融合逻辑约束的全局-局部协同长文本知识抽取方法*

康生国[1,2], 秦继伟[1,2†], 张 莉[1,2]
(1. 新疆大学 计算机科学与技术学院,新疆 乌鲁木齐 830017;
2. 新疆多模态智能处理与信息安全工程技术研究中心,新疆 乌鲁木齐 830017)

摘 要:长文本知识抽取中,语义依赖跨度大、实体指代分散,易导致常规滑动窗口切分破坏跨段语义联系,影响抽取结果的完整性与一致性. 针对这一问题,本文提出一种融合逻辑约束的全局-局部协同长文本知识抽取方法(Global-Local Collaborative Knowledge Extraction,GlocalKE). 该方法首先在全文范围内利用共指消解整合分散的实体表达,形成可跨窗口传递的全局语义索引;随后引入模糊证据匹配,对生成结果与原文支撑片段间的一致性进行校验,以降低生成式模型产生幻觉的风险;最后结合实体对齐与双向软逻辑推理机制,在本体约束下对候选三元组进行统一校验和修正. 在DuIE-Doc和CS-Cn-50两个数据集上的实验表明,所提方法的$F1$值分别达到65.92%和42.55%,均优于对比基线方法,有效提升长文本知识抽取的准确性与逻辑一致性.

关键词:知识抽取;长文本建模;全局-局部协同;逻辑约束;大语言模型

DOI:10.13568/j.cnki.651094.651316.2026.02.07.0002
中图分类号:TP391.1 **文献标识码**:A **文章编号**:2096-7675(2026)05-0546-016
引文格式:康生国,秦继伟,张莉. 融合逻辑约束的全局-局部协同长文本知识抽取方法[J]. 新疆大学学报(自然科学版中英文),2026,43(5):546-561.
英文引文格式:Kang Shengguo, Qin Jiwei, Zhang Li. A logic-constrained global-local collaborative method for long-text knowledge extraction[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2026, 43(5): 546-561.

A Logic-Constrained Global-Local Collaborative Method for Long-Text Knowledge Extraction

Kang Shengguo[1,2], Qin Jiwei[1,2], Zhang Li[1,2]
(1. School of Computer Science and Technology, Xinjiang University, Urumqi Xinjiang 830017, China;
2. Xinjiang Multimodal Intelligent Processing and Information Security Engineering and Technology Research Center, Urumqi Xinjiang 830017, China)

Abstract: In long-text knowledge extraction, long-span semantic dependencies and dispersed entity references often cause conventional sliding-window segmentation to break cross-paragraph semantic links, thereby affecting the completeness and consistency of extraction results. To address this problem, this paper proposes GlocalKE (Global-Local Collaborative Knowledge Extraction), a logic-constrained global-local collaborative method for long-text knowledge extraction. The method first uses coreference resolution over the full text to integrate dispersed entity expressions and form a global semantic index that can be transferred across windows. It then introduces fuzzy evidence matching to verify the consistency between generated results and supporting spans in the original text, reducing the risk of hallucination in generative models. Finally, entity alignment and bidirectional soft logic reasoning are combined to check and revise candidate triples under ontology constraints. Ex-

* **收稿日期**:2026-02-07;**修回日期**:2026-06-07;**录用日期**:2026-06-09.
基金项目:新疆战略人才培养计划一流科技领军人才项目"跨媒体知识聚合方法研究与应用"(XJRC-2025-KJ-PY-KJLJ-046).
作者简介:康生国(1998—),男,博士生,从事知识图谱的研究,E-mail:ksg@stu.xju.edu.cn.
† **通信作者**:秦继伟(1978—),女,博士,教授,主要从事大数据知识工程的研究,E-mail:jwqin@xju.edu.cn.

periments on the DuIE-Doc and CS-Cn-50 datasets show that the $F1$ values of the proposed method reach 65.92% and 42.55%, respectively, and outperform those of the compared baseline methods, improving the accuracy and logical consistency of long-text knowledge extraction.

Key words：knowledge extraction; long-text modeling; global-local collaboration; logic-constrained; large language models

0　引言

知识抽取（Knowledge Extraction，KE）的目标是从非结构化文本中识别实体、关系及相关事实，并将其组织为结构化知识．作为知识图谱构建的基础环节，该任务已被应用于语义搜索[1]、智能问答[2]和决策支持[3]等场景．现有研究多面向句子级或短文本展开，通常默认关系判断所需线索可在有限上下文中获得[4]．但在学术文献、复杂系统手册等长文本中，相关信息往往分散在不同段落，实体间的联系不一定能够在局部语境内完整呈现[5]．如果依赖局部上下文进行抽取，跨段语义线索易被切分过程削弱，进而影响结果的完整性和一致性．因此，长文本知识抽取不仅要识别细粒度实体关系，还需考虑文档层面的语义依赖，以保证抽取结果在语义和事实上的连贯性[6]．

大语言模型（Large Language Models，LLMs）的发展，使知识抽取逐渐从传统判别式标注范式转向生成式建模框架[7]．这类方法在语义理解和知识表征方面表现出更强的能力[8]，在降低标注依赖、支持零样本迁移等方面具有一定优势．但在处理长程依赖捕捉及跨段落信息融合时存在明显结构性短板[9]，若关键语义碎片散落于文档的不同区间，模型难以建立跨段关联，导致信息整合不充分．

一方面，受显存限制和上下文窗口长度约束，长文本通常需要被切分为多个子序列后再分别处理，而这种切分会破坏跨窗口的实体指代链和上下文联系，造成语义信息割裂[10]．另一方面，生成式解码具有一定随机性，在缺少充分证据约束的情况下，模型易产生幻觉（Hallucination），生成原文并未明确支持的错误三元组[11–12]．与此同时，不同窗口内独立完成的抽取结果还可能引发实体归并偏差及结构性冲突，如类型层级混乱或父子关系倒置等问题，进而削弱知识图谱在整体层面的质量与可用性[13]．

针对上述问题，本文提出融合逻辑约束的全局-局部协同长文本知识抽取方法GlocalKE（Global-Local Collaborative Knowledge Extraction）．其模拟人类阅读“全域概览→局部精析→回溯验证”认知过程，首先在全文范围内进行共指消解，将分散出现的实体提及归并到统一标识下，为后续局部窗口提供完整语境信息．然后在局部抽取阶段加入模糊证据匹配机制，根据生成结果与原文片段间的对应关系筛除缺乏支撑的候选项，以提高抽取结果可追溯性并减轻幻觉干扰．在此基础上，利用融合类型惩罚与数值约束的混合度量方案完成实体对齐，并基于预定义本体实施双向软逻辑验证，对候选三元组作进一步修正．

通过上述全局建模、局部生成与后验修正的协同过程，GlocalKE能够对不满足事实约束或逻辑约束的三元组进行识别和修正，提升长文本知识抽取的准确性与整体一致性．本文主要贡献如下：1）提出一种面向长文本知识抽取的闭环式全局-局部协同框架GlocalKE，通过构建跨窗口实体共指索引，将全局语义锚点注入局部窗口推理，并在抽取后进行全文级实体对齐与一致性校验，从机制上缓解长文本切分导致的实体指代链断裂与语义割裂问题；2）提出一种证据感知的生成校验与一致性修正方法，在局部抽取阶段通过模糊证据匹配约束候选关系必须具有原文支撑，在全局阶段结合本体模式设计双向软逻辑推理，对关系方向冲突与类型不一致进行自适应修正，以提升生成结果的事实可靠性与结构一致性；3）实验分别在通用领域数据集DuIE-Doc和垂直领域数据集CS-Cn-50上进行验证，表明GlocalKE在精确率（Precision，P）、召回率（Recall，R）和$F1$值方面均优于基线方法；消融实验也说明全局索引、证据校验和逻辑推理模块对跨窗口语义建模及幻觉抑制具有支持作用．

1　相关工作

1.1　监督式长文本知识抽取方法

早期知识抽取研究主要围绕句子级任务展开，相关方法大体可分为流水线抽取和联合抽取两类[14]．其中流水线方法通常将实体识别与关系分类拆分为两个相互独立的环节，在一定程度上降低了建模难度，但由于缺乏跨子任务的信息共享，前一阶段产生的偏差往往被传递到后续步骤，进而带来较明显的误差累积问题（Er-

ror Propagation)[15]. 针对这一不足,CasRel[16]、TPLinker[17]等联合抽取模型尝试通过共享编码表示或统一标注空间,将实体识别与关系抽取纳入同一框架中完成,在句子级知识抽取场景下取得较好效果.

随着任务场景由句子扩展到篇章,如何建模长距离语义依赖逐渐成为研究重点. 较早的一类代表性方法是基于图神经网络(Graph Neural Network,GNN)的建模框架. LSR[6]、GAIN[18]和DRN[19]等通过构建由实体节点、提及节点及句子节点组成的全局异构图,对实体间可能存在的推理路径进行显式刻画,并借助图卷积网络完成信息传播与关系推断. 这类方法在增强跨句推理能力方面具有一定优势,但图结构构建过程较为复杂,同时对句法分析等外部工具存在较强依赖,因而在实际应用中面临较高的实现与计算成本.

