arXiv:2301.01067v1 [cs.CL] 3 Jan 2023

借助公式化知识实现知识密集型文本到 SQL 语义解析

Longxu Dou[1]***Yan Gao***[2], **Xuqi Liu**[1],**Mingyang Pan**[1],**Dingzirui Wang**[1], **Wanxiang Che**[1], **Min-Yen Kan**[3],**DechenZhan**[1],**Jian-Guang Lou**[2] 哈尔滨工业大学 [1] 微软亚洲研究院[2] 新加坡国立大学 [3] {lxdou, xqliu, mypan, dzrwang, car}@ir.hit.edu.cn, dechen@hit.edu.cn, {yan.gao, jlou}@microsoft.com, kanmy@comp.nus.edu.sg

抽象

本文研究知识密集型文本到 ***SQL***问题：解析专家提出的问题并将其转换为针对领域特定表的 SQL 查询时，需要用到领域知识。我们通过构建新的中文基准测试 KNOWSQL，将这一场景形式化。该基准测试包含涉及多个领域的领域特定问题。随后，我们提出公式化知识来解决这一问题，而不是标注更多数据示例。具体而言，我们构建了一个公式化知识库，作为领域知识库，并提出了一个框架（REGROUP），在解析过程中利用这些公式化知识。使用 REGROUP 进行的实验表明，该方法在 KNOWSQL 上的整体表现提升了 28.2%。

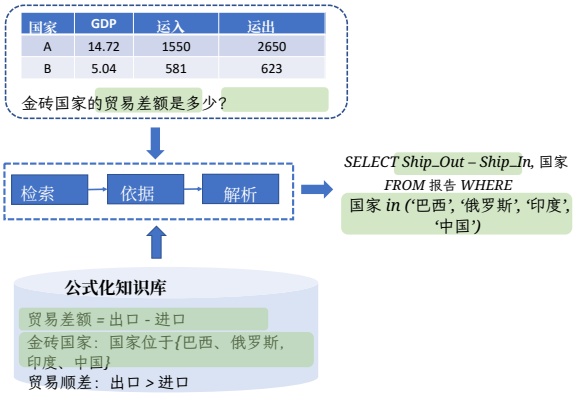

图 1：通过三个步骤利用 REGROUP 和公式化知识，实现知识密集型文本到 SQL：（1）**Re**检索公式化知识；（2）**Grou**将公式化知识中的概念落地；（3）**P**解析问题。

1 引言

文本到 SQL 能将用户查询转换为可执行的 SQL，极大地方便了用户与关系数据库之间的交互。随着大规模基准测试的发布（Zhong 等人，2017年；Yu 等人，2018年，2019a，b）以及模型设计的不断发展（Wang 等人，2020a；Cao 等人，2021年），文本到 SQL 研究如今在学术研究和实际应用中都取得了令人瞩目的成果（Zeng 等人，2020年）。

然而，在文本到 SQL 的专业应用中，例如财务报告数据分析，模型需要借助外部知识，将专家提出的问题映射到领域专用数据库。以一个财务问题为例：沃尔玛的息税前利润是多少[1]？底层数据库中包含可用于计算息税前利润的组成项列。我们将这一问题视为知识密集型文本到 ***SQL***，在这种情况下，解析专家针对领域专用表提出的问题时，领域知识至关重要。这一问题使文本到 SQL 技术无法应用于新兴的专业领域，因而无法协助专家处理数据。

传统方法通常会通过在目标领域标注特定的问题/SQL 对来解决这一问题（Wang 等人，2015年；Herzig 和 Berant，2019年）。随后，这些映射会在训练过程中被诱导出来。这种方法确实有效，但其缺点在于，任何诱导出的信息都既脆弱又高度依赖专业知识：此类知识无法跨领域迁移，并且需要专家知识来构建。

我们提出通过建模非专家如何解决这一问题来应对它。遇到未见过的示例时（如上文的息税前利润案例），他们可能会先从公共资源中查找相关的数学公式，再将公式中提及的概念与特定数据库中提供的模式元素对应起来。这个过程利用了公共资源中已有描述的常见、编码化的公式化知识，例如教程、

*在微软亚洲研究院实习期间的贡献。[1]EBIT 是息税前利润，即 Earnings Before Interest and Tax，计算公式为 收入 – 销售成本 – 营业费用。

t教科书、百科全书和参考资料 s.

受此启发，我们提出通过公式化知识来解决知识密集型文本到 SQL 问题。公式化知识为问题中呈现的领域特定短语如何映射到模式元素上的实际 SQL 操作提供依据。具体而言，我们定义了三类公式化知识的分类体系：计算、并集和条件，每类都对应一段特定的 SQL 代码。随后，我们提出 RE-GROUP，这是一个文本到 SQL 框架（图1），包含三个阶段：(1) 从公式化知识库中检索公式化知识，将其作为外部知识源；(2) 将公式化知识的概念依据到模式元素上；(3) 结合问题、模式和基于事实的公式化知识解析结果。外部公式化知识库为 REGROUP 注入公式化知识，使其具备知识。REGROUP 也具有可扩展性，因为更新公式化知识库不需要重新训练任何模块。

此外，我们构建了中文基准测试 KNOWSQL，以检验 REGROUP 框架的有效性。通过纳入更多 SQL 操作和更具挑战性的领域，它推动了现有知识密集型文本到 SQL 研究的发展，超越了以往工作（Wang 等人，2020b；Zhao 等人，2022年）。实验结果表明，结合公式化知识的 REGROUP 总体性能可提升 23.4%。此外，我们将错误案例分为三类，通过改进 REGROUP 中相应的模块即可解决。最后，我们讨论了未来的潜在研究方向，例如拓展知识范围以及改进 REGROUP 模型设计。

我们的贡献总结如下：

- 据我们所知，我们是首个探索知识密集型文本到 SQL 的团队，并提出了一个具有挑战性的中文基准测试 KNOWSQL，该基准测试需要领域知识。
- 我们提出了一种新颖的框架 REGROUP，通过检索并对齐公式化知识来解决知识密集型文本到 SQL 问题，并且该框架支持知识扩展。
- 实验结果表明，结合公式化知识的 REGROUP 有效，在 KNOWSQL 上实现了 28.2% 的整体提升。

2 知识密集型文本到 SQL

2.1 问题分析

在研究专业数据分析中的真实案例后，我们大致将知识密集型文本到 SQL 所需的知识分为三类：（1）语言学知识 使模型能够适应语言多样性；（2）领域知识 使模型能够理解特定领域的表达和概念；（3）数学知识 能够推导出特定的 SQL 运算（例如，密度 短语对应除法 运算）。这三类知识共同为将问题中的领域特定短语映射为模式元素上的实际 *SQL* 运算提供依据。

