iText2KG: Incremental Knowledge Graphs Construction Using Large Language Models

iText2KG: 使用大型语言模型的增量知识图谱构建

Yassir LAIRGI 1,2[0000−0002−7284−5489] , Ludovic

Yassir LAIRGI 1,2[0000−0002−7284−5489] , Ludovic

MONCLA 1[0000−0002−1590−9546] , Rémy CAZABET 1[0000−0002−9429−3865] , Khalid BENABDESLEM1, and Pierre CLÉAU2

MONCLA 1[0000−0002−1590−9546] , Rémy CAZABET 1[0000−0002−9429−3865] , Khalid BENABDESLEM1, and Pierre CLÉAU2

1 INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621

1 INSA Lyon, CNRS, Claude Bernard Lyon 1 大学, LIRIS, UMR5205, 69621

Villeurbanne {ludovic.moncla, remy.cazabet, khalid.benabdeslem}@liris.cnrs.fr

Villeurbanne {ludovic.moncla, remy.cazabet, khalid.benabdeslem}@liris.cnrs.fr

2 GAUC, Lyon France {yassir.lairgi, pierre.cleau}@auvalie.com

2 GAUC, Lyon France {yassir.lairgi, pierre.cleau}@auvalie.com

Abstract. Most available data is unstructured, making it challenging to access valuable information. Automatically building Knowledge Graphs (KGs) is crucial for structuring data and making it accessible, allowing users to search for information effectively. KGs also facilitate insights, inference, and reasoning. Traditional NLP methods, such as named entity recognition and relation extraction, are key in information retrieval but face limitations, including the use of predefined entity types and the need for supervised learning. Current research leverages large language models' capabilities, such as zero- or few-shot learning. However, unresolved and semantically duplicated entities and relations still pose challenges, leading to inconsistent graphs and requiring extensive post-processing. Additionally, most approaches are topic-dependent. In this paper, we propose iText2KC3 a method for incremental, topic-independent KG construction without post-processing. This plug-and-play, zero-shot method is applicable across a wide range of KG construction scenarios and comprises four modules: Document Distiller, Incremental Entity Extractor, Incremental Relation Extractor, and Graph Integrator and Visualization. Our method demonstrates superior performance compared to baseline methods across three scenarios: converting scientific papers to graphs, websites to graphs, and CVs to graphs.

摘要。大多数可用数据是非结构化的，难以获取有价值信息。自动构建知识图谱 (KG) 对于结构化数据并使其可访问至关重要，从而使用户能够有效检索信息。知识图谱还利于洞察、推理与推断。传统的 NLP 方法，如命名实体识别与关系抽取，是信息检索的关键，但存在局限性，包括使用预定义实体类型和需要有监督学习。当前研究利用大型语言模型的能力，如零样本或少样本学习。然而，未解决的语义重复实体与关系仍然构成挑战，导致图谱不一致并需要大量后处理。此外，大多数方法依赖于主题。本文中，我们提出 iText2KC3 一种用于增量、主题无关且无需后处理的知识图谱构建方法。该即插即用、零样本方法适用于广泛的 KG 构建场景，由四个模块组成: 文档提炼器、增量实体抽取器、增量关系抽取器，以及图谱整合与可视化。我们的方法在三种场景上均优于基线方法: 将学术论文转为图谱、网站转为图谱与简历转为图谱。

Keywords: Knowledge Graph Construction . Large Language Models . Natural Language Processing.

关键词: 知识图谱构建. 大型语言模型. 自然语言处理.

1 Introduction

1 引言

In the contemporary era, most data is unstructured, leading to substantial information loss if not effectively harnessed [3]. This unstructured data lacks a predefined format, posing significant challenges for traditional data processing methodologies. Consequently, organizations must employ advanced text understanding and information extraction techniques to analyze and extract meaningful insights from this data effectively.

在当代，大多数数据是非结构化的，若不能有效利用会导致大量信息损失 [3]。这种非结构化数据缺乏预定义格式，对传统数据处理方法构成重大挑战。因此，组织必须采用先进的文本理解和信息抽取技术来有效分析并从中提取有意义的见解。

Text understanding and information extraction are key tasks in Natural Language Processing (NLP) for automatically processing data from unstructured text documents. The rise of Transformer architectures and pre-trained large language models (LLMs) opens new perspectives for extracting and structuring information from vast amounts of natural language texts [5]. One main aspect deals with Knowledge graphs (KGs) construction. KGs structure representations of knowledge by capturing relationships between entities and hold considerable advantages in analyzing text data collections and inferring knowledge from structured heterogeneous data. For instance, KGs can merge diverse data from multiple sources, offering a cohesive information perspective. They can also give an additional level of explainability to the analysis of text corpora.

[3] The code and the dataset are available at https://github.com/AuvaLab/itext2kg

[3] 代码和数据集可在 https://github.com/AuvaLab/itext2kg 获取

文本理解与信息抽取是自然语言处理 (NLP) 中用于自动处理非结构化文本文件数据的关键任务。Transformer 架构和预训练大型语言模型 (LLMs) 的兴起为从大量自然语言文本中提取和结构化信息开辟了新视角 [5]。一个主要方向是知识图谱 (KG) 构建。知识图谱通过捕捉实体之间的关系来构建知识表示，在分析文本数据集合和从异构结构化数据中推断知识方面具有显著优势。例如，知识图谱可以整合来自多个来源的多样数据，提供一致的信息视角，并可为文本语料分析提供额外的可解释性。

Named Entity Recognition, Relation Extraction, and Entity Resolution are NLP techniques usually utilized to transform unstructured text into structured data, capturing entities, their connections, and associated attributes [9, 10]. However, these methods encounter several limitations [18]. They are frequently restricted to predefined entities and relationships or depend on specific ontologies and mostly rely on supervised learning methods, necessitating extensive human annotation.

命名实体识别、关系抽取与实体消歧是通常用于将非结构化文本转为结构化数据的 NLP 技术，用以捕获实体、它们的连接及相关属性 [9, 10]。然而，这些方法面临若干限制 [18]。它们常常受限于预定义的实体和关系或依赖特定本体，并且多数依赖有监督学习方法，需大量人工标注。

To address these challenges, we aim to leverage LLMs in constructing KGs. Recent advancements in LLMs have shown potential and improved performance across a various range of NLP tasks, including knowledge graph completion, ontology refinement, and question answering, offering promising prospects for KG construction [8]. LLMs also show great ability for few-shot learning, enabling plug-and-play solutions, and eliminating the necessity for extensive training or fine-tuning. They can be used to extract knowledge across diverse domains due to their training in a wide range of information sources [14].

为应对这些挑战，我们旨在利用 LLMs 构建知识图谱。近期 LLM 的进展已在多种 NLP 任务上展示潜力并提升性能，包括知识图谱补全、本体精炼与问答，为 KG 构建提供了有希望的前景 [8]。LLMs 还表现出强大的少样本学习能力，支持即插即用的解决方案，消除大规模训练或微调的必要性。由于它们在大量信息来源上训练，可用于跨多领域提取知识 [14]。

Consequently, recent research has started utilizing advancements in LLMs, especially their capabilities in few-shot learning in KGs construction tasks. However, unresolved and semantically duplicated entities and relations still pose significant challenges, leading to inconsistent graphs that require extensive postprocessing. These inconsistencies can manifest as redundancies, ambiguities, and a real difficulty for graph extension. Additionally, many current approaches are topic-dependent, meaning their effectiveness heavily relies on the specific use case they are designed to handle. This dependency limits the generalizability of these methods across different domains, necessitating customized solutions for each new topic area.

因此，近期研究开始利用大型语言模型的进展，尤其是在知识图谱构建任务中的少量样本学习能力。然而，未解决和语义重复的实体与关系仍然带来重大挑战，导致图谱不一致并需要大量后处理。这些不一致可表现为冗余、歧义，并严重阻碍图谱扩展。此外，许多现有方法依赖于特定主题，其有效性高度依赖于设计所针对的具体用例，这限制了方法在不同领域的泛化性，需要为每个新主题定制解决方案。

In this paper, we propose iText2KG, a zero-shot method to construct consistent KGs from raw documents incrementally, using an LLM. It comprises four modules: 1) Document Distiller reformulates the raw documents, by taking a schema or a blueprint, into predefined and semantic blocks using LLMs. The schema operates like a predefined JSON structure, directing the language model to extract specific textual information associated with particular keys from each document, 2) iEntities Extractor takes the semantic blocks and not only identifies unique semantic entities within the semantic blocks but also resolves any ambiguities, ensuring that each entity is clearly defined and distinguished from others, 3) iRelation Extractor processes the resolved entities along with the semantic blocks to detect the semantically unique relationships. Further details are in the next sections. The final module employs Neo4j 4 to represent these relationships and entities in a graph format visually.

