arXiv:2304.05973v1 [cs.IR] 12 Apr 2023

HiPrompt：通过面向层级的提示实现少样本生物医学知识融合

Jiaying Lu 埃默里大学，美国
jiaying.lu@emory.edu

Jiaming Shen 谷歌研究院，美国
jmshen@google.com

Bo Xiong 斯图加特大学，德国
bo.xiong@ipvs.uni-stuttgart.de

Wenjing Ma 埃默里大学，美国
wenjing.ma@emory.edu

Steffen Staab德国斯图加特大学 英国南安普顿大学
steffen.staab@ipvs.uni-stuttgart.de

Carl Yang 埃默里大学，美国
j.carlyang@emory.edu

摘要

医疗决策过程可以通过全面的生物医学知识库得到增强，而这需要通过统一索引系统融合从不同来源构建的知识图谱。该索引系统通常以层级结构组织生物医学术语，从而为对齐实体提供细粒度的粒度。为应对生物医学知识融合（BKF）任务中监督信号稀缺的挑战，研究人员提出了各种无监督方法。然而，这些方法严重依赖临时的词汇与结构匹配算法，无法捕捉生物医学实体和术语所传达的丰富语义。近年来，神经嵌入模型已被证明在语义丰富的任务中表现有效，但它们依赖充足的标注数据才能得到充分训练。为了弥合标注数据稀缺的 BKF与神经嵌入模型之间的差距，我们提出了 HiPrompt，这是一种高效利用监督信号的知识融合框架，通过面向层次结构的提示激发大语言模型的少样本推理能力。在所收集的 KG-Hi-BKF 基准数据集上的实证结果证明了 HiPrompt 的有效性。

CCS 概念

• **应用计算** → **医疗保健信息系统**； •**信息系统** →检索模型和排序。

关键词

生物医学知识融合、少样本提示、面向资源受限领域的大语言模型、检索与重排序

ACM 引用格式：

Jiaying Lu、Jiaming Shen、Bo Xiong、Wenjing Ma、Steffen Staab和 Carl Yang。2023年。HiPrompt：通过面向层级结构的提示实现少样本生物医学知识融合。载于 第*46*届国际 *ACM SIGIR*信息检索研究与发展大会论文集（*SIGIR '23*），*2023*年7月*23*日至*27*日，台北，台湾。ACM，纽约，纽约州，美国，5页。https://doi.org/10.1145/3539618.3591997

允许为个人使用制作本作品部分或全部内容的数字或纸质副本，或】【。允许用于课堂教学，且无需支付费用，前提是不得为牟利或商业利益而制作或分发副本，并且副本首页须附有本声明及完整引文。本作品中第三方组件的版权必须得到尊重。对于所有其他用途，请联系所有者/作者。*SIGIR '23, 2023*年7月*23*日至*27*日，台北，台湾©2023年 版权所有者/作者所有。ACM ISBN 978-1-4503-9408-6/23/07。https://doi.org/10.1145/3539618.3591997

1 引言

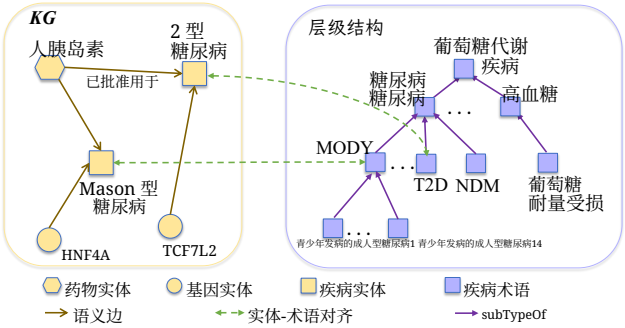


图 1：生物医学知识融合的一个玩具示例，用于查找知识图谱与层级结构之间的实体-术语对齐。左侧：包含生物医学实体的知识图谱。右侧：包含生物医学术语的层级结构。

在生物医学领域，存在大量来自临床实践指南、医疗记录和出版物的知识，这些知识由不同的研究实验室和医疗机构积累而成 [8, 34, 36]。近年来，知识图谱（KGs）已成为一种颇具吸引力的技术，可用于高效地表示、组织和传播知识。生物医学知识图谱存储了生物医学实体的属性及其关系。研究人员不断努力，手动整理生物医学知识图谱，因此涌现出许多特定领域和面向应用的知识图谱。然而，这些经过充分标注的生物医学知识图谱分散在各种数据格式中，妨碍了它们开箱即用。