然而，大多数文本到 SQL 研究聚焦于通用场景（Yu 等人，2018年；Zhong 等人，2017年），这类场景主要需要语言学知识。近年来，Wang 等人（2020b）和Zhao 等人（2022年）通过引入计算类问题，将文本到 SQL 推向了更具挑战性的场景。本文进一步探索知识密集型文本到 SQL，纳入更多运算（例如，计算、并集和条件），并涉及更具挑战性的领域，这些领域需要上述三类知识。

2.2 挑战

尽管预训练语言模型包含语言学知识，但它们缺乏领域知识和数学知识。因此，模型会面临两个问题：(1) 不知道该使用哪些运算：如果某种运算（例如，密度 =总数量 / 空间）从未出现在训练数据中，模型在推理时很少会使用这种未见过的运算；(2) 不知道如何适配运算：模型无法将运算推广到不同领域。例如，模型无法将人口密度（人数 / 土地面积）的计算方法推广到汽车密度（汽车数量 / 停车场面积）。

因此，我们认为 Vanilla 预训练语言模型（1）覆盖面狭窄，因为它仅支持有限的运算；（2）效率低下，因为它无法将运算泛化到不同领域。然而，直接增加标注数据示例的数量既耗时又需要专业知识。相比之下，我们从第3节公式化知识的角度应对这一挑战，这种方法更易于扩展知识。

2.3 KNOWSQL 基准

	#DB	#问题	#公式化
训练集	160	23,157	328
Dev	40	2,731	122
金融	217	1,392	219
房地产	35	749	79
交通运输	36	439	82

表 1：KNOWSQL 数据集统计。

为揭示知识密集型文本到 SQL 问题并推动相关研究，我们构建了一个具有挑战性的中文文本到 SQL 基准，名为 KnowSQL。总体而言，它由两部分组成：基于现有 DuSQL（Wang 等人，2020b）数据集构建的训练集/开发集，以及一个新构建的测试集，其中涵盖了 DuSQL 中发现知识的三个专业领域。

2.3.1 构建 DuSQL 训练集/开发集

我们基于现有的 DuSQL 构建 KNOWSQL 的训练集/开发集。DuSQL 是一个中文多表文本到 SQL 基准数据集。我们将其 200 个数据库划分为 16 个领域，如体育、能源、医疗保健、食品等。鉴于 DuSQL 模式质量高、领域覆盖广，它是构建具有挑战性的知识密集型文本到 SQL 基准数据集的良好起点。然而，DuSQL 对领域特定问题的涵盖并不充分，其中大多数问题无需依赖外部知识即可轻松回答，且只涉及一种 SQL 运算，即计算。因此，我们通过增加更多领域特定问题，并在训练集和开发集中纳入更多运算，对原始 DuSQL 进行了扩展。最终，KNOWSQL 将 DuSQL 训练集的规模从 22,521 扩大到 23,157，将开发集的规模从 2,482 扩大到 2,731。

2.3.2 从零开始构建测试集

为模拟专业的数据分析场景，我们创建了一个具有挑战性的测试集，涵盖三个领域（金融、房地产和交通运输）。在现实生活中，这三个领域对数据分析的要求都很高。与训练集/开发集不同，我们从零开始构建测试集，具体包括：（1）收集领域专用表；（2）标注领域特定问题及其对应的 SQL 查询。

表格收集。 为收集表模式，我们从以下来源收集表格：(1) 公司公开的年度报告 (2) 行业报告 (3) 学术论文 (4) 政府发布的统计报告。为确保表格质量，我们进行了若干预处理步骤。首先，我们将矩阵表（见于年度报告）转换为关系表，使问题能够通过 SQL 回答。接下来，为确保表格数据质量，我们进行数据清洗（例如，筛除无关列以简化表结构，并规范表头以减少噪声）。最后，为避免数据隐私问题，我们进行数值匿名化（例如，移除直接标识符并匿名化地理相关数据）。

问题标注。 如果没有背景知识，标注人员很难提出领域特定问题 [2]。因此，我们首先通过以下方式对标注人员进行领域知识培训：（1）从领域专用开放资源中收集行业术语（即缩写、术语），这些资源被领域专家广泛采用（例如，金融领域的息税前利润），但外行并不熟悉；（2）利用上述材料，模仿领域专家使用行业术语提出问题。

之后，标注人员将依据以下标准标注问题和 SQL：（1）忠实于给定的表（即，不超出表列和表内容的范围）；（2）不能仅凭表中的单个元素直接作答，但可以通过对现有列进行运算得出答案；（3）仅限于一阶运算（即，不包括多跳问题，例如“毛利润是多少？”，因为表中只有“销售额”“平均价格”和“销售成本”，因此模型需要先计算“收入”）。

2.3.3 数据集质量和数据统计

为保证数据质量，我们进行了多轮检查。最终，标注人员间的一致性达到 94.7% [3]。每一轮中，我们都要求每位标注人员按照上述标准审查其他人的标注，并进一步完善不符合标准的标注。如表1所示，测试集包含 ***288***个数据库和***2,580*** 个问题。值得注意的是，测试集中的所有这些具有挑战性的数据示例，都可以由***380*** 条公式化知识涵盖，相关内容将在第3节讨论。

[2]参见第 8节，了解标注者报酬和个人资料。[3]标注者间一致性按重叠投票的比例计算，投票内容是判断问题是否正确且属于领域特定问题。

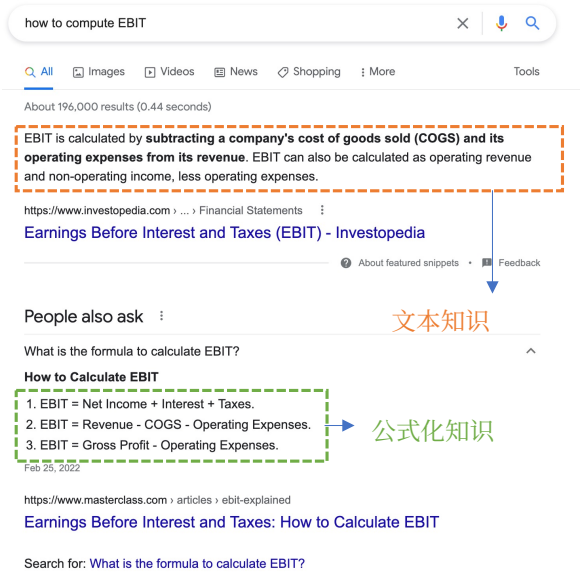

图 2：表达息税前利润的计算的两种知识类型：文本知识 和公式化知识。

3 方法：公式化知识

3.1 动机

遇到未见过的示例时，人们可能会先从教科书或百科全书中查找相关的数学知识或领域知识。如图2所示，关于息税前利润（EBIT）计算的信息以文本和公式两种形式呈现。直观而言，公式形式更可取，因为它：（1）更简洁、精确：例如，a 加 b 乘以 c 比 $a+b*c$ 或 $(a+b)*c$ 更容易产生歧义；（2）易于获取：教科书、教程和学术论文中的大多数计算说明都以公式形式存储；（3）便于 ***SQL*** 解析器处理：公式形式与 SQL 代码片段相近，因此解析器能够轻松生成[4]。

