

中图分类号: TP181

单位代号: 10280

密 级:

学 号: 19721563

上海大学



硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题 目	基于实体关系抽取的材料科学 文本挖掘方法及其应用研究
-----	-------------------------------

作 者 葛献远

学科专业 计算机应用技术

导 师 刘悦

完成日期 2022 年 12 月

姓 名：葛献远

学号：19721563

论文题目：基于实体关系抽取的材料科学文本挖掘方法及其应用研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：

导 师：

答辩日期：

姓 名：葛献远

学号：19721563

论文题目：基于实体关系抽取的材料科学文本挖掘方法及其应用研究

原创性声明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名： 葛献远 日期： 2022.12.11

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密的论文在解密后应遵守此规定）

签 名： 葛献远 导师签名： _____ 日期： 2022.12.11

上海大学工学硕士学位论文

基于实体关系抽取的材料科学 文本挖掘方法及其应用研究

姓 名： 葛献远

导 师： 刘悦

学科专业： 计算机应用技术

上海大学计算机工程与科学学院

2022 年 12 月

A Dissertation Submitted to Shanghai University for the Degree
of Master in Engineering

**Research on Entity and Relation
Extraction based Materials Science
Text Mining Method and Its
Application**

MA Candidate: Xianyuan Ge

Supervisor: Yue Liu

Major: Computer Application Technology

School of Computer Engineering and Science

Shanghai University

December, 2022

摘要

材料科学文献中蕴含着大量有价值的信息。这些信息不仅可以实现追踪材料研究动态，还可以辅助建立材料数据库与指导材料合成。随着已发表科研文献数量呈指数方式增长，仅依靠专家的手工获取方式已经跟不上材料研发的速度。如何从材料大规模文献中抽取出数据与知识以指导材料性能优化仍是亟待解决的问题。

文本挖掘可以从非结构化的文本信息中抽取潜在的、用户感兴趣的重要模式或知识。目前，自然语言处理任务中命名实体识别和关系抽取技术已经成为常用的文本挖掘方法，且在材料领域取得初步进展。但是材料领域文本挖掘面临高质量有监督文本挖掘数据集标注难、材料特殊文本语义特征难以被已有的命名实体识别模型充分融合、材料目标实体及其边界语义信息难以被已有的关系抽取模型感知等问题。针对上述问题，本文分别从高质量有监督材料科学文本挖掘数据集的构建、材料实体识别和关系抽取三个方面研究了基于实体关系抽取的材料科学文本挖掘方法，并应用于 NASICON 型固态电解质材料文献挖掘。

具体研究内容和创新点如下：

(1) 针对材料领域有监督文本挖掘数据集标注难的问题，本文提出了基于数据增强的有监督材料科学文本挖掘数据集构建方法。该方法包括可溯源的文献自动获取、下游任务驱动的文献预处理、材料实体/关系数据标注以及融合材料领域知识的有条件文本数据增强（Conditional Data Augmentation Model Incorporating Materials Domain Knowledge, cDA-DK）。其中，cDA-DK 通过对融合材料领域知识的预训练 DistilRoBERTa 语言模型进行微调，使得其能感知材料领域的特殊性并学习到复杂的上下文语义信息，从而可以实现在有限手工标注可溯源且高质量样本的基础上自动生成文本数据。在 NASICON 型固态电解质和无机材料实体识别数据集的对比实验，证明了 cDA-DK 能有效生成高质量的文本数据。最后，在增强前后的 NASICON 型固态电解质实体识别数据集上分别训练实体识别模型，增强后的模型精确率、召回率和 F1 分别提高了 5%、3% 和 4%。

(2) 针对材料特殊文本语义特征难以被已有的命名实体识别模型充分融合

的问题，本文提出了基于多层语义特征融合的材料命名实体识别方法 MatBERT-BiLSTM-CRF。该方法首先通过构建 MatBERT 模型编码词、位置以及句子嵌入信息来充分提取词级别的语义特征；其次，引入 BiLSTM 模型对句子序列进行建模以捕获词的局部上下文语义特征；再次，利用序列标注分类器 CRF 对单词标签进行预测以获取最优的标签序列来实现材料实体的识别。最后，在 NASICON 型固态电解质和无机材料实体识别数据集上的对比实验结果表明，我们的模型相较于 BiLSTM-CRF、BiLSTM-CNNs-CRF、BERT 模型，F1 性能指标分别提升了 18%、16% 和 9%。此外，应用 MatBERT-BiLSTM-CRF 模型从 1808 篇 NASICON 型固态电解质文献中抽取了 106896 个材料实体；进而提出了基于重要度计算的描述符筛选策略，成功筛选出 408 个激活能相关的候选描述符；在此基础上，利用数据驱动的机器学习进行了激活能预测，模型的 R^2 性能指标达到了 95%。

(3) 针对材料目标实体及其边界语义信息难以被已有的关系抽取模型感知的问题，本文提出了基于实体感知的材料关系抽取方法 MatBERT-BiGRU-Softmax。该方法首先用特殊的封闭标记包裹目标实体词，使得 MatBERT 模型可以充分感知并提取更丰富的目标实体及句子的语义信息；其次，引入 BiGRU 模型对句子序列进行建模以捕获目标实体的局部上下文语义信息；再次，利用 Softmax 函数计算得到候选关系中概率最大的一个来实现材料关系的分类。最后，在 NASICON 型固态电解质和 MatSciRE 材料关系抽取数据集上的对比实验结果表明，我们的模型相较于 WV+CNN+ATT、WV+BiLSTM+ATT、R-BERT 模型，F1 性能指标分别提升了 16%、9% 和 2%。此外，应用 MatBERT-BiGRU-Softmax 模型抽取了 260475 个 NASICON 型固态电解质材料实体关系三元组；进而构建了材料知识图谱和描述符树，成功获取了 24 条 NASICON 型固态电解质构效关系知识；藉此，为领域知识嵌入的特征选择方法提供知识，并进行了 NASICON 型固态电解质激活能预测，在两份数据集上模型的 R^2 性能指标较未嵌入知识的模型提高了 1.4% 和 1.5%。

关键词：文本挖掘；材料科学；命名实体识别；关系抽取

ABSTRACT

There is a wealth of valuable information in materials science literature. This not only enables the tracking of materials research, but aids in the creation of materials databases and guides materials synthesis. With the exponential growth of published scientific literature, manual acquisition fails to meet speed of materials development. Therefore, how to extract data and knowledge from numerous literature becomes urgent, to guide the performance optimization of materials.

The potentially important patterns or knowledge of the interest in users can be extracted from unstructured textual information. Currently, named entity recognition and relation extraction techniques in natural language processing tasks have become popular text mining methods and been used in materials field successfully. However, text mining in field of materials fails to annotate high-quality supervised text mining datasets. Existing named entity recognition models is hard to fully integrate material-specific text semantic features. And existing relation extraction models is hard to perceive materials target entities and their boundary semantic information. To address the above problems, this study focuses on the materials science text mining methods based on entity and relation extraction, which is applied to NASICON-type solid electrolyte materials literature mining. Then, the studies are performed from the aspects of the construction of high-quality supervised text mining datasets for materials science, materials entity recognition and relation extraction. The effectiveness of the above methods is verified with NASICON-type solid electrolyte materials literature mining.

The main research contents and innovations of this study are as follows:

(1) Aiming at the difficulty of annotating supervised text mining datasets for materials science, this study proposes a construction method for supervised materials text mining datasets based on data augmentation. This method includes automatic acquisition of traceable documents, document preprocessing of downstream task-driven, data annotation for materials entity/relation, and conditional text data augmentation

incorporating materials domain knowledge (cDA-DK). Thereinto, the cDA-DK is fine-tuned with the pre-trained DistilRoBERTa model incorporating materials domain knowledge so that can perceive specificities for materials domain and learn contextual semantic information of complex. This enables automatic text data generation based on limited manual annotation of traceable and high-quality samples. The experimental results on NASICON-type solid electrolyte and inorganic materials entity recognition datasets demonstrate that the cDA-DK model is effective in generating high-quality text data. The performance of proposed model with augmented dataset surpasses that of the models with original dataset, where the precision, recall and F1 are improved by 5%, 3% and 4%, respectively.

(2) Existing named entity recognition models are hard to fully integrate the semantic features of material-specific text. Hence, this study proposes a multi-layer semantic feature fusion method MatBERT-BiLSTM-CRF for materials named entity recognition. This method constructs MatBERT to encode word, position and sentence embedding information, to fully extract word-level semantic features by constructing. Then, BiLSTM is introduced to model the sentence sequences so that capture local contextual semantic features for words. To obtain the optimal label sequences for materials entity recognition, the conditional random field (CRF) model is used. Finally, comparative experimental results on NASICON-type solid electrolyte and inorganic materials entity recognition datasets show that the proposed model gain 18%, 16%, and 9% improvement of the F1-Score compared to BiLSTM-CRF, BiLSTM-CNNs-CRF and BERT models, respectively. Furthermore, through MatBERT-BiLSTM-CRF model, 106,896 materials entities are extracted from 1808 NASICON-type solid electrolyte literature. Then, a descriptor screening strategy based on importance-calculated is proposed and 408 candidate descriptors related to activation energy are successfully screened. Following this, machine learning model is used for activation energy prediction, which achieves performance (R^2) of 95%.

(3) Existing relation extraction models fail to perceive materials target entities and

their boundary semantic information. To this end, this study proposes an entity-perceived materials relation extraction method, MatBERT-BiGRU-Softmax. This method wraps the target entity words with special closed tokens, so that the MatBERT can fully perceive the target entities. Then, richer semantic information for target entities and sentences can be extract. To capture the local contextual semantic information around the target entities, the BiGRU is introduced to model the sentence sequences. The Softmax function is used to calculate the highest probability of the candidate relations to achieve the classification of materials relation. The experimental results on NASICON and MatSciRE materials relation extraction datasets show that the proposed model improves the F1-Score by 16%, 9%, and 2% compared to WV+CNN+ATT, WV+BiLSTM+ATT, and R-BERT models, respectively. Through MatBERT-BiGRU-Softmax, 260475 NASICON-type solid electrolyte materials entity-relation triples are extracted. Then, the materials knowledge graph construction and descriptor tree are constructed. Following this, 24 structure-activity relationships knowledge for NASICON-type solid electrolyte materials are successfully obtained. On this basis, the above knowledge is provided for the feature selection method, the machine learning model gained 1.4% and 1.5% performance (R^2) improvement for the prediction of activation energy of NASICON-type solid electrolyte on two datasets compared with data-driven model.

Keywords: Text Mining; Materials Science; Named Entity Recognition; Relation Extraction

目 录

摘要	V
ABSTRACT	VII
目录	X
第一章 绪论	1
1.1 课题来源	1
1.2 课题研究的背景和意义	1
1.3 材料科学文本挖掘国内外研究现状	3
1.3.1 命名实体识别及其在材料科学文本挖掘中的应用	3
1.3.2 关系抽取及其在材料科学文本挖掘中的应用	7
1.4 论文的主要研究内容	12
1.5 论文的内容安排	14
第二章 基于数据增强的有监督材料科学文本挖掘数据集构建方法	16
2.1 问题描述与分析	16
2.2 材料科学文本挖掘数据集构建	19
2.2.1 方法概述	19
2.2.2 可溯源的文献自动获取	20
2.2.3 下游任务驱动的文献预处理	23
2.2.4 材料实体/关系标注	25
2.2.5 融合材料领域知识的文本数据增强	30
2.3 实验	33
2.3.1 实验数据	33
2.3.2 实验设置	34
2.3.3 实验结果与分析	34
2.4 应用	36
2.4.1 NASICON 型固态电解质有监督文本挖掘数据集的构建	37
2.4.2 NASICON 型固态电解质有监督文本挖掘数据集的扩充	40

2.5 小结	41
第三章 基于多层语义特征融合的材料命名实体识别方法	42
3.1 问题描述与分析	42
3.2 方法概述	44
3.3 基于多层语义特征融合的材料命名实体识别	44
3.3.1 基于 MatBERT 的多级别语义特征的融合	45
3.3.2 基于 BiLSTM 的局部上下文语义特征的融合	46
3.3.3 基于 CRF 的实体分类	48
3.4 基于材料命名实体的描述符筛选	49
3.4.1 基于名词库的材料实体存储	49
3.4.2 基于重要度计算的描述符筛选	50
3.5 实验	51
3.5.1 实验数据	51
3.5.2 实验设置	52
3.5.3 评价指标	52
3.5.4 实验结果与分析	53
3.6 应用	57
3.6.1 NASICON 型固态电解质材料实体的存储	57
3.6.2 NASICON 型固态电解质激活能相关描述符的筛选	57
3.7 小结	61
第四章 基于实体感知的材料关系抽取方法	62
4.1 问题描述与分析	62
4.2 方法概述	63
4.3 基于实体感知的材料关系抽取	64
4.3.1 目标实体驱动的实体感知	65
4.3.2 基于实体感知的语义特征提取	66
4.3.3 基于 <i>Softmax</i> 的材料实体关系分类	68
4.4 基于实体关系的材料知识图谱构建与知识获取	69

4.4.1	基于 Neo4j 图数据库的材料知识图谱构建.....	69
4.4.2	基于材料知识图谱的描述符树建立.....	70
4.4.3	基于描述符树的知识获取.....	72
4.5	实验.....	74
4.5.1	实验数据.....	74
4.5.2	实验设置.....	75
4.5.3	评价指标.....	76
4.5.4	实验结果与分析.....	76
4.6	应用.....	79
4.6.1	NASICON 型固态电解质材料的知识图谱构建.....	79
4.6.2	NASICON 型固态电解质材料的描述符树建立.....	81
4.6.3	NASICON 型固态电解质材料的知识获取.....	85
4.7	小结.....	95
第五章	结论与展望.....	96
5.1	本文主要工作	96
5.2	展望	98
	参考文献.....	99
	作者在攻读硕士学位期间公开发表的论文.....	111
	作者在攻读硕士学位期间所作的项目.....	112
	致 谢.....	113

第一章 绪论

1.1 课题来源

本课题来源于国家自然科学基金面上项目“领域知识嵌入的机器学习方法研究镍基单晶高温合金蠕变构效关系”（项目编号：52073169）和国家重点研发计划项目“数据驱动的新型高性能功能材料智能化研发与应用”子课题“功能材料数据质量提升及专用数据库建设”（项目编号：2021YFB3802101），试图以材料科研文献中的文本为研究对象，依据材料四面体准则，通过研究材料命名实体识别及关系抽取文本挖掘方法来自动获取文本中材料实体及其关系等知识，以指导材料性能预测和新型材料的研发进程，并以 NASICON 型固态电解质材料为例进行探索。

