

文章编号: 1003-0077(2024)04-0001-16

表格问答研究综述

张洪虞, 李 韧, 杨建喜, 杨小霞, 肖 桥, 蒋仕新, 王 笛

(重庆交通大学 信息科学与工程学院, 重庆 400074)

摘 要: 表格问答通过自然语言问句直接与表格数据进行交互并得到答案, 是智能问答的主要形式之一。近年来, 研究人员利用以语义解析为主的技术在该领域开展了深入研究。该文从不同表格类型分类及其问答任务问题定义出发, 将表格问答细分为单表单轮、多表单轮、多表多轮式问答三种任务, 并系统介绍了各类表格问答任务的数据集及其代表性方法。其次, 该文总结了当前主流表格预训练模型的数据构造、输入编码以及预训练目标。最后, 探讨当前工作的优势与不足, 并分析了未来表格问答的前景与挑战。

关键词: 表格问答; 语义解析; 自然语言处理; 综述

中图分类号: TP391

文献标识码: A

Researches on Question Answering Over Tables: A Survey

ZHANG Hongyi, LI Ren, YANG Jianxi, YANG Xiaoxia, XIAO Qiao, JIANG Shixin, WANG Di

(School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China)

Abstract: Table question answering (Table QA) directly gets answers form table data through natural language, which is one of the main forms of intelligent question answering. Recently, researchers pay great attention to resolve this task by semantic parsing. In this paper, we divide Table QA tasks into three types: single-table single-turn, multi-table single-turn, and multi-table multi-turn. This paper provides a systematic introduction to datasets and representative methods of various types of Table QA tasks. It also summarizes the data construction, input encoding, and pre-training objectives of the table pre-training models. Finally, we explore the strengths and weaknesses of current work, and discuss the future prospects and challenges of Table QA.

Keywords: table question answering; semantic parsing; natural language processing; survey

0 引言

随着信息化系统在众多行业广泛应用, 电子文档中嵌入的各类型表格及关系数据库表承载了丰富的信息, 成为支撑领域信息交互和智能决策的重要数据基础^[1]。例如, 医疗领域已积累了海量的电子病历文档, 其中蕴含了大量的表格信息。如何准确、快速地从挖掘出有价值的医学信息, 为疾病风险评估、智能辅助诊疗及医疗知识问答等提供支撑, 已成为智慧医疗研究领域的重要任务^[2]。在桥梁管理养护领域, 定期检测报告中存在大量表格, 众多桥梁管理信息系统的数据库表中也记录了细粒度结构病

害、养护建议等信息, 是实现桥梁智能管养的重要基础^[3]。目前, 以 SQL 为典型代表的结构化查询语言是访问关系型数据库表数据的主要途径, 但普通系统用户不具备编写 SQL 语句的技术基础^[4]。因此, 如何使用自然语言等交互方式, 实现表格信息智能检索与问答, 已成为学术界和产业界关注的热点问题之一。现有解决方案主要采用语义解析的方式实现用户问题与表格数据的交互。其中, Test-to-SQL 是其关键技术。该类方法主要实现从自然语言问句到 SQL 查询语句的映射, 并将生成的 SQL 语句输入到 SQL 引擎中检索出对应的答案。因此, Text-to-SQL 映射准确率直接影响了其问答效果。

智能问答作为人工智能研究的一项重要任务,

收稿日期: 2022-07-31 定稿日期: 2023-03-15

基金项目: 国家自然科学基金(62003063); 重庆市教委科学技术研究项目(KJQN202200720); 重庆交通大学研究生科研创新项目(2023s0084)

综合运用了自然语言处理、深度神经网络及知识图谱等多种方法。其中,知识图谱问答以基于三元组结构的语义网络为数据源通过信息检索或语义解析两种模式实现交互式问答^[5]。例如,信息检索式知识图谱问答首先识别问题主题词,并通过实体匹配和答案生成等步骤完成问答任务。以文本段落作为数据源的机器阅读理解任务要求计算机根据给定上下文回答特定问题,也是智能问答领域研究热点之一^[6]。表格信息抽取任务主要从表格数据中抽取语义一致的结构化信息,是知识图谱构建等任务的基础。另外,表格问答以文档表格、数据库表等为数据源,将自然语言问题语句与表格中的结构化或半结构化信息相匹配获取答案。表格问答不仅可以直观地回答自然语言问题,还可以通过排序、聚合等操作推理出复杂答案,具备较强实际应用价值。然而,与机器阅读理解和知识图谱问答相比,表格数据存储形式复杂、多样,在表格的上下文语义理解、问题解析与匹配等方面均面临更大挑战。如图 1 所示,表格问答最初以自然语言-数据库接口形式提出,但由于其复杂程度导致其在很长时间内发展相当缓慢。随着深度学习的发展,以 Text-to-SQL 为主的语义解析成为研究热点。2017 年以来,随着 WiKiSQL^[7]数据集的发布,单表单轮式表格问答获得广泛关注。随着单表单轮式表格问答技术的日渐成熟,从实际应用出发的多表单轮式表格问答逐步兴起。随后,与对话系统结合,表格问答扩展到多轮的形式。2021 年起,随着预训练模型在表格问答领域的探索,各类表格问答任务性能均得到大幅提升。

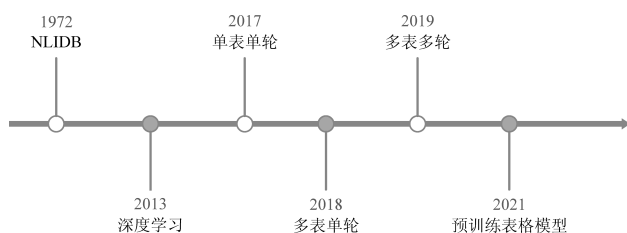


图 1 表格问答研究发展脉络

目前,已有研究人员从不同角度对表格问答相关工作做出了总结回顾。李智等人^[8]从语义解析的角度分析了基于表格知识库的问答模型、算法特点及其相关问题,但该综述仅介绍了传统的语义解析方法,未系统总结目前表格问答领域前沿的预训练模型等内容。Zhang 等人^[9]研究了 Web 表格的提取、搜索、问答等任务,并指出不同任务之间的依赖关系,但该综述仅针对 Web 表格,对其他表格形式

尚未做出深入讨论,且表格问答任务未得到详细分析。Dong 等人^[10]从预训练模型角度出发,介绍了表格预训练模型设计、训练目标和下游任务,探讨了现有表格预训练模型,但未聚焦表格问答任务。Qin 等人^[11]将 Text-to-SQL 语料分为单轮和多轮,分别总结其编码器和解码器架构,并详细阐述了现有预训练表格模型的预训练目标。梁清源等人^[4]从 SQL 生成场景、数据集、模型结构和评估方法层面对基于深度学习的 SQL 生成研究进行分类综述。但该文仅介绍了基于深度学习的 SQL 生成方法,未对传统的方法以及表格问答的发展脉络进行总结。

与上述工作不同的是,本文从不同表格类型分类及其问答任务问题定义为出发点,系统性地总结各类表格问答任务的数据集及其代表性方法,介绍主流表格预训练模型,探讨当前工作的优势与不足,展望表格问答研究的前景与挑战。

1 问题定义

1.1 表格类型分类

表格可以根据其结构化程度分为结构化表格、半结构化表格和非结构化表格^[10]。关系型数据库表是最具代表性的结构化表格,其主键是表中信息的唯一性标识。以 Web 页面表格和 Word 文档表格等为代表的半结构化表格通常由工具软件或标记语言人工构建,并以二维矩阵形式存在,但其缺乏足够的元信息描述表结构和数据类型。除此以外,还有大量只包含原始 RGB 信息的非结构图像表格,例如,书籍中的扫描表、收据图片和手写草稿等。在处理非结构化表格时,首先需要通过光学符号识别和单元格分割对其进行数字化。本文主要针对以结构化和半结构化表格作为数据源的智能问答任务场景。

1.2 表格问答任务分类及形式化定义

参考 Qin 等人^[11]的分类方法,本文将表格问答任务根据问答轮次和涉及表格数量分为单表单轮、多表单轮和多表多轮问答三类,并将其形式化定义为:根据给定的自然语言问句 $Q = \{Q_u\}_{u=1}^U$ 及表格 $T = \{t_j\}_{j=1}^C$,从 T 中根据问题 Q 查询相关信息,并生成答案 $A = \{A_u\}_{u=1}^U$ 。其中, $Q_u = \{q_i\}_{i=1}^{n_u}$, $A_u = \{a_k\}_{k=1}^{m_u}$, U 为问句数量, n_u 和 m_u 分别为第 u 个问句和答案包含的字符数, C 为问答涉及的表格数目。