3.2 面向文本到 SQL 的公式化知识

遵循这一思路，我们聚焦于三类运算（图3）：计算、并集和条件。除了广为人知的计算知识，我们还将分类信息视为联合知识，并将判断标准视为条件知识。这里的设计思路是，左侧部分是知识项的名称，右侧部分则表达其语义含义，即通过对概念进行运算来表示。请注意，所有运算都与 SQL 语法一致，因此更接近 SQL 查询。除了实体，公式化知识的左侧部分也可能是 SQL函数（例如，NOW()）或常量（例如，房地产泡沫的阈值）。

3.3 公式化知识库

我们进一步构建了一个包含 1,954 个公式化知识项的公式化知识库，涵盖 KnowSQL 中涉及的 19 个领域。重要的是，该知识库涵盖了表2所示的 KNOWSQL 的所有这些示例。请注意，该知识库是一种与领域相关的资源，而非与特定数据库绑定。因此，该知识库具有更广泛的适用性，也可用于其他自然语言应用（例如问答）[5]。

标准公式化知识的设计遵循三项标准：(1) 仅考虑一阶（扁平）公式化知识（即公式化知识项中的概念应能与模式元素对齐，而非与另一个公式化知识项对齐）；(2) 存储的公式化知识应既忠实（即得到专家认可）又标准化（即在领域级共享）；(3) 公式化知识应为领域级（即不与特定模式元素绑定）。

收集我们从以下公开资源中收集公式化知识：(1) 百度文库，领域专家通常在该平台分享各个领域的领域知识[6]；(2) 中国知网，中国最大的学术网站[7]；(3) 特定领域的数据分析网站，例如体育领域的 ESPN 和金融领域的雅虎。我们还从英文资源中收集了一些知识，并请标注人员将这些领域知识翻译成中文。

抽象 为了让公式化知识更具通用性，我们提出在领域级而非特定数据库层面累积公式化知识。具体而言，我们在将公式化知识存入知识库之前，先对其概念进行抽象，这意味着运算针对的是概念，而非特定模式。例如，我们会将“2020年中国人口密度 2020 = 2020年中文人数总量 / 中国土地面积”中的公式化知识提取为“人口密度 = 人口总数 / 面积”。

[4]在第6节中，实验结果证明，公式化格式的表现优于文本格式。

[5]参见附录中银行的细粒度统计数据A.1 .

[6]https://wenku.baidu.com/ [7] https://oversea.cnki.net/index/

运算	计算	并集	条件
公式化知识	贸易差额 = 出口 – 进口	金砖国家： 国家属于{巴西、俄罗斯、印度、中国}	贸易顺差：出口 > 进口
抽象	短语 = 模式1 运算符 模式2	短语：模式 在 集合	短语：模式1 或 模式2
示例	中国的贸易差额是多少？ SELECT 出口 -进口 FROM 报告 WHERE 国家=中国	告诉我金砖国家国内生产总值的总和？ SELECT sum(国内生产总值) FROM 报告 WHERE 国家 in (巴西, 俄罗斯, 印度, 中国) GROUP BY 名称	哪个国家存在贸易顺差问题？ SELECT 国家 FROM 报告 WHERE 出口 > 进口

图 3：我们考虑三种公式化知识，以应对知识密集型文本到 SQL。

	#公式化	#计算	#并集	#条件
公式化知识库	1，954	1，102	346	506
涉及 KNOWSQL	891	656	52	183

表 2：公式化知识库的数据集统计及其与 KNOWSQL 的重叠部分。

因此，只需一种公式化知识，就能处理许多模式元素，以计算动物/汽车/商店密度。

K 中的映射NOWSQL 我们进一步考察公式化知识库与 KNOWSQL 基准之间的重叠情况。如第 2.3.3 节所述，KNOWSQL 中的所有问题都涵盖在公式化知识库中。具体而言，知识库中共有 1,954 个知识项，其中 891 个用于回答 KNOWSQL 中的问题，如表 2 所示。值得注意的是，除 KNOWSQL 涉及的内容外，知识库中还额外包含 1,063 个知识项，可支持未来应用公式化知识的相关工作。

4 REGROUP 框架

为解决知识密集型文本到 SQL 问题，我们提出了一种名为 REGROUP 的新框架，包含三个阶段：(1)**检索**公式化知识库中的公式化知识，将其作为外部知识源；(2)**将**公式化知识中的概念与模式元素对齐（例如，出口 对应 发货_运出）；(3) **解**析问题、模式和已对齐的公式化知识，得到结果。如图4所示，REGROUP 包含三个模型：检索器、落地模型和解析器。下面[8]将简要介绍每个模型。

[8]有关模型实现的更多细节，请参见附录 B。

4.1 检索器模型

检索器的目标是从公式化知识库中提取相关的公式化知识项（图4）。难点在于对公式化知识进行细粒度建模，以区分意图相同、但对概念执行的运算不同的知识项，例如以不同方式计算息税前利润。我们直接采用现成的稠密段落检索器（DPR）（Karpukhin 等人，2020年），该检索器最初是为开放域问答设计的。它采用双编码器架构来学习句子和段落的稠密表示，然后计算这些表示之间的点积，作为相似度分数。

为了使 DPR 适用于公式化知识检索任务，我们将公式化知识库视为段落候选项，并将问题与展平模式拼接（以特殊标记“|”分隔），以丰富问题的语义。随后，我们遵循标准的 DPR 训练流程来优化双编码器。具体而言，在训练过程中，我们从 KNOWSQL 标注中获取正向知识项，并从公式化知识库中采样五个负例。在推理过程中，我们首先缓存公式化知识项的嵌入，然后利用 FAISS 算法（Johnson 等人，2017年）对每个公式化知识项进行排序。