1.2 课题研究的背景和意义

长期以来，材料领域已经在经验、理论和计算科学的研究范式中积累了丰富的领域知识。如今，材料领域迎来了数据驱动科学的研究范式^[1]。目前的工作更多地使用数据驱动的机器学习方法来辅助材料科学研究^[2-9]，旨在从历史数据中挖掘出有价值的信息以缩短材料科学的研究周期。然而，受限于样本数据的质量，纯数据驱动的机器学习会获得与领域知识不一致的结果且通常不可解释，如果能将领域知识融入此类模型中，将有利于解决上述问题。因此，材料领域知识对于机器学习的指导至关重要。随着材料数据驱动机器学习的发展，大量的领域知识以非结构化文本的形式存储于文献、专利或报告中。如何快速准确的从文本中挖掘出有价值的信息已经成为材料领域的研究热点之一。

文本挖掘(Text Mining, TM)^[10]因其能够从文本中自动提取出有价值的信息，已经成功应用到材料化学和生物医学^[11-14]等领域。材料科学文本挖掘旨在从非结构化的文本中提取有价值的材料信息和知识，由此可以追踪材料研究动态^[15-17]、指导材料合成^[18]和建立材料数据库^[19, 20]等，其通用工作流程如图 1.1 所示。该流

程包括文本收集与解析、文本预处理、文本分析、信息提取、数据挖掘。

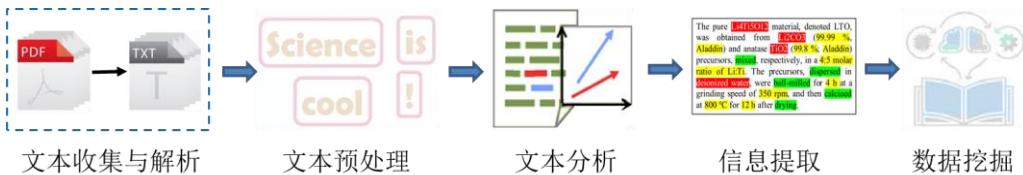


图 1.1 材料科学文本挖掘工作流程

自然语言处理 (Natural Language Processing, NLP) 是计算机领域的一个热点研究方向，提供了大量的文本挖掘方法来从非结构化文本中提取有价值的信息。目前，材料科学文本挖掘主要采用 NLP 任务中的方法从数量庞大且不断增长的科学出版物中快速获取领域知识，进而指导材料相关领域的研究。其中，命名实体识别 (Named Entity Recognition, NER)^[21] 及关系抽取 (Relation Extraction, RE)^[22] 作为 NLP 信息抽取的核心子任务，其相关技术已经被成功应用于材料领域，如有机^[23]和无机材料信息^[19, 24-32]的挖掘，但依然处于起步阶段。NER 擅长从非结构化的文本中抽取关键字及短语等信息，通常是有监督的机器学习分类问题，即模型通过训练学习能自动从文本中识别并抽取相应的实体信息。RE 则擅长对实体间的关系进行分类，目前比较流行的是有监督的 RE 方法，即模型通过对监督数据的训练学习实现对实体间关系的自动分类。例如，在句子 “The introduction of sodium in the cages induces only a small variation of the volume of the unit cell compared to that of the empty compound.” 中，“sodium” 和 “volume” 分别是实体 “成分” 和 “属性” 类实体，二者之间存在 “Cause-Effect”（影响）的关系，即原子会影响其体积的变化。“sodium” 和 “volume” 等实体信息可由 NER 模型识别并抽取。在此基础上，两者之间的关系则可由 RE 模型以实体关系三元组的形式 (sodium, Cause-Effect, volume) 分析并挖掘。

本文旨在开发适用于材料科学文本挖掘的 NER 和 RE 方法来实现对材料知识信息的自动提取，从而达到辅助材料设计的目的。然而，材料领域 NER 和 RE 高质量标注数据集目前依然稀缺。手工构建不仅需要经历繁杂的过程，还会耗费大量的人力和物力。此外，相关研究表明不同领域的文本结构及语言表述具有较大的差异^[33]，由于材料文本复杂且多样，导致设计材料 NER 和 RE 方法面临巨大的挑战。具体来说，需要对材料文本句子结构、句子中实体及实体间的关系进

行深度的语义和语法分析,以构建适用于材料知识信息获取的 NER 和 RE 模型,从而实现相关材料信息的大规模自动提取。

综上所述,构建 NER 及 RE 高质量监督数据集,基于此设计性能好、效率高的 NER 及 RE 模型,并以此实现自动挖掘材料信息以探索材料设计应用是本文研究的主要科学问题,有望实现从文献中大规模抽取材料知识信息,为材料性能预测的机器学习提供数据及先验知识支撑,以加速新材料的探索和研发过程。

1.3 材料科学文本挖掘国内外研究现状

尽管文本挖掘技术已经被成功应用到材料领域,但目前仍然处于起步阶段。现有的材料科学文本挖掘工作大都基于 NLP 任务中命名实体识别和关系抽取方法展开研究。命名实体识别方法旨在从文本中大规模挖掘材料关键字或短语等信息,不仅可以对相关研究热点进行快速分析,还能构建实体语料库或本体库来辅助材料设计的研究。在此基础上,关系抽取方法能够为独立的材料实体建立关联,以获得更复杂的构效关系等知识信息,从而进一步实现辅助材料设计的研究。

1.3.1 命名实体识别及其在材料科学文本挖掘中的应用

命名实体识别技术主要分为三类:早期以基于规则和字典匹配的方法为主。随着机器学习的发展,基于统计机器学习的方法被广泛研究和应用。最近,基于深度学习的方法将其推向高潮。材料领域的命名实体识别旨在识别和分类材料文本中提及的概念来判别具有语义价值的对象。这些对象可以为研究人员映射到属性、找到相似的化合物或纳入注释标记提供帮助。随着命名实体识别技术的发展,材料领域的学者逐渐注意到使用深度学习方法进行命名实体识别的研究优势。

(1) 基于规则匹配和字典的命名实体识别方法

基于规则匹配的方法需要研究者通过语言学家手工构造规则模板从文本中抽取出相关实体。基于字典匹配的方法则需要搜集并建立命名实体词典库,然后通过词典库从文本中匹配以进行实体的抽取。例如, Rau 等人^[34]首次通过设计人工规则并结合启发式搜索思想从财经类新闻文本中自动抽取公司名称类型的命

名实体，准确率达到 95%。Grishma 等人^[35]则通过搜集包括国家、城市及公司的字典，设计了一种基于规则匹配的实体识别系统，该系统可以自动匹配文本中的相关命名实体。Collines 等人^[36]提出了一种基于决策列表来得到更多命名实体规则的方法，从而将得到的规则集合用于从文本中识别命名实体，其核心思想是设置一个种子规则集，然后对该种子进行无监督的训练迭代以生成更多的规则。Swain 等人^[19]搜集了大量已发表文献，并用传统实体识别的技术开发了一个自动提取化学信息的 NLP 工具包 ChemDataExtractor，进而可以从中自动抽取材料性能信息。

基于规则及字典的命名实体识别方法的优点是可以在小数据集上达到一个很高的识别准确率。然而，该方法不适用于处理大规模数据集或全新领域，通常需要重新设置规则或收集新的字典，需要花费大量的人力、物力及财力。此外，不同语言的规则设置也存在差异，因此逐渐被基于统计机器学习方法取代。

（2）基于统计机器学习的命名实体识别方法

随着机器学习的发展，计算自然语言学习会议（Conference On Computation Natural Language Learning, CONLL）渴望找到一种释放人工的、独立于语言的命名实体识别方法，并将机器学习方法解决命名实体识别问题引入大众视野。基于统计的机器学习方法将实体识别任务作为序列标注问题，通过人工标注一定的语料对模型进行有监督训练，从而使模型可以学习到实体及标签的依赖关系，进而可以有效地对实体类别进行预测。与传统分类任务相比，序列标注问题认为当前预测的标签不仅与当前的输入特征有关，还会受到之前预测标签的影响，即预测标签之间是有强相互依赖关系的。

基于机器学习的实体识别方法主要有：隐马尔可夫模型（Hidden Markov Model, HMM）、最大熵（Maximum Entropy, ME）^[37]、最大熵马尔可夫模型（Maximum Entropy Markov Model, HEMM）^[38]、支持向量机（Support Vector Machine, SVM）以及条件随机场（Conditional Random Fields, CRF）^[39]。Bikel 等人^[40]提出了一个基于隐马尔可夫模型的实体识别系统用于自动抽取文本中的名称、日期、时间和数值等实体，并在英语和西班牙等语言上取得了不错的效果。Chi 等人^[41]则利用最大熵模型抽取文本的局部上下文特征同时结合其全局特征

进行实体识别，有效提高了实体识别的分类准确率。Lin 等人^[42]在基于词典和规则方法的基础上，结合最大熵来对生物文本中的实体进行自动识别。他们首先根据已有的生物经验知识设计一系列规则，然后将这些规则融入到最大熵模型中来进行实体识别，大大提高了模型的准确率和召回率。Yamada 等人^[43]则提出了一个基于 SVM 的命名实体识别系统，他们的系统通过训练日语实体识别数据集，在 CONLL-2000 会议上取得了最好的效果。Leaman 等人^[25]开发了一个材料化学命名实体识别器 tmChem 系统，其是通过将两个独立的机器学习模型结合在一起而构建的，该系统在 CHEMDNER 任务中发布的数据集上训练，F1 值达到了 87.4%。

基于统计机器学习的方法无需进行规则或字典的设计及收集，在一定程度上减少了人工的协同。然而，有监督的实体识别工作依然需要专家对文本进行标注，同时机器学习模型尚未考虑句子序列上下文的语义信息。

(3) 基于深度神经网络的命名实体识别方法

近年来，基于深度神经网络（Deep Neural Network, DNN）的方法将命名实体识别的研究推向一个新的高潮，目前绝大多数先进的命名实体识别方法均以 DNN 为基础。深度学习方法由神经元之间的传播来自动提取特征，避免了传统机器学习需要手工构建特征模板的缺点。此外，神经网络引入词向量来表示单词或者句子，不仅可以解决独热编码（One-Hot Encoding）数据稀疏及高维度的问题，而且词向量中还包含了更多的单词及句子语义信息，使得模型实现更精准的实体识别。

Collobert 等人^[44]率先将基于神经网络的方法用于命名实体识别任务的研究，他们设计了基于窗口和句子的 NER 方法。其中前者以当前词的上下文窗口作为输入，然后通过神经网络传播；后者则以整个句子作为当前预测词的输入，同时还加入了句子中每个词的位置特征，然后利用卷积神经网络（Convolutional Neural Network, CNN）提取局部特征从而构造全局特征向量，最后进行实体分类。通过神经网络来提取词向量特征有效地提高了实体分类的准确率，但该方法没有考虑远距离单词间的依赖信息。Mikolov 等人^[45]提出循环神经网络（Recurrent Neural Network, RNN）语言模型来解决序列问题并将其用在语音识别任务中，大幅提升

了识别精度。

上述深度神经网络中，CNN 仅能提取文本序列的局部特征，且需要通过堆叠多层的卷积来增大感受野，无法较好地捕获文本语义信息。RNN 虽然能够提取全局特征，但其是通过逐步递归进行提取的，无法充分提取句子中单词的全局特征。因此，Huang 等人^[46]详细介绍了几种基于长短期记忆网络（Long Short-Term Memory, LSTM）的序列标注模型，并将其用于命名实体识别建模任务，包括 LSTM、Bi-LSTM（Bidirectional-LSTM）、LSTM-CRF、LSTM-CRF 及 Bi-LSTM-CRF。其中，基于 Bi-LSTM-CRF 的网络结构在序列标注任务如词性标注、分块和命名实体识别三个基本任务上的效果都要优于其它模型，且该模型对词嵌入的依赖也较少。需要注意的是，相比于 RNN 当前时刻的输入仅依赖前一时刻的输出而言，LSTM 增加了三个门控单元，即输入门、输出门及遗忘门，可以决定某一时刻要保留或遗忘的信息，最大限度获取语义特征的同时可以避免梯度消失和梯度爆炸等问题。基于此，Weston 等人^[31]将 BiLSTM-CRF 模型应用到材料领域，设计了 7 个材料实体类别标签，手工标注了 800 个材料文献摘要，通过对 BiLSTM-CRF 模型进行训练，从大规模的摘要中抽取了 8000 多万个材料命名实体。He 等人^[32]则从材料合成信息出发，利用 BiLSTM-CRF 模型设计了一个两步的化学实体识别系统，分别从文献中抽取化学合成的前体和目标，为材料化学的合成提供了新思路。然而，材料文本数据冗长、结构复杂，领域词汇普遍存在特殊性、多义性和指代不清等问题，上述材料 NER 方法不能充分捕获单词的语义信息及其与实体标签的依赖关系，从而影响模型分类的准确率。

最近，Vaswani 等人^[47]提出了 Transformer 模型，如图 1.2 所示，该模型采用编码-解码（encode-decode）结构搭建了一个神经网络序列模型框架。Transformer 模型没有考虑递归和卷积的操作，而是完全由注意力机制（Attention）进行编码和解码的操作。其中编码器由 6 个块组成，每个块均由自注意力机制和前馈神经网络组成，解码器也同样由 6 个块组成，每个块由自注意力机制、编码器-解码器（encoder-decoder）和前馈神经网络组成。Transformer 通过多次参数矩阵映射来进行 Attention 的操作，并将其结果拼接起来便可以快速地获取全局语义特征。Yan 等人^[48]通过对 Transformer 模型进行改进，使其可以捕获方向信息及单词的

相对位置特征，提出了 TENER（Transformer Encoder for NER）模型来进行命名实体识别的研究。该模型在 Onto Notes5.0 和 MSRA 实体识别数据集上 F1 值分别达到了 88.43% 和 92.74%，刷新了数据集的最好成绩。2018 年，Google 的 Devlin 等人^[49]基于 Transformer 编码器提出了预训练语言模型 BERT（Bidirectional Encoder Representations from Transformers），刷新了自然语言处理中包括 NER 在内的 11 项任务的最好成绩。因此，越来越多的学者在预训练 BERT 模型的基础上进行命名实体识别的研究。

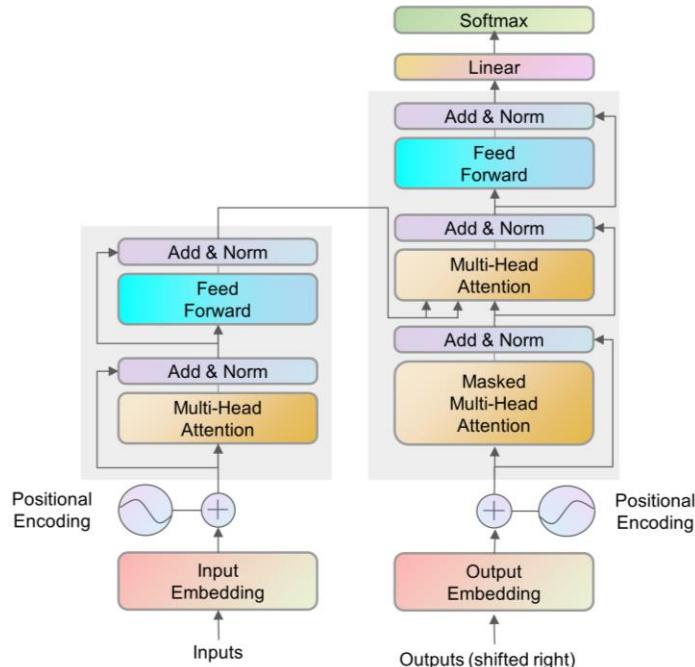


图 1.2 Transformers 模型图

综上所述，材料领域命名实体识别依然处于起步阶段。由于材料领域文本复杂且存在许多一词多义及代词指代现象，现有的实体识别模型无法充分捕获单词的语义信息及其与实体标签的依赖关系，从而影响模型分类的准确率。因此，如何提高实体识别模型捕获材料语义信息的能力，从而打破材料实体识别模型的精度瓶颈成为亟待解决的问题。

