基于大语言模型的绿色低碳领域三元组抽取方法*

王丽君　赵子岩　马丽　蒋慧超　张冉
（国家电网有限公司信息通信分公司，北京 100761）

摘要：三元组抽取旨在提取文本中的实体及其相互关系，从而形成结构化的知识表示，是构建自动化知识图谱的关键技术。尽管基于传统深度学习的三元组抽取方法在拥有充足训练数据时表现出色，但在电力行业绿色低碳领域等垂直场景中，由于缺乏规范化的监督数据，人工标注成本高昂，且论文和专利数据中存在大量专业术语，深度学习抽取方法的识别准确度受限。为了解决这些问题，设计了基于大语言模型的三元组抽取方法，利用闭源大模型标注少量高质量监督数据，结合检索增强技术指导开源模型进行抽取，实现了高质量且自动化的垂直领域抽取。此外，为了提升少样本场景下的抽取效率与精确率，本方法还包含了数据分流与复杂数据划分模块，以抽取难易程度为标准将数据分流，并进一步划分复杂数据来简化抽取，从而提升抽取效果。为了验证模型性能，利用GPT-4自动化标注了一个基于电力领域专利和论文的数据集，并引入了ChatGPT和ChatGLM等知名闭源以及开源大模型作对比，实验结果证明提出的方法具有更好的抽取性能。

关键词：三元组抽取；知识图谱；大语言模型；绿色低碳
中图分类号：C37；N37；P413　DOI：10.3772/j.issn.1673-2286.2025.01.006
引文格式：王丽君，赵子岩，马丽，等. 基于大语言模型的绿色低碳领域三元组抽取方法[J]. 数字图书馆论坛，2025，21（1）：46-54，66.

知识图谱作为一种结构化的语义知识库，在日常生活中有着十分广泛的应用。电商平台的推荐系统、搜索引擎的检索优化以及社交媒体的用户分析都需要依赖知识图谱中丰富的结构与特征信息。在工业界，基于知识图谱的技术应用也有重要意义，例如针对一些新兴领域的技术成熟度分析和潜在创新性技术预见等，可以帮助行业专家从宏观层面对整个领域的发展进行把控。知识图谱技术大多需要一定规模的图数据支撑，例如，知识图谱在电商平台和社交媒体方向的成功应用都建立在海量商品和用户数据之上。因而，构建高质量的图谱数据是知识图谱相关技术在工业新兴领域进一步落地应用的首要任务。

电力行业的绿色低碳领域先进技术在能源转型中具有关键作用。随着可持续发展目标的提出和碳中和政策的实施，绿色低碳技术已成为全球各国能源战略的核心。储氢技术、电动汽车和可控负荷很好地代表电力行业绿色低碳发展的关键技术和方向，体现了能源生产、消费与智能管理的协同创新。通过聚焦这3个领域，能够为绿色低碳电力行业的未来发展提供系统性、全方位的视角，同时为自动化构建高质量知识图谱提供精准的技术和理论支持。本文针对电力行业的绿色低碳领域，聚焦于储氢技术、电动汽车和可控负荷3个关键词相关的公开专利以及论文数据，为当前领域高质量知识图谱的自动化构建提供解决方案。

收稿日期：2024-10-16
*本研究得到国家电网公司总部科技项目“‘双碳’目标下电力绿色低碳关键支撑技术评价方法和专利标准化研究”（编号：1400-202340338A-1-1-ZN）资助。

三元组作为图谱构建的基本单元，能够通过实体及其相互关系形成结构化的知识表示。从非结构化的文本数据中抽取出三元组是构建知识图谱的关键环节。早先三元组抽取方法大多基于特征工程，通过词性以及句法等较为浅层的信息实现对文本的语义理解，在实际应用中抽取效果并不理想[1-2]。近年来，深度学习的发展改变了这一抽取模式[3-4]，以预训练语言模型为代表的深度学习技术将机器自动化抽取三元组等结构化信息的能力提升到了前所未有的高度[5]。尽管迁移学习一定程度上缓解了深度学习方法的数据依赖问题，但当这些方法应用到垂直领域时，仍会因为标注数据稀缺，难以提升微调效果。最近，大语言模型的出现为上述问题提供了可行解。大语言模型是预训练语言模型的模型参数以及训练规模提升至一定高度的产物，具有极强的泛化性能。大量研究表明，大语言模型在包括三元组抽取在内的诸多信息抽取任务上表现出色，即使在零样本场景下，依然能追平监督学习的传统深度学习模型的抽取性能[6]，因而本文选择引入大语言模型进行三元组抽取。

1　相关研究

1.1　传统信息抽取方法

信息抽取是自然语言处理中的一个核心任务，其主要目标是从非结构化文本中提取出有意义的结构化信息，如实体、关系和事件[7]。在早期，基于规则和模板的方法主导了信息抽取研究。这些方法依赖于人工设计的规则和特征，从特定领域的文本中抽取关系，例如基于依存句法分析、关键词匹配和正则表达式等技术实现简单的关系提取。这类方法尽管在特定领域有一定的应用，但表现通常取决于文本格式和语言特征，难以应对复杂多样的语言表达形式[8]。

