软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007370] [CSTR: 32375.14.jos.007370]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

综合实体语义和本体信息的多源中文医疗知识图谱实体对齐*

丁瑞卿[1,2], 赵俊峰[1,2], 王乐业[1,2]

[1](高可信软件技术教育部重点实验室 (北京大学), 北京 100871)

[2](北京大学 计算机学院, 北京 100871)

通信作者: 王乐业, E-mail: leyewang@pku.edu.cn

摘　要: 知识图谱作为结构化的知识表示形式, 在医疗领域具有广泛应用. 实体对齐, 即识别不同图谱中的等价实体, 是构建大规模知识图谱的基础步骤. 尽管已有大量研究关注此问题, 但主要集中在两个图谱的对齐任务上, 一般通过捕捉实体语义和图谱结构信息生成实体的向量表示, 之后计算向量相似度以确定等价实体. 在发现多源图谱对齐过程中存在对齐错误传递的问题的基础上, 考虑到医疗场景对实体对齐的准确性要求较高, 设计综合实体语义和本体信息的多源中文医疗知识图谱实体对齐方法 (MSOI-Align). 该方法首先将多个图谱进行两两组合, 利用表示学习生成实体向量表示, 并且综合实体名称的相似度和本体一致性约束, 借助大语言模型筛选得到候选实体集合. 随后, 基于三元闭包理论结合大语言模型对候选实体集合进行自动化的对齐错误传递识别与纠正. 在 4 个中文医疗知识图谱上的实验结果表明, MSOI-Align 方法显著提升了实体对齐任务的精确性, 与最优的基准方法相比, Hits@1 指标从 0.42 提升至 0.92. 融合后的知识图谱 CMKG 包含 13 类本体、19 万实体和约 70 万三元组. 考虑到版权限制, 开源了受限图谱外的另外 3 个图谱融合的结果——OpenCMKG.

关键词: 中文医疗知识图谱; 多源知识图谱对齐; 大语言模型应用; 本体信息; 实体语义; 对齐错误传递

中图法分类号: TP18

中文引用格式: 丁瑞卿, 赵俊峰, 王乐业. 综合实体语义和本体信息的多源中文医疗知识图谱实体对齐. 软件学报. http://www.jos.org.cn/1000-9825/7370.htm

英文引用格式: Ding RQ, Zhao JF, Wang LY. Multi-source Chinese Medical Knowledge Graph Entity Alignment via Entity Semantics and Ontology Information. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7370.htm

Multi-source Chinese Medical Knowledge Graph Entity Alignment via Entity Semantics and Ontology Information

DING Rui-Qing[1,2], ZHAO Jun-Feng[1,2], WANG Le-Ye[1,2]

[1](Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing 100871, China)

[2](School of Computer Science, Peking University, Beijing 100871, China)

Abstract: Knowledge graph (KG), as structured representations of knowledge, has a wide range of applications in the medical field. Entity alignment, which involves identifying equivalent entities across different KGs, is a fundamental step in constructing large-scale KGs. Although extensive research has focused on this issue, most of it has concentrated on aligning pairs of KGs, typically by capturing the semantic and structural information of entities to generate embeddings, followed by calculating embedding similarity to identify equivalent entities. This study identifies the problem of alignment error propagation when aligning multiple KGs. Given the high accuracy requirements for entity alignment in medical contexts, we propose a multi-source Chinese medical knowledge graph entity alignment method (MSOI-Align) that integrates entity semantics and ontology information. Our method pairs multiple KGs and uses representation learning to generate entity embeddings. It also incorporates both the similarity of entity names and ontology consistency constraints, leveraging a large language model to filter a set of candidate entities. Subsequently, based on triadic closure theory and the large language

* 基金项目: 国家自然科学基金 (U23A20468, 62133004, 72188101)
收稿时间: 2023-12-29; 修改时间: 2024-05-03, 2024-07-22, 2024-09-26; 采用时间: 2024-12-01; jos 在线出版时间: 2025-04-25

model, MSOI-Align automatically identifies and corrects the propagation of alignment errors for the candidate entities. Experimental results on four Chinese medical knowledge graphs show that MSOI-Align significantly enhances the precision of the entity alignment task, with the Hits@1 metric increasing from 0.42 to 0.92 compared to the state-of-the-art baseline. The fused knowledge graph, CMKG, contains 13 types of ontologies, 190 000 entities, and approximately 700 000 triplets. Due to copyright restrictions on one of the KGs, we are releasing the fusion of the other three KGs, named OpenCMKG.

Key words: Chinese medical knowledge graph; multi-source knowledge graph entity alignment; large language model (LLM) application; ontology information; entity semantics; alignment error propagation

知识图谱 (knowledge graph, KG) 是以结构化形式描述的知识元素及其联系的集合, 有广泛的应用场景, 如搜索[1]、推荐[2,3]和问答[4,5]等. 领域知识图谱包含特定领域范围内的知识并面向一个或者多个领域的应用场景, 主要用来解决特定行业或者细分领域的专业问题. 近年来, 很多研究致力于构建领域知识图谱, 例如金融[6]、科研[7]和医疗[8]等. 其中, 医疗知识图谱深刻揭示了医学实体之间的语义网络, 是医疗人工智能的核心. 根据知识图谱, 我们可以通过算法提供医学知识支撑以及生成结果的医学解释, 从而有效提高系统的解释性和准确性[9].

现实场景中, 单一的小型知识图谱往往不够完整, 在支持特定应用时仅能提供有限的知识覆盖率[10,11], 会影响具体任务的表现. 但是, 从头开始构建大规模医学知识图谱是一个复杂且成本高昂的工程, 涉及领域文本的收集整理、命名实体识别、关系抽取, 以及知识修正等多方面内容[12–14]. 目前为止, UMLS[8]是最为人熟知的医疗知识库, 包含大量医学生物相关的实体, 其中绝大部分为英文, 中文占比很少. 而中文医疗知识图谱的构建相对较晚, 没有形成一个大众广泛认可的图谱, 但是存在一些小型的开源图谱[9]. 因此, 融合多个小规模的垂域知识图谱 (比如, 用药图谱和疾病诊疗图谱) 得到规模更大、质量更高的图谱, 相对来说效率更高. 实体对齐 (entity alignment) 是一种广泛采用的知识图谱融合方式, 可以消除异构数据中实体冲突、指向不明等不一致性问题, 从顶层创建一个大规模的统一知识库, 帮助机器理解多源异质的数据[15].

然而, 在使用实体对齐方法进行中文医疗知识图谱对齐时, 需要解决以下挑战.

(1) 如何更好地整合实体语义和本体信息等多维度特征? 现有工作主要关注解决跨语言实体对齐的问题, 基于表示学习方法, 利用知识图谱的三元组信息生成实体嵌入表示, 再计算它们的相似度以筛选出另外一个图谱中匹配的实体. 但是在中文知识图谱的实体对齐中, 实体的名称及本体信息已知的情况下, 如何高效充分地融合实体的名称、本体等信息并且结合图谱结构进行对齐是需要思考的方向.

(2) 如何识别和解决多源知识图谱对齐的错误传递?目前大部分研究主要针对两个图谱间的对齐[15,16], 多图谱对齐的研究很少. 不同于两个图谱的对齐只需要找到两个图谱中等价的实体, 在多源知识图谱对齐的过程中, 对齐错误的传递使得本来不等价的实体相互连接. 这类错误对于构建高质量的图谱来说是想要极力避免的. 因此, 怎么在无需人工干预的情况下自动识别出潜在的错误并加以纠正也是亟待解决的重要问题.

针对以上挑战, 本文设计了综合实体语义和本体信息的多源中文医疗知识图谱实体对齐方法 (MSOI-Align), 融合知识图谱中实体的多种信息以确保更准确地实体对齐结果. 主要贡献包括以下几个方面.

(1) 在知识图谱对齐的过程中, 综合考虑了图谱结构和实体名称的相似度, 以及本体一致性信息, 并利用大语言模型 (large language model, LLM) 中包含的知识对候选的匹配实体对作进一步筛选排序, 形成了综合实体语义和本体信息的两图谱实体对齐方法 (SOI-Align).

(2) 将 SOI-Align 方法扩展到多源图谱对齐问题时, 指出多个知识图谱对齐间存在的错误传递问题, 在多个图谱两两组合进行对齐的基础上, 结合三元闭包理论[17]和大语言模型自动化识别和纠正候选对齐实体对中可能存在的错误传递, 得到完整的多源中文医疗知识图谱实体对齐方法 MSOI-Align.

(3) 将该方法在 4 个中文医疗知识图谱的对齐上进行验证, 发现和常用的知识图谱对齐算法相比, 在精确度上提升了 1 倍以上, 效果明显; 并且将融合后的知识图谱与单个图谱在医疗数据增强的服务[18]上进行比较, 发现融合后的图谱表现更优, 进一步验证了融合后的图谱质量更高.

(4) 融合后的知识图谱 CMKG 包含 13 类 19 万实体, 约 70 万三元组. 考虑到其中一个图谱的版权限制, 将另外 3 个图谱融合的结果开源 (OpenCMKG), 其中包含 7 类 6 万个实体, 35.5 万个三元组, 可见于 https://github.com/Ruiqing

Ding/OpenCMKG.

本文第 1 节介绍知识图谱对齐的相关方法和研究现状. 第 2 节介绍本文问题定义. 第 3 节详述模型框架和方法. 第 4 节进行实验评估和结果分析. 第 5 节总结全文并展望未来研究方向.

1 相关工作

知识图谱实体对齐就是将不同来源的知识进行整合, 识别出表示同一个现实对象的实体, 从而提高知识图谱的完整性和一致性. 作为知识图谱研究领域中的基础问题之一, 该问题多年来一直吸引着众多研究者的关注[19]. 本节总结以往研究中提出的通用实体对齐方法, 并将医疗知识图谱对齐相关研究单独展开介绍, 总结现有研究存在的缺陷或尚未考虑的问题.

1.1 通用的知识图谱实体对齐方法

在深度学习技术流行之前, 大多数方法注重设计合适的相似特征和基于贝叶斯模型的概率估计, 例如, RiMOM[20]将对齐问题转化为最小化决策风险; LogMap[21]通过词法和图匹配, 以及映射关系的修复迭代式地找到对齐实体; PARIS[22]在每次迭代中通过推断实体和关系的等价概率推理扩展了实体和关系的映射关系.

