面向知识图谱的二阶段复杂问句生成框架

张 琨[1,2],王元卓[1,3*],仇韫琦[1,2],白 龙[2,4],江旭晖[1,2],
侯 坤[5],岑建何[3],沈华伟[1,2],程学旗[4]

(1. 中国科学院计算技术研究所数据智能系统研究中心,北京 100190;2. 中国科学院大学计算机科学与技术学院,北京 101408;
3. 中科大数据研究院,河南郑州 450046;4. 中国科学院计算技术研究所网络数据科学与技术重点实验室,北京 100190;
5. 北京工商大学计算机与人工智能学院,北京 100048)

摘 要: 面向知识图谱的问句生成(Question Generation over Knowledge Graph,KGQG)任务是根据知识图谱(Knowledge Graph,KG)子图生成自然语言问句. 现有方法通常是直接将实例化的KG子图转换为问句,并且大多采用教师强制(Teacher-Forcing)的训练策略. 然而,当前方法仍然面临两个主要挑战:(1)实例化的KG子图缺乏确定性查询意图的整合,导致输入与目标输出之间存在语义歧义现象;(2)采用教师强制训练策略训练的生成模型在推理阶段存在曝光偏差问题. 为了缓解语义歧义带来的挑战,本文提出了一个复杂问句生成框架,其包括两个阶段,即事实-查询和查询-问句生成阶段. 在第一阶段,本文设计了一个查询图生成器,将KG子图转换为具有不同查询意图的查询图. 在第二阶段,本文提出了一个问句生成模型,该模型利用密集连接图卷积网络(Densely Connected Graph Convolutional Network,DCGCN)对查询图进行编码,并利用双向自回归变换器(Bidirectional and Auto-Regressive Transformers,BART)模型进行解码以生成问句. 此外,为了减轻曝光偏差问题,本文引入了生成对抗模仿学习对问句生成模型进行训练. 其中,所采用的判别器通过模仿标记数据自适应地学习奖励函数,并指导问句生成模型探索潜在问题空间中的高奖励区域. 本文在三个广泛使用的数据集上进行了大量实验,结果表明所提出的框架具有显著的有效性.

关键词: 问句生成(KGQG);知识图谱(KG);文本生成;曝光偏差;生成对抗模仿学习

基金项目: 河南省重点研发专项(No.241111211900);国家自然科学基金(No.62172393,No.U1836206,No.U21B2046);河南省中原人才计划(No.204200510002)

中图分类号: TP391;TP182 **文献标识码:** A **文章编号:** 0372-2112(2025)06-2104-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240331

A Two-Stage Framework for Complex Question Generation over Knowledge Graph

ZHANG Kun[1,2], WANG Yuan-zhuo[1,3*], QIU Yun-qi[1,2], BAI Long[2,4], JIANG Xu-hui[1,2],
HOU Kun[5], CEN Jian-he[3], SHEN Hua-wei[1,2], CHENG Xue-qi[4]

(1. *Research Center for Data Intelligence Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
2. *School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 101408, *China*;
3. *Big Data Academy, Zhongke, Zhengzhou, Henan* 450046, *China*;
4. *CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
5. *School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing* 100048, *China*)

Abstract: Question generation over knowledge graph (KGQG) aims to generate natural language questions from knowledge graph (KG) facts automatically. Existing methods directly transform an instantiated KG subgraph into a question and usually adopt the teacher-forcing training strategy. However, the current methods still face two major challenges: (1) instantiated KG subgraphs lack the integration of deterministic query intention, resulting in a semantic mismatch between the input and the target output; (2) the teacher-forcing training strategy suffers from exposure bias in the inference stage. To address the challenges posed by semantic ambiguity, this paper proposes a framework for complex question generation consist-

收稿日期:2024-04-15;修回日期:2024-11-15;责任编辑:朱梅玉
*通讯作者:王元卓

ing of two stages, namely, facts-to-query and query-to-question. In the first stage, this paper designs a query graph generator, which converts KG subgraphs into query graphs with different query intentions. In the second stage, this paper proposes a question generation model, which employs densely connected graph convolutional networks (GCN) to encode the query graphs and utilizes the bidirectional and auto-regressive transformers (BART) model for decoding to generate questions. Moreover, to alleviate exposure bias, we train the question generator with generative adversarial imitation learning. The adopted discriminator learns reward functions self-adaptively through imitating the labeled data and guides the question generator to explore the high-reward area of the potential question space. Extensive experiments conducted on three widely-used datasets demonstrate the significant effectiveness of the proposed framework.

Key words: question generation over knowledge graph (KGQG); knowledge graph (KG); text generation; exposure bias; generative adversarial imitation learning

Foundation Item(s): Key Research and Development Program of Henan Province (No. 241111211900); National Natural Science Foundation of China (No. 62172393, No. U1836206, No. U21B2046); Henan Province Zhongyuan Talent Plan (No.204200510002)

1　引言

面向知识图谱的问句生成(Question Generation over Knowledge Graph, KGQG)任务是根据知识图谱(Knowledge Graph, KG)的子图生成自然语言问句,该任务可应用于诸多领域,如教育、情报等. 此外,该任务还可通过生成大量训练语料为问答任务进行数据增强,从而提升问答系统的性能.

当前问句生成方法[1~5]直接将实例化的KG子图作为输入,并采用教师强制(Teacher-Forcing)训练策略,利用编码器-解码器模型生成问句. 然而,这些方法适用于仅依赖一个三元组(主体实体,关系,客体实体)来获取答案的简单问句. 对于涉及多个关系或函数约束的复杂问句,例如比较约束和排序约束类型的问句,当前方法的生成效果则较为有限. 具体而言,本文将复杂问句分为五类:多跳、类型约束、实体约束、比较约束和排序约束类型问句. 在生成复杂问句时,现有模型面临以下两个主要挑战:

挑战一:语义歧义. 现有方法直接将实例化的KG子图输入模型,但这些子图缺乏明确的查询意图,导致输入(即KG子图)与目标输出(即问句)之间存在语义歧义问题. 如图1中的第一个案例所示,现有模型仅将中间节点"叶莉"作为输入,生成了一个单跳的简单问句,而我们的目标是一个多跳问句,如基准问句所示. 同时,在第二和第三个案例中,现有模型在生成带函数约束的复杂问句时表现受限,因为KG子图无法包含此类函数语义信息.

挑战二:曝光偏差. 现有方法主要使用一种称为教师强制(Teacher-Forcing)的策略进行训练,该策略提升了样本的训练效率并保证了训练的稳定性. 然而,这种策略存在曝光偏差(Exposure Bias)[6]问题,即在训练过程中使用了真实的前缀序列,而在推理过程中,无论预测是否正确,都只能使用生成的前缀序列,如果前缀序列中存在生成错误的词,那么就会导致后续生成的序列完全错误,产生错误累积的现象.

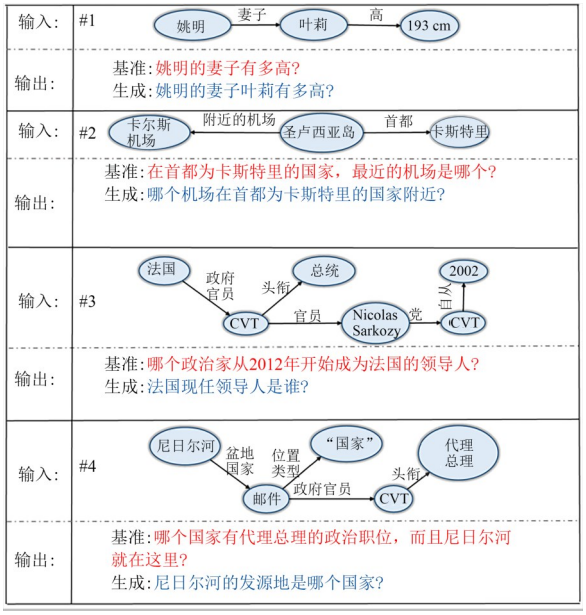

图1　当前复杂问句生成方法面临的挑战示例

为了解决上述挑战,本文提出了一个面向KG的复杂问句生成的两阶段框架,即FaQ2. 在第一阶段,即事实-查询生成阶段,本文设计了一个查询图生成器,将KG子图转换为具有不同查询意图的查询图. 具体而言,首先,本文设计了一组操作符,根据给定的问句类型从KG中检索、添加或修剪相关信息. 然后,本文基于这些操作符构建了六个意图驱动的推理流程,用以生成不同的查询图,每个查询图对应一种类型的复杂问句. 在第二阶段,即查询-问句生成阶段,本文提出了一个问句生成模型,该模型利用密集连接图卷积网络(Densely Connected Graph Convolutional Network, DCGCN)[7]编码查询图,并利用双向自回归变换器(Bidirectional and Auto-Regressive Transformers, BART)[8]根据编码信息解码生成问句.

