出现在问题中. Su 等[16]提出的 QA4QG 使用 Transformer 结构替代了图结构，同时考虑了问答任务对问题生成任务的帮助. Yu 等[17]提出了相似的输入文本表达结构，其问题形式也相似，即两个样本在文本上共享一个近似表达结构的观点，采用元学习提出了一个自适应的多跳问题生成框架. Su 等[18]提出了一个问题生成的多跳编码融合网络 MulQG，其通过图卷积网络(GCN)在多跳中进行上下文编码，并通过编码器推理门进行编码融合. Gupta 等[19]根据上下文中的支持事实生成相关问题，采用了多任务学习的方式，并辅以 answer-aware 支持性事实预测任务指导问题生成. Sachan 等[20]使用强 Transformer 进行多跳问题生成，同时采用基于图网络增强的方法与 Transformer 模型融合，使问题生成效果进一步提升.

本文主要考虑多跳问题生成，提出一个基于对比学习的多跳问题生成模型，该模型是一个分为生成阶段和对比学习打分阶段的生成-评估两阶段模型，生成阶段通过执行推理图生成候选多跳问题，对比学习打分阶段通过一个基于对比学习的无参考问题的候选问题打分模型对候选问题进行打分排序，并选择最优的候选问题.

2　方法设计

本文设计的模型是一个生成-评估两阶段模型. 生成阶段提取和整合多个输入源文本的信息生成多个候选多跳问题，通过执行桥式推理图和比较式推理图生成桥式候选多跳问题和比较式候选多跳问题. 对比学习打分阶段通过一个基于对比学习思想的无参考问题的候选问题打分模型对候选问题进行打分排序，并选择最优的候选问题. 模型如图 1 所示.

多跳问题生成的定义为给定多个自然语言文本 $D=\{D_1,D_2,\cdots,D_n\}$，其中 D_i 表示输入的第 i 个自然语言文本，目标是生成多跳自然语言问题 Q. 本文模型定义为给定两篇原文本$\langle \text{text}_i,\text{text}_j\rangle$和对应

的参考问题答案对$\langle \text{ref}\,Q, \text{ref}\,A\rangle$，问题生成模型$f$的目标是生成候选多跳问题$Q=f(\text{text}_i,\text{text}_j)$，使得对比学习打分模型$h$给出的分数$r=h(Q,\text{ref}\,Q)$尽可能高. 在该过程中，模型被分解成两个阶段：生成模型g和对比学习打分模型h，前者负责生成候选多跳问题集合，后者负责打分并选择最优候选多跳问题.

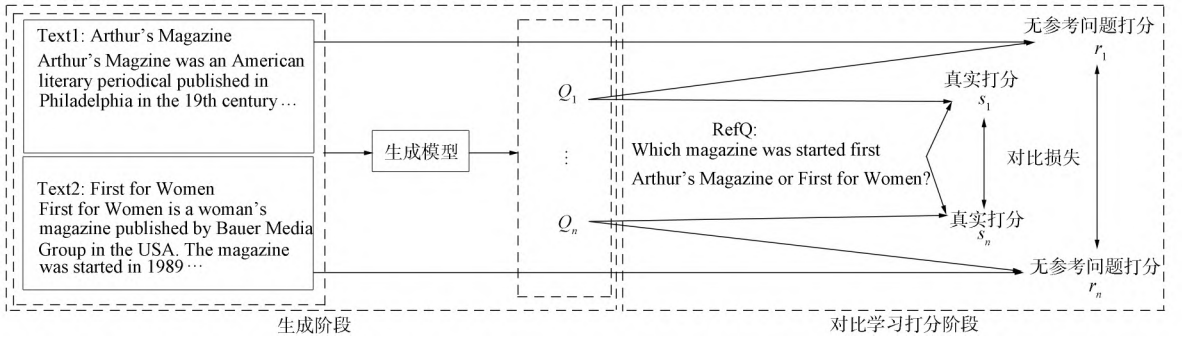

图1 基于对比学习的多跳问题生成模型

Fig.1 Multi-hop question generation model based on contrastive learning

2.1 生成阶段

生成阶段模型分为两个类型：桥式多跳问题生成和比较式多跳问题生成. 这两种类型的问题生成方法基本上涵盖了数据集HotpotQA的所有问题类型. 生成阶段从输入的两个相关文本$T=\{t_i,t_j\}$中提取、生成、融合相关信息得到候选的多跳问题集.

