双语翻译知识图谱的研究构建与应用

林晶晶[1]，彭楷廷[2*]

[1]浙大宁波理工学院外国语学院　浙江宁波
[2]浙大宁波理工学院数据学院　浙江宁波

【摘要】知识图谱作为结构化知识表示的重要技术，在自然语言处理等领域应用广泛。本文系统梳理了该领域的研究现状与发展趋势，针对现有方法在跨语言实体对齐和翻译知识整合上的不足，提出一种基于双语翻译关系建模的 BTRM 模型。该模型通过整合三个模块的协同工作，提升双语知识图谱的构建质量。实验结果表明：BTRM 模型在构建难度、翻译准确性和跨语言应用等方面优于传统知识图谱构建方法，尤其在双语言语义对齐和知识融合方面表现突出。研究不仅为双语翻译知识图谱的构建提供了新的理论框架和技术路径，也为机器翻译、跨语言信息检索等应用提供了更可靠支持。

【关键词】双语翻译；知识图谱；BTRM 模型；语言学习

【基金项目】本文为浙江省“十四五”第二批本科省级教学改革项目“基于知识图谱的英语翻译教学改革与实践探索”（项目编号：JGBA2024692）、浙江省外文学会专题研究项目“翻译地理学视域下的革命文本生产——以甬籍译家群为中心”（项目编号：ZWYB2025024）、浙大宁波理工学院 2023 年度校级教学改革研究项目“知识图谱驱动下的翻译课程体系构建探究”（项目编号：NBTJG-202306）、2025 年宁波市教育科学规划课程“教育家精神引领下高校青年教师专业发展的行动路径研究”（项目编号：2025YGH026）的阶段性研究成果

【收稿日期】2025 年 5 月 15 日　　【出刊日期】2025 年 6 月 16 日　　【DOI】10.12208/j.ssr.20250252

Research, construction and application of bilingual translation knowledge graph

Jingjing Lin[1], Kaiting Peng[2]*

[1]School of International Studies, NingboTech University, Ningbo, Zhejiang
[2]School of Computer Science and Data Engineering, NingboTech Technology, Ningbo, Zhejiang

【**Abstract**】As an important technology for structured knowledge representation, knowledge graphs are widely used in natural language processing and other fields. This paper systematically combs the research status and development trends in this field. Aiming at the shortcomings of existing methods in cross-lingual entity alignment and translation knowledge integration, a BTRM (Bilingual Translation Relationship Modeling) model based on bilingual translation relationship modeling is proposed. The model improves the construction quality of bilingual knowledge graphs through the collaborative work of three integrated modules. Experimental results show that the BTRM model outperforms traditional knowledge graph construction methods in terms of construction difficulty, translation accuracy and cross-lingual applications, especially in bilingual semantic alignment and knowledge fusion. This study not only provides a new theoretical framework and technical path for the construction of bilingual translation knowledge graphs, but also offers more reliable support for applications such as machine translation and cross-lingual information retrieval.

【**Keywords**】Bilingual translation; Knowledge graph; BTRM model; Language learning

引言

在全球化进程持续推进的当下，世界各国在诸多领域的交流合作愈发紧密，跨语言交流的需求正以前所未有的态势增长。Feng 认为知识图谱可以用头实体、关系、尾实体的三元组表示[1]，Wang 认为知识图谱应结合可持续发展化的理念[2]，Hu 认为知识图谱可以结

