数据论文

高质量的材料科学文本挖掘数据集构建方法*

刘悦[1)4)] 刘大晖[1)] 葛献远[1)] 杨正伟[1)]
马舒畅[1)] 邹喆乂[5)] 施思齐[2)3)†]

1) (上海大学计算机工程与科学学院, 上海 200444)
2) (上海大学材料科学与工程学院, 上海 200444)
3) (上海大学材料基因组工程研究院, 上海 200444)
4) (上海市智能计算系统工程技术研究中心, 上海 200444)
5) (湘潭大学材料科学与工程学院, 湘潭 411105)

(2022 年 12 月 5 日收到; 2023 年 2 月 7 日收到修改稿)

科学文献中蕴含的大量历史数据和经验知识, 对材料设计与研发具有重要参考价值. 文本挖掘尽管能高效地探索并利用被存储在海量科学文献中的信息, 但高质量文本数据的获取困难阻碍了其在材料领域更广泛的应用. 本文从品质和数量双视角剖析了材料领域的文本数据质量问题及其相关研究工作, 提出高质量的材料科学文本挖掘数据集构建方法. 该方法通过可溯源的文献自动获取方案确保文本数据的源头可追溯; 以下游任务为驱动对文献进行预处理以提升预标注文本语料的质量; 基于材料四面体准则定义适配全体系的标签注释方案以完成对语料的高品质标注; 利用融合材料领域知识的有条件文本数据增强模型实现材料文本数据量的扩充. 在不同体系数据集上的实验结果表明, 该方法可有效地提升下游文本挖掘模型的预测精度, 其中在 NASICON 型固态电解质材料实体识别任务上的 F1 值达 84%. 本文为文本挖掘在材料领域的深入应用提供理论指导和解决方案, 并有望推进数据与知识双向驱动的材料设计与研发.

关键词: 材料科学文本挖掘, 数据增强, 数据质量
PACS: 07.05.Hd, 07.05.Mh, 88.80.ff, 82.47.Jk **DOI:** 10.7498/aps.72.20222316

1 引 言

发现材料并成功应用是一个极其耗时的过程. 为了加速这一进程, 亟需以一种高效的方式探索并利用被存储在海量科学文献中的历史数据和经验知识[1–3]. 文本挖掘作为一种新兴的信息抽取技术, 能够建立深度学习算法以解释字符序列并从中获取逻辑信息[4]. 材料领域的最新研究表明, 有监督文本挖掘技术已具备从非结构化材料科学文献中提取数据和知识的能力[5–10]. 然而, 文本挖掘的成功依赖于高质量有监督数据, 在材料科学等特定领域场景中, 该类数据通常难以直接获取, 这使得数据集的构建方式变得至关重要.

近年来, 随着文本挖掘与材料科学的密切融合, 研究人员已经意识到数据集构建对文本挖掘模型的重要性. Weston 等[11] 搜集了 327 万篇文献摘要, 并从中筛选出 800 个与无机材料研究密切相关的文本语料, 基于预定义的 7 类实体标签, 手工构建了无机材料实体数据集并成功应用于材料实体识别. Friedrich 等[12] 设计了用于标记科学出版物中固体氧化物燃料电池 (solid oxide fuel cells, SO

* 国家重点研发计划 (批准号: 2021YFB3802101) 和国家自然科学基金 (批准号: 92270124, 52073169, 52102313) 资助的课题.
† 通信作者. E-mail: sqshi@shu.edu.cn

FCs) 实验信息的统一注释方案, 构建了涵盖 45 篇标注文献的 SOFCs 科学语料库, 为 SOFCs 材料文本挖掘研究开创思路. He 等[13] 从 750 篇科学文献中选取了 834 个与无机固相合成反应相关的文本段落, 提出针对无机固相合成反应抽取的文献语料库, 为反应前体挖掘和目标化合物预测提供参考. 上述研究表明, 高质量的有监督材料文本数据集是实现高精度文本挖掘的先决条件. 然而, 实现高质量文本数据集的构建亟需解决两个问题: 1) 如何提升文本数据的“质”? 2) 如何扩充文本数据的“量”?

对于“质”的提升问题, 领域知识可用于文本数据的构建与修正. 通过总结描述符关联关系、反应特性及物理规律等材料领域知识, 并将其融入预处理、标签定义和数据标注流程, 将有助于确保标注样本与专家经验的一致性. 在材料科学研究中, 已有相关工作聚焦结构化数据, 采用统计分析[14] 或机器学习方法[15] 对数据品质问题展开研究. 此外, Liu 等[16] 全面探讨了在预处理阶段嵌入领域知识进行结构化数据品质提升的可行性, 这为非结构化文本数据在构建过程中的“质”提升提供了有效的参考. 聚焦“量”的扩增问题, 可采用数据增强 (data augmentation, DA) 技术对文本数据进行增强与扩充. DA 通过学习训练实例的先验知识而产生更多的新样本以扩充原始数据集[17]. 其中, 无条件 DA[18,19] 通过操纵原始文本的部分单词来创建增强实例以实现对文本语料的无监督扩充. 然而, 由于对标签的不敏感性, 该类方法在对有监督文本语料进行增强时往往会产生与标签信息相悖的噪声样本. 尽管通用领域的有条件 DA[20–22] 以标签为约束保持了生成样本与标签语义的一致, 但受限于材料领域文本的复杂性和高度异构性, 已有方法难以直接迁移至材料文本上进行应用. 据我们所知, 目前尚未有研究提出面向材料领域非结构化文本进行有条件文本数据增强的有效方法. 若能融合领域知识对材料科学文本数据进行构建约束和样本增强, 将有助于提升文本数据的“质”并扩增建模所用数据的“量”, 从而实现高质量材料科学文本挖掘数据集的获取.

因此, 本文提出一种高质量的材料科学文本挖掘数据集构建方法, 旨在降低大规模文本数据在获取、处理、标注和扩充过程中的高昂开销. 首先, 通过可溯源的文献自动获取方案从数据和加工双视角确保文本数据集的源头可追溯; 其次, 以下游任务为驱动利用材料文本特性约束文献预处理流程, 从而提升预标注文本语料的品质; 然后, 基于材料四面体准则定义一套适配全体系的标签注释方案以完成对材料命名实体及其内联关系的标注; 进一步, 建立融合材料领域知识的有条件文本数据增强模型, 以领域知识为约束实现文本数据量的扩充. 在不同体系数据集下的实验结果表明, 本文所提方法凭借少量训练样本即可达到在原始训练样本规模下的预测精度, 有望为高质量材料文本数据集构建提供理论指导和候选方案, 进而为文本挖掘在材料领域的深入应用激发创新思路.

2 有监督材料文本挖掘数据集构建方法

本研究提出一种有监督材料文本挖掘数据集构建方法, 旨在以端到端的管道方式为材料科学文本挖掘数据集的高质量构建提供有效的解决方案, 其整体流程如图 1 所示. 该管道由可溯源的文献自动获取、下游任务驱动的文献预处理、标签定义与数据标注以及融合材料领域知识的文本数据增强四个阶段构成.

2.1 可溯源的文献自动获取

2.1.1 可溯源治理方案

可溯源性评估材料文本数据的获取、分析和应用是否可复现. 可溯源性治理有利于实验过程的调试和文本挖掘在材料领域可靠性的提升, 要求实时记录数据在获取、分析和应用过程中的相关溯源信息, 如数据来源、处理流程和分析结果等[23]. 藉此, 本文提出面向文本挖掘全流程的可溯源处理模型, 以保证文本数据的精确率、召回率和溯源性.

定义 1 可溯源处理模型 (processing model of traceability, PMTra) PMTra 可表示为三元组$\langle O_{\rm tra}, MD\left(O_{\rm tra}\right), M\left(O_{\rm tra}\right)\rangle$, 其中$O_{\rm tra}=\{o_{\rm tra}^1, o_{\rm tra}^2, \cdots, o_{\rm tra}^n\}$为溯源对象; $MD\left(O_{\rm tra}\right)$为描述溯源对象$o_{\rm tra}^n$的获取、分析和应用的元数据; $M\left(O_{\rm tra}\right)$为基于$MD\left(O_{\rm tra}\right)$的溯源机制.

在材料文本挖掘过程中, PMTra 溯源对象被实例化为数据溯源与过程溯源两个部分, 并通过建立对应的多粒度元数据结构来实现二者的可追溯.

