arXiv:2602.01276v1 [cs.AI] 1 Feb 2026

大语言模型驱动的企业知识图谱本体构建

Abdulsobur Oyewale*Liber AI Research* 伦敦，英国
abdulsobur@liberai.org

Tommaso Soru*Liber AI Research*
伦敦，英国
tom@liberai.org

摘要——**企业知识图谱已成为统一异构数据并落实语义治理的关键。然而，其底层本体的构建仍是一项资源密集的人工过程，高度依赖领域专业知识。本文提出 OntoEKG，一种由大语言模型驱动的流水线，旨在加速从非结构化企业数据生成领域特定本体。我们的方法将建模任务分解为两个不同阶段：抽取模块识别核心类与属性，蕴涵模块将这些元素逻辑地组织为层级结构，再将其序列化为标准资源描述框架。针对端到端本体构建缺乏全面基准测试的显著问题，我们采用了一个新的评估数据集，该数据集源自数据、金融和物流领域的文档。实验结果既凸显了该方法的潜力，也揭示了其挑战：在数据域中取得了 0.724 的模糊匹配 F1 分数，同时暴露出范围定义与层级推理方面的局限。**

索引词——**人工智能、大语言模型、语义模型、本体构建、资源描述框架、知识图谱**

I. 引言

过去十年间，企业日益采用语义技术和资源描述框架（RDF），以统一异构数据源、确保语义一致，并支持跨业务领域的互操作分析。这一转变反映出人们日益认识到，企业数据资产需要明确的语义，才能大规模支持治理、数据血缘和下游智能。本体在这些生态系统中发挥着核心作用，它们通过捕捉概念结构、约束词汇表，为企业知识图谱（EKG）提供基础。

与此同时，神经网络模型，尤其是近来兴起的大语言模型（LLM），已开始改变数据工程和集成工作流。它们能够从非结构化内容中提取语义、建议模式结构、执行问答并对齐业务术语，为知识图谱（KG）构建带来了强大的新机遇 [3], [12]。将大语言模型嵌入转换流水线后，有望加速本体工程任务；过去，这些任务依赖大量人工解读和领域专业知识。