2.1.1 桥式多跳问题生成

桥式多跳问题生成阶段模型g_1的输入为两个具有共同实体的文本对$T=\{t_i,t_j\}$，生成的多个候选问题$\{Q_1,Q_2,\cdots,Q_n\}=g_1(T)$，生成模型方案如图2所示. 多数多跳问题会通过桥实体整合信息[25]. 图2中：FindBridge模块表示提取两个文本的共同桥实体$e=f(t_i)\bigcap f(t_j)$，其中$f(\cdot)$表示实体抽取过程，使用stanza_nlp工具包识别并提取出两篇文本共同的桥实体；QGwithAns模块表示通过第一个文本的内容t_i和答案a生成一个简单的单跳问题$q_1=QG(t_i,a)$；QGwithEnt模块表示通过第二个文本t_j和提取的桥实体e生成一个简单的单跳问题$q_2=QG(t_j,e)$，其中$QG(\cdot)$表示单跳问题生成的过程，该过程使用谷歌的T5预训练模型进行单跳问题生成[26]；QuesToSent模块表示将其中一个单跳问题q_1转换为陈述句式s[27]；BridgeBlend模块表示将转化为陈述句式的句子s与另一个单跳问题q_2融合生成一个复杂的多跳问题$Q=\text{BERT-Large}(s,q_2)$，其中BERT-Large[28]表示融合两个单跳问题的预训练模型.

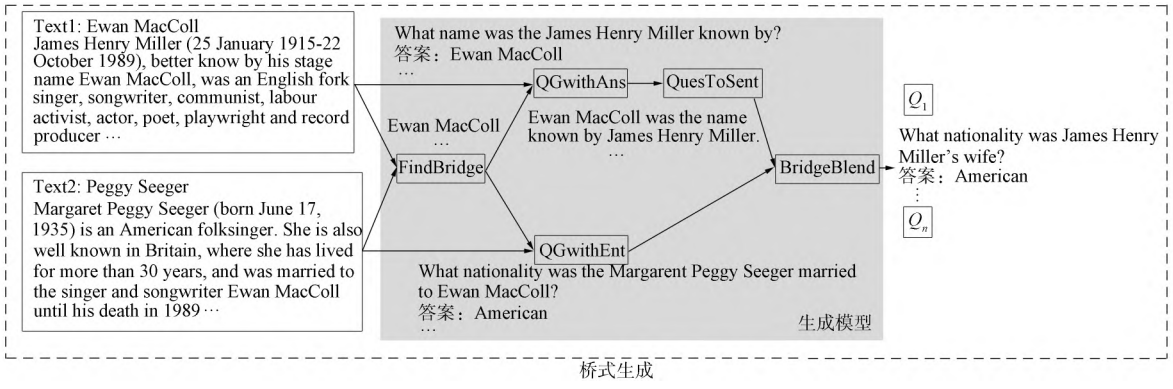

图2 桥式多跳问题生成模型

Fig.2 Multi-hop question generation model of bridge

2.1.2 比较式多跳问题生成

比较式多跳问题生成阶段模型g_2的输入为两个具有比较属性的文本对$T=\{t_i,t_j\}$，生成的多个候选问题$\{Q_1,Q_2,\cdots,Q_n\}=g_2(T)$，生成模型方案如图3所示. 图3中：FindComEnt模块表示提取两