在本文中，我们提出 iText2KG，一种使用 LLM 从原始文档增量构建一致性知识图谱的零样本方法。它由四个模块组成:1)Document Distiller 将原始文档在给定模式或蓝图下重新表述为预定义的语义块，使用 LLM 将文档中与特定键相关的文本信息抽取到类似预定义 JSON 结构的模式指引下；2)iEntities Extractor 接收语义块，不仅识别其中的唯一语义实体，还解决任何歧义，确保每个实体被清晰定义并与其他实体区分开来；3)iRelation Extractor 处理已解析的实体及语义块以检测语义上唯一的关系。进一步细节见后文。最终模块使用 Neo4j 4 以图形方式表示这些关系和实体。

2 Related works

2 相关工作

LLM-based solutions for building KGs can be categorized according to three paradigms: ontology-guided, fine-tuning, and zero- or few-shot learning.

基于 LLM 的知识图谱构建方案可按三种范式分类: 本体引导、微调，以及零/少样本学习。

The AttacKG+ method, a fully automatic LLM-based framework for constructing attack KGs and capturing the progressive stages of cyber attacks, was introduced by [13]. The framework consists of four modules: rewriter, parser, identifier, and summarizer. The rewriter filters out redundant information and organizes report content into sections to preserve key knowledge, pre-cleans data, and sequences events chronologically. Guided by an ontology, the parser extracts threat actions using a triplet model (subject, action, object). The identifier matches these behavior graphs and rewritten sections to the appropriate format. Finally, the summarizer provides an overview of the situation and state at the end of each tactical stage. A theme-specific KG (ThemeKG) was proposed 2, constructed from a theme-specific corpus using an unsupervised framework (TKGCon) to address two main issues: limited information granularity and deficiency in timeliness. This approach generates KGs with accurate entities and relations by leveraging common sense knowledge from Wikipedia and LLMs for ontology guidance. Their model surpasses GPT-4 in performance due to its consistently precise identification of entities and relations

AttacKG+ 方法由 [13] 提出，是一个用于构建攻击知识图谱并捕捉网络攻击进展阶段的全自动 LLM 框架。该框架由四个模块组成:rewriter、parser、identifier 和 summarizer。rewriter 过滤冗余信息并将报告内容组织成章节以保留关键知识，预清洗数据并按时间顺序排列事件。在本体引导下，parser 使用三元组模型 (主体、动作、客体) 抽取威胁行为。identifier 将这些行为图和重写后的章节匹配为适当格式。最后，summarizer 在每个战术阶段结束时提供情况和状态概览。提出了主题特定的知识图谱 (ThemeKG)2，使用无监督框架 (TKGCon) 从主题语料构建，以解决信息粒度有限和时效性不足两大问题。该方法通过利用维基百科的常识知识和 LLM 进行本体引导，生成具有准确实体和关系的知识图谱。由于其对实体和关系的持续精确识别，该模型在性能上优于 GPT-4。

Text2KGBench, a benchmark designed to evaluate the capabilities of language models to generate KGs from natural language text guided by an ontology, was presented by [8]. They define seven evaluation metrics to measure fact extraction performance, ontology conformance, and hallucinations. A semiautomatic method for constructing KGs using open-source LLMs was introduced in recent research [7]. Their pipeline includes formulating competency questions (CQs) and developing an ontology derived from them. To assess the accuracy of the generated answers, they devised a judge LLM, which evaluates the content against ground truth. One major challenge with these proposed methods is their difficulty in generalizing their applicability to diverse KG construction scenarios due to their ontology dependency. The Wikipedia concept graph is also not exhaustive, particularly for country-specific concepts. For instance, it may not adequately cover terms like ”French Research Collaboration Tax Credit”.

Text2KGBench 是由 [8] 提出的一项基准，用于评估语言模型在本体引导下从自然语言文本生成知识图谱的能力。他们定义了七项评估指标以衡量事实抽取性能、本体一致性和幻觉现象。近期研究 [7] 提出一种使用开源 LLM 构建知识图谱的半自动方法，其流程包括制定能力问题 (CQs) 并由此开发本体。为评估生成答案的准确性，他们设计了一个评判 LLM，将内容与真值进行对照评估。这些方法的一个主要挑战是由于对本体的依赖，难以推广到多样的知识图谱构建场景。维基百科概念图也并不完备，尤其是在国家特定概念方面，例如可能无法充分覆盖“法国科研合作税收抵免”等术语。

An LLM was employed for building a KG from unstructured open-source threat intelligence [4]. This approach involves generating a dataset utilizing the zero-shot capability of GPT-3.5. Subsequently, this dataset is utilized for fine-tuning a smaller language model. One major challenge of this method is adapting it to different KG construction scenarios. Especially, the few-shot methods are more resource-efficient than the fine-tuned solutions [12].

有研究使用 LLM 从非结构化的开源威胁情报构建知识图谱 [4]。该方法利用 GPT-3.5 的零样本能力生成数据集，随后用该数据集对较小的语言模型进行微调。该方法的一大挑战是将其适配到不同的知识图谱构建场景。特别地，少样本方法在资源效率上优于微调方案 [12]。

[4] https://neo4j.com/

[4] https://neo4j.com/

An iterative LLM prompting-based pipeline for automatically generating knowledge graphs, which bypasses the need for predefined sets or external ontologies, was proposed by [1]. This pipeline employs a sequence of well-formed LLM prompts for each stage, enabling the identification of relevant entities, extracting their descriptions and types, and identifying meaningful relationships. The authors proposed an approach to entity/relation resolution using semantic aggregation and LLM prompting. It starts with semantic aggregation, calculating similarity scores for entities and relations based on label similarity, entity type similarity, and description similarity using methods like Levenshtein distance and cosine similarity with the Universal Sentence Encoder model. The entities and relations are aggregated if their scores exceed predefined thresholds. Even though the proposed approaches present several advantages, it has certain limitations: (1) The entity/relation resolution phase aggregates nodes and relations having the same meaning, and then the LLM suggests a representative for each cluster based on the cluster elements. This could hinder the precision of the graph, especially if ”bike” and ”motorcycle” need to be separated. Still, the model merges them into ”vehicle.” (2) The latter phase involves post-processing, which could be computationally intensive. (3) The post-processing phase assumes that entities and relations are extracted. Hence, if entities are not resolved before relation extraction, redundant relations from redundant entities could arise, worsening the quality of the relation extraction.

一种基于迭代大语言模型提示的自动生成知识图谱的流水线，被提出以绕过预定义集合或外部本体的需求 [1]。该流水线为每个阶段使用一系列规范化的 LLM 提示，从而识别相关实体、提取其描述与类型并识别有意义的关系。作者提出了使用语义聚合和 LLM 提示的实体/关系解析方法。它从语义聚合开始，基于标签相似度、实体类型相似度和描述相似度 (使用如 Levenshtein 距离和基于 Universal Sentence Encoder 的余弦相似度) 计算实体和关系的相似分数。当分数超过预设阈值时，实体与关系会被聚合。尽管所提出的方法具有若干优点，但存在一定限制: (1) 实体/关系解析阶段将具有相同含义的节点和关系聚合，然后由 LLM 基于簇内元素建议每个簇的代表。这可能削弱图的精确性，尤其当“bike”和“motorcycle”需要区分时，模型仍可能将它们合并为“vehicle”。(2) 后续阶段涉及后处理，可能计算开销大。(3) 后处理阶段假设实体与关系已被提取。因此，如果在关系抽取之前实体未被解析，来自冗余实体的冗余关系可能出现，降低关系抽取的质量。

A comprehensive quantitative and qualitative evaluation of LLMs for KG construction and reasoning was provided [14], using eight diverse datasets across four representative tasks: entity and relation extraction, event extraction, link prediction, and question-answering. Key findings reveal that while GPT-4 performs well in KG construction tasks, it excels even more in reasoning tasks, sometimes surpassing fine-tuned models. The paper also proposes AutoKG, a multi-agent-based approach that utilizes LLMs and external sources for KG construction and reasoning.