然而，企业中的本体构建在很大程度上仍是人工的、迭代的且资源密集的。域利益相关者、数据架构师和语义工程师必须反复

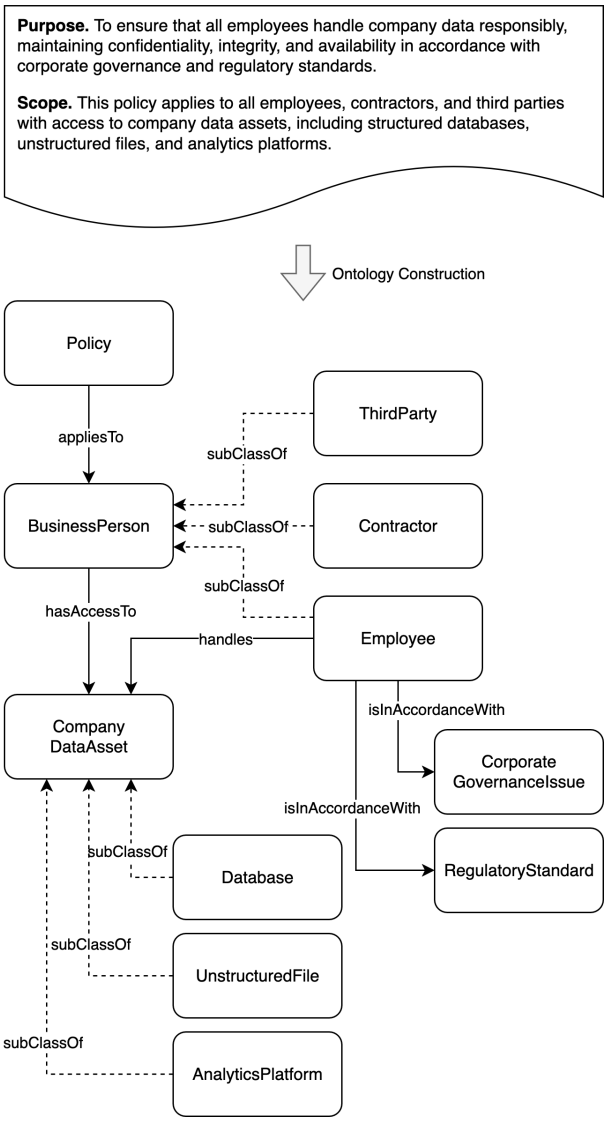

图 1. 本体构建任务示例。

协商概念边界、对齐模式，并记录建模决策。这推动了面向本体设计的人工智能助手的开发——该助手能够与人类专家协作，同时保持严谨性、透明度，

以及治理合规。在本文中，我们探讨大语言模型驱动的工作流如何支持专门用于企业知识图谱的本体创建与演化，在保留语义质量的同时减少建模阻力。

本文组织结构如下：第二节介绍相关工作，第三节描述本文方法，第四节讨论实验结果，最后进行总结。

II. 相关工作

企业本体提取已成为一项研究课题，历时二十余年。该领域最早的研究之一题为从企业内部网半自动获取本体的方法，提出了一套全面的架构和方法，利用企业内部网文本和半结构化资源自动提取并完善领域本体，从而减少对纯人工本体工程的依赖 [6]。

近来，研究者提出了 EOAC-LLM，这是一种基于大语言模型的五步方法，旨在实现领域事件本体的自动生成。与我们不同，作者关注的是一种用于语义时间关系的多维聚合方法 [9]。

论文使用大语言模型生成本体介绍了两种提示技术：无记忆逐个能力问题和 Ontogenia。与我们的方法不同，这两种技术强调多维评估，包括结构标准和专家评估 [8]。

此外，近期一篇论文，从人类专家到机器：一种由大语言模型支持的本体与知识图谱构建方法探讨了如何通过提出能力问题、开发 T-Box、进行知识图谱填充和评估，在极少需要人类专家参与的情况下，借助开源大语言模型半自动地构建知识图谱。该研究聚焦于完整的知识图谱构建，并采用大语言模型评审器自动评估检索增强生成（RAG）生成的答案和提取的概念 [7]。

论文 利用大语言模型实现自动化本体提取与知识图谱生成通过交互式用户界面运用大语言模型。该论文采用迭代式思维链算法，让用户能够根据自己的偏好，反复完善并确认本体 [1]。

使用大语言模型探索本体开发研究了大语言模型根据本体需求、运用各种提示技术生成 OWL 本体的能力，并探讨了如何在多个最先进模型之间比较这些提示技术 [11]。

III. 方法

本工作围绕两项主要贡献展开：

- 一条专门针对企业数据的从文本构建本体流水线；
- 呼吁开发一个用于评估从文本构建本体的综合基准。

本体构建任务的示例如图1所示。以一段关于网络安全行业企业的输入文本为起点构建基于RDF的本体。虚线边表示层次结构，实线边表示类之间的关系，并指定域和值域。

A. 形式化

我们的本体构建任务可形式化如下。给定输入文本 T，我们从语料库中推断类集合 C_T 和属性集合 P_T 。每个类 $c \in C_T$都关联一个标签和一个描述。每个属性$p \in P_T$ 都关联一个标签、一个定义域类和一个值域类，使得 $dom(p) \in C$, $rng(p) \in C$。类也可以存在于层级结构中，其中 $c_1 \subseteq c_2$ 表示属于类 c_1 的所有元素也都属于类 c_2。

用资源描述框架（RDF）的术语来说，每个类 c都是 owl:Class 的实例，每个属性 p 都是 owl:ObjectProperty 的实例。在 OntoEKG 中，数据类型会像*Schema.org*[1]中那样被具体化为各自独立的类。