4.2 落地模型

给定检索到的知识项，落地模型的目标是针对特定模式编辑公式化知-

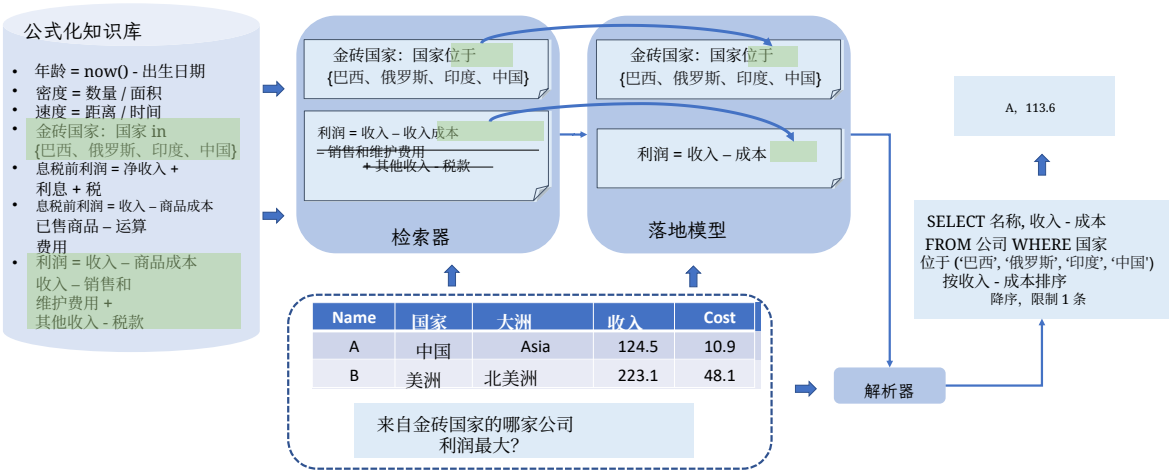

图 4：REGROUP 的流水线：(1) **检索**知识库中的公式化知识项；(2) **将公式化知识的概念对齐**到模式元素；(3) **解析**问题，并结合基于事实的公式化知识将其转换为 SQL。

识项，具体通过（1）移除无关概念，以及（2）使用模式元素实例化概念。主要挑战是对齐的标注（即监督）成本高昂。因此，弱监督的对齐方法会更为合适。具体而言，我们采用了（Liu 等人于2021年提出的 Erasing-then-Awakening（ETA）模型。该模型最初旨在将知识库中的实体与问题中提及的实体进行对齐。ETA 的输出是一个置信度矩阵，表示实体提及与实体之间可能存在的对齐关系。

为了将 ETA 适配到公式化知识对齐任务中，我们将每个知识项视为“问题”，并尝试确定该知识项与哪些特定模式元素相对齐。具体而言，我们通过一个超参数 H 来设定置信度阈值（判断其是否对齐，以及对齐到哪个元素）。如图4所示，我们会筛除置信度低于阈值的概念（划掉的部分），并H 用对齐后的元素（绿色部分）替换这些概念。

4.3 解析器模型

解析器的目标是根据问题和数据库模式预测可执行的 *SQL*。主要挑战在于如何对数据库结构建模，以推断问题中隐含提及的模式，以及如何利用已依据知识（即，知识融合）来发挥已依据知识的作用。我们受到近期进展的启发将大型预训练语言模型用于语义解析问题。例如，(Scholak 等人，2021年；Shin 等人，2021年；Dou 等人，2022年；Xie 等人，2022年)在简单的预训练语言模型框架下，于多项语义解析任务中取得了出色的表现，例如 BART（Lewis 等人，2020年）和 T5（Raffel 等人，2020年）。

鉴于此，我们提议在本研究中采用 UniSAr（Dou 等人，2022年）作为基础解析器。它通过三项非侵入式扩展改进了原版 BART，并在七个文本到 SQL 基准测试上取得了最先进或具有竞争力的性能。具体而言，模型的输入由问题、序列化模式和检索到的公式化知识串联而成。我们认为，解析器在 SQL 生成过程中应正确采用依据充分的公式化知识。

5 实验结果与分析

为评估我们提出的采用外部公式化知识库的 REGROUP 方法，我们在 KNOWSQL 基准上开展了多项实验。我们报告了流水线的整体结果，以及各模块的细粒度结果。我们还进行了错误分析，并将错误案例归纳为三大类。请注意，我们报告的是每种设置在三次运行中的实验结果平均值[9]。

[9]代码和数据可在此链接获取。

模型	Dev	金融	房地产	交通运输	平均值
Vanilla	69.3	8.7	5.7	6.9	22.7
REGROUP（无对齐）	71.7	38.1	25.1	32.7	41.9
REGROUP	74.6	43.7	46.1	39.1	50.9
REGROUP（预言机）	78.4	71.4	84.8	64.7	74.8

表3：不同 KNOWSQL 划分上的总体结果。预言机指使用预言机公式化知识。评估指标为 SQL精确集合匹配。平均值表示前四列的微平均分数。

Data	模型	R@1	R@3	R@10
Dev	BM-25	67.9	89.1	96.5
	REGROUP	73.0	89.8	96.5
金融	BM-25	39.4	66.5	85.9
	REGROUP	46.0	68.1	86.1

表 4：REGROUP 检索器在 KNOWSQL 开发集和金融划分上与 BM-25 的结果对比。评估指标为召回率。

Data	模型	精确率	召回率	F1
Dev	模糊匹配	69.3	62.5	65.7
	REGROUP	71.3	70.4	70.8
金融	模糊匹配	35.3	31.5	33.2
	REGROUP	42.9	44.7	43.8

表 5：REGROUP 落地模型与模糊字符串匹配在 KNOWSQL 开发集和金融划分上的结果对比。

5.1 实验设置

检索器从知识库中返回排名前三的公式化知识项。依据模型进一步将公式化知识中的概念与模式元素对齐，决策阈值 H 为0。6 ，该阈值由经验确定。解析器接收已依据知识、表模式和用户查询作为输入，然后输出 SQL。对于解析基线，我们采用 UniSAr（Dou 等人，2022年）作为 Vanilla 解析器[10]。

5.2 总体结果

如表3所示，我们可以观察到：（1）REGROUP 比原始模型高出 28.2%，这表明使用公式化知识是有效的；（2）将公式化知识与依据进行对齐后，REGROUP 的表现提升了 9.0% ；（3）预言机公式化知识（检索正确且依据对齐正确）使 REGROUP 达到上限 74.8%，这意味着 KNOWSQL 仍有提升空间。

[10]更多实现细节请参见附录B

5.3 细粒度结果

我们在 KNOWSQL 金融领域的开发集和测试集上，将检索器和落地模型与其他基线方法进行比较，以考察它们在通用场景和特定领域场景中的表现。

检索器 我们将 RE-GROUP 的检索器（双编码器）与 BM-25（Robertson 和 Zaragoza，2009年）进行比较。评估指标是检索结果的召回率。我们发现，金融领域比通用领域（开发集划分）更具挑战性，因为其中包含许多同质的公式化知识项：它们左侧表达相同意图，但右侧采用不同的计算方式。例如，图 4 中展示了两种计算“息税前利润”的方法。