此外,为了缓解曝光偏差,本文利用生成对抗模仿学习[9]对问句生成器进行训练. 除了教师强制训练相关的损失函数外,问句生成器还从生成对抗模仿学习中的判别器接收奖励,该判别器通过模仿标记数据自适应地学习奖励函数,并引导问句生成器探索高奖励区域.

为了评估所提出的框架的有效性,本文在三个广泛使用的数据集上进行了大量实验. 实验结果表明:本文的方法在自动评估指标上优于所有采用的基线模型,并在手动评估中展现了比基线模型更强的生成能力.

总的来说,本文的主要贡献如下:

(1)本文提出了一个面向KG的复杂问句生成的两阶段框架FaQ2,并设计了一个查询图生成器来克服语义歧义的挑战.

(2)本文采用生成对抗模仿学习对基于编码器-解码器的问句生成器进行训练,从而缓解生成过程中的曝光偏差问题.

(3)本文在三个基准数据集上进行了大量实验,证明了所提出的框架的有效性,其性能显著优于基线模型.

2 任务描述

本文的任务是给定需要生成的问句类型(查询意图类型),根据结构化的KG子图生成自然语言问句. 针对此任务,本文提出了一个复杂问句生成的两阶段框架. 在第一阶段,即事实-查询生成阶段,本文从KG中抽样得到一个子图K,并根据问句类型(查询意图类型)T将K转换为查询图G. 在第二阶段,即查询-问句生成阶段,本文将查询图G转化为相应的复杂问句Q.

KG子图K是以(s,r,o)形式的三元组集合,其中s、r和o分别表示主体实体、关系和客体实体.

查询图如文献[10,11]所述,查询图G是图表示中的λ演算的受限子集. 因此,它可以转换为可执行的查询语言,例如SPARQL(SPARQL protocol and RDF query language)[12]. 本文的查询图包括三种类型的节点:常量节点、变量节点和答案节点. 常量节点可以是已确定的节点或KG中的实体类型,例如“姚明”. 变量节点和答案节点表示未确定的KG节点或值. 查询图有两种类型的边:关系边和函数边. 关系边表示KG关系,例如“夫妻关系”. 函数边表示部分函数操作,例如比较操作(如“>”)、取最小值操作(如“MinatN”)[13~15].

一个自然语言复杂问句Q对应于KG子图K上的一个查询图G,并涉及多个关系边或函数边. 本文根据问句所对应的查询意图类型来对复杂问句进行划分,主要包括多跳、比较约束、类型约束、实体约束和排序约束. 第3.1.2节中对不同类型的复杂问句进行了说明,每种问句类型对应一类查询意图.

3 二阶段问句生成框架

本节将详细描述FaQ2框架. 如图2所示,该框架包括两个阶段:事实-查询生成阶段和查询-问句生成阶段. 在事实-查询生成阶段,本文提出了一个查询图生成模型,该模型根据指定的查询意图(问句类型)将KG子图转换为相应的查询图. 在查询-问句生成阶段,本文设计了一种基于生成对抗模仿学习的问句生成模型,该模型将查询图转换为自然语言问句. 本文中所提到的复杂问句是指查询图涉及多个关系边或函数边. 由于KG子图存在泄漏节点信息且无法蕴含函数信息的问题,所以事实-查询生成阶段生成指定的查询图,而查询-问句生成阶段则生成对应的自然语言问句.

3.1 第一阶段:事实-查询生成模型

在这个阶段,查询图生成器通过执行与指定问句类型(查询意图)对应的推理流程,将KG子图转换为相应的查询图. 具体而言,本文定义了六种不同的推理流程,每种推理流程由一系列操作符组成. 这些操作符通过启发式规则或基于现有的模型实现,用于对输入的KG子图进行检索、添加或修剪相关信息. 本节将首先介绍这些操作符,然后描述意图驱动的推理流程.

3.1.1 操作符定义

本文定义了七种不同的操作符,具体如下所示:

操作一:挑选非复合值类型(Compound Value Type,CVT)节点. 在KG子图中的节点可以分为CVT节点和非CVT节点两类. CVT节点在Freebase中用于表示数据之间的关系,其中每个条目由多个字段组成,它本身并不表示任何真实值. 而非CVT节点则相反,它表示真实的实体或者值的信息,例如泰勒·斯威夫特. 这个操作符检索KG子图并从中选择一个非CVT节点.

操作二:挑选CVT类型节点. 这个操作符与上面类似,不同之处在于它从KG子图中选择一个CVT节点.

操作三:挑选比较约束节点. 在生成比较约束类型问题时,本文需要找到可以进行比较约束的节点,例如数值节点“12”或日期节点“1980-09-13”等. 该操作从KG子图中识别并提取这些节点.

操作四:挑选类型约束节点. 这个操作符找到表示实体类型的节点,例如“国家”“省市”.

操作五:添加约束函数. 针对问题类型T,该操作符将相应的约束函数应用于KG子图. 例如,对于比较约束类型问句,在选择了比较约束节点和相应的边之后,如“美国,建国日期,1776-07-06”,本文会将比较约束函数添加到日期节点“1776-07-06”上.

操作六:添加排序函数. 在通过之前的操作符选择

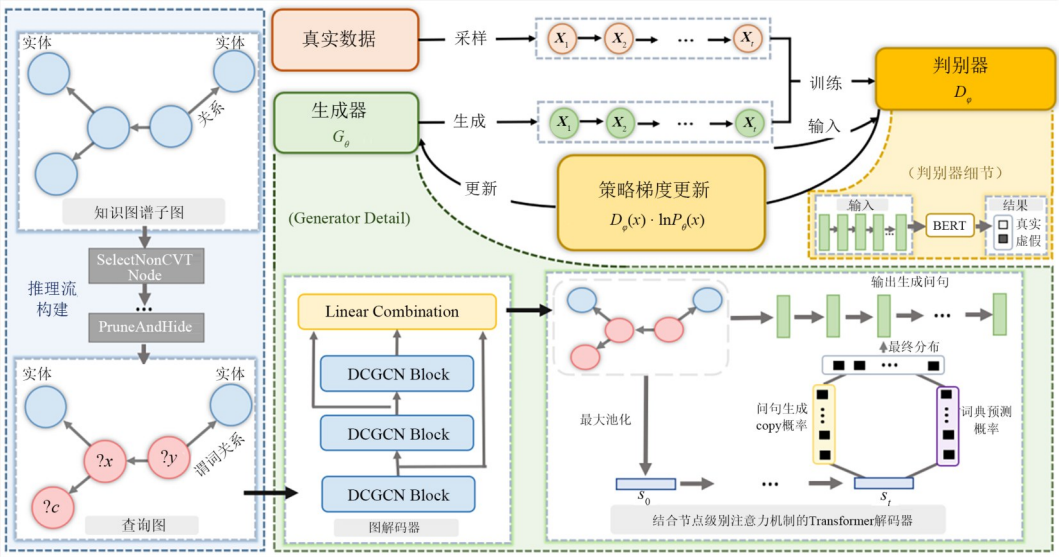

图2　FaQ2的整体框架图

了序数节点和排序函数的边之后,该操作符将相应的排序函数应用于KG子图. 例如,对于子图“Nepal国家代码977”,本文将排序函数应用于节点977上,增加一条排序函数相关的三元组,即“977,MaxatN,? var”. 其中,“? var”表示节点977在排序中的索引.

操作七:修剪并隐藏函数.

最后,本文需要简化KG子图并将其转换为查询图. 该操作符包含两个动作:修剪冗余边和隐藏不必要的节点.

3.1.2　意图驱动的推理流程

基于上述提到的操作符,本文定义了六种类型的推理流程来生成具有对应问句类型的查询图. 每个推理流程$P=n_1,n_2,\cdots$由一系列操作符组成. 每个操作符n_i的输出作为后继操作符n_{i+1}的输入. 每种推理流程的详细信息如图3所示. 本文根据问句类型将每种推理流程命名如下:

单跳类型:该推理流程将KG子图转换为单跳类型的查询图,主要由操作一、操作二和操作七组成;

多跳类型:该推理流程将KG子图转换为多跳类型的查询图,主要由操作一、操作二和操作七组成;

比较约束类型:该推理流程将KG子图转换为比较约束类型的查询图,主要由操作一、操作三、操作五和操作七组成;

类型约束类型:该推理流程将KG子图转换为包含类型约束的查询图,主要由操作一、操作四、操作五和操作七组成;

实体约束类型:推理流程将KG子图转换为包含实体约束的查询,主要由操作一、操作五和操作七组成;

排序约束类型:推理流程将KG子图转换为包含序数约束的查询,主要由操作一、操作三、操作六和操作七组成.