*通讯作者：彭楷廷

合深度强化学习的教育资源[3]。在当今科技发展现状下，将知识图谱与 LLM 大模型相结合无疑是一种流行趋势。“知识图谱与大模型各有优缺点，KG 可以提供额外的、结构化的、高质量的知识，其可解释性知识有助于理解模型输出；而 LLM 可以自动从文本数据中提取知识，从而降低 KG 的构建和维护成本等，两者协同工作可以克服各自缺陷，发挥互补优势”[4]。在新型知识引入层面例如：“时态知识图谱（Temporal Knowledge Graph，TKG）是一种重要的知识表示方法，能够记录包含时间信息的事实，并以结构化形式进行数据存储。它是有效记录现实世界事物变化行为的工具，具备处理互联网中多种类关联数据的能力”[5]。“多模态数据”（如文本、图像、音频等）的丰富性催生了多模态知识图谱（Multi-modalKnowledge Graphs, MMKG）这一新兴领域。多模态知识图谱不仅继承了传统知识图谱连接实体与关系的结构化优势，还通过融合多种模态的信息，如文本描述、图像特征、音频信号等，大大增强了知识图谱的表达能力和实用性”[6]能够揭示事物间复杂的语义关系[7]。双语翻译作为跨越语言障碍、实现顺畅沟通的核心手段，其准确性与效率已然成为影响跨语言交流质量的关键因素。在传统翻译模式中，人工翻译虽能在一定程度上保障翻译质量，但效率较低且成本高昂，难以满足海量信息快速翻译的需求；机器翻译虽能实现快速输出，然而在准确性和语义理解方面仍存在较大提升空间。

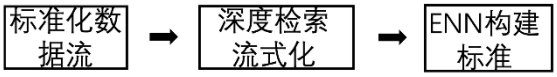
1 研究现状

相较于通用知识图谱在多领域的广泛研究与应用，双语翻译知识图谱研究尚处于起步阶段。曾有学者认为外语学科的知识图谱体现在多语种，多模态，知识与能力并重，专家知识与资源采集结合[8]。Jiang 认为应整合语言、图像、音频等多维信息来创建语料库[9]。在学术数据库中检索发现，专门针对双语翻译知识图谱构建与应用的高质量研究论文数量较少，公开可用的大规模双语翻译知识图谱资源更是稀缺。这主要归因于双语翻译知识图谱构建面临独特且复杂的技术难题[10]，需要秉承学科核心素养的三维条件[11]、跨语言处理、知识表示与推理等多领域技术协同攻关，研究难度较大。

当前，双语翻译知识图谱面临着诸多棘手问题。其一，如何使得双语翻译知识图谱避免结构单一化[12]，能够兼容语句与术语？其二，目前传统的单体系模型默认汉语为一级学科[13]，无法完整输入双语翻译两种语言的数据，难以满足双语翻译知识图谱构建需求；其三，翻译本身的复杂性也带来难题。人们在认识客观事物过程中，“习惯性”地将自身的主体地位定义为主宰地位[14]，并且不同语言文字语法构造不同[15]，这导致单个语言或句子往往对应多种翻译，如何构建一个体系来合理涵盖这些多样的翻译可能性，也是双语翻译知识图谱构建过程中需要解决的关键问题。

2 研究方法

为了针对以上问题提出一个标准化解决方式，此方式能够专用于双语翻译知识图谱的构建以及后续应用层面。本文提出 BTRM 模型（Bilingual Translation Research Model）。如图 1 所示，该模型共分为三个模块：①标准化数据流、②深度检索流式化、③EUN 构建标准。该模型确保了初始数据从输入到完整双语翻译知识图谱体系的构建的标准化。



BTRM构建标准

图 1 BTRM 建构流程

2.1 标准化数据流

标准化数据流将创造语句，术语，语句中的术语，语言 1 系统，语言 2 系统等 5 种数据流分流体系。从广泛收集来的数据中构建单股数据流，采用多层数据筛选机制，首先筛选出术语与语句，再从语句中向量化划分出术语，我们将术语分为单术语、双术语、集成术语。例如：“人工智能领域的创新应用”一句中可以划分出：“人工智能”、“人工智能领域”、“人工智能领域的创新应用”等词，我们细化划分的维度来获取更多术语，在之后的筛查过程中去除重复，噪音等错误数据。去除联系度低的术语来提升精准度。然后为这些从语句中提取的术语和原语句附上父属关系的标签，这样从传统的知识图谱只能建术语库或者语句库到现在的实现两者结合，达到提升检索与联系度的效果。然后从所有的术语和语句中剥离出双语，并归纳为两种系统：语言一系统和语言二系统，这样一来我们将原始数据流分为了术语，语句，语句中的子术语，语言 1，语言 2 系统，这对于一开时就创建数据与节点的数据库有着系统划分明确的优势，在标准化数据流后，我们从该语句中向量切分的词组中提取所有术语，并与原语句创建父属节点关系。