不同于前述依赖图结构的建模思路,基于Transformer的序列建模方法无需显式构造复杂图表示,而是借助自注意力机制在编码过程直接刻画长文本中的上下文依赖,逐渐发展为长文本知识抽取的一条重要技术路径. ATLOP[20]在这一框架下引入自适应阈值机制与局部上下文池化策略,较好缓解多标签分类的阈值选择问题,并成为长文本关系抽取任务中具有代表性的基线模型. 此后,SSAN[21]等进一步围绕注意力机制展开改进,在自注意力层中显式融入实体共指信息与篇章结构特征,增强模型对篇章内部逻辑关联及组织结构的建模能力.

1.2 大语言模型驱动的生成式抽取

随着预训练语言模型不断向生成式建模发展,知识抽取研究逐渐由传统的判别式序列标注框架转向生成式语义映射范式. 相较于早期方法,大语言模型在语义理解和指令响应方面具有更强能力,故在低资源场景下表现出更好的迁移性与泛化潜力,一定程度上缓解传统方法对标注数据依赖较强的问题[22].

在生成式知识抽取的早期研究中,提示工程和上下文学习是使用较多的两种方式. GPT-3等工作表明,只要提供合适的任务说明和少量示例,预训练模型即使在零样本或低资源条件下,也能够完成实体识别、关系判定等基础抽取任务[23]. 之后,思维链(Chain-of-Thought,CoT)也逐渐被用于知识抽取[24]. 其特点是在输出结果前保留中间推理步骤,使模型按“实体定位→语境解析→关系推断”的顺序组织生成过程. 对于嵌套实体和重叠关系这类更易出现歧义的情形,该方法通常更有帮助.

为进一步压缩生成空间并提高输出结果的规范性,后续研究开始更多关注结构化指令微调(Instruction Tuning)和模式约束的引入. UIE[7]等通过构建统一的文本到结构表示映射,将实体抽取、关系抽取和事件抽取纳入同一序列生成框架,实现不同抽取任务间的迁移与共享. 与此同时,CodeKGC[25]、ChatIE[26]等则将知识抽取任务重构为多轮问答或类代码生成过程,借助更清晰的结构组织形式和更严格的语法约束,降低非规范三元组的生成概率. 此外,GoLLIE[27]等还尝试在指令中显式加入本体模式(Schema)定义,以增强模型面对未见实体类型时的零样本适应能力. 表1对UIE、ChatIE、GoLLIE等代表性方法与本文方法进行对比,展示现有方法与本文方法在建模方式和约束机制上的差异.

表 1 代表性方法与本文方法的差异分析

Table 1 Comparison between representative methods and the proposed approach

方法	主要建模对象	跨窗口实体链	原文证据校验	一致性修正	主要优势	主要局限
UIE[7]	统一结构生成	否	否	否	统一任务表达、跨任务共享	面向长文本时缺少跨窗口实体保持与全文规约
InstructUIE[22]	指令调优的统一信息抽取	否	否	否	指令统一建模、多任务迁移	未显式处理长文本跨窗口实体链与全局一致性
ChatIE[26]	多轮问答式抽取	否	否	否	借助交互式分解提升零样本抽取可控性	更偏局部抽取,缺少全文级实体链与一致性约束
GoLLIE[27]	指南驱动的零样本IE生成	否	否	部分	提升对未见任务/指南的零样本泛化能力	约束主要作用于指南遵循与模式理解,未显式解决长文本跨窗口实体链与关系冲突
GlocalKE(本文)	全局索引引导下的局部生成与全局回写	是	是	是	兼顾跨窗口实体链、事实可追溯与结果一致性	需额外构建共指索引与本体约束

不过，上述方法在长文本场景下并不能直接延续短文本任务中的效果．一个比较突出的问题是，生成式模型在解码时更依赖内部参数化知识，而对输入文本中的显式证据利用不足．这样一来，在噪声较多、信息又相对分散的长文本中，关系预测易偏离原文事实，同时缺乏清晰证据支撑[11]．此外，长程语义依赖的建模能力仍然受限于上下文窗口本身．检索增强生成（Retrieval-Augmented Generation，RAG）类方法通过外部检索扩展可利用的信息范围[28–29]，但知识抽取任务所需的往往是跨段落分布的细粒度关系线索．对于这类线索，常规文本片段检索未必能准确捕捉，故在长文本知识抽取中，结果的完整性和精确度易受影响．

1.3　长文本建模与知识一致性

受超长文本计算复杂度和显存资源限制的影响，采用分治式上下文建模仍是当前长文本处理中的常见思路．滑动窗口方法因实现简单且能较好保留局部语义信息，被广泛用于各类长文本任务[10]．然而，这类方法通常需对原始文本进行物理切分，容易破坏自然语言在篇章中的语义连续性．针对跨窗口信息缺失问题，已有研究尝试通过扩大窗口重叠区域，或引入文档级摘要作为补充上下文加以缓解[30]；但这类策略本质上仍属局部补偿，难以从根本上消除上下文切分带来的割裂影响．尤其是在实体共指链跨越窗口边界时，分散在不同段落中的关联线索往往被削弱甚至截断，进而影响模型对跨段落隐含语义关系的准确识别．也有研究尝试利用长上下文大模型（Long-Context LLMs）直接处理整篇文档，但对信息抽取这类信息密度较高的任务，模型仍可能出现“中间迷失（Lost-in-the-Middle）”，表现为对长序列中段内容的感知与利用效率显著降低[9]．

与此同时，现有多数生成式抽取框架倾向于将长文本拆分为若干彼此独立的片段分别处理，缺乏面向全文范围的统一一致性约束[13]．在这种设定下，不同片段的抽取结果在后续合并时易产生多种类型的冲突．一类典型问题是实体层面的指代混淆，即同一实体在不同段落中可能以全称、简称或代词等形式出现，而模型未必能完成准确归并．另一类问题表现为事实层面的结构冲突，如不同窗口可能分别生成“A是B的父类”与“B是A的父类”这类相互矛盾的三元组，但现有方法通常缺少有效的冲突识别与校验机制[31]．此外，模式约束在生成过程中也可能失效，导致模型突破预定义本体的类型边界，输出不符合领域规范的错误知识，最终削弱抽取结果的可信度．

综上，在上下文窗口受限的条件下，如何兼顾局部语义信息的有效利用与跨窗口实体、关系及结构约束的统一建模，仍是长文本知识抽取中的关键问题．基于此，本文围绕抽取前的全局实体索引构建、抽取过程中的证据支撑校验及抽取后的本体驱动一致性回写，设计了GlocalKE框架，形成一种面向长文本知识抽取的全局-局部协同机制．

2　GlocalKE：全局-局部协同知识抽取框架

为缓解长文本知识抽取中长距离依赖难以保留、结构一致性不足等问题，本文设计了GlocalKE框架，通过“全局索引引导-局部生成-跨窗口逻辑校验”的方式组织抽取过程，在无监督或低资源场景下提升模型对跨段落语义关联的建模能力．

2.1　问题形式化定义

对于一篇给定的非结构化长文本序列 $T=\{x_1, x_2, \cdots, x_N\}$，其中：$x_i$ 表示文本中第 i 个字符，N 为文本长度．知识抽取的任务是在预定义本体 $\Omega=(\mathcal{E}_{\text{type}}, \mathcal{R}_{\text{type}})$ 的约束下，从 T 中抽取知识三元组 $\mathcal{T}_{\text{final}}$．为明确区分文本表述与语义对象，本文引入以下定义．

定义 1（实体提及Mention）．提及是文本中具有特定指代意义的连续字符片段．设 $M=\{m_1, m_2, \cdots, m_K\}$ 为文本中识别出的提及集合，其中每个提及 m_i 表示为四元组 $m_i=(start_i, end_i, str_i, t_i)$，分别对应起止位置、文本表述及所属实体类型 $t_i \in \mathcal{E}_{\text{type}}$．

定义 2（实体Entity）．实体指现实世界中唯一的语义对象．在长文本中，一个实体 e_j 对应一组语义共指的提及集合，即 $e_j=\{m \in M \mid \text{Ref}(m)=e_j\}$，其中 $\text{Ref}(\cdot)$ 为指代映射函数．本文的目标之一是通过跨窗口对齐，将分散的提及集合 M 归一化为唯一的实体集合 $\mathcal{E}$．

定义 3（长文本关系抽取Long-Text Relation Extraction）．定义为建立在两个实体 $e_h, e_t \in \mathcal{E}$ 间的语义关

联,最终输出为三元组集合$\mathcal{K}=\{(e_h,\ r,\ e_t)\,|\,e_h,\ e_t\in\mathcal{E},\ r\in\mathcal{R}_{\text{type}}\}$.

相较于句子级抽取,本任务面临两个核心约束:1)跨段落依赖. 关系三元组的头实体e_h和尾实体e_t可能分布在相距较远的文本片段(如不同的段落或滑动窗口)中,其语义证据不局限于单句上下文. 2)零样本推理. 假设在训练阶段不可见特定领域的标注数据,模型需仅依赖大语言模型的通用知识与本体模式引导完成推理.