综上,整体的查询图生成算法流程伪代码如算法1所示,其中,SelectReasoningFlow代表根据一个问句类型选择一个对应的推理流程,ExecuteOperator根据对应的操作符类型对查询图进行改动,更新查询图.

3.2　第二阶段:查询-问句生成模型

在这个阶段,问句生成模型将上述生成的查询图转换为自然语言问句. 具体来说,它利用DCGCN将查询图编码成低维向量. 然后采用BART预训练模型将该向量解码成问句.

为了缓解曝光偏差的问题,本文使用生成对抗模仿学习(Generative Adversarial Imitation Learning,GAIL)策略训练问句生成模型. GAIL利用判别器向问句生成模型提供奖励. 判别器通过模仿学习真实问句样例的特征,自适应地学习奖励函数. 在判别器的指导下,问句生成模型能够探索潜在问句表示空间中的高奖励区域.

3.2.1　问句生成模型

基于上述构建的查询图,本文采用一种图到序列生成模型(Graph-to-sequence)来生成序列. 具体来说,本文利用DCGCN作为图编码器,利用BART作为序列解码器.

接下来,本文将分别详细描述这两个模型. 图神经网络在建模类似网络的数据方面取得了巨大成功. 然而,在使用图神经网络进行基于KG的问句生成的现有研究[13]中,存在着捕捉节点之间非局部交互结构的困难. 例如,KG子图通常包含CVT节点,这些节点用于表示多个其他节点之间的关系. 然而,这些节点在问句中永远不会出现,而生成的问句应该表达由CVT节点连接的多跳路径的含义. 本文将类似的结构称之为非局

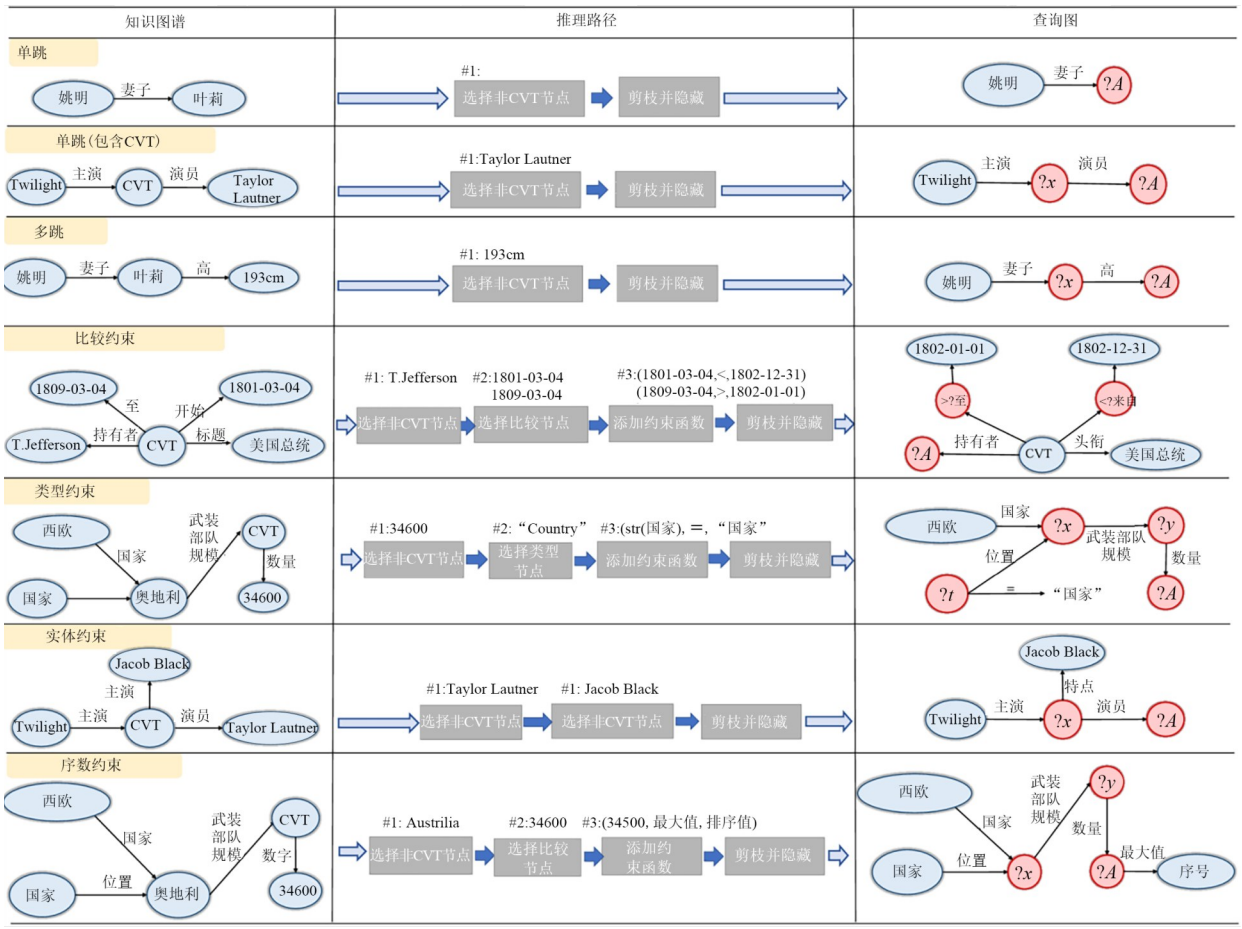

图3 面向查询图生成的六种推理流程图

算法1 查询图生成算法

输入:KG子图K,问句类型T_i

输出:查询图G

1. G←K
2. ReasoningFlow(P_i)←SelectReasoningFlow(T_i)
3. For each operator O_i in P_i do:
4. G←ExecuteOperator (O_i,G)
5. end for
6. return G

部交互.

为了更好地捕捉节点之间的非局部交互结构,本文利用DCGCN作为图编码器.具体来说,它在图卷积网络(Graph Convolutional Network,GCN)层之间应用了密集连接机制.每个DCGCN块由两个子块组成,以捕获不同抽象级别的图结构.每个子块由多个GCN层组成,其中每个GCN层与所有前面的层连接.对于节点u的第l层的输入定义为

$$\boldsymbol{g}_u^{(l)}=\left[\boldsymbol{x}_u;\ \boldsymbol{h}_u^{(1)};\ \cdots;\ \boldsymbol{h}_u^{(l-1)}\right] \tag{1}$$

其中,[·;·]表示向量的串联;$\boldsymbol{x}_u$表示节点u的节点嵌入;$\boldsymbol{h}_u^{(i)}$表示节点u在第i层的输出.

每一个GCN的卷积计算式为

$$\boldsymbol{h}_v^{(l)}=\rho\left(\sum_{u\in\mathcal{N}(v)}\boldsymbol{W}^{(l)}\boldsymbol{g}_u^{(l)}+\boldsymbol{b}^{(l)}\right) \tag{2}$$

其中,$\boldsymbol{W}^{(l)}\in\mathbb{R}^{d_{\text{hidden}}\times d^{(l)}}$表示第$l$层的权重矩阵;$d$表示节点嵌入的维度;$d_{\text{hidden}}$表示GCN层的输出维度;$d^{(l)}=d+d_{\text{hidden}}\times(l-1)$表示第$l$层的输入维度.

此外,在子块之间引入了残差连接.形式上,线性组合层的输出定义为

$$\boldsymbol{h}_{\text{comb}}=\boldsymbol{W}_{\text{comb}}\left(\boldsymbol{W}_{\text{out}}\boldsymbol{h}_{\text{out}}=\boldsymbol{X}\boldsymbol{v}\right)+\boldsymbol{b}_{\text{comb}} \tag{3}$$

其中,$\boldsymbol{h}_{\text{out}}=[h_{(1)};h_{(2)};\cdots;h_{(l)}]$;$\boldsymbol{W}_{\text{comb}}\in\mathbb{R}^{d\times d}$;$\boldsymbol{W}_{\text{out}}\in\mathbb{R}^{d\times(l\times d^{(l)})}$为权重矩阵;$\boldsymbol{b}_{\text{comb}}$为偏置向量.$\boldsymbol{W}_{\text{comb}}$、$\boldsymbol{W}_{\text{out}}$和$\boldsymbol{b}_{\text{comb}}$根据不同的DCGCN层而不同.