当 $C=1$ 且 $U=1$ 时, 为单表单轮式问答; 当 $C>1$ 且 $U=1$ 时, 为多表单轮式问答; 当 $C>1$ 且 $U>1$ 时, 为多表多轮式问答。

单表单轮是指只围绕单个表格进行的单轮问答, 该任务涉及到的问句类型较为简单, 通常为基本的检索式问题。然而在实际应用中, 由于关系数据库表等表格数据通常以外键引用等形式相互关联, 导致多表单轮问答涉及多个表的嵌套或联合查询, 对应的 SQL 语句涉及到 Join、Union 等复杂算子。另外, 单轮表格问答可以进一步扩展为多轮问答, 并结合上下文信息实现对话式信息交互。在多轮问答任务场景下, 问答系统需要关联历史对话信息, 使得多轮问答更具挑战性。

1.3 表格问答方法分类

单表单轮式表格问答任务仅涉及单一表格, 问题多为简单检索式, 传统基于规则的方法可以覆盖大部分简单问题 SQL 模板。另外, 分块机制将 SQL 分解为子句进行预测, 部分研究成果引入类型信息, 辅助模型理解表格内容和结构。本文将单表单轮式表格问答方法分为基于规则的方法、基于分块机制的方法和基于类型信息的方法。

多表单轮式表格问答涉及多张表格上下文交互, 需要更准确的表格信息选择和答案定位。一些研究通过引入多表查询 SQL 语法指导不同的 SQL 组件解码, 具有良好的语法约束力。另外, 图神经网络能较好地编码主、外键信息, 有效提取表格列与自然语言问句中的关联信息, 提升问答准确率。部分成果还通过引入中间表示力争克服自然语言问题与 SQL 之间的不匹配问题。因此, 本文将多表单轮式表格问答方法分为基于 SQL 语法的方法、基于图神经网络的方法和基于中间表示的方法。

多表多轮式表格问答结合上下文信息实现对话式交互, 问答系统需要关联历史对话信息。目前的研究成果表明, 通过实现上下文信息传递和数据增强, 可以有效提升多表多轮式表格问答的性能。本文将多表多轮式表格问答方法分为基于上下文的方法和基于数据增强的方法。

近年来, 预训练模型大幅提升各类自然语言处理任务的性能, 表格预训练也成为业界关注的重点, 是表格问答研究的主流方法和未来趋势。本文从数据构造、输入编码和预训练目标三个方面对现有的表格预训练方法进行总结和分析。

2 单表单轮式表格问答任务

单表单轮式表格问答是该领域研究中的基础任务, 国内外已开展了较多研究。本节首先阐述该任务涉及的公有数据集及其评测标准, 然后分析当前主流方法及其评测效果。

2.1 数据集及评测标准

目前, 业界推出的单表单轮式表格问答数据集有 WiKiSQL^[7]、WikiTableQuestions (WikiTQ)^[12]和 ATIS^[13]等。其中, ATIS 是 1990 年提出的航空旅行领域数据集, 由 27 个表及其查询语句组成, 包含票价、航空公司、航班和城市等出行相关信息。WikiTQ 是 2015 年斯坦福大学提出的弱监督语义解析式数据集, 包含 22 203 个问答句子和 2 108 张表格, 仅提供问答对而不包含逻辑形式。WikiTQ 还包含聚合操作等复杂逻辑推理, 使得该任务更具挑战性。WiKiSQL 包含 24 241 张表、80 654 个问题及 77 840 条 SQL 语句, 其数据样例如图 2 所示。

```
{
  "phase": 1,
  "question": "who is the manufacturer for the order year 1998?",
  "sql": {
    "conds": [
      [
        0,
        0,
        "1998"
      ]
    ],
    "sel": 1,
    "agg": 0
  },
  "table_id": "1-10007452-3"
}
```

图 2 WiKiSQL 数据样例

其中, “phase”表示数据构造阶段, WiKiSQL 的数据构造包括数据生成和验证两个阶段。“question”记录用户提出的自然语言问题。“sql”字段表示问题所对应的 SQL 查询, 其中“conds”为查询条件的三元组列表, 包括所选条件列索引“column_index”、条件运算符索引“operator_index”, 以及条件值“condition”。图 2 中的 0、0、“1998”分别对应表中“year”列索引、“=”索引、条件值“1998”。“sel”字段代表查询所选列索引, 图 2 中的 1 代表“manufacturer”列在表格中的索引。“agg”表示使用的聚合运算符索引, 若未使用聚合运算则值为 0。“table_id”表示该问题所对应表格的 ID。

TableQA^[14]是一个中文表格问答数据集,以财务报告和公共网站为数据源,包含 64 891 个问题以及 20 311 条查询语句,涉及 6 000 多个表。考虑到同一实体或概念有不同的表达方式,TableQA 在标注时强调实体链接问题,约 30% 的数据需要解决实体链接问题。除了公共领域的数据集,一些特定领域也针对其领域特点构建了表格问答数据集。如以财务报表为数据源的金融领域数据集 FinQA^[15]和 TAT-QA^[16],包含数值推理等领域需求。表 1 描述了主流的单表单轮式表格问答数据集相关统计信息。

表 1 单表单轮式表格问答常用数据集

数据集	问题数量	SQL 数量	表格数量	语言
ATIS ^[13]	5 280	947	27	英文
WikiTQ ^[12]	22 203	—	2 108	英文
WiKiSQL ^[7]	80 654	77 840	24 241	英文
TAT-QA ^[16]	16 552	—	—	英文
FinQA ^[15]	8 281	—	—	英文
TableQA ^[14]	64 891	20 311	6 029	中文

单表单轮式表格问答任务的评测指标包括逻辑形式准确率(Logical Form Accuracy)和执行结果准确率(Execution Accuracy)。逻辑形式准确率将模型合成的 SQL 查询与真实的 SQL 查询进行比较,其计算如式(1)、式(2)所示。

$$\text{Score}_{\text{lf}} = \begin{cases} 1, & \text{SQL}' = \text{SQL} \\ 0, & \text{SQL}' \neq \text{SQL} \end{cases} \quad (1)$$

$$\text{Acc}_{\text{lf}} = \frac{1}{N} \sum_{i=1}^N \text{Score}_{\text{lf}}^i \quad (2)$$

其中,SQL'和 SQL 分别代表预测的 SQL 语句和真实的 SQL 语句,Acc_{lf}代表逻辑形式匹配准确率,N 为问题总数。

执行结果准确率将生成的 SQL 语句的执行结果与正确答案进行比较,其计算如式(3)、式(4)所示。

$$\text{Score}_{\text{ex}} = \begin{cases} 1 & Y' = Y \\ 0 & Y' \neq Y \end{cases} \quad (3)$$

$$\text{Acc}_{\text{ex}} = \frac{1}{N} \sum_{i=1}^N \text{Score}_{\text{ex}}^i \quad (4)$$

其中,Y'和 Y 分别代表预测 SQL 语句的执行结果和真实 SQL 语句的执行结果,Acc_{ex}代表执行结果准确率,N 为问题总数。

2.2 单表单轮式表格问答方法

单表单轮式任务是表格问答研究的基础性工作,涉及问题较为简单,目前问答评测效果已经超越人类。本文将单表单轮式表格问答分为基于规则的方法、基于分块机制的方法和基于类型信息的方法。表 2 展示了单表单轮数据集上各类表格问答方法的性能对比。

表 2 单表单轮式表格问答方法性能及其优缺点对比

模型	WiKiSQL		ATIS(Acc _{ex}) /%	WikiTQ(Acc _{ex}) /%	模型优点	模型缺点
	Acc _{lf} /%	Acc _{ex} /%				
SQLNet ^[17]	61.3	68.0	—	—	模型简单,速度快	无法处理复杂 SQL
Coarse2Fine ^[18]	73.0	79.6	95.1	—	共享参数,泛化能力强	无法处理复杂 SQL
TypeSQL ^[19]	66.7	73.5	—	—	特定领域效果好	泛化能力差
TaBERT ^[20]	—	—	—	52.3	表征能力强	所需数据量大
GraPPa ^[21]	—	82.2	—	42.0	表征能力强	所需数据量大
SeaD ^[22]	87.5	93.0	—	—	特定数据集效果好	无法处理复杂 SQL
TAPEX ^[23]	—	89.6	—	57.5	数据集简单	可解释性差
TaPas ^[24]	—	86.4	—	42.6	架构简单	所需数据量大

2.2.1 基于规则的单表单轮式表格问答

由于 SQL 查询语句具有特定的语法结构,早期的表格问答研究多采用基于规则的方法。Woods^[25]通过上下文无关文法解析自然语言问题的语法结构,通过执行语义动作生成结构化查询语句。Das

