一种基于子图转述的问题生成方法

温立强[1],熊冠铭[1*],王 宇[1],陈一朴[1],李伟平[1],赵 文[2]
(1. 北京大学软件与微电子学院,北京 102600;2. 北京大学软件工程国家工程研究中心,北京 100871)

摘 要: 本文提出了一种子图转述的方法用于解决知识图谱问题生成中的未见谓词问题. 传统的问题生成方法主要利用标注的问答数据(问题-逻辑形式对)生成问题,然而标注数据难以覆盖知识图谱中所有的谓词,如何对未见的谓词生成问题依然是一个挑战. 本文提出了一种基于子图结构的语义解耦方法,通过将复杂问题对应的知识图谱子图分解为原子级子图,从而将包含未见谓词的多跳子图拆分为易于处理的单跳子图. 并且本文设计了一种子图转述方法,通过对数据集中的谓词进行采样,得到子图描述文本,并在大规模无监督数据上训练得到子图转述器,能够为包含未见谓词的子图提供自然语言形式的表述,为生成问题提供了有效的信息. 本文定量分析了在不同的难度级别下模型的性能表现,在GrailQA等数据集上的实验结果表明,本文的方法达到了最先进的性能.

关键词: 子图采样;子图转述;未见谓词;问题生成;知识图谱
基金项目: 国家重点研发计划项目(No.2021YFC3340301)
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3578-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221188

A Question Generation Method Based on Subgraph Paraphrase

WEN Li-qiang[1], XIONG Guan-ming[1*], WANG Yu[1], CHEN Yi-pu[1], LI Wei-ping[1], ZHAO Wen[2]
(1. *School of Software and Microelectronics*, *Peking University*, *Beijing* 102600, *China*;
2. *National Engineering Research Center for Software Engineering*, *Peking University*, *Beijing* 100871, *China*)

Abstract: This paper proposes a method based on subgraph rephrasing to solve the problem of unseen predicates in question generation over knowledge graph. Traditional KBQG (Question Generation over Knowledge Base) methods mainly use annotated Q&A (Question and Answer) data (question and logic formal pairs) to generate questions. However, annotated data can't fully cover all predicates in the knowledge graph. It is still a challenge to generate questions with unseen predicates in the knowledge graph. In this paper, we propose a semantic decoupling method based on subgraph structure. By decomposing the subgraph corresponding to a complex question into atomic subgraphs, the multi-hop subgraph containing unseen predicates can be divided into single-hop subgraphs that are easy to handle. In addition, we design a subgraph rephrasing procedure to train a subgraph rewriter on large-scale unsupervised data through sampling the predicates in the dataset by subgraph sampling. The subgraph rewriter will provide natural language form for subgraphs and effective information for generating questions. This paper quantitatively analyzes the performance of the model at different difficulty levels. The experimental results on GrailQA and other datasets show that our method achieves the state-of-the-art performance.

Key words: subgraph sampling; subgraph representation; unseen predicates; question generation; knowledge graph
Foundation Item(s): National Key Research and Development Program of China (No.2021YFC3340301)

1 简介

基于知识的问题生成 Question Generation over Knowledge Base,KBQG)是一项旨在给定知识库的情况下生成自然语言问题的任务. KBQG的应用范围很广,在学术界和工业界都获得了广泛的关注. 在本文中,使用一个特定的知识图谱(Knowledge Graph,KG),Freebase[1]作为知识库.

目前,KBQG的工作主要使用基于序列到序列的神经网络来生成问题,其中主要关注的难点又可细分为:(1)如何处理未见的谓词与实体. 对于未在训练数据中出现的谓词和实体,文献[2,3]提出了一种灵活的拷贝

收稿日期:2022-10-20;修回日期:2023-02-10;责任编辑:梅志强
*通讯作者:熊冠铭

机制从输入中拷贝单词,文献[4~6]从百科数据中引入辅助信息用于丰富谓词的描述,文献[7]使用预训练-微调的方法提升了模型的泛化能力.(2)如何生成复杂的多跳问题.这类方法首先需要使用一种神经网络建模知识图谱子图,然后使用解码器生成问题.其中,文献[8]使用Transformer编码子图三元组,文献[7]提出一种双向图模型编码子图,文献[9]将子图序列化后作为输入.

然而,现有的工作忽略了如何从带有未见谓词的子图中生成复杂问题.复杂问题主要是指出涉及多跳推理、约束关系或数值运算的问题[10].知识图谱谓词是一种较为抽象的语义表述,与自然语言表述之间存在语义鸿沟[11],如何从含有未见谓词的子图中生成复杂问题是对模型生成能力的考验.文献[9]中依据训练集中是否出现谓词将知识库问答数据分为3种难度级别(详见第4.1节),而在问题生成领域,模型在不同难度下的表现还没有被探索.此外,由于知识图谱结构往往较复杂,标注问答数据(问题-逻辑形式对)昂贵且往往难以完全覆盖所有谓词,因此,如何对未见过的谓词生成问题依然是一个挑战.

为了解决上述的问题,本文提出了一种基于子图转述的问题生成方法.

(1)本文提出一种基于子图结构的语义解耦方法.通过将问题对应的知识图谱子图分解为原子级子图,本文将包含未见谓词的多跳子图拆分为容易处理的单跳子图.

(2)本文提出了一种基于子图采样和无监督语料匹配的子图转述方法,能够将子图转述为自然语言描述,为未见的谓词提供辅助信息.该方法通过知识图谱子图采样和大规模语料匹配的方法构建数据,通过训练模型生成子图描述,能够有效提升谓词覆盖率并且充分地挖掘谓词的表述.

(3)本文定量分析了在不同的难度级别下模型的性能表现.本文按照未见谓词的难度构建了两个数据集并进行了实验,结果表明,本文的方法取得了先进的性能.