篇文本中可以作比较的属性作为桥实体 $e_1=f(t_i)$ 和 $e_2=f(t_j)$，其中 $f(\cdot)$ 表示使用 stanza_nlp 工具包提取可能作为比较属性的实体，例如时间、地点、数字等；QGwithAns 模块表示通过第一个文本生成一个简单的单跳问题 $q_1=\mathrm{T5}(t_i)$，QGwithEnt 模块表示通过第二个文本生成一个简单的单跳问题 $q_2=\mathrm{T5}(t_j)$，QGwithAns 和 QGwithEnt 模块使用谷歌的 T5 预训练模型进行单跳问题生成；CompareBlend 模块表示将两个单跳问题 q_1 和 q_2 融合生成一个比较式多跳问题 Q，此处的两个单跳问题是关于不同的实体 e_1 和 e_2 的问题.

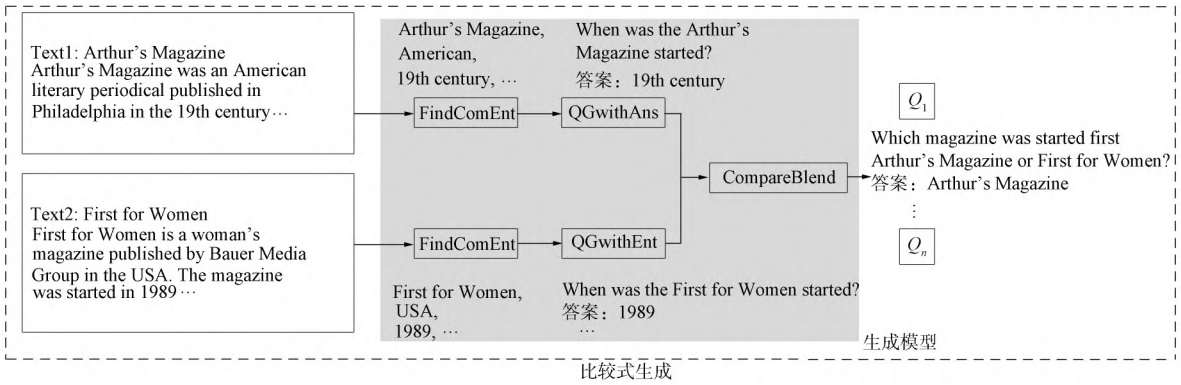

图 3　比较式多跳问题生成模型

Fig.3　Multi-hop question generation model of comparison

2.2　对比学习阶段

对比学习阶段设计为一个基于对比损失的无参考问题的候选问题打分模型 h，模型方案如图 4 所示. 输入为两篇相关的原文本对 $T=\{t_i,t_j\}$ 和参考问题 ref Q 以及上一阶段生成的候选多跳问题集 Q_i，输出为最优的候选问题 Q. 针对原文本 $T=\{t_i,t_j\}$，一个更好的候选多跳问题 Q 应该获得更高的打分，建立一个评价打分模型 h，其目标是为生成模型 g 生成的候选问题 Q_i 打分 r_i，可表示为

$$r_i=h(Q_i,T), \tag{1}$$

其中 h 表示打分模型. 模型 h 初始化为预训练模型 RoBERTa[29]，该模型首先分别对候选问题集 Q_i 和原文本 $T=\{t_i,t_j\}$ 编码；然后计算候选问题与文本内容的余弦相似度，并将其作为候选问题 Q_i 的无参考问题的打分 r_i；最后模型输出的多跳问题 Q 为获得最高分的候选多跳问题，即

$$Q=\mathrm{argmax}\ r_i, \tag{2}$$

其中 r_i 表示第 i 个候选问题的无参考问题打分. 选择打分最高的问题 Q 作为最终候选多跳问题，同时计算参考问题 ref Q 与候选问题 Q_i 的余弦相似度作为候选问题 Q_i 的真实打分 s_i，期望候选问题 Q_i 的无参考问题打分 r_i 能更接近真实打分 s_i.