将来自多个来源的知识图谱融合为准确且全面的知识库，可以极大地支持临床决策 [13, 28]。一种常见做法是将知识图谱中的实体与标准层次索引系统（即生物医学层级结构）对齐 [4, 14, 30, 44]。层级结构能够以更细粒度对实体进行更精确的对齐和分析，从而有利于许多下游任务 [21, 31, 32, 40, 43]。此外，生物医学层级结构会通过定期升级得到良好维护，以纳入新出现的生物医学术语，因此能够支持与多个知识图谱进行可扩展的集成。在这项工作中，我们研究生物医学知识融合（生物医学知识融合，BKF）问题，旨在将生物医学知识图谱中的实体与生物医学层级结构中的术语对齐。图 1 展示了一个 BKF 任务的简单示例。BKF 任务具有以下特点，因此颇具挑战性。首先，不同资源使用的命名词汇并不一致，因

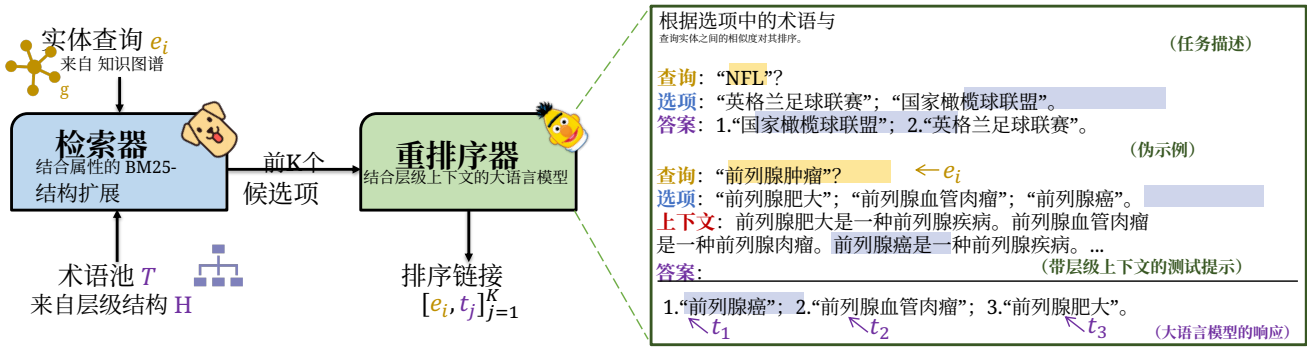


图 2：HiPrompt 框架概览，放大展示基于大语言模型的重排序器 r.

它们由不同的专家团队独立开发。其次，与包含大量已标注实体对作为训练样本的现有知识图谱实体对齐问题 [38, 47]不同，生物医学知识整合缺乏监督信号。第三，知识图谱的拓扑结构与层级结构截然不同：知识图谱是一般图，而层级结构是有向无环图。**现有研究**。生物医学知识融合领域的早期研究主要依赖生物医学术语库来规范词语，并通过词汇匹配建立知识图谱与层级结构之间的对齐 [13, 24, 28, 36]。随后，研究人员探索将一阶逻辑 [15]，概率对齐 [37],或非字面字符串比较 [11] 与词汇匹配相结合，以实现无监督的生物医学知识融合。然而，这些方法无法捕捉实体和术语所传达的丰富语义（例如，同义词、定义、类型），而这些语义对于处理来自多个来源的不一致命名规范至关重要。另一类工作利用神经嵌入模型 [9, 19, 20, 38, 46]，通过语义属性、结构特征和对齐监督，将实体表示为稠密向量。当有足够的训练样本时，这些模型的表现优于无监督模型。然而，生物医学知识融合问题中监督信号的匮乏，导致这些依赖大量数据的神经模型拟合不足。此外，现有方法都没有明确利用生物医学层级结构中术语的层级结构。

本文工作。为应对上述挑战，我们提出**HiPrompt**，一种通过**层**级结构导向**提示**进行少样本生物医学知识融合（BKF）的框架。HiPrompt 利用大语言模型（LLM）生成并提出层级结构中的术语，以便与知识图谱（KG）中的实体进行对齐。其核心洞见是，大语言模型 [7, 10, 39, 48] 可以通过无需梯度的“基于提示的学习” [35, 41]，快速适应未见任务，从而摆脱对特定任务监督信息的依赖。HiPrompt 针对 BKF 任务采用基于提示的学习，并结合精心设计的任务描述，以及由少样本示例生成的少量演示样例。这模拟了人类通过以往经验学习并将所学泛化至新情境，从而完成新任务的过程。此外，我们将层级结构上下文加入提示中，以进一步提升 HiPrompt 的性能。为评估所提出的 HiPrompt 的性能，我们创建了*KG-Hi-BKF*，这是一个新的 BKF 基准，包含两个来自两个生物医学知识图谱 [6, 50] 和一个疾病层级结构 [30] 的数据集，并经过人工核验。实证结果表明，我们的 HiPrompt 框架效果显著，大幅优于传统的无监督词汇匹配模型和神经语义嵌入模型。