为了更好地利用图GCN建模边的标签信息,本文利用了Levi图变换方法将输入的查询图转换为其等价的Levi图[14].在文献[15]的基础上,本文为Levi图添加了逆向和自环边.

为了计算全图的表征,本文利用了池化的方法,该方法将输出节点表征输入到一个全连接的神经网络,并对所有节点的表征进行最大池化操作,以得到图表

示 $\boldsymbol{h}^{\mathcal{G}} \in \mathbb{R}^d$.

本文采用融合注意力机制的BART预训练模型[8]，一次生成一个单词的输出序列. 当前BART已充分应用在摘要生成、问答和长文本生成任务中[8,16]. BART采用更加多样的噪声，通过“破坏”掉有关序列结构的信息，防止模型去过度依赖序列结构信息.

在解码器部分，图嵌入 $\boldsymbol{h}^{\mathcal{G}}$ 被用作解码器的初始输入. 本文仔细遵循了文献[17]中使用的注意力机制.

正如之前的研究所验证的[5,13]，大多数问题中都包含了子图中的实体名称. 因此，本文采用复制机制[18]，直接复制实体名称，以增强生成的问句的质量. 在这里，本文只复制查询图中非变量节点的名称. 在每个时间步 t，给定解码器状态 $\boldsymbol{s}_t$、输入查询图 G 和词汇表 V，生成词 y_t 的概率计算式为

$$\begin{cases} P\left(y_t \middle| \boldsymbol{s}_t, y_{t-1}, \mathcal{G}\right) = p_{\text{genv}} P_{\text{genv}}\left(y_t \middle| \boldsymbol{s}_t, V\right) + p_{\text{cpqg}} P_{\text{cpqg}}\left(y_t \middle| \boldsymbol{s}_t, \mathcal{G}\right) \\ p_{\text{genv}}, p_{\text{cpqg}} = \text{softmax}\left(\boldsymbol{W}_{\text{copy}}\left[\boldsymbol{s}_t, \boldsymbol{y}_{t-1}\right] + \boldsymbol{b}_{\text{copy}}\right) \end{cases} \tag{4}$$

其中，genv 和 cpqg 分别表示词汇生成模式和查询图复制模式；$\boldsymbol{W}_{\text{copy}}$ 表示权重矩阵；$\boldsymbol{b}_{\text{copy}}$ 表示偏置向量. 为了平衡不同模式，本文在上述计算式中采用了一个二维开关概率，其中 $\boldsymbol{y}_{t-1}$ 表示先前生成的单词的嵌入，$P(\cdot|\cdot)$ 表示每种模式下生成目标单词的概率. 这里，$P_{\text{vocab}}(\cdot)$ 是通过将 $\boldsymbol{s}_t$ 与所有单词之间的相似度应用 softmax 函数计算得到的；而 $P_{\text{cpqg}}(\cdot)$ 则是通过将多层感知器网络应用 softmax 函数计算得到的.

3.2.2　生成对抗模仿学习

在所提出的框架中，GAIL学习策略采用判别器 D 来自适应地学习问句生成器的奖励函数. 在每个步骤 t 中，状态 s 是当前生成的标记 $(y_1, y_2, \cdots, y_{t-1})$，动作 a 是候选的下一个单词，动作空间是整个词汇表. GAIL需要找到满足以下目标函数的鞍点：

$$\min_{G_\theta} \max_{D_\phi} \mathbb{E}_{y_r}\left[D_\phi(y_r)\right] + \mathbb{E}_{G_\theta}\left[1 - D_\phi\left(G_\theta(x)\right)\right] \tag{5}$$

其中，θ 和 ϕ 分别表示 G 和 D 中的所有参数；y_r 表示真实问句. 给定查询图 x，生成单词序列 $y_{1:T}$ 的概率计算式为

$$G_\theta\left(y_{1:T} \middle| x\right) = \prod_{t=0}^{T} G_\theta\left(y_t \middle| y_{<t}, x\right) \tag{6}$$

其中，T 表示序列长度；y_t 表示序列在第 t 个步骤生成的单词. 然后，本文使用策略梯度最大化预期奖励 $D_\phi\left(G(x)\right)$：

$$\mathbb{E}_{y\sim G_\theta}\left[\nabla_\theta \ln G_\theta(x) \hat{R}_y\right] \tag{7}$$

其中，$\hat{R}_y$ 表示控制更新的优势项(这里是标准化奖励). 在本文中，一旦获得奖励，则通过使用强化学习策略REINFORCE[19]最大化预期累积奖励来更新生成器 G.

本文利用大规模预训练语言模型 Bert[20] 作为判别器. 判别器接受真实序列及其对应的生成序列作为输入. 问题生成器的奖励 $D_\phi\left(y_g\right) = p_g$ 计算式为

$$\begin{cases} \boldsymbol{s}_r = \text{Bert}(y_r) \\ \boldsymbol{s}_g = \text{Bert}(y_g) \\ p_r, p_g = \text{softmax}\left(W_t\left[\boldsymbol{s}_r; \boldsymbol{s}_g\right]\right) \end{cases} \tag{8}$$

其中，W_t 表示可训练的权重，用于将输出嵌入投影到一个标量对数几率；y_r 表示真实问句；y_g 表示生成问句. 本文通过交叉熵损失来优化判别器，以最大化真实序列的概率 p_r.

3.2.3　训练细节

在训练过程中，本文提出了一个混合目标函数，结合交叉熵损失和策略梯度损失. 在使用交叉熵损失对模型进行预训练后，本文通过优化混合目标函数来训练模型，该损失函数结合了交叉熵损失和策略梯度损失：

$$\mathcal{L} = \gamma \mathcal{L}_P + (1-\gamma)\mathcal{L}_C \tag{9}$$

其中，γ 表示用于平衡两个损失的缩放因子；$\mathcal{L}_P$ 表示策略梯度损失函数；$\mathcal{L}_C$ 表示交叉熵损失函数.

4　实验

为了评估本文提出的FaQ2框架，我们在三个广泛使用的基准数据集上进行了实验，即 WebQuestionsSP(WebQSP)[21]、ComplexWebQuestions(CWQ)[22]和 PathQuestion[23]. 本文采用的基线模型是在采用的数据集上进行过实验的模型. 自动和人工评估的实验结果和消融研究表明了FaQ2的有效性.

4.1　数据集介绍与数据预处理

采用的三个数据集都来自通用领域，WebQSP 和 CWQ 均是基于 Freebase[24] KG 进行构建.

具体而言，WebquestionsSP(WebQSP)包含 4 737 个问句-答案对. 所有问句都是通过 Google Suggest API 收集的，答案则是通过亚马逊 Mechanical Turk 从 Freebase 中获取的. 此外，WebQSP 提供了每个问句对应的 SPARQL 查询.

尽管 WebQSP 被广泛使用，但该数据集中大多数问句都是简单问句，即单跳类型问句，只有少数问句涉及多跳推理和约束推理. 为了解决这个问题，CWQ(Complex Web Questions)对 WebQSP 中的问句添加了更多约束并修改了对应的 SPARQL，然后利用模板和亚马逊 Mechanical Turk 生成对应的自然语言问句. 它总共包含 34 689 个问句，其中大多数是复杂问句.

PathQuestion 包含两个子集，即原始 PathQuestion 和 PathQuestion-Large. 这两个子集合并在一起构成 PQ

数据集,该数据集包含9 731个问句,其中既有一跳问句,也有多跳问句.

对于每个数据集,本文随机选择80%的示例用于训练,10%用于验证,10%用于测试. 对于每个示例,本文将其SPARQL查询转换为KG子图,并根据SPARQL查询的特征指定不同问句的问句类型(查询意图).

4.2 基线模型介绍

本文采用了几种基线模型,具体介绍如下:

L2A[25]:一种基于注意力的Seq-to-Seq模型,用于从开放领域会话系统中的上下文生成自然语言问句. 编码器和解码器都是长短期记忆网络(Long Short-Term Memory,LSTM). 本文将序列化的查询图作为L2A的输入.

Zero-shot[2]:一种基于循环神经网络(Recurrent Neural Network,RNN)的Seq-to-Seq模型,配备有原始的词性复制操作机制,用于生成问句. 在这里,本文将序列化的查询图作为输入.

MHQG[4]:一种基于Transformer的模型,用于自动生成KG上的多跳问句. MHQG将KG子图和答案作为输入,生成自然语言问句,并基于命名实体的流行度进行难度评测.

BiGraph2seq[13]:一种图到序列的生成模型,利用双向门控图神经网络(Bidirectional Graph Neural Network,Bi-GNN)作为图编码器来编码KG子图,并通过复制机制增强RNN解码器.