近年来, 基于嵌入的方法因为其灵活性和有效性成为实体对齐研究的主流方向. TransE[23]最早使用嵌入学习方法来表征关系数据, 使得头实体的表征和关系的表征融合后可以映射到尾实体的表征; 随后, MTransE[24]在 TransE 实现单个知识图谱嵌入学习的基础上, 增加了知识图谱对齐的部分, 此部分的基本思想是根据已知的对齐实体, 学习空间迁移矩阵将一个知识图谱的表征空间映射到另一个知识图谱的表征空间; JAPE[25]提出了基于嵌入的跨语言实体对齐方法并在 DBpedia 数据上进行验证; BootEA[26]针对基于嵌入的实体对齐提出自助方法 (bootstrapping), 迭代地将可能的实体对齐标记为训练数据, 以学习面向知识图谱的实体嵌入; GCN-Align[27]和图匹配网络[28]均使用图卷积神经网络学习图谱中的实体和关系并进行对齐; TransEdge[29]针对图谱的实体之间存在“一对多”和“多对一”的复杂关系类型, 提出以边 (即关系) 为中心的嵌入方法, 根据关系所处的头尾实体的不同决定关系有不同的表示; MultiKE[30]提出了多视图图谱实体对齐方法, 将实体的特征分成 3 个部分 (名称、关系、属性特征) 分别建立表示学习模型, 将多个视图的实体嵌入充分联合以提高对齐模型的性能.

最近, 很多相关的工作从不同的角度切入, 如组合概率模型和嵌入模型的无监督的图谱对齐方法 PRASE[31]和采用自监督学习目标的图谱对齐方法 SelfKG[32]等. 同时, 也有综述文章总结了知识图谱对齐方法遵循 3 步范式: 先基于 TransE 类或者图神经网络类的方法将实体映射到一个向量空间; 之后将两个图谱的向量空间做映射; 最后使用相似度指标 (如, 余弦相似度) 衡量实体在向量空间中的距离, 从而确定当前实体在另一个图谱中的对应实体 (counterpart), 并且开发了工具箱 OpenEA[14]和 EAkit[19], 极大地简化了研究复杂度、提升了相关研究的效率. 其中, DBP15K[25]和 DWY100K[26]是两个常用的实验数据集. DBP15K 包括 4 种不同语言的知识图谱 (中文、英语、法语和日语), DWY100K 中部分实体用 Wikidata 的索引 ID 表示, 因此在这两个数据集中实体名称很难为对齐任务提供有价值信息, 也使得当前大部分的研究都主要依靠知识图谱内的实体关系来学习实体的嵌入表示.

总的来说, 围绕两个图谱对齐问题目前已经展开了较多研究, 但是设计的方法更多集中在关注图结构或者三元组的依赖关系上, 对实体本身语义信息的利用尚不充分. 同时, 验证使用的知识图谱数据相对局限, 对多个图谱实体对齐问题的研究较少. 但是多图谱对齐不能简单地拆分成两两图谱对齐, 我们需要进一步解决多图谱实体对齐时极易出现的对齐实体错误传递问题, 这在之前的研究中尚未提及[33].

1.2 医疗领域知识图谱实体对齐方法

除了通用领域的知识图谱实体对齐任务外, 也有一些研究特别针对医疗知识图谱, 分析实体对齐方法在医疗领域真实场景下的表现并提出新方法. 比较典型的是 Zhang 等人[9]给出了一个医疗评估数据集 MED-BBK-9K, 包括两个中文医疗知识图谱, 分别来自百度百科和实际的医疗企业, 但是实体数量较少, 每个图谱只有 9 000 多个实体. 经实验验证, 大多数模型严重依赖种子数据 (即训练数据中的对齐实体), 不同采样的种子会学习到不同的向量空间映射关系, 最终结果差别也较为明显. 后续, 也有研究关注已有方法忽略了本体信息, 无法利用实体类别的层

次结构信息[34], 极易导致虚假的映射关系. 针对此问题, OntoEA[35]通过联合嵌入学习将本体信息融入模型训练过程, 并且考虑了不同类间的分离约束来提高映射的准确性. 但是在实验结果中发现仍然存在一些本体矛盾的错误, 即模型训练中加入的约束只能起到部分作用, 不能完全避免该问题.

总的来说, 由于医疗领域对准确性要求极高, 实体对齐的精确性可能直接影响到医疗决策和患者安全, 因此我们需要更多关注实体对齐任务的精度表现, 特别是 Hits@1. 同时, 实际的医疗知识图谱缺乏实体对齐标签, 在构建医疗知识图谱对齐任务时, 需要思考如何结合实体的多维度信息提升对齐效果, 特别是已有方法尚未充分利用的语义信息.

2 问题定义

2.1 知识图谱

知识图谱可以表示为 $G=(E, R, O, T)$, 其中, E、R、O、T 分别为实体、关系、本体和三元组的集合. 任意一个三元组 $(h, r, t)\in T$ 包含一个头实体 $h\in E$, 一个关系 $r\in R$ 和一个尾实体 $t\in E$. 本体可以看作实体的类别信息, 每个实体都有对应的本体, 这个关系也可以用三元组表示 $(e, BelongTo, o)$. 图 1 给出了以疾病为核心的医疗知识图谱中广泛存在的本体类型以及这些本体间可能存在的关系, 包括疾病需要的检查、可能的症状、常用或推荐的药品等.

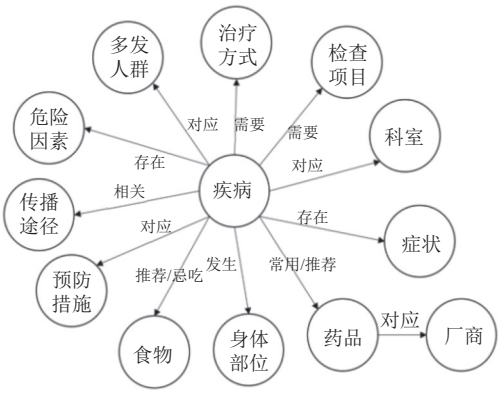


图 1 医疗知识图谱的本体之间可能存在的关系

2.2 知识图谱实体对齐

● 两知识图谱对齐: 已知两个知识图谱 $G_1=(E_1, R_1, O_1, T_1)$ 和 $G_2=(E_2, R_2, O_2, T_2)$, 实体对齐的目标就是找到等价的实体对集合, $Y=\{(e_1, e_2)| e_1\equiv e_2, e_1\in E_1, e_2\in E_2\}$. 同时我们要求对齐实体对需要满足本体一致, 即若 $e_1\equiv e_2$, 且有 $e_1\rightarrow o_1$, $e_2\rightarrow o_2$, 则一定有 $o_1\equiv o_2$.

● 多源知识图谱对齐: 与两知识图谱对齐相似, 已知多个知识图谱 $\{G_1, G_2,\ldots, G_N\}$, 找到等价的实体对集合 $Y=\{(e_i, e_j) \mid e_i\equiv e_j, e_i\in E_i, e_j\in E_j, i\neq j\}$ $(i, j\in\{1, 2,\ldots, n\})$. 不同于两知识图谱对齐, 该问题可能存在错误的对齐传递. 例如, G_1 和 G_2 得到对齐的实体对 (e_1, e_2), G_1 和 G_3 得到对齐的实体对 (e_1, e_3), 根据等价关系, 可以得到 $e_2\equiv e_3$, 但是如果实际上已知 $e_2\neq e_3$, 就说明存在错误传导. 由此, 可以反向推导出 $e_1\neq e_2$ 或者 $e_1\neq e_3$.

3 方 法

为了高效构建大规模高质量的医疗知识图谱, 我们设计了综合实体语义和本体信息的多源知识图谱对齐方法 (multi-source knowledge graph entity alignment via entity semantics and ontology information, MSOI-Align). 该方法主要包括 2 个部分: (1) 将多个来源的知识图谱两两组合, 转化成较为常见的两图谱实体对齐问题. 针对此问题设计 SOI-Align 方法, 充分考虑实体表示向量和名称的相似性, 以及本体一致性的约束, 借助 LLM 将对齐实体进

行初步筛选生成对齐候选集; (2) 在汇总第 (1) 步两两组合的对齐候选集得到多源图谱对齐结果时, 基于三元闭包理论借助 LLM 自动化识别和纠正对齐错误传递, 保证最终对齐结果的精准度.

3.1 两图谱实体对齐 SOI-Align

本文的研究主要针对中文的医疗知识图谱, 每个实体均有自己的名称. 不同于常规的跨语言图谱对齐任务, 可以充分利用实体名称的文本相似度, 例如小儿肠炎和儿童肠炎这两个实体, 只从文本相似度基本就可以判定为对齐实体. 同时, 也希望综合实体语义和实体的本体信息, 确保匹配的准确性和可靠性.

基于以上目标, 本文设计了两图谱实体对齐方法 SOI-Align, 整体框架如图 2 所示, 包括 4 个核心步骤.

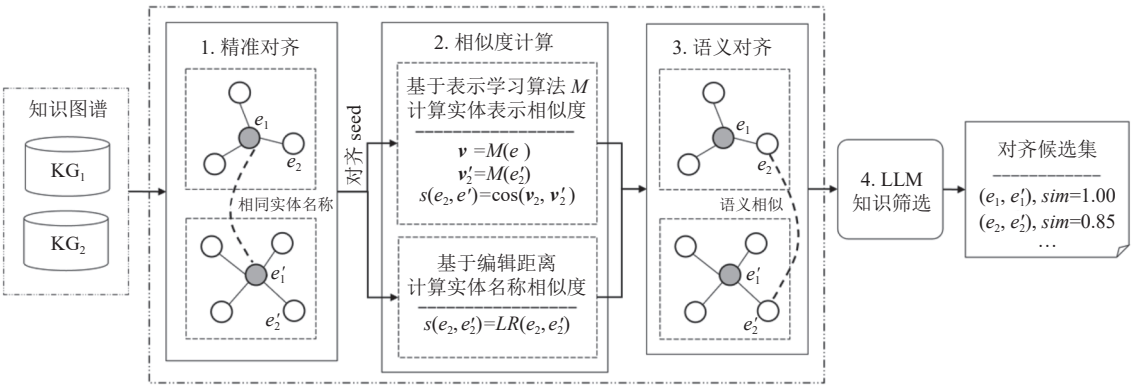


图 2 两图谱对齐 SOI-Align 思路

(1) 精准对齐: 给定两个知识图谱, 若实体名称完全一致则认为两个实体精准对齐, 则将这些对齐的实体对集合作为图谱对齐表示学习方法的种子, 为接下来的表示学习提供有标签的训练样本;

(2) 相似度计算: 为了更全面地获取实体信息, 我们尝试利用表示学习捕捉图谱结构信息. 具体而言, 输入两个知识图谱和精准对齐提供的对齐实体对集合, 根据已有的图谱对齐表示学习方法 $\mathcal{M}$ 得到实体的表示向量并计算余弦相似度. 同时, 考虑到医疗实体名称相对简短, 直接基于编辑距离计算实体名称的相似度;

(3) 语义对齐: 综合实体表示向量和实体名称的相似度, 同时判断两个实体的本体是否一致, 获得实体间的汇总相似度分数. 在相似度高于一定阈值时将其作为实体对齐的初步候选集;

(4) LLM 知识筛选: 考虑到直接利用相似度数值进行排序很可能导致错误, 特别是在多个实体相似度相差较小的情况下. 因此, 我们进一步利用大语言模型编码的隐含知识[36]对得到的对齐候选集做筛选, 获得最终的候选结果.