2　相关工作

基于模板的知识图谱问题生成KBQG在过去几十年中取得了较大的进展.早期的方法大多基于模板生成问题.文献[12]提出了一种无监督系统,能够通过搜索引擎扩充基于模板的问题.文献[13,14]从种子问题中收集结构化三元组,并使用基于模板的方法进行结构化查询.然而,早期的方法大多基于模板生成问题,较少关注未见谓词和复杂语义的问题.随着深度学习技术的发展,KBQG的主流方法转变为给定知识图谱中的一组子图,使用基于序列到序列神经网络生成问题.

未见谓词 现有工作对生成模型的泛化性也做了一定的探索.对于谓词中的稀有词(未登录词),文献[2,15]提出了一种灵活的拷贝机制.对于未知的谓词和实体类型问题,文献[16]在维基百科中收集实体的上下文作为辅助信息,并引入基于词性的拷贝机制来生成问题.然而,文献[4]认为这些辅助信息过于嘈杂,甚至包含错误的信息,于是提出一个复杂的模型整合了各种现成的上下文,并设计了一个答案感知损失函数,以确保生成的问题与最终答案相关联.文献[8]提出了Graph2Seq模型,设计了一种节点级的拷贝机制用于生成问题.文献[7]提出了JointGT,首先在大规模语料上进行预训练,再在问题生成任务中进行微调,实验证明其有效增强了模型的泛化性.

多跳知识图谱问题生成主流方法关注的复杂语义[17],局限于多跳且答案为实体的问题.文献[6,18]提出了一种基于Transformer的端到端的方法,用于在知识图谱子图上生成多跳和难度可控的问题.文献[8]提出了一种图模型,能有效表征较为复杂的知识图谱结构,进而生成多跳问题.文献[7]以序列化的知识图谱子图为输入,直接使用预训练语言模型生成问题.

3　方法

3.1　概览与定义

本文研究了基于知识图谱的问题生成任务.传统的KBQG方法是从知识图谱中的事实三元组生成问题,在大规模知识图谱上,人工标注的数据往往难以覆盖所有的谓词.因此,如何增强模型的泛化能力以处理未见的谓词是一个亟需解决的问题.本文提出一种利用子图转述生成问题的方法.

本文方法的总体流程可见图1.给定S表达式(S表达式是一种适用于知识图谱问答的逻辑形式,能够与SPARQL语句互相转换[9]),基于子图结构的语义解耦方法能够在知识图谱中将该逻辑形式实例化并拆分为原子级子图,对于每一个原子级子图,子图转述器将其转述为自然语言形式的描述.为了训练子图转述器,基于子图采样与无监督匹配的数据构造方法能够自动采样大量数据,进而有效缓解了未见谓词的问题,增加了方法的泛化性(例如,图中标红的谓词"film. film. initial_release_date"为训练集未见的谓词,而右侧的方法能够对其进行大规模采样并生成转述文本,进而增强问题生成效果).

给定一个逻辑形式L和知识库K,本文的目标是生成和L语义一致的自然语言问题Q.整个任务的目标是学习$p(Q|L,K)$.其中,子图转述器与问题生成器分别是两个神经网络模型,其整体定义如下.

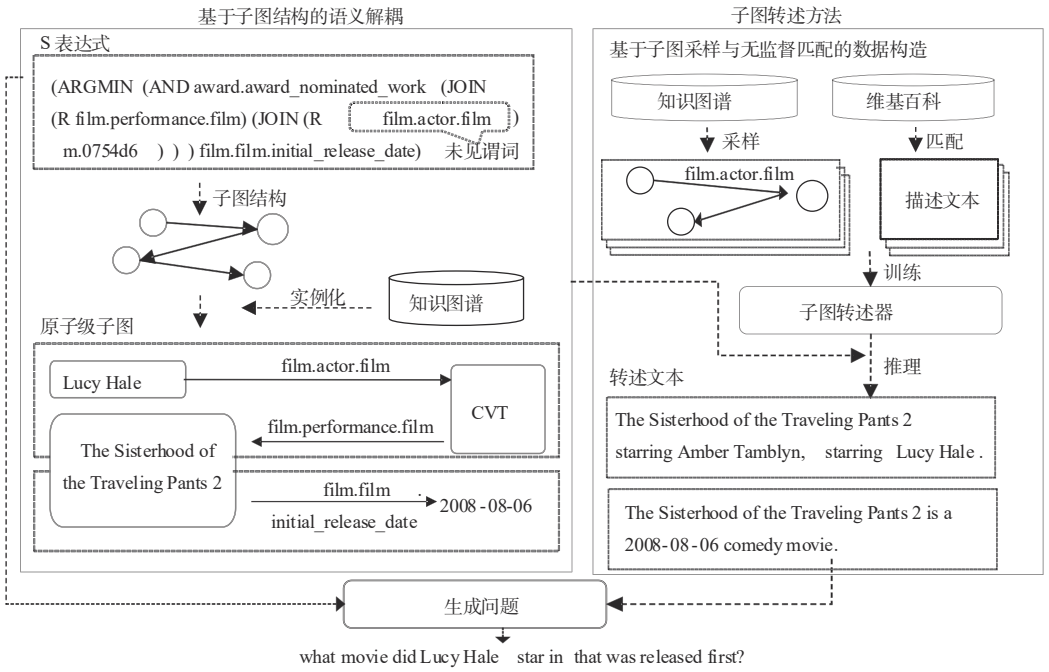

图 1 模型概览

(1)子图转述器. 该模块将逻辑形式拆分为子图集合,对集合中的每一个子图,该模块能够生成一段对应的自然语言形式的转述文本. 总体过程定义如下:

$$T=\mathrm{RP}(L,K,\Theta) \tag{1}$$

其中,T表示转述文本,RP(·)表示子图转述过程,Θ表示模型参数.

(2)问题生成器. 该模块能够从给定的逻辑形式和转述文本中生成问题. 总体过程定义如下:

$$Q=\mathrm{QG}(L,T;\ \Phi) \tag{2}$$

其中,Q表示生成的问题,QG(·)表示问题生成过程,Φ表示模型参数.