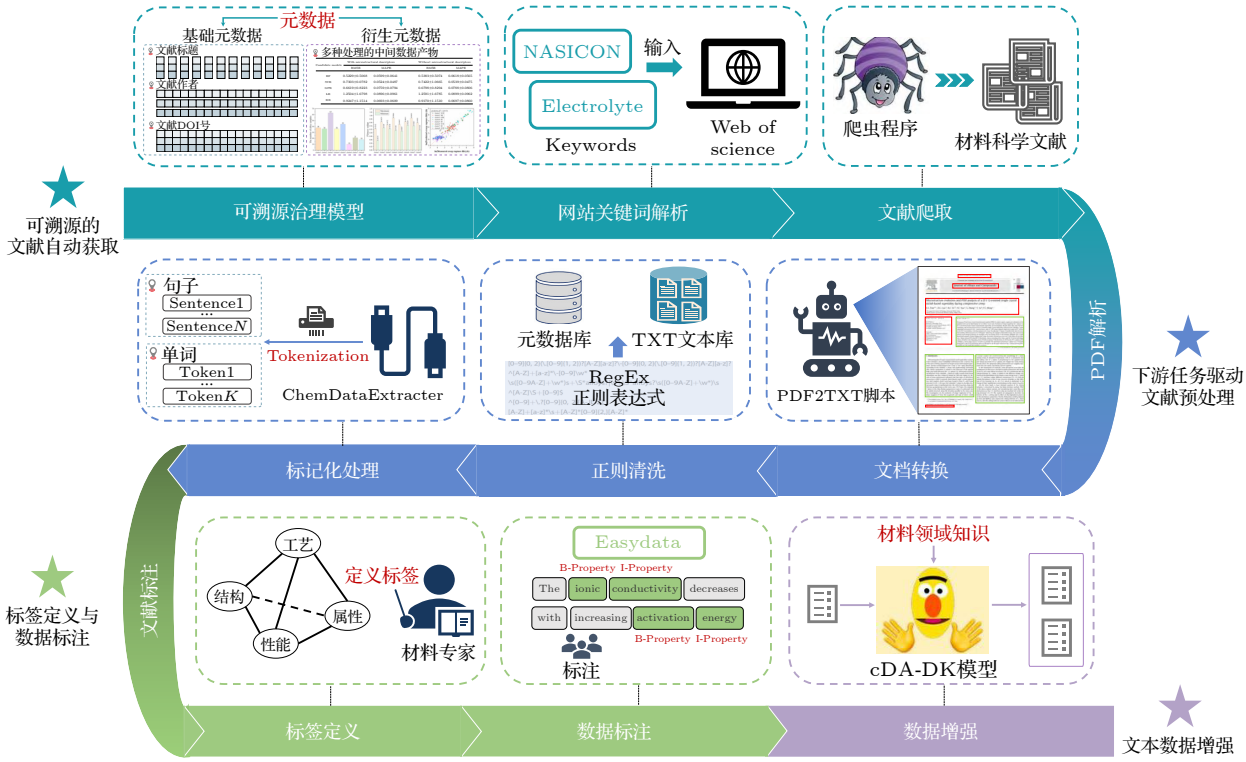

图 1 高质量材料文本挖掘数据集构建管道

Fig. 1. The pipeline for constructing high-quality datasets for materials text mining.

定义 2 基础元数据 (basic metadata, BMD) BMD 用于描述文献的基本元信息, 包括文献标题、作者、摘要、关键字、文本内容、收录期刊、发表日期、影响因子、DOI 和存储路径等.

定义 3 衍生元数据 (derived metadata, DMD) DMD 用于描述数据在处理过程中产生的、记录文献分析和应用过程的衍生元信息, 包括文献加工语料、增强数据、模型输入输出以及知识库等中间结果或衍生数据.

对于文本数据的可溯源性治理, BMD 的存储有助于识别原始数据集中明显存在的错误样本, 从而避免低精度且不合理的文本挖掘任务被执行. 例如, 文本内容可用于检查经预处理后的单条样本在语序和语义上与原文的一致性, 标题、摘要及关键字的组合可快速确定文献研究领域以保证对特定材料术语标注的准确性. 在 BMD 中, DOI 作为数字化对象的标识符, 具有唯一性. 因此, 本文利用文献 DOI 作为主数据的唯一标识符, 建立源文献、文本语料以及衍生数据之间的显式关联, 从而确保文本挖掘任务的数据可追溯.

对于处理过程的可溯源性治理, DMD 的记录和有效组织能够帮助研究人员更透彻地理解当前文本挖掘任务的完整工作流程, 进而验证当前文本挖掘方法的正确性并对其进行修改和完善. 以 DMD 为基础, 将加工操作视为有向边, 加工前后的数据视为头尾结点, 可建立文献分析和应用过程中衍生数据与加工操作之间的顺序关联, 从而确保文本挖掘任务的过程可追溯.

2.1.2 材料科学文本语料获取

语料作为文本挖掘的基石, 是训练数据和被检索信息的重要来源. 材料领域的科学文本语料通常被存储在多种类型的文档中, 如会议论文集摘要、文献、预印本、专利和电子百科全书等. 目前, 获取这些文本语料的方法主要有两种: 1) 利用现有的具备文本挖掘应用程序编程接口 (application programming interfaces, APIs) 或搜索工具的索引数据库进行手动检索获取; 2) 通过网络爬虫自动访问单个或多个发布者的内容.

本文列举了化学和材料科学领域中常见的科学文献数据库, 并对两种语料获取方式进行详细对比, 如表 1 所列. 索引数据库的主要优势在于能够提供格式统一的元数据、便捷的 API 和分析工具, 但其绝大多数出版物严重偏向生物医学和生化学科, 仅少量主题面向物理、有机化学和材料科学. 此外, 索引数据库对文档内容的访问是有限的, 这

表 1 材料科学文本语料获取方式对比

Table 1. Comparison of acquisition methods of materials scientific corpus.

获取方式	数据库	文档类型	访问权限	文档数量	参考
索引数据库 API	CAplus	论文, 专利, 报告	订阅	少	www.cas.org/support/documentation/references
	DOAJ	论文	部分订阅	少	doaj.org
	PubMed Central	论文	开放获取	较少	www.ncbi.nlm.nih.gov/pmc
	Science Direct	论文	订阅	少	dev.elsevier.com/api_docs.html
	Scopus	摘要	开放获取	较少	
	Springer Nature	论文, 书籍	订阅	少	dev.springernature.com/
网络爬虫	网页	论文, 专利, 报告, 书籍	开放获取	多	requests.readthedocs.io, crummy.com/software/BeautifulSoup

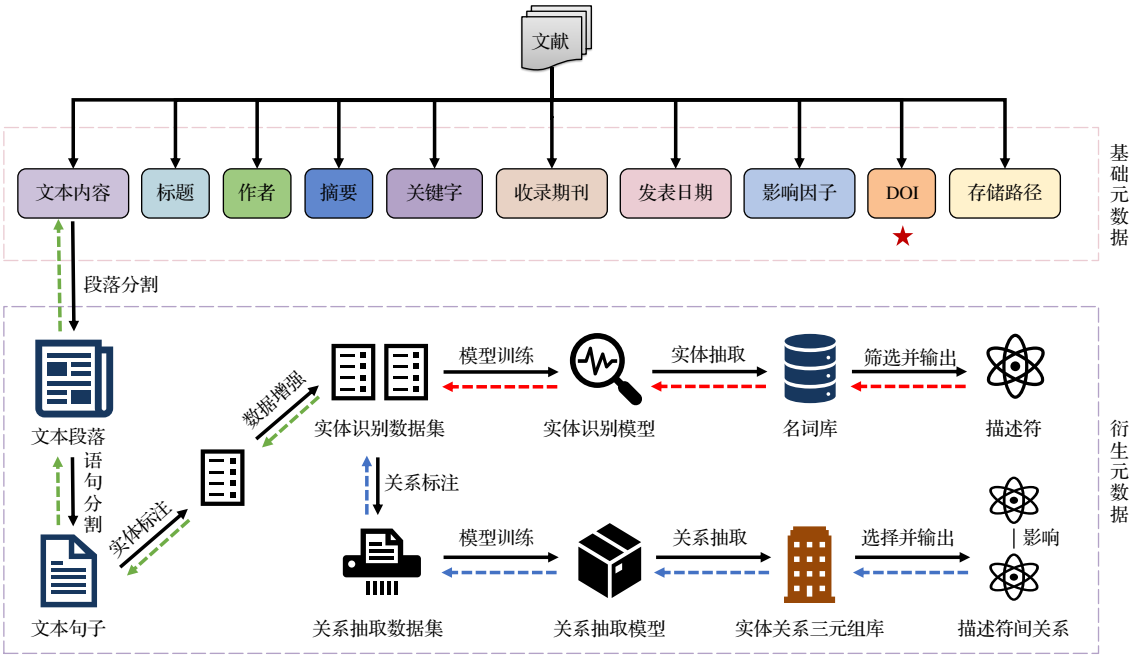

图 2 文献的数据与过程溯源示意图

Fig. 2. The illustration of the traceability of literature data and process.