B. 流水线

图2展示了 OntoEKG流水线的结构。

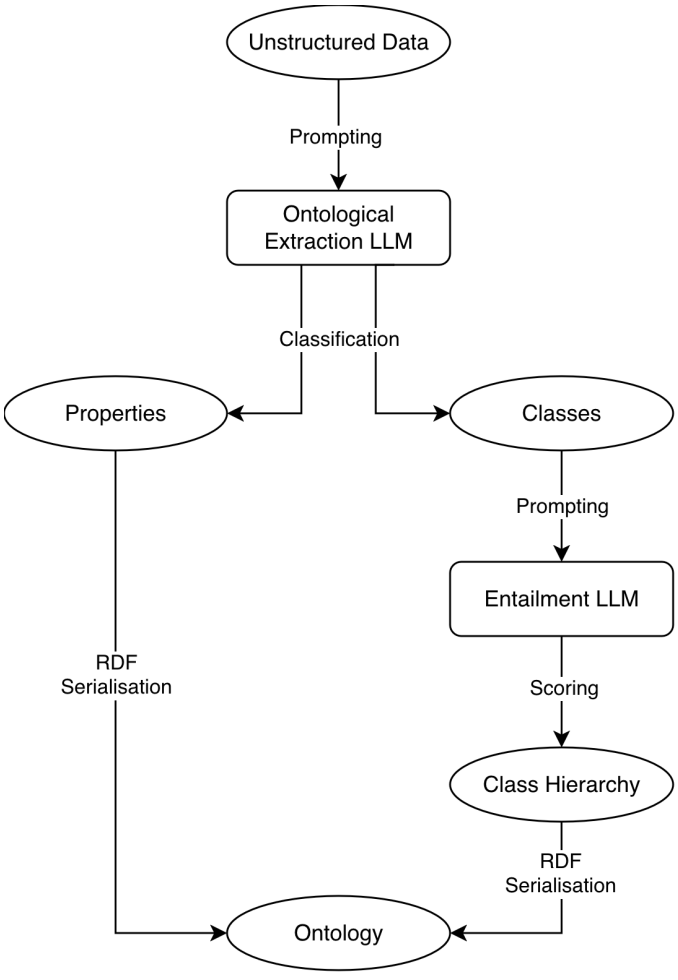

图2. 完整的 OntoEKG流水线。

[1]https://schema.org/DataType

该流水线将非结构化企业文本转换为适用于知识图谱的结构化本体。它采用两步大语言模型流程：第一步提取类和属性，第二步推理这些类之间的层级关系。以下是每个步骤的详细工作流程：

1) **数据摄取**。我们的流水线从将非结构化数据输入系统开始。为确保输出符合预期的结构，我们使用 Pydantic 定义严格的数据模型。这会促使大语言模型输出有效的 JSON，其中包含类、属性、描述、域和值域等特定元数据。

2) **本体元素提取**。本体提取大语言模型会处理原始非结构化数据。流水线使用专门的系统提示词来识别两个核心元素：类（例如“员工”或“车辆”等实体类型）和属性（例如“运营”或“拥有访问权限”等关系）。我们精心设计提示词，使其严格聚焦于提供给它的模式。

3)**基于蕴含的层级结构构建**。对于本体元素提取模块输出的类列表，此步骤会将这些类整理为逻辑分类体系。蕴涵大语言模型会迭代分析提取出的类，以确定继承关系（例如，验证“苹果”是否为“水果”的子类）。它会基于类描述进行逻辑推理，以验证这些关系。

4) **RDF序列化**。最后，系统将提取的属性与构建的层级结构合并为一个形式化图。系统使用 rdflib 库，将数据转换为结构化的资源描述框架三元组（即使用 owl:Class 和 owl:ObjectProperty），并将结果保存到 Turtle文件中，从而创建出机器可读本体。

IV. 评估

A. 基准测试

如第 III 节所述，本文的一个主要目标是呼吁研究界采取行动。具体而言，据我们所知，目前缺少用于评估从文本构建本体的综合基准测试。根据我们的调查，先前的工作要么未能完整解决这项任务，要么未达到所需的质量标准。

OntoURL 是一个综合基准，用于从理解、推理和学习三个维度评估大语言模型处理本体的能力。学习类别中的第二项和第三项任务分别针对类层次结构构建和属性关系构建 [14]。然而，这些任务要求输入为半结构化形式，其中包含类和属性列表。因此，OntoEKG 及其他从纯非结构化数据中提取知识的方法无法使用该基准进行评估。

Text2KGBench 和 OSKGC 等基准测试强调在现有框架内进行实例级提取，将本体视为约束，而非最终产物 [10], [13]。

LLMs4OL 计划举办一项面向基于大语言模型的本体学习方法的挑战，分为四项任务。任务 A（Text2Onto）旨在从文本中提取术语和类型；此处，预期输出是本体的潜在术语列表，而任务 C 则针对分类体系发现 [4]。遗憾的是，这些任务并非按顺序排列，因此无法在其上评估完整的本体构建流水线。此外，任务 C 将类术语与个体混在一起，未区分 T-Box 与 A-Box。