对齐 我们将 REGROUP 的落地模型与基于模糊字符串匹配的方法进行比较 [11]。遵循先前的工作（Lei 等人，2020年；Liu 等人，2021年），我们报告微平均精确率、召回率和 F1分数。我们观察到：（1）基于模型的落地使性能分别提升了 5.1% 和 10.6%；（2）金融等领域专属数据比通用领域带来了更具挑战性的案例，金融领域的表现比开发集低约27.0%。

5.4 错误分析

我们分别从开发集划分中抽取300个案例，以及从测试集中的金融/房地产/交通运输领域抽取100个案例（共600个），用于错误分析。

Vanilla 模型错误 首先，我们比较原始模型错误预测而 REGROUP 正确预测的案例。正如图 5 第一部分中的示例所示，原始模型无法预测训练过程中未见过的运算。相比之下，基于事实的公式化知识使

[11]它枚举了公式化知识中概念的所有 n元语法（$n \leq 5$），并通过模糊字符串匹配将每个 n元语法与模式元素关联起来。



图5：REGROUP 的案例研究。我们首先将其与 Vanilla 解析模型进行比较。然后，我们将 REGROUP 的错误案例分为三类：（1）检索器错误：未能从知识库中获取知识；（2）依据错误：未通过对齐学习知识；（3）解析器错误：生成时未使用已依据知识。

REGROUP 能够正确预测作用于模式元素的运算。

然后，我们将 REGROUP 的错误分为三大类，并在图 5 中列出各类错误所占比例。最后，我们讨论改进 REGROUP 各个部分的潜在未来工作。REGROUP 的一个优势是，其解耦框架可以分别追踪每种错误案例，从而避免灾难性遗忘问题。

检索错误约 43% 的错误归因于检索器：模型未能从知识库中获取正确的知识，因为它无法区分封闭的公式化知识项之间的语义差异。未来的工作应通过细粒度建模（例如注意力机制）来提升其区分能力。Huang 等人（2019年）。

依据错误 约 41% 的错误是由不正确的已依据知识导致的，因为模型未能通过对齐学习知识，无法将概念正确对齐到模式元素。未来的工作应着重研究如何在弱监督下推导出对齐信息或在没有监督的情况下推导。这将大幅减轻特定领域繁重的标注工作。

解析错误仍有 8% 的错误案例由解析导致：公式化知识已被正确检索并依据，但解析器在生成过程中仍未能充分利用已依据知识。未来的工作应显式建模将知识从输入复制到 SQL 片段位置的过程来改进这一点，例如增加门控机制。

其他错误其余 8% 的错误与 SQL 生成有关，例如分组或嵌套 SQL。由于这并非我们的主要关注点，为简洁起见，我们在图5 中忽略了这些案例。

6 讨论

对于文本到 SQL，公式化知识是否优于文本知识? 在第3.1节中，我们直观地认为公式化知识比文本知识更可取。为验证这一点，我们通过以下步骤进行实验：(1) 由标注人员将公式化知识转换为文本知识

；(2) 在与公式化知识相同的实验设置下，使用文本知识训练检索器和解析器。实验结果表明，与表13.6% 相比，文本知识的整体性能有所下降。我们得出结论，REGROUP 更倾向于公式化知识，因为它更接近 SQL 片段或模式表示。此外，公式化知识既精确又简洁。相比之下，文本知识冗余较多，表达相同含义的方式也更加多样。

收集公式化知识的成本是多少？ 在公式化知识库（19 个领域）的收集过程中，我们发现大多数领域都有公共知识资源。此外，与标注数据示例相比，收集公式化知识所需的投入也是可以接受的。例如，我们花了***4*** 小时收集***219*** 条金融领域的公式化知识，其效率远高于标注等量的数据示例。最终，如表 3 所示，无需重新训练模型，公式化知识便能将性能提升 35.0%（从8.7% 提升至 43.7%）。

如何进一步拓展公式化知识的范围？ 本文主要关注领域知识和数学知识，并将其转换为公式化知识格式，以供模型学习。其他类型的知识也有助于进一步提升知识密集型文本到 SQL 的转换能力，例如常识知识（例如，水的冰点：温度=0）或个性化信息（例如，最喜欢的食物：提拉米苏）。因此，未来的工作可以将这些类型的知识封装成公式化格式。

7 相关工作

7.1 文本到 SQL 的领域泛化

为了适用于真实场景，文本到 SQL 模型应能泛化到新的领域，而不依赖昂贵的领域专用标注数据。先前研究表明，当前的文本到 SQL 模型通常无法应对领域泛化场景（Finegan-Dollak 等人，2018年）。近期，研究者通过数据合成（Yin 等人，2021年）、元学习（Wang 等人，2021年）和编码器预训练（Yin 等人，2020年；Herzig 等人，2020年）来解决这一问题。最近，Zhao 等人（2022年）提出采用模式扩展和模式剪枝来预处理表模式。

我们强调，与模式扩展方法相比，我们的方法（结合公式化知识的 REGROUP）的优势在于知识范围广泛：我们不仅考虑计算知识，还考虑并集知识和条件知识。此外，我们的方法可以通过外部且易于维护的公式化知识库进行扩展。

7.2 检索增强语义解析

近期，一种趋势是在各种自然语言处理任务中利用检索增强方法，例如机器翻译（Cai 等人，2019年）和问答（Karpukhin 等人，2020年）。与 REGROUP 类似，先前的工作（Gupta 等人，2022年；Pasupat 等人，2021年）利用检索步骤，为序列到序列模型的学习提供输入上下文示例。

然而，我们的方法在两个方面有所不同：（1）我们的检索对象是经过对齐的公式化知识，其信息比数据示例更为精炼；（2）以往的工作会直接利用检索结果。我们利用落地模型编辑检索到的公式化知识，使其与问题和模式更加相关。

8 结论与未来工作

本文探讨公式化知识，以应对知识密集型文本到 SQL 问题，从而推动文本到 SQL 在专业领域的应用，例如领域专家进行数据分析。首先，我们分析了知识密集型文本到 SQL 面临的挑战，并构建了一个新的高难度基准 KNOWSQL。随后，我们从公式化知识的角度提出解决这一问题的方法。具体而言，我们提出了一个简单的框架 REGROUP，以利用外部公式化知识库。实验结果表明，结合公式化知识的 REGROUP 整体性能提升了 28.2%。

