使用统计分析和大型语言模型对表格数据进行关系检测

Panagiotis Koletsis
Harokopio University
Athens, Greece
pkoletsis@hua.gr

Christos Panagiotopoulos
Harokopio University
Athens, Greece
chris.panagiotop@hua.gr

Georgios Th. Papadopoulos
Harokopio University
Athens, Greece
g.th.papadopoulos@hua.gr

Vasilis Efthymiou
Harokopio University
Athens, Greece
vefthym@hua.gr

Abstract

在过去的几年中，由于表格解释任务的重要性以及该领域新技术和基准的引入，该任务取得了显著的进展。此项工作实验了一种混合方法，用于检测未标记表格数据列之间的关系，该任务被称为 CPA，以知识图谱（KG）作为参考点。该方法利用大型语言模型（LLM）同时运用统计分析减少潜在 KG 关系的搜索空间。减少搜索空间的主要模块包括领域和范围约束检测，以及关系共现分析。实验性评估基于 SemTab 挑战提供的两个基准数据集，评估各模块的影响以及不同量化水平的最新 LLM 的效果。实验不仅在不同的提示技术上进行了，而且公开的研究方法在 github 上可用，被证明在这些数据集上与最新方法具有竞争力。

VLDB 研讨会参考格式：
Panagiotis Koletsis, Christos Panagiotopoulos, Georgios Th. Papadopoulos, and Vasilis Efthymiou. 使用统计分析和大型语言模型对表格数据进行关系检测. VLDB 2024 研讨会：Relational Models.
PVLDB 程序可用性:

源代码、数据和/或其他材料已在 https://github.com/panagiotis-koletsis/cpa4 提供。

1 介绍

理解表格数据的结构、语义和内容是数据集成、信息检索和知识图谱（KG）构建中的基础任务。此过程中一个重要的步骤是确定表格中的列如何相互关联。这一问题即使在表格中存在列标头时（由于数据异构性）已经较为困难，但当表格缺乏有意义的标头时，即便对人类专家而言，也变得特别具有挑战性。SemTab 挑战的列属性注释（CPA）任务 [10] 针对这一问题，旨在识别列之间的语义关系，例如“作者”或“位于”，这些关系在参考知识图谱中有定义。在实践中，这项任务特别具有挑战性，因为缺乏信息丰富的列名称，KG 属性的语义广泛且相互重叠，以及潜在关系的庞大搜索空间（例如，DBpedia 中大约有 3,000 个属性，而 Schema.org 中有 1,500 个属性）。

最近提出了依赖于大型语言模型（LLMs）的语义表格解释方法，并取得了令人鼓舞的结果 [2, 5, 8] 。然而，LLMs 也容易出现幻觉现象，因此所谓的知识注入被建议作为应对这一问题的方法 [15] 。因此，在这项初步工作中，我们探索使用数据分析来检测约束，从而缩小列之间候选关系的搜索空间，从而在提示 LLM 执行 CPA 任务之前减少出错的机会。具体来说，我们从目标知识图谱中检测候选关系的领域和范围限制，在我们的案例中是 Schema.org，并将其与输入表中检测到的领域和范围限制结合起来。此外，我们从训练集中提取共同出现的统计数据，并利用这些数据在为该表检测到一些初始关系时，逐步减少表中的候选关系。最后，我们探索了几种 LLM 和提示技术，并评估了它们的性能。

总之，这项工作的贡献如下：

- 结合 LLMs 和统计分析的 CPA 混合方法。
- 对几个低十亿参数的 LLMs 进行实验评估，使用不同的量化级别和提示技术。
- 消融研究用于确定我们混合方法中每个模块的影响。
- 在 https://github.com/panagiotis-koletsis/cpa4 上公开提供的 CPA 开源解决方案。

几十年来，人们已经认识到理解不当存储的数据的挑战，从而促使探索新的解决方案，这些解决方案主要基于统计和机器学习 [20, 21] ，以及基于查找和基于大语言模型的方法 [4] 。

基于统计的方法。在该领域的基础工作之一 [3] 解决了以下问题：通过统计方法实现模式自动补全、同义词查找和合并模式。通过结合共同出现和计算概率，可以实现模式自动补全。例如，当用户输入制造商名称时，系统会建议型号、年份、价格、里程和颜色。通过知道同义词不可能在同一模式中共同出现并且它们具有相似的内容，可以实现同义词查找。最后，他们通过聚类相似的模式来合并模式。我们的共同出现统计数据受到这项工作的启发。