基于上述发现，我们选择创建自己的数据集，其中包含三个用例——来自数据、金融和物流领域的企业内部政策文本摘录。源代码和数据可在 Github 上的 theOntoEKG 仓库中找到[2]。

B. 实验

我们在 Google Colab[3] 提供的云端机器上运行实验。对于本体抽取步骤，我们选择了 Google Gemini 3 Flash（预览版） [5]；对于蕴含步骤，我们选择了 Anthropic Claude 4.5 Opus [2]。

表 I 本体构建性能，精确匹配。

用例	精确率	召回率	F1
Data	0.083	0.133	0.102
金融	0.000	0.000	0.000
物流	0.040	0.062	0.048

表 II 本体构建性能，模糊匹配。

用例	精确率	召回率	F1
Data	0.656	0.807	0.724
金融	0.095	0.166	0.121
物流	0.366	0.523	0.431

结果见表 I 和表 II。表 I 展示了仅考虑完全匹配的三元组时 OntoEKG 的性能。表 II 中，我们采用基于嵌入的模糊匹配，将预测的三元组与其黄金标准对齐；对于三个用例，我们设置的相似度阈值为 (0.94，0.94，0.95)。最佳表现出现在数据用例中，在模糊匹配设置下，F1分数为 0.724。金融用例最具挑战性，F1 分数仅为 0.121 ；这可能是由于对输入文本的理解不同，尤其是对哪些术语应纳入本体、哪些术语与上下文无关的判断存在差异。

图 3 展示了一个生成示例，该示例由图 1 中的文本处理而来。我们可以看到图谱中出现了两个问题：“Policy”和“GovernanceStandard”被重新声明为彼此的子类，这意味着二者等价；

[2]https://github.com/LiberAI/OntoEKG [3] https://colab.research.google.com

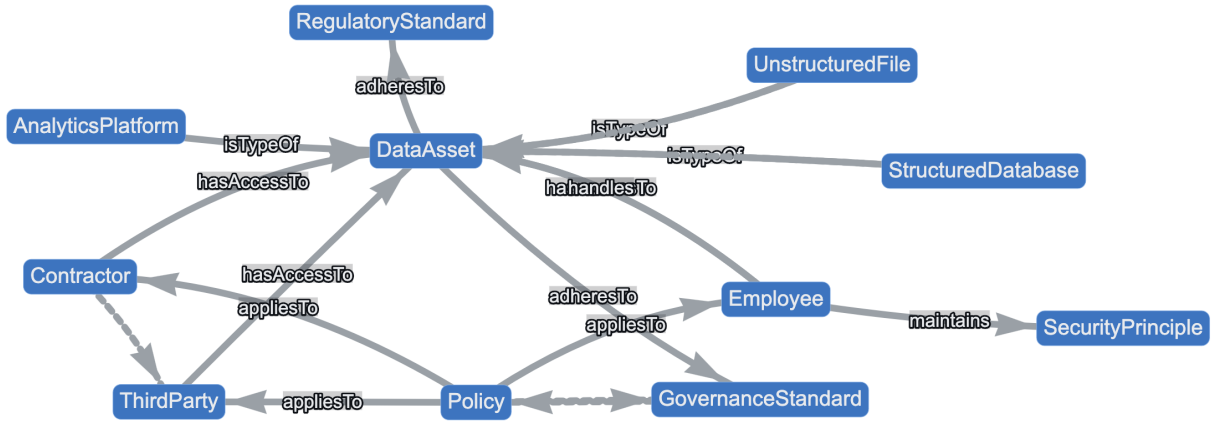

图 3. 为数据用例生成的本体。

此外，还引入了一个“isTypeOf”属性，但从 RDF 的角度看，它究竟对应 rdf:subClassOf 还是 rdf:type，仍存在歧义。

我们使用不同的大语言模型测试了蕴涵任务，包括 Gemini 2.5 Flash、2.5 Pro、3 Flash（预览版）和 Claude 4.5 Sonnet。由于 Gemini 2.5 Pro 的效率不够理想，我们将其排除。其他模型在开发阶段也未达到预期。这些初步结果证实，我们需要专门的基准，以便进行更恰当的评估。