一项针对知识图谱构建与推理的大语言模型的全面定量与定性评估被提供 [14]，使用八个多样数据集覆盖四类代表性任务: 实体与关系抽取、事件抽取、链接预测与问答。主要发现表明，虽然 GPT-4 在知识图谱构建任务中表现良好，但在推理任务中表现更为出色，有时甚至超过微调模型。论文还提出了 AutoKG，一种利用 LLM 与外部资源进行知识图谱构建与推理的多代理方法。

3 Incremental Text2KG

3 增量 Text2KG

This work aims to develop a plug-and-play solution for constructing KGs from documents with resolved entities and relations as output. Adopting a ’zero-shot’ approach is essential to ensure the solution’s applicability across various KG construction scenarios. This approach means that the prompts used to generate the KG do not require prior examples or predefined ontologies.

本工作旨在开发一种即插即用的从文档构建知识图谱的解决方案，输出为已解析的实体与关系。采用“零样本”方法对确保解决方案在各种知识图谱构建场景中的适用性至关重要。该方法意味着用于生成知识图谱的提示不需要先前示例或预定义本体。

3.1 Problem Formulation

3.1 问题表述

A graph can be defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ where $\mathcal{E}$ is the set of nodes and $\mathcal{R}$ denotes the set of edges [5]. Considering the difficulty in merging similar concepts, we defined two constraints for the solution:

图可以定义为 $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ ，其中 $\mathcal{E}$ 是节点集合且 $\mathcal{R}$ 表示边集合 [5]。考虑到合并相似概念的困难，我们为该解决方案定义了两个约束:

(C1) An entity $e_i \in \mathcal{E}$, the set of entities and a relation $r_k \in \mathcal{R}$, the set of relations, should each describe a semantically unique concept.

(C1) 一个实体 $e_i \in \mathcal{E}$ ，实体集合，以及一个关系 $r_k \in \mathcal{R}$ ，关系集合，应各自描述一个语义上唯一的概念。

(C2) The sets of entities and relations should contain semantically unique elements. This means each entity and relation within the knowledge graph must be distinct and unique, with no duplication or semantic overlaps.

(C2) 实体集合与关系集合应包含语义上唯一的元素。这意味着知识图谱中的每个实体与关系必须各自不同且唯一，不允许重复或语义重叠。

These constraints can be mathematically formulated as follows.

这些约束可以按如下数学形式表述。

$$\forall e_i, e_j \in \mathcal{E}, i \neq j \Rightarrow e_i \neq e_j \tag{1}$$

$$\forall r_k, r_l \in \mathcal{R}, l \neq k \Rightarrow r_k \neq r_l \tag{2}$$

3.2 Proposed method

3.2 所提方法

We propose the iText2KG approach composed of four modules (see Figure 1): Document Distiller, Incremental Entities Extractor, Incremental Relations Extractor, and Neo4j Graph Integrator. Each module fulfills a distinct role in constructing the KG. Notably, entity extraction and relation extraction tasks are separated following results described in 1 that positively impact the performance. Further details of modules 1 to 3 are as follows, with the fourth module serving to visualize the graph.

我们提出了由四个模块组成的 iText2KG 方法 (见图 1): 文档提炼器、增量实体提取器、增量关系提取器与 Neo4j 图整合器。每个模块在构建知识图谱中承担不同角色。值得注意的是，实体抽取与关系抽取任务被分离，依据文献 1 中所述的结果，这一分离对性能产生积极影响。模块 1 至 3 的更多细节如下，第四模块用于可视化图谱。

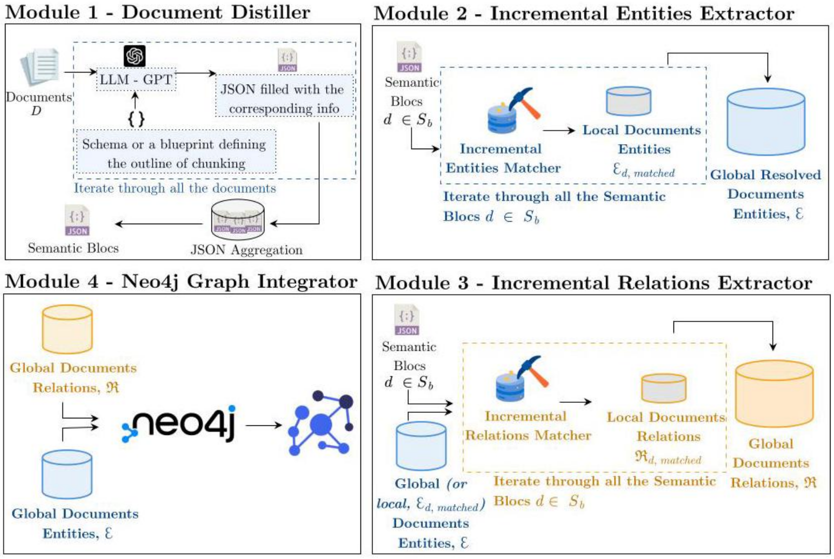
Module 1 - Document Distiller
Documents
D
LLM - GPT
JSON
JSON filled with the corresponding info
{}
Schema or a blueprint defining the outline of chunking
Iterate through all the documents
Semantic Blocs
JSON Aggregation
Module 2 - Incremental Entities Extractor
Semantic Blocs
$d \in S_b$
Incremental Entities Matcher
Local Documents Entities
$\mathcal{E}_{d,\ matched}$
Global Resolved Documents Entities, $\mathcal{E}$
Iterate through all the Semantic Blocs $d \in S_b$
Module 4 - Neo4j Graph Integrator
Global Documents Relations, $\mathfrak{R}$
neo4j
Global Documents Entities, $\mathcal{E}$
Module 3 - Incremental Relations Extractor
Semantic Blocs
$d \in S_b$
Incremental Relations Matcher
Local Documents Relations
$\mathfrak{R}_{d,\ matched}$
Global Documents Relations, $\mathfrak{R}$
Global (or local, $\mathcal{E}_{d,\ matched}$) Documents Entities, $\mathcal{E}$
Iterate through all the Semantic Blocs $d \in S_b$

Fig.1. The overall workflow of the iText2KG modules. Module 3, the Incremental Relations Extractor, operates differently depending on whether global or local document entities are provided as context.

图 1. iText2KG 各模块的总体工作流程。模块 3，增量关系提取器，根据是否提供全局或局部文档实体作为上下文而以不同方式运行。

Module 1 - Document Distiller: This module uses LLMs to rewrite input documents into semantic blocks, considering a predefined schema or blueprint. It is important to note that the schema is not an ontology but a blueprint that biases the LLM towards specific classes while maintaining flexibility in others. Practically, the schema functions like a predefined JSON, instructing the LLM to extract particular values (textual information) for specific keys from each document. Some examples of blueprints are available in the iText2KG Github repository. For each document, we will obtain a JSON semi-filled with the desired information if it exists in the document. Then, we aggregate all these semi-filled JSONs to form the semantic blocks of the documents. We have used Langchain's JSON Parser5 to define the schema along with the documents as context. The main goals of this module are: (a) To improve the signal-to-noise ratio by reducing noise that may pollute the graph with redundant information. (b) To guide the graph construction process using the schema, especially for concept keys. For example, for a scientific article, we could extract the ”title” and the ”authors” and add relations like ”HAS TITLE” and ”HAS AUTHORS” in addition to the semantic information. To ensure the applicability of our solution across various use cases, the schema is an input that depends on user preferences and the particularity of the use case. The idea of reformulating raw documents to enhance the graph construction process has been proven by the following papers 1311]. The two papers as mentioned earlier introduced a rewriter module, but it depends on the article's specific use case. However, our module is adaptable to many use cases.

> 模块 1 - 文档提炼器: 该模块使用大型语言模型将输入文档重写为语义块，参照预定义的模式或蓝图。需注意，该模式不是本体，而是一个偏向特定类别同时对其它类别保持灵活性的蓝图。实际上，模式类似预定义的 JSON，指示 LLM 从每份文档中为特定键提取特定值 (文本信息)。iText2KG Github 仓库中提供了一些蓝图示例。对于每份文档, 我们会得到一个半填充的 JSON, 包含文档中存在的所需信息。然后将所有这些半填充 JSON 聚合为文档的语义块。我们使用了 Langchain 的 JSON Parser5 来定义模式，并把文档作为上下文。该模块的主要目标是:(a) 通过减少可能用冗余信息污染图谱的噪声来提高信噪比；(b) 使用模式引导图谱构建过程，尤其是概念键。例如，对于一篇科学文章，我们可以提取“title”和“authors”，并在语义信息外添加“HAS TITLE”和“HAS AUTHORS”之类的关系。为确保解决方案适用于各种用例, 模式作为输入取决于用户偏好和具体用例的特殊性。通过将原始文档重构以增强图谱构建过程的想法已在以下文献中得到验证 1311]。前述两篇论文引入了一个重写模块，但其依赖于文章的具体用例；而我们的模块可适配多种用例。

Module 2 - Incremental Entities Extractor: The Incremental Entities Matcher (iEntities Matcher) iterates over all the Semantic Blocks and extracts the Global Document Entities. The main algorithm of iEntities Matcher is presented in Figure 2. Initially, entities are extracted from the first semantic block (document) d_0 using an LLM, forming the global entity set $\mathcal{E}$ under the assumption that these entities are pairwise distinct for this first iteration only. Considering the constraint (C1), the LLM is prompted to extract entities representing one unique concept to avoid semantically mixed entities (prompts are presented in the iText2KG GitHub repository).