2.2 GlocalKE整体架构

GlocalKE的整体架构如图1所示. 基本思路是在统一框架内衔接全局语义建模、局部受控生成和结果一致性回写三个环节,构成一个覆盖抽取前、抽取中和抽取后的完整处理流程. 首先在全文范围内构建跨窗口的实体指代链,将分散出现的实体提及组织为可复用的全局语义锚点;在此基础上,再于局部窗口中进行受全局锚点约束的联合生成,并结合证据匹配机制对候选事实进行筛选,使生成结果尽可能具备明确的原文支撑;随后各窗口得到的候选结果被重新汇聚到全文层面,通过实体对齐与本体驱动的逻辑推理模块完成跨窗口归一化、关系冲突检测、方向修正以及类型回写. 经上述过程,GlocalKE将抽取前的全局先验建模、抽取过程中的局部证据约束及抽取后的全文一致性规约连接起来,形成多阶段协同的长文本知识抽取框架. 其核心处理流程可概括为以下四个阶段:1)全局视野层. 利用共指消解技术建立跨越物理窗口的实体指代链,构建全文级的语义索引,为局部抽取提供跨窗口实体先验. 2)局部抽取层. 采用语义边界感知的滑动窗口策略,在每个窗口内并行抽取实体-关系候选集合,并结合模糊证据匹配机制过滤缺乏原文支撑的候选结果. 3)实体对齐层. 对来自不同窗口的候选实体进行跨窗口归一化,将局部提及恢复为全文实体. 4)逻辑推理层. 基于预定义本体模式,对候选三元组进行方向一致性与类型一致性校验,并将修正结果写回全局实体与关系集合,提升知识三元组的结构合法性与事实一致性.

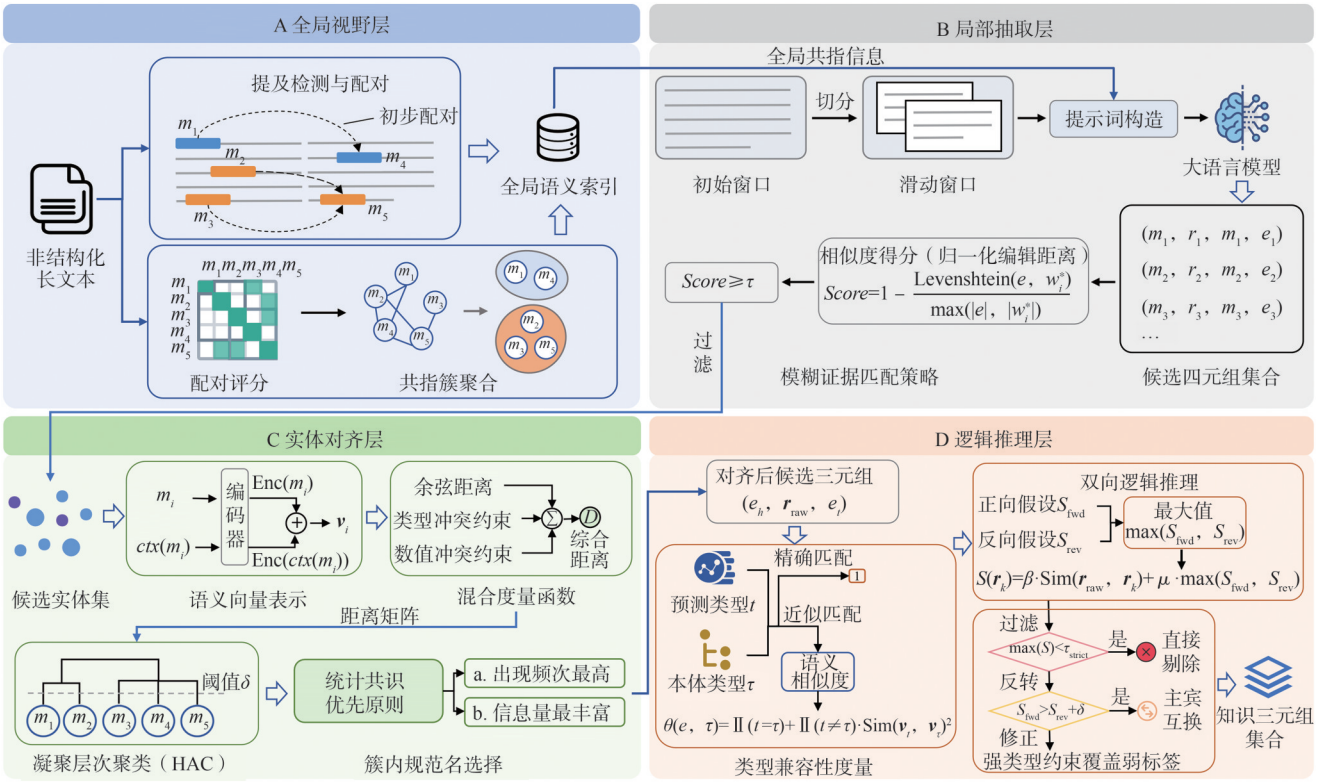


图 1 GlocalKE整体架构

Figure 1 Architecture of GlocalKE

2.3 全局视野层

在长文本处理中,基于滑动窗口的切分方式往往会破坏实体间原本连续的语义联系,尤其不利于跨窗口实体关联的识别. 针对这一问题,本文在局部抽取前引入跨窗口的全局语义索引机制,借助全文范围内的共指信息,为后续分片处理提供文档级的上下文支持.

形式上，给定长文本序列T，首先通过共指消解模块识别文本中的实体提及，将指向同一真实实体的不同表达归并到同一共指簇中．设M表示文档中的实体提及集合，则共指消解过程可表示为映射函数$f: M \to C$，其中$C = \{c_1, c_2, \cdots, c_k\}$为共指簇集合．每个共指簇$c_k \subseteq M$对应一组语义上等价的提及$\{m_{k,1}, m_{k,2}, \cdots, m_{k,n_k}\}$，其中$n_k$表示第$k$个共指簇中的提及数量．

在实现层面，本文采用基于语义表示的神经共指消解方法．先利用预训练语言模型对候选提及进行编码，在此基础上，结合提及间的上下文相似性及局部语境特征，对潜在指代关系进行聚类，进而形成实体级共指链．相较于传统的规则驱动方法，该策略在处理长距离代词回指、简称与全称对应及跨句指代等现象时具有更强的适应能力，更契合长文本场景下分散实体提及的统一表示需求．

在GlocalKE框架中，共指链承担全局语义索引任务，而不直接决定三元组生成．其提供的主要是面向局部抽取阶段的跨窗口实体提示，可视为一种辅助性先验信息．由于共指结果并不直接写入最终输出，即便识别过程中存在少量误差，通常只增加提示噪声，不会直接转化为最终三元组错误．

从作用机制上看，共指结构实际上为全文建立了一种粗粒度的语义拓扑．对于任意落入局部窗口w_i的实体提及m，其所属共指簇c_i可被视为对应的全局锚点．在后续处理中，这一锚点信息被显式注入局部窗口的上下文提示，使模型不仅能利用当前窗口内的信息，还能感知该实体在窗口外的相关表述．如代词“它”可以借助共指链与前文出现的“TCP协议”建立联系．借助这种方式，即使关系三元组的头实体和尾实体分布在不同文本片段中，模型仍可在较一致的实体表示基础上完成推理．

因此，全局视野层输出的不仅是共指链本身，更是一种可跨窗口传递的实体先验．其将因切窗而分散的实体提及重新组织起来，为局部窗口中的关系识别与事实生成提供稳定的全局语义支撑．

2.4　局部抽取层

为缓解大语言模型在上下文长度上的限制，并提升长文本场景下的细粒度抽取能力，局部抽取层采用“分片处理-结果校验”的思路．该层主要包括三个环节：语义边界感知的滑动窗口切分，窗口内的实体关系联合抽取，以及基于证据匹配的结果过滤与幻觉抑制．

考虑到固定长度切分易造成句子截断和语义单元不完整，本文设计一种语义边界感知的滑动窗口策略．对于给定长文本序列T和预设滑动窗口长度L，算法通过回退搜索的方式，优先寻找距离窗口末端最近的自然语言停顿符，如句号、换行符等，并将其位置作为实际截断点L'（满足$L' \leqslant L$）．在此基础上，文本被划分为一组局部子序列$W = \{w_1, w_2, \cdots, w_M\}$．设滑动步长为$S$（满足$S < L$），则相邻窗口间保留一定重叠区域，以保证跨窗口边界附近的实体提及至少能在某一局部窗口中获得相对完整的上下文．

L与S的设定需同时考虑上下文覆盖与资源消耗．L偏小时，部分实体提及与其关系证据可能在窗口边界处被切开，导致相关实体难以在同一局部窗口内共同出现；L偏大时，输入中又会混入较多与目标关系无关的背景内容，从而削弱有效语义线索，并提高推理开销．S也会影响结果稳定性．若S过大，窗口间的重叠不足，跨边界信息更易缺失；若S过小，则带来更多重复计算．基于此，本文未采用固定长度的硬切分，而是尽量结合自然语言中的停顿边界进行分段，以保留更完整的句法和语义结构，减少切分操作对后续关系抽取的干扰．在长文本知识抽取任务中，这种设置更有利于维持跨句关系识别的稳定性．

得到局部窗口w_i后，进一步利用大语言模型的生成能力执行实体-关系联合抽取．不同于将实体识别和关系分类拆分处理的流水线方法，本文将两者统一建模为端到端的序列生成任务，直接输出带有证据信息的候选四元组集合Q_w．每个四元组可形式化表示为$q = (m_h, r_k, m_t, e)$，其中：m_h和m_t分别表示头实体提及和尾实体提及，$r_k \in \mathcal{R}$表示关系类型，e表示支撑该关系预测的原文证据片段．在这一阶段，全局视野层构建的共指信息被进一步注入提示（Prompt）中，用于帮助模型解析局部窗口中的代词或隐式指代，提升局部推理过程的连贯性．

在结果校验阶段，本文采用基于归一化编辑距离的模糊证据匹配方法，对生成证据e与原文窗口w_i间的对应程度进行衡量．考虑到模型生成的证据片段与原文表达间可能存在措辞差异，该策略并不要求两者完全一致，而是在语义基本保持一致的前提下，通过计算相似度得分$Score(e, w_i)$来量化证据的文本支撑度：

$$Score(e, w_i) = 1 - \frac{\text{Levenshtein}(e, w_i^*)}{\max(|e|, |w_i^*|)}, \tag{1}$$

式中:w_i^*表示窗口w_i中与证据片段e最相近的子串. 进一步,本文设置置信下界τ,用于过滤满足$Score(e, w_i) < \tau$的低置信候选项,以保证留下的抽取结果具有较好的原文可追溯性.