1.2　基于深度学习的三元组抽取方法

近年来，随着预训练语言模型的广泛应用，深度学习模型在信息抽取领域取得了显著进展[9]，基于神经网络的模型逐渐取代了基于规则的方法，成为信息抽取的主流技术。预训练语言模型通过在大规模无监督文本上预训练，具备了强大的语言理解能力，能够为下游任务提供更加通用的特征表示。通过利用大规模标注数据和神经网络模型，研究者们可以构建更加鲁棒的三元组抽取系统。这类模型在三元组抽取任务中被广泛使用，研究者们基于这些模型提出了诸多高效的三元组抽取方法。例如，CasRel[10]是基于BERT的三元组抽取框架，采用了级联解码的方式，将关系提取问题转化为序列标注任务，从而实现了实体和关系的高效抽取。TPLinker[11]则通过引入链式连接机制，在关系的多对多映射场景中展现了优异的性能。OneRel[12]进一步提出了统一的实体关系联合三元组抽取框架，通过多任务学习同时完成实体和关系的识别。

这些方法通过引入先进的深度学习技术，有效提升了在三元组抽取任务中的性能，尤其在公共领域数据集上取得了显著效果。然而，尽管深度学习方法取得了一定的成功，但它们往往依赖于大规模的标注数据，这使得这些模型大多集中于通用领域数据集，如新闻、维基百科等，因而难以在数据稀缺的领域中泛化。此外，这些方法在处理长文本时的效果较差，往往无法有效提取出完整的关系[13]。而且由于专利和论文文本的结构复杂、术语多样，现有的深度学习模型往往无法有效处理这些领域的数据。

1.3　基于大语言模型的信息抽取方法

近年来，大语言模型如GPT-3、GPT-4等出现，为信息抽取任务带来了新的可能[14-16]。大语言模型通过在海量无监督数据上进行预训练，具备了强大的生成能力和语言理解能力，能够在少样本甚至零样本学习场景下完成复杂的文本生成和信息抽取任务。相比传统的深度学习模型，大语言模型在无需大规模标注数据的情况下即可生成高质量的抽取结果，尤其是在通用领域的开放式关系抽取任务中表现出色[17]。例如，GPT-RE[18]提出了基于GPT的关系抽取方法，将关系抽取任务转化为问答任务，通过语言模型生成实体和关系的答案。QA4RE[19]则通过重新设计问题格式，引导语言模型以问答的方式进行关系抽取，在零样本学习场景中取得了优异的表现。此外，上下文学习[20]、思维链[21]以及检索增强[22]技术进一步提升了大语言模型在复杂关系抽取任务中的表现，通过引导模型生成中间推理步骤，有效降低了模型的错误率。

尽管如此，大语言模型在专利和学术论文等特定

领域的效果尚未得到充分验证。由于这些领域的文本往往包含大量专业术语和复杂的句法结构，现有的大语言模型通常难以准确理解和抽取出其中的实体和关系。

1.4 基于大语言模型的三元组抽取方法

基于大语言模型的三元组抽取是近年来信息抽取领域的一个重要研究方向。传统的逐步抽取方法通常先进行实体识别，再进行关系提取，这样的两阶段方法容易导致误差传播，尤其是在处理复杂文本时，前一阶段的错误会对后续关系提取任务造成不良影响[23]。为了解决这一问题，研究者们提出了利用大模型联合抽取的方法，将实体识别和关系提取作为一个整体任务同时完成，从而减少信息丢失和错误传播[19,24]。例如，AutoRE[25]和REPLM[26]等模型通过引入联合学习机制，能够在同一框架下同时完成实体和关系的识别任务。Meta In-Context Learning[16]基于大语言模型的上下文学习能力，进一步提升了大语言模型在少样本和零样本抽取任务中的表现。这些三元组抽取方法通过减少任务之间的依赖性和误差传播，在通用领域的数据集上展现了良好的性能[27-28]。然而，现有的大模型三元组抽取技术大多集中于新闻、医疗等领域的数据集[29-30]。由于专利和学术论文文本结构复杂、术语专业，现有模型难以直接适应这些领域的数据。

考虑到大语言模型在少样本条件下的出色表现[16,31]，选择引入大语言模型来完成三元组抽取任务，以更好地应对电力领域中信息抽取训练数据缺失的问题。本文将基于大语言模型的三元组抽取技术应用于绿色低碳领域的专利和学术论文数据，结合检索增强技术实现领域自适应机制，提出了针对绿色低碳领域专利与论文数据的高效抽取方法，有效提升了模型在该领域的适应性和抽取性能。

2 研究思路

大语言模型大多基于通识语料进行预训练，在常识性知识方面表现尚可，但当应用于垂直领域时则会产生幻觉偏差。在相对较少数据上预训练的小规模开源模型中，这类偏差尤其严重，因此需要对语言模型进行领域适配，从而更好地发挥其性能。领域微调是效果最好的领域适配方法，通过在特定数据上进一步训练大语言模型，使得模型可以更好地利用领域知识，且能按照固定格式输出。领域微调需要的数据量较大，且数据的质量很大程度上决定了微调后模型的性能。相比领域微调，基于检索增强的技术并不改变模型参数，而是利用大模型上下文学习，同样可以帮助大模型理解领域知识，并规范其输出格式。该技术对标注数据数量要求低[6]，因此更加契合本文的需求。