> 模块 2 - 增量实体提取器: 增量实体匹配器 (iEntities Matcher) 遍历所有语义块并提取全局文档实体。iEntities Matcher 的主要算法如图 2 所示。最初，使用 LLM 从第一个语义块 (文档) d_0 提取实体，形成全局实体集 $\mathcal{E}$ ，并假定在首次迭代中这些实体两两互不相同。考虑约束 (C1)，LLM 被提示提取代表唯一概念的实体以避免语义混杂的实体 (提示在 iText2KG GitHub 仓库中提供)。

For subsequent documents d in D , the algorithm extracts local entities $\mathcal{E}_d$. It then attempts to match these local entities with the global entities in $\mathcal{E}$. If a local entity e_i is found in $\mathcal{E}$, it is added to the matched set $\mathcal{E}_{d,\text{ matched}}$. If not, the algorithm searches for a similar entity in $\mathcal{E}$ using a cosine similarity measure with a predefined threshold. If no match is found, the local entity is added to $\mathcal{E}_{d,\text{ matched}}$; otherwise, the best matching global entity e_i' (based on maximum similarity) is added. The global entity set $\mathcal{E}$ is then updated by unifying it with $\mathcal{E}_{d,\text{ matched}}$. This process is repeated for each document in D , resulting in a comprehensive global entity set $\mathcal{E}$.

> 对于随后文档中的 d 在 D 中，算法提取局部实体 $\mathcal{E}_d$ 。然后尝试将这些局部实体与 $\mathcal{E}$ 中的全局实体匹配。如果在 $\mathcal{E}$ 中找到某个局部实体 e_i ，则将其加入匹配集合 $\mathcal{E}_{d,\text{ matched}}$ 。若未找到，则算法使用带预定义阈值的余弦相似度在 $\mathcal{E}$ 中搜索相似实体。若仍无匹配，则将该局部实体加入 $\mathcal{E}_{d,\text{ matched}}$ ；否则，选择相似度最高的全局实体 e_i' 加入。随后通过将 $\mathcal{E}$ 与 $\mathcal{E}_{d,\text{ matched}}$ 统一来更新全局实体集 $\mathcal{E}$ 。对 D 中的每份文档重复此过程，最终得到完整的全局实体集 $\mathcal{E}$ 。

[5] https://python.langchain.com/v0.1/docs/modules/model_io/output_parsers/ types/json/

[5] https://python.langchain.com/v0.1/docs/modules/model_io/output_parsers/ types/json/

iEntities Matcher

iEntities Matcher

Initialisation :

初始化:

We extract the entities of the first document d_0 in the

我们从语义块中的第一份文档 d_0 提取实体

semantic blocks S_b . This initializes the set of global entities $\mathcal{E}$.

语义块 S_b 。这初始化了全局实体集合 $\mathcal{E}$ 。

$\mathcal{E}$ = ExtractEntitiesWithLLM $(d_0) = \{e_1, e_2, .., e_n\}$

$\mathcal{E}$ = ExtractEntitiesWithLLM $(d_0) = \{e_1, e_2, .., e_n\}$

And we assume that for this first document d_0 , the entities e_i

并且我们假定对于该第一份文档 d_0 ， 实体 e_i

are pairwise distinct (we prompt the LLM accordinly).

是两两互不相同的 (我们相应地提示 LLM)。

Entity Matching :

实体匹配:

For each document d in S_b except d_0 :

对于 S_b 中除 d_0 之外的每份文档 d :

#Local entities not yet matched with global

# 尚未与全局实体匹配的局部实体

entities in $\mathcal{E}$

```
位于 $\mathcal{E}$ 的实体
$\mathcal{E}_d = \text{ExtractEntitiesWithLLM}\,(d)$
```

$\mathcal{E}_d = \text{ExtractEntitiesWithLLM}\,(d)$

```
#Local entities matched with global entities in $\mathcal{E}$
## 在 $\mathcal{E}$ 中将本地实体与全局实体匹配
$\mathcal{E}_{d,\text{ matched}} = \varnothing$
    - For each entity in $\mathcal{E}_d$ :
    - - 对于 $\mathcal{E}_d$ 中的每个实体:
        - If $e_i \in \mathcal{E}$ :
        - - 如果 $e_i \in \mathcal{E}$ :
            $\mathcal{E}_{d,\text{ matched}} = \mathcal{E}_{d,\text{ matched}} \cup \{e_i\}$
        - Else :
        - - 否则:
            We will search for a match for $\mathbf{e}_i$ in $\mathcal{E}$
            我们将在 $\mathcal{E}$ 中为 $\mathbf{e}_i$ 搜索匹配项
$S_{\mathcal{E}}\,(e_i) = \left\{\text{CosineSim}\left(e_i, e_j\right) > \text{ Threshold } \mid e_j \in \mathcal{E}\right\}$
            - If $S_{\mathcal{E}}\,(e_i) = \varnothing$ :
            - - 如果 $S_{\mathcal{E}}\,(e_i) = \varnothing$ :
                $\mathcal{E}_{d,\text{ matched}} = \mathcal{E}_{d,\text{ matched}} \cup \{e_i\}$
            - Else :
            - - 否则:
                $e_i' = \text{argmax}_{e_j \in \mathcal{E}}\,(S_{\mathcal{E}}\,(e_i))$
                $\mathcal{E}_{d,\text{ matched}} = \mathcal{E}_{d,\text{ matched}} \cup \left\{e_i'\right\}$
 $= \mathcal{E}_{d,\text{ matched}} \cup \mathcal{E}$
```

Fig. 2. The algorithm of iEntities Matcher

图 2. iEntities Matcher 算法

Module 3 - Incremental Relations Extractor: The Global Document Entities $\mathcal{E}$ are provided as context along with each Semantic Block to the Incremental Relations Matcher (iRelations Matcher) to extract the Global Document Relations. The same approach used for iEntities Matcher applies here. We have observed different behaviors in relation extraction depending on whether global or local entities are used as context with the Semantic Block for the LLM. When global entities are provided as context, the LLM extracts both the relations directly stated and implied by the Semantic Block, especially for entities not explicitly present in the Semantic Block. This enriches the graph with potential information but increases the likelihood of irrelevant relations. Conversely, when locally matched entities are provided as context, the LLM only extracts the relations directly stated by the context. This approach reduces the richness of the graph but also lowers the probability of irrelevant relations. The two versions of iRelations Matcher are presented in Figure 3. This result will be further discussed in Section 4.

模块 3 - 增量关系抽取器: 将全局文档实体 $\mathcal{E}$ 与每个语义块一并作为上下文提供给增量关系匹配器 (iRelations Matcher)，以抽取全局文档关系。此处采用与 iEntities Matcher 相同的方法。我们观察到，关系抽取的表现会因将全局还是本地实体作为与语义块的上下文提供给大模型而异。当提供全局实体作为上下文时，模型会抽取语义块中直接陈述的关系以及隐含的关系，尤其是对于在语义块中未明确出现的实体。这会为图谱补充潜在信息，但也增加了无关关系的概率。相反，当提供本地匹配的实体作为上下文时，模型仅抽取上下文中直接陈述的关系。这种方法降低了图谱的信息丰富度，但也减少了无关关系的概率。iRelations Matcher 的两个版本见图 3。本结果将在第 4 节中进一步讨论。

---

iRelations Matcher [Local(or Global) Entities as Context]

iRelations Matcher [以本地 (或全局) 实体为上下文]

Initialisation :

初始化:

We extract the relations of the first document d_0 in the semantic blocks S_b .

我们在语义块 S_b 中抽取第一份文档 d_0 的关系。

We provide the local (or global) matched entities as context. This

我们将本地 (或全局) 匹配的实体作为上下文提供。本条

initializes the set of global relations $\mathfrak{R}$

初始化全局关系集 $\mathfrak{R}$

$\mathfrak{R} = \text{ExtractRelationsWith } LLM\left(d_0, \mathcal{E}_{0,\text{ matched}} \text{ (or } \mathcal{E})\right)$

$\mathfrak{R} = \text{ExtractRelationsWith } LLM\left(d_0, \mathcal{E}_{0,\text{ matched}} \text{ (or } \mathcal{E})\right)$

$= \{r_1, r_2, \dots, r_k\}$

And we assume that for this first document d_0 , the relations r_i

并且我们假设对于这第一份文档 d_0 ， 关系为 r_i

are pairwise distinct (we prompt the LLM accordinly).