严重阻碍了文本挖掘任务所需的大规模文献语料库的直接获取. 网络爬虫利用预编写的网页信息自动获取程序, 能够快速从未提供 API 的网页资源中访问目标主题内容, 而其成本仅在于大量抓取和下载操作会影响出版商服务器的运行. 因此, 为了以较低开销构建材料科学文本语料库, 本文基于 Python 爬虫工具包 Requests 和 Beautifulsoup4 研发网络爬虫程序以实现材料科学文献的自动获取.

此外, 基于 2.1.1 定义的 PMTra 模型, 将可溯源治理方案嵌入材料文献的获取过程, 藉此确保文本挖掘任务的数据与过程可追溯. 如图 2 所示, 可溯源的文献数据获取与处理过程以 BMD 和 DMD 为基础, 通过建立数据之间以及数据与加工操作之间的关联模型从而实现模型训练和推理的源头可追溯. 以命名实体识别任务为例, 在训练阶段, 实体识别模型作为输出结果, 其溯源路径可沿模型训练流表示为: <实体识别模型 $\xrightarrow{\text{训练回溯}}$ 实体识别数据集 $\xrightarrow{\text{DA与标注回溯}}$ 文本语句 $\xrightarrow{\text{句分割回溯}}$ 文本段落 $\xrightarrow{\text{段分割回溯}}$ 文本内容 $\xrightarrow{\text{DOI关联}}$ 文献>; 类似的, 在推理阶段, 描述符作为最终衍生数据, 沿模型推理步骤回溯, 可从细粒度描述符定位到其来源文献, 从而实现衍生数据的源头可追溯. 推理阶段的溯源路径可表示为: <描述符 $\xrightarrow{\text{筛选回溯}}$ 名词库 $\xrightarrow{\text{抽取回溯}}$ 实体识别模型 $\xrightarrow{\text{输入回溯}}$ 文本语句 $\xrightarrow{\text{句分割回溯}}$ 文本段落 $\xrightarrow{\text{段分割回溯}}$ 文本内容 $\xrightarrow{\text{DOI关联}}$ 文献>.

2.2 下游任务驱动的文献预处理

材料科学文本语料的存储格式因其可访问程

度、目标主题和文档种类的差异而大相径庭, 而语料内容仅在以纯文本且可被直接访问的格式呈现时才能被用于后续的抽取任务. 在网络爬虫方式下, 抓取的内容由完整的纸质文件组成, 其中便捷式文档格式 (portable document format, PDF) 是科学文本存储的主要载体. 然而, 嵌入式 PDF 通常具有块结构[4], 其内容按列排版且与表格、图片和公式等信息混合, 严重影响格式转换的可操作性和文本序列的可读性. 因此, 文献预处理阶段的首要任务是对 PDF 格式文档进行解析以提取出可自由访问的纯文本信息; 其次, 为了得到由逻辑构成 (如句子) 或标记 (如单词或短语) 等一系列可表征原始语法结构的预标注语料, 需要借助标记化工具对前一步所得的纯文本内容进行分割, 并将其处理为一个单独或整体的序列信息. 藉此, 本文提出以下游任务为驱动的多阶段文献预处理方法, 利用材料文本特性约束文献预处理流程, 进而提升预标注文本语料的质量.

2.2.1 基于文档转换与正则清洗的前期预处理

预处理前期的目标是获得科学文献中以纯文本格式存储的正文段落. 在此阶段, 首先使用文本处理工具 PDFMiner[24] 对原始文档进行标准化、分割和语法解析, 将 PDF 格式文献转换为 TXT 纯文本数据; 然后, 利用基于正则表达式所定义的语法规则, 清洗纯文本中的乱码、断行、图表等冗余信息, 对摘要以及与实验合成或材料表征相关的正文段落进行分类与存储; 最后, 摘要及正文内容作为文本输入参与后续阶段处理, 标题、作者、参考文献等基础元信息则被存储于元数据库中以确保文献源可追溯.

2.2.2 基于 ChemDataExtracter 的后期预处理

预处理后期的目标是获得文本内容中可表征原始行文逻辑的语句或标记等序列信息, 以构造高质量的预标注文本语料. 标记化 (tokenization) 旨在以特定需求对句子进行逻辑划分, 得到由多个标记 (tokens) 构成的单词或词组序列. 作为文本挖掘的关键步骤, 标记化过程中产生的错误信息往往会沿着管道传播并影响最终结果的准确性. 对于通用英文文本, “. ”, “, ”和“?”等符号是识别句段开始/结束的关键标记, 但其难以直接适配科学文献的标记化需求. 常见表达方式如“Fig. X”, “et al.”和化学式中的句号通常会导致段落的过度分割. 相反, 句末的引文编号反而促进了两个无关语句的合并. 此外, 材料科学文献具有领域特殊性, 针对由多个单词或符号组成的化学术语的标记化处理变得十分复杂. 因此, 亟需借助材料领域特定的文本处理工具, 以实现材料领域复杂文本的标记化处理.

本文对比了化学和材料科学领域常用的自然语言处理工具包, 如表 2 所列. ChemDataExtracter[27] 以基于规则和领域字典的方式解决标记化问题, 且具备材料简单、友好性高、适用范围广等优势, 在多个功能需求中综合表现良好. 因此, 在经过 PDFMiner 转换及正则清洗得到 TXT 纯文本数据后, 本文进一步选择 ChemDataExtracter 作为在预处理后期对其进行标记化处理的工具. 基于该工具完成对文本内容的分段操作, 获得以自然段序列为构成的逻辑结构; 然后以文本段落为基础进行分句和标记化, 得到以句为单位的细粒度标记 (单词) 序列; 最后, 对材料语句进行解析和词性标注, 以识别每个标记的语法属性 (如名词、动词、冠词和形容词等), 进而提升预标注语料的质量.

2.3 标签定义与数据标注

2.3.1 基于材料四面体的标签注释方案

类别标签是影响有监督材料文本挖掘模型性能的先决条件之一, 其在较大程度上决定了数据标注的质量以及文本挖掘的结果. 如表 3 所列, 已有研究大多从特定的材料应用点出发, 设计了适用其目标下游任务的标签类别. 它们往往在单一体系上表现良好, 却难以进行跨领域的迁移和细粒度的材料信息挖掘.

表 2 化学与材料科学中常用的自然语言处理工具

Table 2. Common natural language processing tools in chemistry and materials science.

名称	适用范围	是否开源	版本迭代	功能完备性	难易性	友好性
OSCAR4[25]	化学反应和生物化学	是	快	中	普通	中
ChemicalTagger[26]	化学合成作用和条件	是	慢	中	普通	中
ChemDataExtractor[27]	通用化学和材料科学领域	是	快	高	容易	高

表 3 已有材料文本挖掘研究中的实体标签定义对比

Table 3. Comparison of entity label definitions in previous materials text mining research.

来源	目标	标签数	标签类别	适用领域	应用实例
Weston 等[11]	构建材料领域最新研究结果与历史文献的关联	7	无机材料, 相结构, 描述符, 属性, 应用, 合成方法, 表征方法	无机材料	目标材料检索, 文献搜索与总结, 元信息分析
He等[13]	从无机固相合成反应文献中挖掘反应前体信息	3	材料, 合成反应前体, 目标化合物	无机固相合成反应	固相合成反应前体数据挖掘, 元信息分析
Friedrich等[12]	标注科学出版物中与SOFCs实验相关的信息	4(SOFC) 17(SOFC-slot)	实验, 材料, 数值, 应用等	电池材料	构建SOFCs科学语料库并用于多个实验信息提取任务
Wang 等[10]	从文献中自动挖掘出数据驱动的材料设计模型所需的高质量可靠数据	6	元素, 合金命名实体, 成分含量, 属性描述符, 属性值, 其他	合金材料	钴基单晶高温合金 γ' 相固溶温度预测
Nie等[9]	构建语义表示框架以探索潜在的锂离子电池阴极材料	3	无机材料, 锂离子电池阴极材料, 属性描述符	电池材料	新型锂离子电池阴极材料设计与寻优

表 4 面向通用领域的材料实体类型定义

Table 4. The definition of materials entity types in the general domain.