原始的三元组抽取任务是指抽取头尾实体及其关系，但通过分析目标数据发现，专利以及论文数据中存在的三元组往往以专利或论文本身为头实体，而关系信息在本体构建时就已设定好目标范围，因此抽取重点实际上只有尾实体（值域）。据此，本文提出了一套基于大语言模型的三元组抽取方法，采用综合性能优越的Llama-3-8B作为基座模型，并引入检索增强技术，通过检索与目标文本相近的示例来引导大模型准确地开展三元组抽取，相似度检索所依赖的数据向量化步骤由Bge-m3模型完成。此外，该方法还包含数据分流处理和复杂数据划分两个重要模块。数据分流的核心观念是因地制宜，根据不同数据抽取难度设计相应的抽取策略。具体而言，对于抽取难度较低的数据采用基于正则表达式的规则抽取，对于需要进行语义理解的复杂语句采用基于检索增强的大语言模型抽取，这样的组合抽取模式在保证了模型抽取性能的前提下极大提升了抽取效率。复杂数据划分则主要针对数据分流出的复杂输入数据，用合适的提示指导大模型对专利介绍或论文原文这种存在大量冗余信息的长文本进行划分、筛选，并根据后续抽取需求输出包含目标信息的语段，用于进一步结构化抽取。值得一提的是，为了进一步提升方法的自动化程度，引入了在各项任务上表现出色的闭源大模型GPT-4用于数据标注，标注过程与抽取过程类似。在自动化构建的数据集上进行实验，结果证明本文提出的方法能够很好地应对垂直领域的抽取任务，抽取高效且准确，为领域高质量知识图谱自动化构建提供了可行解。

3 研究方法

根据信息抽取难易程度，将抽取任务分为两个部分，分别设计抽取方案：首先，基于从论文或专利网站

中导出的基本信息抽取出大部分本体涉及的关系三元组，例如作者、机构以及引用等，这些基本信息大多已经过人工或自动化手段的初步结构化，因此抽取难度较低，利用正则表达式即可实现高效抽取；其次，论文原文和专利具体介绍中也存在大量尚未挖掘的关键信息，但由于文本长度长、理解难度高，抽取这部分信息需要模型具有语义理解与总结归纳的能力，因此引入基于大语言模型的方法进行三元组抽取，方法的整体框架如图1所示。针对本体定义进行分析，发现绝大部分定义的关系中专利或论文本身为主语，其余则为宾语，因而在对每篇论文或专利文献进行三元组抽取时，只有一项需要确定，即主语或宾语，据此将这个三元组抽取任务进一步简化为实体抽取任务，实体抽取方法有以下几种。

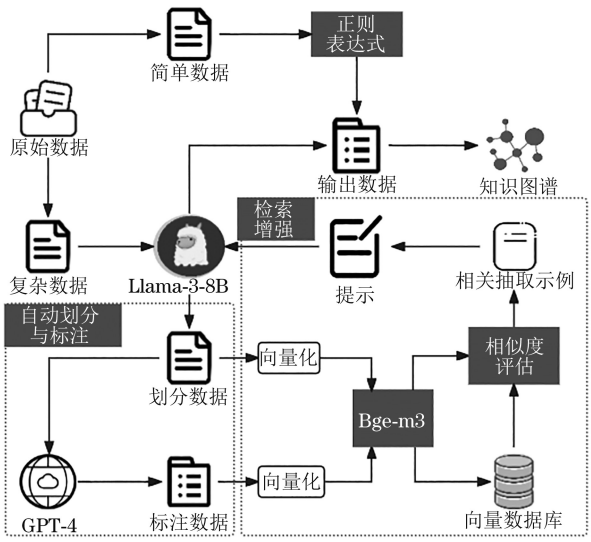

图1　基于大语言模型的三元组抽取方法框架

3.1　基于正则表达式的抽取

本部分抽取方法针对的是从智慧芽导出的专利基本信息，以及从中国知网导出的论文基本信息。其中，从智慧芽中导出的专利基本信息结构化程度较高，原始数据为csv形式的表格，只须根据表头条目选择本体涉及的内容进行简单抽取。从中国知网导出论文基本信息时，须先批量导出refworks形式的数据，再进一步解析为结构化的csv表格数据，才能对表格数据进行类似专利信息抽取的操作。流程中对refworks的结构化处理以及对csv表格数据的进一步抽取都基于正则表达式实现。正则表达式说明以及其抽取示例如表1和表2所示。

表1　正则表达式说明

正则表达式	抽取目标
[\u4e00-\u9fa5]{2，3}（、[\u4e00-\u9fa5]{2，3}）*	人名（2~3个字，顿号隔开）
[\u4e00-\u9fa5]+（公司|学校|院）；	机构名（以公司或学校等结尾，分号隔开）
[\u4e00-\u9fa5]+（学|工程）|	领域名（以学或工程等结尾，“|”符号隔开）
CN\d{5，13}-A（\d）?|	引用专利号（CN开头的为国内专利，一般以A或A+数字结尾）
（19|20）\d{2}-（0[1-9]|1[0-2]）-（0[1-9]|[12][0-9]|3[01]）	日期（YYYY-MM-DD格式）