彼此两两不同 (我们据此提示大型语言模型)。

Relation Matching :

关系匹配:

For each document d in S_b except d_0 :

对于除 d_0 之外的每个文档 d 在 S_b 中:

#Local relations not yet matched with global relations in $\mathfrak{R}$

在 $\mathfrak{R}$ 中尚未与全局关系匹配的本地关系

$\mathfrak{R}_d = \text{ExtractRelationsWithLLM}\left(d, \mathcal{E}_{d,\text{ matched}} \text{ (or } \mathcal{E})\right)$

$\mathfrak{R}_d =$ 使用 LLM 提取关系 $\left(d, \mathcal{E}_{d,\text{ matched}} \text{ (or } \mathcal{E})\right)$

#Local relations matched with global relations in $\mathfrak{R}$

在 $\mathfrak{R}$ 中已与全局关系匹配的本地关系

$\mathfrak{R}_{d,\text{ matched}} = \varnothing$

- For each relation in $\mathfrak{R}_d$:

- 对于 $\mathfrak{R}_d$ 中的每个关系:

- If $r_i \in \mathfrak{R}$:

- 如果 $r_i \in \mathfrak{R}$:

$\mathfrak{R}_{d,\text{ matched}} = \mathfrak{R}_{d,\text{ matched}} \cup \{r_i\}$

- Else :

- 否则:

We will search for a match for r_i in $\mathfrak{R}$

我们将在 $\mathfrak{R}$ 中搜索 r_i 的匹配项

$S_{\mathfrak{R}}(r_i) = \left\{\text{CosineSim}\left(r_i, r_j\right) > \text{ Threshold } \mid r_j \in \mathfrak{R}\right\}$

```
        - If $S_{\mathfrak{R}}(r_i) = \varnothing$ :
        - 如果 $S_{\mathfrak{R}}(r_i) = \varnothing$ :
            $\mathfrak{R}_{d,\text{ matched}} = \mathfrak{R}_{d,\text{ matched}} \cup \{r_i\}$
        - Else :
        - 否则:
            $r_i' = \text{argmax}_{r_j \in \mathfrak{R}}(S_{\mathfrak{R}}(r_i))$
            $\mathfrak{R}_{d,\text{ matched}} = \mathfrak{R}_{d,\text{ matched}} \cup \{r_i'\}$
    , $\mathfrak{R} = \mathfrak{R}_{d,\text{ matched}} \cup \mathfrak{R}$
Return $\mathfrak{R}$
返回 $\mathfrak{R}$
```

Fig. 3. The two versions of iRelations Matcher

图 3. iRelations Matcher 的两个版本

Module 4 - Graph Integrator: The Global Document Entities and the Global Document Relations are fed into Neo4j to construct the knowledge graph.

模块 4 - 图整合器: 全局文档实体和全局文档关系被输入 Neo4j 以构建知识图谱。

4 Experiments

4 实验

We chose GPT-4 in all our experiments due to its performance in KG construction and reasoning capabilities, as demonstrated by [14]. Notably, GPT-4 achieves near fine-tuned state-of-the-art performance, even in zero-shot scenarios. To validate our method, it is essential first to evaluate Module 1 to ensure the concordance of the extracted information with the schema and the semantics of the input documents. Moreover, evaluating modules 1 and 2 regarding the extracted triplets and the quality of entity/relation resolution is also important. To ensure the applicability of our method across different KG construction scenarios, we have adopted three use cases: website to KG, scientific article to KG, and Curriculum Vitae to KG.

我们在所有实验中均选择了 GPT-4，基于其在知识图谱构建和推理能力方面的表现 (见 [14])。值得注意的是，GPT-4 即便在零样本场景下也能达到接近微调后的最先进性能。为验证我们的方法，首先必须评估模块 1，以确保提取信息与模式以及输入文档的语义一致。此外，对模块 1 和模块 2 提取的三元组及实体/关系消解质量进行评估也很重要。为保证方法在不同知识图谱构建场景中的适用性，我们采用了三种用例: 网站到知识图谱、科研文章到知识图谱和简历到知识图谱。

We have adapted the metrics proposed by [13] for Module 1 to our use cases. Hence, we propose the following metrics:

我们已将 [13] 提出的度量指标针对模块 1 调整到我们的用例。因此，我们提出以下指标:

- Schema consistency: Evaluate whether the content of the rewritten text matches the input schema (the blueprint). For each key presented in the schema, we define $C_s(k)$ as the number of elements correctly matched to the schema related to the key $k.I_s(k)$ as the number of elements that were added but did not belong to the schema. The consistency score for a key in the schema is:

- 模式一致性: 评估重写文本的内容是否与输入模式 (蓝图) 匹配。对于模式中呈现的每个键，我们定义 $C_s(k)$ 为正确匹配到该键相关模式的元素数量，$k.I_s(k)$ 为被添加但不属于该模式的元素数量。该键在模式中的一致性得分为:

$$\mathrm{SC}\,(\mathrm{k}) = \frac{C_s(k) - I_s(k)}{T_s(k)} \tag{3}$$

Such as:

例如:

$T_s(k)$: The total elements in the schema corresponding to the key k .

$T_s(k)$: 模式中对应该键的元素总数 k 。

If $C_s(k) < I_s(k), SC(k) = 0$.

若 $C_s(k) < I_s(k), SC(k) = 0$ 。

Hence, the schema consistency score is :

因此，模式一致性得分为:

$$\mathrm{SC} = \sum_{k \in K} \frac{SC(k)}{\mathrm{card}\,(K)} \tag{4}$$

Where K is the set of keys of the schema.

其中 K 是模式的键集合。

- Information consistency: Evaluate whether the rewritten text's content matches the original report's semantics, categorized as follows: very different (<30%), medium (30-60%), largely consistent (60-90%), and fully consistent (>90%).

- 信息一致性: 评估重写文本的内容是否与原报告的语义相符，分为: 非常不同 (<30%)、中等 (30–60%)、大体一致 (60–90%) 和完全一致 (>90%)。

For the second and third modules, it is important to ensure that the extracted entities and relations are resolved and that the extracted triplets are relevant to the input documents. Therefore, we propose the following metrics:

对于第二和第三模块，重要的是确保提取的实体和关系被消解，且提取的三元组与输入文档相关。因此，我们提出以下度量:

- Triplet Extraction Precision: Evaluate the consistency of the triplets with the corresponding text regardless of the entity/relation resolution process. It is important to note that a relevant triplet is implied and not necessarily directly stated by the text. We define the precision score as the number of extracted relevant triplets divided by the total number of extracted triplets.

- 三元组提取精准度: 评估三元组相对于对应文本的一致性，而不考虑实体/关系消解过程。需注意，相关三元组可以是隐含的，并不必然在文本中直接陈述。我们将精准度定义为提取到的相关三元组数量除以提取到的三元组总数。

- Entity/Relation Resolution False Discovery Rate: Evaluate the proportion of unresolved (false positive) entities or relations among the total extracted entities or relations. Specifically, we calculate the ratio of unresolved entities or relations to the total number of extracted entities or relations. This metric provides a clear indication of the reliability of the entity and relation extraction process by highlighting the proportion of errors (unresolved entities/relations) within the total extractions.

- 实体/关系消解假发现率: 评估在提取的实体或关系总数中未被消解 (假阳性) 的比例。具体地，我们计算未被消解的实体或关系数与提取的实体或关系总数之比。该指标通过突出总提取中的错误比例 (未消解的实体/关系) 来明确表示实体和关系提取过程的可靠性。

4.1 Datasets and Baseline Methods

4.1 数据集与基线方法

To evaluate Document Distiller, we have generated 5 CVs using GPT-4, selected 5 company websites, and 5 scientific articles. It is important to note that we have extracted the textual information from websites, which will serve as input to our model.

为评估 Document Distiller，我们使用 GPT-4 生成了 5 份简历，选择了 5 个公司网站和 5 篇科研文章。需说明的是，我们已从网站中抽取文本信息，作为模型的输入。

To evaluate the consistency of triplets extracted by iEntities Extractor and iRelations Extractor, we used the annotated dataset from [6]. We observed that this dataset is not exhaustive for triplet extraction, leading us to conduct manual checks for triplets not present in the dataset. This manual check combined with the aforementioned dataset composes the ground truth. To assess the False Discovery Rate of the entity/relation resolution process, we performed the KG construction process using different baseline methods.