T5[26]:即Text-to-Text Transfer Transformer,是一种先进的自然语言处理(Natural Language Processing,NLP)模型,它通过将所有文本相关任务转换为文本到文本的格式,使用统一的框架来处理语言理解和生成任务.

JOINGT[27]:JOINGT将结构感知的语义聚合模块整合到基础预训练语言模型(Pre-trained Language Models,PLMs)中,以保留图结构,并设计了三种预训练任务来学习图与文本之间的对齐.

DSM[28]:文献[28]在2022年提出的DSM模型致力于研究子图的多样性,并采用元学习方法来构建和理解这些多元化的子图结构.

ChatGPT:即GPT-3.5-turbo,直接采用该模型进行生成.

SGSH:一种基于大模型的问句生成模型. 该模型通过结合骨架启发式激活GPT-3.5的丰富语义知识,从而在KGQG任务中实现最佳的性能.

4.3 实现细节介绍

本文使用PyTorch进行了实现. 在训练过程中,本文调整了特征维度d,取值范围为{120,180,240,300,360},最终设置为300. 对于WebQSP和CWQ数据集,本文使用了来自OpenKE[29]的预训练的50维向量[30]作为KG中实体和关系的表征. 对于PQ数据集,本文直接采用了预训练的100维GloVe[31]作为KG中实体和关系的表征. 对于WebQSP和CWQ数据集,本文设置答案标记表征维度大小为64;对于PQ数据集,本文设置答案标记表征维度大小为32. 特征向量由KG中实体和关系向量、答案标记向量以及其余维度的随机值拼接而成.

本文采用Adam优化器,初始学习率设置为0.000 3. 在解码器层中,本文采用了0.3的丢弃比例,以防止过拟合. 在解码过程中,本文使用了束搜索,束大小为10. 此外,本文在验证集上的困惑度在连续30个epoch中没有改善时停止训练迭代. 所有参数都是根据验证集的性能进行自动调整.

本文所提出的FaQ2框架采用4块16 G的V100进行训练,在训练5 h后收敛.

4.4 实验结果

本文通过一组基于N-gram的度量标准来衡量方法的性能:BLEU-4[32]、METEOR[33]和ROUGE-L[34]. 准确地说,BLEU-4反映了修改后的4-gram的精度,ROUGE-L使用基于输出和目标之间最长公共子序列的精度和召回的调和平均值来计算F_1分数. METEOR修改了BLEU中的精度和召回计算,用映射单元和用于错误词序的惩罚函数的加权F_1分数替换它们.

表1展示了FaQ2和采用的基线模型在机器评价指标上的结果,粗体表示FaQ2的结果. 为了公平比较,所有的基线模型都使用Facts-to-Query阶段生成的查询图作为输入. 本文还进行了仅使用KG子图作为输入的消融实验,将在下文讨论. 结果显示,FaQ2在三个基准数据集上优于所有的基线模型. 具体来说,相对于SGSH(ChatGpt),FaQ2在CWQ上将BLEU-4分数提升了2.27个百分点,在WebQSP上提高了3.84个百分点,在PQ上提高了1.30个百分点. 与此同时,FaQ2在METEOR和ROUGE-L这两个指标上也显著超过了基线模型.

本文还注意到,Zero-shot比L2A表现更好,因为它在生成阶段利用了复制机制. MHQG模型比Zero-shot和L2A模型表现更好,因为它利用了基于Transformer的模型,该模型应用了自注意力机制来解决RNN中的远距离依赖问题. 与Seq-to-Seq模型[2,4,25]相比,基于GNN的编码器在建模查询图方面具备优势,因为基于RNN的编码器和基于Transformer的编码器均无法很好地建模了查询图的图结构信息. 因此,FaQ2和BiGraph2seq在两个基准测试集上表现出了较大的优势. BiGraph2seq中使用Bi-GNN,本文则采用DCGCN来捕获中节点之间的非局部交互. 同时,FaQ2在解码过程中通过生成对抗模型学习得到全局的奖励信号. 因此,

表1　在三个数据集上的机器评测指标结果　　单位:%

方法	CWQ			WebQSP			PQ		
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L
L2A	4.01	13.78	30.59	8.01	19.45	32.58	18.23	20.56	51.45
Zero-shot	6.37	16.32	32.10	9.45	21.52	34.78	20.55	22.76	53.32
MHQG	9.35	19.42	35.78	13.34	24.88	39.14	26.76	33.27	59.27
MHQG w/o Stage1	8.27	18.32	31.67	12.27	24.58	37.43	25.99	33.16	58.94
BiGraph2seq	26.01	28.12	53.58	27.86	30.24	62.77	59.23	44.57	75.38
BiGraph2seq w/o Stage1	24.01	25.67	52.13	25.11	27.23	60.13	58.15	43.78	74.82
T5	24.22	24.08	51.26	25.58	28.48	60.36	57.82	43.52	74.62
T5 w/o Stage 1	22.28	22.04	49.23	24.62	27.48	59.77	56.70	42.62	73.48
JOINGT	25.02	25.42	52.42	27.52	27.93	60.76	59.32	44.39	74.82
JOINGT w/o stage 1	22.87	23.83	51.36	26.72	26.72	59.94	58.11	43.26	73.62
DSM	27.23	27.57	53.92	28.94	29.58	61.22	61.02	45.57	76.33
DSM w/o Stage 1	26.11	26.48	53.04	28.13	28.85	60.76	60.33	44.43	75.68
ChatGPT	24.42	24.59	51.63	25.01	25.96	60.12	57.55	44.36	75.22
ChatGPT w/o Stage 1	23.02	23.62	50.32	24.34	24.86	59.46	56.68	43.95	74.84
SGSH(ChatGpt)	28.23	30.44	54.82	30.26	33.32	64.73	62.47	46.92	77.24
SGSH w/o Stage1	27.36	28.26	53.66	28.92	31.25	63.55	60.78	44.32	76.18
FaQ2	**30.50**	**32.43**	**57.07**	**34.10**	**35.23**	**67.84**	**63.77**	**48.23**	**78.69**
w/o Stage1	27.03	28.73	53.98	29.31	30.98	63.85	61.38	46.67	76.67
w/o GAIL	27.89	29.87	54.11	29.87	31.45	64.01	59.70	44.78	75.48

本文的模型在很大程度上优于BiGraph2Seq.

在移除stage 1,也就是查询图生成器后,多个模型的性能均出现了下降,这也进一步证明了查询图生成器的有效性. 此外,将GPT-3.5直接应用于知识图谱问答生成(Question Generation over Knowledge Base,KBQG)并未能取得良好的性能. 与现有的最先进的PLMs方法(即DSM)相比,本文注意到ChatGPT在几个数据集上的表现并不好. 这些表现并不符合GPT-3.5引人注目的能力,这可以解释为直接使用GPT-3.5配合一个基础提示仅提供了粗略的指导,而不能提供具体和准确的指导方向,导致生成的问题质量不佳.

如上所述,简单问句即单跳问句,复杂问句可分为五种类型. 本文将CWQ和WebQSP数据集中的问句按照图2中所示的六种问句类型进行划分,并分析了FaQ2在不同问句类型上的实验结果,如表2所示. 可以看出,FaQ2可以稳定地生成不同类型、高质量的复杂问句. 此外,相对于FaQ2在复杂问句上的生成效果,FaQ2在简单问句生成上的效果提升并不明显. 一个可能的原因是单跳问句在CWQ数据集中仅占约1%,这导致了FaQ2在训练中存在欠拟合的问题.

4.5　消融实验

为了深入分析FaQ2各个模块的效果,本文进行了消融实验,即分别移除了查询图生成器和生成对抗模仿学习模块,如表1所示.

表2　FaQ2在不同类型问句上的生成效果对比　　单位:%

问句类型	CWQ			WebQSP		
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L
单跳	18.39	25.65	47.16	31.03	32.54	65.48
多跳	26.31	30.56	53.24	34.57	34.37	67.02
实体约束	29.61	32.17	55.56	39.59	37.42	73.93
类型约束	25.62	30.89	53.35	31.53	33.46	56.92
比较约束	31.69	31.52	57.85	33.13	35.84	71.85
序数约束	32.53	34.46	56.92	25.30	27.90	58.66

(1)查询图生成器效果分析. 如上所述,现有的模型直接将实例化的KG子图作为输入并生成问句. 为了验证查询图生成器的效果,本文遵循这个设置,并分析FaQ2和三个基线模型(MHQG、BiGraph2seq和SGSH)的结果,如表1所示. 显然,这四个模型在移除查询图生成器后,在CWQ和WebQSP数据集上的性能都有显著下降. 这验证了查询图生成器有效性,该模块缓解了以往模型存在的语义歧义问题. 然而,这些模型在PQ数据集上性能下降的并不明显. 一个可能的原因是PQ数据集中的问句是由手动的模板构建的,主要由单跳或多跳这两种类型问句构成,输入的KG子图和目标问句之间的查询意图差距并不明显.