3.2 基于子图结构的语义解耦

本文提出一种基于子图结构的语义解耦的方法,将包含未见谓词的多跳子图拆分为容易处理的单跳子图.

具体地,给定一个可执行的图状逻辑形式L,定义原子级子图g. 在Freebase中,原子级子图类型一共有两种,即type(g)∈{CVT, Single},函数type(g)返回子图的类型. 其中,CVT(Compound Value Type, CVT)是Freebase中的一种复合类型的数据结构,通常用于表示一个事件,每个条目由一个中心节点和多个CVT谓词组成. 其中,定义"CVT谓词"为尾实体是CVT节点的谓词. Single表示一跳关系,通常用于表示一个事实,可以形式化为(主语,谓语,宾语)三元组. 定义"一跳关系谓词"为主语和宾语都是非CVT节点的谓词,例如头实体为命名实体,尾实体为值(数值、日期或字符等)类型. 具体而言,使用Virtuoso存储Freebase,并且按照Google提供的CVT类型的谓词集合来区分某一个谓词的类型.

例如,给定问题"what movie did lucy hale star in that was released first?"和逻辑形式,将其拆解为图2转述器推理部分所示的两个子图,分别表示"lucy hale star in the movie"和"the movie's release date is 2008-08-06".

3.3 基于子图采样和无监督匹配的数据构造

转述过程RP(·)从逻辑形式L中生成转述文本T. 图1中展示了子图转述的所有流程.

给定一个可执行的图状逻辑形式L,首先将其中的变量节点在知识库K中实例化,得到一张完整的图G之后,使用上述语义解耦方法将图G拆为原子级别的子图集合,对于每个子图$g\in G$,将其序列化后,使用转述器转述为子图转述文本. 将所有的子图转述文本拼接后得到逻辑形式L对应的转述文本T.

正式定义如下:

$$T=\mathrm{CONCAT}(\{B(p_\Theta(d_i|d_{<i},g))|g\in G\}) \tag{3}$$

$$G=f(S,K) \tag{4}$$

其中,函数CONCAT(·)表示拼接所有元素,函数B(·)表示使用集束搜索(beam search)进行推理(本文只保留得分最高的一个生成结果),d_i表示第i个生成的词,函数f(·)表示图采样过程,函数p_Θ(·)表示以Θ为参数的生成模型.

本方法的优越性在于,转述文本能根据输入子图

中的实体信息变化而变化,而传统方法只能使用统计方法保留硬匹配的结果,例如对于谓词“person.place_of_birth”,匹配结果为“is birthplace of”.

3.3.1 子图采样

在转述器的训练阶段,给定某一个谓词,需要从知识库采样一系列包含该谓词的原子级子图. 给定一个一跳关系谓词,以该谓词为中心,采样对应的头实体和尾实体,即采样一系列三元组(头实体,谓词,尾实体);给定一个CVT谓词,首先采样该谓词的尾实体集合(CVT节点集合),再对集合中的每一个节点,以该节点为中心,实例化一个CVT子图,以此获得一系列的子图. 例如,如图2转述器训练部分所示,给定CVT谓词“film. actor. film”,能够采样到一个子图:(Margaret Hamilton→ film. actor. film→CVT→film. performance. film→ The Wizard of Oz). 由于资源有限,本文没有采样Freebase中的所有谓词,而是对本文所用数据集(见第4.1节)中涉及到的谓词进行了采样.

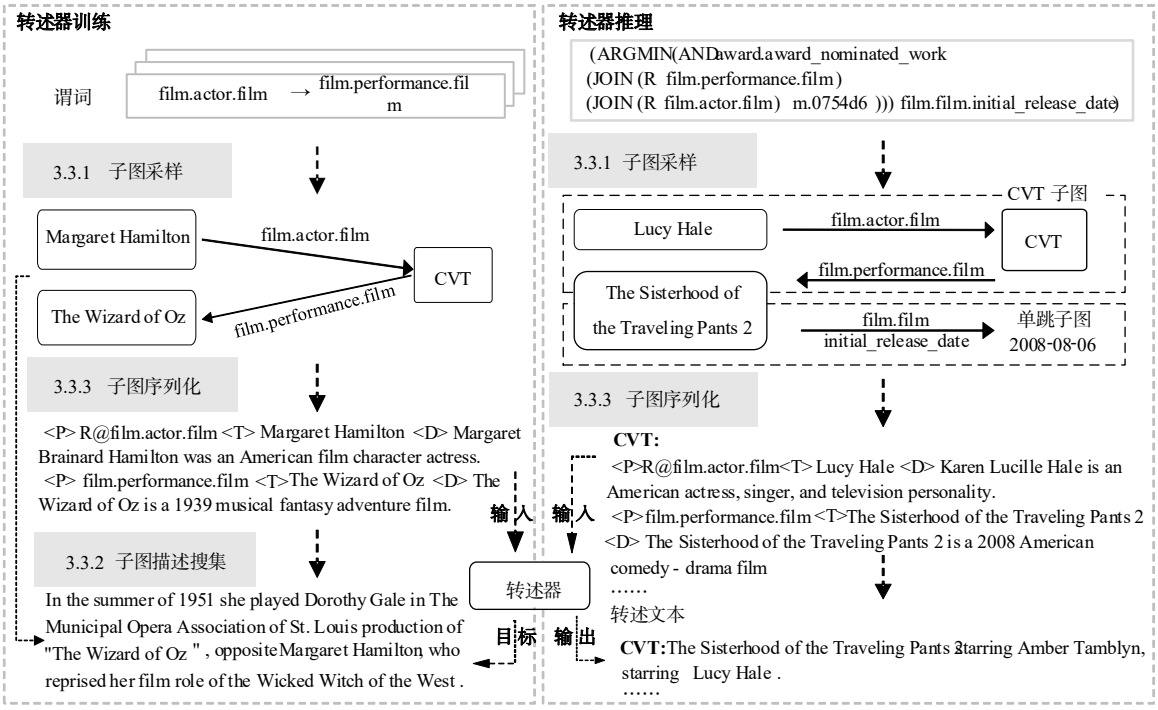

图 2　子图转述器的训练与推理过程

在转述器的推理阶段,子图采样的过程略有不同. 如图2转述器推理部分所示,给定一个可执行的图状逻辑形式L,首先将其转为SPARQL,然后将其中的变量节点在知识库K中实例化,得到一张完整的图G之后,将图G拆为原子级别的子图.