2 生物医学知识融合

2.1 问题定义

BKF 旨在将现有的专门生物医学知识图谱整合到一个统一的生物医学索引系统中，并以层级结构表示。我们将生物医学知识图谱和层级结构定义如下：生物医学知识图谱是一个多关系图 $\mathrm{G} = (E, R, RT)$，其中E，R，RT分别是一组不同类型的实体、一组关系名称，以及关系三元组的集合。生物医学层级结构是一个有向无环图（DAG）$\mathrm{H} = (T, TP)$，其中 T是一组术语，$TP \in T \times T$是一组上下位术语对。知识图谱与层级结构之间的拓扑差异，使我们的 BKF 任务有别于其他相关任务（例如，实体对齐、知识图谱整合）。此外，实体 E和术语 T都包含丰富的相关语义属性（例如，定义、同义词）。最后，我们将任务定义如下：

定义 2.1（生物医学知识融合）。给定一个生物医学知识图谱 G、一个生物医学层级结构 H、一组预对齐实体-术语对 $[e_a, t_a]_{a=1}^{M}$，以及一组未对齐实体 $[e_1, e_2, \cdots e_N] \in \mathrm{G}$。目标是将每个未对齐实体链接到层级结构 $LK = \{(e_i, t_j) | e_i \in \mathrm{G}\ t_j \in \mathrm{H}\}$，使得 t_j是知识图谱中实体 e_i对应的层级结构中最具体的术语。在我们的工作中，我们重点关注小样本设置，其中样本量 M非常小，以反映生物医学领域普遍存在的标注数据稀缺问题。

2.2 HiPrompt 的技术细节

图 2 展示了我们提出的 HiPrompt 框架的整体架构。为了在训练样本有限的情况下完成生物医学知识融合任务，我们的核心思路是通过面向层级结构的提示来利用大语言模型。然而，大语言模型无法处理过长的输入提示（例如，GPT-3 最多只支持 4096 个词元），而这些提示包含所有候选术语及其层级上下文。一种可行的替代方案是针对给定的查询实体，穷举检查每个候选术语，但推理成本将会非常高 [23]。因此，我们提出使用检索与重排序 [12, 22, 42] 方法来解决上述挑战。

检索模块。检索器为粗粒度候选筛选提供了一种高效方案，从而降低了 HiPrompt 的整体推理成本。给定来自知识图谱 G 的一个实体查询 e_i 以及来自层级结构 H 的所有候选术语 T，检索器会生成一个粗略排序的候选列表（t_{r1}，t_{r2}，$\cdots$ t_{rK}），以避免基于大语言模型的重排序器进行不必要的计算。HiPrompt 框架具有灵活性，可使用任意无监督排序函数（例如，词频-逆文档频率 [27]，潜在狄利克雷分配 [3]）来生成排序后的

设置	模型	SDKG-DzHi Hits@1	Hits@3	nDCG@1	nDCG@3	WuP	MRR	repoDB-DzHi Hits@1	Hits@3	nDCG@1	nDCG@3	WuP	MRR
零样本	编辑距离	65.51	70.39	68.08	50.82	85.53	68.69	68.69	71.37	71.71	54.15	85.21	70.71
	BM25	73.07	87.40	77.56	63.01	91.97	81.06	59.38	74.75	70.33	64.51	90.71	68.84
	LogMap	75.75	79.06	76.97	54.82	85.06	77.38	86.60	87.73	87.38	60.79	91.68	87.09
	PARIS	22.68	22.68	23.15	16.13	43.85	22.68	6.35	6.35	6.42	4.44	32.28	6.35
	AML	OOM	OOM	OOM	OOM	OOM	OOM	78.00	78.56	78.67	54.90	86.02	78.26
	SapBERT	69.61	87.24	76.38	63.86	93.78	78.97	75.04	90.69	81.24	73.51	94.25	83.61
	SelfKG	57.95	69.45	58.98	47.29	74.25	64.70	72.78	81.10	75.95	63.78	88.41	77.71
	HiPrompt	**90.79**	**93.08**	**91.57**	**77.00**	**96.74**	**92.13**	**88.01**	**91.26**	**90.70**	**82.85**	**97.06**	**90.64**
单次提示	SapBERT	69.56	87.22	76.34	63.84	93.29	78.93	75.00	90.68	81.21	73.51	94.13	83.59
	MTransE	0.0	0.16	0.0	0.05	35.09	0.16	0.0	0.28	0.14	0.27	28.89	0.37
	HiPrompt	**92.11**	**95.11**	**93.53**	**77.63**	**97.25**	**93.91**	**88.28**	**91.53**	**90.61**	**81.31**	**96.39**	**90.28**