综上,局部抽取层在尽量保留细粒度上下文感知能力的同时,生成带有原文证据支撑的候选四元组,为跨窗口实体归一和逻辑一致性约束提供可进一步校验的中间结果.

2.5 实体对齐层

实体对齐层的目标是把分散在不同局部窗口中的候选提及统一到全文层面的实体表示上,以便后续逻辑推理可围绕实体本身展开全局校验. 然而,仅依赖实体名称或单纯依靠上下文往往难以应对复杂的指代消歧问题. 为此,本文提出一种融合实体本体与局部语境的混合表示策略.

对于候选提及m_i,将其实体嵌入与上下文嵌入进行加权组合,以构造语义表示$\boldsymbol{v}_i$:

$$\boldsymbol{v}_i = \alpha \cdot \mathrm{Enc}(m_i) + (1 - \alpha) \cdot \mathrm{Enc}(ctx(m_i)), \tag{2}$$

式中:$\mathrm{Enc}(\cdot)$表示预训练语义编码器;$ctx(m_i)$表示m_i所处的上下文窗口;$\alpha \in [0, 1]$作为调节系数,用于平衡实体表层特征与上下文语义信息的相对作用,避免实体表示单纯依赖实体名称本身或局部上下文,增强实体表征的鲁棒性与判别力.

仅依赖向量空间相似性时,一些类型不同但语境接近的实体(如“导演”和“演员”),或字面形式相近的数值实体(如“2 000”与“4 000”),往往表现出较高的余弦相似度,干扰后续聚类结果. 针对这一问题,本文设计一种结合语义距离与规则约束的混合距离度量函数,将提及m_i与m_j间的综合距离定义为:

$$D(m_i, m_j) = d_{\cos}(\boldsymbol{v}_i, \boldsymbol{v}_j) + \lambda_{\mathrm{type}} \cdot \mathbb{I}(t_i \neq t_j) + \Phi_{\mathrm{rule}}(m_i, m_j), \tag{3}$$

式中:$d_{\cos}$为余弦距离;λ_{type}为类型惩罚系数,用于抑制不同类型实体在聚类过程中被错误合并;$\mathbb{I}(\cdot)$为类型冲突指示函数,当两个实体类型不一致时取为1,否则为0. 函数$\Phi_{\mathrm{rule}}(m_i, m_j)$则表示基于实体表面形式引入的规则惩罚项,用于进一步提高实体归一化过程的鲁棒性.

本文主要考虑两类规则约束:1)数值实体约束. 当m_i与m_j均属数值型实体(如年份、数值或百分比)且二者表面形式并不相同时,对其距离施加较强惩罚,避免不同数值被错误归并. 2)字符串结构约束. 当两个提及间存在明显的子串关系时,如“阿凡达”与“阿凡达2”,则引入轻量级字符串惩罚,以降低续作名称、扩展名称或相似简称被误合并的可能性. 通过加入上述规则,模型可在一定程度上弥补纯语义嵌入在处理数值实体和字符串变体时的不足,从而提升实体对齐的准确性.

获得提及间的距离矩阵后,本文采用凝聚层次聚类(Hierarchical Agglomerative Clustering,HAC)算法完成实体归一化,将距离小于阈值δ的提及归并到同一实体簇c_k中. 对于簇内规范名称的确定,本文采用“统计共识优先”的原则,优先选择簇内出现频次最高的提及作为代表,若多个提及频次相同,则进一步选取字符更长、信息更完整的表述作为规范实体名称. 该策略能够利用全文层面的统计信息,对局部窗口中可能出现的拼写偏差或非规范简称起到一定平滑作用.

2.6 逻辑推理层

实体对齐模块解决的是跨窗口提及标准化归并问题,但并不意味着候选三元组已完全可靠. 由于各局部窗口是相对独立完成生成的,结果中仍可能保留语义或逻辑层面的偏差,如关系的定义域、值域与本体约束不符,或主客体方向出现颠倒. 为此,本文在全局层面设置逻辑推理模块,引入基于本体模式的双向逻辑推理机制,对候选三元组进行一致性校验,并在需要时作出修正.

为支持实体层面的类型校验,首先设计一种非线性的类型兼容性度量函数,用于刻画实体预测类型与本体预定义类型间的一致程度. 考虑到嵌入空间中不同细粒度类型间并非完全割裂,如“电影”与“影视作品”间通常存在层级上的接近关系,故对于任意实体e、其预测类型t及本体中的标准类型τ,定义兼容度$\theta(e, \tau)$为:

$$\theta(e, \tau) = \mathbb{I}(t = \tau) + \mathbb{I}(t \neq \tau) \cdot \mathrm{Sim}(\boldsymbol{v}_t, \boldsymbol{v}_\tau)^2, \tag{4}$$

式中:$\mathrm{Sim}(\cdot)$表示余弦相似度;$\mathbb{I}(\cdot)$为指示函数;$\boldsymbol{v}_t$和$\boldsymbol{v}_\tau$分别为类型t与τ的语义向量. 该定义将精确匹配与近似匹配区分开来,当类型一致时,直接赋予满置信度;当类型不完全一致时,则利用语义相似度进行柔性衡量,并通过平方项进一步拉大高相似类型与低相似噪声类型间的差异,以增强类型校验的判别性.

除实体类型约束外,关系表达本身也需纳入规范化约束. 在生成式抽取任务中,大语言模型生成的关系

描述通常具有一定的语义合理性，但未必能严格映射到预定义本体的关系集合中．若仅依赖实体类型进行约束，易忽略关系文本本身包含的语义信息，故难以兼顾关系表达的语义准确性与结构合法性．基于此，本文在统一评分框架下同时引入关系语义相似性与类型一致性两类信号，前者用于衡量模型生成的关系表述与标准关系间的语义接近程度，后者用于检验候选三元组是否满足该关系对应的头尾实体类型约束．通过融合上述两类信息，可在保留生成式表达灵活性的同时，进一步借助本体模式对三元组的逻辑结构进行有效约束．

设标准关系集合为$\mathcal{R}$，则对其中任意候选关系$\boldsymbol{r}_k$，本文定义其联合评分函数为：

$$S(\boldsymbol{r}_k)=\beta\cdot\mathrm{Sim}(\boldsymbol{r}_{\mathrm{raw}},\boldsymbol{r}_k)+\mu\cdot\max(S_{\mathrm{fwd}},S_{\mathrm{rev}}),\tag{5}$$

式中：$\boldsymbol{r}_{\mathrm{raw}}$表示大语言模型生成的原始关系文本的语义向量；$\boldsymbol{r}_k\in\mathcal{R}$表示第$k$个预定义标准关系的向量；$S_{\mathrm{fwd}}$和$S_{\mathrm{rev}}$分别对应正向假设$(e_h\rightarrow e_t)$与反向假设$(e_t\rightarrow e_h)$下的类型兼容度得分．若关系$\boldsymbol{r}_k$要求的头实体和尾实体类型分别记为$\tau_{\mathrm{head}}$和$\tau_{\mathrm{tail}}$，则正向得分表示为：

$$S_{\mathrm{fwd}}=\frac{1}{2}\left(\theta(e_h,\tau_{\mathrm{head}})+\theta(e_t,\tau_{\mathrm{tail}})\right).\tag{6}$$

反向得分S_{rev}的计算方式与此相同，只是头尾实体的位置互换．参数β与μ分别控制关系语义相似性与结构一致性约束在总体评分中的权重．前者主要反映模型对关系含义的语义判断，后者则体现本体模式对三元组结构的约束作用；二者结合后，可为关系归一化和逻辑修正提供更稳定的依据．

得到联合评分结果后，本文进一步采用如下拓扑逻辑修正策略．若$\max(S_{\mathrm{fwd}},S_{\mathrm{rev}})<\tau_{\mathrm{strict}}$，认为该三元组与本体模式严重不符，直接予以过滤．当反向得分明显高于正向得分，即满足$S_{\mathrm{rev}}>S_{\mathrm{fwd}}+\delta$（$\delta>0$为方向敏感阈值）时，判定该候选结果可能存在主宾倒置问题，此时交换三元组中的头尾实体位置，即$(e_h,e_t)\leftarrow(e_t,e_h)$，以修正关系方向错误．对于通过校验的候选三元组，本文使用得分最高的标准关系$\boldsymbol{r}^*$对应的强类型约束，对原始生成阶段给出的弱类型标签进行回写和细化．如可将较为泛化的“人物”进一步修正为本体中更精确的“作曲家”，以减轻局部生成过程带来的类型噪声．

