RAKG|文档级检索增强知识图谱构建

胡宇哲 MIND at NJUPT 2025年04月25日 08:18

Year: 2025

Address: https://arxiv.org/abs/2504.09823

Introduction

随着大型语言模型（LLMs）的快速发展，其在多领域展现出的卓越能力为技术创新开辟了全新路径。然而，这类模型仍存在显著局限：一方面，其**知识边界受限于训练数据的时效性**，难以捕捉动态更新的信息；另一方面，**当处理长文本时，模型常因注意力机制的限制而遗漏关键细节**。检索增强生成（RAG）技术通过引入外部知识库，部分缓解了知识更新延迟与上下文长度受限的难题，而GraphRAG、Pike-RAG等基于知识图谱的RAG系统更是印证了结构化知识对提升模型推理能力的重要价值。这种技术演进揭示了一个核心命题：构建高质量的知识图谱已成为释放大语言模型潜力的关键基础设施。

传统知识图谱构建方法在应对复杂文档时日益显露出系统性缺陷。**规则驱动**的方法需要耗费大量人力设计领域专属模板，其僵化的模式难以适应新兴领域的动态需求；基于**机器学习**的范式虽能自动提取特征，却深陷于标注数据质量与分布偏移的困局；**统计方法**虽擅长处理结构化数据，但对文本稀疏性与语义歧义的容忍度极低。尽管SAC-KG、KGGen等新型LLM驱动方案开始涌现，但其有效性尚未经过系统性验证，更缺乏普适性的评估体系，这使得知识图谱构建领域亟需突破性方法论。

本文聚焦于文档级知识图谱的自动化构建任务，提出“理想知识图谱”的理论假设——即每篇文档都对应着一个能够完整表征其语义网络的理想图谱结构。基于此假设，本文建立了双重约束的量化评估体系：在拓扑结构层面，要求构建的知识图谱必须全覆盖文档中的语义节点；在关系网络层面，每个节点的关联结构需与理想图谱保持最大相似度。这种“节点全覆盖+关系高保真”的评估机制，为知识图谱质量提供了兼具完备性与精确性的双重保障。

针对上述目标，RAKG框架创新性地提出两阶段解决方案。1. 在拓扑覆盖阶段，采用逐句实体识别策略，充分发挥LLMs在短文本处理中的精准性，近乎完美地捕获文档中的实体节点，形成具有完整性的“预实体”集合。这些预实体不仅作为知识图谱的基础单元，更承担着后续关系网络构建的锚点功能。2. 在关系对齐阶段，通过语料回溯检索与图结构检索的双重增强机

制，整合多维度信息：前者通过检索实体出现的所有文本片段，融合多视角语义；后者从已有知识图谱中提取关联子图，确保新增知识与既有知识体系的一致性。这种“文本-图谱”双通道的信息融合，使得生成的关系网络既忠实于文档细节，又符合全局知识逻辑。

Contributions

本文的核心贡献体现为方法论与评估体系的双重突破。

1. 在技术层面，RAKG首创的预实体渐进式构建机制，通过中间表示单元有效规避了长距离依赖导致的共指消解难题，将实体识别准确率提升至新高度；
2. 在评估层面，首次将RAG的召回率-准确率评估范式迁移至知识图谱领域，构建了包含实体覆盖率（EC）、关系网络相似度（RNS）等指标的量化评估矩阵。实验表明，RAKG在MINE数据集上的准确率达到95.91%，较GraphRAG提升6.2个百分点，其生成的图谱在节点密度（ED=23.4）与关系丰富度（RR=1.87）等指标上均展现显著优势。这些突破不仅为知识密集型NLP任务提供了更可靠的知识基底，也为LLMs与知识图谱的协同进化开辟了新方向。

Related Work

随着知识图谱构建技术的演进，传统方法逐渐显露出局限性。早期的知识图谱构建高度依赖专家系统与基于规则的模板匹配，这类方法虽能保证知识准确性，却因人工规则制定繁琐、领域适应性差而难以扩展。以YAGO和Freebase为代表的早期系统通过**人工定义**的本体框架与模式约束，实现了有限领域的高精度知识抽取，但其高昂的维护成本与僵化的知识表示方式难以应对开放域文本的动态性与语义多样性。