基于查找的方法。MTab [17] 曾在 SemTab 挑战赛的头几年占据主导地位，它通过在目标知识图谱（例如 DBpedia）上建立的本地索引进行实体查找。此外，MTab 引入了一种文字匹配方法，用于将表格单元格值与知识图谱的相应属性进行对齐 [17] 。DAGOBAH [9] 是一个后来在 SemTab 挑战赛中占据主导地位的系统，进一步探索了上述逻辑，引入了预处理步骤，例如方向检测、表头检测、关键列检测和列原始类型的识别，同时也引入了预评分算法。其中一些想法在 SemTab 之前的相关研究中已经被提出（例如，[7, 13, 19] ）。

基于 LLM。随着近年来 LLM 的各种进展，研究人员探索了其在表格解释任务中的能力。例如，[2] 使用了 GPT3 模型结合少量样本学习和零样本学习来解决 SemTab 任务。更先进的工作进一步探索了 LLM 的微调以用于表格解释任务，[14] 甚至采用了基础模型的方法用于表格解释 [22] 。

混合。有趣的是，一些混合型工作尝试利用上述方法的优势，例如结合 LLM 和基于查找的技术 [8] 。TorchicTab [5] 采用了一种用于语义表注释的双系统方法，将启发式与分类相结合。启发式组件 TorchicTab-Heuristic 使用 Wikidata 等 RDF 图进行候选查找，并通过多种搜索策略和语义相似性进行排名，通过上下文分析和多数票策略处理多样的表结构。分类组件 TorchicTab-Classification 将列类型和属性注释视为使用 DODUO 语言模型（一个微调过的 BERT [6] ）以及子表采样来管理标记限制和保持上下文的多分类任务。这种结合的方法在 SemTab 2023 挑战中取得了顶级表现。

VLDB 基金会论文集。ISSN 2150-8097。

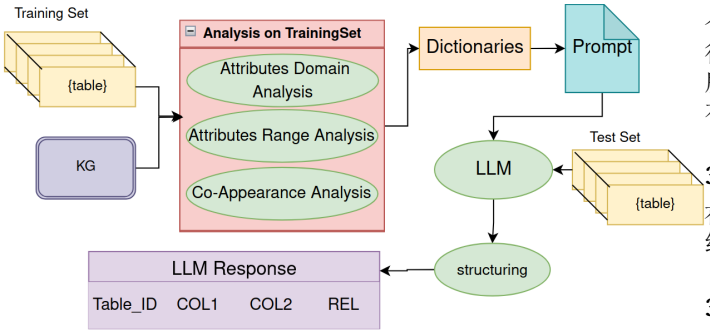

Figure 1: 方法论概述。

2 方法论

在本节中，我们描述了我们的方法，遵循图 1 中提出的概述。

基础。在基础架构中，表格、感兴趣的列以及可能关系的完整列表被解析到模板中，然后由 LLM 处理（跳过图 1 的分析和词典）。如果输出不符合所需格式（即，关系列表中的一个单词），通过解析第一次提示的输出并请求一个感兴趣的单词应用一种简单但有效的技术。如果再次失败，我们就用初始提示重新询问不同的模型。

领域属性分析。领域属性分析不只是传递 schema.org 中发现的可能关系的整个列表，而是专注于为每种实体类型识别一个更小、更相关的关系集。这是可行的，因为训练集表中的每个数据点都从对应的领域开始。为了实现我们的目标，我们扫描训练集，在包含所有属性类型的字典中附加相应的关系到字典中的对应键。

范围属性分析。基于初步的想法，我们通过根据它们的原始类型进行过滤，进一步缩小了关系的搜索空间。所提出的用于属性范围分析的模块扫描训练集，并读取数据点中相应列的表格的第一个元素，然后使用解析器和正则表达式将其分类为四种类型之一 - 字符串、数字、日期和 URL。生成的字典有一个主要缺点：由于数据集的不一致性和日期解析过程的原因，它包含了一些异常值。例如，一些关系可能出现在超过 2-3 种原始类型中。为了改进这个模块，我们引入了一个阈值来排除很少使用的关系。这个阈值丢弃了所有出现频率低于最频繁的关系的 5 % 的关系。这样一来，数据集中包含的一些异常值就不会被 LLM 考虑，因此不会对 LLM 的提示添加噪声。