3.3.2 子图序列化

为了利用先进的预训练语言模型,本方法需要对子图进行序列化处理. 对于“单跳子图”,使用如下模板:“<H>[头实体名称]<D>[头实体描述]<P>[谓词]<T>[尾实体名称]<D>[尾实体描述]”,特殊字符“<H>, <D>, <P>和<T>”用于分割不同的语义部分. 对于“CVT子图”,为了区分谓词的方向,在CVT谓词(指向一个CVT节点)的前面加上特殊字符“R@”. 将CVT子图按边拆分为二元组(具有相同的头实体)集合,对其中每一个二元组,使用如下方法序列化:“<P>[R@][谓词]<T>[尾实体名称]<D>[尾实体描述]”. 最后将CVT子图对应的二元组集合中所有的二元组序列拼接为一个字符串,最为最终的序列化结果. 图2中给出了一个例子.

3.3.3 描述文本搜集

转述器的作用是给定原子级子图g并生成对应的转述文本. 例如,给定CVT子图(m.01d1st film.actor.film ? y. ? y film.performance.film ? x.)以及实体m. 01d1st的名称“Nick Cannon”,标注员需要理解这些实体和谓词的含义,并且写下对应的描述:“Nick Cannon star in film [? x]”,即该子图对应的语义是“star in”. 这样细粒度的标注需要耗费极大的人工成本,大规模标注是不切实际的. 因此,本文提出一种利用大规模无监督语料增强谓词语义信息的方法. 不同于之前使用规则匹配(例如基于实体名称共现[16]),本文将远程监督的思想和生成式预训练模型相结合,提出了一种软生成的方法.

具体而言,本方法需要搜集原子级子图-自然语言描述文本对作为训练数据,这里将搜集到的描述文本记为 d. 为了匹配描述文本,本方法利用英文维基百科(2018-12-20版)作为语料库. 维基百科中,每一个页面都有标题和正文,正文中有许多段落,本文使用Spacy工具重新分词并使用Elasticsearch建立索引. 对于“单跳子图”,本方法在维基百科中寻找头实体和尾实体同时出现的句子. 对于“CVT子图”,本方法在维基百科中寻找所有实体同时出现的段落,并删去段落中没有出现任何实体的句子.

3.4 转述文本

转述文本是对子图的一种自然语言形式的描述,是对谓词语义信息的一种补充,准确的转述文本能够有效缓解未见谓词的问题. 如图2所示,转述器从给定的CVT子图中生成了“starring”,成功表述了“出演”的语义. 即给定一种逻辑形式 L,第3.3.1节将其实例化为图 G;第3.3.2节对子图 $g \in G$ 进行序列化;然后使用转述器生成转述文本 T;最后将 L 和 T 作为输入,生成问题. 转述文本难以使用自动化的指标进行评价,因为其没有标注的答案,因此本文增加了人工评价.

转述器基于一个先进的预训练语言模型BART[19]实现. 为了使得训练过程更加平稳,本文将上述两种子图类型的数据合并,并随机打乱,作为转述器的训练数据. 训练转述器的损失函数如下:

$$\text{Loss} = -\sum_{M}\sum_{i=1}^{|\bar{d}|}\log p_{\Theta}(\bar{d}_i|\overline{d_{<i}}, g) \tag{5}$$

其中,$\bar{d}_i$ 是 $\bar{d}$ 的第 i 个词,M 是训练数据的长度.

3.5 问题生成

问题生成部分QG(·)以逻辑形式 L 和转述文本 T 作为输入,生成问题 $\bar{Q}$. 本方法的问题生成模型同样基于BART微调. 训练和推理的过程和经典的微调BART方法是一致的. 总体过程如图3所示.

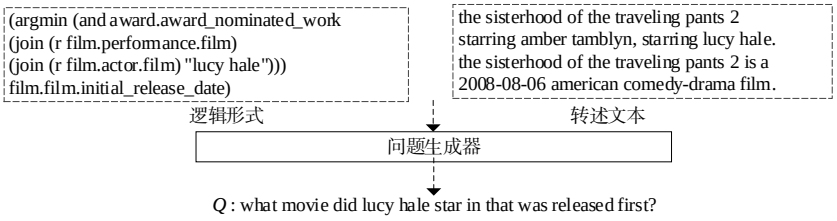

图3 问题生成

正式定义如下,给定训练样本 $(L, T, \bar{Q})$,目标函数 Loss′为:

$$\text{Loss}' = -\sum_{N}\sum_{i=1}^{|\bar{Q}|}\log p_{\Phi}(\bar{Q}_i|\bar{Q}_{<i}, S, T) \tag{6}$$

其中,$\bar{Q}_i$ 是标注的问题 $\bar{Q}$ 中的第 i 个词,N 是训练数据的长度,Φ 表示问题生成模型的参数. 使用标准的集束搜索推理,具体的超参设定参见第4.2节.

4 实验

本节讨论实验并与现有研究对比,以评估本方法的有效性.

4.1 数据集

GrailQA-QG GrailQA数据集是一个关注模型泛化能力的知识库问答数据集,共包含64 331个样本,其中训练、验证和测试集分别包含44 337、6 763和13 231个样本. 验证集由三种泛化级别的数据构成,分别是“独立同分布”(independent identically distributed)、“组合”(compositional)和“零样本”(zero-shot),其难度递增. “独立同分布”表示样本中的涉及的要素:谓词、实体类型和函数操作(聚合、比较级、最高级等,以下简称函数)和训练集是同分布的;“组合”表示训练集包含上述三要素,但是不包含其组合;“零样本”则表示既不包含要素,也不包含组合. 但是,原始的测试集只提供了问题,没有提供标注的逻辑形式,因此本文对原始数据重新划分,得到支持问题生成任务的数据集:GrailQA-QG. 具体地,本文使用原始的验证集作为问题生成任务的测试集,将原始的训练集随机划分10%作为验证集.