尽管结果令人鼓舞，我们仍发现该方法存在若干局限：

- 对于大语言模型而言，自主确定模型的最优范围并不容易。为确保准确性，必须明确提示模型相关类和属性的边界。
- 有时，大语言模型倾向于提出个体而非类。同样，这是因为没有明确规定目标抽象层级。
- 在蕴含阶段，我们发现大语言模型混淆了层级关系的方向性，并采用了宽泛的包含关系定义，从而影响了RDF模型的逻辑一致性。

V. 结论

本文介绍了OntoEKG，这是一种由大语言模型驱动的、面向企业知识图谱的本体构建方法。初步结果表明，语义建模这类繁琐且资源密集型的任务可以借助自动化技术完成。我们还强调了建立一个用于从非结构化数据中进行本体构建的综合基准的必要性。

未来的工作将包括实现一种端到端方法，将文本转换为基于资源描述框架（RDF）的语义模型。我们将集成对命名个体的处理能力，并直接从文本中提取实体元数据，例如保留来源信息。此外，我们预计能够逐步构建企业本体具体做法是将现有的提议模型连同输入文本一起提供给 OntoEKG 本身，从而使模型在不同源文档之间保持一致。我们计划邀请研究社区共同开发一个全面的端到端本体构建基准。

参考文献

[1] Mohammad Sadeq Abolhasani和Rong Pan。利用大语言模型进行自动本体提取与知识图谱生成。*arXiv*预印本 *arXiv:2412.00608*，2024年。[2] Anthropic。Claude 4.5 opus。https://www.anthropic.com/claude，2025年。访问日期：2025-12-31。[3] Haonan Bian。大语言模型赋能的知识图谱构建：综述。*arXiv*预印本，2025年。arXiv:2510.20345。[4] H. Giglou、J. D’Souza、N. Mihindukulasooriya和S. Auer。LLMs4ol 2025概览：第二届大语言模型本体学习挑战赛。载于*Open Conference Proceedings*，第6卷，2025年。[5] Google。Gemini 3 flash。https://gemini.google.com/，2025年。访问日期：2025-12-31。[6] J. Kietz、A. Maedche和R. Volz。从企业内部网半自动获取本体的方法。载于*EKAW-2000 Workshop “Ontologies and Text”，Juan-Les-Pins*，法国，*2000*年*10*月，2000年。[7] V. Kommineni、B. K¨onig-Ries和S. Samuel。从人类专家到机器：一种由大语言模型支持的本体与知识图谱构建方法。*arXiv*预印本 *arXiv:2403.08345*，2024年。[8] A. Lippolis、M. Saeedizade、R. Keskis¨arkk¨a、S. Zuppiroli、M. Ceriani、A. Gangemi、E. Blomqvist和A. Nuzzolese。使用大语言模型生成本体。载于欧洲语义网会议，第321–341页。Springer，2025年。[9] Zhenhai Lu、Jungang Yang、Chang Guo、Chao Zhang和Yao Li。EOAC-LLM：由大语言模型驱动的事件本体自动构建系统。可见于*SSRN 5775146*，2025年。[10] N. Mihindukulasooriya、S. Tiwari、C. Enguix和K. Lata。Text2KGBench：一个用于从文本生成本体驱动知识图谱的基准。载于国际语义网会议，第247–265页。Springer，2023年。[11] Mohammad Javad Saeedizade和Eva Blomqvist。使用大语言模型探索本体开发。载于欧洲语义网会议，第143–161页。Springer，2024年。[12] Tommaso Soru、Edgard Marx、Diego Moussallem、Gustavo Publio、Andr´e Valdestilhas、Diego Esteves和Ciro Baron Neto。SPARQL作为外语。载于*SEMANTiCS*，2017年。[13] Dali Wang和Mizuho Iwaihara。OSKGC：一个用于从文本进行基于本体模式的知识图谱构建的基准。载于*CEUR WorkshopProceedings*，第4041卷。CEUR-WS，2025年。[14] Xiao Zhang、Huiyuan Lai、Qianru Meng和Johan Bos。Ontourl：一个用于评估大语言模型在符号化本体理解、推理和学习方面能力的基准。*arXiv*预印本 *arXiv:2505.11031*，2025年。