1.3.2 关系抽取及其在材料科学文本挖掘中的应用

关系抽取作为信息抽取任务的一个重要环节，已经被业内学者广泛研究。国内的哈工大信息检索实验室和清华大学自然语言处理团队在关系抽取的研究上

取得了许多显著的成果，并公布了多个关系抽取数据集。国外的谷歌自然语言处理团队和 OpenAI 研究所在大规模预训练语言模型上取得了重大的突破，刷新了包括关系抽取在内的多个自然语言处理任务的最好成绩。根据关系抽取对监督数据的依赖程度，该任务分为有监督关系抽取和远程监督关系抽取。材料领域关系抽取工作旨在标识和分类给定材料文本中提到的实体对之间关系，不仅可以将实体与其属性相关联，还可以建立与其它实体的共现关系，从而实现材料文献中领域知识的自动获取。然而，该领域关系抽取工作进展缓慢，研究者往往从所研究的材料对象出发，利用传统关系抽取方法来从材料文本中获得领域知识。

（1）有监督关系抽取方法

有监督关系抽取方法需要标注一定数量的训练语料，然后基于特征工程构建模型提取目标实体及句子的语义信息，同时训练模型使其学习目标实体与其对应关系的依赖，从而实现关系精准分类。此方法可以分为基于规则、基于特征向量、基于核函数及基于深度学习的方法。

基于规则的方法利用语言学和自然语言处理学等相关知识，分析对应文本、总结出相应的关系规则并利用规则在非结构化文本中匹配，从而提取实体关系三元组。Miller 等人^[50]通过语法解析器对已有实体进行分析并生成规则，然后基于匹配的方式从文本中抽取相应的实体关系三元组。邓擘等人^[51]则提出了一种中文的实体关系抽取技术，他们利用基于规则的模式匹配结合词汇语义匹配技术对中文文本中的实体关系进行抽取，实验证明该方法要优于单独使用规则匹配抽取的方法。Kuniyoshi 等人^[52]将规则的方法应用到材料领域，它们为了提取隐藏在科学文献中的无机材料的合成过程，由基于深度学习的序列标记器和一个简单的基于启发式的规则关系提取器开发了一个自动机器读取系统，并构造了一个固态电池合成过程的语料库，通过对该系统的训练可以实现自动从文献中提取材料合成过程。基于规则的关系抽取方法具有较多局限性，比如严重依赖语言学家制定的规则、不同的领域或不同的语言则需要重新制定规则等，使得该方法具有较差的可移植，同时耗费大量人力物力。

基于特征向量的方法通过训练模型提取句子中的词汇、句法及语义等信息来构造特征向量，然后基于特征向量的相似度选择关系分类器来进行关系的分类，

最后实现从非结构化文本中抽取实体关系三元组。Kambhatla 等人^[53]构建了一个最大熵模型进行关系分类。他们通过结合词汇、句法及语义特征等信息来构建特征向量以训练模型，实验证明该方法可以很好的对关系进行分类。甘丽新等人^[54]则通过融入句子的句法关系及动词特征构建特征向量提出了一种基于句法语义信息的中文关系抽取方法，并选取 SVM 分类器进行关系的分类，并通过实验证明了其方法的有效性。基于特征向量的方法尽管一定程度上提升了关系抽取的性能，但是由于特征的选择需要人工依靠经验来判断，且特征组合方式是有限的，因此很难进一步提升关系分类的效果。

基于核函数的方法能够解决基于特征向量关系抽取方法存在的问题。该方法不需要构造特征向量，取而代之的是将原始输入映射到一个新的特征空间。其可以捕获句子的远距离特征，同时结合特征间的顺序及结构等信息进行关系的分类。Zelenko 等人^[55]率先将核函数的方法应用到关系抽取任务。他们定义并设计了核函数以及其相应算法，同时将核函数与支持向量机结合进行关系的分类，实验证明该方法明显优于基于特征的算法。Zhou 等人^[56]则提出了最短路径包含树核函数的方法进行关系抽取。他们首先通过句法分析对句子进行解析并得到其最短路径包含树，然后与核函数及特征空间结合，最后将其特征结合来进行关系的分类，有效的提升了分类准确率。基于核函数的关系抽取方法以树为基础对象，通过对子树间的相似程度进行关系分类，可以进一步提升关系抽取的性能。然而其速度相对较慢，存在错误传播的问题，同时在大规模数据集上效果并不明显。

目前越来越多学者通过设计深度神经网络模型来进行关系抽取的研究。基于深度学习的关系抽取方法不需要设置规则、不需要预先选择和抽取特征以及不需要设计核算法，其通过神经网络对句子中目标实体词及句子向量进行语义信息的提取，然后由分类层进行关系的预测。

Socher 等人^[57]率先将 RNN 应用到关系抽取任务，他们首先利用句法依存分析将原始文本语料库中的句子转化为树的结构，然后为树的每个节点设置一个向量和矩阵来学习单词及其相邻词的含义，从而可以捕获到单词间的远距离语义信息来进行关系的分类。Zhou 等人^[58]提出了一个 Bi-LSTM 结合注意力机制的模型进行关系抽取，其中 LSTM 可以有效缓解 RNN 在传播过程中梯度消失及梯度爆

炸的问题，并利用注意力机制为目标实体词分配更高的权重，最后得到向量表示以预测关系的类别。Liu 等人^[59]最早将 CNN 应用于关系抽取任务。他们以单词的 One-Hot 编码作为输入，通过 CNN 模型来提取目标实体词及句子的语义特征，并最终利用全连接层结合 *Softmax* 函数对关系进行分类，在 ACE2005 数据集上的 F1 值比基于核函数的方法提高了 9%。Xu 等人^[60]则通过 CNN 提取目标实体最短依存路径的特征，对于目标实体距离较远时依存分析树引入的噪声问题，采用负采样策略来应对，取得了不错的效果。Zhang 等人^[61]结合图卷积神经网络（Graph Convolutional Network, GCN）提取依存句法树中的依存特征来进行关系抽取任务的研究。他们还设计剪枝策略对依存句法树进行操作以过滤掉无关信息，最后在关系分类上取得了不错的效果。

随着预训练 BERT 在大量 NLP 任务上刷新了最好的成绩，越来越多学者利用预训练 BERT 模型对关系抽取任务进行研究。Wu 等人^[62]设计了 R-BERT 关系抽取模型，他们利用预训练 BERT 模型提取单词级别及句子级别的语义信息，并将目标实体及句子语义信息进行拼接，然后通过全连接层及 *Softmax* 函数进行关系分类。在 SemEval-2010 task 8 关系抽取数据集上刷新了最好成绩。Huang 等人设计了^[63]D-BERT 关系抽取模型，他们考虑每个单词和目标实体之间依赖关系的高级句法特征，将其纳入预训练 BERT 语言模型中，并利用 BERT 的中间层来获取不同层次的语义信息，在此基础上对上述特征融合得到多粒度特征用于最终的关系分类，在多个数据集上的效果都有所提升。

近年来，实体关系联合抽取由于能将实体识别及关系抽取任务联合起来执行文本挖掘工作以此减少误差传播，已经被越来越多的学者研究。魏晓等人^[64]将联合抽取任务应用到材料领域，提出基于双向门控循环单元-图神经网络-条件随机场（BiGRU-GNN-CRF）的材料实体关系联合抽取方法和基于改进 TextRank 算法的材料工艺知识抽取方法，实现了从专利、论文等材料文献中自动获取材料实体、关系、工艺流程等领域知识，并在其所构建的非调制特殊钢、铝基复合材料、热障陶瓷涂层材料三个材料领域知识图谱上进行了初步应用探索，验证了知识图谱为材料构效关系研究提供知识支撑的可能性。尽管实体关系联合抽取可以使两个子任务进行交互从而减少误差传播，但是该任务不易于实现且灵活度不够，更重

要的是其无法利用实体信息，且由于材料文本十分复杂，信息抽取工作具有很大的不确定性，因此不利于材料科学文本挖掘的初步探索。

(2) 远程监督关系抽取方法

远程监督的思想率先由 Graven 等人提出。考虑到手工标注数据过于耗费人力物力，因此他们希望通过已有的名词库来自动标注数据以构建生物学名词库。Mintz 等人^[65]率先将远程监督的思想用于关系抽取任务。他们将已有的 Freebase 名词库中包含实体对及其关系的句子作为目标对象，然后收集维基百科中的文本并将其中有相同目标实体对的句子提取出来，最后提取获得句子的特征并训练机器学习分类器对关系进行抽取。

采用远程监督进行关系抽取往往基于一个假设：如果名词库中的两个目标实体具有某种关系，则就认为名词库句子中只要包含该实体对均可能有相同关系类型。然而，无论是何种自然语言，其表达形式无疑具有多样性，相同实体对出现在不同的句子中很有可能具有不同的意思。从上述假设中我们可以得知，远程监督的假设过于肯定，其得到的标注数据也很大程度上存在噪声数据，从而影响关系抽取的效果。因此，远程监督关系抽取存在的最大问题便是噪声问题。Ridel 等人^[66]为了缓解远程监督关系抽取的噪声问题，首次将多示例学习引入远程监督关系抽取，并进一步提出了 at-least-one 的假设：如果两个目标实体已经有某种关系，则名词库所有包含该目标实体对的句子中至少有一个句子可以表示该类关系。他们将具有相同实体对的多个句子看成一个包并为其打上关系类别标签，如果包中至少有一个句子具有该类关系则标注为正例，否则就标注为负例。实验结果证明多示例的远程监督关系抽取一定程度上缓解了噪声问题。基于此，Wang 等人^[31]则将远程监督关系抽取方法应用到材料领域，构造了一个管道模型来自动捕获化学成分和性质等数据。其核心思想是通过搜集文献构建实体数据集及定义关系规则，接着分别设计基于 BiLSTM-CRF 的实体识别模型及基于 Snowball 的远程监督关系抽取方法，通过对模型的训练以实现自动从材料文献中抽取相关数据，并由抽取的数据实现对材料性能的预测。

尽管上述方法一定程度缓解了远程监督关系抽取的噪声问题，但依赖于人为设计特征且未能充分提取句子语义信息。因此，噪声问题仍未彻底解决。随着深

度学习在有监督关系抽取的成功应用，越来越多学者利用深度神经语言模型来获取词及句子级别嵌入向量。此类模型不需要人为设计特征且可以由模型自动捕获单词及句子语义向量，因此一定程度上可以减少远程监督关系抽取的噪声问题。Zeng 等人^[67]率先将深度学习应用到远程监督关系抽取任务中，并提出了分段卷积神经网络（Piecewise Convolutional Neural Network, PCNN）关系抽取模型。该模型在构建词嵌入特征向量时将词向量和相对位置向量结合作为模型的输入，并通过 CNN 来进行句子级别语义的提取；在此基础上，通过目标实体将句子分割为三部分，分别对每部分进行卷积及最大池化操作来提取每部分的局部语义特征；最后由 *Softmax* 函数进行关系分类。需要注意是，该模型是根据 at-least-one 假设^[66]并以多示例学习的方式进行训练的，实验结果表明其在准确率和召回率上得到了很大的提升。为了进一步提取句子的语义信息同时减少监督关系抽取的噪声影响，Jat 等人^[68]通过双向门控单元（Bidirectional Gate Recurrent Unit, BiGRU）来捕获句子中单词的上下文信息，并利用单词级别的注意力机制来获得句子级别的向量表示，然后利用句子级别的注意力机制加权获得包的向量表示，最后接 *Softmax* 函数进行关系分类，进一步提高了远程监督关系抽取的准确率和召回率。

综上所述，材料领域普遍采用规则匹配的方式进行关系抽取的研究，因此依然处于起步阶段。材料文本中的关系十分复杂，一句话中可能存在多种重叠关系，使得现有的材料关系抽取方法未能充分考虑目标实体的语义信息及其对关系分类的影响，也未能充分考虑单词的上下文信息及句子的语义信息，严重影响模型分类的准确性。如何提高关系抽取模型感知目标实体，并充分捕获实体感知的材料文本语义特征，从而提升材料关系抽取模型的分类准确率成为亟待解决的问题。

1.4 论文的主要研究内容

如图 1.3 所示，本文探究从材料科研文献出发到材料实体及关系抽取的全流程，开展基于数据增强的有监督数据集构建方法、基于多层语义特征融合的材料命名实体识别方法和基于实体感知的材料关系抽取方法研究。具体研究内容如下：



图 1.3 研究内容总体框架图

(1) 研究基于数据增强的有监督材料科学文本挖掘数据集构建方法

针对材料高质量有监督文本挖掘数据集标注难的问题，本文拟研究基于数据增强的有监督材料科学文本挖掘数据集构建方法，为材料领域的研究人员更快速地构建文本挖掘数据集提供指导。首先，研究基于可溯源的文献自动获取，将溯源机制嵌入网络爬虫程序中以有效地获取目标文献数据源；其次，研究下游任务驱动的文献预处理，以材料文本特性为约束设计文本预处理方式从而获取预标注干净的材料文本数据；再次，研究基于专家经验指导的数据标注，分析已有的材料科学文本挖掘数据标注场景，在材料专家的指导下设计有监督数据标签并选择合适的工具进行标注；最后，研究融合材料领域知识的有条件文本数据增强，将具备领域特性的材料文本知识融合到预训练语言模型中，使得其充分感知材料文本的复杂语义信息，以动态生成高质量的材料文本数据。

(2) 研究基于多层语义特征融合的材料命名实体识别方法

针对材料特殊文本语义特征难以被已有的命名实体识别模型充分融合的问题，本文拟研究基于多层语义特征融合的材料命名实体识别方法以实现材料实体的准确识别。首先，研究基于多层语义特征融合的材料命名实体识别模型 MatBERT-BiLSTM-CRF 的构建。其中，设计 MatBERT 同时编码词嵌入、位置嵌入及句子嵌入信息，以捕获含丰富材料信息的 token 及句子级别的语义特征并将其融合形成单词的向量表示；引入 BiLSTM 对句子序列建模，以进一步捕获单词的局部上下文语义特征，最终获得 token 级别的语义特征向量；利用 CRF 并基于

token 级别的语义特征向量对单词或短语进行标签预测以获取最优的标签序列。其次，研究 MatBERT-BiLSTM-CRF 模型在材料领域的应用，包括基于名词库的材料实体存储和基于重要度计算的筛选策略。具体地，构建名词库将大规模材料实体信息进行分类存储；设计重要度计算策略用于筛选与特定目标材料性能相关的高质量描述符实体，通过构建简单机器学习模型实现相关材料性能的预测。