WQCWQ-Unseen WebQuestionsSP[3]数据集和ComplexWebQuestions 1.1数据集[20]是知识库问答领域常用的数据集,其中包含问题和对应的SPARQL表达式. 前人的工作[6-8]将WebQuestionsSP和ComplexWebQuestions(旧版)合并,构建了适用于问题生成的数据集,但是他们并没有保留原始的SPARQL,而是将原始SPARQL从SELECT改为CONSTRUCT,在知识库中查询后仅保存返回的RDF三元组. 这么做丢失了语义信息[28],且旧版的数据已难以获得,因此本文重新处理了数据. 除了ComplexWebQuestions 1.1测试集未标注SPARQL外,本文合并了其他的数据,记为WQCWQ数据. 为了测试泛化能力,仿照GrailQA,将测试集划分为“独立同分布”和“未见”(unseen)两部分,得到数据集:WQCWQ-Unseen. 这里的“未见”和不同于上文中的“零样本”:只要测试集中的样本包含训练集中未见的谓词

即为“未见”.

表1展示了数据集的统计信息,包括各种划分下样本的个数、涉及的谓词总数以及问题的平均长度.

表1　数据集统计

数据集	谓词数	样本数 (Train/Dev/Test)	合计	问题长度
GrailQA-QG	2 973	39 903/4 434/6 763	51 100	10.50
独立同分布	1 266	—/—/1 593(23.55%)	6 763	10.35
组合	145	—/—/1 514(22.39%)		10.63
零样本	305	—/—/3 656(54.06%)		11.38
WQCWQ-Unseen	931	24 477/3 525/6 965	34 967	14.12
独立同分布	503	—/—/3 814(54.76%)	6 965	14.23
未见	768	—/—/3 151(45.24%)		13.80

4.2　实现

本文所提出的转述器和问题生成器均基于生成式预训练语言模型BART[19]. 使用HuggingFace的Transformers[21]库加载BART权重,使用Adam[22]作为训练的优化器. 由于计算资源有限,对于训练转述器,本文直接选择了一套超参,在两个NVIDIA RTX A6000(48 GB) GPU上训练了20轮,花费50小时;对于训练问题生成器,本文在不同数据集上进行了细致的调参. 表2中展示了训练转述器的参数以及训练问题生成器的超参搜索空间. 对于问题生成器,本文以BLEU-4为标准逐个遍历所有可能的参数组合并保存最佳参数. 表3中列出了在不同数据集上问题生成器使用的超参.

4.3　基线方法

本文选择如下的模型作为本文对比的基线:基于预训练模型 本文选择JointGT[7]作为使用预训练语言模型方法的对比基线. JointGT使用BART从知识图谱生成文本,其设计了一种结构感知语义聚合模块,对每个Transformer层增加图结构的输入,然后在大规模的知识图谱—文本语料上进行预训练,最后在问题生成任务上进行微调.

非预训练模型 本文选择CopyNet[23]和Graph2Seq[8]作为不使用预训练语言模型的对比基线. CopyNet是一个序列到序列模型,其设计了一个拷贝机制,能够直接生成输入中的未登录词,是序列生成任务中常见的基线模型. Graph2Seq引入了一个图神经网络对KG中的子图进行编码,并提出了一个基于图节点的拷贝机制用于生成问题.

本文使用这些方法发布的代码和参数,在问题生成数据集上进行了复现.

4.4　评价指标

参照文献[5~8, 24],本文使用BLEU-4(B-4)[25],METEOR(ME)[26]和ROUGE-L(R-L)[27]作为评价指标. BLEU-4和METEOR最初被设计用于评估机器翻译系统,而ROUGE-L用于评估文本摘要系统.

表2　超参搜索空间

模型超参	转述器设定	搜索空间
Learning Rate	3×10^{-5}	[3×10^{-5}, 5×10^{-5}, 10^{-4}]
Warmup Proportion	0.1	[0.1, 0.2, 0.3]
Batch Size	64	[16, 24, 32, 64]
Beam Size	10	[5, 10]
Length Penalty	—	[1.0, 1.2]
Input Length	512	512
Output Length	512	128
Warmup Epoch	10	50
Early Stop Patience	—	10
Maximum Gradient Norm	1.0	1.0
Optimizer	Adam	Adam
Epsilon (for Adam)	10^{-8}	10^{-8}

表3　问题生成器超参设定

数据集	GrailQA		WQCWQ	
	QG	Constrain	Unseen	Constrain
Warmup Proportion	0.2	0.1	0.1	0.1
Learning Rate	3×10^{-5}	3×10^{-5}	5×10^{-5}	3×10^{-5}
Batch Size	32	32	16	32
Length Penalty	1.2	1.0	1.2	1.2

4.5　结果分析

表4中展示了针对测试集中不同泛化级别的问题,本文的方法和基线方法之间的表现. 可以看到本文的方法在两个数据集各种级别的问题上都要好于其他方法. 总体来看,相比于含有逻辑形式的CopyNet,本文方法的BLEU-4在两个数据集上分别提升了2.84/6.33. 这充分说明本文的方法既能够保留语义,又能在生成问题中补充未见谓词的信息.