(2)生成对抗模仿学习效果分析. 本文在三个数据集上评估了移除生成对抗模仿学习模块的FaQ2框架.

如上所述,由于曝光偏差,早期生成中存在的错误将在后期不断累积,好比滚雪球. 表1所示的结果表明,生成对抗模仿学习训练策略可以提升问句生成的整体性能,因为它可以在生成问句时提供全局指导,并减轻曝光偏差的问题.

此外,本文的实验结果表明:在删除生成对抗模型学习模块后,FaQ2和BiGraph2seq模型之间的主要区别是图编码器,且它们的输入一致. 考虑到所有指标,FaQ2在没有生成对抗模型学习模块的情况下,性能仍然优于BiGraph2seq. 这说明FaQ2的图编码器可以更好地捕获查询图中节点之间的非局部交互,更适合于复杂问句的生成.

4.6 人工实验分析

除了上述自动度量标准外,本文还进行了人工实验分析,分别评估整体框架FaQ2和查询图生成器的性能.

(1)FaQ2整体框架评估. 从CWQ测试集中随机选择100个问句进行手动评估. 本文将FaQ2和SGSH在测试集中生成的问句配对,并进行随机排序. 两名人类标注者被要求从三个方面判断哪个更好,自然度即生成的问句的流畅性和可理解性,正确性评估了问题在语法上的表达,语义度量了问题中包含的语义信息. 结果如表3所示. 结果显示,本文的模型明显效果更好,在所有三个数据集上,FaQ2在流畅度、正确性和语义匹配度三个方面表现更优的实例更多,尤其是在WebQSP和CWQ数据集上. 结果表明:FaQ2从流畅度、正确性和语义匹配程度三个方面提高了生成问句的质量.

表3 FaQ2与当前最先进模型SGSH(ChatGPT)的手动评估结果

指标	CWQ			WebQSP			PQ		
	流畅度	正确性	语义匹配度	流畅度	正确性	语义匹配度	流畅度	正确性	语义匹配度
Win	19	37	28	35	35	29	7	10	5
Tie	79	59	69	59	59	64	90	86	92
Lose	2	4	3	6	6	7	3	4	3

(2)查询图生成器性能评估. 为了验证查询图生成器在不同规模和复杂度的KG上的效率与准确性,本文采用人工评估方式进行验证. 另外在生成查询图方面,近两年工作较少,本文以ChatGPT为基准模型,通过设计Prompt将序列化KG子图转化为对应的查询图,作为本文模型的对比模型.

两名人类标注者被要求从生成查询图的准确性角度判断哪个更好,依然按照上述的Win、Tie和Lose指标进行评价,在生成查询图准确性方面,本文模型优于ChatGPT,则为Win,反之为Lose. 同时,为了探索在不同规模KG上的表现,本文也分别采样了CWQ数据集对应KG子图的50%、75%和100%,人工评估对比在不同规模下生成查询图的准确性,如表4所示.

表4 查询图生成器与ChatGPT在KG规模不同时的手动评估结果

指标	CWQ_KG(50%)	CWQ_KG(75%)	CWQ_KG(100%)
Win	50	78	96
Tie	123	98	82
Lose	27	24	22

由表4可知,随着KG规模逐渐变大,本文模型在查询图生成器准确性方面的表现逐渐优于ChatGPT,这是因为随着KG规模增大,本文模型可以更有效地定位和采样所需的节点及关系,而对于ChatGPT而言,引入的噪声可能增加,从而导致其生成查询图的准确性逐渐降低. 这也说明了本文所提出的查询图生成器能够较好地适配当前任务.

4.7 泛化能力分析

为了探索FaQ2在不同数据集大小、数据分布和不同领域上的泛化性能,本文分别进行了实验,如表5所示,粗体表示FaQ2的结果.

表5 FaQ2在训练数据集大小不同时的性能对比 单位:%

方法	CWQ(25%)			CWQ(50%)		
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L
BiGraph2Seq	22.98	22.26	49.59	25.00	27.12	52.77
T5	23.28	24.86	50.82	25.82	27.04	53.10
SGSH(ChatGpt)	25.12	25.84	51.66	27.18	29.32	53.62
FaQ2	**27.68**	**28.82**	**54.07**	**29.32**	**31.23**	**55.97**

(1)数据集大小. 在数据集方面,本文分别从CWQ数据集中采样了25%、50%和全部的数据. 并测试本模型和部分基线模型在不同训练集大小下,在测试集上的性能.

从表5可得,FaQ2和基线模型在训练数据集逐步减少后,效果性能下降,在BLEU-4指标上,相较于基线模型SGSH,本文框架在50%和25%下降的百分比较少,说明本文框架具备一定的泛化能力.

(2)不同数据集分布. 在数据分布方面,本文将CWQ中的数据按照问句类型进行分类,并对训练集中每一类的样本抽取随机数的样本,组合成数据集,组成比例如表6所示.

表6　不同数据集的问句类型组成　　单位:%

训练集	单跳	多跳	类型约束	实体约束	序数约束	比较约束
训练集1	45	17	15	10	10	3
训练集2	20	25	15	20	10	10

本文分别在数据样本1和数据样本2上进行训练，并在原始的测试集上测试，得到的效果如表7所示，粗体表示FaQ2的结果.

由表7可知，当训练集中的数据集发生变化时，FaQ2和基线模型效果也会随之改变，而相对于当前最优的基线模型SGSH(ChatGPT)，FaQ2的性能变化相对较小. 说明FaQ2框架具备一定的泛化能力.

表7　FaQ2在数据分布不同时的表现　　单位:%

方法	训练集1			训练集2		
	BLEU-4	METEOR	ROUGE-L	BLEU-4	METEOR	ROUGE-L
BiGraph2seq	26.21	28.18	53.86	25.82	27.92	53.48
T5	24.33	24.34	51.78	23.58	23.64	50.26
SGSH(ChatGPT)	28.32	30.54	55.78	27.26	29.57	54.01
FaQ2	**30.74**	**32.55**	**57.62**	**30.02**	**32.03**	**56.57**

(3)不同领域数据. 为了验证FaQ2框架解决语义歧义问题在不同领域上的可扩展性，本文在不同领域数据集上进行了验证，如面向电影KG的数据集MetaQA，结果如表8所示，粗体表示FaQ2的结果.

表8　FaQ2在MetaQA数据集上的表现　　单位:%

方法	MetaQA		
	BLEU-4	METEOR	ROUGE-L
T5	30.988	33.69	57.92
T5 w/o Stage1	29.330	32.72	57.13
ChatGPT	28.620	28.12	56.88
ChatGPT w/o Stage1	27.530	27.17	55.13
SGSH(ChatGPT)	34.250	30.68	60.92
SGSH(ChatGPT)w/o stage1	33.270	29.35	59.44
FaQ2	**36.440**	**38.25**	**62.36**
w/o stage 1	34.220	35.83	60.92

由表8可知，FaQ2的性能超越了所有基线模型，这说明FaQ2模型在解决语义歧义方面较为有效，并且在不同领域上具备可扩展性和泛化能力. 而针对语义歧义问题，本文主要通过利用查询图生成器将KG子图转化为更接近问句意图的查询图来解决. 在移除查询图生成器后，表1和表8中均可以看到效果明显下降. 这不仅说明了查询图生成器的有效性，同时，也证明了本文在解决语义歧义问题上的有效性.

4.8　案例研究

(1)FaQ2整体框架在不同问句类型上的案例研究. 如图4所示，本文模拟了一个在线服务过程，以检验整体框架的有效性. 本文随机抽样了一些在采用的数据集中没有出现的KG子图，并将它们作为FaQ2的输入. 这些子图的问句类型被随机分配，并由FaQ2生成相应的问句. 抽样的问句类型包括比较约束(comparative constraint)、实体约束(entity constraint)、类型约束(type constraint)和排序约束(Ordinal Constraint). 如图4所示，所提出的框架可以生成高质量的复杂问句.