表 1：主要实验结果（百分比）。

列表。实践中，我们选择无监督 BM25 [26] 作为排序函数。由于实体和概念包含丰富的属性信息和结构信息，我们进一步利用这两类信息来扩展 [2] 查询实体和候选术语。

重排序模块。给定查询实体 e_i 和粗排后的候选列表 $(t'1, t'2, \cdots t'_K)$，我们通过无梯度的基于提示的学习，请求大语言模型将该列表重新排序为 $(t_1, t_2, \cdots t_K)$，其中 t_1 是 e_i 中最具体的术语。图 2 展示了输入提示及重排序器响应的示例。输入提示由以下部分组成：(1) 精心编写的文本任务描述；(2) 少样本示例中的说明性演示；(3) 根据查询实体和粗排列表构造的测试提示。基于大语言模型的重排序器本质上通过估计条件概率来处理生物医学知识融合任务：$P_{LLM}(w_1, w_2, \ldots, w_n|prompt)$，其中 $(w_1, \ldots, w_n)$ 是长度可变的输出词序列。所需的重排序列表可通过简单的映射函数从输出序列转换得到：$(t_1, t_2, \cdots t_K) = f(w_1, w_2, \ldots, w_n)$。

对于示例模板，我们使用查询实体构成问题字符串“Query: $\{e_i\}$”，使用粗粒度候选列表构成选项字符串“Choices: $\{t'1; t'2; \ldots t'_K\}$”，并使用真实标签构成答案字符串“Answer: $\{t_1; t_2; \ldots, t_K\}$”。虽然零样本设置中不存在这样的真实标签样本，但我们提出了伪示例技术，采用域外实体—术语对来展示规范的格式。真实示例和伪示例对于以一致格式 [16, 29] 生成输出序列都至关重要。对于测试提示，我们使用与示例相同的模板，同时将答案字符串保留为“Answer:”，让大语言模型预测接下来的内容。为了进一步引导大语言模型遵循候选术语的层级约束和依赖关系，我们提出了带有层级上下文的新型测试提示，其中每个候选术语的上位词都包含在上下文字符串中。更具体地说，我们遍历生物医学层级结构 T，以定位候选术语 t'_i 的上位词术语 $t'_{i,p_1}, \cdots t'_{i,p_j}$。因此，上下文字符串构造为“Contexts: $\{t'1$ isA $t'1_{,p}; \ldots; t'_K$isA $t'_{K,p}\}$”。

3 实验

基准数据集。 我们使用以下数据来源来创建我们的 KG-Hi-BKF 基准[1]：（1）SDKG [50]：以疾病为中心的涵盖五种癌症和六种非癌症疾病的知识图谱。（2）repoDB [6]：我们采用其原始三元组，并通过查询 DrugBank [44] 和UMLS元词库 [4]生成实体属性。（3）DzHi [30]：一种源自广泛使用的疾病本体 [30]的层级结构，其深度为 13。我们首先使用资源自身已有的映射，这会导致两个知识库之间形成多对多链接。随后，我们进一步人工验证多对多链接的正确性，并将数据集整理至正确阶段。表2展示了所创建基准的统计数据。可以看出，这些链接遵循一对一假设 [38], , 且带标签的实体-术语对的规模非常小。

数据集	来源	#疾病	#实体	#链接
SDKG-DzHi	SDKG	841	19,416	635
	DzHi	11,159	11,159	635
repoDB-DzHi	repoDB	2,074	3,646	709
	DzHi	11,159	11,159	709

表2：KG-Hi-BKF 基准的统计信息。

对比模型。我们将 HiPrompt 与以下两组基线模型进行比较：(a) 非神经传统模型：(a.1) **编辑距离** [25] 通过名称的编辑距离来量化实体与术语之间的距离。(a.2) **BM25** [26] 根据每篇文档中出现的查询词元对文档进行排序。(a.3) **LogMap** [15]通过逻辑约束和语义特征匹配实体与术语。(a.4) **PARIS** [37] 提供一种由免参数调优概率模型驱动的开箱即用融合工具。(a.5) **AML** [11] 基于非字面字符串比较算法。是一种基于概率估计的概率匹配系统。(b)神经嵌入模型：(b.1)**SapBERT** [17] 通过 Transformer 学习对齐同义生物医学实体。(b.2) **MTransE** [9] 通过轴校准和线性变换，将平移式知识图谱嵌入方法 TransE [5] 扩展到多语言系统实体对齐。(b.3) **SelfKG** [18] 设计了一种自负采样策略，在没有标注正样本对时，将采样得到的负样本对彼此推远。