具体地,本文根据输入数据的形式,将方法划分为两大类,一类的输入是RDF图,另一类含有逻辑形式. 其中,Graph2Seq和JointGT具有相同的输入,观察到在GrailQA-QG上,Graph2Seq的BLEU-4相比于JointGT高出2.27,然而在WQCWQ-Unseen数据集中,前者比后者低了4.91. 这是由数据集的特点导致的,在GrailQA数据集中,给定S表达式,标注员需要将其中的答案实体类型、命名实体和谓词转述为问题,这会导致标注员倾向于直接从中复制单词;而在WQCWQ数据集中,给定SPARQL语句,标注人员的转述更加口语化,且不要求转述答案类型. 如表5所示,问句中的“teach at”需要转述而不是直接从逻辑形式中获取,这更加贴合真实场景,也为模型生成增加了难度.

JointGT首先在KG-to-text语料上预训练,然后在QA数据上微调,因此其泛化能力完全依赖于学习的参数. 实验结果表明,在GrailQA-QG的零样本设定下,其

表现没有带有拷贝机制的Graph2Seq好. 然而在WQCWQ-Unseen数据集的“未见”设定下,JointGT相比于Graph2Seq有了大幅的提升,B-4/ME/R-L分别提升了4.91/5.9/5.54. 此外,对于CopyNet,本文对比了RDF图和逻辑形式两种输入对生成效果的影响,可以看到以逻辑形式为输入能够显著增强模型的表现.

表4 在GrailQA-QG和WQCWQ-Unseen上的实验结果

数据集	输入形式	模型	参数量	独立同分布			组合			零样本			总体		
				B-4	ME	R-L	B-4	ME	R-L	B-4	ME	R-L	B-4	ME	R-L
GrailQA-QG	RDF图	CopyNet	1.7×10^7	40.51	37.91	58.28	24.77	29.11	47.35	19.67	24.89	43.59	25.68	28.64	47.89
	RDF图	Graph2Seq	1.3×10^7	41.63	39.08	60.47	25.11	29.34	48.30	26.92	31.09	49.11	30.21	32.45	51.61
	RDF图	JointGT	1.6×10^8	41.08	39.02	58.82	26.14	31.27	47.93	23.40	31.26	47.44	27.94	32.94	50.23
	逻辑形式(S表达式)	CopyNet	1.7×10^7	42.98	40.00	60.64	27.34	32.18	50.93	22.49	27.92	47.51	28.38	31.45	51.37
	逻辑形式(S表达式)+文本	Ours	1.39×10^8	**44.83**	**39.46**	**59.83**	**30.79**	**32.72**	**51.54**	**25.35**	**30.93**	**48.10**	**31.22**	**33.14**	**51.63**
WQCWQ-Unseen	RDF图	CopyNet	1.7×10^7	30.31	32.27	57.97	—	—	—	20.97	25.69	48.61	26.18	29.30	53.73
	RDF图	Graph2Seq	1.3×10^7	30.94	31.76	57.74	—	—	—	18.69	21.65	43.69	25.69	27.13	51.38
	RDF图	JointGT	1.6×10^8	35.02	35.37	60.23	—	—	—	25.37	30.20	52.91	30.60	33.02	56.92
	逻辑形式(SPARQL)	CopyNet	1.7×10^7	30.43	32.55	58.08	—	—	—	21.44	26.25	48.76	26.46	29.70	53.86
	逻辑形式(SPARQL)+文本	Ours	1.39×10^8	**36.51**	**35.76**	**61.24**	—	—	—	**28.09**	**30.52**	**53.80**	**32.79**	**33.39**	**57.88**

表5 数据集特点

数据集	类别	数据样例
GrailQA	逻辑形式	(AND opera.opera (JOIN opera.opera.librettist "Karl Haffner"))
	标注问题	which opera has the librettist of karl haffner?

WQCWQ	逻辑形式	PREFIX ns: <http://rdf.freebase.com/ns/>
		SELECT DISTINCT ?x
		WHERE {
		?c ns:book.author.book_editions_published "The World As I See It" .
		?c ns:people.person.employment_history ?y .
		?y ns:business.employment_tenure.company ?x .
		?x ns:common.topic.notable_types "College" .
		}

	标注问题	What colleges did the author of The World As I See It teach at?

4.6 消融实验

本节设计了消融实验以分析不同成分的贡献. 首先,移除了转述文本,观察输入只有逻辑形式的效果;其次,将转述文本分别替换为命名实体的实体描述和使用实体名称共现的匹配文本[16],以验证转述文本的有效性. 为减少实验随机性导致的误差,每个设定都运行了5次并取平均值,括号内给出了标准差. 实验结果如表6所示,可以看到,我们的方法显著地优于基线方法,因此本方法是有效的. 此外,生成式方法的评价比较困难,因为同样的含义具有多种表述,特别是B-4值达到到30%后生成的句子已经与原句难以区分,本文的方法稳定超过基线,验证了实验的有效性.

4.7 人工评价

本小节分别对转述器和问题生成器的结果进行了人工评估. 由于这两个模块的目标本质上是一致的,因此使用相同的评估指标.

参照文献[16],本文使用3名评价人员,并要求每人依据如下两个指标进行打分.

(1)谓词覆盖度:生成的文本是否表达了给定逻辑形式(或子图)中所有的谓词. (2)自然度:根据流畅性

表6　消融实验

数据集	方法	独立同分布			组合			零样本			总体(标准差)		
		B-4	ME	R-L	B-4	ME	R-L	B-4	ME	R-L	B-4	ME	R-L
GrailQA-QG	Ours	44.46	39.25	59.01	30.84	32.71	50.96	25.56	30.97	47.63	31.25(0.33)	33.11(0.24)	51.65(0.88)
	无转述文本	43.21	38.19	57.98	27.55	31.37	49.68	22.45	28.05	45.02	29.33(0.82)	31.55(0.81)	49.69(1.27)
	替换为实体描述	41.86	36.81	56.25	28.28	30.65	48.75	23.80	28.57	45.36	29.08(0.81)	30.78(0.78)	48.68(1.12)
	替换为匹配文本	41.02	35.88	54.85	27.70	30.07	47.62	23.28	27.73	44.44	28.52(0.51)	29.83(0.54)	46.87(1.06)
WQCWQ-Unseen	Ours	36.38	35.66	60.37	—	—	—	28.22	30.57	53.16	32.77(0.24)	33.35(0.32)	57.11(0.76)
	无转述文本	35.17	34.49	59.22	—	—	—	26.73	29.20	51.51	31.42(0.57)	32.09(0.53)	55.73(1.01)
	替换为实体描述	34.60	34.17	58.79	—	—	—	26.12	28.73	50.97	30.83(0.48)	31.70(0.59)	55.25(0.99)
	替换为匹配文本	34.33	33.30	57.18	—	—	—	26.96	29.02	50.65	30.69(0.28)	31.91(0.33)	54.17(0.61)

和可读性给生成的文本打分,从1到5分,其中5分为非常清晰自然,3分为语法正确但拗口,1分为完全无法理解.