共现分析。共现模块背后的理念是，如果某些关系在训练集中的同一个表中没有共现，那么在测试集中的同一个表中共现的可能性也不大。对于特定领域的每个关系，我们构建一个相应的词典（例如，我们为属性 Book.name 构建一个词典，并为 Person.name 构建另一个词典）。这个词典将包含在同一个表中与初始关系共现的其他关系，希望能进一步缩小搜索空间。此外，在解析表的列时，LLM 对前几列的预测也会从下一列的候选集合中移除，因为我们假设一个关系在每个表中只出现一次。这种技术有一个主要缺点：为了成功，模型的初始预测必须是正确的。否则，下一列的候选关系的搜索空间会越来越小，从而导致错误的传播，并要求模型从不包含正确关系的列表中选择一种关系。在这个初步的方法中，我们从左向右逐个扫描列，不过其他的方法（例如，从最有可能被正确标注到最不可能的顺序）也值得探索。

推理阶段。在推理阶段，首先提取表格的领域（即 SemTab 的表格主题检测 - TD - 任务）。然后，对于每个真实值中的列，通过上述过程 (R) 确定其类型。最后，在提示模板中提供这两个列表的交集。值得一提的是 (a) 我们从每个表格中抽取 500 行样本进行分析，并且 (b) 提供的脚本会动态运行并自动生成所有字典，无需人工操作。这个步骤离线进行，使用单独的脚本，因为每个数据集只需运行一次。

3 实验

在本节中，我们描述实验设置，并提供和分析我们研究的实验结果。

3.1 实验设置

硬件-软件。所有实验都在单个 GPU（RTX 3080 Ti，12GB VRAM）上进行。通过 Ollama（v.0.6.6）和 Langchain 实现对模型的访问。

模型。重点主要在于来自 Ollama 的最新模型，这些模型从 120 亿到 720 亿参数不等，并进行了 2 位到 4 位的量化。测试了 DeepSeek-r1、Gemma3、phi4、Llamma3.3 等模型，并对 `Qwen2.5:32B-Instruct-Q3_K_L` 进行了更详细的评估。

数据集。在我们的实验中，我们使用 Round 1 和 Round 2 的 SemTab 2023 SOTAB [12] 数据集进行 CPA 任务，schema.org 作为其目标知识图谱。Round 1 数据集由来自 7 个不同领域的 50 个预定义关系构成。Round 2 数据集的格式与 R1 数据集完全相同，但它由来自 17 个不同领域的 105 个关系构成。此外，它的数据点数量几乎是 R1 的三倍。

评价指标。我们采用标准的 SemTab 评估指标，即准确率、召回率、Macro_F1 和 Micro_F1，以及以秒为单位的执行时间。为了确保所有方法的评估一致，SemTab 挑战赛提供了一个 Python 评估脚本。

基础。测试了具有不同量化水平的多种大型语言模型（LLM）。如 Table 1 所示，Qwen 系列模型表现出色。这可能是因为 Qwen 模型专门被训练来理解和生成结构化格式。Qwen-32B 的失败迭代次数较少，表明其输出格式更好。出乎意料的是，尽管 LLaMA 3.3 体积更大，却未能表现出竞争力。Qwen2.5-72B Q2 与 Qwen2.5-32B Q3 相比，结果不一致且失败迭代次数显著增多，这表明蒸馏影响了其处理结构化输出的能力。其他模型没有竞争力，并且由于激活了结构化组件而导致执行时间增加。

领域（D）。在两个数据集（Table 1 和 Table 2 ）中，领域提取显示了最显著和清晰的改进，表明缩小和更相关的搜索空间有助于 LLM 更好地执行。此外，两个数据集均表现出较低的失败率和执行时间。在 R1 的情况下，每次迭代的执行时间相比基础减少了 2.5 %。这表明更少的选择使模型能够进行更高效的决策。在 R1 数据集中，Micro_F1 分数提高了 16.7 %，而在 R2 中，同一指标提高了 17.3 %。

范围（R）。在 R1 数据集上，单独使用该方法相比基线提高了所有发现的指标。最显著的是精确度提高了 6 % 。在 R2 数据集上，所有观察到的指标都下降了。然而，类似于领域方法，在这两个数据集上都注意到了执行时间的减少。