总体来看，逻辑推理层将本体约束转化为可计算的软约束信号，在全文范围内对局部候选结果进行统一校验和必要修正．通过这一过程，GlocalKE可将抽取前的全局建模、抽取中的局部生成、抽取后的全文一致性规约衔接起来，形成完整的“全局-局部-全局”协同处理链路．

2.7　整体算法描述

综合上述各核心模块，将GlocalKE框架的完整执行流程形式化为算法1．该算法以非结构化长文本T和预定义本体模式Ω为输入，通过多阶段协同处理，最终输出结构化的知识三元组集合$\mathcal{T}_{\mathrm{final}}$．

算法 1

输入：长文本T，本体模式Ω，滑动窗口长度L，滑动步长S，证据匹配阈值τ，平衡系数α，类型惩罚系数λ_{type}，聚类阈值δ，关系语义权重β，逻辑一致性权重μ

输出：知识三元组集合$\mathcal{T}_{\mathrm{final}}$

// 阶段一：全局视野与局部生成
$C\leftarrow\mathrm{CoreferenceResolution}(T)$ ▷构建全局共指链
$W\leftarrow\mathrm{SemanticSlidingWindow}(T,L,S)$ ▷语义边界感知的窗口切分
$\mathcal{T}_{\mathrm{cand}}\leftarrow\varnothing$
for $\boldsymbol{w}_i\in W$ **do**
　$P_{\mathrm{context}}\leftarrow\mathrm{InjectGlobalIndex}(\boldsymbol{w}_i,C)$
　$\mathcal{T}_{\mathrm{local}}\leftarrow\mathrm{LLM_JointExtraction}(\boldsymbol{w}_i,\Omega,P_{\mathrm{context}})$
　for $t=(m_h,r,m_t,e)\in\mathcal{T}_{\mathrm{local}}$ **do**
　　$s_{\mathrm{ground}}\leftarrow 1-\dfrac{\mathrm{Levenshtein}(e,w_i^*)}{\max(|e|,|w_i^*|)}$ ▷模糊证据匹配校验
　　if $s_{\mathrm{ground}}\geqslant\tau$ **then** $\mathcal{T}_{\mathrm{cand}}.\mathrm{add}(t)$
　end for
end for

13. // 阶段二:跨窗口实体对齐
14. $M \leftarrow \mathrm{ExtractMentions}(\mathcal{T}_{\mathrm{cand}})$
15. $\boldsymbol{V} \leftarrow \{\, \alpha \cdot \mathrm{Enc}(m) + (1-\alpha) \cdot \mathrm{Enc}(ctx(m)) \,|\, m \in M \,\}$
16. 计算距离矩阵 $\boldsymbol{D}$,其中 $D_{ij} = d_{\cos}(\boldsymbol{v}_i, \boldsymbol{v}_j) + \lambda_{\mathrm{type}} \cdot \mathbb{I}(t_i \neq t_j) + \lambda_{\mathrm{num}} \Phi_{\mathrm{num}}(m_i, m_j)$
17. $\mathrm{Clusters} \leftarrow \mathrm{HAC_Clustering}(\boldsymbol{V}, \boldsymbol{D})$
18. $\mathcal{E}_{\mathrm{map}} \leftarrow \varnothing$
19. **for** $c_k \in \mathrm{Clusters}$ **do**
20. $\quad e_{\mathrm{norm}} \leftarrow \mathrm{SelectByFreqAndLen}(c_k)$ ▷频次优先的规范化
21. $\quad \mathcal{E}_{\mathrm{map}}.\mathrm{update}(\{\, m: e_{\mathrm{norm}} \,|\, m \in c_k \,\})$
22. **end for**
23. $\mathcal{T}_{\mathrm{aligned}} \leftarrow \mathrm{MapMentions}(\mathcal{T}_{\mathrm{cand}}, \mathcal{E}_{\mathrm{map}})$
24. // 阶段三:逻辑推理与自修正
25. $\mathcal{T}_{\mathrm{final}} \leftarrow \varnothing$
26. **for** $t = (e_h, r, e_t) \in \mathcal{T}_{\mathrm{aligned}}$ **do**
27. $\quad S_{\mathrm{fwd}}, S_{\mathrm{rev}} \leftarrow \mathrm{CalcSchemaCompat}(t, \Omega)$
28. $\quad$ **if** $\max(S_{\mathrm{fwd}}, S_{\mathrm{rev}}) < \tau_{\mathrm{strict}}$ **then continue** ▷严重冲突过滤
29. $\quad$ **if** $S_{\mathrm{rev}} > S_{\mathrm{fwd}} + \delta$ **then**
30. $\quad\quad t \leftarrow (e_t, r, e_h)$; $\mathrm{UpdateType}(t, \Omega)$ ▷方向反转修正
31. $\quad$ **else** $\mathrm{UpdateType}(t, \Omega)$
32. $\quad \mathcal{T}_{\mathrm{final}}.\mathrm{add}(t)$
33. **end for**
34. **return** $\mathcal{T}_{\mathrm{final}}$.

3 实验与分析

本文从对比实验、消融实验及参数敏感性三个方面,对GlocalKE在长文本复杂语境中的表现进行分析.所有实验均部署在NVIDIA RTX 4060 GPU平台上.为观察该方法在不同大语言模型基座下的适配情况,分别采用Qwen3-Max、GLM-5和DeepSeek-V3.2作为基础生成模型.与此同时,考虑到生成式解码存在随机性,将温度参数(Temperature)固定为0.1,以降低随机因素对实验结果的干扰.

实验设置覆盖通用领域和垂直领域两类数据场景,主要关注全局索引、模糊证据匹配和逻辑推理等模块在长距离语义建模中的作用.另需说明的是,本文研究的是面向长文本知识抽取的生成式框架,故比较的重点在于不同协同机制带来的性能变化,而非与需要额外监督训练的文档级判别式方法进行严格对等的横向比较.

3.1 数据集与评价指标

3.1.1 数据集

DuIE-Doc(通用领域).由句子级关系抽取数据集DuIE 2.0[32]扩展而来.构建过程中,将原本彼此独立、但在共指或语义上相关的短句重新组织为连贯文档,并补充跨段落关联标注,模拟真实的长文本知识抽取场景,并检验模型对长距离语义依赖的处理能力.

CS-Cn-50(垂直领域).面向《计算机网络》课程内容构建的垂直领域长文本数据集,语料主要选自经典教材与开源技术文档.相比通用领域文本,这一数据集在概念组织、逻辑嵌套和术语使用上更为集中和严格.其标注采用“机器预标-人工校核”的方式完成,先借助Gemini 3 Pro与GPT-5.2生成候选三元组,再由人工对术语统一性、关系覆盖情况及跨段落关联进行核对与修正,从而保证标注质量.

3.1.2 评价指标

评价体系沿用信息抽取任务标准的P、R与$F1$值.首先对每篇文档分别基于预测三元组集合与标注三元

组集合计算P、R和$F1$值，仅当头实体、关系类型和尾实体均匹配时，判定该三元组为正确抽取；随后对全部文本的对应指标取算术平均（宏平均），作为最终评价结果．在长文本知识抽取任务中，远距离依赖、跨段干扰及事实覆盖不均衡等现象较为常见，单一指标往往不足以完整呈现模型性能．考虑到$F1$值能够兼顾P和R，本文将其作为主要评价指标，并以P和R作为补充，分析模型在不同方面的表现．

3.2　主实验

3.2.1　对比基线

本文选取以下五种具有代表性的方法作为基线．

1）Vanilla（Zero-shot）．用于表征大语言模型在不引入额外辅助机制时的基础抽取能力．该设置仅依赖模型原生上下文窗口，对超长文本采用硬截断方式处理，不额外加入跨窗口记忆模块或后处理步骤，观察模型在缺少外部协同支持的情况下，仅凭参数化知识完成端到端生成时的表现．2）CoT（Chain-of-Thought）．用于表征显式推理提示对生成式抽取的影响．该方法在指令中加入“通读-分析-判定”的思维链，引导模型将长文本抽取过程拆解为若干连续步骤，分析在无监督条件下，显式推理引导是否能改善局部上下文中的关系判断能力．3）UIE-style．参照统一信息抽取方法UIE[7]的思想，将实体、关系及三元组抽取统一转化为结构化文本生成任务．该基线通过统一输出模板约束模型生成结果，使不同类型的抽取对象能够在同一生成框架下表达．4）ChatIE-style．参照ChatIE[26]的交互式抽取思想，将知识抽取过程拆分为多轮问答式步骤，依次完成命名实体识别、实体对组合生成和关系类型判定．5）GoLLIE-style．参照GoLLIE[27]的指南驱动抽取思想，在提示词中显式加入实体类型、关系模式及约束说明，引导模型按照预定义本体生成候选三元组．该基线主要用于分析本体指南对生成结果规范性的影响，尤其关注模式约束是否能提升垂直领域和零样本场景下的抽取稳定性．