实体标签	定义	示例
Composition	与化学式有关的内容; 描述材料内部与含量相关的内容等.	NaCl, $CaCl_2$; Na concentration, Electrons charge carriers.
Structure	晶体结构; 相; 用于刻画晶体结构的名称等.	Fcc, Phase; Bottleneck, Channel, Path.
Property	带单位的可度量值; 材料表现出来定性的性质或现象; 描述材料产生物理/化学行为或物理/化学机制的名词等.	Conductivity, Activation, Radius; Ferroelectric, Metallic; Phase transition, Ionic reaction.
Processing	材料合成技术或加工工艺; 材料改性手段等.	Solid state reaction, Annealing; Doping.
Characterization	用于表征材料的任何实验、理论、模型或公式等.	XRD, STM, Photoluminescence, DFT; Bethe-Salpeter equation.
Application	任何高级的应用; 任何特定的器件、系统等.	Cathode, Photovoltaics; Battery Management System.
Feature	样品类型、形状的特殊描述.	Single crystal, Bulk, nanotube, Quantum dot.
Condition	描述材料所处的环境或外部条件.	980 °C , 1000 MPa.

材料四面体即材料学四要素, 其旨在研究材料的成分、结构、加工工艺、性能以及它们之间的关系[28]. 这四个要素彼此紧密联系, 形成了构效关系研究和新材料设计的根本基础, 被视为材料科学领域的研究范式. 藉此, 以加工工艺-结构-成分-性能四面体为准则, 本文首先对通用领域下材料文本挖掘的目标内容进行总结. 在此基础上, 通过对已有内容进行高度抽象, 提出 8 个适配全体系的描述符实体类型, 如表 4 所列.

为了捕获材料实体之间的潜在关联, 本文进一步定义通用领域下的材料实体关系类型, 如表 5 所列. 上述基于材料四面体准则的实体关系标签注释方案满足通用领域场景下的材料文本挖掘需求, 可实现对大规模材料文本语料的标注. 此外, 该方案允许对特定材料类别进行细粒度优化, 以满足特殊场景下的挖掘需求.

2.3.2 实体关系标注

高质量的有监督文本挖掘数据集的构建仍然依赖领域专家的手工标注. 为了尽可能降低标注开销, 选择简单易用且适配的标注工具显得非常必要. 表 6 对比了目前常用的文本标注工具, 本文从中选取综合表现较好的 EasyData 进行实体关系标注. 具体流程如图 3 所示. 首先, 基于预处理后期所得的标记化序列, 对语句中的关键材料术语(由一个或多个标记构成) 赋予对应的实体类型以完成粗粒度的材料实体标注. 其次, 采用 BIO[29] 序列标注方法, 对材料语句中的实体和非实体信息进行细粒度的划分. 在这种标注模式中, 实体的开头(B-begin)、内部 (I-inside) 和外部 (O-outside) 均被特殊标记覆盖. 这对于解释多词实体 (如“thin film”) 而言是必要的. 最后, 为表征两个实体之间的潜在关系, 引入〈e1〉〈/e1〉和〈e2〉〈/e2〉等特殊

表 5 面向通用领域的材料实体关系类型定义

Table 5. The definition of materials relation types in the general domain.

关系标签 (A to B)	定义	可能存在此关系的实体类型
Cause-Effect	A对B有影响	Property-Property, Composition-Structure, Structure-Property, ...
Component-Whole	A是B的部分	Composition-Composition, ...
Feature-Of	A是B的特征	Feature-Composition, Feature-Application, ...
Located-Of	A占据了B位置	Composition-Structure, ...
Instance-Of	A是B的实例	Composition-Composition, Structure-Structure, Property-Property, ...
Condition-On	A的条件是B	Processing-Condition, ...
Method-Of	A的表征方法是B	Property-Characterization, ...
Other	A与B存在除上述关系类型外的其他关系	—

表 6 常用文本标注工具对比

Table 6. Comparison of common tools for text annotation.

标注工具	适配任务	文本要求	角色管理权限	难易性	友好性	可扩展性	参考
Label Studio	多模态信息标注	严格	不完善	普通	中	中	labelstud.io
Brat	关系标注	一般	完善	普通	中	低	github.com/nlplab/brat
Doccano	文本分类	严格	较完善	普通	低	低	github.com/doccano
EasyData	实体与关系标注	一般	完善	容易	高	高	ai.baidu.com/easydata/

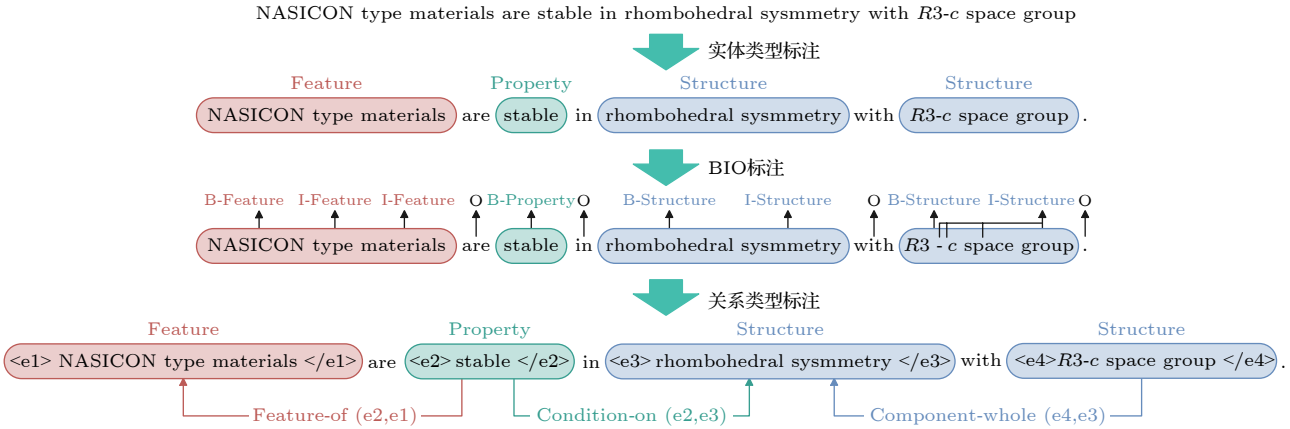

图 3 实体关系标注流程示意图

Fig. 3. The process of annotation on entities and relations.

标记包裹三元关系中的主体或客体, 并在指定关系标签后赋予 (e1, e2) 或 (e2, e1) 标记以表示关系的方向性.

2.4 融合材料领域知识的文本数据增强

大规模高质量的文本挖掘数据集的构建成本极高. 尽管 DA 已被广泛应用于解决深度学习的小样本问题, 但受限于材料文本的特殊性, 如 $(Y, In)BaCo_3ZnO_7$, $(La_{0.8}Sr_{0.2})_{0.97}MnO_3$ 等复杂化学式以及“X-ray powder diffraction (XRD)”, “bond valence sums (BVS)”等缩写形式存在, 导致一般方法难以直接应用于材料文本数据的增强任务. 藉此, 本文提出融合材料领域知识的有条件文本数据增强模型 (conditional data augmentation model of incorporating materials domain knowledge, cDA-DK), 将材料领域知识融入预训练语言模型中, 通过微调使其学习材料领域词汇特征, 从而动态生成高质量的材料文本数据.

图 4 展示了 cDA-DK 的处理流程. cDA-DK 采用预训练语言模型 DistilRoBERTa(Roberta[30] 的知识蒸馏版本) 以大规模扩增材料文本数据. 由于模型的原始词汇表偏向通用领域, 导致其在对材料文本进行处理时会造成语句被过度分割, 从而丢失领域词汇的特殊语义. 此外, 过度分割会导致待