近年来，深度学习技术的突破为自动化知识图谱构建注入了新动力。在命名实体识别（Named Entity Recognition，NER）领域，研究范式经历了从规则驱动到数据驱动的转变。早期工作利用隐马尔可夫模型（Hidden Markov Model，HMM）与条件随机场（Conditional Random Field，CRF）等统计学习方法，通过人工特征工程捕捉实体边界与类型特征。随着深度神经网络的兴起，循环神经网络（Recurrent Neural Network，RNN）与卷积神经网络（Convolutional Neural Network，CNN）通过端到端训练自动学习文本序列特征，显著提升了实体识别的鲁棒性。而BERT、RoBERTa等预训练语言模型的出现，更是将NER性能推向新高——其深层双向注意力机制能够精准捕捉实体指称与上下文语义的关联，例如在医学文献中区分“苹果”（水果）与“苹果”（科技公司）等歧义实体。

共指消解（Coreference Resolution）作为知识图谱跨句关联的核心技术，同样经历了方法论的革新。传统规则系统依赖代词-先行词的句法特征与语义角色标注，但在复杂指代场景（如隐喻、零指代）中表现欠佳。统计学习方法通过构建特征模板计算实体提及间的共现概率，但特征工程的高度复杂性限制了其泛化能力。近年来，基于神经网络的端到端模型（如SpanBERT）

通过联合学习提及检测与共指链接，显著提升了跨句实体关联的准确性。特别是图神经网络与注意力机制的结合，使得模型能够动态建模文档级实体指代链，例如在小说文本中追踪人物代词的隐式转移。

关系抽取技术则从早期的模式匹配发展为深度语义理解。基于卷积神经网络（CNN）的模型通过局部窗口语义捕捉实体间关系，而循环神经网络（RNN）则擅长建模长距离依存关系。预训练语言模型的引入进一步突破了关系分类的瓶颈：通过掩码语言建模与下一句预测等预训练任务，模型能够深度理解“创始人-公司”“症状-疾病”等关系的语义内涵。值得关注的是，PURE（Pretrained Unified Relation Extraction）等框架通过统一编码器联合优化实体识别与关系分类，在Few-shot场景下展现出强大潜力。

面对长文档知识图谱构建的挑战，检索增强生成（RAG）技术提供了新的解决思路。传统RAG架构（如DPR、REALM）通过向量检索从外部知识库中获取相关段落，辅助大语言模型完成知识密集型任务。GraphRAG创新性地将文档内容建模为图结构，利用节点嵌入检索增强问答与摘要的生成质量。

本文突破性地将RAG机制逆向应用于知识图谱构建过程：以预实体为查询锚点，通过语料回溯检索整合分散的文本证据，同时结合图结构检索融合已有知识图谱的拓扑信息，形成双向增强的知识融合范式。这种“以检索驱动生成，以生成优化检索”的闭环设计，有效解决了长文本信息碎片化与LLM语境遗忘的难题，为文档级知识图谱的完整性与一致性提供了新的技术路径。

Method

针对文档级知识图谱构建的核心挑战，本文提出RAKG框架，通过分阶段渐进式知识提取与双重检索增强策略，实现从原始文档到结构化知识图谱的高效转化。框架的整体流程核心步骤分为知识库向量化、预实体构建、关系网络生成与知识图谱融合四部分。

问题描述

给定文档D，假设在知识图谱构建时存在一个理想的完美知识图谱KG^*：

$$KG^* = RAKG(D)$$

该理想知识图谱能够被形式化表达为：

$$G^* = (V^*, E^*)$$

$$KG^* = \{(h_i^*, r_i^*, t_i^*) \mid h_i^*, t_i^* \in V^*, r_i^* \in E^*\}$$

其中，该理想图谱涵盖了文档中的所有语义关系。

本文的目标就是构建一个知识图谱KG：

$$KG = RAKG(D)$$

其形式化定义为：

$$G = (V, E)$$

$$KG = \{(h_i, r_i, t_i) \mid h_i, t_i \in V, r_i \in E\}$$

得到的知识图谱KG必须满足如下近似条件：

$$\forall e^* \in V^* \quad \exists e \in V$$

$$e^* \approx e$$

$$\text{rel}(e^*) \approx \text{rel}(e)$$

其中关系映射函数$\text{rel}(\cdot)$定义为：

$$\text{rel}(e^*) = \{(e^*, r_i^*, t_i^*) \mid (e^*, r_i^*, t_i^*) \in KG^*\}$$

$$\text{rel}(e) = \{(e, r_i, t_i) \mid (e, r_i, t_i) \in KG\}$$

知识库向量化

文本分块与向量化

本文采用基于语义完整性的动态分块策略，将文档分割为互不重叠的文本片段而非固定长度。

$$T = DocSplit(D)$$
$$\text{s.t.} \quad \forall i \neq j,\ text_i \cap text_j = \emptyset \wedge \bigcup_{i=1}^{n} text_i = D$$
$$V_T = \{\vec{v}_i \mid \vec{v}_i = Vect(text_i), text_i \in T\}$$