(3) 研究基于实体感知的材料关系抽取方法

针对材料目标实体及其边界语义信息难以被已有的关系抽取模型感知的问题，本文拟研究基于实体感知的材料关系抽取方法以实现材料关系的精准分类。首先，研究基于实体感知的材料实体关系抽取 MatBERT-BiGRU-Softmax 的构建，包括目标实体驱动的实体感知、基于实体感知的语义特征提取和基于 *Softmax* 的材料实体关系分类。其中，实体感知设计特殊标记“[]”和“{}”对两个目标实体词进行包裹，使得模型清晰地感知目标实体及其边界信息，以此作为下一阶段的输入属性；语义特征提取由 MatBERT 和 BiGRU 模型组成，MatBERT 用于提取句子级别语义特征和包含句子嵌入、单词嵌入及位置嵌入的单词级别语义特征，BiGRU 用于进一步对句子序列建模以提取句子及目标实体的局部上下文语义特征；利用 *Softmax* 函数计算得到候选关系中概率最大的一个来实现材料关系的分类。其次，研究 MatBERT-BiLSTM-CRF 模型在材料领域的应用，包括基于 Neo4j 图数据库的材料知识图谱的构建、基于材料知识图谱的描述符树的构建和基于描述符树的知识获取。其中，Neo4j 图数据库用于存储模型抽取的材料实体关系三元组信息和构建材料知识图谱；在此基础上，建立材料描述符树并对其进行填充。通过将描述符树与知识图谱结合以推理获得材料构效关系知识并对其进行表示，利用材料知识嵌入的特征选择方法对知识的有效性进行验证。

1.5 论文的内容安排

本文内容组织如下：

第一章为绪论部分。首先，分析了材料领域文本挖掘的研究背景及意义；其次，概述了材料科学文本挖掘的国内外研究现状，并分析了命名实体识别及关系抽取技术用于材料科学文本挖掘研究中存在的问题；再次，提出了本文主要的研

究目的和内容；最后，简介了本文的结构安排。

第二章研究了基于数据增强的有监督文本挖掘数据集构建方法。首先，介绍有监督材料科学文本挖掘数据集构建和数据增强方法的研究现状及其存在的问题；其次，提出有监督材料科学文本挖掘数据集构建的全流程，并详细叙述了所包含的关键技术；再次，通过实验对数据增强模型的性能、鲁棒性及可用性进行验证；最后，将提出的流程应用于 NASICON 型固态电解质材料文献的获取和文本挖掘数据集的构建。

第三章主要研究了基于多层语义特征融合的材料命名实体识别方法。首先，介绍材料实体识别任务研究现状及存在的问题；其次，提出一种多层语义特征融合的材料命名实体识别模型，并详细叙述了所包含的关键技术；再次，在 NASICON 型固态电解质和无机材料实体识别数据集上进行实验及结果分析来验证实体识别模型的有效性；最后，将提出的材料 NER 方法应用于 NASICON 型固态电解质材料实体信息的抽取，同时设计筛选策略从中选择描述符，并构建机器学习进行激活能预测的研究来验证所选描述符的有效性。

第四章主要研究了基于实体感知的材料关系抽取方法。首先，介绍目前材料关系抽取任务研究现状及存在的问题；其次，提出一种实体感知的材料关系抽取模型，并详细叙述了所包含的关键技术；再次，在 NASICON 型固态电解质和电池材料关系抽取数据集上进行实验及结果分析来验证模型的有效性；最后，将提出的材料关系抽取方法应用于 NASICON 型固态电解质材料实体关系三元组信息的抽取，利用 Neo4j 图数据库实现材料知识图谱的构建，在此基础上建立描述符树并进行填充以获取材料构效关系知识，最终利用知识嵌入特征选择方法对知识的有效性进行验证。

第五章，对全文的工作进行总结，指出可以继续探讨研究的内容，以及对未来工作的展望。

第二章 基于数据增强的有监督材料科学文本挖掘数据集构建方法

标注数据是有监督文本挖掘模型训练的基础。高质量有监督材料科学文本挖掘数据集的构建对于材料科学文本挖掘研究起着至关重要的作用。然而，数据集构建通常要经历文献语料库获取、文本预处理、标签定义和人工标注等一系列复杂流程，面临着高昂的人工开销。文本数据增强作为一种从现有训练样本中生成新的训练样本的技术，可减少有监督数据集的人工标注开销，还能够有效提高文本挖掘模型的鲁棒性。然而，该技术在材料科学应用领域目前仍缺乏相关研究。因此，本章研究基于数据增强的有监督材料科学文本挖掘数据集构建方法。首先，介绍有监督材料科学文本挖掘数据集构建和数据增强方法的研究现状及其存在的问题；其次，提出有监督材料科学文本挖掘数据集构建的全流程，并详细叙述所包含的关键技术；再次，通过实验对数据增强模型的性能、鲁棒性及可用性进行验证；最后，将本章所提出的流程应用于 NASICON 型固态电解质材料文献的获取和文本挖掘数据集的构建。

2.1 问题描述与分析

近年来，随着文本挖掘技术日趋成熟，越来越多的学者将其应用到材料领域以实现自动从文本中挖掘材料信息，并取得了良好的成果^[31, 32, 52, 69-74]。然而，数据质量决定机器学习模型性能的上限，材料领域有监督的文本挖掘任务同样需要高质量文本标注数据^[75]。

随着文本挖掘在材料领域的广泛应用，研究人员已经认识到高质量数据集对文本挖掘模型的重要性。例如，Weston 等人^[31]通过相关期刊和出版商搜集了 327 万个材料文献摘要并从中选取 800 个具有代表性个体，然后基于材料文本特性选择合适的预处理工具对文本进行预处理，最后通过定义的 7 类实体标签对文本进行手工标注，从而构建了一份高质量无机材料命名实体识别数据集。He 等人^[32]从出版商收集了 4061814 篇材料文献并从中选取 371850 个与材料合成实验相关

主题的自然段，然后基于材料文本预处理工具对其进行预处理并定义 2 类实体标签，最后手工标注了一份构建高质量材料合成实验命名实体识别数据集。上述研究表明，高质量有监督文本挖掘数据集的构建流程通常需要经历文献语料库获取、文本预处理、标签定义和人工标注等一系列复杂的操作^[19, 31, 32, 52, 76]。这不仅要求标注人员具备一定的领域知识，而且导致大规模有监督数据集的构建成本极高，使得材料科学文本挖掘的研究难以快速得到突破。

数据增强（Data Augmentation, DA）^[77-80]是一种被广泛应用于扩充训练样本规模的技术。增加训练样本能够有效解决机器学习模型在小样本学习时产生的过拟合问题并增强模型鲁棒性。DA 已经逐渐成为解决数据不足的有效方法之一。例如，合成少数过采样技术（Synthetic Minority Oversampling Technique, SMOTE）^[81, 82]被用于解决机器学习分类数据集的不平衡问题。针对回归数据稀缺问题的 SMOTE（Synthetic Minority Oversampling Technique for Regression, SMOTER）^[81]和具有高斯噪声的合成少数过采样（Synthetic Minority Oversampling with Gaussian Noise, SMOGN）^[83]则被用于解决机器学习回归数据集的增强问题。在此基础上，Zhang 等人^[84]设计了插值过采样（Oversampling with Interpolation, OSIP）和高斯过采样（Oversampling with Gaussian noise, OSGN）两种基于统计的增强方法对大块金属玻璃数据集进行增强以预测其最大直径，并通过实验证明了其设计的机器学习模型更能准确的对玻璃直径进行预测。随着深度学习的发展，以生成对抗网络（Generative Adversarial Network, GAN）^[85]为代表的数据增强方法通过学习样本中潜在的联合概率分布，可以生成高质量的数据来增强真实数据。Naaz 等人^[86]设计了一种基于 GAN 的数据增强方法对目前公开的两个电池参数数据集进行扩充，避免了昂贵和乏味的重复实验收集数据，同时实现了对锂离子电池的充电状态和健康状态的预测，因而验证了 GAN 生成电池数据的真实性与可用性。上述数据增强方法有效地缓解了材料机器学习模型由于数据的不平衡导致预测精度较低的问题。然而，上述方法均仅适用于结构化数据的大规模扩充，无法解决非结构化数据如文本语料的扩增。

现有的文本数据增强方法可以归纳为有条件和无条件增强两种类型。其中，无条件增强只作用于文本数据而不引入标签的信息，其往往通过操纵原始实例中

的几个单词来创建增强实例。例如，单词或短语的替换^[79, 87]、不同语言的回译^[88-90]及对抗方式增强^[91]等。无条件增强方法可以对无监督数据集进行大规模增强，然而由于其对标签的不敏感性，导致在对有监督数据集增强时往往会产生与标签相违背的噪声数据。有条件增强则需强制引入标签信息到模型中，使得生成的数据会考虑数据与标签的一致性。其主要是基于深度生成模型^[92]或预训练语言模型^[80, 93]生成增强数据。有条件的增强方法弥补了无条件增强对标签的不敏感性，因此一定程度上减少了其产生噪声数据的影响。然而，受限于材料文本的特殊性，即材料文献中有大量由多个词、符号和其它类型结构实体组成的术语，如 $(Y, In)BaCo_3ZnO_7$ 、 $(La_{0.8}Sr_{0.2})_{0.97}MnO_3$ 和 $(1 - x)Pb(Zr_{0.52}Ti_{0.48})O_{3-x}BaTiO_3$ ，此外，还有许多独特的材料领域词汇及其缩写，如“X-ray powder diffraction(XRD)”、“bond valence sums(BVS)”及“Scanning Electron Microscope(SEM)”等，使得现有的基于深度生成模型或预训练语言模型的有条件文本数据增强方法对材料文本不敏感，在进行增强时会将特殊词汇一般化处理，从而导致模型生成的自然语言文本质量较差，因此无法直接应用于材料文本数据的增强任务。

综上所述，材料领域仍然缺乏有监督的文本挖掘数据集及有效的文本数据增强方法来减少监督数据标注过程的复杂人工开销。此外，现有的文本数据增强技术在扩充数据时会产生较大的噪声数据，从而影响生成数据的质量。因此，本章提出一种基于数据增强的有监督材料科学文本挖掘数据集构建方法，包括可溯源的文献自动获取、下游任务驱动的文献预处理、材料实体/关系数据标注以及融合材料领域知识的有条件文本数据增强四个阶段。该方法前三个阶段可以快速构建有限的、可溯源且高质量样本，在此基础上，第四阶段将自然语言处理中的预训练语言模型迁移到材料领域，针对材料文本数据增强任务进行训练和微调，使其能充分学习并捕获材料文本的复杂特性，进而动态生成高质量的材料文本增强数据，最终以较低的成本实现高质量材料科学文本挖掘数据集的构建。

2.2 材料科学文本挖掘数据集构建

2.2.1 方法概述

本章提出一种基于数据增强的有监督材料科学文本挖掘数据集构建方法，设计了基于管道模式的高质量文本数据集的构建流程，总体流程框架如图 2.1 所示。该方法由文献获取、预处理、文献标注及数据增强四部分组成。其中，基于可溯源的文献自动获取将溯源机制嵌入网络爬虫程序中，以有效地获取目标文献数据源，包含输入关键字、网站解析和文献爬取三个步骤；基于下游任务驱动的文献预处理在材料文本特性的驱动下设计文本预处理方式，提取出文献中纯文本数据，包含文本获取、文本剖析和文本加工三个步骤，以获取预标注干净的材料文本数据；基于专家经验指导的数据标注分析已有的材料科学文本挖掘数据标注场景，在材料专家的指导下设计有监督数据标签并选择合适的工具进行标注，包括标签定义和数据标注两个步骤，以获得文本挖掘样本数据；融合材料领域知识的有条件文本数据增强将具备领域特性的材料文本知识融合到预训练语言模型中，使得其充分感知材料文本的复杂语义信息，以动态生成高质量的材料文本数据。



图 2.1 高质量材料文本数据集构建的全流程

2.2.2 可溯源的文献自动获取

为了实现可溯源的文献自动获取,本节提出了基于网络爬虫的文献采集和基于元数据的文献数据溯源。其中,前者设计网络爬虫程序,通过对网页解析、页面获取、信息及字段爬取等操作来实现对 PDF 文献的采集;后者则构建可溯源处理模型来实现材料文献语料库和文本挖掘结果的溯源机制。

(1) 材料文献语料获取

文献获取第一步是开发和获得一个目标主题语料库,使得材料信息可以从中检索。然而,材料文献语料内容因可访问程度、目标主题文章语料库数量和文档种类的不同而不同,这些内容只有在以纯文本且可访问的格式呈现时才能在后续的模型中被学习,但这种格式具有多样性的特点^[94],例如论文集的摘要、研究类文章、技术报告、预印本、专利、电子百科全书等。目前,获得这些格式文本语料主要有两种方式:(1)通过使用现有、可用的文本挖掘应用程序编程接口(Application Programming Interfaces, APIs)的索引数据库和搜索工具;(2)通过爬虫访问单个发布者的内容。表 2.1 对两种方式进行了详细的比较。

表 2.1 材料语料获取方式对比

获取方式	数据库名称	文档类型	材料文档数量	付费情况	签订协议
索引数据库 API	CAplus	论文、专利、报告	少	订阅付费	-
	DOAJ	论文(仅限开放获取)	少	部分付费	-
	PubMed Central	论文	少	公开免费	-
	Science Direct	论文	少	订阅付费	-
	Scopus	摘要	少	公开免费	-
爬虫访问	Springer Nature	论文、书籍	少	订阅付费	-
		论文、专利、报告、书籍	多	公开免费	文本和数据挖掘协议

我们期望一种文献获取方式能以较低的成本获取较全面的材料文献语料。基于 API 索引的数据库尽管能提供统一的格式元数据及方便的 API,但其大多数出版物严重偏向于生物医学和生化学科,只有小部分属于物理、有机化学和材料科学,此外其内容的访问是有限的:一方面其需要付费订阅才能获得,另一方面其

只提供公开访问出版物的检索。而基于爬虫的方式则能够从不提供 API 的资源中访问内容，其可以方便得到大量的材料文献且是免费的。尽管在大多数情况下，下载和访问重要发布者内容需要遵守文本和数据挖掘协议，但该抓取和大量下载只会影响出版商服务器的操作。