等人^[26]提出了一个基于统计学的概率框架语义模型,在上下文中提取关键词并预测框架语义结构,根据文本语义谓词参数结构回答问题。给定一个问题,Popescu 等人^[27]首先判断其在语义上是否可处理,如果可以,则输出对应的 SQL 查询,如果不行,

则回答无法匹配,以防止因错误回答导致用户体验不佳。Popescu 等人^[27]首先根据英文问题分类提出了以“wh”开头的问题类型集合{"what", "which", "where", "who", "when"},然后将语义相近的单词加入同一数据库元素集合,通过图匹配将问题单词映射到数据库并通过模板合成最终 SQL 查询。基于规则的语义解析方法结构清晰,可解释性强,对特定领域效果较好。然而产生式、规则集合等模板的编写依赖于领域专家,人力资源消耗大,且领域规则无法推广至其他领域,泛化性较差。

2.2.2 基于分块机制的单表单轮式表格问答

基于分块机制的问答方法将 SQL 语句分解为不同子句,以分块预测的方式将子句的预测结果拼接成目标 SQL。考虑到 Seq2Seq 模型处理同语义 SQL 的局限性,SQLNet^[17]提出了一种基于草图的方法,以槽填充的方式实现语义解析。草图为生成的部分结构化语句,本质上属于问答任务的中间信息。如图 3 所示,方框表示待预测的插槽,有向边表示其依赖关系,SQLNet 只需要在插槽中填充内容,无需考虑输出语法。为实现内容填充,Xu 等人^[17]额外提出了一个 Seq2Set 模型来预测无序的约束子集,进一步简化了生成任务,并通过列注意机制来捕获草图中预定义的依赖关系。基于草图的方法通用性高,生成目标灵活,针对不同问题可复用性强。

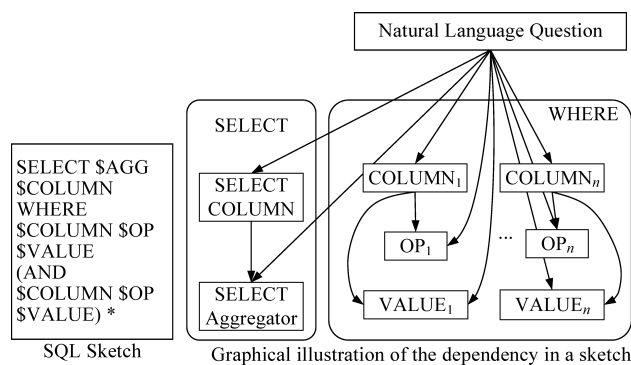


图 3 SQLNet 模型图^[17]

Dong 等人^[18]提出了一种结构感知的神经网络,将语义解析过程分解为粗粒度和细粒度两个阶段。该模型首先生成隐藏底层信息的语义草图,然后预测缺失细节,最终获取完整的语义表示。与 SQLNet 等一次性解码的模型相比,从粗到细两阶段解码将高层语义信息与低层语义信息分离,使得模型可以共享粗粒度结构知识,更容易适应不同的领域,且提高了任务间的性能。

2.2.3 基于类型信息的单表单轮式表格问答

基于类型信息的问答方法通过引入实体类型信息辅助模型更好地理解自然语言问题和表格数据,以提升问答精度。例如,ColNet 模型^[28]基于列类型标注的框架,将实体列类型信息嵌入向量空间,帮助模型理解表格数据类型,以提升性能。Gol 等人^[29]将单元格语义和上下文信息嵌入低维单元空间,并通过基于 RNN 的分类器提升复杂文档中单元格类型的分类性能。另一方面,TypeSQL^[19]利用类型信息辅助理解自然语言问题中的稀有实体和数字,以槽填充的方式实现语义解析。TypeSQL 将问题中的数字和日期分为四个类别:整数、浮点数、日期和年份。为了识别命名实体,TypeSQL 设置了个人、地点、国家、组织和运动五种类型实体,涵盖了数据集中大多数实体,最大限度地融合了文本类型。虽然引入类型信息能有效提升问答效果,但同时需要额外的数据标注和编码空间。

3 多表单轮式表格问答任务

3.1 数据集及评测标准

为契合真实应用场景,表格问答数据集从单表问答扩展为多表联合问答。Spider^[30]是耶鲁大学在 2018 年提出的一个大规模跨领域复杂语义解析数据集,由 10 181 个问题和 5 693 个复杂 SQL 查询组成,涉及 200 个数据库和多个表,内容涵盖了 138 个不同的领域。2019 年,Min^[31]等人将 Spider 翻译为中文并命名为 CSpider,是第一个大规模跨领域中文语义解析数据集。DuSQL^[32]是一个大规模中文跨领域 Text-to-SQL 数据集,由 23 797 个问题-SQL 对组成,涵盖 200 个数据库和 813 个表格,包含了大量涉及行列计算的 SQL 语句。表 3 展示了多表单轮式表格问答数据集信息。

表 3 多表单轮式表格问答数据集

数据集	问题规模	SQL 规模	表格规模	数据库规模	语言
Spider ^[30]	10 181	5 693	4 504	200	英文
CSpider ^[31]	9 691	5 263	5 693	200	中文
DuSQL ^[32]	23 797	23 797	813	200	中文

多表单轮表格问答的评测指标包括精确集合匹配率(Exact Set Match)和执行结果准确率(Execution

Accuracy)。精确集合匹配率衡量预测 SQL 查询在语义上是否等价于真实 SQL 查询,其计算方式如式(5)、式(6)所示。

$$\text{Score}_{\text{em}} = \begin{cases} 1 & Y' = Y \\ 0 & Y' \neq Y \end{cases} \quad (5)$$

$$\text{Acc}_{\text{em}} = \frac{1}{N} \sum_{i=1}^N \text{Score}_{\text{em}}^i \quad (6)$$

其中, $Y' = \{(k'_i, v'_i), i \in (1, m)\}$, $Y = \{(k_i, v_i), i \in (1, m)\}$, 分别代表预测 SQL 语句和真实 SQL 语句的组件集合, k 表示 SQL 组件元素, v 代表其对应值, m 为组件数量。 Acc_{em} 代表精确集合匹配率, 可以解决字符串匹配中语义相同 SQL 顺序的误判问题。

执行结果准确率需要在给定的真实值列表中选择某一值并将其填充到预测 SQL 的正确插槽中, 比较生成 SQL 语句的执行结果与真实结果, 如式(7)、式(8)所示。

$$\text{Score}_{\text{ex}} = \begin{cases} 1 & V' = V \\ 0 & V' \neq V \end{cases} \quad (7)$$

$$\text{Acc}_{\text{ex}} = \frac{1}{N} \sum_{i=1}^N \text{Score}_{\text{ex}}^i \quad (8)$$

其中, V' 和 V 分别代表预测 SQL 语句的执行结果和真实 SQL 语句的执行结果, Acc_{ex} 代表执行结果准确率。执行结果准确率可能会产生错误评估, 因为当预测 SQL 与真实 SQL 在语义上不同时返回的结果可能相同(如 NULL)^[30]。

3.2 多表单轮式表格问答方法

多表单轮式表格问答联合多张表格进行交互, 对问题所链接的表格信息的选择依赖性更强。本文将多表单轮式表格问答方法分为基于 SQL 语法的方法、基于图神经网络的方法和基于中间表示的方法, 表 4 展示了现有方法在 Spider 数据集上的性能对比。

表 4 多表单轮式表格问答方法在 Spider 数据集上的性能及其优缺点对比

模型	All	Easy	Medium	Hard	Extra Hard	优点	缺点
SyntaxSQLNet ^[31]	27.2	48.0	27.0	24.3	4.6	符合 SQL 语法	时间人力成本高
Context-GNN ^[32]	39.4	—	—	—	—	有效编码主外键信息, 提升复杂 SQL 生成能力	无法推理
Global-GNN ^[33]	47.4	—	—	—	—	零样本 SQL 生成	限定处理预定义 SQL 类型
Adaptive-GNN ^[34]	60.1	—	—	—	—	自适应编码	图谱构建未考虑 WHERE 条件
RAT-SQL ^[35]	57.2	74.8	60.7	53.6	31.5	扩展细化关系边	约束多
IRNet ^[36]	46.7	70.1	49.2	39.5	19.1	扩展语法树	约束多
IRNet(BERT) ^[36]	54.7	77.2	58.7	48.1	25.3	扩展语法树	约束多
GAP ^[37]	69.7	87.2	75.1	63.7	41.2	表征能力强	所需数据量大
TABERT ^[20]	65.2	—	—	—	—	表征能力强	所需数据量大
GraPPa ^[21]	69.6	—	—	—	—	表征能力强	所需数据量大
SDCUP ^[38]	70.7	—	—	—	—	表征能力强	所需数据量大