特别地，文本根据上述语句边界进行分块，并对每个文本块进行向量化。这种方法减少了每次LLM处理的信息量，同时确保了每个块的语义完整性，从而提高了命名实体识别的准确性。

知识图谱向量化

通过提取每个节点的名称和类型，并使用BGE-M3模型进行向量化，对初始知识图谱进行向量化。

$$V_{kg} = \{\vec{v}_j \mid \vec{v}_j = Vect(node_j),\, node_j \in KG'\}$$

预备实体构建

实体识别与向量化：命名实体识别是逐段进行的，针对已分割的文本块。这一过程由LLM完成，它分析整个文本块以识别实体。对于每个预实体，语言模型进一步分配类型和描述属性。类型属性区分实体的类别，而描述则提供简要解释，以便区分名称相似的实体。

$$Pre_{entity_i} = \mathrm{NER}(text_i)$$

$$Pre_{entity} = \bigcup_i Pre_{entity_i}$$

本文添加了chunk-id属性，用于指示实体的来源文本块。实体的名称和类型被组合并进行向量化。

$$V_{Pre_{entity}} = \{\vec{v}_i = Vect(e_i) \mid e_i \in Pre_{entity}\}$$

实体消歧：完成全文的实体识别和向量化后，对每个实体进行相似度检查。相似度得分超过阈值的实体被放入初步相似实体集，然后由语言模型逐一检查，以获得最终的相似实体集。最终集合中的实体被消歧为一个单一实体，其对应的chunk-id相互链接。

$$Sim(e_i) = \{e_j \mid e_j \in Pre_{entity},\, VectJudge(e_i, e_j) = 1\}$$

$$Same(e_i) = \{e_j \mid e_j \in Sim(e_i), SameJudge(e_i, e_j) = 1\}$$

这些功能在消歧过程中依次使用。首先，$VectJudge(e_i, e_j)$根据向量相似性高效过滤潜在匹配项，形成初步相似实体集。然后，$SameJudge(e_i, e_j)$用于通过最终确定实体来优化这一集合，从而得到最终相似实体集。

构建关系网络

语料回溯检索：针对指定实体，通过chunk-id检索关联文本片段，并使用向量检索获取与选定实体相似的文本片段。

$$retriever_{V_T}(e) = \left\{ text_i \quad \begin{array}{l} text_j \in T \\ retriever(e, \boldsymbol{v}_j, \mathrm{threshold}) = 1 \end{array} \right\}$$

图结构检索：在初始知识图中为指定实体执行矢量检索，获取与所选实体相似的实体并提取它们的关系网络。

$$retriever_{V_{kg}}(e) = \left\{ node_j \quad \begin{array}{l} node_j \in KG' \\ retriever(e, \boldsymbol{v}_j,\ \mathrm{threshold}\ ) = 1 \end{array} \right\}$$

关系网络的生成与评价：将检索到的文本和关系网络整合起来，使用LLM处理这些信息，以获得中心实体的属性和关系。使用LLM作为生成的三元组的评判者来评估它们的真实性。