为了评估 iEntities Extractor 与 iRelations Extractor 提取三元组的一致性，我们使用了文献 [6] 中的带注释数据集。我们发现该数据集并不穷尽三元组提取的所有情况，因此对数据集中未包含的三元组进行了人工核查。该人工核查与前述数据集共同构成了金标准。为评估实体/关系解析过程的误报率 (False Discovery Rate)，我们使用不同的基线方法执行了知识图谱构建流程。

We have compared our method against baseline methods including Graph Construction using OpenAI Function Method [6, Langchain 7] and LlamaIndex 8

我们将本方法与包括使用 OpenAI Function 方法构建图谱 [6, Langchain 7] 及 LlamaIndex 8 在内的基线方法进行了比较

4.2 First Module Evaluation Results

4.2 第一个模块的评估结果

Schema Consistency Table 1 demonstrates that Document Distiller achieves high schema consistency across various document types. Scientific articles and CVs exhibit the highest schema consistency scores, indicating the module's capability to handle structured information, particularly for documents where the data is primarily organized using titles. While still achieving a strong score of 0.94, websites present a slightly lower consistency, which may be attributed to web content's varied and less structured nature. These results highlight the robustness and adaptability of Document Distiller in processing and extracting structured information from diverse document types.

模式一致性表 1 显示 Document Distiller 在多种文档类型上实现了较高的模式一致性。学术文章和简历的模式一致性得分最高，表明该模块能处理结构化信息，尤其适用于以标题组织数据的文档。网站的得分虽仍较高 (0.94)，但略低，可能因网页内容更为多样且结构性较差。上述结果突显了 Document Distiller 在处理并提取多种文档类型结构化信息方面的稳健性与适应性。

Table 1. The Schema Consistency Score for the different types of documents.

表 1. 不同类型文档的模式一致性得分。

Documents	CVs	Scientific Articles	Websites
Schema Consistency Score	0.97 ± 0.09	0.98 ± 0.04	0.94 ± 0.13

文件	简历	学术文章	网站
模式一致性得分	0.97 ± 0.09	0.98 ± 0.04	0.94 ± 0.13

Information Consistency Figure 4 illustrates the information consistency across different types of documents: CVs, scientific articles, and websites. For CVs, the majority of the information (74.5%) is fully consistent, with 25.5% being largely consistent and no medium consistency. This indicates that the rewritten text closely matches the semantics of the original content for CVs. This is because CVs are primarily written in clear and concise phrases, making it easier for the LLM to capture the semantics. In the case of scientific articles, 57.1% of the information is fully consistent, and 42.9% is largely consistent, showing a high degree of accuracy in preserving the original semantics, though slightly less than CVs. This is predictable, especially since scientific articles are written in scientific English with more complex phrases. Websites have 56.0% of information fully consistent, 24.0% largely consistent, and 20.0% medium consistency. This may be due to the unstructured nature of web content, which poses a greater challenge for accurate semantic rewriting.

信息一致性图 4 展示了不同类型文档 (简历、科学论文和网站) 之间的信息一致性: 在简历中，大部分信息 (74.5%) 完全一致，25.5% 基本一致，且无中等一致性。这表明重写文本与原文语义高度匹配，因为简历主要以清晰简练的短语书写，LLM 更易捕捉其语义。科学论文中，57.1% 的信息完全一致，42.9% 基本一致，表明在保留原始语义方面精确度高，但略低于简历，这在意料之中，尤其是科学论文使用更复杂的学术英文短语。网站中，56.0% 的信息完全一致，24.0% 基本一致，20.0% 为中等一致性。这可能由于网页内容结构不规范，给准确的语义重写带来更大挑战。

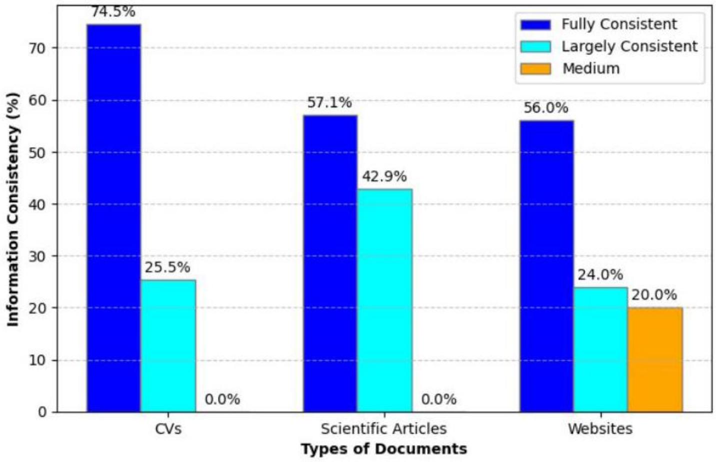

Fig. 4. Bar Plot of the Information Consistency Scores for the different types of Documents

图 4. 不同类型文档信息一致性分数的条形图

4.3 Second and Third Modules Evaluation Results

4.3 第二与第三模块的评估结果

[6] https://github.com/tomasonjo/blogs/blob/master/llm/openaifunction_ constructing_graph.ipynb

6 https://github.com/tomasonjo/blogs/blob/master/llm/openaifunction_ constructing_graph.ipynb

https://python.langchain.com/v0.1/docs/use_cases/graph/constructing/

[8] https://docs.llamaindex.ai/en/stable/examples/property_graph/property_ graph_basic/

[8] https://docs.llamaindex.ai/en/stable/examples/property_graph/property_ graph_basic/

Triplet Extraction Table 2 shows different behaviors in relation extraction depending on whether global or local entities are used as context with the Semantic Block for the LLM. The precision of relevant triplets when global entities are fed as context is 10% lower than that of relevant triplets when local entities are fed as context. When global entities are used as context, the LLM extracts relations explicitly mentioned and implied within the Semantic Block. This results in a richer graph with more potential information and a higher chance of irrelevant relations. On the other hand, using locally matched entities as context leads the LLM to extract only the directly stated relations, resulting in a less enriched graph but with a lower likelihood of irrelevant relations.

三元组抽取表 2 显示了在关系抽取中，根据在语义块中使用全局实体或局部实体作为上下文的不同行为。当以全局实体作为上下文时，相关三元组的精确率比以局部实体作为上下文时低 10%。当使用全局实体作为上下文时，LLM 会抽取语义块中明确提到的和暗示的关系，导致图谱更丰富、潜在信息更多，但也更可能包含无关关系。另一方面，使用局部匹配实体作为上下文会使 LLM 仅抽取直接陈述的关系，生成的图谱不那么丰富，但包含无关关系的可能性较低。

Table 2. Precision scores for relevant triplets across two datasets: music and computer science. The scores are presented for Global Entities as Context and Local Entities as Context.

表 2. 在两个数据集 (音乐与计算机科学) 上相关三元组的精确率。分列展示了以全局实体为上下文和以局部实体为上下文的得分。

	Global Entities	Local Entities
Computer Science Dataset	0.83 ± 0.06	0.94±0.06
Music Dataset	0.81 ± 0.05	0.9±0.07

	全局实体	本地实体
计算机科学数据集	0.83 ± 0.06	0.94±0.06
音乐数据集	0.81 ± 0.05	0.9±0.07

This presents a trade-off that depends on the use case. We leave it to the user to decide whether to accept a 10% decrease in precision in exchange for an enriched graph or to gain 10% precision with a less enriched graph.

这提出了一个取舍，取决于使用场景。我们由用户自行决定是接受在精度上 10% 的下降以换取更丰富的图，还是以较少的丰富度换取 10% 的精度。

Entity/Relation Resolution To the best of our knowledge, LlamaIndex constructs unconnected sub-graphs with edge-level and node-level textual information for retrieval-augmented generation (RAG); hence, we did not evaluate Lla-maIndex against our method. From Table 3 and Table 4, we conclude that our method delivers superior results for the entity and relation resolution process across three different KG construction scenarios: scientific articles to KG, CVs to KG, and websites to KG. Additionally, the results indicate that when the number of input documents is small and they are structured with clear, non-complex phrases, the LLM performs well in entity and relation resolution, as demonstrated by the CVs to KG process.