3.1.1 实体相似度计算及语义对齐

近年来, 知识图谱表示方法发展迅速, 主要使用知识图谱表示学习或者图神经网络模型为每个实体学习一个低维向量表示, 之后计算向量的相似度以找到等价实体对. 简而言之, 给定一个基于表示学习的实体对齐方法 $\mathcal{M}$ (MtransE[24]等), 通过训练可以得到两个图谱中所有实体的向量表示. 假设实体 e_1 属于知识图谱 G_1, 实体 e_2 属于知识图谱 G_2, 它们的向量表示分别为 $\boldsymbol{v}_1=\mathcal{M}(e_1)$, $\boldsymbol{v}_2=\mathcal{M}(e_2)$, 则可以计算两个向量表示间的余弦相似度, 即:

$$sim(e_1,e_2)=\cos(\boldsymbol{v}_1,\boldsymbol{v}_2)=\frac{\boldsymbol{v}_1\cdot\boldsymbol{v}_2}{|\boldsymbol{v}_1|\times|\boldsymbol{v}_2|} \tag{1}$$

设定一个相似度阈值 λ (我们在实验中设定为 0.75), 当 $sim(e_1, e_2)\geqslant\lambda$ 时, 则认为两个实体等价.

只使用向量表示计算相似度较为方便, 但是不可忽略的是很容易出现各类问题. 首先就是配对的实体本体不一致, 比如某个属于症状的实体和某个属于检查的实体实际上不能匹配, 但是只凭向量相似度大于某个阈值, 很容易出现错误匹配, 即本体矛盾. 同时, 我们也发现实体中是否包含否定词代表着完全不同的意义, 如, 不完全性肠梗阻和完全性肠梗阻是两个不同的疾病. 但是在知识图谱中这两个实体的三元组关系较为相似, 包括对应的症状、推荐使用的药物等. 所以, 它们学习到的向量表示相似度也较高. 这种情况需要融合其他维度信息: (1) 本体一致性: 判断 2 个实体的本体类型是否相同, 如果本体不同则表示一定不匹配; (2) 否定词检索: 给定否定词列表 N (不,

非, 否, 无, 没, 异, 假), 若两个实体中有且只有一个实体名称包含否定词, 则这两个实体一定不匹配; (3) 实体名称的莱文斯坦比 (Levenshtein ratio, *LR*): 实体名称已知情况下, 可以匹配的实体在名称上存在一定的相似性[9]. 考虑到编辑距离是常用的字符串差异比较方法, 本文采用基于编辑距离转换的莱文斯坦比作为名称相似度的计算方法. 由此, 实体语义相似度的计算公式可以改进为:

$$sim(e_1,e_2)=\mathcal{I}_o\times(1-\mathcal{I}_n)\times(\alpha\cdot LR(e_1,e_2)+(1-\alpha)\cdot\cos(\boldsymbol{v}_1,\boldsymbol{v}_2)) \tag{2}$$

其中, $\mathcal{I}_0$=1 表示 e_1 和 e_2 的本体一致, 否则 $\mathcal{I}_0$=0; $\mathcal{I}_n$=1 表示 e_1 和 e_2 中只有一个实体名称出现否定词, 否则 $\mathcal{I}_n$=0. 在实体匹配时, 可以根据新的相似度方法计算并选出候选的匹配实体. 算法 1 中详细列出了基于多维信息的实体匹配过程.

算法 1. 实体匹配算法 *Match*.

输入: 2 个实体集合 E_1、E_2 及对应的实体向量 V_1、V_2; 否定词列表 N; 相似度阈值 λ; 文本相似度占比 α;
输出: 两个实体集合中匹配的实体对 *dict_match*.

1. *dict_match*←{} //初始化匹配实体对字典
2. **FOR** e_1 in E_1 **DO**
3. *dict_sim*←{}
4. **FOR** e_2 in E_2 **DO**
5. 根据公式 (2) 计算相似度 *dict_sim*[e_2]
6. **END FOR**
7. 根据 value 值降序排列 *dict_sim*
8. **FOR** e_2 in *dict_sim* **DO**
9. **IF** *dict_sim*[e_2]≥λ **DO** //相似度达到阈值
10. *dict_match*[(e_1, e_2)] = *dict_sim*[e_2]
11. **END IF**
12. **END FOR**
13. **END FOR**
14. **RETURN** *dict_match*

但是当两个图谱规模较大、实体数量较多时, 两两实体名称之间分别计算莱文斯坦比的计算量巨大、效率极低. 因此, 怎样降低计算量、提升效率也需要加以解决. 实际上, 两两实体分别计算名称间的相似度时, 绝大部分的相似度均为 0, 所以很多的计算是不必要的. 也就是, 需要快速检索出可能匹配的实体, 并在可能匹配的实体集合内做计算即可. 基于这个想法, 我们将三元组的关系作为重要知识源, 从两个图谱名称相同的实体对出发, 根据相同的三元组关系, 找到对应的尾实体集合. 之后, 只需要在这两个实体集合中寻找可能匹配的实体对. 这样, 就可以极大地压缩搜索空间, 避免不必要的计算. 算法 2 给出了搜索过程的实现方式, 最终可以得到所有可能的匹配实体对候选集 *dict_match*.

算法 2. 基于三元组关系的实体检索及匹配 *Match*2.

输入: 两个知识图谱 G_1, G_2; 实体对齐模型 $\mathcal{M}$; 相似度阈值 λ; 文本相似度占比 α;
输出: 实体对齐候选集 *dict_match*.

1. S←{(e_1, e_2)∈E_1×E_2|e_1=e_2} //初始化对齐实体对
2. V_1, V_2 = $\mathcal{M}$(G_1, G_2, S) //使用 $\mathcal{M}$ 生成实体集合 E_1, E_2 的向量表示 V_1, V_2
3. O=O_1∩O_2 //两知识图谱共有的本体集合
4. **FOR** o in O **DO**

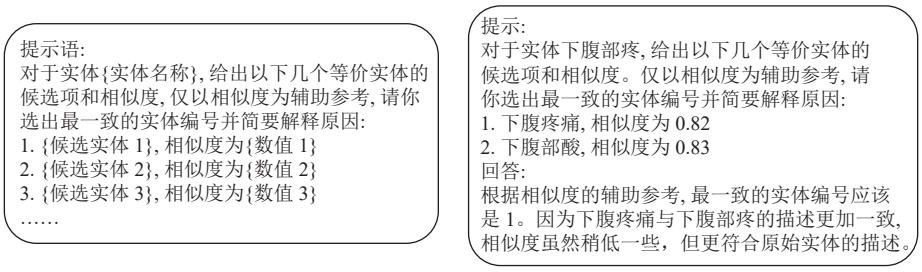
5. $D_1=\{e|e\in E_1 \& (e, BelongTo, o)\in T_1\}$
6. $D_2=\{e|e\in E_2 \& (e, BelongTo, o)\in T_2\}$
7. $dict\_match\_cur \leftarrow Match(D_1, D_2, \lambda, \alpha)$
8. $dict\_match.add(dict\_match\_cur)$
9. **END FOR**
10. **RETURN** *dict_match*

直观来说, 假设两个知识图谱 G_1 和 G_2 的实体集合分别是 E_1 和 E_2, 对应实体数量是 N_1 和 N_2. 如果直接将实体两两组合计算相似度的复杂度为 $O(N_1\times N_2)$. 但从名称相同的实体对出发, 记为$\{(e_1, e_2)|name(e_1)\equiv name(e_2), e_1\in E_1, e_2\in E_2\}$(数量为 M 对), 根据相同的三元组关系 r 找到的尾实体集合 $Tail(e_1, r)$ 和 $Tail(e_2, r)$, 数量分别为 n_1 和 n_2. 可知 $n_1 \ll N_1$, $n_2 \ll N_2$. 所以计算复杂度可以从 $O(N_1\times N_2)$ 降低到 $O(M\times n_1\times n_2)$. 以本文使用的知识图谱规模为例, N_1、N_2 为 10^4 数量级, M 为 10^4 数量级, 而 n_1、n_2 基本为个位数或十位数. 因此复杂度至少可以降低 2 个数量级.

3.1.2 LLM 知识筛选

在基于语义计算实体相似度时我们先设定阈值 λ 进行筛选生成候选集. 但是对一个实体来说, 可能有多个相似度高于 λ 的实体存在, 因此存在多个匹配. 考虑到实际应用中对医疗知识图谱质量要求较高, 我们的目标是找到最相似的实体做对齐, 而不是保留所有可能对齐的实体. 最简单的方法是直接选择相似度最高的实体, 但是当几个实体之间的相似度差别很小时, 相似度数值最大的很可能并不是最优选择. LLM 已被证明隐式编码大量知识, 可以作为内置的搜索引擎[37]. 也有研究使用 LLM 辅助知识图谱构建[38,39]和知识图谱补全[40]. 因此, 我们借助 LLM 的知识进行进一步筛选.

构建好的提示 (prompt) 是 LLM 效果的重要保证[41], 受到 SuperICL 将本地小模型作为 LLM 插件以提升文本预测任务性能的启发[42], 我们将本地计算的实体相似度作为提示的一部分, 相当于提供更多参考信息以确保 LLM 得出更合理的推理及解释. 具体的提示模版如图 3(a) 所示, 将算法 2 得到的实体匹配词典 *dict_match* 作为输入, 对有多个匹配实体候选的情况使用 LLM 做最终选择. 实验中, 通过 OpenAI 的 API 调用 ChatGPT (GPT-3.5-Turbo) 获得反馈. 图 3(b) 给出了一个 LLM 知识筛选的示例, 对于实体下腹部疼, 有两个可能匹配的实体下腹疼痛和下腹部酸, 语义相似度分别为 0.82 和 0.83. 在提示下, LLM 可以兼顾相似度和自身知识, 选择出更一致的实体: 下腹疼痛, 并且给出合理的解释: 下腹疼痛与下腹部疼的描述更加一致, 相似度虽然稍低一些, 但更符合原始实体的描述. 通过 LLM 知识筛选, 两图谱实体对齐的候选集中, 每个实体只保留了一个最可能匹配的实体.