定量评估。 我们主要关注零样本和单样本设置，并利用剩余的带标签样本作为测试集来报告定量结果。采用了若干严格的和宽松的 评估指标。对于仅认可完全正确预测的严格指标，我们采用**Hits@k** 和平均倒数排名

[1]KG-Hi-BKF 基准可在 https://doi.org/10.6084/m9.figshare.21950282 获取。

（**平均倒数排名**）。对于也会奖励近似命中的宽松评估指标，我们采用**nDCG@k**，并使用指数衰减 [1] 和基于层级结构的术语相关性分数**WuP** [45]。所有对比基线均使用其推荐的超参数运行。对于所有非神经传统模型，由于它们属于无监督方法，我们仅报告零样本结果。对于神经嵌入方法，我们报告使用已发布模型权重（SapBERT）或进行自监督训练（SelfKG）得到的零样本结果，同时报告通过在一个示范训练样本上微调这些模型（SapBERT、MTransE）得到的单次提示结果。对于 HiPrompt，我们使用 GPT-3 [7] 作为重排序器的大语言模型，并将其温度超参数设为 0，以降低生成结果的随机性。由于初步探索表明不同模板对模型性能没有显著影响，因此使用单个提示模板已足够。由于训练数据有限，我们不采用自动提示生成技术 [33, 49]。**主要结果**。表 1 展示了零样本和单次提示设置下的定量结果。在这两种设置下，HiPrompt 在所有评估指标上均大幅优于其他所有方法，证明了所提出的面向层级结构的提示方法是有效的。在零样本设置下，非神经无监督基线 LogMap 的表现位居第二。所有受测模型均能成功生成预测，但 AML 在 SDKG-DzHi 数据集上出现了内存不足（OOM）错误。PARIS 在零样本设置下表现最差，因为它无法为每个查询实体预测对齐术语，而是基于自身的临时阈值生成对齐结果。MTransE 在单次提示设置下表现最差，因为仅使用一个训练样本会导致欠拟合。比较相同模型（SapBERT、HiPrompt）在零样本和单次提示设置下的结果，我们发现性能差异微乎其微，这表明，有效激发自适应推理能力是解决监督数据稀缺的生物医学知识融合问题的关键因素之一。

扩展。	*SDKG-DzHi*			*repoDB-DzHi*		
	Hits@5	Hits@10	Hits@20	Hits@5	Hits@10	Hits@20
Name	88.66	89.61	90.55	85.05	88.72	90.27
+Atr.	94.96	96.85	98.11	89.00	92.52	95.20
+策略	90.08	90.71	91.81	88.15	90.27	92.24
+属性+策略	**96.85**	**97.64**	**98.74**	**91.11**	**93.65**	**95.63**

表3：采用不同扩展策略的检索器。

LLMs	SDKG-DzTaxo Hits@1	Hits@3	MRR	repoDB-DzTaxo Hits@1	Hits@3	MRR
	单次提示（不含 Hi. Context 的提示）					
GPT-3	**91.80**	**94.32**	**93.45**	**87.85**	**91.24**	**89.92**
GPT-JT	75.08	86.44	81.80	58.33	69.77	66.42
OPT-6.7B	68.93	80.44	76.38	60.73	73.59	69.33
	单次提示 (提示词含 Hi. Context)					
GPT-3	**92.11**	**95.11**	**93.91**	**88.28**	**91.53**	**90.28**
GPT-JT	80.76	93.69	87.45	69.07	82.91	77.24
OPT-6.7B	72.40	84.86	79.64	63.70	77.68	72.41

表 4：使用不同大语言模型和提示词的重排序器。

消融研究。我们进一步开展消融研究，以评估层级导向技术的影响。表3比较了 HiPrompt 检索模块采用的不同扩展策略。可以看出，如果扩展知识图谱实体和层级术语】【。

同时具备属性特征和结构特征（“+属性。+结构。”变体）时，检索器可以取得最佳的 Hits@k 性能。表 4 比较了不同大语言模型以及 HiPrompt 重排序模块所使用的不同提示。在所考察的大语言模型中，拥有 1750 亿参数的 GPT-3凭借其庞大的参数空间，超过了拥有 6B参数的GPT-JT [39]和拥有 6.7B参数的 OPT-6.7B [48]。将所提出的层级上下文添加到仅包含名称的提示中后，每个大语言模型在所有指标上都取得了更好的性能，从而证明了显式层级导向信息的重要性。我们还观察到，GPT-JT 和 OPT-6.7B 的提升幅度比 GPT-3 更为显著，因为 GPT-3 可能已经编码了此类层级信息。