2.2 深度检索流式化

在数据流分流后，我们引入 LLM 大模型与 Light RAG 检索来创新性提高构建效率，以此来提供更准确

的关系映射[16]。为了保证数据质量与适配前文提到的标准数据流规则，本研究分别采用两种规范化流式的处理方式，分别为开源的 Light RAG 深度检索流，以及传统的数据本体库构建流，学者可根据不同需求来选择合适的模型。传统的数据本体库构建流本文将不做探讨。检索增强生成（RAG）检索与生成式模型 Light RAG 从外部检索相关信息作为上下文输入模型以提高内容准确性和相关性，解决生成式模型在知识准确性、实时性及上下文衔接断层问题，适用于需大量事实性知识场景；Graph RAG 在 RAG 基础上增加图像处理且有较大优势；Light RAG 是 Graph RAG 的轻量级优化版本，通过优化检索过程、减少计算资源消耗提高检索效率，能在保证质量前提下更快获取信息并兼容大规模数据集。目前 Light RAG 采用 LLM 大模型与 embedding 模型作为接口，Embedding 模型是一种将高维数据（如单词、句子、图像等）映射到低维连续向量空间的技术。在自然语言处理（NLP）、计算机视觉（CV）、推荐系统等多个领域都有重要应用。

为了计算出构建双语翻译知识图谱的最佳接口配置，我们对目前流行的 LLM 大模型与 embedding 模型进行了检测，包括对语言处理、向量分块、语言检索、响应时间、流式输出、节点数量、模型稳定性作为参考方向，以最后输出结果逻辑与层级度作为考量。将不同模型进行不同搭配批量化测试 20 次后，我们得到了以下数据如图 2 所示。

从中我们可以得到：采用 API 调用 deepseek-reasoner 模型结合本地 nomic-embed-test 模型呈现双语翻译知识图谱精准度最高，在参数层面，我们采用 LightRAG 官方给出的数据，如图 3 所示。

在初步自动化处理后，我们可以将 Light RAG 构建的本体库拆分。结合实际来将去噪去重后的数据按照传统知识图谱本体结构进行构建和组织存储。随后进行标准化操作，例如统一数据格式、术语命名等。这一过程能够确保数据的一致性和规范性，为后续知识图谱的构建奠定本体和关系基础。

如图 4 所示：从图中官方传统知识图谱模型与适配双语翻译知识图谱模型的单体传统知识图谱可视化中我们可以看出：逻辑性显著提升，节点间的关联层级更加清晰。使双语翻译的检索更为方便。在后续研究中，学者可以根据具体需求灵活选择这两种流式处理方式，以满足不同场景下双语翻译知识图谱构建的要求。