因此，为了以较低的成本获取较全面的材料文献语料库，本节基于 Python 爬虫工具包研发网络爬虫程序以实现 PDF 文献语料库的获取，并将爬取的结果以公式（2.1）的形式存储。

```
Doc={"Ctt":{“Prg1”:[“Stc11”,“Stc12”,…,“Stc1n”],“Prg2”:[“Stc21”,“Stc22”
,...,“Stc2n”],...,“Prg3”:[“Stc31”,“Stc32”,...,“Stc3n”]},“Tt”:"",“Aut”:""
,“Abst”:"",“Kwd”:"",“Jnl”:"",“Isd”:"",“IF”:"",“Doi”:"",“Spt”:""}  
(2.1)
```

其中，“Doc”表示文献，“Ctt”表示文本内容，“Prg”表示段落，“Stc”表示句子，“Tt”表示标题，“Aut”表示作者，“Abst”表示摘要，“Kwd”表示关键字，“Jnl”表示发表期刊，“Isd”表示发表日期，“IF”表示影响因子，“Spt”表示存储路径。

（2） 基于元数据的文献数据溯源

在基于网络爬虫获取文献语料时，可溯源对于文献数据和下游文本挖掘任务具有十分重要的作用，其不仅可以确保文献数据的来源是可追溯的，从而能够评估材料文献数据的获取、分析和应用是否可复现，而且还能确保文本挖掘结果是可靠的，从而能够提升文本挖掘在材料领域的研究进展。因此，本节提出基于元数据的文献数据溯源，其通过构建可溯源处理模型（Processing Model of Traceability, PMTra）来实现材料文献数据、文本挖掘过程及其结果的溯源，以实时确保研究对象的真实性。

定义 2.1 可溯源处理模型(Processing Model of Traceability, PMTra): PMTra 可表示为 $\langle O_{tra}, MD(O_{tra}), M(O_{tra}) \rangle$ 的三元组形式，其中 $O_{tra} = \{o_{tra}^1, o_{tra}^2, \dots, o_{tra}^n\}$ 为溯源对象； $MD(O_{tra}) = \{D_{tra}^1, D_{tra}^2, \dots, D_{tra}^n\}$ 为描述溯源对象的元数据； $M(O_{tra})$ 为溯源机制。

本节的可溯源处理模型包括数据和过程溯源两个部分。因此，PMTra 的溯源对象实际表示为 $O_{tra} = \{o_{tra}^1, o_{tra}^2\}$ ，对应的描述溯源对象的元数据实际表示为

$MD(O_{tra}) = \{D_{tra}^1, D_{tra}^2\}$ 。其中， D_{tra}^1 为文献数据， D_{tra}^1 为对文献数据进行溯源的基础元数据； D_{tra}^2 为文献挖掘的过程， D_{tra}^2 为对文献挖掘过程进行溯源的衍生元数据。 $M(O_{tra})$ 为数据和过程溯源机制。

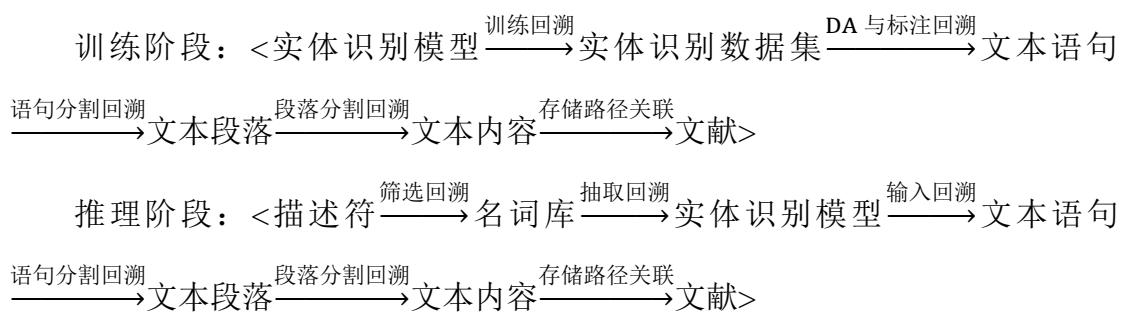
定义 2.2 基础元数据：用于描述文献的基本信息，包括文献的标题、作者、摘要、关键字、文本内容、收录期刊、发表日期、影响因子、DOI 和存储路径等信息。基础元数据在 PDF 文献检索后的界面由网络爬虫程序爬取，一定程度上保证了文献的真实性，其通常比文献中的文本内容更具有结构性。

基础元数据中存储路径是唯一的。因此，本文通过文献的存储路径唯一标识，建立文献和文本挖掘任务数据/过程的关联关系，通过文本内容可以得到文献的存储路径，从而保证了文本挖掘任务数据/过程的源头可追溯。

定义 2.3 衍生元数据：是指文献挖掘过程中产生的、记录文献分析和应用过程的数据，包括模型的输入数据、数据的加工、加工后的结果和加工的顺序等信息，其在文本挖掘过程中被记录。衍生元数据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工程的完整工作流程，从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。

衍生元数据以加工为中心，并建立了数据与加工之间的关系，可表示为<输入数据 i，加工 p，加工结果 o>三元组的形式。其中，前一个流程加工的结果是后一个加工流程的输入数据。以基础元数据和衍生数据为基础，建立数据/加工与原始文献的关系以及数据与加工之间的三元组关系，为数据与过程的溯源机制 $M(O_{tra})$ 的设计和实现奠定了基础。

图 2.2 展示了文献文本挖掘的数据和过程溯源流程，我们基于衍生数据三元组构建了如图 2.2 所示的红色箭头和绿色箭头指向的溯源路径，包括模型训练和推理阶段的溯源路径：



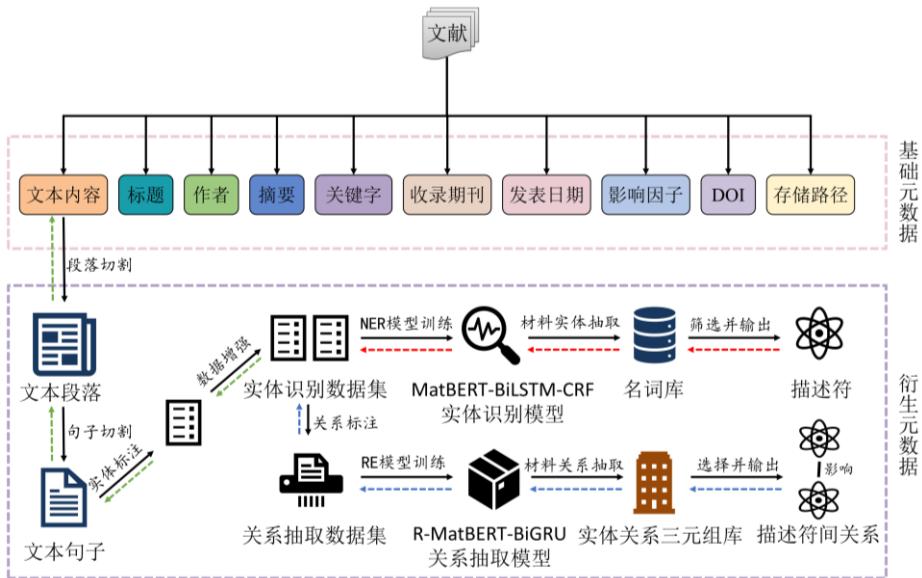


图 2.2 文献数据和过程溯源示意图，黑色实线表示文本挖掘的加工过程，图标表示文本挖掘过程中产生的衍生元数据，红色虚线表示描述符的回溯路径，蓝色虚线表示描述符间关系的回溯路径，绿色虚线表示二者共同的回溯路径。

对于描述符的溯源，我们沿着推理阶段的溯源路径溯源，即从描述符定位到名词库中实体的信息，接着通过实体信息再定位到句子，之后通过句子定位到段落，继而通过段落找到所属的文本内容，最后由文本内容关联到文献的基础元数据，即可定位到文献的存储路径，找到描述符所属的文献，从而实现数据的源头可溯源。当发现实体类别有误时，则需要沿着训练阶段的溯源路径进行溯源，即根据溯源定位到数据集中的数据，接着定位到所属文献的段落句子，从而检查当前实体标注及实体识别方法的正确性，并寻找解决方案对其进行调整。描述符间关系的溯源机制亦是类似地构建基于元数据的溯源路径来进行，其溯源路径如图 2.2 中蓝色箭头和绿色箭头所示。

2.2.3 下游任务驱动的文献预处理

文献预处理是有监督材料数据集构建过程中一个必要环节，它可以对下游文本挖掘任务产生重大的影响^[27]。文献预处理的手段根据文本挖掘方法及其每个阶段使用工具的差异而有所不同。其中，预处理初期旨在得到材料纯文本数据。PDF 科研文献是文本数据存储的主要载体之一，具有易获得、更全面的特质。因此，

大规模提取文本信息仍然需要将 PDF 转化为纯文本格式，即需要对爬取的 PDF 文档进行解析以从中提取出文本数据；在预处理后期，为了得到干净的单词、短语、语句或自然段等预标记语料，需要借助符号化工具等技术对前一步处理的文本进行分割，以将其处理为一个单独或整体的序列信息。基于上述目的，本节提出下游任务驱动的文献预处理方法。

- **预处理初期任务驱动的基于 Python 的 PDF2TXT 脚本程序。**通过该程序实现从 PDF 文档中提取材料纯文本数据。首先，对 PDF 文献进行解析，将 PDF 格式文献输出为 TXT 格式的文本数据；然后，在该过程中，利用正则表达式等相应技术进行去除断行、参考文献、图、表等不利于文献挖掘的内容；最后，将 PDF 文献中剩余的大量纯文本内容存储于 TXT 文本库中。

- **预处理后期任务驱动的基于 ChemDataExtractor 的文本处理。**在预处理后期，需要对纯文本数据进行处理以得到可以用于标注的数据。在英文文本中，标点符号是识别句子较为明显的方式之一，然而材料文本具有领域特殊性，即材料科学领域的语言常常因为由多个词、符号和其它类型结构实体组成的术语而变得复杂。例如， $(Y, In)BaCo_3ZnO_7$ 、 $(La_{0.8}Sr_{0.2})_{0.97}MnO_3$ 和 $(1-x)Pb(Zr_{0.52}Ti_{0.48})O_{3-x}BaTiO_3$ 等。因此，材料领域需要专门的文本处理工具，其对材料科学文本挖掘的成功十分重要。表 2.2 总结了目前常用的材料文献符号化处理方式，这些方式均是基于命名实体识别方法开发的自然语言处理工具，且其通常采用由字典、手工制作的规则或模式或词性标注等技术的组合实现。

表 2.2 材料学科中常用的符号化处理工具

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 ^[95]	局限于化学反应和生物	是	快	中	普通	中
ChemicalTagger ^[96]	局限于化学合成作用和条件	是	慢	中	普通	中
ChemDataExtractor ^[19]	适用于通用材料化学	是	快	高	容易	高

表 2.2 对比了 OSCAR4^[95]、ChemicalTagger^[109]、ChemDataExtractor^[19]材料学科中常用的符号化工具。我们期望选择一种合适的工具，其能够以较完备的功

能处理更多类型的材料文本。ChemDataExtractor 的功能在上述三种工具中的功能最完备，且其操作简单，对用户友好，同时可以处理通用领域的材料化学文本。此外其版本迭代快，表明该工具会实时将最新技术融入进来，未来具有较大的竞争优势。因此，我们将 ChemDataExtractor 作为材料文本符号化处理工具，以实现对材料领域复杂文本进行文本分段、分句及分词等操作以得到干净的能进行标注的半结构化文本数据。

2.2.4 材料实体/关系标注

(1) 基于材料四面体的实体关系标签定义

表 2.3 材料实体标签设计的对比

设计人	设计目标	标签类别数	标签类别	适用领域	可解决问题
Swain 等	开发能自动从大规模非结构化材料文献中挖掘化学信息的工具	3	属性、关系和测量	通用材料	大规模化学数据库快速创建问题
Weston 等	将材料发现的新结果与已发表文献联系起来	7	无机材料、相、描述符、性能、应用、合成方法和表征方法	无机材料	材料查找、指导文献搜索与总结及回答简单的元问题
Wang 等	从文献中自动挖掘出数据驱动材料设计所需高质量可靠数据	4	γ' 耐热温度、密度、固相及液相温度	合金材料	具有高 γ' 耐热温度的钴基单晶高温合金预测问题
Pan 等	构建语义表示框架用于锂离子阴极的文献挖掘	3	无机材料信息、锂离子阴极和描述符	电池材料	锂离子阴极开发的候选材料寻找问题

类别标签设计是影响有监督材料科学文本挖掘模型性能优劣的先决条件之一，在一定程度上决定了数据标注的质量以及文本挖掘的结果（即期望从文本中挖掘出何种类型的材料信息）。材料命名实体识别文本挖掘工作发展至今，不同领域专家从不同的下游任务出发，设计了适用于其独特下游任务的实体标签。我

们对比了 Swain 等人^[19]、Weston 等人^[31]、Wang 等人^[97]和 Pan 等人^[71]设计的实体标签，如表 2.3 所示。从表中可以看出，标签的设定需要从材料的应用点出发，同时要确定标签的可适用领域及其可以解决的问题。本节以通用领域描述符的自动识别为研究目标，期望通过挖掘出不同类型的描述符信息来实现对材料性能预测及构效关系的研究。然而，特定材料的属性会受到多种类型描述符的影响。例如，NASICON 型固态电解质激活能会受到如成分、结构、工艺及性能等描述符的影响。材料四面体即材料学四要素，旨在研究材料的成分、结构、制造、性能以及它们之间的关系^[98]，体现了对材料间构效关系的研究。故在处理工艺-结构-属性-性能四面体准则的驱动下，对需要挖掘的内容进行了总结，如表 2.4 所示。

表 2.4 文本挖掘内容的分析

挖掘内容	描述
成分信息	材料文献中经常会对所研究的材料成分进行概述，包括组成的元素以及元素的含量。
结构信息	材料的结构可能会受到其成分的影响，且其可能会影响材料的属性。
属性信息	材料的属性是材料设计研究中十分重要的特性，它可能受到材料结构的影响；此外，材料的不同属性也会有一定的影响，且不同的材料性能往往决定着其材料的应用，因此属性信息广泛的存在于各类材料文献中。
处理工艺	材料的处理工艺即材料的加工方法，在材料的实验部分存在着大量的处理工艺的信息。
实验条件	材料的实验条件是研究者在对材料进行实验时的另一重要参数，实验的成功与否与实验条件息息相关。
表征方法	材料表征方法是研究材料化学成分、内部组织结构和材料基本特性的检测、分析技术。
应用信息	材料的应用具体是指材料在生活中的应用场景。
特别描述	材料的特别描述表示的是对材料的形容或者对某种材料的高度抽象总结。