3.2.1 基于 SQL 语法的多表单轮式表格问答

多表联合问答涉及到的 SQL 语句较为复杂, 包括主键、外键等信息, 很难将其拆解为分类问题进行预测。主流方法在解码的过程中引入抽象语法树结构, 利用树解码的方式建模 SQL 语法规则。SyntaxSQLNet^[31]提出一种基于 SQL 特定语法的树解码器, 以递归的形式来管理 SQL 的生成, 并提供 SQL 路径的历史记录来帮助预测下一个 SQL 单词。然而, SyntaxSQLNet 需要通过对 SQL 模式进行分组, 然后手动编辑选定的 SQL 模式及其相应的

问题列表来实现数据增强, 这将耗费大量时间和人力资源。Huo 等人^[39]通过图卷积神经网络共享列名跨领域连接所有表和数据库, 以实现更好的泛化, 有效改进 SyntaxSQLNet。基于 SQL 语法的表格问答方法具有良好的语法约束力, 生成的 SQL 可执行率较高。然而, SQL 语法模块分类需要额外的人工成本投入, 且递归解码方式时间复杂度较大。

3.2.2 基于图神经网络的多表单轮式表格问答

图神经网络能够有效提取表格列与自然语言问句中的关联信息, 提升多表单轮式表格问答准确率。

曹合心等人^[34]提出了一种自适应的图构建和图编码方式,通过对问句和数据库的拼接词向量进行卷积操作生成图神经网络初始化权重,并以图邻接矩阵的方式编码关系信息,大幅提升了 Spider 数据集上的准确率。如图 4 所示,Bogin 等人^[32]将表格结构编码为图的形式,并通过图神经网络为每个节点提供全局表示。模型编码时以表名和列名作为结点,将表格与所包含的列用双向边连接作为第一类边,主键指向外键的有向边为第二类边,外键指向主键的有向边为第三类边。然后根据问题输入对图表示进行修剪,通过计算问题单词与模式项之间的相关度筛选有价值的结点。解码阶段采用基于语法的 LSTM 解码器,通过语法规则进行输出,并引入自注意力机制链接相关性最高的表。此外,针对零样本多表问答,Bogin 等人^[33]提出全局图神经网络

对问题所对应的数据库进行评分,并通过重排器选择最终结果,进一步提升了 Spider 数据集的准确率。

Bogin 等人^[32]通过图神经网络对模式外键关系有向图进行编码,解决了模式表示问题。然而,该方法无法将模式编码与问题联系起来,从而很难对模式链接进行推理。此外,该模型将模式编码过程中的信息传播限制为预定义的外键关系图,无法覆盖所有情况。为了解决上述问题,RAT-SQL^[35]提出基于关系感知的自注意力机制,将对模式实体和问题单词的全局推理与对预定义模式关系的结构化推理进行结合,解决了 Text-to-SQL 任务中编码器的模式编码、模式链接和特征表示问题。基于图神经网络的方法能够更好地表示 SQL 间的主、外键关系,适应更加复杂的结构。

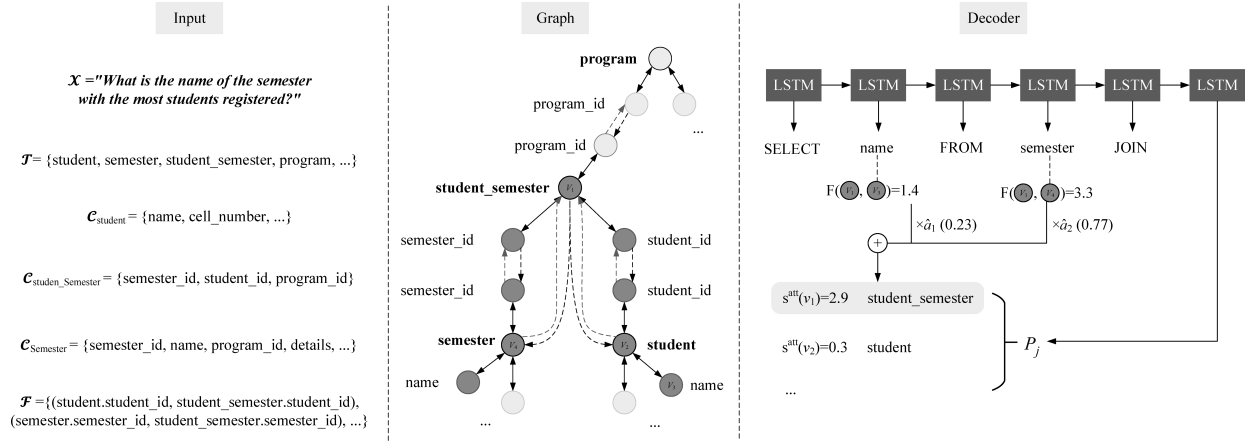


图 4 图神经网络模型^[32]

3.2.3 基于中间表示的多表单轮式表格问答

由于自然语言与 SQL 语言存在不匹配问题,一些研究通过引入中间表示来连接自然语言与 SQL。Guo 等人^[36]提出的 IRNet 将 SQL 合成过程分为三个阶段,通过引入中间表示实现问题到表格内容的映射。IRNet 首先在问题和模式之间进行链接,识别问题中提到的列和表,并根据问题中提到的列的方式为列分配不同的类型。然后,IRNet 设计了一种树形结构查询 SemQL 作为连接自然语言和 SQL 的中间表示。如图 5 所示,SemQL 由 Select 和 Filter 两部分组成。其中,Select 部分表示查询列,Filter 子树整合了 WHERE 和 HAVING 等筛选条件。最后,IRNet 从具有领域知识的 SemQL 查询中推断出 SQL 查询,有效提升了在 Spider 数据集上的问答准确率。在利用 BERT 对问题、模式和模

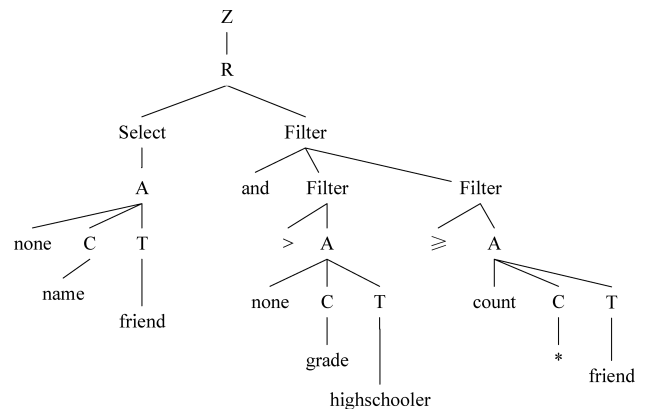


图 5 SemQL 说明示例^[36]

式链接结果进行编码后,借助预训练模型强大的特征表示能力,IRNet 在 Spider 数据集下的效果提升了 8%。基于中间表示的表格问答方法有效缓解了

自然语言与 SQL 之间的不匹配问题,但仍存在一些限制,如不支持 FROM 子句的自连接。

4 多表多轮式表格问答任务

4.1 数据集及评测标准

2019 年,表格问答从单轮扩展为多轮,并和对话系统进行结合。SPaRC^[40]是一个具有复杂上下文依赖的多表多轮表格问答数据集,由 4 298 个问题及其对应 SQL 语句组成,包含 138 个领域的复杂数据库交互。SequentialQA (SQA)^[41]是微软为探索对话式问答提出的弱监督语义解析数据集,包含 17 553 个问题-答案对。CoSQL^[42]是一个跨领域对话式数据集,涵盖了大多数类型的对话式数据库查询交互,包括模糊问题和无法回答的问题。在中文语境下,Guo 等人^[43]提出跨领域多轮交互 Text-to-SQL 中文数据集 CHASE,由 17 940 个问题及其 SQL 查询语句组成。CHASE 数据集问题序列如图 6 所示,每个问题都有相应的 SQL 语句,后续问题可能会依赖前文信息。