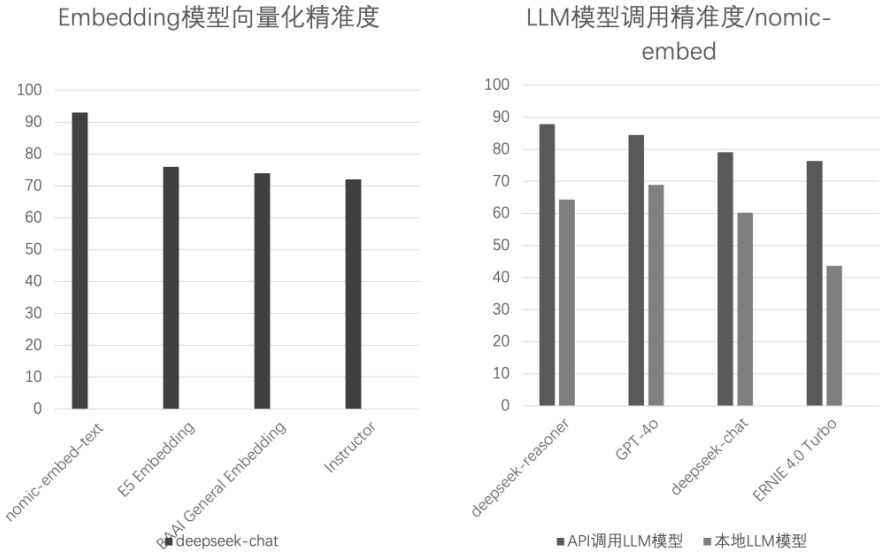


图 2 模型测试对比

Embedding模型选择	embedding_dim	max_token_size	运行框架
nomic-embed-text	768	8192	Ollama

图 3 模型调用参数

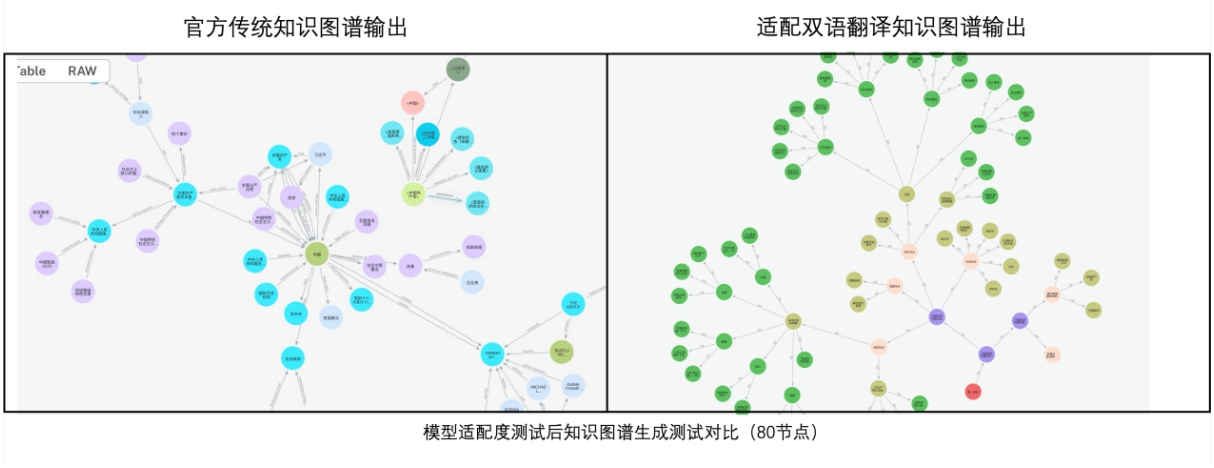


图 4 生成逻辑对比

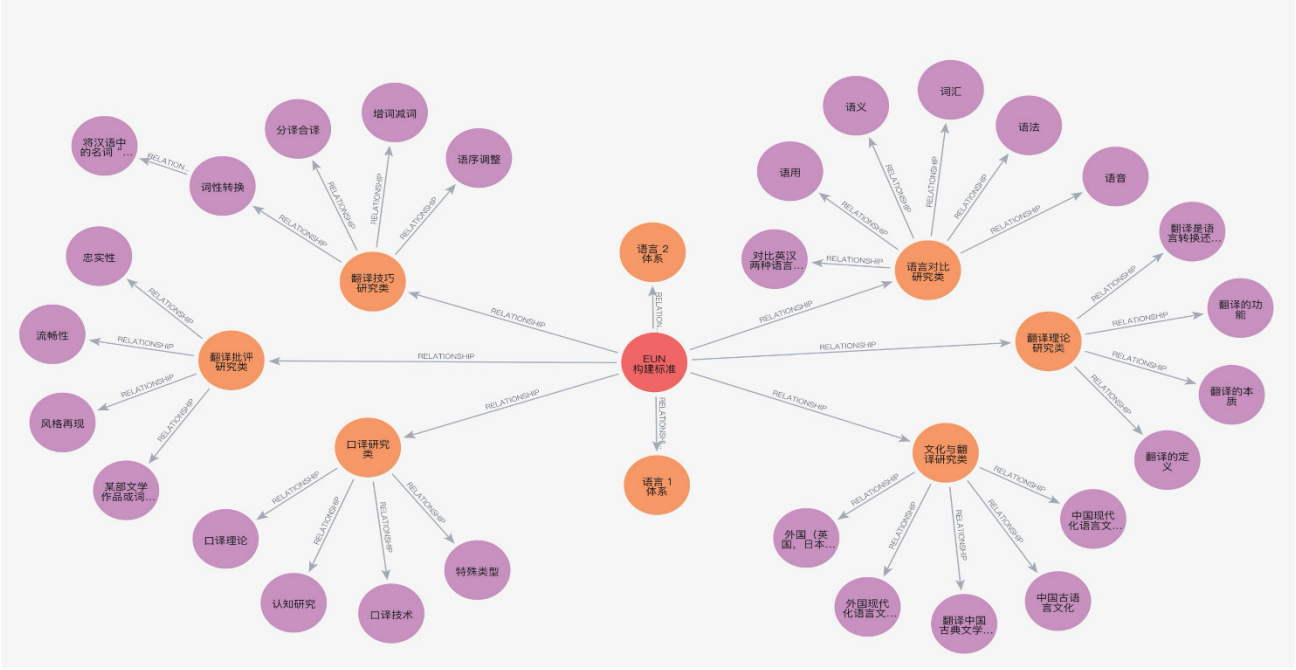


图 5 EUN 构建标准

2.3 EUN 构建标准

为了使双语翻译知识图谱的本体构建流程规范化统一化[17]，我们提出了 EUN 构建标准（Exact Unification Neo4j 构建标准），该标准确保了双语翻译体系的紧密配合和检索。该标准分为理论部分与应用部分。在构建理论部分中，我们对双语翻译知识图谱的构建流程与构建结果输出进行统一规范化管理，确保使用该模型系统构建双语翻译知识图谱的用户，其数据具有兼容性，实现知识图谱数据统一化。统一术语定义与命名规则，避免因术语不一致导致节点重复、语句属性不匹配等问题。如图 5 所示：EUN 构建标准定义了构建的两大体系：语言 1 体系，语言 2 体系。六大研究门类[18]构成双语翻译知识图谱构建实践系统。其中六大研究门类分为：翻译理论研究类，语言对比研究类，翻译技巧研究类，翻译批评研究类，口译研究类，文化与翻译研究类。

在构建应用部分中，双语翻译知识图谱构建实践系统可以简化数据库的导入流程，实现一键导入效果。在该系统下，我们开发了专门数据导入工具(CUN)，其工作原理如图 6 所示。

学者只要将经过处理和标准化的数据整理成指定格式，通过提示操作即可将数据快速导入到知识图谱

构建系统中。在未来，我们将持续优化此工具，通过增加 LLM 的 API 调用接口和联合接入 Light RAG 接口达到可以通过直接上传数据流而无需整理数据流分流就能输出双语翻译可视化图谱的成果。