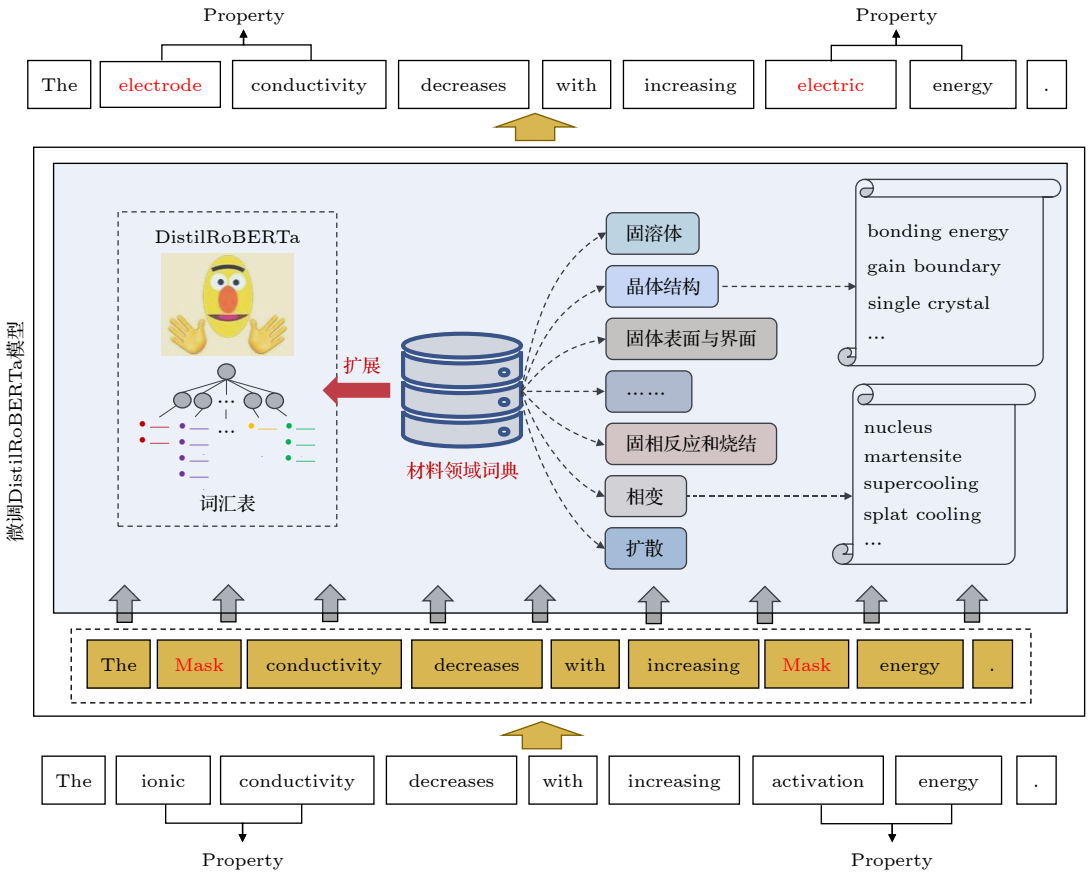

图 4 基于 cDA-DK 的材料文本数据增强

Fig. 4. Materials textual data augmentation based on cDA-DK.

处理序列过长, 这进一步限制原模型的微调性能. 为加速微调效率并提高对材料特殊词汇的感知能力, 本文通过搜集大量材料科学术语以建立材料领域词典, 将其作为外部知识融入 DistilRoBERTa 模型. 具体地, 检查材料领域词典中的科学词汇是否存在于 DistilRoBERTa 的原始词汇表, 并将不存在于其中的材料术语通过原模型“add_tokens”方法添加分词以扩展词汇表. 基于此, 通过“resize_token_embeddings”方法为新增词汇训练与已有词汇相同维度的嵌入向量并添加至原有的词嵌入矩阵中, 由此便获得材料领域知识指导下的 DistilRoBERTa 模型.

为使得增强模型更契合下游任务, 以材料文本为输入对领域知识指导下的 DistilRoBERTa 模型展开无监督训练, 实现其对文本数据增强任务的微调. 在此阶段, cDA-DK 的分词器不会过度分割材料领域特殊词汇, 其分词结果保留了原词汇的语义信息, 减少了模型所需处理的文本序列长度, 进而提高微调效率.

最后, 利用微调后的 DistilRoBERTa 模型捕获输入语句的上下文语义, 实现有监督材料文本的数据扩充. 具体地, 以待增强的文本数据及其对应的标签信息作为输入属性, DistilRoBERTa 随机遮掩语句中的部分单词并记录对应标签信息. 通过建立词汇与标签之间的依赖关系, 模型产生语义丰富的嵌入向量以表征每个词的信息. 在此基础上, 利用嵌入向量对被遮掩位置进行预测以生成候选的增强词汇表, 并从中选择与原始语义及标签信息最为相似的单词作为增强后的文本数据. 值得注意的是, 增强数据完全依赖于微调 DistilRoBERTa 模型所学习的材料语义知识, 这可以有效地减少噪声数据的产生.

cDA-DK 的伪代码如算法 1 所示. $D_{\text{train}} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$为原始数据集, 其中 $x_i = \{t_1, t_2, \cdots, t_n\}$表示长度为 n的句子, $y_i = \{l_1, l_2, \cdots, l_n\}$则为句中每个标记对应的标签信息; $P_{\text{DistilRoBERTa}}$表示预训练语言模型 DistilRoBERTa, $F_{\text{DistilRoBERTa}}$为其微调后的版本; $C = \{w_1, w_2, \ldots, w_m\}$表示词汇数为 m 的材料领域词典; $D_{\text{synthetic}}$ 为增强后的材料文本数据集.

算法1 数据增强方法cDA-DK

输入 原始数据集 $D_{\text{train}}=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$
预训练语言模型模型 $P_{\text{DistilRoBERTa}}$
材料领域词典 $C=\{w_1,w_2,\ldots,w_m\}$

输出 增强数据集 $D_{\text{synthetic}}$

1: 开始

2: **for** $w_i \in C$ **do**

3: w_i 输入至 $P_{\text{DistilRoBERTa}}$ 的词汇表并训练其对应的词向量

4: 在下游任务文本数据增强上微调 $P_{\text{DistilRoBERTa}}$ 得到 $F_{\text{DistilRoBERTa}}$

5: 初始化 $D_{\text{synthetic}}=\{\}$

6: **for** $\{x_i,y_i\} \in D_{\text{train}}$ **do**

7: $(\widehat{x}_i,\widehat{y}_i)=F_{\text{DistilRoBERTa}}(x_i,y_i)$ // 生成新的样本

8: $D_{\text{synthetic}}=D_{\text{synthetic}}\cup(\widehat{x}_i,\widehat{y}_i)$ // 生成样本加入增强数据集

9: 结束

3 结果与讨论

3.1 NASICON 有监督文本挖掘数据集的构建与扩充

钠超离子导体 (NASCION) 材料因其表现出高离子电导率, 具有优异的化学及热力学稳定性以及快速简单的合成工艺等优点, 在二次电池固态电解质/电极材料研究中受到广泛关注[31,32]. 本文以 NASICON 型固态电解质材料文本挖掘任务为例, 验证所提方法的有效性.

3.1.1 数据集构建

基于有监督材料文本挖掘数据集构建管道, 三位领域专家共同处理并标注了 55 篇与 NASICON 型固态电解质相关的材料科学文献, 藉此构建面向 NASICON 体系的实体关系抽取数据集. 如表 7 所列, 与通用领域的公开数据集 CoNLL-2004 相比, NASICON 实体关系数据集在标签类别上更为丰富, 可适配特定领域下细粒度的材料信息挖掘目标. 此外, NASICON 数据集的标签规模与 CoNLL-2004 相当, 而样本数量几乎是后者的一倍. 这是因为前者在标注过程中出现较多冗余数据, 本文将这部分信息作为负样本得以保留, 可藉由对比学习方式提升下游模型对硬负样本的识别准确率.

表 7 NASICON 实体关系数据集与 CoNLL-2004 数据集的对比

Table 7. Comparison of the NASICON dataset with the CoNLL-2004 dataset.

数据集	样本数	实体类型	实体数	关系类型	关系数
CoNLL-2004	1, 441	4	5, 347	5	2, 020
NASICON	2, 434	8	4, 857	8	2, 297

分析标注示例 (见表 S1 和 S2), 由于科学文献中领域词汇的专业性和语法结构的复杂性, 标注样本中存在大量一词多义和关系重叠等问题. 此外, 对比图 5 中 NASICON 和 CoNLL-2004 中样本关系三元组个数和语句长度的分布情况可以发现, 材料文本中的语句一般较长 (超过 20 个单词), 且句中隐含的实体关系较为复杂 (超过一半以上的样本其三元组个数大于 2—3 个), 样本量整体呈现正态分布且存在严重的长尾分布现象, 这印证了材料语句结构的复杂特性.

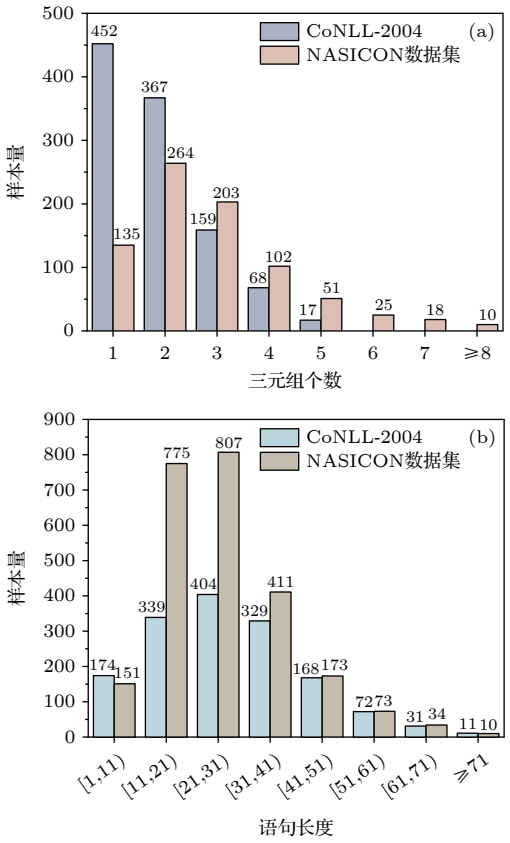

图 5 两份数据集的样本统计情况对比 (a) 三元组个数分布情况; (b) 语句长度分布情况

Fig. 5. Comparison of sample statistics of two datasets: (a) The distribution of numbers of triplets; (b) the distribution of length of sentence.