我们进一步分析不同类型的错误案例，探讨改进 REGROUP 的三个方向：（1）迭代填补公式化知识库中的空白；（2）通过改进落地模型，缩小公式化知识与特定模式之间的差距；（3）推动解析器充分利用更复杂的公式化知识（例如常识）。

伦理考量

本研究推出了 KnowSQL，这是一个免费开放的数据集，供研究社区研究知识密集型文本到 SQL 问题。KnowSQL 中的数据基于 DuSQL（Wang 等人，2020b）构建；DuSQL 是一个免费开放的跨数据库中文文本到 SQL 数据集。我们还从中国知网[12]和百度文库[13]，收集了公式化数据和表格数据，这些资源也可免费开放用于学术研究。表格内容经过匿名化处理，以解决隐私问题。为标注 KnowSQL，我们招募了 3 名中国大学生（1 名女性、2 名男性）。每位学生标注（SQL，问题）对可获得 4 元（0.6 美元），收集公式化知识项可获得 2 元（0.3 美元）。这项报酬是参照以往类似数据集的构建工作确定的（Guo 等人，2021年）。由于所有问题序列均从开放访问的数据库或公共表格中收集，因此不存在隐私问题。

局限性

（1）KNOWSQL 基于 DuSQL 构建，DuSQL 是一个中文大规模文本到 SQL 数据集。因此，本文的语言覆盖范围仅限于中文。我们将其他语言的扩展留待未来工作。（2）对于公式化知识的范围，我们主要处理与每个 SQL 短语相关联的三类知识：计算、并集和条件。一些类型的知识尚未得到充分探索，例如常识知识。（3）对于 REGROUP 的模型设计，我们在改进许多现有工作的基础上构建了该模型。尽管取得了颇具前景的评估结果，但案例研究表明，在检索、对齐或解析过程中仍存在许多挑战。

致谢

我们感谢所有匿名审稿人提出的建设性意见。Wanxiang Che 获得了 2020年AAA0106501 项目以及国家自然科学基金项目 62236004 和 61976072 的资助。Dechen Zhan 是通讯作者。

参考文献

Deng Cai、Yan Wang、Wei Bi、Zhaopeng Tu、Xiaojiang Liu 和 Shuming Shi。2019年。检索-引导的对话回复生成：一种从匹配到生成的框架。载于*2019*年自然语言处理经验方法会议暨第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第1866–1875页。计算语言学协会。

Ruisheng Cao、Lu Chen、Zhi Chen、Yanbin Zhao、Su Zhu 和 Kai Yu。2021年。LGESQL：融合局部与非局部关系的线图增强型文本到 SQL 模型。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第2541–2555页。计算语言学协会。

Yiming Cui、Wanxiang Che、TingLiu、Bing Qin 和 Ziqing Yang。2021年。中文 BERT 的全词掩码预训练。*IEEE/ACM* 音频、语音与语言处理汇刊，29:3504–3514。

Longxu Dou、Yan Gao、Mingyang Pan、Dingzirui Wang、Wanxiang Che、Dechen Zhan 和 Jian-Guang Lou。2022年。UniSAr：一种用于文本到 SQL 的统一结构感知自回归语言模型。arXiv。

Catherine Finegan-Dollak、Jonathan K. Kummerfeld、Li Zhang、Karthik Ra manathan、Sesh Sadasivam、Rui Zhang 和 Dragomir Radev。2018年。改进文本到 SQL 评估方法。载于第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第351–360页。计算语言学协会。

Jiaqi Guo、Ziliang Si、Yu Wang、Qian Liu、Ming Fan、Jian-Guang Lou、Zijiang Yang 和 Ting Liu。2021年。Chase：用于跨数据库上下文相关的文本到 SQL 的大规模实用中文数据集。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第2316–2331页，线上。计算语言学协会。

Vivek Gupta、Akshat Shrivastava、Adithya Sagar、Armen Aghajanyan 和 Denis Savenkov。2022年。RetroNLU：检索增强的面向任务语义解析。载于 第*4*届对话式人工智能自然语言处理研讨会论文集，第184–196页。计算语言学协会。

Jonathan Herzig 和 Jonathan Berant。2019年。不要改述，要检测！快速有效地收集语义解析数据。载于《*2019*年自然语言处理经验方法会议暨第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集》，第3810–3820页，中国香港。计算语言学协会。

[12] https://oversea.cnki.net/index/ [13] https://wenku.baidu.com/

Jonathan Herzig、Pawel Krzysztof Nowak、Thomas Müller、Francesco Piccinno 和 Julian Eisenschlos。2020年。TaPas：通过预训练进行弱监督表格解析。载于《计算语言学协会第*58*届年会论文集》，第4320–4333页。计算语言学协会。

Hsin-Yuan Huang、Eunsol Choi 和 Wen-tau Yih。2019年。Flowqa：把握对话式机器理解中的历史信息流。载于第七届国际学习表征会议（*ICLR*）。

Jeff Johnson、Matthijs Douze 和 Hervé Jégou。2017年。利用图形处理器进行十亿规模的相似性搜索。arXiv。

Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。用于开放域问答的密集段落检索。载于*2020* 年自然语言实证方法会议（*EMNLP*）论文集，第6769–6781页。

Wenqiang Lei、Weixin Wang、Zhixin Ma、Tian Gan、Wei Lu、Min-Yen Kan 和 Tat-Seng Chua。2020年。重新审视模式链接在文本到SQL中的作用。载于*2020* 年自然语言实证方法会议（*EMNLP*）论文集，第6943–6954页，线上。计算语言学协会。

Mike Lewis、Yinhan Liu、Naman Goyal、Mar- jan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于计算语言学协会第*58*届年会论文集，第7871–7880页，线上。计算语言学协会。

Qian Liu、Dejian Yang、Jiahui Zhang、Jiaqi Guo、Bin Zhou 和 Jian-Guang Lou。2021年。从预训练语言模型中唤醒用于语义解析的潜在对齐。载于计算语言学协会论文集：*ACL-IJCNLP 2021*，第1174–1189页。

Myle Ott、Sergey Edunov、Alexei Baevski、Angela Fan、Sam Gross、Nathan Ng、David Grangier 和 Michael Auli。2019年。fairseq：用于序列建模的快速、可扩展工具包。载于*2019*年北美洲计算语言学协会分会年会论文集（演示），第48–53页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Panupong Pasupat、Yuan Zhang 和 Kelvin Guu。2021年。通过检索增强实现可控语义解析。载于 *2021*年自然语言处理实证方法会议论文集，第 7683–7698 页。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。2019年。PyTorch：一种命令式风格、高性能的深度学习库。载于 神经信息处理系统进展第 *32* 卷，第 8024–8035 页。