CUN 工具构建流程

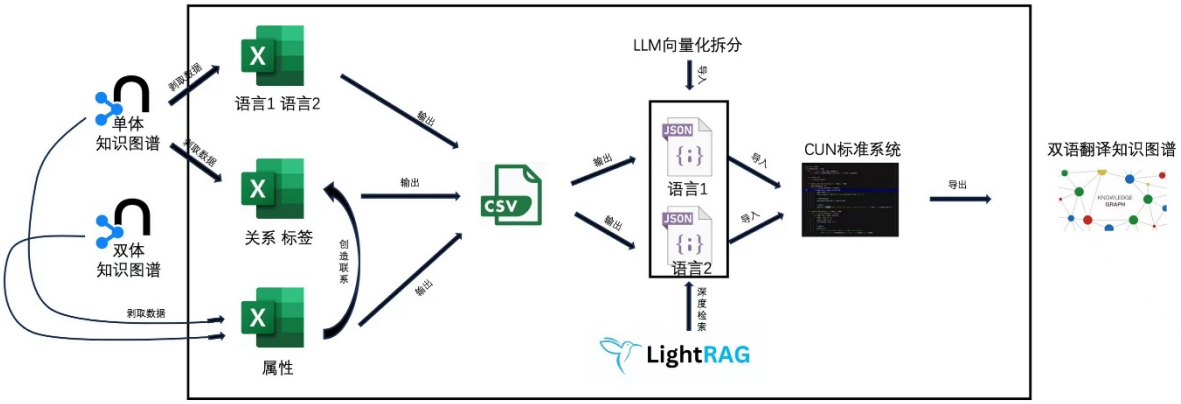


图 6 CUN 工具构建流程

3 研究结论

BTRM 模型为双语翻译知识图谱的构建与应用提供格式化框架，通过优化流程、简化复杂度，通过数据输入模块确保数据全面完整，深度检索模块清洗数据提升质量，联合知识图谱构建模块创新结构增强知识关联，EUN 构建标准模块实现数据统一、便捷导入及整合，形成完整高效解决方案推动构建工作。未来，该模型在跨语言信息检索中可提升学者获取多语言信息的效率与准确性，在机器翻译辅助场景为系统提供双语知识与技术支持、改善专业领域翻译质量，在语言学习与教学领域可作为直观系统的双语知识资源助力学习。

参考文献

[1] 冯志伟. 自然语言处理的重要资源:“知识图谱”[J]. 外语学刊,2021,(05):1-9.

[2] 王逸飞,于水. 可持续发展报告研究热点与前沿趋势——基于 Cite Space 知识图谱的可视化分析 [J]. 商业会计,2025,(05):105-109.

[3] 胡志敏. 基于知识图谱和深度强化学习的智慧教育资源推荐研究 [D]. 青岛大学,2024.

[4] 李晓理,刘春芳,耿劭坤. 知识图谱与大语言模型协同共生模式及其教育应用综述 [J/OL]. 计算机工程与应用,1-15 [2025-04-14]. https://kns-cnki-net.webvpn.nbt.edu.cn/kcms/detail/11.2127.TP.20250403.1313.006.html.

[5] 张强,高颖,任豆豆. 融合时态信息和图结构动态演化的知识图谱补全模型研究 [J/OL]. 数据分析与知识发现,1-14 [2025-04-14]. https://kns-cnki-net.webvpn.nbt.edu.cn/kcms/detail/10.1478.g2.20250403.1349.004.html.

[6] 崔好丽,赵震.多模态知识图谱补全研究综述 [J/OL]. 计算机技术与发展,1-10 [2025-04-14]. https://doi.org/10.20165/j.cnki.ISSN1673-629X.2025.0055.

[7] 杨道宇,王灿辉.论学科实践的教育学意义及其实现[J]. 中国教育学刊,2025,(04):51-56.

[8] 李佐文.外语学科知识图谱的特征、建构与应用[J]. 中国外语,2023, 20 (02):70-76.

[9] 蒋丽平,陈霞.粤剧汉英多模态双语语料库的构建与翻译应用研究[J]. 五邑大学学报（社会科学版), 2025, 27 (01):71-75+94.

[10] 原伟.基于多语种可比语料库的语言学领域多词术语抽取研究 [J].语料库研究前沿,2020,(00):139-148.

[11] 陈家宝.基于知识图谱和深度学习的课程推荐应用研究[D]. 西北师范大学,2024.

[12] 史威.论学科核心素养下的 “三维要件”[J]. 天津市教科院学报,2024, 36 (05):56-65.

[13] 冯志伟,李宇明,冉永平.中国语言学学科建设三问 [J]. 中国语言战略,2024, 11 (01):1-8.

[14] 钱佳,田晓明.论 “学科交叉” 与 “交叉学科”[J]. 江苏高教,2024(09):59-66.

[15] 陆俭明.关于中国语言学话语体系建设之思考 [J]. 长江学术,2024(02):6-9.

[16] 张为,肖巧玲,刘海江,等.大模型辅助的汉语文化负载词抽取与知识图谱构建 [J]. 数字图书馆论坛,2025, 21 (01):33-45.

[17] 刘敏,江文.教育知识图谱的概念模型设计研究 [J]. 中国多媒体与网络教学学报 (中旬刊), 2024(08):221-224.

[18] 赵小云,杨先卫,陈秋紫,等.知识结构视角下的教育知识图谱模型构建研究 [J]. 数字教育,2025, 11 (01):23-30.

版权声明：©2025 作者与开放获取期刊研究中心（OAJRC）所有。本文章按照知识共享署名许可条款发表。
http://creativecommons.org/licenses/by/4.0/

OPEN ACCESS