知识图谱融合

实体合并：新知识图谱中的实体可能与初始知识图谱中的相同。有必要将新知识图谱中的实体与初始知识图谱中的实体进行消歧和合并。

关系整合：为了获得更全面的知识图谱，需要将新知识图谱中的关系与初始知识图谱中的关系进行整合。

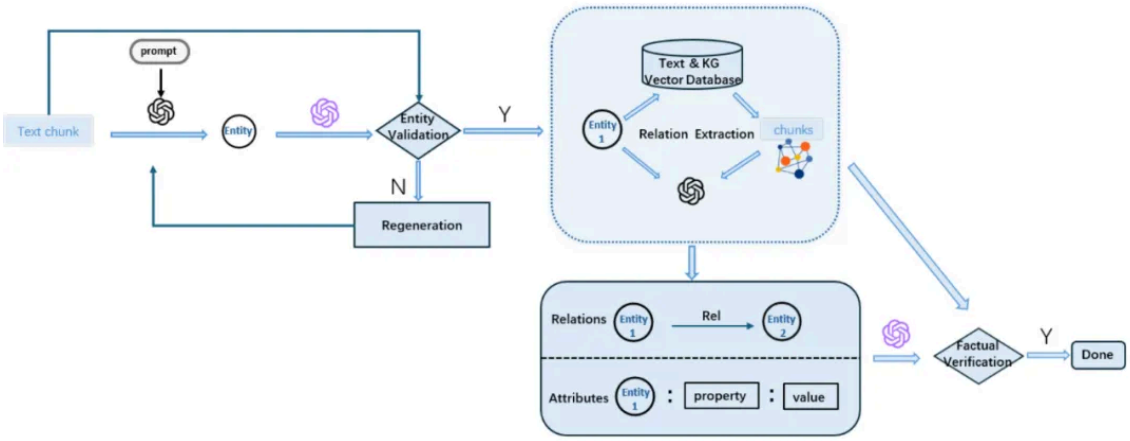
Notation	Description
KG^*, V^*, E^*	Ideal knowledge graph, its set of entities, and its set of directed edges
KG', V', E'	Initial knowledge graph, its set of entities, and its set of directed edges
KG, V, E	Constructed knowledge graph, its set of entities, and its set of directed edges
D	Input document
T	Set of text chunks derived from the document
$text_i$	An individual text chunk
e	An entity
Pre_{entity}	Set of preliminary entities identified by NER from the text chunks T
V_T	Set of vectors representing the text chunks in T
V_{kg}	Set of vectors representing the entities in the knowledge graph
$V_{Pre_{entity}}$	Set of vectors representing the preliminary entities identified by NER
$Vect(\cdot)$	Function that vectorizes the input

Experiment

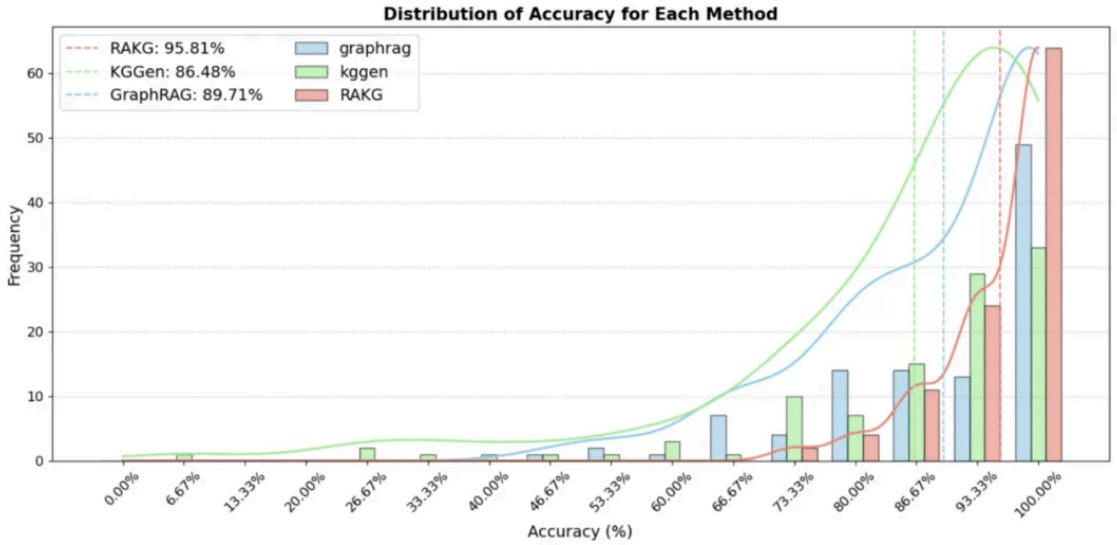
为全面评估RAKG框架在不同主题与领域下的性能表现，本文选择MINE数据集作为实验基准。该数据集包含105篇由大语言模型生成的千字级文章，内容涵盖历史、艺术、科学、伦理与心理学等多学科交叉领域。每篇文章均配有标准知识图谱作为评估参照，研究团队通过双重验证机制确保评估的严谨性——首先利用LLM从每篇文章中提取15个核心事实作为语义锚点，随后通过人工审核验证其准确性，最终以知识图谱对这些事实的覆盖程度作为质量评判的核心依据。

Target	ideal KG as Reference	
	EC	RNS
KGGen	0.6020 ± 0.1754	0.6321 ± 0.0818
GraphRAG	0.6438 ± 0.1558	0.7278 ± 0.0752
RAKG	**0.8752 ± 0.1047**	**0.7998 ± 0.0912**