共现（C）。单独使用时，它对性能有明显的负面影响，因为由 LLM 产生的任何初始错误都会在后续迭代中传播。结果证实了这种行为，显示了性能指标的显著下降和失败率的大幅上升。

范围 & 域（RD）。尽管单独的类型分析影响不一，类型和域分析的结合在两个数据集中都实现了最佳结果。在 R1 数据集中，相比仅使用域分析，Micro_F1 额外提高了 1.2 %，相比基础则提高了 18.2 %。在 R2 数据集中，Micro_F1 又进一步提高了 5 %，同时精确度提高了 6.5 %。值得注意的是，Micro_F1 相比基础提高了 23 %。虽然所提出的方法并没有达到 SOTA 性

Table 1: CPA 在 R1 数据集上。对于方法，D：域，R：范围，C：共现，RD：范围 & 域，RDC：范围 & 域 & 共现，RDC_p：范围 & 域 & 共现（精度 =1）。对于 LLMs，L：llama3.3:70b-instruct-q2_K，G：gemma3:12b，Q32：qwen2.5:32b-instruct-q3_K_L，Q72：qwen2.5:72b-instruct-q2_K，R1：deepseek-r1:14b，Φ：phi4:14b。

Approach	Macro_F1	Micro_F1	P	R	Time(s)	LLM
Kepler-aSI [1]	-	0.235	0.230	-	-	-
Base	0.346	0.435	0.483	0.363	9.6	L
Base	0.638	0.703	0.687	0.649	11.3	Q72
Base	0.309	0.372	0.440	0.310	1.3	G
Base	0.313	0.381	0.455	0.288	10.9	R1
Base	0.202	0.453	0.249	0.206	6.46	Φ
Base	0.634	0.687	0.706	0.662	3.9	Q32
R	0.679	0.694	0.748	0.681	3.8	Q32
D	0.771	0.802	0.801	0.792	3.8	Q32
C	0.054	0.232	0.092	0.051	4.1	Q32
RD	0.792	0.812	0.834	0.801	3.8	Q32
RDC	0.770	0.796	0.803	0.769	3.9	Q32
RDC_p	0.802	0.820	0.845	0.810	3.7	Q32
RDC_p	0.441	0.575	0.544	0.428	1.6	G
RDC_p	0.185	0.339	0.248	0.179	7.5	R1
RDC_p	0.227	0.566	0.269	0.216	11.0	Φ

能，但其相较于 TorchicTab 和 MUT2KG 有一个重要优势。这个优势在于，该方法无需任何微调，而上述方法则需要。

范围 & 域 & 共现（RDC）。共现模块的负面影响延续到类型、域和共现的组合中，导致所有观察指标的性能下降，并且在两个数据集中执行时间和失败率都增加。

基于精确度（RDC_p）的范围 & 域 & 共同出现性。在这个实验中，我们测试了仅在模型能够以 100 % 的精确度识别的关系上激活共同出现模块。这是有意义的，因为 100 % 的精确度意味着每当模型识别到该关系时，它都能正确识别。为此，在验证集中计算了每个类别的得分。然后手动仅对 LLM 识别出的这些关系进行激活。在 R1 数据集中，micro_f1 得分提高了 1 %，表明这种方法有效。不幸的是，在 R2 数据集中，这种方法未能提高结果，但与使用完整共同出现性的方式相比，micro_f1 提高了 15 %。

提示消融。在 R1 数据集上，出乎意料的是，排除 CoT 得到的 micro_F1 分数最高。排除示例也产生了积极的效果。虽然这最初被认为是意外的，但可以通过 R1 比 R2 更简单，额外的信息可能会让模型困惑来解释。通过提示工程观察到 micro_F1 分数总体改善了 4.5 %。在 R2 数据集上，包含角色、示例和 COT 的提示取得了最佳结果。每个提示部分的排除也产生了负面影响。总的来说，使用提示工程使 micro_f1 分数增加了 5.4 %。尽管 LLM 在零样本性能上有所改善，提示工程仍是不费力提升结果的方法。然而，它应该被验证，因为在提示中引入一个看似有益的组件实际上可能产生负面影响。

大型语言模型的比较。不同的模型在初始和最终设置上，以不同的量化级别和参数进行了测试。在 R1 阶段，DeepSeek 和 Phi-4 在结构化方面表现不佳，激活结构化组件导致执行时间和失败率增加。Gemma3 在执行时间方面表现较好，但仍不如所提出的模型具有竞争力。在 R2 阶段，所有模型的结果均有所下降，表明数据集复杂度更高。虽然 Phi-4 在输出质量方面优于 DeepSeek，但其执行时间超过四倍。Gemma3 由于频繁的意外标记错误而无法完成数据集测试，结果为空白。