3.2.2　实验结果分析

表2列出了GlocalKE与各基线方法在DuIE-Doc和CS-Cn-50上的实验结果．在相同的生成式推理设置下，GlocalKE的P、R和$F1$值在两个数据集上均为最优．结合结果可知，在加入全局索引、证据校验和逻辑规约后，模型对跨窗口信息的处理更为完整，长文本抽取过程中由局部切分带来的不一致问题得到一定缓解．

表 2　不同模型在DuIE-Doc与CS-Cn-50上的性能对比

Table 2　Performance comparison of different models on DuIE-Doc and CS-Cn-50

评价指标	DuIE-Doc			CS-Cn-50		
	P/%	R/%	$F1$值/%	P/%	R/%	$F1$值/%
Vanilla	22.41	26.45	21.01	18.52	12.35	14.82
CoT	34.29	37.09	36.42	28.45	25.10	26.67
UIE-style	55.07	55.45	52.13	40.86	32.74	34.95
ChatIE-style	33.09	47.42	37.40	30.12	29.68	28.54
GoLLIE-style	62.40	54.25	55.24	44.73	34.82	37.62
GlocalKE（本文）	**66.49**	**69.29**	**65.92**	**47.61**	**38.46**	**42.55**

从基线结果看，不同方法在长文本场景下的表现差异明显．Vanilla依赖固定窗口切分文本，连续语义关系易在窗口边界处被截断，对跨窗口关系的识别能力相对有限．CoT通过推理提示改善局部分析过程，但缺少全文层面的统一约束，在实体指代漂移和关系冲突出现时不够稳定．UIE-style和GoLLIE-style借助统一结构模板或本体指南增强生成结果的规范性，整体表现优于Vanilla和CoT，但这类方法主要约束生成形式和模式遵循，尚未显式维护跨窗口实体链，也缺少基于原文证据的后验校验．ChatIE-style通过多轮问答拆分抽取过程，可降低一次性生成复杂度，但多轮交互仍主要围绕局部片段展开，难以保证全文层面的结果一致性．相比之下，GlocalKE在DuIE-Doc上的$F1$值较CoT提高29.50个百分点，较GoLLIE-style提高10.68个百分点，说明长文本知识抽取不仅需要局部推理和模式约束，还需要额外的全局-局部协同与证据回溯机制．

与DuIE-Doc相比,CS-Cn-50上的结果略低,这与垂直领域文本本身的特点有关. 该数据集包含较多专业术语、缩写表达和层级化概念,信息分布更为碎片化,关系判断也更依赖领域背景知识. 传统方法在此类场景下因上下文缺失或本体利用不足而表现不佳. GlocalKE在此类场景下的表现仍优于基线模型,凸显该方法在处理长程依赖与逻辑约束时的有效性,相关误差来源及后续改进方向将在3.6.3节进一步讨论.

3.3 消融实验

本文在DuIE-Doc数据集上进行消融实验,分析GlocalKE各核心组件在长文本知识抽取中的作用. 具体做法是分别移除全局索引(Global Index)、证据匹配校验(Evidence Matching)、跨窗口实体对齐(Entity Alignment)和逻辑推理(Logic Reasoner),并比较不同变体在P、R和$F1$值上的变化,结果如表3所示.

表 3 GlocalKE在DuIE-Doc数据集上的消融实验结果

Table 3 Ablation study results of GlocalKE on the DuIE-Doc dataset

模型变体	P/%	R/%	$F1$值/%	Δ$F1$值/%
-全局索引(w/o Global Index)	57.49	55.11	52.37	−13.11
-证据匹配(w/o Evidence Matching)	52.50	66.04	58.50	−6.98
-实体对齐(w/o Entity Alignment)	64.29	59.96	62.15	−3.33
-逻辑推理(w/o Logic Reasoner)	66.46	61.86	64.10	−1.38
GlocalKE(本文)	**68.11**	**66.72**	**65.48**	—

由表3可知,不同组件对模型性能的影响并不相同,其中全局索引的作用最为明显,去除该模块后,模型$F1$值下降13.11%,说明在滑动窗口处理长文本的条件下,如果缺少跨窗口实体和语义线索的统一组织,模型较难维持稳定的长距离语义关联,也表明全局索引在框架中发挥连接局部窗口与全文语义信息的重要作用. 证据匹配模块移除后,模型的R仍为66.04%,而P降至52.50%,表现出高召回、低精度的特点. 造成这一现象的原因在于,缺少原文证据约束时,生成模型往往保留更多语义上看似合理的候选关系,但其中一部分并不能在文本中找到充分支撑,故更易形成错误三元组. 加入证据匹配后,系统可利用生成结果与原文片段间的对应程度,对低置信候选项进行筛选. 虽然R会受到一定影响,但抽取结果在可信性和可追溯性方面更有保障. 实体对齐模块和逻辑推理模块对应的$F1$值增益分别为3.33%和1.38%. 这两部分带来的指标变化相对有限,但作用主要集中在结果整理和后验修正阶段. 前者用于处理跨窗口场景下同一实体存在不同表述的问题,有助于提高实体归一化质量;后者在候选结果生成后引入本体约束,对类型不匹配或存在结构冲突的三元组进行检查与修正. 二者的价值不只体现在指标增幅上,更体现在结果表达是否规范、整体结构是否一致上.

整体来看,GlocalKE的性能提升并不是由某一单独模块带来的,而是来源于多个组件间的配合. 全局索引和证据匹配分别对应长文本知识抽取中较突出的跨窗口语义割裂和事实幻觉问题,实体对齐与逻辑推理则在此基础上进一步改善结果的规范性与一致性.

3.4 参数敏感性分析

L与τ分别影响局部上下文的覆盖范围及候选结果的过滤强度,是本文框架中较为关键的两个超参数. 为考察其对长文本知识抽取效果的影响,进一步分析不同参数设置下模型性能的变化,结果如图2所示.

L影响局部抽取阶段可覆盖的上下文范围. 由图2(a)可知,随着L从256逐步增大到768,模型的$F1$值持续提高,并在$L = 768$时达到最大. 其原因在于,窗口过小时,跨段落实体间的语义联系更易在切分过程中被打断,一些依赖较长上下文的信息难以在同一窗口内完整保留. 而适当增大窗口后,关系判断所需的语境信息更加充分,抽取结果也更完整. 当L进一步增加到1 024及以上时,性能反而下降. 说明在长文本抽取中,扩大上下文范围虽可补充信息,也将引入更多无关内容,增加特征筛选和关系判断的难度. τ主要决定证据校验模块对候选三元组的筛选强度. 由图2(b)可知,$F1$值随τ的变化呈现较明显的倒"U"形趋势,并在$\tau = 0.75$时取得最佳结果. 当阈值设置过低时,过滤条件相对宽松,一些缺乏原文充分支撑的候选结果被保留下来,从而影响P;相反,当阈值设置过高时,部分语义上成立、但表述形式与原文不完全一致的合法三元组也可能被误删,进而对R造成不利影响.

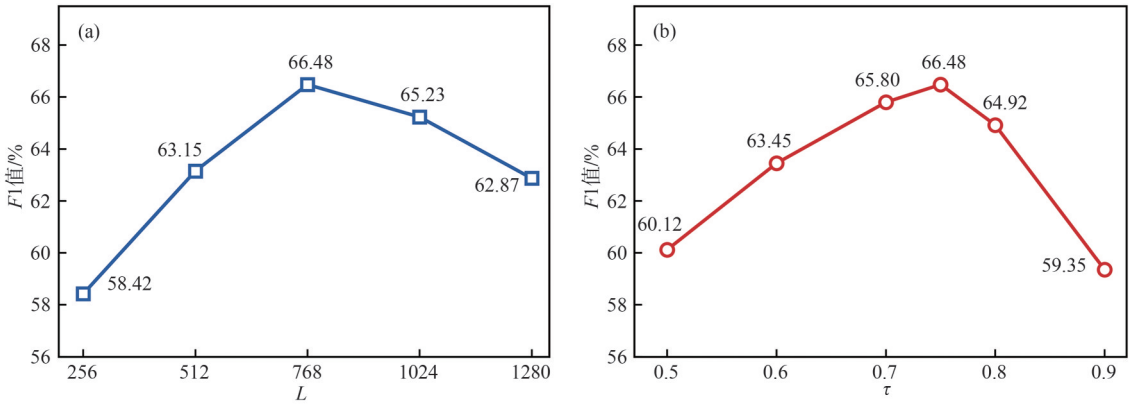


图 2　滑动窗口长度与证据匹配阈值对模型性能的影响

Figure 2　Impact of sliding window length and evidence matching threshold on model performance

除上述两个参数外，α 与 λ_{type} 也会影响模型表现．前者用于调节实体提及语义与上下文信息在表示中的融合比例，后者用于约束不同类型实体在聚类过程中被错误合并的风险．预实验结果表明，在本文所用数据集上，实体提及本身通常已具有较强的区分能力，而局部上下文中往往包含一定比例的背景性描述，故当 α 取较大值时，模型整体表现更为稳定．基于这一观察，本文最终将 α 设为0.95，使实体表示以提及本身的语义信息为主，同时保留少量上下文作为补充．对于 λ_{type}，若取值过低，类型约束不足，不同类别实体更易在聚类中被错误合并；若取值过高，又可能对本应保留的语义相似实体形成过强抑制．综合实验结果，本文将 λ_{type} 统一设为1，在语义相似性与类型一致性间取得相对合理的平衡．此外，针对数值实体和字符串变体，本文通过规则惩罚项进行补充约束，提高实体归一化的稳定性．