对于评估转述器,本文分别从两个数据集中筛选了100个原子级子图;对于评估问题生成器,本文分别从两个数据集的测试集中抽取了100个样例. 本文选择同为基于预训练语言模型的JointGT作为比较对象. 评估结果见表7,可以看到,转述器能够准确并且流畅地转述谓词,问题生成器能够生成高质量的问题. 值得注意的是,问题生成器在GrailQA-QG数据集上生成文本的可读性比标注的答案还要好,这进一步验证了上述GrailQA数据集标注的问题不够自然的观察. 总体来说,本文的方法能够合理转述谓词,并且基于转述的文本能够生成高质量的问题.

4.8　案例分析

本小节以WQCWQ中的数据为例,直观展示了本方法的流程并和基线模型进行了对比. 如图4所示,在选取的样例中,谓词"film. director. film"是未见的. 第1步,在知识图谱中执行并实例化整个SPARQL,获得完整的RDF子图;第2步,将RDF子图拆解;第3步,对于每一个原子级子图,按照上述规则将其序列化;第4步,使用训练好的子图转述器获得每个子图的转述文本;最后,给定逻辑形式和转述文本,问题生成器生成最终的问题. 表8列出了不同方法的生成结果. Graph2Seq对于未见的谓词泛化能力有限,其错误地拷贝了实体;JointGT则没有表达出"导演"的语义:Ron Howard"导演(direct)"了该电影,而不是"参演(star in)". 相比之下,本文的方法成功生成了高质量的问题. 并且,在转述文本中,模型完整地生成了"被导演(is directed by)"的语义,这再一次证明了转述文本能够为模型提供有效的信息.

表7　人工评价

评价模块		GrailQA-QG		WQCWQ-Unseen	
		谓词覆盖度	自然度	谓词覆盖度	自然度
转述器(子图类型)	single	78% (4.9)	4.11 (0.19)	82% (5.4)	4.35 (0.16)
	CVT	69% (5.4)	3.89 (0.22)	71% (7.3)	3.65 (0.25)
问题生成器(模型)	JointGT	83% (3.8)	4.28 (0.19)	84% (3.3)	4.27 (0.18)
	Ours	85% (4.2)	4.30 (0.16)	88% (3.9)	4.85 (0.17)
答案	标注的问题	98% (1.9)	4.20 (0.22)	95% (2.1)	4.81 (0.17)

表8　案例分析

模型	生成的问题
Graph2Seq	what clay animation did clay animation direct?
JointGT	what clay animation movie did ron howard star in ?
Ours	what clay animation movie was directed by ron howard?

4.9　错误分析

表9中展示了一些生成错误的案例,为了方便展示,本文将表中逻辑形式的变量和实体ID替换为对应的名称. 转述器最常见的错误是生成不存在的事实,例如表9中的样例,庞培剧院(the Theatre of Pompey)是在罗马共和国后期由庞培大帝(Pompey the Great)建造

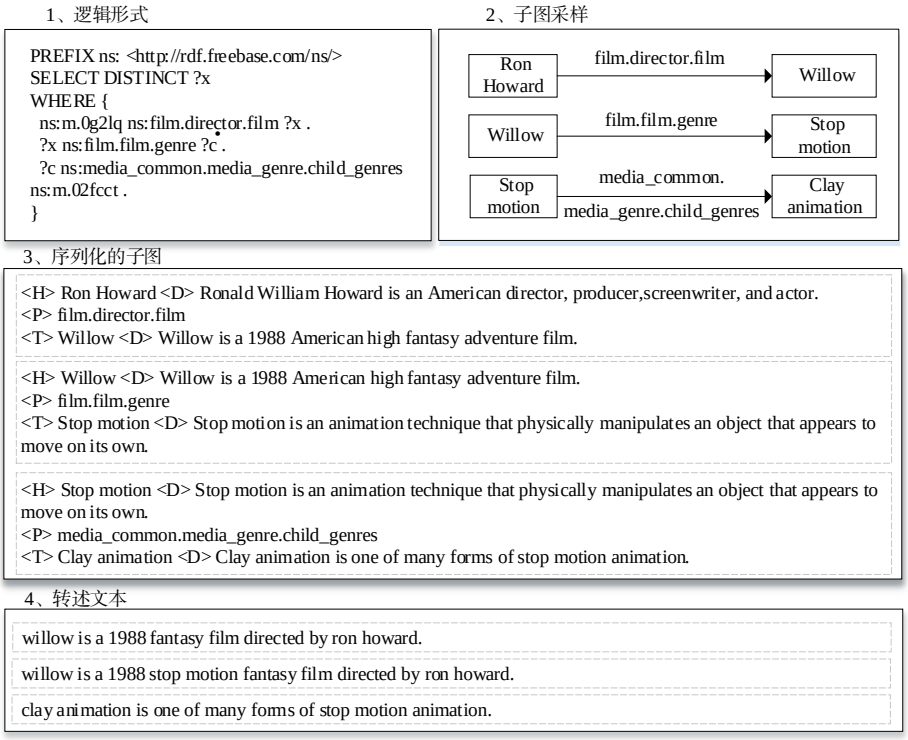

图 4 案例分析

表 9 错误分析

逻辑形式	SELECT DISTINCT ?x WHERE {
	"Julius Caesar" people.deceased_person.place_of_death "The Theatre of Pompey" (?x).
	}

转述文本	The Theatre of Pompey was built during the reign of Julius Caesar.
标注的问题	Where was Caesar when he was stabbed?
Ours	Where did Julius Caesar die?