表2　正则表达式抽取示例

原始文本	抽取结果
华南理工大学|广东省特种设备检测研究院（广东省特种设备事故调查中心）	[“华南理工大学”，“广东省特种设备检测研究院”]
机械工程|工程学|储氢|物理学	[“机械工程”，“工程学”，“物理学”]
CN101602485A|CN111013593A|CN113200515A|CN1380136A|CN1522224A|US20090246575A1	[“CN101602485A”，“CN111013593A”，“CN113200515A”，“CN1380136A”，“CN1522224A”]
申请日期为2015-12-31	[“2015-12-31”]

3.2　基于数据自动划分的细粒度信息抽取

针对复杂专利信息与论文原文这些长文本，引入具有较强文本理解能力以及总结归纳能力的大语言模型。同时考虑到没有充足的标注数据进行监督模型的训练，设计了一套适用于少样本场景的实体抽取方案，补充抽取更深层的目标信息。该方案主要是基于检索增强生成的框架，所采用的大模型为在中文数据上微调的Llama-3-8B模型，其核心思想是利用大模型的语义理解能力对复杂文档进行划分，并根据目标信息进行结构化抽取。

（1）基于大语言模型的数据自动化划分与标注策略。首先，利用大模型对数据进行初步划分。基于本体中的目标关系筛选目标文本，过滤掉无用信息使得后续模型的输入相对简洁精炼，从而减少推理时的计算资源并提升抽取的效果，计算公式如式（1）所示。

$$\left[T_1^i, T_2^i, \cdots, T_n^i\right] = \mathrm{LLM}\left(P_1^i\right) \tag{1}$$

式中：P_1^i 代表包含原始文本T^i的用于划分数据的提示；下标n代表需要抽取的实体类型的数量；LLM在这里指Llama-3-8B模型。

然后，对数据进行自动化标注。为提升效率，采用效果好但闭源的GPT-4模型对每种目标类型的实

体进行一定数量的标注，构建一个高质量的少样本数据集，用于指导后续抽取任务。基于划分出的数据块 $\left[T_1^i,T_2^i,\cdots,T_n^i\right]$，自动标注的计算公式如式（2）所示。

$$\left[\left(T_1^i,a_1^i\right),\left(T_2^i,a_2^i\right),\cdots,\left(T_n^i,a_n^i\right)\right]=\mathrm{LLM}\left(P_2^i\right) \quad (2)$$

式中：P_2^i 为包含 $\left[T_1^i,T_2^i,\cdots,T_n^i\right]$ 的用于自动化标注的提示；a_j^i 为大模型对于 T_j^i 的标注。

（2）基于检索增强的细粒度信息抽取策略。首先，将上述标注好的数据集进行向量化存储，选择北京智源人工智能研究院发布的Bge-m3作为向量化模型，将每一条标注数据作为一个文本块进行向量化，便于后续细粒度筛选相似数据指导抽取。向量化过程如式（3）所示。

$$\left[\boldsymbol{H}_1^i,\boldsymbol{H}_2^i,\cdots,\boldsymbol{H}_n^i\right]=\mathrm{Encoder}\left(\left[T_1^i,T_2^i,\cdots,T_n^i\right]\right) \quad (3)$$

式中：$\boldsymbol{H}_k^i$ 为对应数据块 T_k^i 的向量表示；Encoder为编码器。

然后，利用基于k近邻方法的检索器从数据库中检索与目标文本相似的数据，检索的实现主要基于余弦相似度，计算查询文本与所有数据库中文本在向量空间的相似度值，计算公式如式（4）所示。

$$f(i)=\mathrm{Similarity}\left(\boldsymbol{H}_j^i,\boldsymbol{H}_j^{\mathrm{input}}\right) \quad (4)$$

式中：$\boldsymbol{H}_j^{\mathrm{input}}$ 为输入文本基于第j种实体类型划分后的向量表示。相似度计算公式可简化为余弦函数。对于k近邻方法的应用，k为一个预设的常数，需要根据相似度顺序找到相似度排名前k的文档作为候选，计算公式如式（5）~式（6）所示。

$$X=\left\{x_1,x_2,\cdots,x_k\middle|k\leqslant S,f\left(x_1\right)\geqslant f\left(x_2\right)\geqslant\cdots\geqslant f\left(x_k\right)\geqslant f\left(x\_\right)\right\} \quad (5)$$

$$E_j=\left\{\left(T_x^i,a_x^i\right)\middle|\forall x\in X\right\} \quad (6)$$

式中：S为数据库中所有标注数据的数量；$x_$为数据库中不在X内的其他索引。

返回检索到的相似度排名前k的数据块及其标注数据，将它们作为抽取示例包装进提示中，用于指导大模型进行少样本抽取，抽取任务提示模板示例如图2所示。提示中还包括对任务的描述，这里将抽取任务重构为一个文本生成任务，使其以json格式返回目标实体。