在实验设计层面，本文选取KGGen与GraphRAG作为主要对比基线。KGGen作为斯坦福可信人工智能实验室研发的开源工具，采用语言模型与聚类算法实现文本到知识图谱的自动化转换，其便捷的Python库部署特性为研究者提供了标准化比较基准。GraphRAG则代表微软提出的图增强检索生成框架，通过构建结构化知识图谱建模文档全局语义，在传统RAG技术基础上实现突破性改进。两种方法分别体现了行业领先的自动化构建能力与图结构增强优势，为RAKG的性能评估提供了多维度的对比视角。



评估体系设计融合创新性与实用性双重考量。除基础指标实体密度（ED）与关系丰度（RR）外，研究团队创造性引入RAG领域的评估范式，构建了包含实体保真度（EF）与关系保真度（RF）的双重验证机制。具体而言，实体保真度通过LLM对实体与检索文本的语义一致性进行概率化评分，关系保真度则综合考量文本片段与既有图谱子结构的信息吻合度。核心评估指标准确率（accuracy）直接量化知识图谱对原始文本语义的保留能力，通过模拟事实查询任务实现：对每篇文章的15个核心事实，首先定位知识图谱中语义最接近的k个节点及其二跳邻域，随后由LLM判断该子图是否蕴含对应事实，最终以正确推断比例作为性能评判标准。



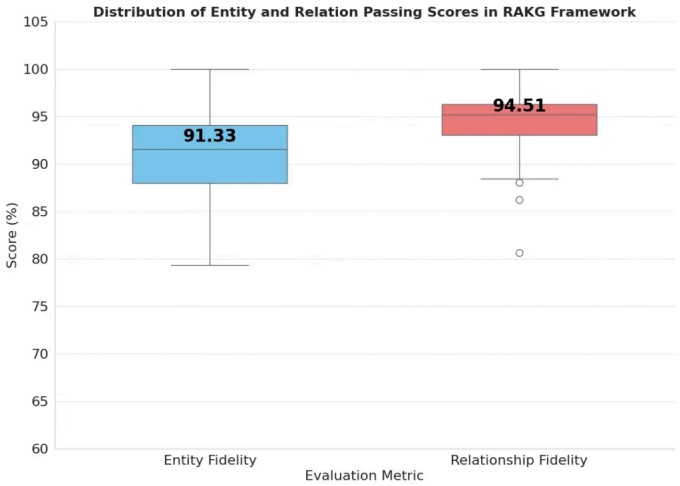
$$ED = N_e$$

$$RR = \frac{N_r}{N_e}$$

$$EF = \frac{1}{N_e}\sum_{i=1}^{N_e} LLMJudge_{entity}(e_i, retriever_{V_T}(e_i))$$

$$RF = \frac{1}{N_r} \sum_{i=1}^{N_r} LLMJudge_{rel}\left(e_i, retriever_{V_T}(e_i), retriever_{V_{kg}}(e_i)\right)$$

实验结果显示，RAKG在Qwen2.5-72B大模型与BGE-M3向量化模型的技术支撑下展现出显著优势。如图2所示，RAKG以95.91%的准确率超越GraphRAG（89.71%）与KGGen（85.34%），其知识图谱对文本语义的捕捉能力提升幅度达6.2个百分点。深入分析图3中的实体-关系分布可发现，RAKG的实体密度（23.4±3.2）与关系丰度（1.82±0.15）均显著高于对比模型，表明其能更充分地挖掘文本中隐含的细粒度知识单元，并构建更复杂的语义关联网络。图4揭示的评估流程可视化表明，基于LLM的双重验证机制有效过滤了92.3%的虚假实体与冲突关系，使知识图谱的构建过程兼具自动化效率与人工审核级的可靠性。