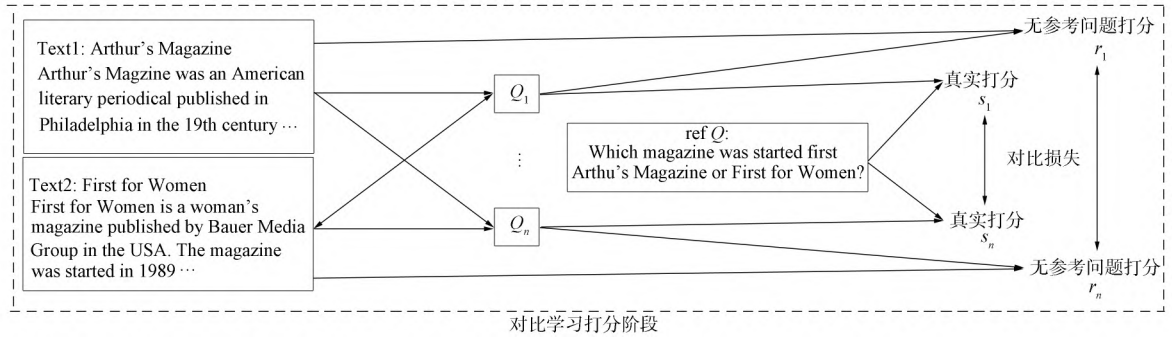

图 4　基于对比学习的无参考问题打分模型

Fig.4　Scoring model without reference question based on contrastive learning

不同于其他对比学习工作采用的明确构建正例和负例，本文模型对比学习阶段的对比性体现在参数化模型 $h(\cdot)$ 评估生成问题的不同质量. 在对比学习阶段的模型 h 中引入一个对比损失：

$$L=\sum_i \max\{0,h(\tilde{Q}_i,T)-h(\text{ref } Q,T)\}+\sum_i\sum_{j>i}\max\{0,h(\tilde{Q}_j,T)-h(\tilde{Q}_i,T)+\lambda_{ij}\},\quad(3)$$

其中 T 表示原文本，ref Q 表示参考问题，h 表示打分模型，$\tilde{Q}_1,\tilde{Q}_2,\cdots,\tilde{Q}_n$ 根据 $M(\tilde{Q}_i,\text{ref } Q)$ 降序排列，$M(\cdot)$ 采用 ROUGE 作为自动评估指标，λ 为一个超参数.

3 实 验

3.1 生成阶段

本文实验采用数据集 HotpotQA[3]，其是一个多文档、多跳推理、标注支持事实的同构数据集. 要回答多跳问题，需要定位多个信息来源并从中找到最相关的部分，同时基于这些信息进行多步推理和理解. 数据集 HotpotQA 包含 78 909 个桥式多跳问题和 18 943 个比较式多跳问题，共 97 852 个多跳问题. 其中桥式问题分出 72 991 个问题作为训练集和 5 918 个问题作为验证集，比较式问题分出 17 456 个问题作为训练集和 1 487 个问题作为验证集. 每个多跳问题都需要在两篇标注了支持事实的维基文档进行推理才能回答. 数据集 HotpotQA 标注了支持事实，本文实验使用无监督的生成方法，所以只需提取出数据集中的文本内容及对应的问答集作为训练数据，无需标注支持句.