(a) 提示模版　　(b) 根据提示得到 LLM 反馈的示例

图 3　LLM 知识筛选提示

3.2　多源图谱实体对齐方法 MSOI-Align

多源图谱实体对齐实际上就是对两知识图谱对齐方法 SOI-Align 的扩展, 但是需要解决实体对齐的错误传递问题, 尽可能保证对齐的精准度. 实体对齐的错误传递如问题定义部分所述, 这里给出一个具体的例子辅助理解:

已知当前 KG_1 和 KG_2 有候选实体对 (白斑, 皮肤白斑); KG_1 和 KG_3 有候选实体对 (白斑, 舌白斑). 如果只是将两两图谱通过 SOI-Align 方法得到的对齐结果拼接到一起, 以 KG_1 的实体白斑作为中介, 可以得到 KG_2 的皮肤白斑和 KG_3 的舌白斑也是匹配的, 但是明显它们对应着不同的身体部位, 皮肤和舌头, 是不相同的实体. 因此, 在多源知识图谱对齐时, 我们需要识别出此类问题并加以纠正.

基于以上观察, 本文以 SOI-Align 为基础, 针对多图谱对齐问题完善得到 MSOI-Align 方法. 具体流程如图 4 所示. 首先, 将多个知识图谱两两组合, 拆分为多个两知识图谱对齐问题, 通过 SOI-Align 算法得到多个对齐候选集; 之后, 将多个对齐候选集整合到一起, 检索是否某个实体在其他多个图谱上均有对齐的实体, 即进行三元闭包性质检查, 该过程由 LLM 进行辅助以减少人工干预, 实现自动化识别. 若发现多个实体之间的对齐关系存在错误传递, 则进一步纠正, 生成最终的实体对齐结果.

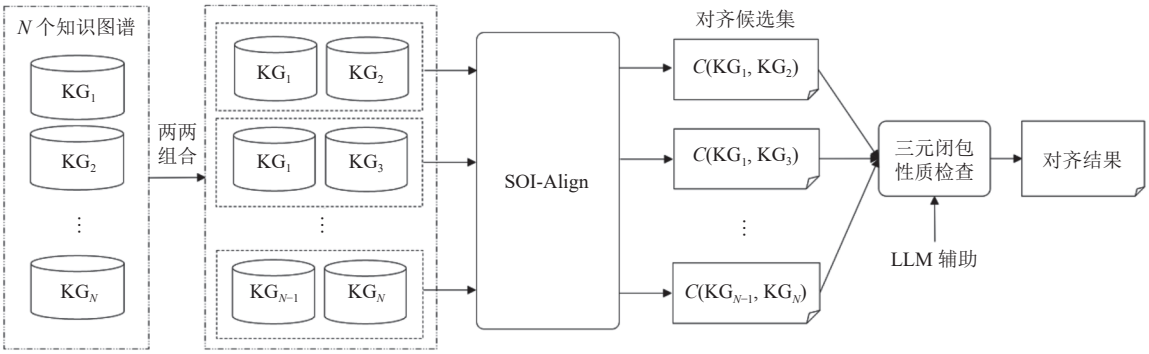


图 4 多源知识图谱对齐方法 MSOI-Align 流程

定理 1. 三元闭包 (triadic closure) . 在社交网络中, 由 *A*、*B*、*C* 这 3 个节点所组成的三元组, 如果 *A* 与 *B*、*A* 与 *C* 之间存在强联系, 则 *B* 与 *C* 之间也仅存在强联系.

在社交网络分析中, 三元闭包理论有广泛应用[17], 如定理 1 所示. 该理论是指如果两个人在社交网络中有共同的朋友, 那么他们之间很可能也存在相应的关系. 具体来说, 如果 *B* 和 *C* 是社交网络中的两个人, 而 *A* 是他们的共同朋友, 那么 *B* 和 *C* 之间很可能存在社交关系, 可能是朋友、亲戚等. 反向推导, 如果我们发现 *B* 和 *C* 是相互敌对的或者完全不存在联系, 那么 *A* 和 *B*、*A* 和 *C* 的朋友关系很可能不成立. 实际上, 根据以上说明我们可以确定 3 个节点之间存在 3 种平衡状态: 1) 3 个节点之间完全没有关联; 2) 只有两个节点之间有边; 3) 3 个节点之间均存在关联的边. 也就是 3 个节点仅存在两条边是不平衡的, 如图 5 所示.

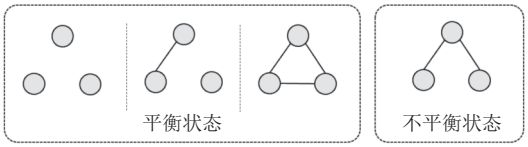


图 5 基于三元闭包理论的三节点平衡状态

由此, 如果发现存在 3 个实体相互对齐的情况, 需要确定是否处于平衡状态. 若不平衡, 说明存在错误传递问题, 可进一步修正. 以图 6 列出的两个场景为例, 已知经过 SOI-Align 进行两图谱实体对齐可得, KG_1 和 KG_2 有候选实体对: 抽动症和小儿抽动症; KG_1 和 KG_3 有候选实体对: 抽动症和多动症. 此时, 只有两个对齐关系, 处于不平衡状态, 需要转成平衡状态. 因此要验证小儿抽动症和多动症是否匹配, 若它们不匹配, 遵循严格的实体对齐标准, 可以认为 3 个实体之间不存在对齐关系, 不考虑只存在一个对齐关系的平衡状态. 同理, 针对白斑和皮肤白斑、白斑和舌白斑这两组实体对齐的情况, 也进一步验证皮肤白斑和舌白斑是否匹配. 在不匹配时, 可知白斑、皮肤白斑和舌白斑均不对齐. 该验证过程由 LLM 进行判断, 保证整个过程可以自动化实现. 以图 6 的第 1 个情况为例, 在发

现存在不平衡状态的 3 个实体后, 根据提示“作为医学专家, 小儿抽动症和多动症是等价的吗? 用等价或者不等价回答”, 得到 LLM 的反馈为“不等价”, 因此这 3 个实体之间存在对齐错误传递, 于是将它们的关系修改为均不对齐. 相比于对 3 个对齐关系都使用 LLM 重新做判断, 只对没有对齐关系的这一条边进行验证, 在符合三元闭包理论检验要求的同时, 和 LLM 的交互也更高效.

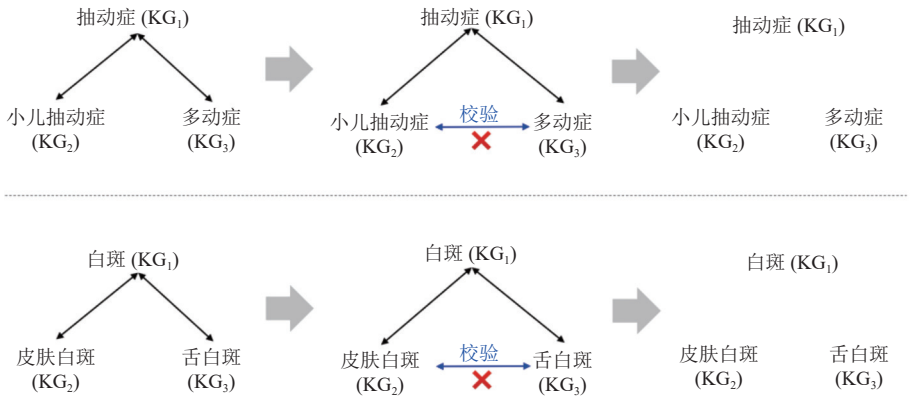


图 6　三元闭包理论应用示例

算法 3 给出了多源图谱对齐的详细流程. 通过这个方法, 能够从两图谱对齐的候选集中过滤掉所有不正确的匹配, 进一步提升对齐的准确性.

算法 3. 多源知识图谱对齐 MSOI-Align.

输入: 知识图谱集合$\{G_1, G_2,\ldots, G_N\}$; 大规模语言模型 *LLM*;
输出: 匹配实体对 *Res*.

1. *Combinations*=$\{(G_1, G_2),\ldots, (G_{N-1}, G_N)\}$ //知识图谱两两组合
2. **FOR** (G_i, G_j) in *Combinations* **DO**
3. C_{ij} = SOI-Align(G_i, G_j) //根据算法 2 得到两图谱对齐候选集
4. **END FOR**
5. 合并所有 C_{ij} 得到全部对齐候选集 *C*
6. 检索对齐候选集 *C* 包含的不平衡状态的三实体集合 $U=\{(e_i, e_j, e_k)|e_i, e_j, e_k$ are from different $G_s\}$
7. **FOR** (e_i, e_j, e_k) in *U* **DO** //不平衡状态下进行三元闭包检验
8. **IF** $(e_i, e_j) \notin C$ **DO** //当 (e_i, e_j) 不在对齐候选集
9. **IF** *LLM*(e_i, e_j) **DO** //当 LLM 判断等价时将其加入对齐候选集
10. *C.add*(e_i, e_j)
11. **ELSE** //否则从对齐候选集中删除其他 2 个
12. *C.remove*(e_i, e_k)
13. *C.remove*(e_j, e_k)
14. **END IF**
15. **END IF**
16. **END FOR**
17. *Res*←*C*
18. **RETURN** *Res*

4 实验分析

4.1 实验数据

本文收集了 4 个中文医疗知识图谱, 基本情况如下.

(1) CPubMedKG (https://cpubmed.openi.org.cn/graph/wiki, 2022 年 2 月 7–9 日获取): 通过 API 接口获取 CPubMed-KG1.1 中常见疾病及其相关的三元组, 共计 11 类本体、15.7 万实体, 约 46 万三元组;

(2) QAKG (https://github.com/liuhuanyong/QASystemOnMedicalKG, 2021 年 7 月 8 日获取): 以垂直型医药网站为数据来源, 以疾病为核心, 包含 7 类本体、4.4 万实体, 约 29 万三元组;

(3) OwnThink (https://www.ownthink.com/docs/kg, 2022 年 4 月 23 日获取): 包含 8 类本体、3.5 万实体, 约 21 万三元组;

(4) CHIP2021 (http://cips-chip.org.cn/2021/eval1, 2022 年 4 月 20 日获取): 从第 7 届中国健康信息处理会议 (CHIP2021) 的测评任务一数据集中抽取, 包含 2 类本体、1.8 万实体, 约 1.7 万三元组.