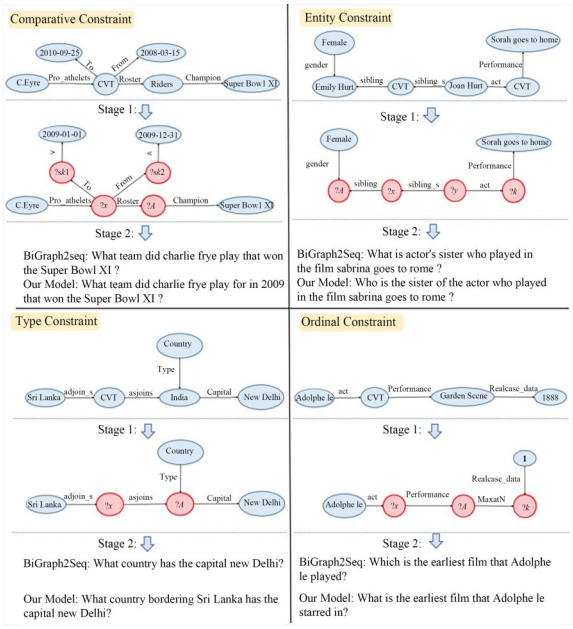

图4　FaQ2在不同问句类型上的案例分析

(2)FaQ2在处理不同类型语义歧义的案例研究. 本文面向KG生成问句领域中常见的语义歧义现象，对语义歧义类型进行归纳，可以分为三类：节点信息歧义、边信息歧义和约束信息歧义. 节点信息歧义包括生成问句中泄漏节点信息或未充分表达节点信息. 边信息歧义主要指生成问句未充分表达边上所代表的关系信息. 约束信息歧义指由于KG子图不包含函数操作信息，如最大最小和排序等函数操作信息，故生成问句无法表达相关信息.

为了验证FaQ2在解决不同类型语义歧义的有效性，本文进行了详细的案例研究，如图5所示.

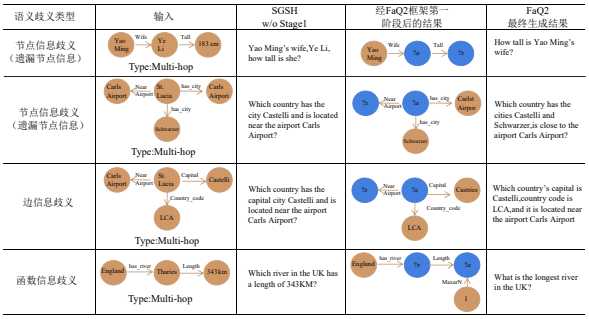

图5 FaQ2在面对不同语义歧义时的案例研究

由图5可知，根据不同类型的语义歧义，本文分别进行了案例研究. 在第一个案例中，由于输入的KG子图泄漏了“Ye Li”节点，导致SGSH w/o stage 1生成的节点暴露了该节点信息，从而最终生成了单跳问句，与目标多跳（Multi-hop）不符. 在第二个案例中，基线模型遗漏了节点信息“Schwarzer”，而FaQ2所采用的图神经网络模型更加充分地建模结构和内容信息，最终生成信息更加丰富的问句. 类似地，在第三个案例中，基线模型未能全面建模“country_code”关系信息，导致生成的问句遗失相关信息. 在最后一个案例中，由于KG子图本身不蕴含函数操作信息，故无法生成带函数约束类型信息，比如排序、比较等，本文提出的查询图生成器可以对KG子图添加相关函数操作信息，最终生成需要的带函数约束的问句.

5 相关工作

5.1 问句生成

KGQG的早期模型[35]主要通过从候选结构化查询中重构问句文本，并将其与原始问句进行比较. 通过对候选查询进行评价，可以重构预定义的模板. 一些方法[36~39]利用生成的问句在双向学习或半监督学习框架中训练问答模型. 近两年，许多研究集中在问句生成而不是问答的增强上. 这些工作主要采用编码器-解码器模型，并通过丰富输入信息优化生成效果. 在一些工作[40]中引入了RNN，用于针对KG中的事实生成自然语言问句. 在文献[3]中，问句是基于RNN的模型生成的，模型输入由相应的三元组和实体类型构成. 为了解决未见过的谓词和实体类型的挑战，文献[2]利用WiKidata语料库中的信息，在编码器-解码器架构中引入了词性复制动机制，以优化生成问句的效果. 然而，上下文无法涵盖所有关系类型. 因此，文献[5]提出了一个神经编码器-解码器模型，整合了多样化的上下文. 为了解决语义漂移问题，文献[1]提出了一个知识丰富、类型受限和语法引导的KGQG模型. 然而，这些模型仅关注从链式的KG子图中生成单跳或多跳问句，而采用的基于RNN的模型无法处理图结构数据. 文献[4]提出了一个用于在KG上生成复杂的多跳问句的模型. 此外，文献[13]应用了Bi-GNN模型对KG子图进行编码. 然而，在当前的任务设置中，由于KG子图不包含函数操作信息，因此他们的方法不能生成具有函数约束的复杂问句，包括比较约束和排序约束. 文献[41]提出了一种子图转述的方法用于解决KG问句生成领域的未见谓词问题.

在近年来的研究中，大型语言模型（Large Language Models，LLMs）在问句生成任务方面的应用得到了广泛的探索与发展. 首先，文献[42]提出了SGSH框架，该框架旨在通过结合模板启发式方法激活GPT-3.5的丰富语义知识，从而在KGQG任务中实现最新的性能. 这种方法通过构建一个模板训练数据集，并使用软提示策略训练BART模型，生成与每个输入关联的骨架，从而提供了一种细粒度的指导来激励LLMs生成更优的问句. 随后，文献[43]开发的ToolQA数据集针对LLMs在问答任务中的幻觉和弱数理推理问题，提出了利用外部工具增强LLMs的问答能力. 通过为LLMs提供与外部知识交互的专用工具，ToolQA旨在准确评估LLMs使用外部工具进行问答的能力. 进一步地，文献[44]提出了一个利用多角色LLMs代理来解决基于知识图谱的问答（Knowledge-Based Question Answering，KBQA）任务的统一框架. 该框架通过将LLMs代理分配为多个角色（包括通才、决策者和顾问）来处理KBQA子任务，证明了LLMs在解决复杂任务方面的灵活性和高效性. 最后，文献[45]探索了如何通过集成领域特定数据来增强LLMs在问答系统中的表现，尤其是当数据以混合格式（文本和半结构化表格）存在时. 这些研究不仅彰显了在问句生成及问答任务中利用大型语言模型的巨大潜力，也揭示了通过精细化的方法和框架设计，如何有效地挖掘和应用LLMs的知识，从而在各种NLP任务中达到前所未有的性能水平.

5.2 模仿学习

模仿学习侧重于从示例中学习策略，这在优化强化学习中的奖励模型方面取得了显著成果. 传统方法是通过逆强化学习（Inverse Reinforcement Learning，IRL）[46,47]找到最优的奖励函数来解释专家行为. 然而，IRL需要在学习循环中解决强化学习问题，在大型环境中运行需要昂贵的试错代价. 因此，研究者提出了生成对抗模仿学习[9]，它使用生成对抗网络（Generative Adversarial Network，GAN）[48]来学习专家策略，消除了中间的IRL步骤. GAIL最近在NLP的许多领域中被应用，如知识推理[49]、实体抽取[50]等. 本文尝试将生成对抗模仿学习引入问题生成任务中，旨在减轻暴露偏差.

6 结论

本文专注于KG上的问句生成任务,并提出了一个针对复杂问句生成的两阶段框架,即FaQ2,以应对语义歧义的挑战. 在第一阶段,本文设计了一个查询图生成器,利用基本操作符构建推理流程,将KG子图转换为带有不同查询意图的查询图. 在第二阶段,本文提出了一个问句生成器,该生成器利用DCGCN对查询图进行编码,并采用带有节点级复制机制的BART模型来生成问句. 此外,为了缓解曝光偏差,本文引入了GAIL来训练问句生成器. 除了教师强制损失外,问句生成器还从一个基于Bert的判别器中获得奖励. 该判别器通过模仿标记数据自适应地学习奖励函数,并指导问句生成器探索潜在问句空间中的高奖励区域.

本文在三个广泛使用的基准数据集上评估了所提出的FaQ2框架的有效性,结果表明其性能超越了所有基线模型. 在未来的工作中,本文计划探索在生成过程中控制问题复杂度的方法.

参考文献

[1] BI S, CHENG X Y, LI Y F, et al. Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases[EB/OL]. (2020-10-23) [2024-04-25]. https://arxiv.org/abs/2010.03157v3.

[2] ELSAHAR H, GRAVIER C, LAFOREST F. Zero-shot question generation from knowledge graphs for unseen predicates and entity types[EB/OL]. (2018-02-19) [2024-04-25]. http://arxiv.org/abs/1802.06842.

[3] INDURTHI S R, RAGHU D, KHAPRA M M, et al. Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia: ACL, 2017: 376-385.