最后，将提示输入大模型，获取回答，并对大模型的回答进行后处理，通过正则表达式等手段删掉输出中的多余文本，规范格式后整合为json文件，用于后续评估。

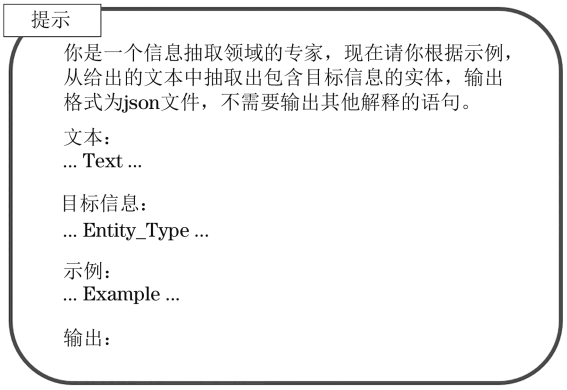

图2　抽取任务提示模板示例

4　实验与分析

4.1　数据集构建

实验数据包含论文与专利两类，论文数据来自中国知网，专利数据来自智慧芽。关键词设定为储氢技术、电动汽车以及可控负荷，文献符合其一即可，同时要求具体内容围绕某种具体技术或方法，最终分别选择符合要求的论文和专利中发表年份较近的20篇作为原始数据。

原始数据包括两类：一类是包含各类基本信息的数据，以高度结构化的表格形式存在；另一类是描述论文和专利具体内容的文本信息，即论文原文和专利的各项具体介绍，论文原文主要以pdf格式存在，专利具体内容则为文本格式。对pdf格式的数据采用OCR方法识别，获取文本格式的数据，并对该数据进行简单的清洗，提升数据质量。采用PP-OCRv4模型作为OCR的主要模型，它集成了SRN、NRTR模型进行版面对象的检测与分割，根据论文排版的特性将文本切分为块进行识别，同时在后处理中加入信息过滤，滤除那些无关紧要的冗余信息。

最终采用GPT-4进行自动化标注，同时在后期进行人工校对以保证数据质量。在自动化标注流程中，采用和抽取方法类似的方案：先利用大模型对文本进行划分，再对划分后的文本进一步标注。

为了避免标注出的数据存在长尾问题，基于关键词匹配的方式筛选可能包含各类信息的文本，使得各类文本的占比尽可能平均，最终标注获得的数据集信息如表3所示。

表3　数据集信息

关系类别	数量/个	值域平均长度/字
化学成分	149	4.20
装置组成	153	5.01
技术特点	119	12.06
核心技术	109	7.48
解决问题	180	9.37

4.2　抽取结果示例

实验基于两个RTX 4090实现，最终抽取出的专利和论文结果如图3所示。

一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 化学成分, 氢气
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 储氢瓶
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 加氢装置
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 气瓶仓
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 快装公接头
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 快装母接头
一种储氢瓶与加氢装置连接　脱离方法以及充氢方法, 装置组成, 制氢模块
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 电解槽
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 开合滑套
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 装置组成, 锁紧和脱离装置
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 解决问题, 如何设计一种小型化制氢加氢装置
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 解决问题, 满足用户高频率、低容量的加氢需求
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 解决问题, 保证充氢方法和充氢过程具有简单、
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 解决问题, 保证整个充氢过程的顺利进行
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 核心技术, 制氢加氢
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 核心技术, 储氢瓶
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 核心技术, 加氢装置
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 核心技术, 充氢
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 技术特点, 快速、方便连接和分离
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 技术特点, 实时监控制氢过程
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 技术特点, 及时发现问题并进行处理
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 技术特点, 有效控制气瓶仓的温度
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 技术特点, 提高安全性
一种储氢瓶与加氢装置连接、脱离方法以及充氢方法, 技术特点, 改善自动化程度

(a)专利抽取结果示例

一步法制备球形活性炭及其储氢性能研究, 化学成分, 聚苯乙烯树脂
一步法制备球形活性炭及其储氢性能研究, 化学成分, KOH
一步法制备球形活性炭及其储氢性能研究, 化学成分, 液氮
一步法制备球形活性炭及其储氢性能研究, 装置组成, X射线粉末衍射仪
一步法制备球形活性炭及其储氢性能研究, 装置组成, 场发射扫描电子显微镜
一步法制备球形活性炭及其储氢性能研究, 装置组成, 物理吸附仪
一步法制备球形活性炭及其储氢性能研究, 装置组成, 低温储氢装置
一步法制备球形活性炭及其储氢性能研究, 核心技术, 一步炭化-活化法
一步法制备球形活性炭及其储氢性能研究, 解决问题, 缺乏高效的氢气储运方式
一步法制备球形活性炭及其储氢性能研究, 技术特点, 大比表面积
一步法制备球形活性炭及其储氢性能研究, 技术特点, 多孔结构
一步法制备球形活性炭及其储氢性能研究, 技术特点, 高质量储氢密度
一步法制备球形活性炭及其储氢性能研究, 技术特点, 低成本

(b)论文抽取结果示例

图3　抽取结果示例

4.3　评估指标

为了测试方案基于大模型部分的抽取性能，采用精确率、召回率以及F1值3个指标对模型的抽取结果进行评估。精确率计算的是所有抽取结果中与标注一致的结果占比，召回率计算的是抽取正确的结果占所有标注结果的比例，F1值则是综合精确率与召回率得出的评分。

4.4　对比实验

为了验证方案的抽取性能，引入ChatGPT和文心一言两个闭源模型以及ChatGLM-3-6B和基座模型Llama-3-8B两个开源模型对数据直接抽取，进行对比实验。随机选择1/4的部分作为测试集，其余作为模型可以检索并参考的数据集，重复实验多次后对各个指标取平均值，对比实验结果如表4所示。