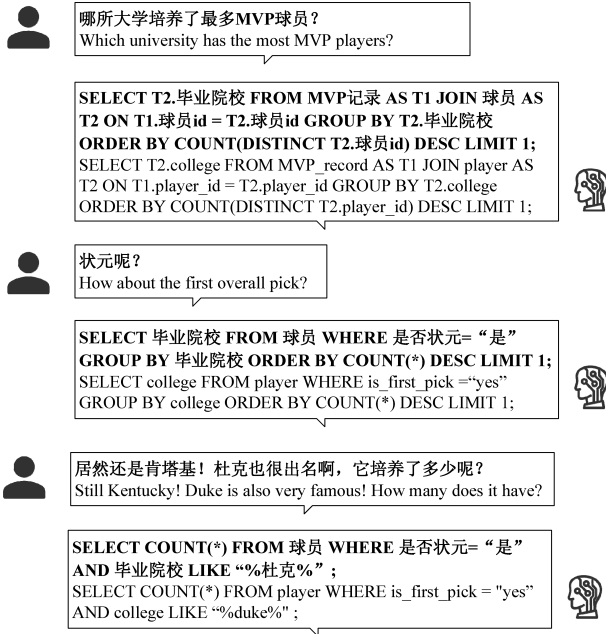


图 6 CHASE 数据集问题序列^[44]

如第二个问题“状元呢?”省略了前文隐含的询问主语“大学”,完整的问题应该为“哪所大学培养了最多状元?”,第三个问题省略了查询条件“状元”。HybridQA^[44]是第一个基于表格和文本数据的多跳异构混合问答数据集,需要对异构信息进行推理。多表多轮式表格问答数据集具体信息如表 5 所示。

表 5 多表多轮式表格问答数据集

数据集	问答对规模	表格规模	语言
SPaRC ^[40]	12 726	1 020	英文
SQA ^[41]	17 553	982	英文
CoSQL ^[42]	15 598	1 020	英文
CHASE ^[43]	17 940	1 280	中文
HybridQA ^[44]	69 611	13 000	英文

多表多轮表格问答方法的评测指标包括问题匹配率 (Question Match) 和交互匹配率 (Interaction Match)。问题匹配率计算所有问题的精确集合匹配分数,其计算如式(9)、式(10)所示。

$$\text{Score}_{\text{em}} = \begin{cases} 1 & Y' = Y \\ 0 & Y' \neq Y \end{cases} \quad (9)$$

$$\text{Acc}_{\text{qm}} = \frac{1}{N} \sum_{i=1}^N \text{Score}_{\text{em}}^i \quad (10)$$

其中, Y' 和 Y 分别代表预测 SQL 语句和真实 SQL 语句的组件集合,当所有预测 SQL 子句均正确时分数为 1。 Acc_{qm} 代表问题匹配率, N 表示问题总数。

交互匹配率计算轮次交互的精确集合匹配分数,当交互中的每个问题都符合精确集合匹配时分数为 1,如式(11)、式(12)所示。

$$\text{Score}_{\text{interaction}} = \begin{cases} 1 & \prod_{i=1}^t \text{Score}_{\text{em}}^i = 1 \\ 0 & \prod_{i=1}^t \text{Score}_{\text{em}}^i = 0 \end{cases} \quad (11)$$

$$\text{Acc}_{\text{im}} = \frac{1}{Q} \sum_{q=1}^Q \text{Score}_{\text{interaction}}^q \quad (12)$$

其中, t 表示每个交互中的轮次数量, Q 表示交互总量, Acc_{im} 代表交互匹配率。为了提升语义准确性,Zhong 等人^[45]提出了测试套件准确率,通过区分相近查询并手动检查其对预测查询的判断来确保测试套件的质量。

4.2 多表多轮式表格问答方法

多表多轮式表格问答涉及到历史轮次信息,其历史 SQL 语句包含了丰富的上下文信息。多表多轮式表格问答包括基于上下文的方法和基于数据增强的方法,各类模型在多表多轮数据集上的性能对比如表 6 所示。

表 6 多表多轮式方法性能及其优缺点对比

模型	SParC		CoSQL		SQA		优点	缺点
	问题 匹配率	交互 匹配率	问题 匹配率	交互 匹配率	问题 匹配率	序列 匹配率		
GNN ^[46]	—	—	—	—	55.1	28.1	模型架构简单	初始问题实体匹配效果不佳
RichContext ^[47]	52.6	29.9	41.0	14.0	—	—	方法全面有效	上下文表征有待提升
EditSQL ^[48]	47.9	25.3	40.8	13.7	—	—	复用历史 SQL	容易产生误差传播
HIE-SQL ^[49]	64.6	42.9	53.9	24.6	—	—	有效编码上下文信息	容易产生误差传播
CQR-SQL ^[50]	67.3	43.9	58.3	27.4	—	—	泛化能力强	容易产生误差传播
Eisenschlos ^[51]	—	—	—	—	71.0	44.8	速度快	精度下降
TableFormer ^[52]	—	—	—	—	72.3	47.5	鲁棒性强	训练时间长
SCORE ^[53]	62.4	38.1	51.6	21.2	65.4	38.5	泛化能力强	无法追踪远距离信息

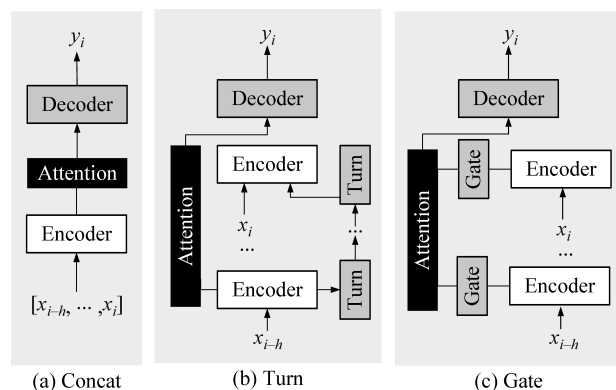
4.2.1 基于上下文的多表多轮式表格问答

相较于单轮式表格问答,多表多轮式表格问答方法需要解决上下文问题模式表示问题及历史 SQL 查询编码问题。如图 7 所示,Liu 等人^[47]研究了不同上下文信息编码方式对解析性能的影响。第一种方式是将每个轮次的问题进行串联拼接后编码,第二种方式是依次对每个轮次的问题进行编码,并通过轮次注意力关联信息,第三种方式是分别对每个轮次的问题进行编码,并通过门控机制平衡每个轮次问题的重要性。实验结果表明,拼接和轮次的方法优于门控机制。考虑到多轮问答中相邻的自然语言问题通常具有相关性,并且对应的 SQL 查询存在部分重叠,Zhang 等人^[48]通过编辑历史 SQL 查询来利用交互信息提高 SQL 的生成质量。该模型通过文本和表格编码器分别编码每个回合的信息,并提出轮次注意力机制捕获交互历史。解码器根据问题上下文、表格模式以及历史 SQL 查询进行解码输出。Müller 等人^[46]将表格和问题编码为图

神经网络,通过指针网络从编码图中选择答案,在 SQA 上证明了其有效性。

4.2.2 基于数据增强的多表多轮式表格问答

对于以问题-答案对训练的弱监督数据集,执行伪 SQL 同样能得到正确答案,即欺骗查询问题。由于伪 SQL 不能准确表示问题含义,以伪 SQL 为正例进行训练会极大地影响语义解析的性能。为了缓解欺骗查询问题,Misra 等人^[54]使用策略重塑方法引入轻量级的领域知识,将搜索过程向与文本更兼容的语义解析引导。CQR-SQL^[50]进一步考虑模式与对话之间的依赖性,通过辅助对话问题重构来显式利用模式并解耦 SQL 的上下文依赖。CQR-SQL 首先提出了一种模式增强的递归对话问题重构方法来生成领域相关的自包含问题。其次,通过模式一致性和树状 SQL 解析一致性两个任务训练模型,将多轮问题和自包含问题的语义映射到同一空间,通过充分的上下文理解增强 SQL 解析能力。基于数据增强的表格问答方法能有效传递多轮问答中的上下文信息,提升语义解析性能。

图 7 多表多轮问题编码策略^[47]

5 表格预训练模型

Transformer^[55]的提出使得自然语言处理研究进入预训练语言模型时代。随后 BERT^[56]、GPT^[57]等预训练模型大幅提升了各类自然语言处理任务的性能,表格问答领域也逐渐开始探索基于预训练模型的方法。本文从数据构造、输入编码、预训练目标三个方面介绍表格预训练模型。

5.1 数据构造

预训练模型的效果很大程度上依赖于微调时高质量的训练数据,而问题-SQL 对的构建将耗费大量人力资源,一些预训练模型类方法致力于研究微调训练数据的构造。例如, GraPPa^[21] 首先从标注 Text-to-SQL 语料中抽象模式组件创建 SQL 语法模板,然后使用同步上下文无关文法通过替换查询列、操作符、条件值等数据合成大量高质量问题-SQL 对。TAPEX^[23] 通过采样表格中的列名和单元格值填充 SQL 模板以获得大规模语料库。训练过程中的数据均由模型自动合成,无需额外的人工标注,大幅降低了训练成本。Score^[53] 通过 SParC^[40] 中的 500 个样本推导出由同步问题-SQL 模板和上下文问题模板组成的问题-SQL 生成语法,并合成约 435 000 个多轮 Text-to-SQL 数据集。