3.5　不同基座模型的泛化性分析

为考察GlocalKE在不同大语言模型基座下的适用性，本文在统一的推理流程和提示结构下，补充比较Qwen3-Max、GLM-5和DeepSeek-V3.2三种生成模型．通过不同模型条件重复实验，进一步观察该框架在不同语义建模能力下的表现是否稳定．除底层生成模型外，其余设置均保持一致，包括提示模板、滑动窗口策略、实体对齐方式及逻辑约束规则等，以尽可能保证不同模型间结果的可比性，具体如表4所示．

表 4　不同基座模型下GlocalKE性能对比

Table 4　Performance comparison of GlocalKE with different foundation models

基座模型	数据集	P/%	R/%	$F1$值/%
Qwen3-Max	DuIE-Doc	66.49	69.29	65.92
GLM-5		55.43	64.73	57.62
DeepSeek-V3.2		61.12	66.48	62.21
Qwen3-Max	CS-Cn-50	47.61	38.46	42.55
GLM-5		39.27	33.14	35.94
DeepSeek-V3.2		43.85	36.71	39.92

由表4可知，GlocalKE在不同大语言模型基座下均能保持较稳定的抽取效果．虽然不同模型在 P、R 和 $F1$ 值上存在一定差异，但整体变化趋势较为一致，说明该框架并不会因底层基座模型的变化而失去作用．从两个数据集的实验结果来看，GlocalKE在不同基座模型下的 $F1$ 值变化不大，说明方法本身具有一定的稳定性．其性能提升并非单纯建立在某个基座模型之上，更多源于框架设计本身．全局索引、证据匹配校验、跨窗口实体对齐等共同作用，对长文本场景中易出现的上下文割裂问题形成补充，故在生成能力不同的基座模型下，方法仍能保持较为一致的抽取表现．

3.6　案例与错误分析

选取CS-Cn-50数据集中“TCP/IP分层模型与数据封装过程”的文本片段作为案例，对GlocalKE在垂直领域长文本中的工作过程进行分析．该文本具有典型的专业领域特征，一方面，核心概念及其具体实例分散在

不同段落中;另一方面,数据单元间的封装关系具有较明确的方向约束和层级逻辑. 图3展示了GlocalKE在该案例中进行跨窗口关联和逻辑自修正的处理过程.

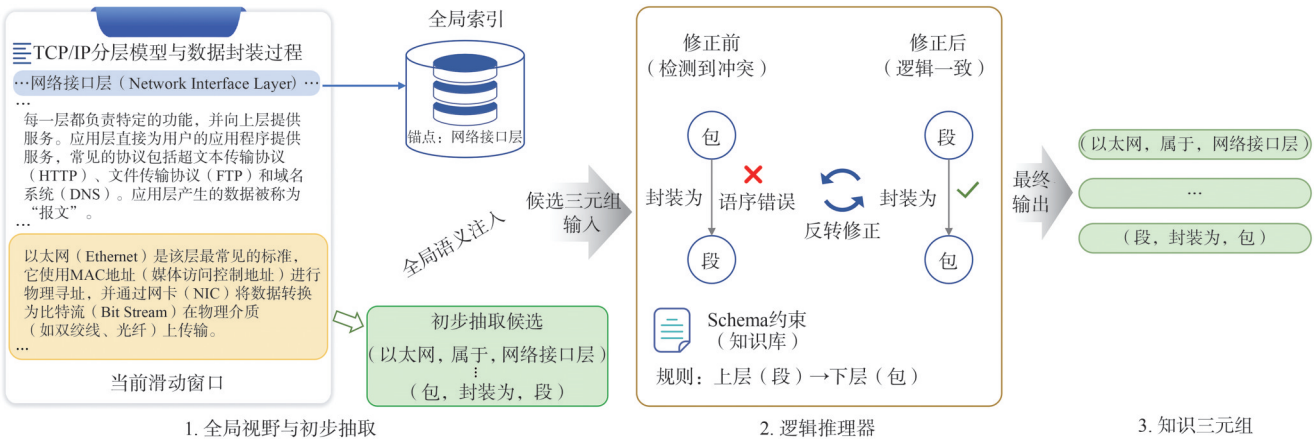


图 3 GlocalKE跨窗口关联与逻辑自修正工作流程

Figure 3 Workflow of GlocalKE for cross-window association and logical self-correction

3.6.1 基于全局索引的跨段落语义消解

在长文本知识抽取中,滑动窗口虽能缓解输入长度限制,但也易打断原本连续的上下文. 对于上述案例,层级概念“网络接口层”出现在文本前部,其对应的具体实现标准“以太网”则位于靠后的段落,中间还穿插关于应用层、传输层和网络层的大量说明性内容. 若仅依赖局部窗口进行抽取,模型较难稳定建立两者间的关联.

针对这一问题,GlocalKE在处理初期先构建全文范围内的全局语义索引,将“应用层”“传输层”“网络接口层”等高频核心实体组织为可复用的语义锚点. 当窗口滑动到末段并处理“以太网”相关内容时,模型可借助这些先验知识,将当前局部语境与前文出现的层级概念联系起来. 即使实体分布跨越多个窗口,模型仍能在较一致的全局语义背景下恢复首尾实体之间的从属关系,减少中间冗余描述对关系判断的干扰.

3.6.2 基于本体模式约束的逻辑冲突自修正

在专业领域文本中,术语集中出现、句式嵌套较多是常见现象,使生成模型更易在抽取时误判语义角色. 以“IP协议将传输层的数据段封装成‘包’”为例,句子的关键不在于表层词序,而在于“段”到“包”的封装方向. 若模型没有准确把握这一层级关系,而是更多依赖表面表达形式,可能在头实体和尾实体的判定上出现偏差. 实验中,基线模型生成(包,封装为,段)这一结果. 单看文字形式,该三元组似乎具备一定合理性;但从计算机网络知识本身来看,这一方向与封装过程并不相符,属于较典型的方向性错误.

针对这一问题,GlocalKE的逻辑推理层发挥约束与修正功能. 根据预定义本体模式,“封装”关系被限定为遵循“上层数据单元→下层数据单元”的层级流向. 在推理过程中,系统识别出候选三元组(包,封装为,段)存在逻辑不一致,头实体“包”对应网络层,尾实体“段”对应传输层,与本体约束要求的方向关系并不一致. 在此基础上,逻辑推理模块进一步结合类型兼容度判定结果,对候选项进行方向修正,最终得到(段,封装为,包). 这一案例说明,在生成结果之外引入显式领域约束,能在一定程度上减少由词序偏置或局部语境误判引发的结构性错误.

3.6.3 错误类型分析与局限性讨论

尽管前文实验和案例分析表明GlocalKE在跨窗口关联和一致性修正方面具有一定效果,但从具体预测结果看,该方法在一些复杂情形下仍会出现错误,主要包括以下三类:1)复杂嵌套关系场景. 当目标事实依赖多层语义嵌套或需同时结合多个从句、多个段落的信息进行判断时,模型可能出现关系方向判断错误或关系遗漏. 2)低频实体与专业术语场景. 在垂直领域长文本中,低频术语、缩写形式及边界不够清晰的专业实体较难形成稳定表示,为实体识别和归一化带来困难. 3)跨窗口证据分散场景. 当关键证据被大量无关内容隔开并分布在相距较远的段落时,即使有全局索引参与辅助,局部窗口仍可能无法覆盖完整证据链,造成部分事实遗漏.

这些问题主要与长距离依赖保持、低频术语表示及跨窗口证据覆盖不足有关，未来计划引入外部领域知识库作为辅助先验，同时结合对比学习等手段进一步增强低频术语的语义特征表征，提升模型在缺少大规模预训练语料支持下的零样本识别能力，缩小通用领域与垂直领域间的性能差距.

3.7　时间与空间复杂度分析

本节对GlocalKE框架各处理阶段的时间与空间复杂度进行理论分析，讨论其在长文本处理中对计算资源的敏感性. 设输入长文本长度为N、滑动窗口长度为L、滑动步长为S，则窗口数量可近似表示为：

$$W = \left\lceil \frac{N - L}{S} \right\rceil + 1. \tag{7}$$

设候选实体提及数量为M、候选三元组数量为K、本体关系类型数量为N_r、语义向量维度为d，在该设定下，GlocalKE各模块的主要时间与空间复杂度如表5所示.