根据评估的结果，可以明显看出所提方法的优势，总体F1值相较闭源模型ChatGPT以及文心一言分别取得了约11.1个百分点和6.9个百分点的提升。它抽取出的实体具有较高的准确性，并且包含这些实体的三元组构成的图谱具有较强的覆盖性、完整性。部分关系对应的三元组抽取精确率相对较低，但经过后期人工二次评估，发现实际精确率高于表4所示的基于GPT-4标注的评估结果，这说明部分情况下所提方案的抽取表现甚至优于GPT-4，实验结果证明所提方法是有效的。

利用ChatGPT进行抽取时“化学成分”与“装置组成”抽取效果较差，分析发现该模型容易混淆两类实体，因而输出里包含大量错误答案，导致精确率远远低于召回率，影响了抽取性能；文心一言对这两类信息的区分能力较好，这很可能归功于其在中文数据上的大量训练，对中文的理解更加准确。两个开源大模型由于参数量和训练量有限，表现一般，本文选择它们中的Llama作为基座模型并在性能上取得了较大的提升，F1值提升约16.2个百分点。

4.5　消融实验

为了探究本文提出的方案各部分对最终抽取性能以及抽取效率的影响，对自动化数据划分模块与相关文档检索模块进行消融实验。消融实验结果如表5所示，推理效果如表6所示。

实验结果表明，原始数据划分模块能从原始数据

表4　对比实验结果

模型	关系类别	精确率	召回率	F1值
ChatGPT	化学成分	0.493	0.696	0.577
	装置组成	0.504	0.703	0.587
	技术特点	0.738	0.737	0.738
	核心技术	0.652	0.729	0.688
	解决问题	0.765	0.701	0.732
	总体	0.630	0.711	0.668
文心一言（ERNIE-3.5）	化学成分	0.650	0.687	0.668
	装置组成	0.699	0.732	0.715
	技术特点	0.731	0.715	0.723
	核心技术	0.649	0.722	0.684
	解决问题	0.758	0.736	0.747
	总体	0.701	0.719	0.710
ChatGLM-3-6B	化学成分	0.538	0.586	0.561
	装置组成	0.579	0.582	0.580
	技术特点	0.623	0.574	0.597
	核心技术	0.517	0.508	0.512
	解决问题	0.610	0.631	0.621
	总体	0.577	0.583	0.580
基座模型（Llama-3-8B）	化学成分	0.606	0.589	0.597
	装置组成	0.648	0.639	0.643
	技术特点	0.635	0.611	0.623
	核心技术	0.561	0.578	0.569
	解决问题	0.653	0.618	0.635
	总体	0.625	0.609	0.617
所提模型	化学成分	0.733	0.789	0.760
	装置组成	0.803	0.827	0.815
	技术特点	0.819	0.766	0.792
	核心技术	0.680	0.791	0.731
	解决问题	0.807	0.754	0.780
	总体	0.773	0.785	0.779

中过滤无关文本，并根据关系类型的不同对长文本初步分类，细化后续模型抽取的目标文本，能够通过减少输入长度降低硬件需求，在抽取效率升至原先3倍的同时保证抽取信息的高质量；相关文档检索模块可以帮助模型规范输出格式，并指导模型更准确地抽取部分语义复杂的样本，对模型抽取性能有较大提升作用，单此模块就可将F1值提升约10.3个百分点。

表5　消融实验结果

模型	精确率	召回率	F1值
原模型	0.773	0.785	0.779
删去数据划分模块模型	0.742	0.699	0.720
删去数据划分和相关检索模块模型	0.625	0.609	0.617

表6　推理效果对比

模型	推理速度/秒	GPU数/个
原模型	213.3	2
删去数据划分模块模型	684.8	4

5　结语

本文提出了一种基于大语言模型的三元组抽取方法，结合检索增强技术实现了自动化的数据标注与抽取，且通过简化任务、分流数据以及切分复杂数据等辅助手段，实现了效率高、抽取准的效果。本文主要有如下3个方面的贡献。①基于电力领域公开的专利和论

文信息，用GPT-4自动化标注了一个抽取数据集，涵盖5种不同的目标信息类型，并且分布均匀。②提出了基于Llama大语言模型的专利和论文三元组抽取方法，实现了复杂输入的自动切分与领域数据的自动标注，同时结合检索增强技术实现领域自适应机制，可以解决垂直领域中标注数据稀少而原始文本复杂的问题。③在标注的数据集上对本文提出的方法进行了对比实验和消融实验，实验结果证明这一方法在少样本场景下具有很好的抽取性能。

本文立足于电力行业绿色低碳领域的实际场景，使用开源的基座模型，便于本地部署，消除了数据安全等隐患；引入自动化标注模块，减少人工成本，提高了自动化程度；利用大语言模型辅助数据处理，降低了硬件依赖程度，提升了计算效率与抽取精确率。从数据安全、成本、自动化程度以及抽取性能等角度综合分析，发现本文提出的基于大语言模型的三元组抽取方法适应了当前领域的各种需求，推动了领域图谱的自动化构建，对其他垂直领域的大模型应用也有一定的启示意义。

值得一提的是，根据本文方法得到的抽取输出可以作为新一轮任务的标注输入，这样一种可以循环迭代的模式非常有利于后续使用时性能的提升以及方法的更新。增加的数据量意味着更多的标注数据可以用于指导大模型的上下文学习，因而抽取效果的提升是必然的；当数据量积攒到一定程度时，还可以引入指令微调来帮助大模型更好地适配当前领域。这种模式一定程度上赋予了模型自动学习的能力，未来可以基于此实现更加自动化、智能化的新方法。

参考文献

[1] 赵奇猛，王裴岩，冯好国，等. 面向中文专利的开放式实体关系抽取研究[J]. 计算机工程与应用，2015，51（1）：125-129，171.