3.1.2 数据集扩充

本节基于 cDA-DK 方法对 NASICON 实体关系数据集进行增强, 以实现有监督文本挖掘数据集的大规模扩充. 增强前后的数据对比如表 8 所列. 经过数据增强后, 原数据集在样本量、实体和关系

表 8 NASICON 实体关系数据集在增强前后的数据示例对比

Table 8. Comparison of samples before and after augmentation of NASICON dataset.

数据集	样本数	实体数	关系数	示例
原始数据集	2434	4857	2297	The (O) **ionic** (B-Property) conductivity (I-Property) decreases (O) with (O) increasing (O) **activation** (B-Property) energy (I-Property) . (O)
cDA-DK 增强数据集	4846	9714	4594	The (O) **electrode** (B-Property) conductivity (I-Property) decreases (O) with (O) increasing (O) **electric** (B-Property) energy (I-Property) . (O)

表 9 实验数据集信息

Table 9. The details of experimental datasets.

数据集名称	应用领域	重命名	样本量	语料规模	来源
NASICON 实体识别数据集	NASICON 型固态电解质	Dataset 1	2, 434	55篇文献	领域专家标注
		Dataset 2	2, 434	—	数据增强
		Dataset 3	305	35篇文献	非专业人员标注
Matscholar[11]	无机材料	Dataset 4	5, 459	800份摘要	领域专家标注
		Dataset 5	5, 459	—	数据增强

规模上均扩充了一倍. 此外, 对比二者示例可知, cDA-DK 产生的增强词汇在材料语义及标签特性上均与原始领域词汇保持一致, 且增强样本在语义和语法结构上基本符合材料领域专家认知. 藉此, 实现了由小规模标注样本到大规模高质量文本挖掘数据集的扩充.

3.2 cDA-DK 有效性检测实验

为进一步检验增强数据集的质量, 本节以材料命名实体识别作为目标任务, 在 NASICON 数据集和公开的材料实体识别数据集上展开对比实验, 以评估 cDA-DK 的有效性及其在不同材料体系下的迁移能力.

3.2.1 实验数据

实验数据来源于 3.1 节构建并扩充的 NASICON 实体识别数据集以及 Weston 等 [11] 公开发布的实体识别数据集 Matscholar, 详细信息如表 9 所列. 其中, 本文将 3.1.1 节构建的 NASICON 原始数据集记为 Dataset 1, 将 cDA-DK 在该数据集上的增强部分记为 Dataset 2; 类似地, Matscholar 的原始数据记为 Dataset 4, 增强部分记为 Dataset 5. 此外, 本节另标注了 35 篇 NASICON 相关文献, 以作为小样本实体识别数据集 Dataset 3. 值得注意的是, Dataset 3 是由非材料领域研究者在专家指导下完成标注, 因此其质量和数量均劣于 Dataset 1.

3.2.2 实验设置

cDA-DK 增强模型的实验设置如下: 随机遮掩输入语句中的 3 个单词; 利用微调的 DistilRoBERTa 模型生成与被遮掩单词语义相近的 5 个候选单词; 通过计算余弦相似度选择与原始单词相似度最大的候选单词作为增强数据.

针对下游命名实体识别任务, 采用课题组近期提出的基于多层语义特征融合的材料实体识别模型 MatBERT-BiLSTM-CRF[33] 作为基线, 以评估实验数据集在具体任务上的性能表现. 模型的实验参数设置如表 S3 所列, 数据集划分比例为 8:1:1. 此外, 本文采用精确率 (Precision)、召回率 (Recall) 和 F1 值 (F1-score) 来评估模型的预测精度. 其中, F1 值作为精确率和召回率的调和平均数, 在性能评估中起主导作用.

3.2.3 实验结果与分析

1) cDA-DK 模型性能验证

本节在原始数据集和增强数据集上分别建立命名实体识别模型, 以验证 cDA-DK 模型对材料文本挖掘数据集的增强性能. 表 10 展示了实体识别模型基于不同实验数据集训练并在相同测试集上进行评估后的实验结果. 其中, 在 Dataset 2 及 Dataset 5 增强数据集上, 实体识别模型的 F1 值分别可达 0.70 和 0.77. 这表明仅由增强样本训练所得的模型已经能以较高的准确度从材料文献中挖掘预定义的实体信息. 尽管该得分略低于在领域专家标注数据集上的结果, 但前者的优势在于, 2 和 5 这两份数据在构建过程中无人工开销而完全由 cDA-DK 主导产生. 此外, 在由增强样本和少量标注样本组合而成的数据集 (Dataset 2+3) 上, 实体识别模型的预测精度相较于 Dataset 1 提

表 10 实体识别模型在不同材料数据集上的实验结果

Table 10. The results of NER model on various materials datasets.

数据集	材料类别	样本量	Precision	Recall	F1-score
Dataset 1	NASICON 型固态电解质	2, 434	0.78	0.83	0.80
Dataset 2		2, 434	0.68	0.72	0.70
Dataset 2+3		2, 739	0.83	0.85	0.84
Dataset 4	无机材料	5, 459	0.86	0.90	0.88
Dataset 5		5, 459	0.75	0.78	0.77

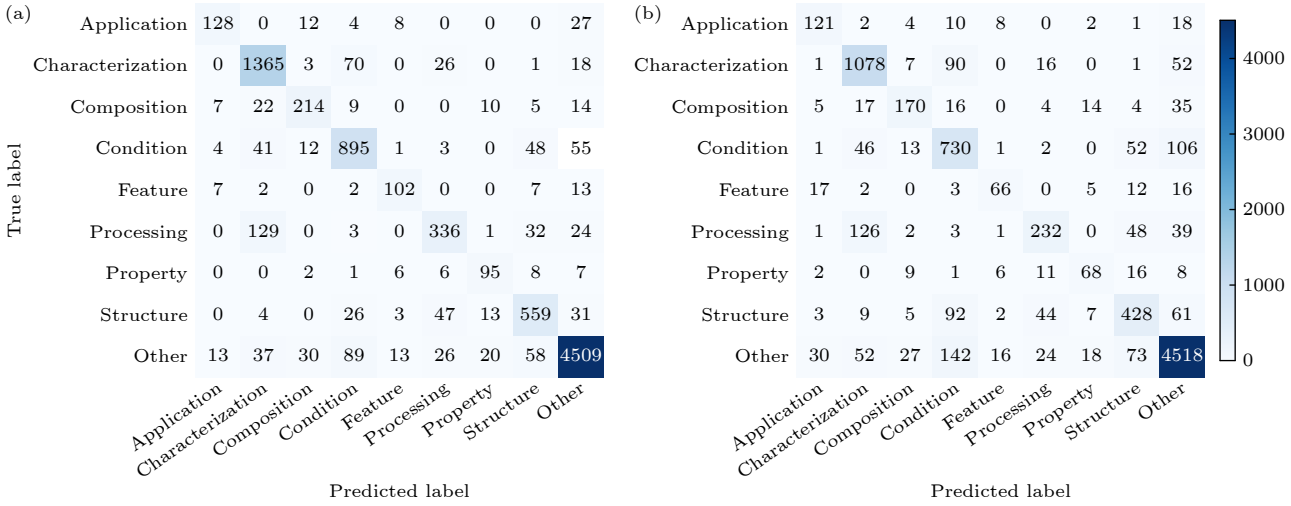

图 6 实体识别模型在不同数据集上的混淆矩阵 (a) Dataset 1 的混淆矩阵; (b) Dataset 2 的混淆矩阵

Fig. 6. Confusion matrix of NER model on various datasets: (a) The confusion matrix of Dataset 1; (b) the confusion matrix of Dataset 2.

升了 5%, 而前者的样本规模与后者几乎相当. 这表明模型仅凭借少量训练样本即可达到其在原始样本规模下的预测精度, 进一步印证了 cDA-DK 对文本数据增强的有效性, 能够充分减少有监督材料文本挖掘数据集的构建成本. 同时, 在 NASICON 型固态电解质和无机材料两个不同类别数据集上的实验结果表明, cDA-DK 具备一定的鲁棒性和迁移能力.