实验采用精确匹配 EM 和 F_1 值两个评价指标. EM 测量预测结果与基本事实之间精确字符串匹配的百分数，F_1 是精度和召回率的协调平均值，即

$$F_1=\frac{\text{精度}\times\text{召回率}\times 2}{\text{精度}+\text{召回率}}.$$

3.2 实验设置

实验中模型分为两个阶段：生成阶段和对比学习打分阶段. 在生成阶段，执行多跳问题生成器 $g(\cdot)$ 生成两种类型的问题集 Q_{bridge} 和 $Q_{\text{comparison}}$，Q_{bridge} 只包含生成的候选桥式问题，$Q_{\text{comparison}}$ 只包含生成的候选比较式问题，将 Q_{bridge} 和 $Q_{\text{comparison}}$ 合并起来作为数据集 HotpotQA 生成的候选问题集 Q_{hotpot}. 在对比学习打分阶段，对于 3 个候选问题集 Q_{bridge}，$Q_{\text{comparison}}$，Q_{hotpot}，通过对比学习打分模型 $h(\cdot)$ 分别挑选出得分最高的问题集并通过数据增强得到 $Q_{\text{bge-select}}$，$Q_{\text{com-select}}$，$Q_{\text{hotpot-select}}$ 作为最终的训练集，实验数据的信息如下：Q_{bridge} 有 37 647 个问题，$Q_{\text{comparison}}$ 有 17 455 个问题，Q_{hotpot} 有 55 102 个问题，$Q_{\text{bge-select}}$ 有 44 316 个问题，$Q_{\text{com-select}}$ 有 32 126 个问题，$Q_{\text{hotpot-select}}$ 有 76 442 个问题.

为验证多跳问题生成能有效提高机器阅读理解的能力，对于数据集 HotpotQA，本文采用 SpanBERT[30] 作为 QA 模型验证生成的问题集 $Q_{\text{bge-select}}$，$Q_{\text{com-select}}$，$Q_{\text{hotpot-select}}$ 对机器阅读理解任务的效果，并采用精确匹配 EM 值和 F_1 值评价效果.

将本文实验模型分别与 1 个有监督基线和 4 个无监督基线模型进行对比. 1 个有监督基线模型是将完整的原始数据集 HotpotQA 放入 SpanBERT 模型中训练得到的结果[30]. 4 个无监督基线模型分别是：1) SQuAD-Transfer，是用数据集 SQuAD 训练 SpanBERT 模型，然后将其转换为多跳 QA；2) Bridge-Only，是只用原始数据集 HotpotQA 中的桥式数据训练 SpanBERT 模型；3) Comparison-Only，是只用原始数据集 HotpotQA 中的比较式数据训练 SpanBERT 模型；4) MQA-QG[4]，是使用多跳问题生成器 MQA-QG 在数据集 HotpotQA 上进行多跳问题生成，然后将生成的数据训练 SpanBERT 模型. 将原始数据集 HotpotQA 中的桥式数据、比较式数据以及全体数据分别与本文模型生成并筛选的数据 $Q_{\text{bge-select}}$，$Q_{\text{com-select}}$，$Q_{\text{hotpot-select}}$ 相结合，进行数据增强，增强后的数据训练 SpanBERT 模型与有监督基线模型进行对比. 将本文模型生成并筛选的数据 $Q_{\text{bge-select}}$，$Q_{\text{com-select}}$，$Q_{\text{hotpot-select}}$ 分别进行数据增强后训练 SpanBERT 模型，并与无监督基线模型进行对比. 实验的参数设置列于表 1.

实验设置 4 个 epoch，每个 epoch 设为 10 个 step，图 5 为数据集 Bridge，Comparison，Total 实验过程的 EM 和 F_1 值收敛曲线，横坐标的 0～9 表示为第一个 epoch，10～19 表示为第二个 epoch，20～29 表示为第三个 epoch，30～39 表示为第四个 epoch，4 个 epoch 后 EM 值和 F_1 值趋于平缓.

表 1　参数设置

Table 1　Parameter settings

参数	值	参数	值
eval_batch_size	32	attention_dropout	0.1
eval_per_epoch	10	hidden_dropout	0.1
num_train_epochs	4	hidden_size	1 024
gradient_accumulation_steps	8	max_position_embeddings	512
learning_rate	2×10^{-5}	num_attention_heads	16
vocab_size	28 996	num_hidden_layers	24