Table 2: 针对 R2 数据集上的 CPA。对于方法，D：域，R：范围，C：协应用，RD：范围 & 域，RDC：范围 & 域 & 协应用，RDC_p：范围 & 域 & 协应用（精度 =1）。对于 LLM，G：gemma3:12b，Q32：qwen2.5:32b-instruct-q3_K_L，R1：deepseek-r1:14b，Φ：phi4:14b。

Approach	Macro_F1	Micro_F1	P	R	Time(s)	LLM
TSOTSA [11]	-	0.235	0.434	-	-	-
Anu	-	0.623	0.791	-	-	-
TorchicTab [5]	-	0.871	0.880	-	-	-
MUT2KG [16]	-	0.793	0.848	-	-	-
DREIFLUSS [18]	-	0.174	0.320	-	-	-
Base	0.559	0.630	0.655	0.575	4.2	Q32
R	0.542	0.608	0.622	0.565	4.1	Q32
D	0.704	0.739	0.756	0.717	4.0	Q32
C	0.026	0.146	0.042	0.026	4.3	Q32
RD	0.756	0.776	0.797	0.763	3.9	Q32
RDC	0.478	0.654	0.525	0.474	4.7	Q32
RDC_p	0.685	0.749	0.734	0.695	4.1	Q32
RD	-	-	-	-	-	G
RD	0.137	0.296	0.184	0.132	7.9	R1
RD	0.200	0.548	0.239	0.192	27.5	Φ

Table 3: 关于提示的消融研究。RD：范围 & 领域，RDC_p：范围 & 领域 & 共同应用（精度 =1），E：示例，E+COT：示例 +COT。

Approach	Macro_F1	Micro_F1	P	R	Time(s)	Without
Dataset: R1						
RDC_p	0.766	0.799	0.816	0.773	4.0	ALL
RDC_p	0.813	0.835	0.844	0.818	3.7	COT
RDC_p	0.786	0.809	0.839	0.794	3.7	Role
RDC_p	0.805	0.826	0.845	0.816	3.7	E
RDC_p	0.782	0.796	0.834	0.791	3.7	E+COT
RDC_p	0.802	0.820	0.845	0.810	3.7	-
Dataset: R2						
RD	0.728	0.736	0.798	0.729	3.8	ALL
RD	0.691	0.729	0.755	0.698	4.0	COT
RD	0.722	0.763	0.775	0.729	3.9	Role
RD	0.688	0.718	0.758	0.697	3.9	E
RD	0.711	0.724	0.786	0.716	3.9	E+COT
RD	0.756	0.776	0.797	0.763	3.9	-

在这项工作中，探索了一种基于混合大型语言模型（LLM）和统计分析的方法用于 CPA 任务。所使用的 LLM 在参数大小和量化级别上有所不同。为减少关系的搜索空间，探索了三个主要组件：域、范围和共现。每个组件分别基于表的域、列的基本类型，以及共同出现的关系来减少关系。最后，对构建的提示进行了消融研究，展示了正确使用提示的重要性。域和范围方法的结合被证明在 SemTab 中与最先进的方法具有竞争力。

未来，我们计划通过选择关系最少的列作为 LLM 的起始点来增强共同出现的方法，以减少错误传播。此外，我们还将探索一种使用预训练嵌入来分类每一列的替代方法。在这种方法中，目标列中每一行的嵌入将被聚合，然后将结果向量与每一个可能关系的预训练嵌入进行比较。在嵌入空间中选择最接近的匹配项，理想情况下能捕捉到正确的语义含义。

```
PROMPT_TEMPLATE = """
    You are an expert in table interpretation, specialized in relations identification among columns.
    ----Knowing that this table is about {domain}----{table}----
    On the provided table i want you to answer this question:
    What is the relation between column 0 and column {num}.
    ----***All the possible relations are in this list {relations}. Examine all the possible provided
    relations and keep the most suitable one!Be careful to only use the list of relations!
    ----Example-------
    Knowing that this table is about Book----
    On the table below i want you to answer this question:
    What is the relation between col 0 and col 1
    |0|1|2|3|4|
    |Schaum's Outline of Operating|McGraw-Hill Education|20,39|EUR|9780071364355|
    |Le petit bonhomme de pain d'épices. Niveau 1|Black Cat-Cideb|6,80|EUR|9788853010858|--
    Desired output --publisher-------
    Think step by step -1. Locate the 2 columns -2. Extract context from the other columns
    -3. Identify the most suitable relation from the relation list
    Provide only the desired output (1 word) and nothing more!
    Dont explain yourself and don't provide extra comments! -----"""
```