[2] 吕学强，董志安. 基于SAO结构的中文专利文本实体关系抽取方法：CN109933781A[P]. 2019-06-25.

[3] LÜ X Q，LÜ X R，YOU X D，et al. Relation extraction toward patent domain based on keyword strategy and Attention+BiLSTM model（short paper）[M]//WANG X H，GAO H H，IQBAL M，et al. Collaborative Computing：Networking，Applications and Worksharing. Cham：Springer International Publishing，2019：408-416.

[4] 王熙，吕佳高. 利用机器学习对生物医药文献命名实体识别和关系抽取研究[J]. 机器人技术与应用，2020（2）：42-48.

[5] WANG R Z，XU G F，QI G L，et al. Attributed triple extraction by combination under contrastive learning[M]//ONIZUKA M，LEE J G，TONG Y X，et al. Database Systems for Advanced Applications. Singapore：Springer Nature Singapore，2024：402-414.

[6] GUO Q，GUO Y，ZHAO J. KBPT：knowledge-based prompt tuning for zero-shot relation triplet extraction[J]. PeerJ Computer Science，2024，10：e2014.

[7] 丁睿祎，王玉琢，章成志. 基于学术论文全文内容的特定领域算法实体抽取研究[J]. 数字图书馆论坛，2022（3）：2-14.

[8] 周峰，吴斌，石川. 复杂网络构建中信息抽取技术综述[J]. 数字图书馆论坛，2008（6）：28-33.

[9] DEVLIN J，CHANG M，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Stroudsburg：Association for Computational Linguistics，2019：4171-4186.

[10] WEI Z P，SU J L，WANG Y，et al. A novel cascade binary tagging framework for relational triple extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：Association for Computational Linguistics，2020：1476-1488.

[11] 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]//Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg：International Committee on Computational Linguistics，2020：1572-1582.

[12] SHANG Y M，HUANG H Y，MAO X L. OneRel：joint entity and relation extraction with one module in one step[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2022，36（10）：11285-11293.

[13] 黄政，张学福. 一种基于网页信息抽取的OA期刊资源采集方法研究[J]. 数字图书馆论坛，2017（5）：25-32.

[14] LI Z X，ZENG Y T，ZUO Y X，et al. KnowCoder：coding structured knowledge into LLMs for universal information extraction[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg：Association for Computational Linguistics，2024：8758-8779.

[15] PENG L T，WANG Z L，YAO F，et al. MetaIE：distilling a meta model from LLM for all kinds of information extraction tasks

[EB/OL]. [2024-06-16]. https://arxiv.org/abs/2404.00457v1.

[16] LI G Z, WANG P, LIU J J, et al. Meta in-context learning makes large language models better zero and few-shot relation extractors[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2024: 6350-6358.

[17] XU D R, CHEN W, PENG W J, et al. Large language models for generative information extraction: a survey[J]. Frontiers of Computer Science, 2024, 18(6): 186357.

[18] WAN Z, CHENG F, MAO Z Y, et al. GPT-RE: in-context learning for relation extraction using large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2023: 3534-3547.

[19] WADHWA S, AMIR S, WALLACE B C. Revisiting relation extraction in the era of large language models[J]. Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2023, 2023: 15566-15589.

[20] MIN S, LEWIS M, ZETTLEMOYER L, et al. MetaICL: learning to learn in context[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022: 2791-2809.

[21] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. [2024-06-16]. https://arxiv.org/abs/2201.11903v6.

[22] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[23] XIE T Y, LI Q, ZHANG Y, et al. Self-improving for zero-shot named entity recognition with large language models[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2024: 583-593.

[24] ZHANG K, GUTIERREZ B J, SU Y. Aligning instruction tasks unlocks large language models as zero-shot relation extractors[C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg: Association for Computational Linguistics, 2023: 794-812.

[25] LI G Z, WANG P, KE W J. Revisiting large language models as zero-shot relation extractors[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: Association for Computational Linguistics, 2023: 6877-6892.

[26] ZHANG M M, ZHU S C, ZHANG J M, et al. Entity relation joint extraction with data augmentation based on large language model[M]//SHI Z Z, TORRESEN J, YANG S X. Intelligent Information Processing XII. Cham: Springer Nature Switzerland, 2024: 207-214.

[27] ZHANG Y H, DU T W, MA Y S, et al. AttacKG+: boosting attack knowledge graph construction with large language models[EB/OL]. [2024-12-16]. https://arxiv.org/abs/2405.04753v1.

[28] ZHAO X Y, DENG Y, YANG M, et al. A comprehensive survey on relation extraction: recent advances and new frontiers[J]. ACM Computing Surveys, 2024, 56(11): 1-39.