因为我们关注临床疾病相关的知识图谱, 没有扩大到生物和化学等基础医学信息相关的实体, 所以没有从 UMLS[8]抽取中文实体的子图, 也没有考虑 BIOS 图谱[11]. 使用到的 4 个知识图谱的本体类型和实体数量以及三元组相关的详细统计 信息分别列在表 1 和表 2 中.

表 1 知识图谱实体数量统计

本体	CPubMedKG	QAKG	OwnThink	CHIP2021
疾病	15859	8807	8616	6046
药物	33133	3828	3546	—
食物	—	4870	249	—
科室	255	54	48	—
生产商	—	17201	15898	—
症状	13264	5998	6336	12682
治疗方式	45007	—	598	—
检查	30425	—	—	—
多发人群	298	—	—	—
身体部位	2715	—	—	—
传播途径	30	—	—	—
风险因素	13998	—	—	—
预防措施	2690	—	—	—
总计	157674	40758	35281	18728

表 2 知识图谱实体数量统计

三元组关系	CPubMedKG	QAKG	OwnThink	CHIP2021
科室所属关系	—	37	37	—
并发症	13266	12024	24712	—
疾病所属科室	1669	8806	8615	—
疾病一般用药	107965	14647	13332	—
疾病可吃食物	—	22230	21749	—
疾病风险因素	34646	—	—	—
疾病对应症状	38234	54709	49093	—
疾病需要检查	115527	39418	—	—
疾病需要治疗	136174	—	20607	—
疾病忌吃食物	—	22239	5458	—
疾病推荐药物	—	59465	53296	—
疾病推荐食物	—	40221	—	—
疾病发生部位	7702	—	—	—
疾病传播路径	53	—	—	—
疾病多发人群	779	—	—	—
疾病预防措施	4783	—	—	—
疾病诱发疾病	1775	—	—	—
药物生厂商	—	17315	15956	—
同义实体	—	—	—	17719
总计	462573	291111	212855	17719

考虑到每个实体都有对应的名称, 而且本文方法将名称一致的实体作为对齐的种子, 即基于表示学习方法训练的正样本, 也统计了这些知识图谱中各类本体类型中名称一致的实体数量. 如后文表 3 所示, 可以发现这些实体占比较大, 也从一定程度上保证了在训练时有较为充足的正样本, 能够提供丰富的监督学习信号.

4.2 基准方法

在实验中, 采用 6 个知识图谱对齐任务常用的基准模型.

(1) MtransE[24]: 将 TransE 模型作为实体表示学习模型, 进一步根据对齐的三元组, 学习空间迁移矩阵实现两个知识图谱的映射;

(2) BootEA[26]: 为增加训练难度, 采样区分度更小的三元组作为负样本; 同时采用 bootstrapping 策略, 在更新实体表示的过程中把可能对齐的实体样本加入训练数据;

(3) GCN-Align[27]: 将图卷积神经网络 (GCN) 作为实体特征编码器以学习知识图谱的图结构信息;

(4) TransEdge[29]: 针对一对多、多对多等复杂的三元组关系, 改进了 TransE 模型;

(5) OntoEA[35]: 综合本体和实体信息进行联合学习, 避免对齐的实体出现本体矛盾的错误;

(6) FuAlign[43]: 首先将对齐种子看做一个实体从而聚合两个知识图谱, 之后通过多视图表示学习方法, 整合实体名称、实体邻域上下文和拓扑结构等不同类型信息生成实体表示.

表 3 知识图谱间名称一致的实体数量统计

实体类型	名称一致的实体数量
疾病	9429
药物	3945
生产商	15761
症状	6708
治疗方式	212
科室	54
食物	247
总计	36356

因为 MSOI-Align 中基于语义的相似度评价、LLM 知识筛选和对齐错误传递识别方法和具体使用的表示学习方法无关, 所以这 6 个基准方法都可以作为 MSOI-Align 的实体表示学习的选择方案, 进而影响实体间相似度的大小. 因此, 我们也开发了多种变体 MSOI-Align-Mix_{x}, 在根据公式 (2) 计算实体相似度的时候, 综合考虑实体名称的莱文斯坦比和实体表示的余弦相似度, 其中 α 设为 0.25, x 可表示为以上 6 种基准方法的任意一个.

4.3 实验设置

4.3.1 模型参数设置

在训练所有基准方法时, 将两个知识图谱中名称相同的实体对均作为正样本, 保证充分学习到知识图谱的三元组信息以及两个图谱对齐实体的信息, 由此生成更好的实体表示. 在开源知识图谱工具箱 Eakit (https://github.com/THU-KEG/EAkit) 和 OntoEA (https://github.com/ZihengZZH/OntoEA) 的代码基础上完善并进行实验. 为了公平比较, 所有基准方法均训练 1000 个 epochs, 学习率为 0.005, 生成实体的表示向量维度为 100. 其他参数采用 EAkit 和 OntoEA 代码中的默认设置.

4.3.2 评估指标

实体对齐任务常用的评价指标包括 Hits@k 和 MRR (平均倒数排名, mean reciprocal ranking)[19]. Hits@k 表示与某一实体正确对齐的实体排在前 k 个的占比, MRR 衡量给定的查询 (即实体) 中, 模型返回的对齐结果中排名第 1 的平均倒数. 因此, 这两个指标的取值范围在 0–1 之间, 越接近于 1 说明方法的性能越好. 考虑到医疗领域知识图谱更要求精准性, 在对齐过程中, 我们更加关注 Hits@1 指标, 同时也补充新指标, 实体正确匹配收益. 正确匹配收益是用对齐结果中正确的数量减去错误的数量, 越大越好. 若为负值, 说明对齐实体中的错误数量占比更多, 该结果较差.

对齐实体的标注, 因为使用的 4 个知识图谱没有已知的对齐实体 (当实体名称不同时), 无法直接评价模型的对齐结果, 需要一定的人工标注. 标注由有医学知识项目背景的学生完成. 在尽量减少工作量并保证评估的有效性的前提下, 从得到结果的每种类型 (如, 疾病、症状等) 中随机抽取 100 个实体对进行标注, 预估出大致的 Hits@1 和对齐数量. 同时为了保证准确标注, 当不确定两个实体是否匹配时标注人员查询外部资料进一步确认. 标注数据可见于 https://github.com/RuiqingDing/OpenCMKG/blob/main/manual_annotation.csv.

4.3.3 实验环境

实验均在单张 NVIDIA GeForce RTX 3090 GPU 进行, 主要环境配置包括 Python 3.6、PyTorch 1.8.0、CUDA

11.1 和 PyTorch Geometric (PyG) 2.0.3.

4.4 实验结果与分析

4.4.1 实体对齐结果

表 4 列出了基于表示学习的模型和 MSOI-Align 这两大类方法的实体对齐数量和 Hits@1, 分别对应了每个方法的两行数据. 汇总统计了所有实体类型的对齐数量之和以及加权平均的 Hits@1, 收益为对齐正确数量减去错误数量. 根据表 4 中的数据, 可以得出以下结论.

表 4 多源知识图谱对齐的实体数量、Hits@1 和收益

类别	方法	指标	疾病	药物	生产商	症状	治疗方式	科室	食物	汇总统计	收益
表示学习模型	MtransE	匹配数量	488	518	569	277	172	1	198	2223	−2081
		Hits@1	0.01	0.00	0.10	0.02	0.02	0.00	0.00	0.03	
	TransEdge	匹配数量	511	407	697	674	110	1	609	3009	−2577
		Hits@1	0.09	0.08	0.16	0.03	0.00	0.00	0.01	0.07	
	GCN-Align	匹配数量	552	416	542	435	407	4	645	3001	−2173
		Hits@1	0.14	0.00	0.55	0.07	0.02	0.00	0.00	0.14	
	BootEA	匹配数量	725	824	1261	544	319	1	362	4036	−2752
		Hits@1	0.12	0.03	0.37	0.06	0.04	0.00	0.05	0.16	
	OntoEA	匹配数量	372	334	403	316	128	0	34	1587	−607
		Hits@1	0.25	0.43	0.38	0.23	0.19	—	0.10	0.31	
	FuAlign	匹配数量	435	362	518	564	197	0	11	2087	−341
		Hits@1	0.32	0.51	0.44	0.47	0.28	—	0.09	0.42	
MSOI-Align-Mix	GCN-Align	匹配数量	583	113	198	188	130	0	10	1222	838
		Hits@1	0.85	0.87	0.90	0.79	0.82	—	0.30	0.84	
	BootEA	匹配数量	377	312	276	241	113	0	1	1320	**1106**
		Hits@1	0.82	0.93	0.99	0.96	0.96	—	1.00	**0.92**	
	OntoEA	匹配数量	251	55	108	239	32	0	0	685	555
		Hits@1	0.85	0.92	1.00	0.91	0.94	—	—	0.90	
	FuAlign	匹配数量	329	204	285	372	107	0	0	1297	997
		Hits@1	0.81	0.90	0.94	0.88	0.95	—	—	0.88	

● MSOI-Align-Mix 明显优于基于表示学习的基准方法: 对比 MSOI-Align 的各类变体和 5 种基于表示学习的模型, 可以发现 MSOI-Align 的 Hits@1 和收益均有明显提升. 而且所有基准方法的收益均为负值, 即对齐错误数量高于正确的数量.. 因为基准方法只计算表征向量相似度, 没有充分考虑实体本身的名称相似度和类别的信息, 而 MSOI-Align-Mix 对实体本体 (即类别) 一致性做了强制约束, 保证只有实体的本体一致时才能匹配, 同时也考虑了实体本身名称的相似度, 综合了更多维度的信息. 其次, 基准模型中 MtransE 和 TransEdge 的 Hits@1 仅为 0.03 和 0.07, 说明这 2 种方法基本没有找到等价的实体. 在标注它们的结果时, 也发现很多实体对完全不存在相似之处, 说明学习到的向量表示质量较差. 因此, 在 MSOI-Align-Mix 的实验中, 我们去除了这 2 种方法, 只对其他 4 种方法进行了进一步验证.

● MSOI-Align-Mix 类型的方法中 Mix_BootEA 表现最优: Mix_BootEA 的 Hits@1 为 0.92, 收益为 1106, 在所有 MSOI-Align 类型方法中均最大. 相对于直接使用 BootEA 得到的表示向量的相似度, Hits@1 从 0.31 提升至 0.92. 当方法的 Hits@1 和收益越大, 说明对齐结果中正确实体的数量和占比越多, 进一步也保证了对齐后大规模知识图谱的质量. 同时, 也能发现尽管 Mix_OntoEA 和 Mix_FuAlign 的 Hits@1 都表现较好, 分别为 0.90 和 0.88, 但是得到的匹配实体数量较少, 导致收益较低, 分别为 555 和 997.