5.2 输入编码

在表格问答任务中,输入包括自然语言问题和表格模式两部分,二者具有不同的结构和格式。自然语言问题为文本序列,表格模式通常分布于二维

矩阵中。编码表格时,需要将二维表格数据转化为一维线性化序列,常用的方法是将表格按行展开。TaBERT^[20] 通过联合自然语言文本和结构化表格进行编码,获取自然语言问题与表格的语义表示。如图 8 所示, TaBERT 首先根据问题筛选出与之相关的 K 行作为内容快照,然后将问题与内容快照中的每一行通过串联进行拼接,作为 Transformer 模型的输入。在对所选行进行线性化的过程中,表格信息包括列名、列类型以及单元格值。考虑到经过 BERT 输出的行表示是独立的,而实际上表格各行之间相互关联,为了允许信息在不同行的单元格中流动, TaBERT 提出了垂直注意力机制,在不同行的对齐向量间计算,以此聚合跨行信息。TaPas^[24] 将问题和序列化后的表格串联作为输入,两个分类层分别预测所选的单元格子集和聚合操作符,在该单元格子集上执行相应的聚合操作得到最终答案。TaPas 架构简单,能够处理包括聚合操作在内的复杂问答。然而,由于 TaPas 将整张表放入内存作为上下文处理,模型无法处理大型表格或包含多张表格的数据库。为突破内存限制,可以进行表格压缩或信息过滤,以便只对相关内容进行编码,提升模型效率。

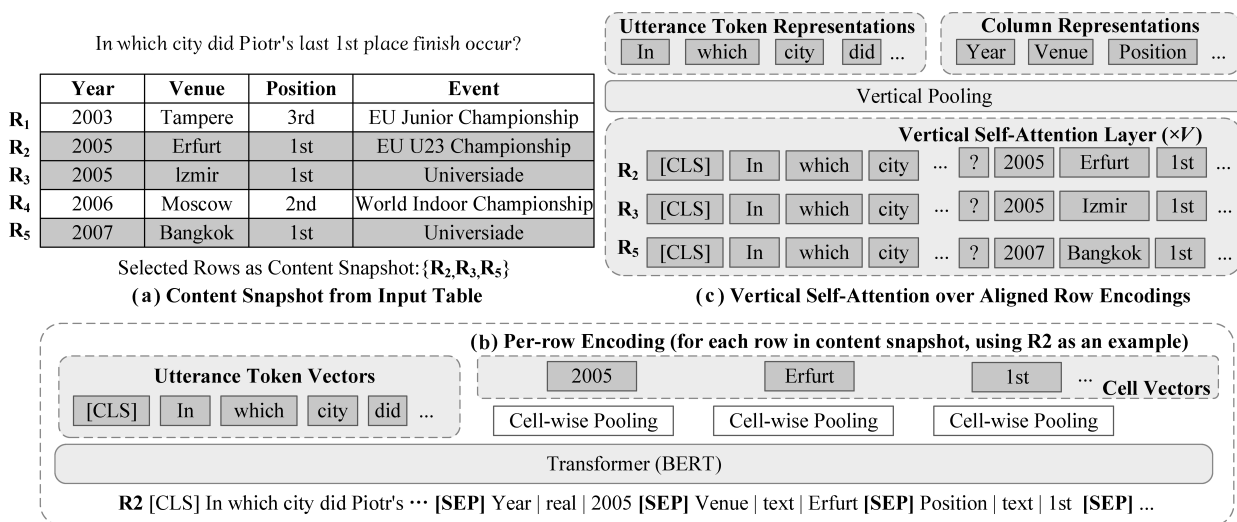


图 8 TaBERT 模型图^[20]

5.3 预训练目标

5.3.1 单表单轮式表格问答预训练目标

单表单轮表格问答任务较为简单,预训练目标主要针对表格结构和内容理解,多采用 Masked Language Model (MLM) 的不同变体。TaBERT^[20] 在自然语言问句上沿用了 BERT 的 MLM 策略,并针对表格模式提出了 Masked Column Prediction

(MCP) 和 Cell Value Recovery (CVR) 训练目标。MCP 通过训练模型恢复被 Mask 的列名和数据类型学习表格的列表示, CVR 预测内容快照中的原始单元格值。TaBERT 虽然性能较好,但需要大量语料进行训练,高质量训练数据的缺失影响该方法在垂直领域的应用。TaPas^[24] 将 BERT 的 MLM 训练目标扩展到结构化数据,通过选择单元格子集及其聚合操作来预测答案。除了 MLM, GraPPa^[21] 提

出了 SQL Semantic Prediction(SSP)目标预测表格每列在 SQL 中的语法角色。TAPEX^[23]通过预训练来模拟 SQL 执行器,其思想是如果一个模型被训练成能够忠实地“执行”SQL 查询,那么其对表格结构必将有深刻的理解。TAPEX 以 SQL 查询在表格上执行的真实结果作为监督信号去训练模型生成的结果。Wang 等人^[58]提出了一种执行引导策略,在解码过程中通过部分生成的 SQL 来检测并排除错误程序,进一步提高了性能。该机制可用于任何自回归生成模型,辅助大多数语义解析预训练模型进一步提升性能。SeaD^[22]模型通过感知去噪目标来训练 Seq2Seq 模型,并优化子句敏感的执行引导解码策略,以克服生成模型执行引导解码的局限性。

5.3.2 多表单轮式表格问答预训练目标

多表单轮表格问答任务联合多表交互,其关键在于链接问题与表格以及表格间的连接方式,预训练目标主要针对模式链接。Shi 等人^[37]根据 Text-to-SQL 任务的难点提出列预测、列恢复、SQL 生成等预训练目标,并利用生成模型合成预训练数据来联合学习自然语言和表模式的表示。列预测任务预测模式中的每一列是否被问题提及,提高模型检测问题所提及列的能力。列恢复任务将列名随机替换为其单元格值并要求模型恢复原始列名,强化模型

根据问题中具体值推断查询列的能力。SQL 生成任务通过生成给定问题和模式的 SQL 查询提高模型合成复杂 SQL 的能力。StruG 模型^[59]基于文本表格平行语料库,提出三个基于结构的预训练目标,通过识别自然语言文本中对于列和列值的引用,同时将之对应到正确的模式上,有效捕捉文本表格的对齐关系。基于语义解析的预训练表格模型的最终目标是提升下游 Text-to-SQL 任务的效果。如图 9 所示,问题中的“ID”对应模式中的查询列“Student ID”,问题中的“height”和“tallest”对应模式中的查询列“Height”,问题中的“boy”对应模式中的条件列“Sex”。这种自然语言问题与数据库模式之间存在的显式交互关系是揭示不同成分之间语义和结构对应关系的关键因素。为了解决模式链接问题,SDCUP^[38]通过双仿射网络预测问题和表格模式之间的细粒度依赖边,并提出实体扰动恢复预训练目标去学习问题和表格相关列之间的关系。考虑到数据构造和自动标记依赖引起的实例级噪声会影响收敛稳定性和预训练性能,SDCUP 引入了模式感知的课程训练策略,通过评估数据难度和模型接受能力渐进式地输入训练数据。基于预训练模型的问答方法针对多表单轮式数据集的特点设计模式依赖等预训练目标,极大地提升了在 Spider 数据集下的问答性能。

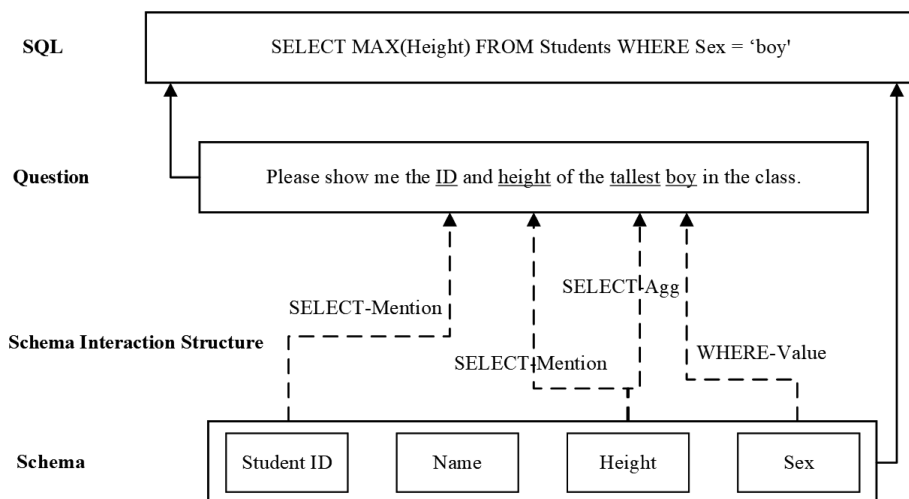


图 9 模式链接示例^[38]