[4] KUMAR V, HUA Y C, RAMAKRISHNAN G, et al. Difficulty-Controllable Multi-Hop Question Generation from Knowledge Graphs[M]//The Semantic Web-ISWC 2019. Cham: Springer International Publishing, 2019: 382-398.

[5] LIU C, LIU K, HE S Z, et al. Generating questions for knowledge bases via incorporating diversified contexts and answer-aware loss[EB/OL]. (2019-10-29) [2024-04-25]. https://arxiv.org/abs/1910.13108v1.

[6] RANZATO M, CHOPRA S, AULI M, et al. Sequence level training with recurrent neural networks[EB/OL]. (2016-05-06)[2024-04-25]. https://arxiv.org/abs/1511.06732v7.

[7] GUO Z J, ZHANG Y, TENG Z Y, et al. Densely connected graph convolutional networks for graph-to-sequence learning[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 297-312.

[8] LEWIS M, LIU Y H, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[EB/OL]. (2019-10-29) [2024-04-25]. https://arxiv.org/abs/1910.13461v1.17.

[9] HO J, ERMON S, HO J, et al. Generative adversarial imitation learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 4572-4580.

[10] YIH W T, CHANG M W, HE X D, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C/OL]. (2015-05-01) [2024-04-05]. https://www. microsoft. com/en-us/research/publication/semantic-parsing-via-staged-query-graph-generation-question-answering-with-knowledge-base/.

[11] QIU Y Q, ZHANG K, WANG Y Z, et al. Hierarchical query graph generation for complex question answering over knowledge graph[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 1285-1294.

[12] HOMMEAUX E P. SPARQL query language for RDF[J/OL]. (2013-03-26) [2024-04-05]. http://www.w3.org/TR/rdf-sparql-query/.

[13] 仇韫琦, 王元卓, 白龙, 等. 面向知识库问答的问句语义解析研究综述[J]. 电子学报, 2022, 50(9): 2242-2264.
QIU Y Q, WANG Y Z, BAI L, et al. A survey of question semantic parsing for knowledge base question answering[J]. Acta Electronica Sinica, 2022, 50(9): 2242-2264. (in Chinese)

[14] 张元鸣, 姬琦, 徐雪松, 等. 基于知识图谱关系路径的多跳智能问答模型研究[J]. 电子学报, 2023, 51(11): 3092-3099.
ZHANG Y M, JI Q, XU X S, et al. Knowledge graph relation path network for multi-hop intelligent question answering[J]. Acta Electronica Sinica, 2023, 51(11): 3092-3099. (in Chinese)

[15] 高留杰, 赵文, 张君福, 等. G2S: 基于语义块的知识图谱问答语义解析[J]. 电子学报, 2021, 49(6): 1132-1141.
GAO L J, ZHAO W, ZHANG J F, et al. G2S: Semantic segment based semantic parsing for question answering over knowledge graph[J]. Acta Electronica Sinica, 2021, 49(6): 1132-1141. (in Chinese)

[16] LI Z, WANG Z J, TAN M, et al. DQ-BART: Efficient sequence-to-sequence model via joint distillation and quantization[EB/OL]. (2022-03-21) [2024-04-05]. https://arxiv.org/

abs/2203.11239v1.

[17] LEVI F W. Ordered groups[C]//Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP. Beijing: Association for Computational Linguistics, 2015: 15.

[18] BECK D, HAFFARI G, COHN T. Graph-to-sequence learning using gated graph neural networks[EB/OL]. (2018-06-26)[2024-04-25]. https://arxiv.org/abs/1806.09835v1.

[19] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[M/OL]. (2016-05-19) [2024-04-05]. http://arxiv. org/abs/1409.0473.

[20] TU Z P, LU Z D, LIU Y, et al. Modeling coverage for neural machine translation[EB/OL]. (2016-08-06) [2024-04-25]. https://arxiv.org/abs/1601.04811v6.

[21] GU J T, LU Z D, LI H, et al. Incorporating copying mechanism in sequence-to-sequence learning[EB/OL]. (2016-06-08)[2024-04-25]. https://arxiv.org/abs/1603.06393v3.

[22] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8(3): 229-256.

[23] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24) [2024-04-25]. https://arxiv.org/abs/1810.04805v2.

[24] YIH W T, RICHARDSON M, MEEK C, et al. The value of semantic parse labeling for knowledge base question answering[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Redmond: Microsoft Research, 2016: 201-206.

[25] TALMOR A, BERANT J. The web as a knowledge-base for answering complex questions[EB/OL]. (2018-03-18)[2024-04-25]. https://arxiv.org/abs/1803.06643v1.

[26] 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(1): 5485-5551.

[27] KE P, JI H Z, RAN Y, et al. JointGT: Graph-text joint representation learning for text generation from knowledge graphs[EB/OL]. (2021-06-19) [2024-04-05]. https://arxiv.org/abs/2106.10502v1.

[28] GUO S S, ZHANG J, WANG Y L, et al. DSM: Question generation over knowledge base via modeling diverse subgraphs with meta-learner[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022: 4194-4207.

[29] ZHOU M T, HUANG M L, ZHU X Y. An interpretable reasoning network for multi-relation question answering[EB/OL]. (2018-06-01) [2024-04-25]. https://arxiv.org/abs/1801.04726v3.

[30] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: A collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 1247-1250.

[31] DU X Y, SHAO J R, CARDIE C. Learning to ask: Neural question generation for reading comprehension[EB/OL]. (2017-04-29)[2024-04-25]. https://arxiv.org/abs/1705.00106v1.

[32] HAN X, CAO S L, LV X, et al. OpenKE: An open toolkit for knowledge embedding[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels: Association for Computational Linguistics, 2018: 139-144.

[33] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2013: 2787-2795.

[34] 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 (EMNLP). Stanford: Stanford University, 2014: 1532-1543.

[35] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[J]. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002, 2002: 311-318.

[36] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Pittsburgh: Carnegie Mellon University, 2005: 65-72.

[37] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Los Angeles: Association for Computational Linguistics, 2004: 74-81.

[38] BERANT J, LIANG P. Semantic parsing via paraphrasing [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

Papers). Stanford: Stanford University, 2014: 1415-1425.
[39] JIA R, LIANG P. Data recombination for neural semantic parsing[EB/OL]. (2016-06-11)[2024-04-25]. https://arxiv.org/abs/1606.03622v1.
[40] KOČISKÝ T, MELIS G, GREFENSTETTE E, et al. Semantic parsing with semi-supervised sequential autoencoders[EB/OL]. (2016-09-29) [2024-04-25]. https://arxiv. org/abs/1609.09315v1.
[41] CHEN Y, WU L F, ZAKI M J. Toward subgraph-guided knowledge graph question generation with graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(9): 12706-12717.
[42] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
[43] LI R P, CHENG X. DIVINE: A generative adversarial imitation learning framework for knowledge graph reasoning[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019: 2642-2651.
[44] ZHANG T T, JI H, SIL A. Joint entity and event extraction with generative adversarial imitation learning[J]. Data Intelligence, 2019, 1(2): 99-120.
[45] GUO S S, LIAO L Z, ZHANG J, et al. SGSH: Stimulate large language models with skeleton heuristics for knowledge base question generation[EB/OL]. (2024-04-02)[2024-04-05]. https://arxiv.org/abs/2404.01923v1.
[46] HU S, ZOU L, ZHU Z X. How question generation can help question answering over knowledge base[M]//Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2019: 80-92.
[47] CAO R S, ZHU S, LIU C, et al. Semantic parsing with dual learning[EB/OL]. (2019-07-24)[2024-04-25]. https://arxiv.org/abs/1907.05343v2.
[48] SERBAN I, GARCÍA-DURÁN A, GÜLÇEHRE Ç, et al. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus[EB/OL]. (2016-05-29)[2024-04-25]. https://arxiv.org/abs/1603.06807v2.
[49] RUSSELL S. Learning agents for uncertain environments (extended abstract)[C]//Proceedings of the Eleventh Annual Conference on Computational Learning Theory. New York: ACM, 1998: 101-103.
[50] NG A Y, RUSSELL S. Algorithms for inverse reinforcement learning[C]//Proceedings of the 17th International Conference on Machine Learning. Stanford: Morgan Kaufmann, 2000: 663-670.

作者简介

张　琨　男,1997年生. 现为中国科学院计算技术研究所博士研究生. 主要研究方向为知识库问答、问句生成.
E-mail: zhangkun18z@ict.ac.cn

王元卓　男,1978年生. 现为中国科学院计算技术研究所研究员,博士生导师,中科大数据研究院院长. 主要研究方向为网络大数据分析、开放知识计算、社交网络演化计算.
E-mail: wangyuanzhuo@ict.ac.cn