Figure 2: 我们在这项工作中使用的主要提示，及其不同部分以不同颜色突出显示（请参阅提示消融）。

Acknowledgments

This work has received funding from the EU Horizon Europe research and innovation programme through the GANNDALF project (GA No. 101084265). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the EU.

References

[1] Wiem Baazouzi, Marouen Kachroudi, and Sami Faïz. 2023. Kepler-aSI at SemTab 2023. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3557)*. 85–91.

[2] Jean Petit Bikim, Carick Appolinaire Atezong Ymele, Azanzi Jiomekong, Allard Oelen, Gollam Rabby, Jennifer D'Souza, and Sören Auer. 2024. Leveraging GPT Models For Semantic Table Annotation. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3889)*. 43–53.

[3] Michael J. Cafarella, Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. WebTables: exploring the power of tables on the web. *Proc. VLDB Endow.* (2008), 538–549.

[4] Marco Cremaschi, Blerina Spahiu, Matteo Palmonari, and Ernesto Jiménez-Ruiz. 2024. Survey on Semantic Interpretation of Tabular Data: Challenges and Directions. *CoRR* abs/2411.11891 (2024).

[5] Ioannis Dasoulas, Duo Yang, Xuemin Duan, and Anastasia Dimou. 2023. TorchicTab: Semantic Table Annotation with Wikidata and Language Models. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3557)*. 21–37.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[7] Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. 2017. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. In *ISWC*. 260–277.

[8] Viet-Phi Huynh, Yoan Chabot, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. 2022. From Heuristics to Language Models: A Journey Through the Universe of Semantic Table Interpretation with DAGOBAH. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3320)*. 45–58.

[9] Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Frédéric Deuzé, Thomas Labbé, Pierre Monnin, and Raphaël Troncy. 2021. DAGOBAH: Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3103)*. 19–31.

[10] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, and Kavitha Srinivas. 2020. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In *ESWC*, Vol. 12123. 514–530.

[11] Azanzi Jiomekong, Uriel Melie, Hippolyte Tapamo, and Gaoussou Camara. 2023. Semantic Annotation of TSOTSATable Dataset. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3557)*. 15–20.

[12] Keti Korini, Ralph Peeters, and Christian Bizer. 2022. SOTAB: The WDC Schema.org Table Annotation Benchmark. In *SemTab@ISWC*, Vol. 3320. CEUR-WS.org, 14–19.

[13] Oliver Lehmberg and Christian Bizer. 2016. Web table column categorisation and profiling. In *WebDB*. 4.

[14] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2023. Table-GPT: Table-tuned GPT for Diverse Table Tasks. *CoRR* abs/2310.09263 (2023).

[15] Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In *ESWC*. 182–185.

[16] Shervin Mehryar and Remzi Celebi. 2023. Semantic Annotation of Tabular Data for Machine-to-Machine Interoperability via Neuro-Symbolic Anchoring. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3557)*. 61–71.

[17] Phuc Nguyen, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. 2019. MTab: Matching Tabular Data to Knowledge Graph using Probability Models. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 2553)*. 7–14.

[18] Vishvapalsinhji Ramsinh Parmar and Alsayed Algergawy. 2023. DREIFLUSS: A Minimalist Approach for Table Matching. In *SemTab@ISWC (CEUR Workshop Proceedings, Vol. 3557)*. 50–60.

[19] Dominique Ritze, Oliver Lehmberg, and Christian Bizer. 2015. Matching HTML Tables to DBpedia. In *WIMS*. 10:1–10:6.

[20] Yalin Wang and Jianying Hu. 2002. A machine learning based approach for table detection on the web. In *WWW*. 242–250.

[21] Richard Zanibbi, Dorothea Blostein, and James R. Cordy. 2004. A survey of table recognition. *Int. J. Document Anal. Recognit.* (2004), 1–16.

[22] Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024. TableLlama: Towards Open Large Generalist Models for Tables. In *NAACL*. 6024–6044.