Peng Qi、Yuhao Zhang、Yuhui Zhang、J ason Bolton 和 Christopher D. Manning。2020年。Stanza：适用于多种人类语言的 Python 自然语言处理工具包。计算语言学协会第*58*届年会：系统演示论文集，第101–108页。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索迁移学习的极限：统一的文本到文本转换器。机器学习研究期刊，21(140):1–67。

Stephen Robertson 和 Hugo Zaragoza。2009年。概率相关性框架：*BM25*及其扩展。《信息检索基础与趋势》。

Torsten Scholak、Nathan Schucher 和 Dz mitry Bahdanau。2021年。PICARD：针对语言模型的受约束自回归解码进行增量解析。载于*2021*年自然语言处理经验方法会议论文集，第9895–9901页。

Richard Shin、Christopher H. Lin、Sam Thomson、Charles Chen、Subhro Roy、Emmanouil Antonios Platanios、Adam Pauls、Dan Klein、JasonEisner 和 Benjamin Van Durme。2021年。受约束的语言模型带来少样本语义解析器。*CoRR*，abs/2104.08768。

Bailin Wang、Mirella Lapata 和 Ivan Titov。2021年。面向语义解析领域泛化的元学习。载于计算语言学协会北美分会 *2021*年年会论文集：人类语言技术，第 366–379 页。计算语言学协会。

Bailin Wang、Richard Shin、Xiaodong Liu、Oleksandr Polozov 和 Matthew Richardson。2020a。RAT-SQL：面向文本到SQL解析器的关系感知模式编码与链接。载于计算语言学协会第 *58* 届年会论文集，线上。计算语言学协会。

Lijie Wang、Ao Zhang、Kun Wu、Ke Sun、Zhenghua Li、HuaWu、Min Zhang 和 Haifeng Wang。2020b。DuSQL：大规模且实用的中文文本到 SQL 数据集。发表于*2020*年自然语言处理实证方法会议论文集

处理（*EMNLP*），第 6923–6935 页，线上。计算语言学协会。

Yushi Wang、Jonathan Berant 和 Percy Liang。2015年。一夜之间构建语义解析器。载于计算语言学协会第*53*届年会和第*7*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第1332–1342页，中国北京。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformer：最先进的自然语言处理。载于*2020*年自然语言处理经验方法会议：系统演示论文集，第38–45页。

Tianbao Xie、Chen Henry Wu、Peng Shi、Ruiqi Zhong、Torsten Scholak、Michihiro Yasunaga、Chien-Sheng Wu、Ming Zhong、Pengcheng Yin、Sida I.Wang、Victor Zhong、Bailin Wang、Chengzu Li、Connor Boyle、Ansong Ni、Ziyu Yao、Dragomir Radev、Caiming Xiong、Lingpeng Kong、Rui Zhang、Noah A. Smith、Luke Zettlemoyer 和 Tao Yu。2022年。Unifiedskg：使用文本到文本语言模型统一并开展多任务结构化知识落地。*arXiv*预印本 *arXiv:2201.05966*。

Pengcheng Yin、Graham Neubig、Wen-tau Yih 和 Se- bastianRiedel。2020年。TaBERT：面向文本数据和表格数据联合理解的预训练。载于第*58*届计算语言学协会年会论文集，第8413–8426页。计算语言学协会。

Pengcheng Yin、John Wieting、Avirup Sil 和 Graham Neubig。2021年。论构建有效零样本语义解析器的要素。arXiv。

Tao Yu、Rui Zhang、Heyang Er、Suyi Li、Eric Xue、Bo Pang、Xi Victoria Lin、Yi Chern Tan、Tianze Shi、Zihan Li、Youxuan Jiang、Michihiro Yasunaga、Sungrok Shim、Tao Chen、Alexander Fabbri、Zifan Li、Luyao Chen、Yuwen Zhang、Shreya Dixit、Vin- cent Zhang、Caiming Xiong、Richard Socher、Walter Lasecki 和 Dragomir Radev。2019a。CoSQL：一项旨在实现面向数据库的跨领域自然语言接口的对话式文本到SQL挑战。载于*2019*年自然语言处理实证方法会议与第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第1962–1979页，中国香港。计算语言学协会。

Tao Yu、Rui Zhang、Kai Yang、Michihiro Yasunaga、Dongxu Wang、Zifan Li、James Ma、Irene Li、Qingning Yao、Shanelle Roman、Zilin Zhang，和 Dragomir Radev。2018年。Spider：一个大规模、由人工标注的数据集，用于复杂的跨领域语义解析和文本到 SQL 任务。载于*2018*年自然语言处理实证方法会议论文集，第 3911–3921 页。

Tao Yu、Rui Zhang、Michihiro Yasunaga、Yi Chern Tan、Xi Victoria Lin、Suyi Li、Heyang Er、Irene Li、Bo Pang、Tao Chen、Emily Ji、Shreya Dixit、David Proctor、Sungrok Shim、Jonathan Kraft、Vin- cent Zhang、Caiming Xiong、Richard Socher 和 Dragomir Radev。2019年。SParC：语境中的跨领域语义解析。载于 第*57*届计算语言学协会年会论文集，第4511–4523页，意大利佛罗伦萨。计算语言学协会。

Jichuan Zeng、Xi Victoria Lin、Steven C.H. Hoi、Richard Socher、Caiming Xiong、Michael Lyu 和 Irwin King。2020年。Photon：一种稳健的跨领域文本到 SQL 系统。载于 第*58*届计算语言学协会年会论文集：系统演示，第204–214页。计算语言学协会。

Chen Zhao、Yu Su、Adam Pauls 和 E mmanouil An- tonios Platanios。2022年。通过模式扩展弥合文本到 SQL 解析中的泛化差距。载于《第*60*届计算语言学协会年会论文集》（第*1*卷：长篇论文），第5568–5578页。计算语言学协会。

Victor Zhong、Caiming Xiong 和 Richard Socher。2017年。Seq2sql：利用强化学习，从自然语言生成结构化查询。*CoRR*，abs/1709.00103。

A 公式化知识库详情

A.1 知识来源

我们遵循 KNOWSQL，在 19 个领域和 1 个其他领域构建公式化知识库。其他领域存储 KNOWSQL 中不常见或通用的知识项，例如密度和速度的计算。接下来，我们将简要分析收集到的知识库。

A.2 跨领域统计

不同领域在线公开可用的数据量各不相同。如图所示：6，毫不意外，金融和房地产拥有最丰富的公开资源。

A.3 领域内分布

我们还观察到不同领域的知识分布各不相同。如果领域侧重于计算（例如，金融报告和基金），我们认为数据分析往往更加客观，因此模型更容易学习。如果领域侧重于条件（例如，房地产和奖项），我们认为数据分析往往更加主观，因为学习语义更具挑战性。