[29] EFEOGLU S, PASCHKE A. Relation extraction with fine-tuned large language models in retrieval augmented generation frameworks[EB/OL]. [2024-12-16]. https://arxiv.org/abs/2406.14745v2.

[30] TIAN S B, JIN Q, YEGANOVA L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health[J]. Briefings in Bioinformatics, 2023, 25(1): bbad493.

[31] YE J J, XU N, WANG Y K, et al. LLM-DA: data augmentation via large language models for few-shot named entity recognition[EB/OL]. [2024-12-16]. https://arxiv.org/abs/2402.14568v1.

作者简介

王丽君，女，硕士，高级工程师，研究方向：数据分析、人工智能、电力信息通信技术，E-mail：wlj_lisa@126.com。
赵子岩，男，博士，正高级工程师，研究方向：电力光通信技术、数字孪生技术、电力标准化技术。
马丽，女，硕士研究生，高级工程师，研究方向：电力信息系统建设及运维工作。
蒋慧超，女，硕士，中级工程师，研究方向：电子与通信技术、人工智能、网络安全。
张冉，女，硕士，中级工程师，研究方向：网站运维、网络安全。

（下接第66页）

2021，41（18）：80-86.

[29] 赵旸. 基于引用语境分析的科研贡献点识别方法研究[D]. 北京：中国科学院文献情报中心，2024.

[30] MATSUI A，CHEN E，WANG Y W，et al. The impact of peer review on the contribution potential of scientific papers[J]. PeerJ，2021，9：e11999.

[31] 叶继元，郭卫兵. 生成式人工智能参与学术评价的反思[J]. 中国社会科学评价，2024（1）：37-48，158.

作者简介

刘晓娟，女，博士，教授，研究方向：信息计量与科学评价，E-mail：lxj_2007@bnu.edu.cn。
于姚，女，硕士研究生，研究方向：信息计量与科学评价。
沈嘉宁，女，硕士研究生，研究方向：信息计量与科学评价。

Analysis of Open Peer Review Comments for High-Level Papers and Enlightenment: Taking H1 Connect as an Example

LIU XiaoJuan YU Yao SHEN JiaNing
(School of Government, Beijing Normal University, Beijing 100875, P. R. China)

Abstract: Open peer review provides new perspectives for evaluating papers. Analyzing the characteristics of review comments on high-level papers can aid in their assessment and optimize the academic evaluation system. Taking papers on COVID-19 that are recommended by experts of H1 Connect, a globally authoritative open peer review platform in the biomedical field, and published on top journals as typical cases of high-level papers, we collect open peer review comments of these papers on H1 Connect. We analyze the attention and recognition, emotional characteristics, and value characteristics reflected in the expert review comments of high-level papers from both indicator and textual perspectives. The findings reveal that high-level papers can quickly attract the attention of peer reviewers, but only a few spark heated discussions. The open peer review texts often contain emotional sentences reflecting the reviewers' attitudes, with more positive comments addressing the overall performance of the papers and fewer negative comments mainly addressing the research methods used. The open peer review comments indicate that the value of papers is reflected in multiple dimensions including academic fields and practical applications, with an especially strong emphasis on practical contributions. Through the analysis of open peer review comments on high-level papers, it is recommended that open peer review data sources be introduced into the academic evaluation system to enhance the scientificity and effectiveness of academic evaluation. In practical applications, it is still necessary to further improve the peer review indicators, deeply explore the key elements in the review texts, strengthen resource construction and technological integration, and utilize cutting-edge technology to intelligently analyze the deep value of review texts.

Keywords: Open Peer Review; High-Level Paper; Academic Evaluation; H1 Connect

（责任编辑：管清濛）

（上接第54页）

Triplet Extraction Method for Green and Low-Carbon Field Based on Large Language Models

WANG LiJun ZHAO ZiYan MA Li JIANG HuiChao ZHANG Ran
(State Grid Information & Telecommunication Co., Ltd., Beijing 100761, P. R. China)

Abstract: Triplet extraction aims to extract entities and their relationships from text to form structured knowledge representations, which is a key technology for building automated knowledge graphs. Although traditional deep learning-based triplet extraction methods perform well when sufficient training data is available, in vertical scenarios such as the green and low-carbon sector of the power industry, the lack of standardized supervised data, high cost of manual annotation, and the presence of many specialized terms in papers and patents limit the recognition accuracy of these methods. To address these issues, this paper proposes a triplet extraction method based on large language models. By using proprietary large models to annotate a small amount of high-quality labeled data and combining retrieval-augmented techniques to guide open-source models for extraction, high-quality and automated vertical domain extraction has been achieved. Moreover, to improve extraction efficiency and precision in few-shot scenarios, this method also includes a data streamlining and complex data segmentation module, which divides the data based on the difficulty level of extraction and further divides complex data to simplify the extraction process, thereby improving the extraction effect. To verify the performance of the model, we automatically annotate a dataset of patents and papers in the power field using GPT-4, and introduce comparisons with well-known proprietary and open-source large models such as ChatGPT and ChatGLM. The experimental results demonstrate that our method achieves better extraction performance.

Keywords: Triplet Extraction; Knowledge Graph; Large Language Model; Green and Low-Carbon

（责任编辑：王玮）