● 本体一致性和融合语义信息对于实体对齐的结果有重要影响: 对比所有基准模型, 可以发现 OntoEA 和 FuAlign 相关方法的 Hits@1 都相对更高. 这是因为 OntoEA 模型在训练时对本体的一致性加了约束, 虽然不能完全保证对

齐实体的本体一样, 但是比其他方法有明显改善, 也证明了本体信息的价值. FuAlign 在表示学习的过程中融合多维度信息, 其中包括根据实体名称转换的向量表示. MSOI-Align 对本体一致性做了强制约束, 同时直接使用莱文斯坦比衡量实体名称相似度, 以简单有效的方式保证对齐的准确率.

4.4.2 验证融合语义信息有效性的消融实验

在两图谱对齐的过程中, 我们综合考虑了 通过表示学习方法得到的实体向量表示和实体名称信息的相似度. 为了验证这种融合语义信息的有效性, 设计了两组消融实验: (1) MSOI-Align-Literal: 计算实体相似度的时候只考虑实体名称的莱文斯坦比, 即公式 (2) 中的 α=1; (2) MSOI-Align-Embed_{x}: 计算实体相似度的时候只考虑实体表示的余弦相似度, 即公式 (2) 中的 α=0, 实体表示通过基准方法 x 获得. 和主实验保持一致, x 可为 GCN-Align、BootEA、OntoEA 和 FuAlign.

实验结果如表 5 所示, Mix_BootEA 在所有方法中仍然优势明显. 相对于只考虑实体名称相似度的 Literal 方法, Hits@1 提升了 17.95%, 收益增加 5.03%. 同时, 可以发现尽管 Embed_OntoEA 和 Mix_OntoEA 的 Hits@1 都表现较好, 分别为 0.77 和 0.90, 但是得到的匹配实体数量较少, 导致收益较低, 仅为 574 和 555. FuAlign 也存在类似的情况. 也发现名称实体信息十分重要, Literal 的 Hits@1 为 0.78, 高于所有 4 种 MSOI-Align-Embed 的方法, 这说明只考虑实体名称相似度就能获得较好的表现. 因此, 在可以获取实体名称的情况下应该充分利用该信息.

表 5 考虑不同维度实体信息的对齐结果

类别		指标	疾病	药物	生产商	症状	治疗方式	科室	食物	汇总统计	收益
MSOI-Align_Literal		匹配数量	443	756	128	496	41	2	15	1 881	1 053
		Hits@1	0.66	0.79	0.95	0.83	1.00	1.00	0.07	0.78	
MSOI-Align-Embed	GCN-Align	匹配数量	686	376	310	253	159	0	4	1 788	626
		Hits@1	0.63	0.77	0.65	0.74	0.61	—	0.00	0.68	
	BootEA	匹配数量	613	211	591	453	154	3	6	2031	755
		Hits@1	0.68	0.53	0.71	0.76	0.63	1.00	0.00	0.69	
	OntoEA	匹配数量	343	110	228	301	94	0	0	1 076	574
		Hits@1	0.74	0.83	0.82	0.72	0.81	—	—	0.77	
	FuAlign	匹配数量	412	258	374	443	145	0	0	1 632	766
		Hits@1	0.69	0.87	0.76	0.68	0.72	—	—	0.73	
MSOI-Align-Mix	GCN-Align	匹配数量	583	113	198	188	130	0	10	1 222	838
		Hits@1	0.85	0.87	0.90	0.79	0.82	—	0.30	0.84	
	BootEA	匹配数量	377	312	276	241	113	0	1	1 320	**1 106**
		Hits@1	0.82	0.93	0.99	0.96	0.96	—	1.00	**0.92**	
	OntoEA	匹配数量	251	55	108	239	32	0	0	685	555
		Hits@1	0.85	0.92	1.00	0.91	0.94	—	—	0.90	
	FuAlign	匹配数量	329	204	285	372	107	0	0	1 297	997
		Hits@1	0.81	0.90	0.94	0.88	0.95	—	—	0.88	

4.4.3 验证 LLM 有效性的消融实验

为了验证 LLM 在实体对齐任务中作为筛选工具的有效性, 从 MSOI-Align-Mix 中分别去除掉使用 LLM 的两个部分, 以及完全不使用 LLM, 即: (1) MSOI-Align (–LLM 知识筛选), 在两图谱实体对齐时直接选择相似度最高的实体, 不用 LLM 辅助筛选; (2) MSOI-Align (–三元闭包检验), 多源图谱对齐时, 在没有 LLM 判断的情况下不对三元闭包性质进行检验, 即不解决对齐错误传递问题; (3) MSOI-Align-Mix (–All_LLM), 同时去除 LLM 知识筛选和三元闭包检验. 采用表现较好的 BootEA、FuAlign 这 2 个方法进行实验. 结果如表 6 所示, 去除 LLM 知识筛选后, BootEA、FuAlign 的 Hits@1 从 0.92、0.88 下降至 0.87、0.81, 收益从 1 106、997 下降至 1 015、934, 说明只根据相似度的数值大小选择匹配的实体是不够的, LLM 能够帮助从相差不大的候选结果中选择更匹配的实体, 提升对齐的准确性. 同时, 去掉三元闭包检验, 说明本方法保留了一些不合理的匹配结果, 使得保留的实体对数量更

多, 但是 BootEA、FuAlign 的 Hits@1 也是从 0.92、0.88 下降至 0.88、0.83, 收益从 1 106、997 下降至 1 074、909. 这一结果充分证明了利用三元闭包理论解决对齐错误传递的必要性.

表 6　不同阶段去除 LLM 辅助的对齐结果

类别	方法	指标	疾病	药物	生产商	症状	治疗方式	科室	食物	汇总统计	收益
MSOI-Align-Mix (–LLM知识筛选)	BootEA	匹配数量	404	323	287	265	102	0	2	1 383	1 015
		Hits@1	0.79	0.91	0.94	0.85	0.88	—	0.50	0.87	
	FuAlign	匹配数量	359	275	353	416	114	0	3	1 520	934
		Hits@1	0.76	0.77	0.88	0.82	0.79	—	0.33	0.81	
MSOI-Align-Mix (–三元闭包检验)	BootEA	匹配数量	395	337	298	266	124	0	2	1 422	1 074
		Hits@1	0.83	0.85	0.98	0.86	0.90	—	0.50	0.88	
	FuAlign	匹配数量	317	254	302	381	116	0	3	1 373	909
		Hits@1	0.79	0.83	0.90	0.81	0.85	—	0.33	0.83	
MSOI-Align-Mix (–All_LLM)	BootEA	匹配数量	514	408	631	354	206	0	10	2 123	−585
		Hits@1	0.48	0.27	0.45	0.15	0.36	—	0.10	0.36	
	FuAlign	匹配数量	385	436	392	470	135	0	6	1824	244
		Hits@1	0.52	0.64	0.53	0.61	0.44	—	0.17	0.57	
MSOI-Align-Mix	BootEA	匹配数量	377	312	276	241	113	0	1	1 320	**1 106**
		Hits@1	0.82	0.93	0.99	0.96	0.96	—	1.00	**0.92**	
	FuAlign	匹配数量	329	204	285	372	107	0	0	1 297	997
		Hits@1	0.81	0.90	0.94	0.88	0.95	—	—	0.88	

4.4.4　超参数 α 敏感度分析

在 MSOI-Align-Mix_{x}模型中, 通过参数 α 调整实体名称相似程度在实体相似度计算中的占比, 即公式 (2). 这里以 Mix_BootEA 方法为基础, 进一步调整 α={0, 0.25, 0.5, 0.75, 1.0}, 对比其 Hits@1 和收益. 如图 7 所示, 随着 α 增大, Hits@1 呈现出先上升后下降的趋势, 在 0.25 时达到最高; 收益出现一定的波动, 因为该指标同时受到 Hits@1 和匹配数据量双方面的影响. 综合来看, 仍然是在 α=0.25 时取得较好的结果.

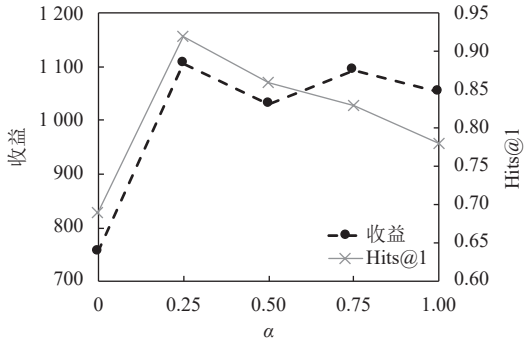


图 7　不同 α 下 MSOI-Align-Mix_BootEA 的收益及 Hits@1

4.4.5　Hits@k 和 MRR 指标评价结果比较

在 MSOI-Align 的两图谱对齐过程中, 针对图谱中的实体, 使用 LLM 从另一个图谱中筛选保留了最可能对齐的一个实体, 因此主要计算了 Hits@1 指标. 这里可以放松条件, 通过生成候选对齐序列计算 Hits@k 和 MRR 指标. 具体来说, 在两图谱对齐过程中只要求 LLM 将选出的可能匹配的实体进行排序. 之后在多源图谱对齐时, 针对任一实体, 可将来自其他不同图谱的对齐候选集放在一起进行请求 LLM 进行二次排序. 此时, 根据正确对齐的实体出现的顺序可以计算得到 Hits@k 指标和 MRR 指标. 以 BootEA 作为表示学习基准方法, 对比了它与 MSOI-Align-Mix_BootEA 的结果.

如表 7 所示, 可以发现 MSOI-Align-Mix_BootEA 在 3 个 Hits@k (k=1, 3, 5) 和 MRR 指标上均有明显提升, 特别是 Hits@5 为 0.96, 说明在 Top-5 个候选对齐实体列表中基本已经可以包含正确对齐的实体. 而且相比 BootEA 方法本身, MSOI-Align-Mix_BootEA 通过融合实体名称信息和借助 LLM 进行排序调整取得明显效果.