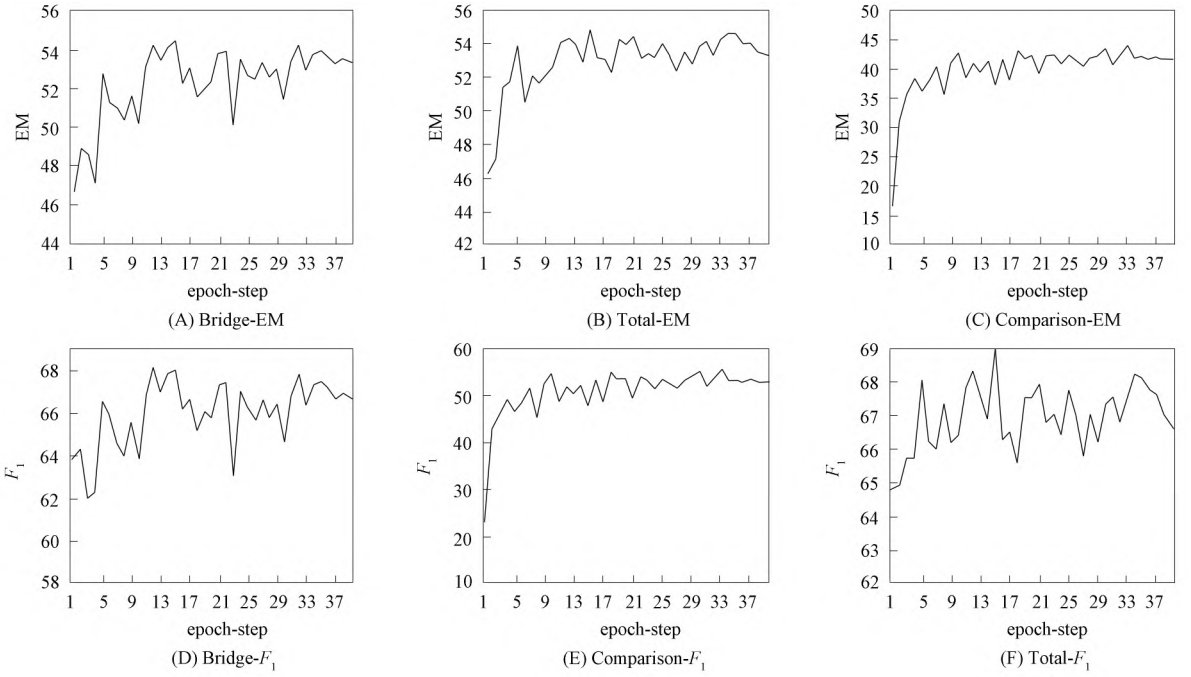

图 5　实验过程中 EM 和 F_1 值的收敛过程

Fig.5　Convergence process of EM and F_1 values during experimental process

3.3　实验结果分析

不同模型的实验结果列于表 2. 由表 2 可见，对于有监督基线 SpanBERT，本文模型的生成数据和原始数据的组合数据集对于部分机器阅读理解任务的效果有一定提升，侧面反映了模型扩充的数据集质量较好. 基于有监督的数据集 Bridge 和 Comparison 的实验结果表明，由于 Bridge 桥式数据的推理链是顺序链，需要找到共同的桥实体才能有效关联两个文本，而 Comparison 对比式数据的推理链需要比较两个实体的同一属性，因此二者对于训练集的数据量有一定要求，特别是 Comparison 对比式数据对训练集的数据量更敏感.

表 2　各模型的实验结果

Table 2　Experimental results of each model

基线类型	模型	Bridge		Comparison		Total	
		EM	F_1	EM	F_1	EM	F_1
有监督	SpanBERT[30]	68.2	83.5	74.2	80.3	69.4	82.8
	本文	65.1	78.4	53.6	67.0	69.5	82.7
无监督	Bridge-Only	55.4	71.4	12.4	19.1	46.7	60.9
	Comparison-Only	9.8	14.5	38.2	45.0	15.5	20.9
	SQuAD-Transfer	54.6	69.7	25.3	35.2	48.7	62.8
	MQA-QG[4]	56.5	72.2	48.8	54.4	54.9	68.6
	本文	54.2	68.1	43.8	55.5	54.8	69.0