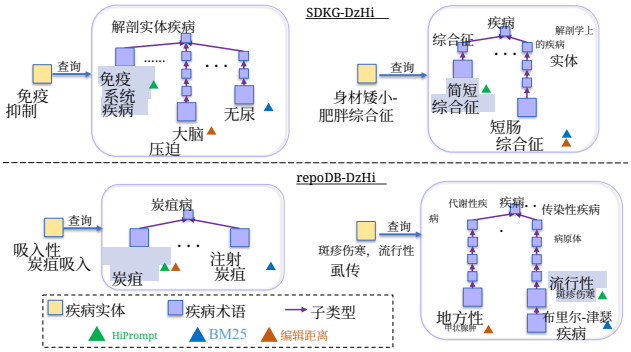


图 3：未标注数据上的案例研究。紫色高亮部分表示查询实对齐结果。

案例研究。图 3 展示了 BM25、Edit-Dist 和 HiPrompt 的融合结果。总体而言，HiPrompt 能为查询实体找到层级结构中最具体的术语，同时满足语义相似性和层级约束。例如，对于查询“免疫抑制”，HiPrompt 能识别出“免疫系统疾病”最为合适，而不是过于笼统的上位词“解剖实体疾病”，或过于具体的下位词，例如“免疫系统癌症”或“过敏性疾病”。另一方面，EditDist 只考虑词汇匹配，因此忽略了同一生物医学概念的不同命名惯例。BM25 也主要依赖词汇匹配，但在匹配时会纳入生物医学术语的名称、定义和同义词，因此在处理各种名称时表现更佳。然而，BM25 忽略了层级信息，导致对齐术语的粒度不恰当（例如，术语“流行性斑疹伤寒”对于查询实体“虱传流行性斑疹伤寒”而言过于宽泛）。

4 结论

本文研究如何在标注数据稀缺的情况下，自动将知识图谱融合到标准层级索引体系中。我们提出了新颖框架 HiPrompt，利用面向层次结构的提示激发大语言模型的少样本推理能力，并将其设计为一种高效利用监督信号的方法。在新收集的 KG-Hi-BKF 基准（包含两个数据集）上的性能比较证明了 HiPrompt 的有效性。生物医学知识融合未来值得探索的方向包括：(1) 探索自动生成层次感知提示的方法，以进一步减少人工干预；(2) 扩大生物医学知识融合的范围，使层级结构能够随着对齐实体动态增长。

致谢

本研究获得埃默里大学计算机科学系提供的内部经费和 GPU 服务器支持。Bo Xiong 获得了国际马克斯·普朗克智能系统研究生院（IMPRS-IS）的支持，并由欧盟“地平线 2020”研究与创新计划根据玛丽·斯克沃多夫斯卡-居里资助协议（编号：860801）提供资助；此外，还获得德国研究联合会（DFG）依据德国卓越战略提供的资助——EXC 2075-390740016（SimTech）。