5.3.3 多表多轮式表格问答预训练目标

多表多轮表格问答任务涉及上下文轮次信息,预训练目标包括历史信息传递以及鲁棒性提升。Eisenschlos 等人^[51]提出反事实陈述和句子合成两个预训练任务改进了 TaPas,在精度适度下降的情

况下,模型速度提高了两倍。反事实数据增强为每一个正例创建一个差异最小的反例,合成数据的目标是改进针对数值数据的运算和比较任务。TableFormer^[52]进一步改善了 TaPas 性能容易受到行或列顺序扰动影响的问题,缓解了模型倾向于选择表

尾数据作为答案的情况,提高了模型的鲁棒性。除此之外, TaPas 只能预测部分正确答案,而 TableFormer 可以在有多个答案的情况下预测全部答案。由于自然语言和 SQL 之间存在不匹配问题, HIE-SQL^[49] 将自然语言和 SQL 视为两种模态,并提出了一个历史信息增强的双模态预训练模型,捕捉自然语言与 SQL 之间的语义连接并减小二者之间的差距。SQLBERT^[45] 沿用了 MLM 预训练目标对 SQL 查询进行 Mask 操作,使模型更好地学习 SQL 结构。Score^[53] 提出了 Column Contextual Semantics(CCS)和 Turn Contextual Switch(TCS)两个预训目标,其中, CCS 学习自然语言问句与数据库模式的对齐关系, TCS 捕获多轮对话中上下文传递的信息。Score 适应于大多数多轮语义解析任务,但其追踪远距离轮次对话信息时的效果有待提升。

6 前景与挑战

总体而言,表格问答是一个较新的研究方向。但随着深度神经网络和预训练模型的发展,各类表格问答任务性能均得到大幅提升,为高效利用表格信息提供技术支撑。如何提升问答准确度、优化问答性能、提升模型泛化性是未来表格问答任务的挑战,主要体现在以下几个方面。

6.1 纯表格问答

随着预训练模型极大地提升了自然语言处理的技术水平,预训练方法被逐步应用于联合训练表格和文本表示。然而,许多实际问题仅涉及对表格数据的语义理解,而无需额外的文本输入,如检索表格中相似的列、填充缺失单元格信息等任务。现有的预训练模型联合文本、表格进行编码,在处理没有任何相关文本的表格时表现不佳。Iida 等人^[60]设计了一个损坏单元格检测的预训练目标,判断每个单词是原始文本的一部分还是已被破坏,以此来避免错误的信息提取。TABBIE 只从表格数据中学习,大幅降低了计算成本,并在一系列基于表格的预测任务中达到最先进的水平。针对不包括相关文本信息的纯表格数据,如何提升模型性能并降低计算成本,值得未来进一步开展相关研究工作。

6.2 跨领域表格问答

随着表格问答服务的日益普及,实际应用中来

自不同领域的客户也逐步增加,跨领域的表格问答成为了新的挑战。Chemmengath 等人^[61]研究发现目前基于 transformer 的表格问答模型在面對训练时没有见过的主题时,模型的性能下降 5%~6%。为了解决这个问题,Chemmengath 等人进一步提出了 T3QA 框架,以提高模型跨领域的鲁棒性。T3QA 首先将特定领域的主题词汇注入 BERT,然后通过 Text-to-Text 文本生成器生成特定领域的训练数据,最后通过逻辑形式重排器产生更精准的预测。跨领域表格问答致力于提升模型通用性,具有实际工程应用价值。

6.3 答案生成问题

现有表格问答模型得到的结果仍为表格信息,而表格中的内容一般比较简短,经常是以实体词的方式呈现,可读性较差。为了提升用户体验,表格问答系统还需进一步结合上下文进行自然语言答案生成。近年来预训练生成模型如 GPT-3^[62]、T5^[63]等在文本生成相关任务中取得了显著性能提升,在 Table-to-Text 方向,KGPT^[64]和 TableGPT^[65]等模型也进一步关注基于结构化表格数据生成对应的文字描述。在答案生成任务实际落地的过程中,由于生成模型的解码速度一般相对较慢,推理效率需要大幅度优化。除此之外,由于垂直领域的特殊表达需要有一定的泛化能力,因此对生成模型的效果提出了更高的要求。

6.4 高效表格问答

高效的性能在系统实际运用中至关重要,现有的模型通过参数共享、池化、稀疏等方式提升模型性能。MATE^[66]采用稀疏注意力机制允许头部有效地关注表中的行或列,让模型的计算速度和内存消耗随序列长度的增加呈线性增长。Transformer-XL^[67]通过段级递归和位置编码解决学习依赖超越固定长度的问题。ETC^[68]引入带有相对位置编码的全局-局部注意机制扩展输入长度,编码结构化输入。但计算开销较大,且局部稀疏注意力的 Mask 机制限制了其处理所有单词的能力。对于模型的内存消耗、运行速度、可解决问题类型等性能方面的提升是未来的一大研究方向。

6.5 面向不同垂直领域实际场景的表格问答

现有的 Text-to-SQL 模型在迁移到实际应用表格问答任务时,由于不同垂直领域存在问题数和数

数据集领域特性强、数据规范性和格式统一性亟待加强、标注语料可解释性有待提升等不足,会出现模型精确匹配率降低的问题^[69]。实际应用中,还存在用户问句中关键信息与数据库表名、列名信息表示不一致,以及问句口语化表达、领域性简称导致的隐含语义等问题。另外,由于中文语境下问句表达更加灵活多变,语义信息丰富,导致对问句理解更具挑战性。如何提升模型在面向不同垂直领域实际场景应用时的准确性,实现表格问答进一步赋能行业应用是未来一大挑战。

7 结论

随着智能时代的来临,如何快速从海量数据中获取目标信息成为了不同行业的智能问答需求。表格作为一种高价值的信息载体,是问答系统的主要数据源之一。本文从任务角度系统介绍了各类表格问答的模型方法、算法特点及其优势与不足,并归纳总结表格预训练模型的数据构造、输入编码以及各项任务下的预训练目标。最后,针对当前表格问答的局限性,本文提出了未来表格问答的前景与挑战。

参考文献

- [1] 高良才,李一博,都林,等.表格识别技术研究进展[J].中国图象图形学报,2022,27(06): 1898-1917.
- [2] 马浩,戴国琳,刘新遥,等.医学知识图谱自动构建研究[J].医学信息,2022,35(04): 10-12.
- [3] YANG J, XIANG F, LI R, et al. Intelligent bridge management via big data knowledge engineering[J]. Automation in Construction, 2022, 135: 104118.
- [4] 梁清源,朱琪豪,孙泽宇,等.基于深度学习的 SQL 生成研究综述[J].中国科学(信息科学),2022, 52(08): 1363-1392.
- [5] 郑泳智,朱定局,吴惠彝,等.知识图谱问答领域综述[J].计算机系统应用,2022,31(04): 1-13.
- [6] THERASA M, MATHIVANAN G. Survey of machine reading comprehension models and its evaluation metrics[C]//Proceedings of the 6th International Conference on Computing Methodologies and Communication, 2022: 1006-1013.
- [7] ZHONG V, XIONG C, SOCHER R. Seq2sql: Generating structured queries from natural language using reinforcement learning[J]. arXiv preprint arXiv:1709.00103, 2017.
- [8] 李智,王震,杨赋庚,等.基于表格的自动问答研究与展望[J].计算机工程与应用,2021,57(13): 67-76.
- [9] ZHANG S, BALOG K. Web table extraction, retrieval and augmentation: A survey[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(2): 1-35.
- [10] DONG H, CHENG Z, HE X, et al. Table pretraining: A survey on model architectures, pretraining objectives, and downstream tasks[J]. arXiv preprint arXiv:2201.09745, 2022.
- [11] QIN B, HUI B, WANG L, et al. A survey on text-to-sql parsing: Concepts, methods, and future directions[J]. arXiv preprint arXiv:2208.13629, 2022.
- [12] PASUPAT P, LIANG P. Compositional semantic parsing on semi-structured tables[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1470-1480.
- [13] PRICE P. Evaluation of spoken language systems: The atis domain[C]//Proceedings of the 3rd Darpa Speech & Natural Language Workshop, 1990: 24-27.
- [14] SUN N, YANG X, LIU Y. TableQA: A large-scale Chinese text-to-sql dataset for table-aware sql generation[J]. arXiv preprint arXiv:2006.06434, 2020.
- [15] CHEN Z, CHEN W, SMILEY C, et al. FinQA: A dataset of numerical reasoning over financial data [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 3697-3711.
- [16] ZHU F, LEI W, HUANG Y, et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 3277-3287.
- [17] XU X, LIU C, SONG D. Sqlnet: Generating structured queries from natural language without reinforcement learning[J]. arXiv preprint arXiv:1711.04436, 2017.
- [18] DONG L, LAPATA M. Coarse-to-fine decoding for neural semantic parsing[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2018: 731-742.
- [19] YU T, LI Z, ZHANG Z, et al. TypeSQL: Knowledge-based type-aware neural text-to-sql generation [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2018: 588-594.
- [20] YIN P, NEUBIG G, YIP W, et al. TaBERT: Pre-