表 7　实体对齐任务 Hits@k 和 MRR 指标结果

方法	Hits@1	Hits@3	Hits@5	MRR
BootEA	0.17	0.31	0.58	0.29
MSOI-Align-Mix_BootEA	0.73	0.90	0.96	0.82

4.4.6　使用不同 LLM 的结果比较

在以上实验均采用 GPT-3.5-Trubo 进行辅助. 本节尝试测试不同的 LLM 进行实体对齐任务的表现. 考虑到虽然有很多工作尝试使用 LLM 与知识图谱进行组合[44], 但是针对 LLM 的幻觉问题[45]和是否能够学到长尾知识[46], 仍然存在较多担忧. 因此, 我们先对比了目前主流的 LLM 模型在医疗实体对齐标注问题上和人工标注的一致性. 具体来说, 将所有人工标注的对齐实体数据聚合起来, 通过提示 (prompt) 用 LLM 再次标注. 如表 8 所示, 对比了 5 个 LLM 的结果. 其中, GPT-3.5-Turbo 的一致性为 88.85%, 在实体对齐任务中已经取得了较好的表现. 在国内的 LLM 中, Qwen-2.5 和 DeepSeek-V2 的一致性均超过了 GPT-3.5-Turbo, 特别是 DeepSeek-V2 的一致性最高, 达到 91.70%. 同时, 也发现 GPT-4.0-Turbo 相对较差, 这可能和标注的数据是中文且为医疗专业领域相关.

表 8　不同 LLM 进行实体对齐标注与人工标注的一致性 (%)

方法	与人工标注的一致比例
GPT-3.5-Turbo	88.85
GPT-4.0-Turbo	87.67
GLM3-Turbo	84.84
Qwen-2.5	90.47
DeepSeek-V2	**91.70**

据此结果, 本文选择了一致性最优, 且调用 API 价格最低的 DeepSeek-V2 作为 LLM 进行实体对齐实验的验证. 针对表现最优的 MSOI-Align-Mix_BootEA 方法, 对比 GPT-3.5 和 DeepSeek-V2 的结果, 如表 9 所示. 可以发现 DeepSeek-V2 对齐的 Hits@1 有少量提升 (从 0.92 提升至 0.93), 收益提升较明显 (从 1 006 提升至 1 155), 这也说明 DeepSeek-V2 在实体对齐判断时纠正了一些 GPT-3.5 判断出错的情况.

表 9　采用 GPT-3.5-Turbo 和 DeepSeek-V2 作为辅助进行实体对齐结果

LLM	指标	疾病	药物	生产商	症状	治疗方式	科室	食物	汇总统计	收益
GPT-3.5-Turbo	匹配数量	377	312	276	241	113	0	1	1 320	1 006
	Hits@1	0.82	0.93	0.99	0.96	0.96	—	1.00	0.92	
DeepSeek-V2	匹配数量	365	317	294	238	132	0	1	1 347	**1 155**
	Hits@1	0.84	0.93	0.98	0.97	0.98	—	1.00	**0.93**	

4.4.7　知识图谱应用示例

KnowledgeDA[18]提出了一种基于领域知识图谱实现文本数据增强的统一范式, 其核心思路就是识别出任务数据集的医疗实体, 并根据知识图谱的实体和三元组关系设计相应的增强策略, 从而提升小样本情况下的具体文本任务的表现. 同时, 它也强调了图谱质量对于文本增强效果的影响. 特别地, 它使用到了我们的一个知识图谱 QAKG 进行实验, 主要是对 CMID 和 KUAKE-QIC 这两个中文医疗问诊数据进行分类. 为了验证对齐后的知识图谱 CMKG 具有更强的能力, 分别使用 4 个独立的知识图谱和对齐后的 CMKG 作为 KnowledgeDA 方法内置的图谱, 然后对比在 CMID 和 KUAKE-QIC 这 2 个分类任务上的准确率.

结果分别如图 8 和图 9 所示, 纵坐标表示是在分类任务上的准确率, 横坐标表示 KnowledgeDA 使用的知识图谱, None 表示没有进行增强. 两个图对比来看, 使用图谱进行文本增强后的表现都优于没有增强的表现. 因为 CHIP2021 图谱数据包含的实体数量较少, 质量相对不高, 增强的效果相对较差. 对齐之后的图谱相对更大, 包含的实体和三元组信息更丰富, 增强后的准确率相对来说也更高. 我们开源的 OpenCMKG 是除了 CPubMedKG 外 3 个图谱融合得到, 可以发现相对于 4 个图谱融合的 CMKG 增强效果略有下降, 但仍然优于单个图谱的表现. 这也证明了进行多个图谱对齐的意义和价值.

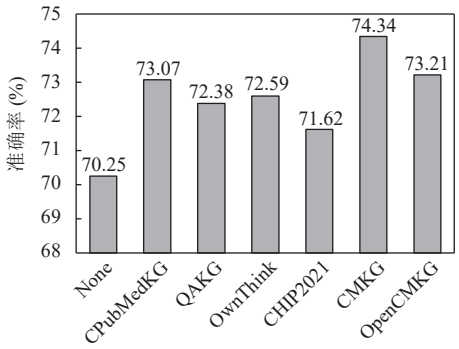


图 8 CMID 基于不同 KG 的数据增强表现

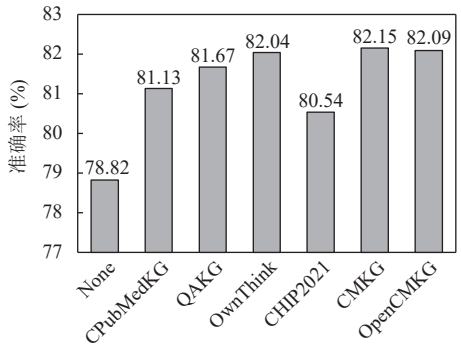


图 9 KUAKE-QIC 基于不同 KG 的数据增强表现

4.5 实例分析

本文找出正确和错误匹配的实体对, 分别列于表 10 和表 11 中. 在表 10 中, 发现存在一对名称完全不同, 但是等价的实体: 赛乐特和盐酸帕罗西汀片, 它是一种治疗抑郁症的药物. 同时, 详细分析了一些错误匹配的例子, 可以看到表 11 中的错误基本具有一定的相关性, 可能属于同一疾病的两个子类, 或存在比较相似的三元组关系, 比如不同药物 (拉坦前列素滴眼液和曲伏前列素滴眼液), 但是用于治疗同一类疾病. 后续可进一步分析错误样例完善方法.

表 10 正确匹配的实体对示例

实体类型	实体1	实体2
疾病	HCV感染	传染性肝炎
	重度脑性昏迷	重度昏迷
	不全性肠梗阻	不完全肠梗阻
症状	腹部胀痛	腹胀
	强制性思维	强迫思维
	心跳缓慢	心率缓慢
治疗方式	针灸推拿康复治疗	针刺配合推拿治疗
	康复训练物理治疗	物理康复治疗
	超声洁治或者刮治	超声洁治
药物	葡醛酸钠注射液	注射用葡醛酸钠
	匹伐他汀钙片	匹伐他汀钙
	赛乐特	盐酸帕罗西汀片

表 11 错误匹配的实体对示例

实体类型	实体1	实体2
疾病	气管支气管结核	气管支气管炎
	近端肾小管酸中毒	远端肾小管酸中毒
	原发性心脏淋巴瘤	继发性心脏淋巴瘤
症状	骨损	骨疼
	全血细胞减少	白细胞减少
治疗方式	药物溶栓	手术取栓
	介入动脉化疗栓塞	肝动脉化疗栓塞
药物	诺氟沙星滴眼液	氧氟沙星滴眼液
	拉坦前列素滴眼液	曲伏前列素滴眼液

5 总 结

本文针对多源医疗知识图谱对齐问题, 综合实体语义和本体信息计算实体间相似性, 并提出了多图谱对齐的错误对齐累积问题, 基于三元闭包理论使用 LLM 进行辅助筛选. 相对于之前的方法, MSOI-Align 确保了更高的实体对齐准确率, 有效地保证了可以获得对齐的大型图谱的质量. 与 OntoEA 相比, 我们验证了本体一致的强约束效

果明显. 对于中文知识图谱, 还发现实体本身名称具有很强的价值, 而常用的方法往往忽略了这一信息, 从而导致了极大的损失. 总的来说, MSOI-Align 在多医疗知识图谱对齐中具有一定的优势, 并为解决相关问题提供了有价值的思路.

当然, MSOI-Align 也存在一定的局限性. 首先, 在计算相似度时考虑了实体名称的相似度, 但存在一些医学实体本身名称完全不同, 但表达相同含义的情况. 例如, COVID-19 和新冠病毒名称完全不同, 但指代的是同一种疾病. 因此, MSOI-Align 可能比较容易遗漏这种匹配关系. 而且, 本文通过两两知识图谱的组合进行对齐, 如果知识图谱数量较多, 则组合数量也会增加, 导致模型训练需要更长的时间. 这可能限制了方法的可扩展性和效率. 在未来的研究中, 可以进一步改进和优化这些方面, 以提高对齐结果的准确性和效率. 随着 LLM 的飞速发展和更新迭代, 特别是经过大量高质量中文、甚至是医疗领域专业数据训练的 LLM 的出现和应用, 本方法的性能也会得到较大提升.

References:

[1] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 2017, 8(3): 489–508. [doi: 10.3233/SW-160218]

[2] Guo QY, Zhuang FZ, Qin C, Zhu HS, Xie X, Xiong H, He Q. A survey on knowledge graph-based recommender systems. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(8): 3549–3568. [doi: 10.1109/TKDE.2020.3028705]

[3] Li FL, Chen HH, Xu GH, Qiu T, Ji F, Zhang J, Chen HQ. AliMeKG: Domain knowledge graph construction and application in e-commerce. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. ACM, 2020. 2581–2588. [doi: 10.1145/3340531.3412685]

[4] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 793.

[5] Yang ZL, Qi P, Zhang SZ, Bengio Y, Cohen W, Salakhutdinov R, Manning CD. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 2369–2380. [doi: 10.18653/v1/D18-1259]

[6] Elhammadi S, Lakshmanan LVS, Ng R, Simpson M, Huai BX, Wang ZF, Wang LJ. A high precision pipeline for financial knowledge graph construction. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: ACL, 2020. 967–977. [doi: 10.18653/v1/2020.coling-main.84]

[7] Tang J, Zhang J, Yao LM, Li JZ, Zhang L, Su Z. ArnetMiner: Extraction and mining of academic social networks. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Las Vegas: ACM, 2008. 990–998. [doi: 10.1145/1401890.140200]

[8] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 2004, 32(S1): D267–D270. [doi: 10.1093/nar/gkh061]

[9] Zhang ZH, Liu HL, Chen JY, Chen X, Liu B, Xiang YJ, Zheng YF. An industry evaluation of embedding-based entity alignment. In: Proc. of the 28th Int'l Conf. on Computational Linguistics: Industry Track. ACL, 2020. 179–189. [doi: 10.18653/v1/2020.coling-industry.17]

[10] Färber M, Bartscherer F, Menne C, Rettinger A. Linked data quality of dbpedia, freebase, OpenCyc, wikidata, and YAGO. Semantic Web, 2018, 9(1): 77–129. [doi: 10.3233/sw-170275]

[11] Demartini G. Implicit bias in crowdsourced knowledge graphs. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 624–630. [doi: 10.1145/3308560.3317307]

[12] Yu S, Yuan Z, Xia J, Luo SX, Ying HY, Zeng SH, Ren JY, Yuan HY, Zhao ZY, Lin YC, Lu KM, Wang J, Xie YT, Shum HY. BIOS: An algorithmically generated biomedical knowledge graph. arXiv:2203.09975, 2022.