参考文献

[1] KrisztianBalog和 Robert Neumayer。2012年。面向实体查询的分层目标类型识别。在 CIKM。 [2] BodoBillerbeck和Justin Zobel。2005年。文档扩展与查询扩展在即席检索中的比较。在《第十届澳大拉西亚文档计算研讨会论文集》。 [3] David M Blei、AndrewYNg和Michael IJordan。2003年。潜在狄利克雷分配。JMLR（2003年）。 [4] Olivier Bodenreider。2004年。统一医学语言系统（UMLS）：整合生物医学术语。《核酸研究》（2004年）。 [5] Antoine Bordes、Nicolas Usunier、Alberto García-Durán、JasonWeston 和 Oksana Yakhnenko。2013年。用于建模多关系数据的嵌入翻译。在 NeurIPS。[6] Adam SBrown 和 Chirag J Patel。2017年。用于药物再定位的标准数据库。《科学数据》（2017年）。[7] Tom Brown、BenjaminMann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell等。2020年。语言模型是少样本学习者。NeurIPS（2020年）。 [8] Payal Chandak、Kexin Huang 和 Marinka Zitnik。2022年。构建知识图谱以实现精准医学。bioRxiv（2022年）。 [9] Muhao Chen、YingtaoTian、Mohan Yang和 CarloZaniolo。2017年。用于跨语言知识对齐的多语言知识图谱嵌入。在 IJCAI。 [10] Xiangjue Dong、Jiaying Lu、JianlingWang和 James Caverlee。2023年。通过对比学习进行闭卷问题生成。在 EACL。 [11] Daniel Faria、Catia Pesquita、Emanuel Santos、Matteo Palmonari、IsabelF Cruz 和 Francisco M Couto。2013年。AgreementMakerLight 本体匹配系统。在 ODBASE。 [12] Michael Glass、Gaetano Rossiello、Md Faisal Mahbub Chowdhury、Ankita Naik、PengshanCai 和Alfio Gliozzo。2022年。Re2G：检索、重排序、生成。在 NAACL。 [13] DanielScott Himmelstein、Antoine Lizee、Christine Hessler、Leo Brueggeman、Sabrina L Chen、Dexter Hadley、Ari Green、Pouya Khankhanian 和SergioE Baranzini。2017年。生物医学知识的系统整合为药物再利用确定优先级。eLife（2017年）。 [14] ShuaiJiang、QihengQian、Tongtong Zhu、Wenting Zong、YunfeiShang、Tong Jin、Yuansheng Zhang、MingChen、Zishan Wu、YuanChu 等。2023年。细胞分类学：一个经过整理、具有多维表征的细胞类型资源库。《核酸研究》（2023年）。 [15] Ernesto Jiménez-Ruiz 和 BernardoCuenca Grau。2011年。LogMap：基于逻辑且可扩展的本体匹配。在 ISWC。 [16] Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo和Yusuke Iwasawa。2022年。大语言模型是零样本推理器。在 ICML 2022 知识检索与语言模型研讨会。 [17] FangyuLiu、Ehsan Shareghi、Zaiqiao Meng、MarcoBasaldella 和Nigel Collier。2021年。面向生物医学实体表示的自对齐预训练。在NAACL。 [18] Xiao Liu、Haoyun Hong、XinghaoWang、Zeyi Chen、Evgeny Kharlamov、YuxiaoDong 和 Jie Tang。2022年。SelfKG：知识图谱中的自监督实体对齐。在 TheWeb Conference。 [19] Jiaying Lu、XiangjueDong 和CarlYang。2023年。通过任务引导的图翻译进行弱监督概念图生成。IEEE TKDE（2023年）。 [20] JiayingLu和 Carl Yang。2022年。开放世界分类法与知识图谱协同学习。在 AKBC。 [21] Wenjing Ma、JiayingLu 和 HaoWu。2023 年。Cellcano：用于单细胞 ATAC-seq 数据的监督式细胞类型识别。《自然通讯》（2023年）。 [22] Yoshitomo Matsubara、ThuyVu 和 Alessandro Moschitti。2020年。基于Transformer 的高效答案选择重排序。在 SIGIR。[23] David Patterson、Joseph Gonzalez、Quoc Le、Chen Liang、Lluis-Miquel Munguia、Daniel Rothchild、DavidSo、MaudTexier 和 JeffDean。2021年。碳排放与大型神经网络训练。arXiv 预印本 arXiv:2104.10350（2021年）。[24] Xiang Ren、JiamingShen、Meng Qu、XuanWang、ZeqiuWu、Qi Zhu、Meng Jiang、Fangbo Tao、Saurabh Sinha、David Liem、Peipei Ping、Richard M. Weinshilboum、以及 Jiawei Han。2017。Life-iNet：一种面向生命科学的基于结构化网络的知识探索与分析系统。发表于 ACL。[25] Eric Sven Ristad 和 Peter N Yianilos。1998。字符串编辑距离的学习。TPAMI（1998）。[26] Stephen Robertson、Hugo Zaragoza 等。2009。概率相关性框架：BM25 及其后续发展。《信息检索基础与趋势》（2009）。[27] Gerard Salton 和 Chris Buckley。1988。自动文本检索中的词项加权方法。《信息处理与管理》（1988）。