- training for joint understanding of textual and tabular data[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 8413-8426.
- [21] YU T, WU C, LIN X, et al. GraPPa: Grammar-augmented pre-training for table semantic parsing [C]//Proceedings of the International Conference on Learning Representations, 2020.
- [22] XUAN K, WANG Y, WANG Y, et al. SeaD: End-to-end text-to-sql generation with schema-aware denoising[J]. arXiv preprint arXiv:2105.07911, 2021.
- [23] LIU Q, CHEN B, GUO J, et al. TAPEX: Table pre-training via learning a neural sql executor[C]//Proceedings of the International Conference on Learning Representations, 2021: 1-16.
- [24] HERZIG J, NOWAK P, MUELLER T, et al. TaPas: Weakly supervised table parsing via pre-training[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 4320-4333.
- [25] WOODS W. The lunar sciences natural language information system[R]. US: BBN Report, 1972: 1-388.
- [26] DAS D, CHEN D, MARTINS A, et al. Frame-semantic parsing[J]. Computational Linguistics, 2014, 40(1): 9-56.
- [27] POPESCU A, ETZIONI O, KAUTZ H. Towards a theory of natural language interfaces to databases [C]//Proceedings of the 8th International Conference on Intelligent User Interfaces, 2003.
- [28] CHEN J, JIMÉNEZ RUIZ E, HORROCKS I, et al. Colnet: Embedding the semantics of web tables for column type prediction [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 29-36.
- [29] GOL M, PUJARA J, SZEKELY P. Tabular cell classification using pre-trained cell embeddings[C]//Proceedings of the IEEE International Conference on Data Mining. IEEE, 2019: 230-239.
- [30] YU T, ZHANG R, YANG K, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020: 3911-3921.
- [31] YU T, YASUNAGA M, YANG K, et al. SyntaxSQLNet: Syntax tree networks for complex and cross-domain text-to-sql task[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 1653-1663.
- [32] BOGIN B, BERANT J, GARDNER M. Representing schema structure with graph neural networks for text-to-sql parsing[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4560-4565.
- [33] BOGIN B, GARDNER M, BERANT J. Global reasoning over database structures for text-to-sql parsing [J]. arXiv preprint arXiv:1908.11214, 2019.
- [34] 曹合心, 赵亮, 李雪峰. 图神经网络在 Text-to-SQL 解析中的技术研究[J]. 计算机科学, 2022, 49(4): 110-115.
- [35] WANG B, SHIN R, LIU X, et al. RAT-SQL: Relation-aware schema encoding and linking for text-to-sql parsers [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7567-7578.
- [36] GUO J, ZHAN Z, GAO Y, et al. Towards complex text-to-sql in cross-domain database with intermediate representation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4524-4535.
- [37] SHI P, NG P, WANG Z, et al. Learning contextual representations for semantic parsing with generation-augmented pre-training [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13806-13814.
- [38] QIN B, WANG L, HUI B, et al. SDCUP: Schema dependency-enhanced curriculum pre-training for table semantic parsing[J]. arXiv preprint arXiv:2111.09486, 2021.
- [39] HUO S, MA T, CHEN J, et al. Graph enhanced cross-domain text-to-sql generation[C]//Proceedings of the 13th Workshop on Graph-Based Methods for Natural Language Processing, 2019: 159-163.
- [40] YU T, ZHANG R, YASUNAGA M, et al. SParC: Cross-domain semantic parsing in context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4511-4523.
- [41] IYYER M, YIH W, CHANG M. Search-based neural structured learning for sequential question answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1821-1831.
- [42] YU T, ZHANG R, ER H, et al. CoSQL: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 1962-1979.
- [43] GUO J, SI Z, WANG Y, et al. CHASE: A large-scale and pragmatic Chinese dataset for cross-data-

- base context-dependent text-to-sql[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 2316-2331.
- [44] CHEN W, ZHA H, CHEN Z, et al. HybridQA: A dataset of multi-hop question answering over tabular and textual data[G]//Proceedings of the Association for Computational Linguistics, 2020: 1026-1036.
- [45] ZHONG R, YU T, DAN K. Semantic evaluation for text-to-sql with distilled test suites[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 396-411.
- [46] MÜLLER T, PICCINNO F, NICOSIA M, et al. Answering conversational questions on structured data without logical forms[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5902-5910.
- [47] LIU Q, CHEN B, GUO J, et al. How far are we from effective context modeling?: An exploratory study on semantic parsing in context[C]//Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence, 2021.
- [48] ZHANG R, YU T, ER H, et al. Editing-based sql query generation for cross-domain context-dependent questions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 2020: 5338-5349.
- [49] ZHENG Y, WANG H, DONG B, et al. HIE-SQL: History information enhanced network for context-dependent text-to-sql semantic parsing[G]//Proceedings of the Association for Computational Linguistics, 2022: 2997-3007.
- [50] XIAO D, CHAI L, ZHANG Q, et al. CQR-SQL: Conversational question reformulation enhanced context-dependent text-to-sql parsers[J]. arXiv preprint arXiv: 2205.07686, 2022.
- [51] EISENSCHLOS J, KRICHENE S, MUELLER T. Understanding tables with intermediate pre-training[G]//Proceedings of the Association for Computational Linguistics, 2020: 281-296.
- [52] YANG J, GUPTA A, UPADHYAY S, et al. TableFormer: Robust transformer modeling for table-text encoding[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 528-537.
- [53] YU T, ZHANG R, POLOZOV A, et al. Score: Pre-training for context representation in conversational semantic parsing[C]//Proceedings of the International Conference on Learning Representations, 2020: 1-16.
- [54] MISRA D, CHANG M, HE X, et al. Policy shaping and generalized update equations for semantic parsing from denotations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 2442-2452.
- [55] 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, 2017.
- [56] DEVLIN J, CHANG M, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [57] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. <https://cdn.openai.com/research-covers/Language-unsupervised/Language-understandly-paper.pdf>[2022-04-15]
- [58] WANG C, TATWAWADI K, BROCKSCHMIDT M, et al. Robust text-to-sql generation with execution-guided decoding[J]. arXiv preprint arXiv:1807.03100, 2018.
- [59] DENG X, HASSAN A, MEEK C, et al. Structure-Grounded Pretraining for Text-to-SQL[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 1337-1350.
- [60] IIDA H, THAI D, MANJUNATHA V, et al. TABBIE: Pretrained representations of tabular data[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 3446-3456.
- [61] CHEMMENGATH S, KUMAR V, BHARADWAJ S, et al. Topic transferable table question answering [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 4159-4172.
- [62] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 33: 1877-1901.
- [63] 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: 1-67.
- [64] CHEN W, SU Y, YAN X, et al. KGPT: Knowl-

- eDGE-grounded pre-training for data-to-text generation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 8635-8648.
- [65] GONG H, SUN Y, FENG X, et al. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 1978-1988.
- [66] EISENSCHLOS J, GOR M, MUELLER T, et al. MATE: Multi-view attention for table transformer efficiency [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 7606-7619.
- [67] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2978-2988.
- [68] AINSLIE J, ONTANON S, ALBERTI C, et al. ETC: Encoding long and structured inputs in transformers[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 268-284.
- [69] 吕剑清,王先兵,陈刚,等.面向工业生产的中文 Text-to-SQL 模型[J].计算机应用,2022;1-8.
- [70] MIN Q, SHI Y, ZHANG Y. A pilot study for Chinese sql semantic parsing [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 3652-3658.
- [71] WANG L, ZHANG A, WU K, et al. DuSQL: A large-scale and pragmatic Chinese text-to-sql dataset [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020.



张洪虞(1998—),硕士研究生,主要研究领域为自然语言处理、表格问答。

E-mail: hongyizhang@mails.cqjtu.edu.cn



杨建喜(1977—),博士,教授,主要研究领域为桥梁智能管养与数据分析。

E-mail: yjx@cqjtu.edu.cn



李韧(1985—),通信作者,博士,副教授,主要研究领域为自然语言处理、知识图谱技术。

E-mail: renli@cqjtu.edu.cn