实体/关系解析据我们所知，LlamaIndex 为检索增强生成 (RAG) 构造了带有边级和节点级文本信息但不互联的子图；因此我们没有将 LlamaIndex 与我们的方法进行评估。根据表 3 和表 4，我们得出结论: 在三种不同的知识图谱构建场景 (学术文章到 KG、简历到 KG、网站到 KG) 中，我们的方法在实体与关系解析过程中表现更优。此外，结果表明当输入文档数量较少且结构上含有清晰、非复杂短语时，大型语言模型在实体与关系解析方面表现良好，这一点在简历到 KG 的过程中得以体现。

Table 3. False Discovery Rates of unresolved entities for entity resolution process across three KG construction scenarios.

表 3. 三种知识图谱构建场景中实体解析过程未解析实体的错误发现率 (FDR)。

	OpenAI Function	Langchain	LlamaIndex	Our Method
Scientific Articles	0.11 ± 0.04	0.14 ± 0.08	-	0.01 ± 0.01
CVs	0	0	-	0
Websites	0.31 ± 0.05	0.29 ± 0.06	-	0

	OpenAI 功能	Langchain	LlamaIndex	我们的方法
科学文章	0.11 ± 0.04	0.14 ± 0.08	-	0.01 ± 0.01
简历	0	0	-	0
网站	0.31 ± 0.05	0.29 ± 0.06	-	0

Table 4. False Discovery Rates of unresolved relations for relation resolution process across three KG construction scenarios.

表 4. 在三种知识图谱构建场景下，关系解析过程中未解析关系的错误发现率。

	OpenAI Function	Langchain	LlamaIndex	Our Method
Scientific Articles	0.07 ± 0.01	0.06 ± 0.01	-	0.01 ± 0.01
CVs	0	0	-	0
Websites	0.15 ± 0.01	0.14±0.02	-	0

	OpenAI 功能	Langchain	LlamaIndex	我们的方法
科学论文	0.07 ± 0.01	0.06 ± 0.01	-	0.01 ± 0.01
简历	0	0	-	0
网站	0.15 ± 0.01	0.14±0.02	-	0

Moreover, the False Discovery Rates of unresolved entities and relations for websites to KG are higher than in the other KG construction scenarios. This is due to the larger number of documents (chunks) and the unstructured nature of website textual information. Consequently, without an effective resolution process, the LLM struggles to map similar entities or relations. Therefore, as long as the number of documents (chunks) is large and the text is unstructured with complex language, the entity/relation resolution process becomes crucial for building consistent KGs.

此外，网站到知识图谱的未解析实体和关系的错误发现率高于其他知识图谱构建场景。这是由于文档 (块) 数量更多且网站文本信息无结构化所致。因此，若没有有效的解析流程，LLM 很难将相似的实体或关系映射起来。因此，只要文档 (块) 数量较大且文本无结构且语言复杂，实体/关系解析过程对构建一致的知识图谱就至关重要。

Threshold Estimation To estimate the threshold for merging entities and relationships based on cosine similarity, a dataset of 1,500 similar entity pairs and 500 relationships, inspired by various domains (e.g., news, scientific articles, HR practices), was generated using GPT-4 and is available in the iText2KG GitHub repository. Entities and relationships were vectorized using the pre-trained model text-embedding-3-large The mean and standard deviation of cosine similarity for these datasets were then calculated (Table 5). An upper threshold (e.g., 0.7) was chosen to ensure high precision, while a lower threshold reduced resolution specificity.

阈值估计为了基于余弦相似度估计合并实体和关系的阈值，使用 GPT-4 生成了一个包含 1,500 对相似实体和 500 条关系的数据集，灵感来自多个领域 (例如新闻、学术文章、人力资源实践)，并可在 iText2KG GitHub 仓库中获取。使用预训练模型 text-embedding-3-large 对实体和关系进行了向量化。然后计算了这些数据集的余弦相似度的均值和标准差 (表 5)。选择了一个较高的上阈值 (例如 0.7) 以确保高精确率，而较低的阈值则降低了解析的特异性。

Table 5. Cosine Similarities of the Two Datasets for Entity and Relationship Resolution.

表 5. 实体和关系解析的两个数据集的余弦相似度。

Entities Dataset	Relationships Dataset
0.6±0.12	0.56±0.1

实体数据集	关系数据集
0.6±0.12	0.56±0.1

To illustrate the results of KG construction, Figure 5 presents a comparison between baseline methods and iText2KG across three distinct scenarios. The observations are as follows:

为说明 KG 构建结果，图 5 对基线方法与 iText2KG 在三种不同场景下进行了比较。观察结果如下:

- The baseline methods reveal the presence of isolated nodes without relations in all three KG construction scenarios. This phenomenon may be attributed to the simultaneous execution of entity extraction and relations extraction, which can induce hallucinatory effects in language models, leading to a ”forgetting” effect. This observation supports the findings of [1], which suggest that separating the processes of entity and relation extraction can enhance performance.

- 在三种 KG 构建场景中，基线方法均出现无关系的孤立节点。此现象可能源于实体抽取与关系抽取同时进行，导致语言模型产生幻觉效应，从而出现“遗忘”效应。此观察支持 [1] 的结论，即将实体抽取与关系抽取分离可提升性能。

- From the ’Website to KG’ scenario, an increase in the volume of input documents is associated with the emergence of noisy nodes within the graph. This underscores the critical need for Module 1 to effectively refine and distill the input data.

- 在“网站到 KG”场景中，输入文档量的增加伴随图中噪声节点的出现，强调了模块 1 需有效提炼与精炼输入数据的重要性。

- The iText2KG method demonstrates improved entity and relation resolution across the three KG construction scenarios. According to the data in Table 3 and Table 4 when input documents are fewer and composed of straightforward, non-complex phrases, the language model shows high efficiency in entity and relation resolution, as evidenced in the ’CVs to KG’ process. Conversely, the challenges increase with more complex and voluminous data sets, as shown in the ’Website to KG’ scenario.

- iText2KG 方法在三种 KG 构建场景中均表现出更好的实体与关系解析能力。根据表 3 与表 4 的数据，当输入文档较少且由简单、非复杂短语组成时，语言模型在实体与关系解析上效率很高，如“简历到 KG”过程所示；相反，随着数据更复杂且量更大，挑战随之增加，如“网站到 KG”场景所示。

Moreover, it is important to highlight the effect of the chunking size of the input document and the threshold on KG construction. Input documents to the Document Distiller can be independent documents or chunks. If the chunk size is smaller, the semantic blocks will capture more specific details from the documents, and vice versa.

此外，需要强调输入文档的分块大小与阈值对 KG 构建的影响。送入文档蒸馏器的输入可以是独立文档或分块。若分块较小，语义块将捕捉到文档中更具体的细节，反之亦然。

5 Conclusion

5 结论

In this paper, we introduced iText2KG, an approach for incremental KG construction leveraging the zero-shot capabilities of LLMs. Our methodology addressed limitations inherent in traditional KG construction processes, which typically depend on predefined ontologies and extensive supervised training.

本文提出了 iText2KG，一种利用大模型零样本能力进行增量式 KG 构建的方法。我们的方法解决了传统 KG 构建流程的局限性，传统方法通常依赖预定义本体与大量监督训练。

[9] https://platform.openai.com/docs/guides/embeddings/embedding-models

[9] https://platform.openai.com/docs/guides/embeddings/embedding-models

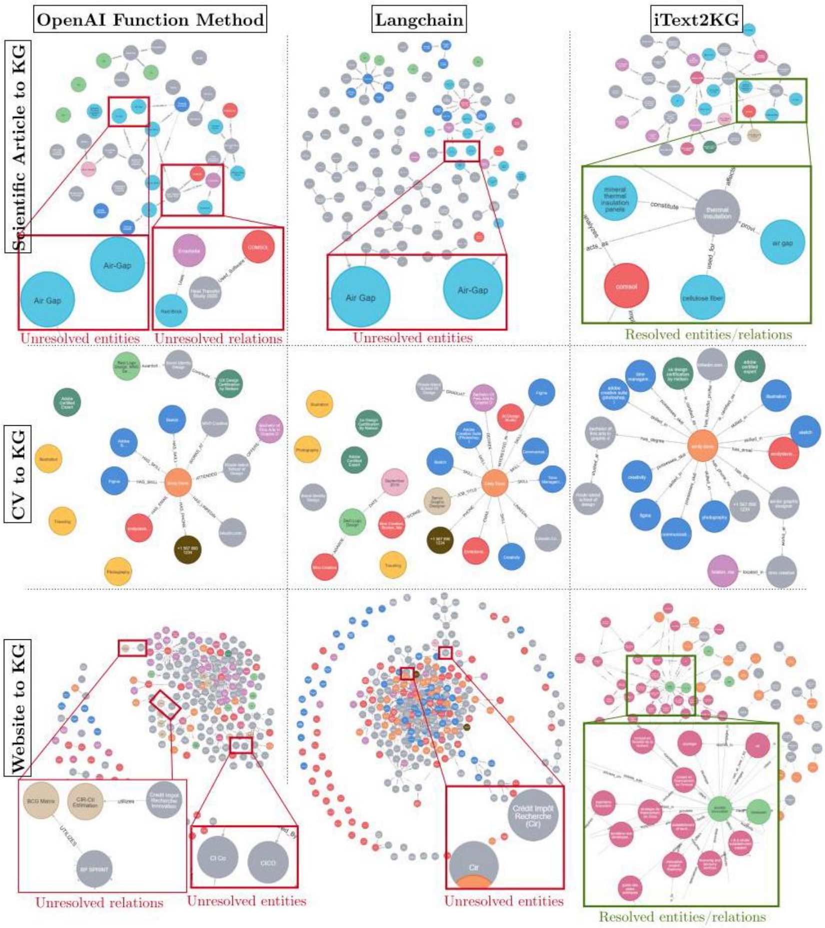

Fig. 5. Comparison of KG construction across three scenarios between baseline methods and our method, iText2KG.