[13] Yan JH, Zong CQ, Xu JA. Nested entity recognition approach in Chinese medical text. Ruan Jian Xue Bao/Journal of Software, 2024, 35(6): 2923–2935 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6927.htm [doi: 10.13328/j.cnki.jos.006927]

[14] Yang YJ, Xu B, Hu JW, Tong MH, Zhang P, Zheng L. Accurate and efficient method for constructing domain knowledge graph. Ruan Jian Xue Bao/Journal of Software, 2018, 29(10): 2931–2947 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5552.htm [doi: 10.13328/j.cnki.jos.005552]

[15] Sun ZQ, Zhang QH, Hu W, Wang CM, Chen MH, Akrami F, Li CK. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc. of the VLDB Endowment, 2020, 13(12): 2326–2340. [doi: 10.14778/3407790.3407828]

[16] Zhang TC, Tian X, Sun XH, Yu MH, Sun YH, Yu G. Overview on knowledge graph embedding technology research. Ruan Jian Xue

Bao/Journal of Software, 2023, 34(1): 277–311 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6429.htm [doi: 10.13328/j.cnki.jos.006429]

[17] Huang H, Dong YX, Tang J, Yang HX, Chawla NV, Fu XM. Will triadic closure strengthen ties in social networks? ACM Trans. on Knowledge Discovery from Data, 2018, 12(3): 30. [doi: 10.1145/3154399]

[18] Ding RQ, Han X, Wang LY. A unified knowledge graph augmentation service for boosting domain-specific NLP tasks. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023. 353–369. [doi: 10.18653/v1/2023.findings-acl.24]

[19] Zeng KS, Li CJ, Hou L, Li JZ, Feng L. A comprehensive survey of entity alignment for knowledge graphs. AI Open, 2021, 2: 1–13. [doi: 10.1016/j.aiopen.2021.02.002]

[20] Tang J, Li JZ, Liang BY, Huang XT, Li Y, Wang KH. Using Bayesian decision for ontology mapping. Journal of Web Semantics, 2006, 4(4): 243–262. [doi: 10.1016/j.websem.2006.06.001]

[21] Jiménez-Ruiz E, Cuenca Grau B. LogMap: Logic-based and scalable ontology matching. In: Proc. of the 10th Int'l Semantic Web Conf. on the Semantic Web. Bonn: Springer, 2011. 273–288. [doi: 10.1007/978-3-642-25073-6_18]

[22] Suchanek FM, Abiteboul S, Senellart P. PARIS: Probabilistic alignment of relations, instances, and schema. Proc. of the VLDB Endowment, 2011, 5(3): 157–168. [doi: 10.14778/2078331.2078332]

[23] Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2787–2795.

[24] Chen MH, Tian YT, Yang MH, Zaniolo C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: AAAI Press, 2017. 1511–1517.

[25] Sun ZQ, Hu W, Li CK. Cross-lingual entity alignment via joint attribute-preserving embedding. In: Proc. of the 16th Int'l Semantic Web Conf. on the Semantic Web. Vienna: Springer, 2017. 628–644. [doi: 10.1007/978-3-319-68288-4_37]

[26] Sun ZQ, Hu W, Zhang QH, Qu YZ. Bootstrapping entity alignment with knowledge graph embedding. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 4396–4402. [doi: 10.24963/ijcai.2018/611]

[27] Wang ZC, Lv QS, Lan XH, Zhang Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 349–357. [doi: 10.18653/v1/D18-1032]

[28] Xu K, Wang LW, Yu M, Feng YS, Song Y, Wang ZG, Yu D. Cross-lingual knowledge graph alignment via graph matching neural network. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 3156–3161. [doi: 10.18653/v1/P19-1304]

[29] Sun ZQ, Huang JC, Hu W, Chen MH, Guo LB, Qu YZ. TransEdge: Translating relation-contextualized embeddings for knowledge graphs. In: Proc. of the 18th Int'l Semantic Web Conf. on the Semantic Web. Auckland: Springer, 2019. 612–629. [doi: 10.1007/978-3-030-30793-6_35]

[30] Zhang QH, Sun ZQ, Hu W, Chen MH, Guo LB, Qu YZ. Multi-view knowledge graph embedding for entity alignment. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI, 2019. 5429–5435. [doi: 10.24963/ijcai.2019/754]

[31] Qi ZY, Zhang ZH, Chen JY, Chen X, Xiang YJ, Zhang NY, Zheng YF. Unsupervised knowledge graph alignment by probabilistic reasoning and semantic embedding. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. Montreal: IJCAI, 2021. 2019–2025. [doi: 10.24963/ijcai.2021/278]

[32] Liu X, Hong HY, Wang XH, Chen ZY, Kharlamov E, Dong YX, Tang J. SelfKG: Self-supervised entity alignment in knowledge graphs. In: Proc. of the 2022 ACM Web Conf. Lyon: ACM, 2022. 860–870. [doi: 10.1145/3485447.3511945]

[33] Sun ZQ, Cui YN, Hu W. Lifelong representation learning of multi-sourced knowledge graphs via linked entity replay. Ruan Jian Xue Bao/Journal of Software, 2023, 34(10): 4501–4517 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6887.htm [doi: 10.13328/j.cnki.jos.006887]

[34] Zhang JD, Li J. Knowledge graph embedding combining with hierarchical type information. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3331–3346 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6295.htm [doi: 10.13328/j.cnki.jos.006295]

[35] Xiang YJ, Zhang ZH, Chen JY, Chen X, Lin ZX, Zheng YF. OntoEA: Ontology-guided entity alignment via joint knowledge graph embedding. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. ACL, 2021. 1117–1128. [doi: 10.18653/v1/2021.findings-acl.96]

[36] Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu YX, Miller A. Language models as knowledge bases? In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 2463–2473. [doi: 10.18653/v1/D19-1250]

[37] Ziems N, Yu WH, Zhang ZH, Jiang M. Large language models are built-in autoregressive search engines. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023. 2666–2678. [doi: 10.18653/v1/2023.findings-acl.167]

[38] Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y. COMET: Commonsense Transformers for automatic knowledge graph construction. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 4762–4779. [doi: 10.18653/v1/P19-1470]

[39] West P, Bhagavatula C, Hessel J, Hwang J, Jiang LW, Le Bras R, Lu XM, Welleck S, Choi Y. Symbolic knowledge distillation: From general language models to commonsense models. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: ACL, 2022. 4602–4625. [doi: 10.18653/v1/2022.naacl-main.341]

[40] Choi B, Ko Y. Knowledge graph extension with a pre-trained language model via unified learning method. Knowledge-based Systems, 2023, 262: 110245. [doi: 10.1016/j.knosys.2022.110245]

[41] Zhao XD, Ouyang SQ, Yu ZG, Wu M, Li L. Pre-trained language models can be fully zero-shot learners. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Toronto: ACL, 2023. 15590–15606. [doi: 10.18653/v1/2023.acl-long.869]

[42] Xu CW, Xu YC, Wang SH, Liu Y, Zhu CG, McAuley J. Small models are valuable plug-ins for large language models. In: Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 283–294. [doi: 10.18653/v1/2024.findings-acl.18]

[43] Wang CX, Huang ZH, Wan Y, Wei JY, Zhao JZ, Wang PH. FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs. Information Fusion, 2023, 89: 41–52. [doi: 10.1016/j.inffus.2022.08.002]

[44] Pan SR, Luo LH, Wang YF, Chen C, Wang JP, Wu XD. Unifying large language models and knowledge graphs: A roadmap. IEEE Trans. on Knowledge and Data Engineering, 2024, 36(7): 3580–3599. [doi: 10.1109/TKDE.2024.3352100]

[45] Ji ZW, Lee N, Frieske R, Yu TZ, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12): 248. [doi: 10.1145/3571730]

[46] Kandpal N, Deng HK, Roberts A, Wallace E, Raffel C. Large language models struggle to learn long-tail knowledge. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 641.

附中文参考文献:

[13] 闫璟辉, 宗成庆, 徐金安. 中文医疗文本中的嵌套实体识别方法. 软件学报, 2024, 35(6): 2923–2935. http://www.jos.org.cn/1000-9825/6927.htm [doi: 10.13328/j.cnki.jos.006927]

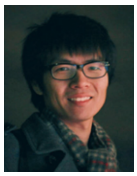
[14] 杨玉基, 许斌, 胡家威, 仝美涵, 张鹏, 郑莉. 一种准确而高效的领域知识图谱构建方法. 软件学报, 2018, 29(10): 2931–2947. http://www.jos.org.cn/1000-9825/5552.htm [doi: 10.13328/j.cnki.jos.005552]

[16] 张天成, 田雪, 孙相会, 于明鹤, 孙艳红, 于戈. 知识图谱嵌入技术研究综述. 软件学报, 2023, 34(1): 277–311. http://www.jos.org.cn/1000-9825/6429.htm [doi: 10.13328/j.cnki.jos.006429]

[33] 孙泽群, 崔员宁, 胡伟. 基于链接实体回放的多源知识图谱终身表示学习. 软件学报, 2023, 34(10): 4501–4517. http://www.jos.org.cn/1000-9825/6887.htm [doi: 10.13328/j.cnki.jos.006887]

[34] 张金斗, 李京. 一种结合层次化类别信息的知识图谱表示学习方法. 软件学报, 2022, 33(9): 3331–3346. http://www.jos.org.cn/1000-9825/6295.htm [doi: 10.13328/j.cnki.jos.006295]

丁瑞卿(1999－), 女, 博士, 主要研究领域为医疗数据挖掘, 知识图谱应用.



王乐业(1987－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为普适计算, 群智感知, 城市计算.

赵俊峰(1974－), 女, 博士, 研究员, CCF 高级会员, 主要研究领域为大数据分析, 大语言模型, 知识工程, 数据治理, 城市计算.