2) cDA-DK 生成数据质量验证

为进一步验证 cDA-DK 生成的数据质量, 本节以 NASICON 数据集为例, 通过构建实体识别模型在 Dataset 1 和 Dataset 2 上的混淆矩阵, 评估在每个实体类别下增强样本与原始样本的质量差距. 图 6 为混淆矩阵结果, 矩阵中的数值表示行标签实体类别被预测为列标签实体类别的个数, 而对角线元素则表示被模型正确预测的个数. 由图可知, 无论是人工标注数据还是 cDA-DK增强数据, 实体识别模型在 8 个实体类别上的正确预测占比均为最高, 这表明 cDA-DK 产生的增强数据质量与领域专家标注数据的质量分布高度吻合. 模型在人工标注数据集上训练后的预测精度普遍优于增强数据集, 这是必然结果. 因为 Dataset 1 是由多名领域专家标注并评估后构建的高质量数据, 而 cDA-DK 在有条件数据增强过程中会受到难以避免的标签分布不均匀以及噪声数据等问题的影响, 导致其对部分实例的处理能力下降. 但值得注意的是, cDA-DK 在无需人工标注的情况下产生的文本数据质量已无限接近领域专家标注数据. 此外, 实体识别模型在两份数据上对“Other”类型实体的正确预测比例几乎完全一致, 印证了 cDA-DK 对原始数据中硬负样本的强识别性能.

3) cDA-DK 对下游模型训练效率的影响

图 7 展示了 MatBERT-BiLSTM-CRF 在人工标注及其增强数据集上训练和验证过程中损失函数 (loss function) 值的变化曲线. 由图 7(a) 和图 7(c) 可知, 在人工标注数据集上, 模型分别在第 14 次及第 12 次训练迭代过程中收敛; 由图 7(b) 和图 7(d) 可知, 在对应的两份增强数据集上模型分别在第 11 次及第 9 次训练迭代过程中收敛, 即模型在增强数据集上的训练和收敛速率有所提升.

这是由于 cDA-DK 并未直接使用预训练 Distil-RoBERTa 模型进行数据增强, 而是通过融合材料语料库中的领域词汇对原模型进行微调. 在该过程中, 材料特殊术语不会被分词器过度分割, 从而减小了下游模型所需处理的序列长度, 进而提升了下游模型的训练效率. 此外, 融入材料领域知识不仅可以加快语言模型的微调进程, 还能更好地捕获材料领域词汇所处的上下文语义信息, 为高质量材料文本数据的生成奠定基础.

3.3 应用初探

本文以 NASICON 型固态电解质材料的激活能构效关系研究为例初步探索了所建数据集的潜在用途. 具体地, 首先利用基于 NAISOCN 实体关系数据集训练所得的实体识别模型从 1808 篇 NASICON 材料科学文献中提取出 106896 个描述符实体; 其次, 以激活能预测为目的从中筛选出 408 个高质量的性能驱动描述符; 据此, 建立了 6 个不同的机器学习模型对目标属性激活能进行预测, 其中最优预测模型的决定系数 (R^2) 可达 96%(详见表 S4)[33]. 本文进一步分析了预测模型中对结果起决定作用的部分描述符, 进而为研究激活能驱动的材料构效关系奠定基础. 如图 8 所示, 以“Temperature”为例的 Condition 类实体、

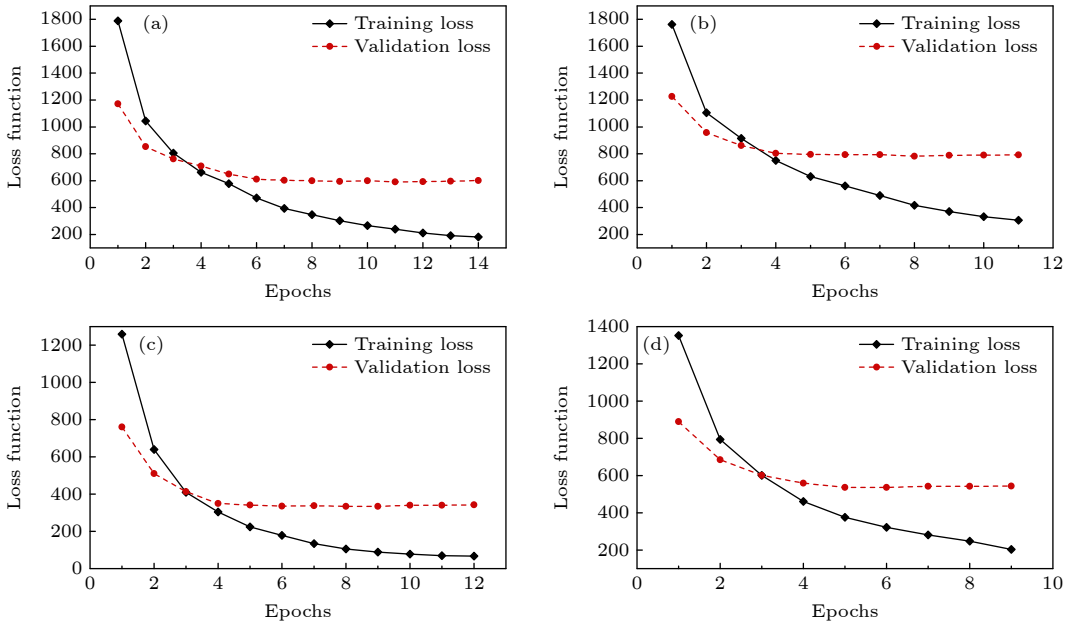

图 7 MatBERT-BiLSTM-CRF 在不同数据集上的训练及验证 Loss 变化曲线 (a) Dataset 1 上的 Loss 变化曲线; (b) Dataset 2 上的 Loss 变化曲线; (c) Dataset 4 上的 Loss 变化曲线; (d) Dataset 5 上的 Loss 变化曲线

Fig. 7. The training and validation loss function of MatBERT-BiLSTM-CRF on various datasets: (a) The loss function on Dataset 1; (b) the loss function on Dataset 2; (c) the loss function on Dataset 4; (d) the loss function on Dataset 5.

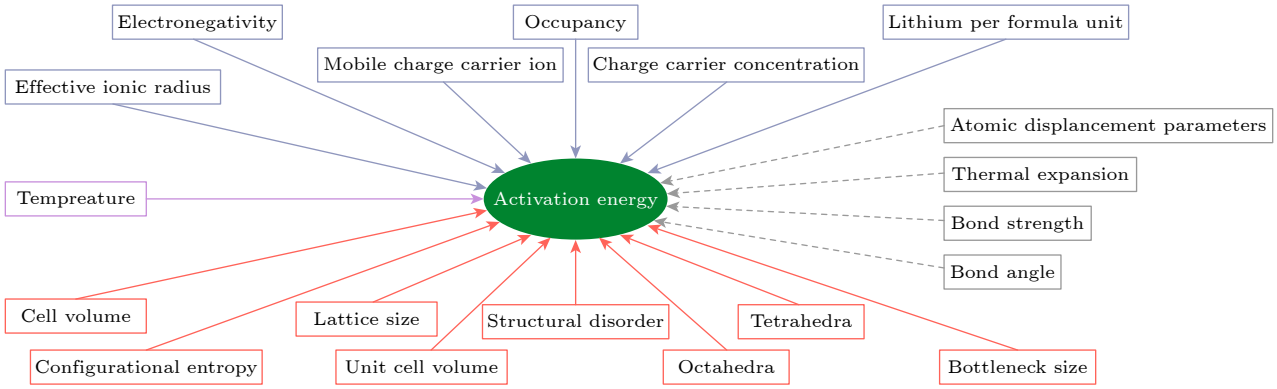

图 8 对激活能预测起关键影响的部分描述符, 其中虚线表示尚未被研究的潜在描述符[33]

Fig. 8. Partial descriptor entities that are critical for predicting activation energy, of which dotted lines indicate potential ones still to be developed[33].

以“effective ionic radius”和“Occupancy”等为例的 Composition 类实体以及以“octahedra”和“lattice size”等为例的 Structure 类实体均对 NASICON 型固态电解质材料的激活能产生关键影响. 此外, 在所筛选的描述符实体中, 本文还发现了一些尚未被研究的且与激活能预测具有潜在关联的候选描述符. 例如, “bond strength”可能是有用的, 因为离子输运性能可能与 M—O 和 X—O 之间的键强度有关. 上述应用示例为理解性能驱动的构效关系提供了重要见解, 有望促进材料的设计与发现.