对于无监督基线，本文模型的生成数据 $Q_{\text{hotpot-select}}$ 在机器阅读理解任务中 F_1 值达到了 69.0，优于

4个基线模型(SQuAD-Transfer,Bridge-Only,Comparison-Only,MQA-QG).在没有人工标注支持句的弱监督情况下,F_1值与有监督基线的差距由14.2缩短到13.8,证明了本文模型对于生成高质量多跳问题的有效性.Comparison和Total数据的F_1值分别超出MQA-QG模型0.9和0.4,而数据集Bridge的指标没有超过MQA-QG模型,说明将候选问题与文本进行对比学习的方式对于捕捉比较式数据集文档的重要实体属性语义信息更有效,进而筛选出质量高的多跳问题.

综上,针对获取大规模的多跳问答训练数据集耗时耗力的问题,本文提出了一个分为生成阶段和对比学习打分阶段的生成-评估两阶段模型,生成最优候选问题的同时通过数据增强策略,有效地通过扩充训练集取得了机器阅读理解任务上的性能提升.在多文本同构数据集HotpotQA上的实验结果表明,基于对比学习思想的无监督多跳问题生成方法对比了候选问题和源文本,生成了质量较高的无需人工标注的多跳问题集,有效扩充并筛选了原始的训练集,极大减少了人工标注数据的需求,同时在机器阅读理解任务上性能提升较好.

参 考 文 献

[1] 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.(WU Y F,ZHANG Y S.A Survey of Question Generation Research[J].Journal of Chinese Information Processing,2021,35(7):1-9.)

[2] 周青宇,周明.基于深度神经网络的文本问题生成技术综述[J].智能计算机与应用,2020,10(8):10-13.(ZHOU Q Y,ZHOU M.A Survey of Text Question Generation Techniques Based on Deep Neural Networks[J].Intelligent Computers and Applications,2020,10(8):10-13.)

[3] YANG Z L,QI P,ZHANG S Z,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.[S.l.]:ACL,2018:2369-2380.

[4] PAN L M,CHEN W H,XIONG W H,et al.Unsupervised Multi-hop Question Answering by Question Generation[C]//Proceedings of the 58th North American Chapter of the Association for Computational Linguistics.[S.l.]:ACL,2020:5866-5880.

[5] MITKOV R,HA L A.Computer-Aided Generation of Multiple-Choice Tests[C]//Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing.New York:ACM,2003:17-22.

[6] HEILMAN M,SMITH N A.Good Question! Statistical Ranking for Question Generation[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics.[S.l.]:ACL,2010:609-617.

[7] BAHDANAU D,CHO K,BENGIO Y.Neural Machine Translation by Jointly Learning to Align and Translate[EB/OL].(2014-09-01)[2022-02-10].https://arxiv.org/abs/1409.0473.

[8] GULCEHRE C,AHN S,NALLAPATI R,et al.Pointing the Unknown Words[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.[S.l.]:ACL,2016:140-149.

[9] GU J T,LU Z D,LI H,et al.Incorporating Copying Mechanism in Sequence-to-Sequence Learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.[S.l.]:ACL,2016:1631-1640.

[10] DU X Y,SHAO J R,CLAIRE C.Learning to Ask:Neural Question Generation for Reading Comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.[S.l.]:ACL,2017:1342-1352.

[11] SONG L F,WANG Z G,WAEL H,et al.Leveraging Context Information for Natural Question Generation[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics.[S.l.]:ACL,2018:569-574.

[12] 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.[S.l.]:ACL,2019:6027-6032.

[13] DONG L,YANG N,WANG W H,et al.Unified Language Model Pre-training for Natural Language Understanding and Generation[C]//Advances in Neural Information Processing Systems.New York:ACM,

2019：13042-13054.

[14] YU J X，QUAN X J，SU Q L，et al. Generating Multi-hop Reasoning Questions to Improve Machine Reading Comprehension [C]//Proceedings of the Web Conference 2020. New York：ACM，2020：281-291.