为了实现从材料文献中自动获取描述符信息，通过对上述信息进行高度的抽象，本节设计了 8 个描述符类别实体标签，分别为：成分（Composition）、结构（Structure）、属性（Property）、工艺（Processing）、表征（Characterization）、应用（Application）、特别描述（Feature）及外界环境（Condition），其可以概括绝大多数的材料文献的描述符信息。表 2.5 展示了每个标签的定义及示例。同时，为了实现自动从文献中抽取出描述符间的关联关系，本节设计了 8 种实体类型之间的 8 种关系类型，分别为：原因-影响（Cause-Effect）、部分-整体（Component-Whole）、特征（Feature-Of）、位置（Located-Of）、实例（Instance-

Of)、条件 (Condition-On)、方法 (Method-Of) 及其它 (Other) 类, 其具体定义如表 2.6 所示。

表 2.5 材料领域 8 种实体类型定义

实体标签	定义	例子
Composition	任何与化学式有关; 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、相的名称; 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、化学过程行为, 或者物理、化学机制的名词等。	Conductivity, Activation, Radius; Ferroelectric, Metallic; Phase transition, Ionic reaction
Processing	任何合成材料的技术; 任何合成材料的技术; 材料改性的手段等。	Solid state reaction, Annealing; Doping
Characterization	用来表征材料, 实验或理论的任何方法; 也可以是一个模型或者是公式等。	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、形状的特殊说明等。	Single crystal, Bulk, nanotube, Quantum dot
Condition	描述材料所处的环境 (材料的外部条件)。	Temperature, Pressure

表 2.6 材料领域 8 种关系类型定义

关系标签	定义 (A 和 B 为描述符实体)	可能存在关系的实体类型
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	Property-Property 、 Composition-Property
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 无明显关系	-

(2) 基于 EasyData 的实体/关系数据标注

● **标注工具的选择。**为了获得用于模型训练学习的有监督材料实体识别及关系抽取数据集，需要研究者手工标注部分样本。选择合适的标注工具有利于提高标注效率。因此，本节总结了目前常用的文本标注工具，如表 2.7 所示。

表 2.7 文本标注工具

标注工具	擅长标注任务	导入文本要求	角色管理权限	难易性	友好性	扩展性
Label Studio	多模态标注	严格	不完善	一般	中	中
EasyData	实体标注	一般	完善	容易	高	高
Brat	关系标注	较严	完善	一般	中	低
Doccoano	文本分类	严格	较完善	一般	低	低

从表中可以看出，“Label Studio”适合多模态数据的标注，尽管其可以用作文本标注，但是角色权限管理没有“Doccano”完善；“Brat”擅长关系标注，在构建知识图谱数据集时较有优势，但其界面较为粗糙，对用户不是特别友好；“Doccano”提供了文本分类及序列标注的标注功能，尽管其有较完善的角色权限管理和美观的画面风格，但其对文本数据及标签的要求较为严格，如对于预标注标签，其下标是不计算空格的，且多文件导入平台时可能会出错；“EasyData”则功能完善、角色权限管理成熟且操作简单，特别适合命名实体识别的标注，对新手标注人员友好且扩展性高。

不同的文本挖掘方法有着不同格式的数据需求，导致相应的标注流程也有所差异。对于材料命名实体识别任务，本节选择 EasyData 工具进行材料实体识别数据集的标注。通过对定义的描述符实体及关系标签分析可知，材料文本不同实体类型间可能会有重叠关系，而目前已有的标注工具大都只能针对单一实体类型间的关系进行标注。因此对于材料关系抽取任务，本节没有选择标注工具，而是对实体识别后的数据直接进行操作来实现关系数据集的标注。

● **实体识别数据集的标注。**材料实体识别数据集的标注首先需要设置标注模式，实体识别是一个序列标注的问题，即语料库中的每一个句子都被视为一个序列，序列中的每一个单词被看作一个项（token），实体识别的结果需要对序列中每一个项进行分类。此外，实体识别还需设定标注模型以解决一个实体有多个

单词组成的情况，因此本节采用“BIO”^[99]的标注模式。在这种标注模式下，对于单个单词的实体，用“B-”接实体类型进行标注；而对于多个单词的实体，用 B 表示一个实体的开始，将实体的首个单词标注为“B-”接实体类型，其余单词一律看作中间部分 I，用“I-”接实体类型进行标注；对于不属于实体类型的单词，则用 O 进行标注。将预处理好的材料文本以 TXT 格式文件传入 EasyData 实体标注平台，并将本节设计的实体标签与之关联起来。通过材料专家的经验知识来对 TXT 文件段落或句子中相应的实体进行标记。标记结束后，EasyData 导出标记内容——JSON 文件格式，即在对应段落或句子中已标记实体与其实体类型键值对的字典形式，同时还记录了其位置索引信息。为了构建如表 2.8 所示的实体识别数据集格式，需要对 JSON 文件进行进一步转换（后处理），即用 ChemDataExtractor 工具对段落进行句子切分，随后对句子进行项的切分，同时结合 BIO 标注模式及当前位置项的实体标签，最终得到处理后的实体识别模型可直接训练的数据集。

表 2.8 命名实体识别的数据集格式

项	标签
Materials	O
with	O
the	O
general	O
formula	O
AxBB'(PO4)3	B-Composition
consisting	O
of	O
a	O
BB'(PO4)3	B-Composition
framework	O
built	O
up	O
...	...

- **关系抽取数据集的标注。**关系抽取则是一个实体对之间关系的分类任务，其数据的标注需要在实体识别的基础上进行标注。具体地，在获得标注的实体后，将样本数据处理为每个句子只包含两个实体的格式，通过手工选择的方式完成关系标注。

在此基础上，将关系标注的数据进行处理转换为关系抽取模型可以接受的输入格式，如表 2.9 所示。每一条数据包括“句子”、“关系标签”及“特殊说明”三行内容。其中“Sentence”指的是输入的句子，其中含有潜在关系的两个实体分别用“<e1></e1>”和“<e2></e2>”包裹；“Relation”指的是句子中两个实体的关系类型，由于具有某种关系的两个实体间往往是有方向的，因此在关系类别后用“（e1,e2）”或“（e2,e1）”来表示具有该种关系的两个实体间的方向；“Comment”则表示该条关系数据的注释。

表 2.9 关系抽取模型读取的数据格式

描述	数据
Sentence	The linear increases of <e1>M1-O(2) distance</e1> versus A cation ionic radii , in ASnFe(PO4)3 phases , is accompanied with a regular increase of <e2>chex parameters</e2>.
Relation	Cause-Effect(e1, e2)
Comment	-

2.2.5 融合材料领域知识的文本数据增强

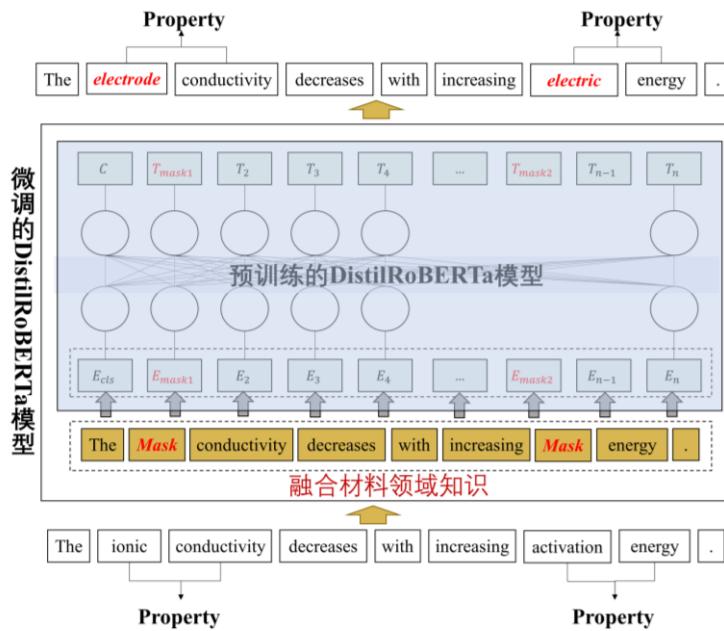


图 2.3 cDA-DK 模型示意图

由于大规模高质量文本挖掘数据集的构建成本较高，因此本章引入数据增强以期通过少量标注样本生成更多高质量数据。然而，将通用领域流行的基于深度

生成模型或预训练语言模型的增强方法直接应用于材料文本数据的扩充时会影响生成数据质量。因此，本节提出融合材料文本知识的有条件文本数据增强模型（cDA-DK 模型），如图 2.3 所示。该模型主要采用自然语言处理 DistilRoBERTa 模型（Roberta^[100]的知识蒸馏版本）对材料文本数据进行扩充。其通过将材料领域文本知识融入预训练 DistilRoBERTa 模型，并对其进行微调来学习材料文本领域特征，从而实现动态地生成高质量的材料文本数据，具体步骤如下：

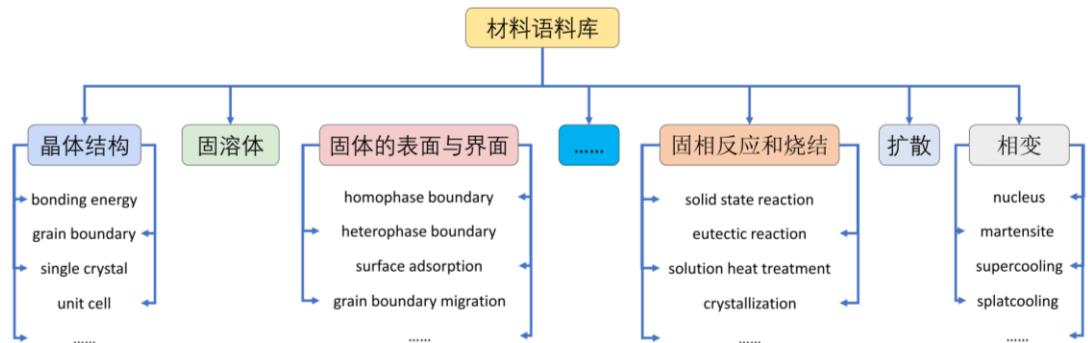


图 2.4 材料语料库示意图

首先，将材料领域知识融入到 DistilRoBERTa 模型。DistilRoBERTa 模型是在 OpenWebTextCorpus^[101]（约 38GB 的网页文本数据）上预训练的语言模型，其有属于自己训练的词汇表及分词器。然而，词汇表中的材料领域特有的词汇并不密集，使得其分词器在遇到这种情况时会将材料专业词汇分割成更小的项来处理，因而导致模型需要处理的序列变长且很难捕获该类领域词汇的特殊含义，同时也会使模型微调的速度变慢。为了加速 DistilRoBERTa 模型的微调速度同时使其能够更好的捕捉材料领域文本的语义信息，我们搜集了大量材料领域独特术语（单词或词组）构成材料语料库如图 2.4 所示，包括材料晶体结构、晶体结构缺陷、晶体结构缺陷-固溶体、熔体结构、固体的表面与界面、相图、扩散、相变、固相反应和烧结等专业词汇，并将其作为材料领域知识融入到 DistilRoBERTa 模型的词汇表中。具体地，检查材料领域特殊词汇是否存在与 DistilRoBERTa 模型的词汇表，并将不存在词汇表中的材料术语通过 DistilRoBERTa 分词器的“add_tokens”方法添加分词并扩展词汇表；通过“resize_token_embeddings”方法将新增词汇的嵌入向量添加到 DistilRoBERTa 模型的嵌入矩阵中。需要注意的是，新增词汇的嵌入是随机初始化为与模型词汇表中单词具有相同维度的嵌入向

量。由此便得到材料语料库指导下的 DistilRoBERTa 模型。

接着，对预训练的 DistilRoBERTa 语言模型进行微调。为了使新加入的领域知识具备更多材料语义，我们结合下游任务（文本增强）对材料语料库指导下的预训练 DistilRoBERTa 模型进行微调，即对于输入的材料文本进行无监督的训练。在此过程中，模型的分词器在遇到材料领域特殊词汇时不会将其分割，从而可以捕获到更多材料文本的特殊含义语义特征，同时减小了模型所需处理的文本序列，进而提升了模型微调的效率。

最后，通过微调后 DistilRoBERTa 模型来实现文本数据的增强。具体地，将待增强的文本数据及其对应的监督标签输入到微调后 DistilRoBERTa 模型中，模型会随机遮掩句子中的一些单词，并记住被遮掩词汇的标签信息，即建立该词汇与其标签的依赖关系，同时模型通过学习到的上下文语义信息生成含丰富材料语义信息的向量形式；接着通过语义向量特征预测被遮掩位置的单词以生成候选的增强词汇，并选择与其语义最为相似的词汇生成增强后的文本数据；最后将增强后的数据存储并输出。值得注意的是，生成的高质量数据完全是基于微调后 DistilRoBERTa 模型所学到的材料语义知识，不仅可以减少噪声数据，而且还与材料领域密切相关。cDA-DK 的伪代码如算法 2.1 所示。

算法 2.1：数据增强方法

输入：待增强数据集 $D_{train} = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$ ，其中， x_i 表示数据集中第 i 个句子， y_i 是 x_i 中每个单词对应的标签；预训练的 DistilRoBERTa 语言模型 $G_{DistilRoBERTa}$ ；材料语料库 $C = \{c_1, c_2, \dots, c_j, \dots, c_m\}$ ，其中， c_j 表示第 j 个单词或短语；增强后的数据集 $D_{synthetic} = \{\}$ 。

输出： $D_{synthetic}$

- 1: 开始
 - 2: **foreach** $c \in C$ **do**
 - 3: c 输入到 $G_{DistilRoBERTa}$ 模型中 // 材料领域知识的融合
 - 4: $F_{DistilRoBERTa} \leftarrow$ 微调 $G_{DistilRoBERTa}$ // 预训练语言模型的微调
 - 5: 初始化 $D_{synthetic}$
 - 6: **foreach** $\{x_i, y_i\} \in D_{train}$ **do**
 - 7: $\{\hat{x}_i, \hat{y}_i\} = F_{DistilRoBERTa}(x_i, y_i)$ // 生成新的样例
 - 8: $D_{synthetic} = D_{synthetic} \cup \{\hat{x}_i, \hat{y}_i\}$ // 样例加入增强数据集合
 - 9: 结束
-

2.3 实验

2.3.1 实验数据

为了验证 cDA-DK 对不同材料体系有监督文本数据集的有效性和迁移能力，我们分别在手工标注及公开的实体识别数据集进行实验，实验数据简介如表 2.10 所示。

表 2.10 cDA-DK 实验数据集

应用材料	数据集	描述	数据量
NASICON 型 固态电解质	Dataset 1	手工标注 55 篇文 献的数据集	2434 个句子共计 65690 条数据
	Dataset 2	模型生成 Dataset 1 的增强数据集	2434 个句子共计 65690 条数据
	Dataset 3	手工标注 35 篇文 献的数据集	305 个句子共计 6980 条数据
	Dataset 4	手工标记 800 个摘 要	5459 个句子共计 142730 条数据
	Dataset 5	模型生成 Dataset 4 的增强数据集	5459 个句子共计 142730 条数据