图 5. 在三种场景中基线方法与我们的方法 iText2KG 的 KG 构建比较。

A key advantage of the iText2KG approach is its flexibility, which stems from the use of a user-defined blueprint that outlines the key components to extract during KG construction. This allows the method to adapt to a wide range of scenarios, as there is no universal blueprint for all use cases; instead, the design varies depending on the specific application. Moreover, The iText2KG method achieves document-type independence by using a flexible, user-defined blueprint to guide the extraction process, allowing it to handle both structured and unstructured texts.

iText2KG 方法的一大优势是灵活性，来源于用户定义的蓝图，该蓝图列出在 KG 构建过程中需抽取的关键组件。此设计使方法能适应多种场景，因为不存在通用蓝图；相反，设计随具体应用而变。此外，iText2KG 通过使用灵活的用户定义蓝图来引导抽取过程，从而实现对文档类型的独立性，能够处理结构化与非结构化文本。

Empirical evaluations across diverse contexts, such as scientific documents, web content, and CVs, demonstrated the superior performance of the iText2KG approach compared to established baseline methods. The method achieves enhanced schema consistency and high precision in entity and relation extraction, effectively mitigating issues related to semantic duplication and unresolved entities, which are prevalent in traditional methodologies.

在科学文献、网页内容与简历等不同语境下的实证评估表明，iText2KG 相较于现有基线方法表现更优。该方法在模式一致性上有所提升，并在实体与关系抽取上具有高精度，有效缓解了传统方法中常见的语义重复与未解析实体问题。

Future research will focus on enhancing metrics such as cosine similarity for advanced entity and relation matching, eliminating the necessity to define a threshold as a hyperparameter, and integrating the entity type as a parameter of the matching process.

未来研究将着重提升余弦相似度等用于高级实体与关系匹配的度量，消除将阈值作为超参数的必要性，并将实体类型整合为匹配过程的一个参数。

References

参考文献

1. Carta, S., Giuliani, A., Piano, L., Podda, A.S., Pompianu, L., Tiddia, S.G.: Iterative zero-shot LLM prompting for knowledge graph construction. arXiv preprint arXiv:2307.01128 (2023)

1. Carta, S., Giuliani, A., Piano, L., Podda, A.S., Pompianu, L., Tiddia, S.G.: Iterative zero-shot LLM prompting for knowledge graph construction. arXiv preprint arXiv:2307.01128 (2023)

2. Ding, L., Zhou, S., Xiao, J., Han, J.: Automated construction of theme-specific knowledge graphs. arXiv preprint arXiv:2404.19146 (2024)

2. Ding, L., Zhou, S., Xiao, J., Han, J.: Automated construction of theme-specific knowledge graphs. arXiv preprint arXiv:2404.19146 (2024)

3. Eberendu, A.C., et al.: Unstructured data: an overview of the data of big data. International Journal of Computer Trends and Technology 38(1), 46-50 (2016)

3. Eberendu, A.C., et al.: Unstructured data: an overview of the data of big data. International Journal of Computer Trends and Technology 38(1), 46-50 (2016)

4. Hu, Y., Zou, F., Han, J., Sun, X., Wang, Y.: LLM-Tikg: Threat intelligence knowledge graph construction utilizing large language model. Available at SSRN 4671345 (2023)

4. 胡毅, 邹飞, 韩杰, 孙翔, 王勇: LLM-Tikg: 利用大型语言模型构建威胁情报知识图谱。发表于 SSRN 4671345 (2023)

5. Jin, B., Liu, G., Han, C., Jiang, M., Ji, H., Han, J.: Large language models on graphs: A comprehensive survey. arXiv preprint arXiv:2312.02783 (2023)

5. 金斌, 刘光, 韩成, 蒋明, 纪豪, 韩杰: 图上的大型语言模型: 一项综合综述。arXiv 预印本 arXiv:2312.02783 (2023)

6. Kabal, O., Harazallah, M., Guillet, F., Ichise, R.: Enhancing domain-independent knowledge graph construction through OpenIE cleaning and llms validation (G-T2KG). In: 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024) (2024), to appear

6. Kabal, O., Harazallah, M., Guillet, F., Ichise, R.: 通过 OpenIE 清洗和 LLMs 验证 (G-T2KG) 增强领域无关的知识图谱构建。在: 第 28 届基于知识与智能信息与工程系统国际会议 (KES 2024) (2024), 待刊

7. Kommineni, V.K., König-Ries, B., Samuel, S.: From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. arXiv preprint arXiv:2403.08345 (2024)

7. Kommineni, V.K., König-Ries, B., Samuel, S.: 从人类专家到机器: 一种由 LLM 支持的本体与知识图谱构建方法。arXiv 预印本 arXiv:2403.08345 (2024)

8. Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K.: Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In: International Semantic Web Conference. pp. 247-265. Springer (2023)

8. Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K.: Text2kgbench: 用于从文本生成本体驱动知识图谱的基准。在: 国际语义网会议。第 247-265 页。Springer (2023)

9. Nasar, Z., Jaffry, S.W., Malik, M.K.: Named entity recognition and relation extraction: State-of-the-art. ACM Computing Surveys (CSUR) 54(1), 1-39 (2021)

9. Nasar, Z., Jaffry, S.W., Malik, M.K.: 命名实体识别与关系抽取: 技术现状。ACM 计算综述 (CSUR) 54(1), 1-39 (2021)

10. Singh, S.: Natural language processing for information extraction. arXiv preprint arXiv:1807.02383 (2018)

10. Singh, S.: 用于信息抽取的自然语言处理。arXiv 预印本 arXiv:1807.02383 (2018)

11. Sun, Z., Ting, Y.S., Liang, Y., Duan, N., Huang, S., Cai, Z.: Knowledge graph in astronomical research with large language models: Quantifying driving forces in interdisciplinary scientific discovery. arXiv preprint arXiv:2406.01391 (2024)

11. 孙振, 丁雅诗, 梁研, 段宁, 黄思, 蔡哲: 在天文学研究中结合大型语言模型的知识图谱: 量化跨学科科学发现的驱动力。arXiv 预印本 arXiv:2406.01391 (2024)

12. Wornow, M., Lozano, A., Dash, D., Jindal, J., Mahaffey, K.W., Shah, N.H.: Zero-shot clinical trial patient matching with LLMs. arXiv preprint arXiv:2402.05125 (2024)

12. Wornow, M., Lozano, A., Dash, D., Jindal, J., Mahaffey, K.W., Shah, N.H.: 使用 LLMs 的零样本临床试验患者匹配。arXiv 预印本 arXiv:2402.05125 (2024)

13. Zhang, Y., Du, T., Ma, Y., Wang, X., Xie, Y., Yang, G., Lu, Y., Chang, E.C.: AttacKG+: Boosting attack knowledge graph construction with large language models. arXiv preprint arXiv:2405.04753 (2024)

13. 张洋, 杜涛, 马悦, 王轩, 谢勇, 杨光, 陆云, 张恩慈: AttacKG+: 利用大型语言模型提升攻击知识图谱构建。arXiv 预印本 arXiv:2405.04753 (2024)

14. Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., Zhang, N.: LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. arXiv preprint arXiv:2305.13168 (2023)

14. 朱远, 王晓, 陈俊, 乔帅, 欧阳, 姚跃, 邓莎, 陈浩, 张楠: 用于知识图谱构建与推理的 LLMs: 近期能力与未来机会。arXiv 预印本 arXiv:2305.13168 (2023)