表 5　GlocalKE各模块时间与空间复杂度分析

Table 5　Time and space complexity analysis of GlocalKE modules

模块	主要操作	时间复杂度	空间复杂度
语义边界感知切分	按滑动窗口长度与自然句界切分文本	$O(N)$	$O(W)$
局部生成抽取	对W个窗口分别进行大语言模型推理	$O\left(WC_{\mathrm{LLM}}(L)\right)$	$O\left(S_{\mathrm{LLM}}(L)\right)$
全局共指索引	提及编码、相似度计算与共指聚合	$O\left(M^2 d\right)$	$O\left(Md + M^2\right)$
模糊证据匹配	候选证据与窗口文本进行近似匹配	$O(KL)$	$O(K)$
跨窗口实体对齐	距离矩阵计算与实体聚类	$O\left(M^2 d\right)$	$O\left(Md + M^2\right)$
逻辑推理修正	本体匹配与方向校验	$O(KN_r d)$	$O(K + N_r d)$

GlocalKE的主要计算开销集中在局部生成抽取、全局共指索引构建和跨窗口实体对齐三个部分. 局部生成阶段需对W个窗口分别进行大语言模型推理，其总开销与窗口数量、单窗口推理成本相关. 由于每次推理仅处理长度为L的局部窗口，而不是一次性输入完整长文档，故单次推理的显存压力主要受窗口长度控制，不随全文长度N直接增长. 全局共指索引和实体对齐阶段涉及候选实体提及间的相似度计算，是GlocalKE相较于普通滑动窗口抽取方法引入的主要额外开销. 若直接构建两两距离矩阵，其时间复杂度约为$O\left(M^2 d\right)$、空间复杂度约为$O\left(Md + M^2\right)$. 不过，候选实体提及数量M通常远小于全文长度N，可通过实体类型预分组、低置信候选过滤和语义向量缓存等方式降低实际开销. 总体来看，GlocalKE以一定的后处理成本换取跨窗口实体一致性、证据可追溯性和逻辑约束能力，适合在长文本知识抽取场景中采用窗口并行和候选剪枝等方式进行部署优化.

4　总结与未来工作

本文围绕长文本知识抽取中上下文窗口受限及生成式模型事实偏差较为突出的问题，提出并验证了GlocalKE协同抽取框架. 该框架通过全局视野层建立跨窗口实体索引，尽量减少机械切分对长距离语义关联造成的破坏；同时，在后续推理过程中引入领域本体约束，对候选三元组进行一致性校验和必要修正. 实验结果表明，在本文采用的生成式推理设置下，GlocalKE能在一定程度上提高跨段落关系识别的准确性，并改善抽取结果在逻辑层面的整体一致性.

GlocalKE的计算代价主要集中在大语言模型推理与语义编码两个部分. 由于文本是通过局部滑动窗口送入模型的，系统无需一次性接收超长上下文，因而单次推理的显存压力相对更稳定，出现显存溢出的风险也相应降低. 相比之下，一致性校验阶段带来的附加开销较为有限，不会明显改变整体计算结构. 实际部署时可根据任务需求和资源条件灵活调整窗口大小及批次规模，从而兼顾抽取性能与计算代价.

后续工作将进一步关注多源异构数据条件下的知识融合与知识推理问题，并考察该框架在教育等应用场景中构建动态知识图谱时的少样本适应能力. 同时，针对跨领域长文本中的实体规范化和复杂关系建模，仍有必要开展更细致的研究，提升方法在真实复杂环境中的泛化能力和稳定性.

参考文献:

[1] Liu Z H, Xiong C Y, Sun M S, et al. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval[C/OL] //ACL, 2018: 2395-2405. https://aclanthology.org/P18-1223/.

[2] Saxena A, Tripathi A, Talukdar P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings[C/OL] //ACL, 2020: 4498-4507. https://aclanthology.org/2020.acl-main.412/.

[3] Wang X, He X N, Cao Y X, et al. KGAT: Knowledge graph attention network for recommendation[C/OL] //KDD, 2019: 950-958. https://doi.org/10.1145/3292500.3330989.

[4] Zeng D J, Liu K, Chen Y B, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C/OL] //ACL, 2015: 1753-1762. https://aclanthology.org/D15-1203/.

[5] Yao Y, Ye D M, Li P, et al. DocRED: A large-scale document-level relation extraction dataset[C/OL] //ACL, 2019: 764-777. https://aclanthology.org/P19-1074/.

[6] Nan G S, Guo Z J, Sekulić I, et al. Reasoning with latent structure refinement for document-level relation extraction[C/OL] //ACL, 2020: 1546-1557. https://aclanthology.org/2020.acl-main.141/.

[7] Lu Y J, Liu Q, Dai D, et al. Unified structure generation for universal information extraction[C/OL] //ACL, 2022: 5755-5772. https://aclanthology.org/2022.acl-long.395/.

[8] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C/OL] //NeurIPS, 2020: 1877-1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_source=transaction&utm_medium=email&utm_campaign=linkedin_newsletter.

[9] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.

[10] Zhu C, Ping W, Xiao C W, et al. Long-short Transformer: Efficient Transformers for language and vision[C/OL] //NeurIPS, 2021: 13681-13693. https://proceedings.neurips.cc/paper/2021/hash/9425be43ba92c2b4454ca7bf602efad8-Abstract.html.

[11] Ji Z W, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 248.

[12] Li J Y, Cheng X X, Zhao X, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models[C/OL] //EMNLP, 2023: 6449-6464. https://aclanthology.org/2023.emnlp-main.397/.

[13] Pan S R, Luo L H, Wang Y F, et al. Unifying large language models and knowledge graphs: A roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580-3599.

[14] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(6): 1083-1106.

[15] Li Q, Ji H. Incremental joint extraction of entity mentions and relations[C/OL] //ACL, 2014: 402-412. https://aclanthology.org/P14-1038/.

[16] Wei Z P, Su J L, Wang Y, et al. A novel cascade binary tagging framework for relational triple extraction[C/OL] //ACL, 2020: 1476-1488. https://aclanthology.org/2020.acl-main.136/.

[17] Wang Y C, Yu B W, Zhang Y Y, et al. TPLinker: Single-stage joint extraction of entities and relations through token pair linking[C/OL] //ACL, 2020: 1572-1582. https://aclanthology.org/2020.coling-main.138/.

[18] Zeng S, Xu R X, Chang B B, et al. Double graph based reasoning for document-level relation extraction[C/OL] //EMNLP, 2020: 1630-1640. https://aclanthology.org/2020.emnlp-main.127/.

[19] Xu W, Chen K H, Zhao T J. Discriminative reasoning for document-level relation extraction[C/OL] //ACL-IJCNLP, 2021: 1653-1663. https://aclanthology.org/2021.findings-acl.144/.

[20] Zhou W X, Huang K, Ma T, et al. Document-level relation extraction with adaptive thresholding and localized context pooling[C/OL] //AAAI, 2021: 14612-14620. https://doi.org/10.1609/aaai.v35i16.17717.

[21] Xu B F, Wang Q, Lyu Y J, et al. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction[C/OL] //AAAI, 2021: 14149-14157. https://doi.org/10.1609/aaai.v35i16.17665.

[22] Wang X, Zhou W K, Zu C, et al. InstructUIE: Multi-task instruction tuning for unified information extraction[PP/OL]. V1. arXiv(2023-04-17)[2026-01-07]. https://doi.org/10.48550/arXiv.2304.08085.

[23] Shen Y L, Tan Z Q, Wu S H, et al. PromptNER: Prompt locating and typing for named entity recognition[C/OL] //ACL,

2023：12492-12507．https：//aclanthology.org/2023.acl-long.698/．

［24］Wei J，Wang X Z，Schuurmans D，et al．Chain-of-thought prompting elicits reasoning in large language models［C/OL］//NeurIPS，2022：24824-24837．https：//proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html．

［25］Bi Z，Chen J，Jiang Y N，et al．CodeKGC：Code language model for generative knowledge graph construction［J］．ACM Transactions on Asian and Low-Resource Language Information Processing，2024，23（3）：45．

［26］Wei X，Cui X Y，Cheng N，et al．ChatIE：Zero-shot information extraction via chatting with chatGPT［PP/OL］．V2．arXiv（2024-05-27）［2026-01-07］．https：//doi.org/10.48550/arXiv.2302.10205．

［27］Sainz O，García-Ferrero I，Agerri R，et al．GoLLIE：Annotation guidelines improve zero-shot information-extraction［C/OL］//ICLR，2024：16853-16870．https：//proceedings. iclr. cc/paper_files/paper/2024/hash/cda04d7ea67ea1376bf8c6962d8541e0-Abstract-Conference.html．

［28］Lewis P，Perez E，Piktus A，et al．Retrieval-augmented generation for knowledge-intensive NLP tasks［C/OL］//NeurIPS，2020：9459-9474．https：//proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html．

［29］Chevalier A，Wettig A，Ajith A，et al．Adapting language models to compress contexts［C/OL］//EMNLP，2023：3829-3846．https：//aclanthology.org/2023.emnlp-main.232/．

［30］Pan L M，Wu X B，Lu X Y，et al．Fact-checking complex claims with program-guided reasoning［C/OL］//ACL，2023：6981-7004．https：//aclanthology.org/2023.acl-long.386/．

［31］Geng S B，Josifoski M，Peyrard M，et al．Grammar-constrained decoding for structured NLP tasks without finetuning［C/OL］//EMNLP，2023：10932-10952．https：//aclanthology.org/2023.emnlp-main.674/．

［32］Li S J，He W，Shi Y B，et al．DuIE：A large-scale Chinese dataset for information extraction［C/OL］//NLPCC，2019：791-800．https：//link.springer.com/chapter/10.1007/978-3-030-32236-6_72．

责任编辑：张自强