具体的，在本章爬取到的 1898 篇 NASICON 型固态电解质科研文献中，分别选取其中 55 篇和 35 篇含大量描述符信息的文献并进行预处理，随后进行了小样本实体识别数据集的标注工作，最终得到两份 NASICON 型固态电解质材料实体识别数据集，记为“Dataset 1”和“Dataset 3”，其中前者包含 2434 个句子共计 65690 条数据，后者包含 305 个句子共计 6980 条数据。在此基础上，通过 cDA-DK 模型对“Dataset 1”进行数据增强，将纯增强后的数据集记为“Dataset 2”，其包含 2434 个句子共计 65690 条数据且完全是由模型生成的没有人工干预的增强数据集。

为了验证 cDA-DK 模型在公共数据集上的表现，即模型在不同材料数据集上的鲁棒性，我们选取了 Ceder 等人^[31]公开发表的无机材料数据集，记为“Dataset 4”，其包含 5459 个句子共计 142730 条数据，以及利用 cDA-DK 模型对其增强，生成纯增强后的数据集，记为“Dataset 5”，其包含 5459 个句子共计 142730 条数据。

2.3.2 实验设置

cDA-DK 增强模型的参数设置如下：随机遮掩句子中 3 个单词；利用 DistilRoBERTa 模型预测生成语义相近的 5 个被遮掩位置的单词；最后，从 5 个预测的词中通过余弦相似度计算选择语义最接近原始数据的一个以动态生成增强数据。值得注意的是，本章验证生成数据的质量均由实体识别实验进行，且选择有相同参数配置的实体识别模型（MatBERT-BiLSTM-CRF），其训练集和测试集比例为 8:2。

2.3.3 实验结果与分析

(1) cDA-DK 模型性能验证

本节在 NASICON 型固态电解质和无机材料实体识别数据集上进行实验，结果如表 2.11 所示。

表 2.11 不同材料数据集上的实验结果

应用材料	数据集	P	R	F1
NASICON 型 固态电解质材料	Dataset 1	0.78	0.83	0.80
	Dataset 2	0.68	0.72	0.70
	Dataset 2 + Dataset 3	0.83	0.85	0.84
无机材料	Dataset 4	0.86	0.90	0.88
	Dataset 5	0.75	0.78	0.77

从表中可以看出，实体识别模型在完全由 cDA-DK 模型生成的 NASICON 型固态电解质（“Dataset 2”）和无机材料（“Dataset 5”）增强数据集上训练的 F1 分别为 0.70 和 0.77。尽管上述得分低于其在手工标注数据集上（“Dataset 1”和“Dataset 4”）训练的效果（0.80 和 0.88），但其训练的数据集没有人工干预，即完全是由数据增强模型生成的，且训练后的模型已经能以较高的准确度用于从材料文献中挖掘出相关实体知识。cDA-DK 模型增强的 NASICON 型固态电解质和无机材料数据集上的实体识别实验结果证明了其有效性，同时该模型在上述不同的数据集上都能生成较高质量的材料数据也表明了其具有一定的鲁棒性。

此外，与“Dataset 1”相比，实体识别模型在“Dataset 2 + Dataset 3”的各方

面性能均要更优 (P 、 R 及 $F1$ 分别提高了 5%、3%、4%)，即以少量标注数据及增强数据结合训练的实体识别模型能够以更高的准确度挖掘材料文本中的实体知识。由此表明，本章的 cDA-DK 模型可以有效减少材料有监督文本数据标注的开销。这是因为没有直接使用预训练的 DistilRoBERTa 模型获取材料语义信息，而是通过融合材料文本知识对其进行微调，使得其学习了材料文本的复杂特征，因而能够动态生成高质量的领域文本数据。

(2) cDA-DK 模型生成数据质量验证

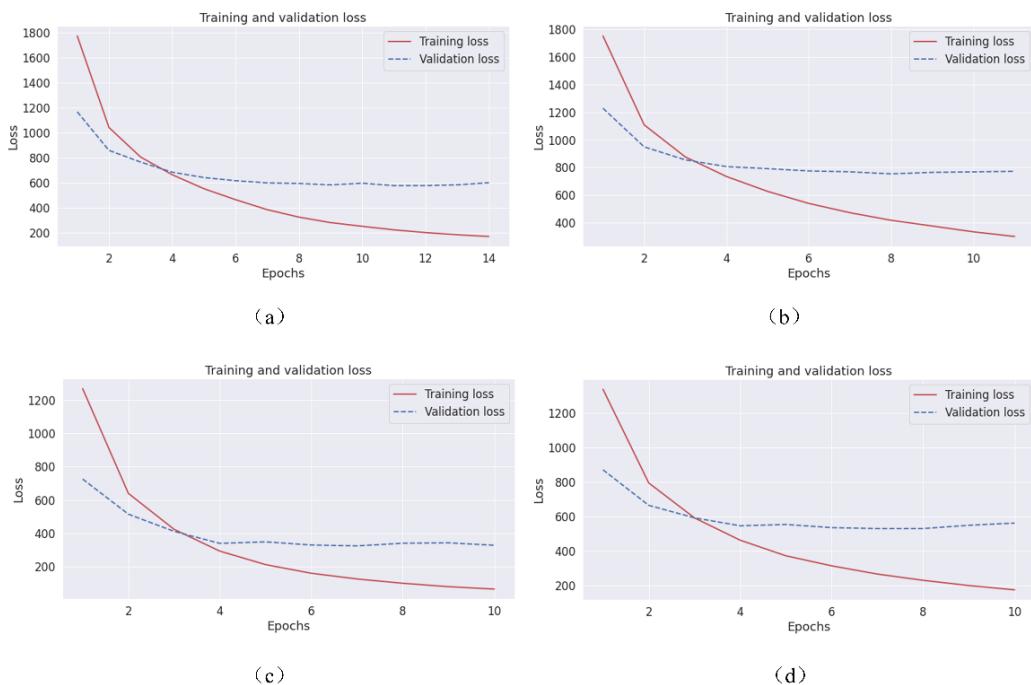


图 2.5 实体识别模型在不同数据集上 loss 曲线。（a）模型在“Dataset 1”训练集和验证集上 loss 曲线。（b）模型在“Dataset 2”训练集和验证集上 loss 曲线。（c）模型在“Dataset 4”训练集和验证集上 loss 曲线。（d）模型在“Dataset 5”训练集和验证集上 loss 曲线。

图 2.5 展示了实体识别模型在手工标注及其增强数据集上的训练过程中损失函数 (loss) 曲线。从中可以看出，图 2.5a (手工标注数据集) 和 2.5b (增强数据集) 模型分别在第 14 及 11 次训练过程中收敛，图 2.5c (手工标注数据集) 和 2.5d (增强数据集) 分别在第 10 及 9 次收敛，即模型在增强数据集上的训练速度更快，且收敛速度也更快。这是 cDA-DK 模型将其预训练模型词汇表中缺乏的材料特殊术语加入其分词器及模型词汇表中一起对 DistilRoBERTa 模型进行微调。在该过程中，材料特殊术语不会被分词器切分，因而减小了模型需要处理的

序列，从而提升了模型的效率。此外，融入材料领域知识不仅可以使模型微调的速度加快，还能更好地捕捉材料数据中的信息。

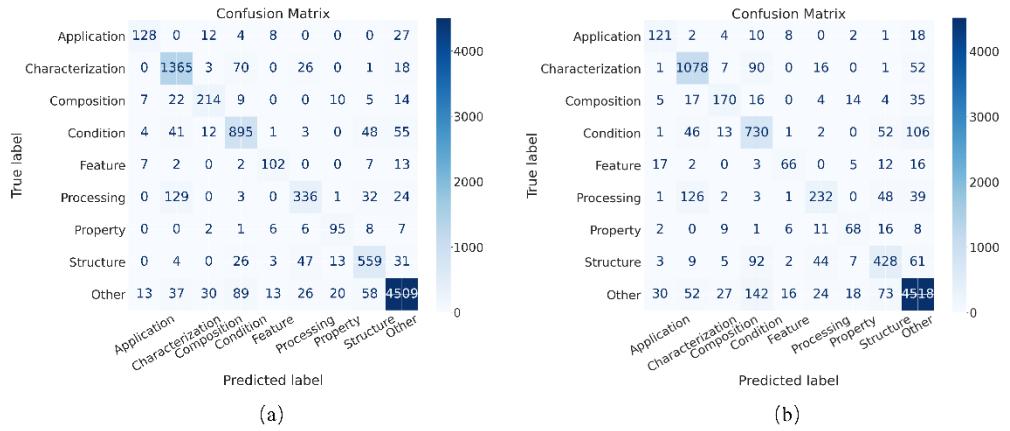


图 2.6 实体识别模型在 NASICON 型固态电解质材料手工标注及其增强数据集上训练的混淆矩阵。（a）模型在手工标注数据集上的混淆矩阵。（b）模型在增强数据集上的混淆矩阵。

图 2.6 展示了实体识别模型分别在“Dataset 1”及“Dataset 2”上训练的混淆矩阵。从图中可以看出，无论是在手工标注还是数据增强数据集上，实体识别模型在八个实体类别上预测正确标签所占的比重均是最大的，表明 cDA-DK 模型生成的材料文本数据的可靠性。尽管如此，实体识别模型在纯手工标注的数据集上训练后的预测结果（图 2.6a）要优于增强数据集上训练后的预测（图 2.6b）。这是因为纯手工标注的有监督文本数据是在领域专家指导下构建的，其每条样本的正确性已经过多名专家评估确认，cDA-DK 模型在有条件数据增强过程中会受到无法避免的样本分布不均衡等问题的影响，导致模型对这些实例的处理能力有限。值得注意的是，cDA-DK 在无需人工标注的情况下生成材料文本数据的质量接近于专家指导下的人工标注数据。

2.4 应用

在电池材料研究中，NASICON 型化合物因具有较好的热稳定性、化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材料研究中已受到广泛关注^[102-105]。固态电池的离子导电性能取决于离子在固态电解质中的扩散，迁移离子的激活能是衡量 NASICON 型固态电解质材料离子输运性能的关键指标之一。因此，实现 NASICON 型固态电解质激活能的精准预测，能够加速新型高

性能固态电解质材料的发现过程。

随着文本挖掘技术的日趋成熟，越来越多的材料学家将其应用于材料领域以从文献中自动挖掘领域知识，从而达到辅助材料设计的目的。然而，目前尚未有 NASICON 型固态电解质文献挖掘的相关研究，因而缺乏相应的有监督文本挖掘数据集。本节以 NASICON 型固态电解质为例，将 2.2 节的方法应用于 NASICON 型固态电解质材料有监督文本挖掘数据集的构建，包括有监督文本挖掘数据集的手工构建和有监督文本挖掘数据集的扩充。

2.4.1 NASICON 型固态电解质有监督文本挖掘数据集的构建

基于 2.2.2-2.2.4 节材料实体识别和关系抽取任务的标注流程，我们手工标注了 NASICON 型固态电解质材料实体识别及关系抽取数据集共 2434 个句子、4857 个实体及 2297 个关系。为了证明手工标注材料数据的可用性，本节进一步将手工标注的数据集与公开的 CoNLL-2004 数据集进行对比，分别对数据集的句子数量、实体数量和关系数量进行比较，结果如表 2.12 所示。

表 2.12 NASICON 型固态电解质材料实体关系数据集与 CoNLL-2004 数据集的对比

数据集	CoNLL-2004 数据集	NASICON 型固态电解质材料数据集
句子数	1441	2434
实体数	5347	4857
关系数	2020	2297

从表中可以看出，手工标注的 NASICON 型固态电解质材料数据集整体句子数量几乎是 CoNLL-2004 数据集句子数量的一倍，实体数略少于公开数据集的实体数量，关系数则多于公开数据集，在定量方面初步证明 NASICON 型固态电解质材料数据集的可用性。此外，在实体识别及关系抽取任务中，句子的长度及其复杂性往往会影响模型的训练效果，因此本节进一步将两个数据集的句子长度也进行了比较。如图 2.7 所示为两个数据集句子长度的对比图，从中可以看出 NASICON 型固态电解质材料数据集的句子长度分布与公开的 CoNLL-2004 数据集大体相同，均呈正态分布趋势。其中，句子长度在 11 个单词到 30 个之间的明显多于公开数据集，而材料实体识别及关系抽取数据集的 10 个单词以下的句子数量占比较少。表明手工标注数据集的句子长度普遍较长，从侧面也说明材料文

本较为复杂。其余句子长度与公开数据集句子长度数量十分接近，综上证明了手工标注数据集的可用性。

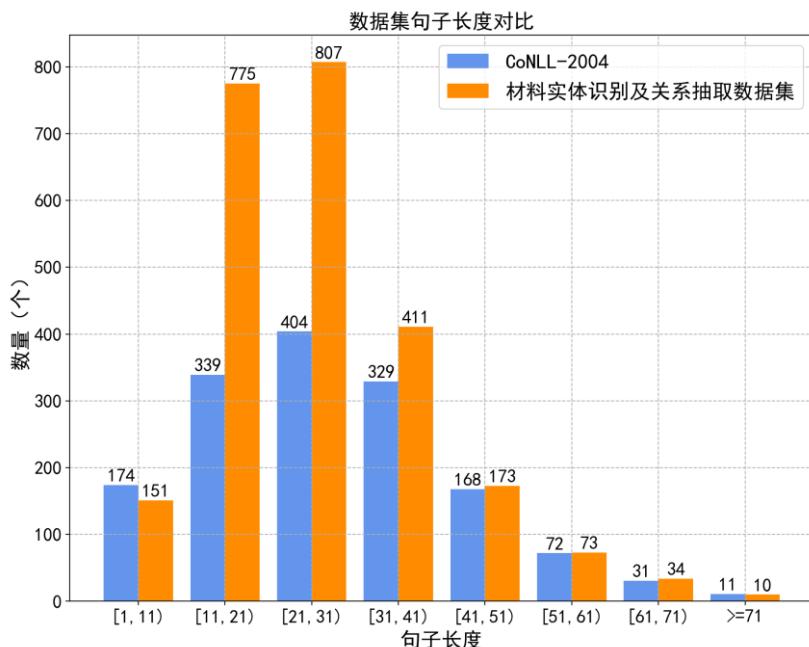


图 2.7 不同数据集句子长度对比

表 2.13 材料实体识别数据集实体类型示例

实体类型	示例
成分	The bond valence sum (BVS) of 1.08 is in good agreement with the value expected for Na+ .
结构	The results from both measurements can be correlated only, if at room temperature Na+ is located in the interstitial sites and if the transition of these ions to the Na2 site occurs at high temperature.
属性	Conductivity measurements in the range 30-350 s^{-1} reveal an activation energy of 0.3 eV for Na+ conduction but conductivity values were found to change with temperature of sample preparation.
加工工艺	Regrinding and reheating the mix results in very slow incorporation of the free ZrO ₂ , probably by replacement of Na+ from the zirconium sites.
应用	It is shown that Na ₃ TiP ₃ O ₉ N can reversibly cycle Na-ions in a manner suitable for secondary batteries , and that the volume changes on Na removal are remarkably small (<1%) relative to other known Na-ion.
表征方法	The XRD diffraction pattern of Na ₃ MnTi(PO ₄) ₃ can be indexed into a rhombohedral NASICON type unit cell with the R3c space group.
特殊描述	Single crystal x-ray analysis was used to identify the composition NaZr ₂ P ₃ O ₁₂ and to refine its structure, which has rhombohedral.
外部条件	The anisotropy of the thermal vibrations of sodium atoms in NaSn ₂ (PO ₄) ₃ at room temperature is described by two different flattened ellipsoids.