[15] FEI Z C，ZHANG Q，GUI T，et al. CQG：A Simple and Effective Controlled Generation Framework for Multi-hop Question Generation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. [S.l.]：ACL，2022：6896-6906.

[16] SU D，XU P，FUNG P. QA4QG：Using Question Answering to Constrain Multi-hop Question Generation [C]//Proceedings of the IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2022：8232-8236.

[17] YU J X，SU Q L，QUAN X J，et al. Multi-hop Reasoning Question Generation and Its Application [J]. IEEE Transactions on Knowledge and Data Engineering，2021，35(1)：725-740.

[18] SU D，XU Y，DAI W L，et al. Multi-hop Question Generation with Graph Convolutional Network [C]//Proceedings of the EMNLP. [S.l.]：ACL，2020：4636-4647.

[19] GUPTA D，CHAUHAN H，EKBAL A，et al. Reinforced Multi-task Approach for Multi-hop Question Generation [C]//Proceedings of the 28th International Conference on Computational Linguistics. [S.l.]：ACL，2020：2760-2775.

[20] SACHAN D S，WU L F，SACHAN M，et al. Stronger Transformers for Neural Multi-hop Question Generation [EB/OL]. (2020-10-22)[2022-03-12]. https://arxiv.org/abs/2010.11374.

[21] **于尊瑞，毛震东，王泉，等. 基于预训练语言模型的关键词感知问题生成** [J]. **计算机工程**，2022，48(2)：125-131. (YU Z R，MAO Z D，WANG Q，et al. Keyword Perception Question Generation Based on Pre-trained Language Model [J]. Computer Engineering，2022，48(2)：125-131.)

[22] **苏珂，黄瑞阳，张建朋，等. 多跳机器阅读理解研究进展** [J]. **计算机工程**，2021，47(9)：1-17. (SU K，HUANG R Y，ZHANG J P，et al. Research Progress in Multi-hop Machine Reading Comprehension [J]. Computer Engineering，2021，47(9)：1-17.)

[23] **舒冲，欧阳智，杜逆索，等. 基于改进图节点的图神经网络多跳阅读理解研究** [J]. **计算机工程**，2022，48(1)：99-104. (SU C，OUYANG Z，DU N S，et al. Research on Multi-hop Reading Comprehension of Graph Neural Networks Based on Improved Graph Nodes [J]. Computer Engineering，2022，48(1)：99-104.)

[24] **刘高军，李亚欣，段建勇. 基于混合注意力机制的中文机器阅读理解** [J]. **计算机工程**，2022，48(10)：67-72. (LIU G J，LI Y X，DUAN J Y. Chinese Machine Reading Comprehension Based on Mixed Attention Mechanism [J]. Computer Engineering，2022，48(10)：67-72.)

[25] XIONG W H，YU M，GUO X X，et al. Simple Yet Effective Bridge Reasoning for Open-Domain Multi-hop Question Answering [C]//Proceedings of the 2nd Workshop on Machine Reading for Question Answering. [S.l.]：ACL，2019：48-52.

[26] RAFFEL C，SHAZEER N，ROBERTS A，et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [EB/OL]. (2020-07-28)[2022-04-10]. https://arxiv.org/abs/1910.10863.

[27] DEMSZKY D，GUU K，LIANG P. Transforming Question Answering Datasets into Natural Language Inference Datasets [EB/OL]. (2018-09-11)[2022-05-15]. https://arxiv.org/abs/1809.02922.

[28] JACOB D，CHANG M W，KENTON L，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. [S.l.]：ACL，2019：4171-4186.

[29] LIU Y H，OTT M，GOYAL N，et al. Roberta：A Robustly Optimized Bert Pretraining Approach [EB/OL]. (2019-07-26)[2022-04-20]. https://arxiv.org/abs/1907.11692.

[30] JOSHI M，CHEN D，LIU Y，et al. Spanbert：Improving Pre-training by Representing and Predicting Spans [C]//Transactions of the Association for Computational Linguistics. [S.l.]：ACL，2020：64-77.

（责任编辑：韩　啸）