的,而不是朱利叶斯·凯撒(Julius Caesar). 这是远程监督方法的常见问题.

5 结论

本文提出了一种基于子图转述的问题生成方法,能够表征复杂语义,且对于未见的谓词也有较好的效果. 本文使用SPARQL或者S表达式作为模型输入,利用预训练语言模型在大规模无监督数据上训练了子图转述器,能够对包含未见谓词的子图提供多样的表述,用于生成问题. 实验结果表明,本文的方法达到了最先进的性能.

参考文献

[1] 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.

[2] BAO J W, TANG D Y, DUAN N, et al. Text generation from tables[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019, 27(2): 311-320.

[3] 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). Stroudsburg: Association for Computational Linguistics, 2016: 201-206.

[4] LIU C, LIU K, HE S Z, et al. Generating questions for knowledge bases via incorporating diversified contexts and answer-aware loss[C]//Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP. Stroudsburg: Association for Computational Linguistics, 2019: 2431-2441.

[5] BI S, CHENG X Y, LI Y F, et al. Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases[C]//Proceedings of the 28th International Conference on Computational Linguistics. Strouds-

burg, International Committee on Computational Linguistics, 2020: 2776-2786.

[6] KUMAR V, HUA Y C, RAMAKRISHNAN G, et al. Difficulty-controllable multi-hop question generation from knowledge graphs[C]//The Semantic Web-ISWC 2019. New York: ACM, 2019: 382-398.

[7] KE P, JI H Z, RAN Y, et al. JointGT: Graph-text joint representation learning for text generation from knowledge graphs[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg, Association for Computational Linguistics, 2021: 2526-2538.

[8] CHEN Y, WU L F, ZAKI M J. Toward subgraph guided knowledge graph question generation with graph neural networks[EB/OL]. (2020-03-13)[2022-05-20]. http://arxiv.org/abs/2004.06015.

[9] GU Y, KASE S E, VANNI M, et al. Beyond I.I.D.: Three levels of generalization for question answering on knowledge bases[C]//Proceedings of the Web Conference 2021. New York: ACM, 2021: 3477-3488.

[10] LAN Y S, HE G L, JIANG J H, et al. A survey on complex knowledge base question answering: Methods, challenges and solutions[EB/OL]. (2021-05-25)[2022-05-20]. http://arxiv.org/abs/2105.11644.

[11] 肖仰华, 徐波, 林欣, 等. 知识图谱: 概念与技术[M]. 北京: 电子工业出版社, 2020.
XIAO Y H, XU B, LIN X, et al. Knowledge Graph[M]. Beijing: Publishing House of Electronics Industry, 2020. (in Chinese)

[12] SONG L F, ZHAO L. Question generation from a knowledge base with Web exploration[EB/OL]. (2016-10-12)[2022-05-20]. http://arxiv.org/abs/1610.03807.

[13] SEYLER D, YAHYA M, BERBERICH K. Generating quiz questions from knowledge graphs[C]//Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015: 113-114.

[14] SEYLER D, YAHYA M, BERBERICH K. Knowledge questions from knowledge graphs[C]//Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. New York: ACM, 2017: 11-18.

[15] BAO J W, TANG D Y, DUAN N, et al. Table-to-text: Describing table region with natural language[EB/OL]. (2018-05-29)[2022-05-20]. http://arxiv.org/abs/1805.11234.

[16] ELSAHAR H, GRAVIER C, LAFOREST F. Zero-shot question generation from knowledge graphs for unseen predicates and entity types[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 218-228.

[17] 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)

[18] 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.

[19] LEWIS M, LIU Y H, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the ACL. Stroudsburg: Association for Computational Linguistics, 2020: 7871-7880.

[20] TALMOR A, BERANT J. The web as a knowledge-base for answering complex questions[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 641-651.

[21] WOLF T, DEBUT L, SANH V, et al. Transformers: state-of-the-art natural language processing[C]//Proceedings of the 2020 Conference on EMNLP: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2020: 38-45.

[22] Kingma D, Ba J. Adam: A method for stochastic optimization[EB/OL]. (2014-12-22) [2022-05-20]. http://arxiv.org/abs/1412.6980.

[23] 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 (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2016: 1631-1640.

[24] 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 [J]. Journal of Chinese Information Processing, 2021, 35 (7): 1-9. (in Chinese)

[25] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A

method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2002: 311-318.

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

[27] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74-81.

[28] 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)

作者简介

温立强 男,1991 年 9 月出生于山西省孝义市. 现为北京大学软件与微电子学院前沿工程博士生,主要研究领域为软件工程、知识图谱问答.
E-mail: wenlq@pku.edu.cn

熊冠铭 男,1996 年 8 月出生于福建省三明市. 毕业于北京大学软件与微电子学院. 主要研究知识图谱问答.
E-mail: gm_xiong@qq.com

王 宇 男,1978 年出生,辽宁沈阳人. 北京大学软件与微电子学院博士研究生,主要研究方向为知识图谱构建和自然语言处理.
E-mail: wangyu_cn@stu.pku.edu.cn

陈一朴 男,1992 年出生,河南禹州人. 现为北京北大软件工程股份有限公司数据智能研究院算法工程师,主要研究领域为推荐系统、自然语言处理.
E-mail: eap@buaa.edu.cn

李伟平 男,1973 年 3 月,辽宁凌源人,北京大学教授,主要研究方向为大数据分析,信息抽取.
E-mail: wpli@ss.pku.edu.cn

赵 文 男,1967 年出生,辽宁大连人. 现为北京大学软件工程国家工程研究中心研究员、博士生导师,主要研究领域为软件工程、软件安全.
E-mail: zhaowen@pku. edu. cn