4 总 结

文本挖掘因其能高效地探索并利用被存储在海量科学出版物中的数据与知识而被逐渐应用于材料科学研究. 尽管研究人员已经意识到数据构建对材料文本挖掘建模的重要性, 但仍然缺乏对数据质量内涵的深入理解和高质量文本数据集构建的有效策略. 本文厘清了数据构建全流程中“质”与“量”的关联. 基于此, 提出了有监督材料文本挖掘数据集构建管道, 通过融合材料领域知识对文本数据进行品质提升和样本增强. 该管道包含可溯源的文献自动获取、下游任务驱动的文献预处理、标签定义与数据标注以及 cDA-DK 增强四个步骤. 前三步在材料领域知识介入下确保构建过程与专家经验的一致性从而提高文本数据的“质”, 后一步融合领域知识进行有条件文本数据增强来扩增建模所用数据的“量”. 在两份不同体系数据集的实验结果表明, cDA-DK 凭借少量训练样本即可超过在原始训练样本规模下的预测精度. 其中, 在 NASICON 实体识别任务上的 F1 值达 84%. 本方法可降低大规模有监督材料科学文本数据集在构建过程中的高昂开销, 还能有效地提高下游文本挖掘模型的预测精度, 对进一步提升材料文本挖掘的普适性、准确性和实用价值具有十分重要的意义.

附 录

作者声明支撑本研究成果的主要数据可在正文及其补充材料中查询. 其他实验数据和源代码可根据合理要求从通讯作者 (Email: sqshi@shu.edu.cn) 处获得.

参考文献

[1] Gupta T, Zaki M, Krishnan N M A, Mausam 2022 *npj Comput. Mater.* **8** 102

[2] Olivetti E A, Cole J M, Kim E, Kononova O, Ceder G, Han T Y J, Hiszpanski A M 2020 *Appl. Phys. Rev.* **7** 041317

[3] Venugopal V, Sahoo S, Zaki M, Agarwal M, Gosvami N N, Krishnan N M A 2021 *Patterns* **2** 100290

[4] Kononova O, He T, Huo H, Trewartha A, Olivetti E A, Ceder G 2021 *iScience* **24** 102155

[5] Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E 2017 *Chem. Mater.* **29** 9436

[6] Mysore S, Jensen Z, Kim E, Huang K, Chang H S, Strubell E, Flanigan J, McCallum A, Olivetti E 2019 *Proceedings of the 13th Linguistic Annotation Workshop* Florence, Italy, August 1, 2019 p56

[7] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G, Jain A 2019 *Nature* **571** 95

[8] Vaucher A C, Zipoli F, Geluykens J, Nair V H, Schwaller P, Laino T 2020 *Nat. Commun.* **11** 3601

[9] Nie Z, Zheng S, Liu Y, Chen Z, Li S, Lei K, Pan F 2022 *Adv. Funct. Mater.* **32** 2201437

[10] Wang W R, Jiang X, Tian S H, Liu P, Dang D P, Su Y J, Lookman T, Xie J X 2022 *npj Comput. Mater.* **8** 9

[11] Weston L, Tshitoyan V, Dagdelen J, Kononova O, Trewartha A, Persson K A, Ceder G, Jain A 2019 *J. Chem. Inf. Model.* **59** 3692

[12] Friedrich A, Adel H, Tomazic F, Hingerl J, Benteau R, Maruscyk A, Lange L 2020 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* Seattle, Washington, July 5–10, 2020 p1255

[13] He T, Sun W, Huo H, Kononova O, Rong Z, Tshitoyan V, Botari T, Ceder G 2020 *Chem. Mater.* **32** 7861

[14] Beal M S, Hayden B E, Le Gall T, Lee C E, Lu X, Mirsaneh M, Mormiche C, Pasero D, Smith D C, Weld A, Yada C, Yokoishi S 2011 *ACS Comb. Sci.* **13** 375

[15] Rajan A C, Mishra A, Satsangi S, Vaish R, Mizuseki H, Lee K R, Singh A K 2018 *Chem. Mater.* **30** 4031

[16] Liu Y, Zou X X, Yang Z W, Shi S Q 2022 *J. Chin. Ceram. Soc.* **50** 863 (in Chinese) [刘悦, 邹欣欣, 杨正伟, 施思齐 2022 硅酸盐学报 **50** 863]

[17] Zhao K L, Jin X L, Wang Y Z 2021 *J. Software* **32** 349 (in Chinese) [赵凯琳, 靳小龙, 王元卓 2021 软件学报 **32** 349]

[18] Wei J, Zou K 2019 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9 th International Joint Conference on Natural Language Processing* Hong Kong, China, November 3–7, 2019 p6382

[19] Morris J X, Lifland E, Yoo J Y, Grigsby J, Jin D, Qi Y 2020 *Proceedings of the 2020 EMNLP (Systems Demonstrations)* Punta Cana, Dominican Republic, November 16–20, 2020 p119

[20] Malandrakis N, Shen M, Goyal A, Gao S, Sethi A, Metallinou A 2019 *Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019)* Hong Kong, China, November 4, 2019 p90

[21] Wu X, Lü S W, Zang L J, Han J Z, Hu S L 2019 *Computational Science–ICCS 2019* (Cham: Springer Nature Switzerland AG) p84

[22] Kumar V, Choudhary A, Cho E 2021 arXiv: 2003.02245 [cs. CL]

[23] Xu X, Lei Y, Li Z 2020 *IEEE Trans. Ind. Electron.* **67** 2326

[24] Shinyama Yhttps://euske.github.io/pdfminer/ [2022-11-20]

[25] Jessop D M, Adams S E, Willighagen E L, Hawizy L,

Murray-Rust P 2011 *J. Cheminf.* **3** 41
[26] Hawizy L, Jessop D M, Adams N, Murray-Rust P 2011 *J. Cheminf.* **3** 17
[27] Swain M C, Cole J M 2016 *J. Chem. Inf. Model.* **56** 1894
[28] Sun C C 2009 *J. Pharm. Sci.* **98** 1671
[29] Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D 1999 *Natural Language Processing Using Very Large Corpora* (Berlin: Springer) pp157–176
[30] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V 2019 arXiv: 1907.11692 [cs. CL]
[31] Chen S, Wu C, Shen L, Zhu C, Huang Y, Xi K, Maier J, Yu Y 2017 *Adv. Mater.* **29** 1700431
[32] Xiao R J, Li H, Chen L Q 2018 *Acta Phys. Sin.* **67** 128801 (in Chinese) [肖睿娟, 李泓, 陈立泉 2018 物理学报 **67** 128801]
[33] Liu Y, Ge X Y, Yang Z W, Sun S Y, Liu D H, Avdeev M, Shi S Q 2022 *J. Power Sources* **545** 231946

DATA PAPERS

A high-quality dataset construction method for text mining in materials science*

Liu Yue [1)4)] Liu Da-Hui [1)] Ge Xian-Yuan [1)] Yang Zheng-Wei [1)]
Ma Shu-Chang [1)] Zou Zhe-Yi [5)] Shi Si-Qi [2)3)†]

1) (*School of Computer Engineering and Science, Shanghai University, Shanghai* 200444, *China*)
2) (*School of Materials Science and Engineering, Shanghai University, Shanghai* 200444, *China*)
3) (*Materials Genome Institute, Shanghai University, Shanghai* 200444, *China*)
4) (*Shanghai Engineering Research Center of Intelligent Computing System, Shanghai* 200444, *China*)
5) (*School of Materials Science and Engineering, Xiangtan University, Xiangtan* 411105, *China*)

(Received 5 December 2022; revised manuscript received 7 February 2023)

Abstract

Numerous data and knowledge generated and stored as text in peer-reviewed scientific literature are important for materials research and development. Although text mining can automatically explore this information, the barriers of acquiring high-quality textual data prevent its general application in materials science. Herein, we systematically analyze the issues of textual DATA QUALITY and related research from the perspectives of data quality and quantity. Following this, we propose a pipeline to construct high-quality datasets for text mining in materials science. In this pipeline, we utilize the traceable automatic acquisition scheme of literature to ensure the traceability of textual data. Then, a data processing method driven by downstream tasks is used to generate high-quality pre-annotated corpora conditioned on the characteristics of material texts. On this basis, we define a general annotation scheme derived from materials science tetrahedron to complete high-quality annotation. Finally, a conditional data augmentation model incorporating material domain knowledge (cDA-DK) is constructed to augment the data quantity. Experimental results on datasets with various material systems demonstrate that our method can effectively improve the accuracy of downstream models and the F1-score towards the named entity recognition task in NASICON-type solid electrolyte material reaches 84%. This study provides an important insight into the general application of text mining in materials science, and is expected to advance the material design and discovery driven by data and knowledge bidirectionally.

Keywords: text mining in materials science, data augmentation, data quality

PACS: 07.05.Hd, 07.05.Mh, 88.80.ff, 82.47.Jk **DOI:** 10.7498/aps.72.20222316

* Project supported by the National Key Research and Development Program of China (Grant No. 2021YFB3802101), and the National Natural Science Foundation of China (Grant Nos. 92270124, 52073169, 52102313).

† Corresponding author. E-mail: sqshi@shu.edu.cn