B REGROUP 的实现细节

检索器 我们根据 Karpukhin 等人 (2020年)[14]的代码实现检索器。我们采用 Chinese BERT-wwm-ext（Cui 等人，2021年）作为预训练编码器[15]。它会返回排名前三的检索到的公式化知识。未来的工作可以参照 Karpukhin 等人 (2020年)，通过批内采样或基于 BM25 的采样改进负采样。

落地模型 提交本文时，ETA[16] 的代码尚未发布。我们依据论文使用 PyTorch 重新实现了 ETA 模型（Paszke 等人，2019年）。我们在 SPIDER-L（Lei 等人，2020年）上将自行实现的模型与原始模型进行评估，以检验重新实现的模型是否有效。我们的模型取得了 82.1% 的列 F1 分数，而Liu 等人 （2021年）报告的结果为 82.5%。在 KNOWSQL 上进行的实验也使用了中文 BERT。

解析器 我们基于以下论文的代码构建本文：Dou 等人 (2022年)[17] 我们选择 mBART- CC 25[18] 作为基础模型进行微调。遵循原始模型，我们将解析器的输入构建如下： *[模式]* | *[对齐后的公式化知识]* | *[问题]*,其中 '|' 是不同部分之间的分隔符。

资源与工具 在分词方面，考虑到 Stanza（Qi 等人，2020年）表现出色，我们采用了 Stanza。对于检索器和落地模型，我们使用 Transformer 库（Wolf etal.，2020年）导入 BERT 模型。对于解析器模式，我们对数据进行预处理，并使用 fairseq 框架（Ott 等人，2019年）微调 mBART

设备与训练时间 我们在一块 NVIDIA TESLA V100-32GB GPU 上完成了所有这些实验。检索器、落地模型和解析器的训练分别耗时约 4 小时、3 小时和 8 小时。微调 mBART 所需的最低设备配置为 NVIDIA TESLA P100-16G。

超参数 所有超参数均与各模型引用论文中的设置保持一致。唯一的区别是检索器和落地模型的批量大小；我们将其调至可在 NVIDIA TESLA V100-32G GPU 上运行的最大值。

[14]DPR 检索模型代码[15]中文 BERT-wwm 模型[16]ETA 对齐模型代码

[17]UniSAr解析器代码[18] mBART 模型

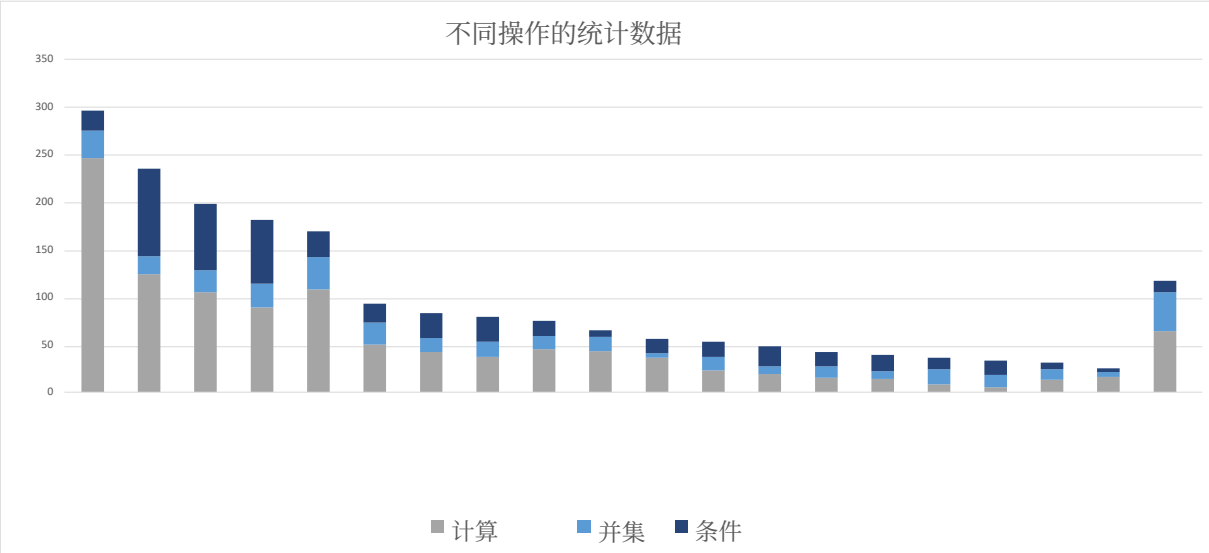

图6：不同领域中三种操作的统计

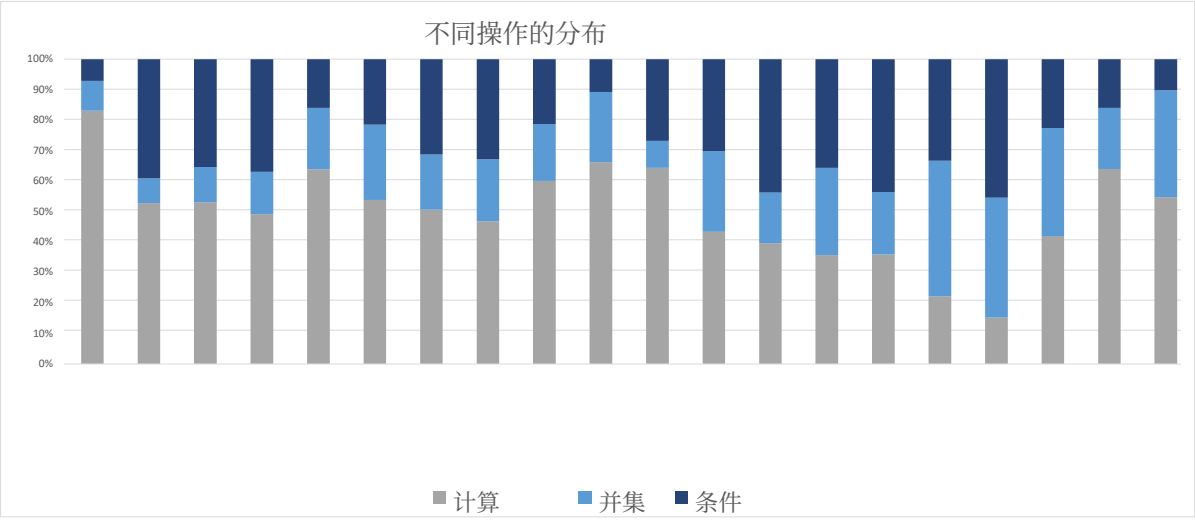

图7：不同领域中三种操作的分布