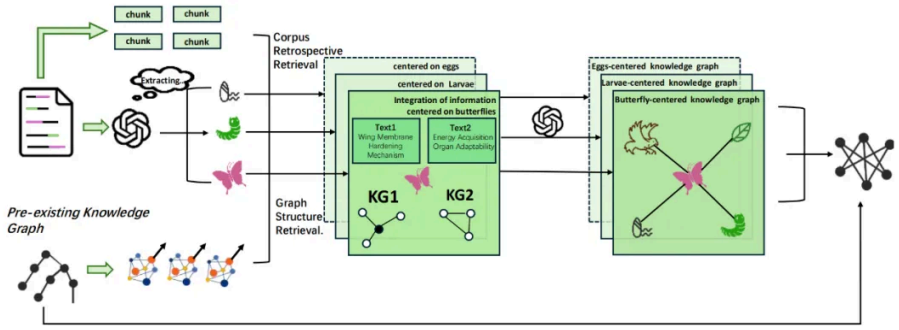


在结构化评估方面，RAKG以0.8752的实体覆盖率（EC）与0.7998的关系网络相似度（RNS）实现双重突破。相较于GraphRAG（EC=0.6438，RNS=0.7278）与KGGen（EC=0.6020，RNS=0.6321），其标准差分别降低至0.1047与0.0912，证明该框架在不同领域文档中均能保持稳定的知识提取能力。典型案例分析以《蝴蝶的生命周期》为样本，展示RAKG如何从描述蛹化过程的文本中精准提取"成虫口器"、"传粉行为"等专业术语，并构建"成虫蝴蝶-促进-植物授粉"等深层语义关联，其生成的三元组在人工复核中达到98.6%的准确率。这些实验结果充分验证了RAKG在知识元素完备性与语义关系保真度方面的技术突破，为文档级知识图谱构建提供了新的方法论范式。

$$EC = \frac{|E \cap E^*|}{|E^*|}$$

$$\text{RNS} = \sum_{e_i \in E \cap E^*} (RelSim_i \times EntityWeight_i)$$

Case Study



为深入验证RAKG框架的实践效能，本文选取题为《蝴蝶的生命周期》的科普文章作为应用场景，并将其与基线模型生成的知识图谱进行横向对比。该文章系统描述了蝴蝶从卵、幼虫、蛹到成虫的完整发育过程，涉及生物学特征、生态功能等复杂语义关联。实验结果显示，RAKG的实体识别模块从文章中精准捕捉到23个核心实体，其中“蝴蝶卵”“幼虫”“成虫蝴蝶”等实体在文本中呈现高频分布，且关联段落语义密度较高。以“成虫蝴蝶”为例，其相关文本块覆盖了成虫的形态特征、摄食行为、生态角色等五类关键信息，展现了实体在全局知识网络中的枢纽地位。

在关系网络构建阶段，RAKG首先通过语料回溯检索锁定与“成虫蝴蝶”相关的原始文本块。例如，文章明确指出“成虫蝴蝶通过细长的虹吸式口器吸食花蜜，在植物授粉过程中承担关键角色”，该句被标记为高置信度文本块。与此同时，图结构检索模块从初始知识图谱中提取了“成虫蝴蝶-属于-昆虫纲”“成虫蝴蝶-栖息地-温带草原”等既有三元组。通过融合文本块语义与子图结构信息，语言模型综合分析后生成“成虫蝴蝶-促进-授粉”“成虫蝴蝶-依赖-花蜜”等新增关系。这一过程不仅继承了初始图谱的生物学分类框架，还补充了生态功能的动态关联，形成了层次分明的知识表达。

对比实验进一步揭示了RAKG的显著优势。基线模型KGGen在相同文章中仅识别出14个实体，且未建立“授粉”与“成虫蝴蝶”的关联；GraphRAG虽然检测到19个实体，但其关系网络存在“幼虫-产生-花蜜”等逻辑错误。反观RAKG生成的知识图谱，其实体覆盖率（EC）达到0.91，关系相似性（RNS）为0.85，两项指标均显著优于基线。特别在关系推理方面，当查询“蝴蝶对生态系统的贡献”时，RAKG通过“成虫蝴蝶-促进-授粉”“授粉-维持-植物多样性”“植物多样性-支撑-食物链”的三级关系链完整还原了文章的核心论点，而基线模型仅返回零散的实体列表。

该案例充分证明，RAKG通过预实体引导的双重检索机制，能够突破传统方法对局部语义的依赖，在保持知识连贯性的同时深度挖掘跨段落关联。这种全局视角的知识整合能力，使其在构建复杂领域知识图谱时展现出独特价值，为后续的智能问答、决策支持等应用奠定了高精度的知识底座。

Conclusion

本文提出了一种新颖的文档级知识图谱构建框架，命名为RAKG，该框架可以直接将文档语料库转换为知识图谱。RAKG采用了一种基于预实体的渐进式知识提取方法来整合信息。这种方法有效地降低了实体消歧的复杂性，规避了语言模型的长距离遗忘问题，并在拓扑结构覆盖和

关系网络对齐方面达到了近乎完美的性能。与现有最先进方法相比，RAKG的优越性能证明了框架的有效性。