[28] Alberto Santos、Ana R Colaço、Annelaura B Nielsen、Lili Niu、Philipp E Geyer、Fabian Coscia、Nicolai J Wewer Albrechtsen、Filip Mundt、Lars Juhl Jensen 和 Matthias Mann。2020。临床知识图谱将蛋白质组学数据整合到临床决策中。bioRxiv（2020）。[29] Timo Schick 和 Hinrich Schütze。2021。利用完形填空问题进行少样本文本分类和自然语言推理。发表于 EACL。[30] Lynn M Schriml、James B Munro、Mike Schor、Dustin Olley、Carrie McCracken、Victor Felix、J Allen Baron、Rebecca Jackson、Susan M Bello、Cynthia Bearer 等。2022。人类疾病本体 2022 年更新。《核酸研究》（2022）。[31] Jiaming Shen 和 Jiawei Han。2022。自动化分类体系发现与探索。Springer Nature。[32] Jiaming Shen、Wenda Qiu、Yu Meng、Jingbo Shang、Xiang Ren 和 Jiawei Han。2021。TaxoClass：仅使用类别名称的层次化多标签文本分类。发表于 NAACL。[33] Taylor Shin、Yasaman Razeghi、Robert L Logan IV、Eric Wallace 和 Sameer Singh。2020。AutoPrompt：利用自动生成的提示从语言模型中引出知识。发表于 EMNLP。[34] Dibakar Sigdel、Vincent Kyi、Aiden Zhang、Shaun P Setty、David Liem、Yu Shi、Xuan Wang、Jiaming Shen、Wei Wang、Jiawei Han 和 Peipei Ping。2019。生物医学出版物中用户定义的短语与类别关联的云端短语挖掘与分析。JoVE 144（2019）。[35] Karan Singhal、Shekoofeh Azizi、Tao Tu、S Sara Mahdavi、Jason Wei、Hyung Won Chung、Nathan Scales、Ajay Tanwani、Heather Cole-Lewis、Stephen Pfohl 等。2022。大语言模型编码临床知识。arXiv 预印本 arXiv:2212.13138（2022）。[36] Chang Su、Yu Hou、Suraj Rajendran、Jacqueline RMA Maasch、Zehra Abedi、Haotan Zhang、Zilong Bai、Anthony Cuturrufo、Winston Guo、Fayzan F Chaudhry 等。2021。通过综合性生物医学知识中心（iBKH）开展生物医学发现。medRxiv（2021）。[37] Fabian M Suchanek、Serge Abiteboul 和 Pierre Senellart。2011。PARIS：关系、实例与模式的概率式对齐。VLDB（2011）。[38] Zequn Sun、Qingheng Zhang、Wei Hu、Chengming Wang、Muhao Chen、Farahnaz Akrami 和 Chengkai Li。2020。基于嵌入的知识图谱实体对齐基准研究。VLDB（2020）。[39] Together。2023年。GPT-JT-6B。https://huggingface.co/togethercomputer/GPT-JT-6B-v1。访问日期：2023年2月14日。[40] George Tsatsaronis、Georgios Balikas、Prodromos Malakasiotis、Ioannis Partalas、Matthias Zschunke、Michael R Alvers、Dirk Weissenborn、Anastasia Krithara、Sergios Petridis、Dimitris Polychronopoulos 等。2015。BIOASQ 大规模生物医学语义索引与问答竞赛概述。《BMC 生物信息学》（2015）。[41] Boshi Wang、Sewon Min、Xiang Deng、Jiaming Shen、You Wu、Luke Zettlemoyer 和 Huan Sun。2022。迈向理解思维链提示：关于关键因素的实证研究。arXiv 预印本 arXiv:2212.10001（2022）。[42] Lidan Wang、Jimmy Lin 和 Donald Metzler。2011。一种用于高效排序检索的级联排序模型。发表于 SIGIR。[43] Lu Wang、Ruiming Tang、Xiaofeng He 和 Xiuqiang He。2022。通过子目标表示学习实现动态治疗推荐的层次化模仿学习。发表于 WSDM。[44] David S Wishart、Yannick D Feunang、An C Guo、Elvis J Lo、Ana Marcu、Jason R Grant、Tanvir Sajed、Daniel Johnson、Carin Li、Zinat Sayeeda 等。2018。DrugBank 5.0：DrugBank 数据库的 2018 年重大更新。《核酸研究》（2018）。[45] Zhibiao Wu 和 Martha Palmer。1994。动词语义与词汇选择。发表于 ACL。[46] Bo Xiong、Nico Potyka、Trung-Kien Tran、Mojtaba Nayyeri 和 Steffen Staab。2022。用于 EL++知识库的忠实嵌入。发表于 ISWC。[47] Chengjin Xu、Fenglong Su、Bo Xiong 和 Jens Lehmann。2022。利用时间关系注意力进行时间感知实体对齐。发表于 WWW。[48] Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin 等。2022。Opt：开放式预训练 Transformer 语言模型。arXiv 预印本 arXiv:2205.01068（2022）。[49] Zhuosheng Zhang、Aston Zhang、Mu Li 和 Alex Smola。2023年。大语言模型中的自动思维链提示。发表于第十一届国际学习表征会议。[50] Chaoyu Zhu、Zhihao Yang、Xiaoqiong Xia、Nan Li、Fan Zhong 和 Lei Liu。2022。基于知识图谱嵌入的特定疾病多模态推理。《生物信息学》（2022）。