NASICON 型固态电解质材料实体识别及关系抽取数据集共包含 8 种实体和 8 种关系类型。为了进一步分析手工标注数据集的特点，本节给出了每种实体及关系类型的实例。如表 2.13 所示为材料实体识别数据集每种实体类型的实例，表 2.14 则是材料关系抽取数据集每种关系类型的实例。从表 2.13 的实体类型示例中可以看出，材料文本中有许多材料领域特殊的词汇，例如化学式、表征方法等。此外，相同的单词在不同语境下可能表达不同的含义，故而其实体类型也同。例如，当“Na⁺”单独出现时属于成分实体类型，而当其与“site”同时出现时则表示结构实体类型；“bottleneck”在传统语境下表示事情发展状态停滞不前，而在材料文本大部分语境下则表示的是晶体的结构信息。总之，材料文本数据冗长、结构复杂，领域词汇普遍存在特殊性、多义性和指代不清等问题。

表 2.14 材料关系抽取数据集关系类型示例

关系类型	示例
造成-影响	<p>1. (成分-属性、属性-属性)In the system Na_(1+x)Zr₂Si_xP_(3-x)O₁₂ , on the other hand, the introduction of excess <u>Na ions</u>⁽¹⁾ introduces electrostatic <u>Na+-Na+ interactions</u>⁽²⁾ that can lower the <u>activation energy</u>⁽³⁾ even though transport must be via a Na₁ site.</p> <p>2. (结构-属性)The increase of the <u>M1 site size</u>⁽¹⁾ in Na₂SnFe(PO₄)₃ is accompanied by oxygen displacements perpendicular to the c axis which give rise to <u>rotation</u>⁽²⁾ of the PO₄ tetrahedra and leads to a <u>distortion</u>⁽³⁾ of the Sn(Fe)(1-x)(PO₄)₃ framework.</p>
部分-整体	<p>1. (成分-成分)The original Na super ionic conductors NASICON materials were solid solutions derived from <u>NaZr₂P₃O₁₂</u>⁽¹⁾ by partial replacement of <u>P</u>⁽²⁾ by <u>Si</u>⁽³⁾ with extra <u>Na</u>⁽⁴⁾ to balance the charges.</p> <p>1 (成分-属性)The calculated BVS value 5.22 shows that the <u>As₅ cation</u>⁽¹⁾ is also slightly <u>over bonded</u>⁽²⁾.</p>
特征	<p>2 (成分-结构)The present paper reports on the <u>crystal structure</u> and vibrational spectra of <u>NaZr₂(AsO₄)₃</u>.</p>
位置	<p>1. (成分-结构)The <u>Na cation</u>⁽¹⁾ is located on the <u>6b position</u>⁽²⁾ with a trigonal antiprismatic coordination and enhanced anisotropic displacement parameters.</p> <p>1. (属性 - 属性)The monochromator is a crystal of Ge that selects a <u>wavelength</u>⁽¹⁾ of <u>1.594 Å</u>⁽²⁾.</p>
实例	<p>2. (特殊描述 - 特殊描述)X-ray powder diffraction shows that the <u>phosphates</u>⁽¹⁾ belong to the <u>NZP type</u>⁽²⁾.</p> <p>3. (结构-结构)The four <u>P-O bonds</u>⁽¹⁾ in the near regular <u>tetrahedron</u>⁽²⁾ (point symmetry2) range from 1.524-1.525 Å, with O-P-O angles deviating by no more than 1.5 <sYm> from the ideal 109.5 <sYm> tetrahedral angle (Table III).</p>

条件	1. (工艺-条件)For each crystal structure determination, data were collected using Mo Ka radiation ⁽¹⁾ up to 29-650 ⁽²⁾ . 2. (属性-条件)The disorder ⁽¹⁾ is larger at 100 K ⁽²⁾ than at 295 K ⁽³⁾ .
方法	1. (表征-属性)The Rietveld plots ⁽¹⁾ represent a good structure fit between observed and calculated intensity with satisfactory R-factors ⁽²⁾ . 2. (表征-属性)Its anisotropic thermal expansion ⁽¹⁾ , has been calculated from high temperature X-ray diffraction ⁽²⁾ , and it is linear in the range from room temperature up to 800 <sYm>. 3. (工艺 - 成分) Good crystals ⁽¹⁾ could, however, be obtained after tempering ⁽²⁾ in platinum crucible for several weeks at 11000 <sYm>.
其它	1. (条件-属性)At room temperature ⁽¹⁾ no diffuse intensity ⁽²⁾ was observed.

从表 2.14 关系类型示例中可以看出材料关系抽取数据集的复杂性, 即存在着重叠关系抽取。具体的, 材料文本句子中实体及关系之间存在一对一、一对多及多对多的重叠关系, 即一个实体对可能对应一种关系类型, 也可能对应两个及以上不同的关系类型, 此外相同或不同的实体对在不同的材料语境下可能对应不同的关系。例如, 材料的属性间存在着互相影响的关系; 材料的成分会影响材料的结构; 材料的结构会影响材料的属性; 不同的实体类型具有相同的关系等。总之, 材料领域具有十分复杂的文本特性及关系结构, 导致材料关系抽取数据集中具有大量的重叠关系。

2.4.2 NASICON 型固态电解质有监督文本挖掘数据集的扩充

表 2.15 NASICON 型固态电解质材料科学文本挖掘数据集增强前后数据示例对比

数据	示例
原始数据	The (O) ionic (B-Property) conductivity (I-Property) decreases (O) with (O) increasing (O) activation (B-Property) energy (I-Property). (O)
增强数据	The (O) electrode (B-Property) conductivity (I-Property) decreases (O) with (O) increasing (O) electric (B-Property) energy (I-Property). (O)

为了构建大规模的 NASICON 型固态电解质材料有监督文本挖掘数据集, 同时最大程度的减少人工标注的开销, 我们将 2.2.5 节提出的 cDA-DK 文本数据增强模型应用于此进行数据的扩充。表 2.15 展示了 NASICON 型固态电解质材料科学文本挖掘数据集增强前后数据示例对比。其中, 我们对原始数据中发生改变的词用红色高亮显示, 从中可以看出, 增强后的数据均是和材料相关且于原文相近的文本。

最终，我们将 2.2.5 提出的 cDA-DK 数据增强模型作用于 NASICON 型固态电解质材料有监督文本挖掘数据集，其增强前后数据量对比如表 2.16 所示。

表 2.16 NASICON 型固态电解质材料科学文本挖掘数据集增强前后数量对比

数据集	原始数据集	增强数据集
句子数	2434	4846
实体数	4857	9714
关系数	2297	4594

从表 2.16 中可以看出，增强后的文本数据在句子、实体和关系数量上均扩充了一倍。即我们基于 cDA-DK 模型实现了大规模高质量材料有监督文本挖掘数据集的自动生成，从而减少了传统数据人工标注的复杂开销，为材料科学文本挖掘的研究提供了数据开发支撑。

2.5 小结

本章主要研究了基于数据增强的有监督材料数据集的构建方法，首先阐述了有监督数据集的构建流程和数据增强的研究现状，并分析了存在的问题；然后针对材料有监督文本挖掘数据集标注难的问题，本文提出了基于数据增强的有监督材料科学文本挖掘数据集构建方法，包括可溯源的文献自动获取、下游任务驱动的文献预处理、材料实体/关系数据标注及融合材料领域知识的有条件文本数据增强（cDA-DK 模型）。其中，cDA-DK 模型通过对融合材料领域知识的预训练 DistilRoBERTa 模型进行微调，使得其能感知材料领域的特殊性并学习到复杂的上下文语义信息，从而可以实现在有限手工标注可溯源且高质量样本的基础上自动生成文本数据。通过对 NASICON 型固态电解质和无机材料实体识别数据集进行增强实验，证明了 cDA-DK 模型不仅可以通过少量标注的数据生成高质量文本数据，而且还能提高下游文本挖掘模型的鲁棒性。

第三章 基于多层语义特征融合的材料命名实体识别方法

命名实体识别（Named Entity Recognition, NER）方法可以自动从文本中挖掘出关键字或短语等信息，在材料领域已取得初步成效。然而，材料文本数据冗长、结构复杂，领域词汇普遍存在特殊性、多义性和指代不清等问题，使得现有的 NER 方法不能充分捕获材料单词的语义信息及其与实体标签的依赖关系，从而影响模型分类的准确率。本章针对上述问题展开材料领域 NER 方法的研究。首先，介绍材料 NER 的研究现状及其存在的问题；其次，提出一种多层语义特征融合的材料 NER 方法，并详细叙述所包含的关键技术；再次，将在 NASICON 型固态电解质和无机材料实体识别数据集上进行对比及消融实验来验证模型的有效性；最后，将提出的材料 NER 方法应用于 NASICON 型固态电解质激活能相关描述符的抽取，同时设计筛选策略进一步选择描述符，并结合机器学习进行激活能预测的研究。

3.1 问题描述与分析

材料 NER 方法能够通过识别和分类文本中提及的概念来挖掘具有语义价值的实体对象。这些实体对象不仅可以映射到材料性能上，还能为研究人员找到相似的化合物或纳入注释标记提供巨大的帮助。然而，材料领域 NER 的研究依然处于起步阶段，使用的技术以传统基于字典、规则、机器学习的单一或组合方法偏多。例如，Lowe 等人^[106]将字典和基于模式的技术组合进行 NER 的研究，使用命名约定规则研发了生物材料 NER 工具 LeadMine。Lezan 等人^[96]通过解析化学文本实验合成部分的化学标记，基于模式规则和字典的组合技术进行材料化学 NER 的研究并开发了一个材料化学 NER 工具 ChemicalTagger。Swain 等人^[19]提出了基于字典和机器学习（CRF）组合技术的化学 NER 方法，实现了化学实体及其相关属性、测量和关系标签自动抽取的 ChemDataExtractor 工具的研发。上述材料 NER 方法可以在小数据集上达到一个很高的识别准确率，但是在面对大

规模的数据集或其它领域时便不适用，往往需要重新设置新的规则或者收集新的字典。此外，手工设计规则及收集字典的步骤往往需要花费大量的人力、物力及财力。

近年来，深度学习技术已经被应用于材料领域进行材料 NER 的研究，该技术通过构建深度学习模型来自动提取材料文本间的语义特征来实现实体信息的抽取。例如，Weston 等人^[31]设计了 7 个实体标签：无机材料（MAT）、对称/相标签（SPL）、样本描述符（DSC）、材料性能（PRO）、材料应用（APL）、合成方法（SMT）和表征方法（CMT），并将 Word2vec 词嵌入模型（提取语义信息）和深度命名实体识别 BiLSTM-CRF 模型引入材料领域来对材料 NER 进行研究。他们通过对模型的训练最终实现了从摘要中自动抽取相关命名实体。He 等人^[32]为了提取材料合成信息，将 Word2vec 词嵌入模型引入材料领域以提取语义信息，并利用 BiLSTM-CRF 深度学习模型进行材料 NER 的研究。他们设计了一个两步的化学实体识别系统，其通过训练可以实现从文献中抽取化学合成的前体和目标，为材料化学的合成提供了新思路。上述深度学习材料 NER 技术极大的推动了材料科学文本挖掘的研究。然而，材料文本数据冗长、结构复杂，领域词汇普遍存在特殊性、多义性和指代不清等问题。针对上述问题，材料领域常用的 Word2vec 词嵌入模型^[107]仅通过将每个单词映射为高维空间的词嵌入向量无法动态地捕获含丰富语义信息的单词及句子级别嵌入向量，因而制约着实体识别模型的准确率。此外，Yimam 等人^[108]通过大量的对比实验发现，与预训练的 BERT 模型相比，无监督的 Word2vec 方法的确很难完全理解领域词汇。因此，本章拟将自然语言预处理中预训练 BERT 模型引入材料领域对材料命名实体识别方法进行研究。

综上所述，材料领域仍缺乏有效的深度学习 NER 模型来自动从文献中挖掘材料信息。同时，现有的材料 NER 模型使用的词嵌入模型生成的是与语境无关（静态嵌入）、不具备复杂特征（如语法、语义）的词嵌入向量，严重影响实体识别的准确率。因此，本章将 BERT 模型迁移到材料领域进行材料 NER 方法的研究，并提出多层语义特征融合的实体识别模型 MatBERT-BiLSTM-CRF。该模型通过 MatBERT 来同时编码材料词嵌入、位置嵌入及句子嵌入信息从而动态捕

获材料复杂文本间的深层语义特征，以缓解材料复杂句子的一词多义及代词指代的编码问题，并利用 BiLSTM 对句子序列进行建模，以抽取材料局部上下文语义特征，最后采用 CRF 对句子最优的标签序列进行预测，以实现材料实体的精准分类。

3.2 方法概述

本章研究了面向材料领域的材料命名实体识别方法，提出了多层次语义特征融合的材料命名实体识别方法 MatBERT-BiLSTM-CRF。该方法包括基于 MatBERT 的多级别语义特征的融合、基于 BiLSTM 的局部上下文语义特征的融合和基于 CRF 的材料实体分类，以快速抽取材料实体信息。进一步，提出了基于材料命名实体的描述符筛选，包括基于名词库的材料实体存储和基于重要度计算的描述符筛选策略，以从文献中筛选出适用于目标材料性能预测的高质量描述符。整个方法以管道的形式进行材料命名实体识别模型构建与应用，具体流程如图 3.1 所示。



图 3.1 多层语义特征融合的材料命名实体识别方法及应用流程

3.3 基于多层次语义特征融合的材料命名实体识别

图 3.2 展示了基于多层次语义特征融合的材料命名实体识别模型 MatBERT-BiLSTM-CRF 的结构，由基于 MatBERT 的多层次语义特征的融合、基于 BiLSTM 的局部上下文语义特征的融合和基于 CRF 的材料实体分类三部分组成。其中，

MatBERT 同时编码词嵌入、位置嵌入及句子嵌入信息，以捕获含丰富材料信息的 token 及句子级别的语义特征并将其融合形成单词的向量表示；在此基础上，BiLSTM 对句子序列建模，以进一步捕获单词的局部上下文语义特征，从而最终获得 token 级别的语义特征向量；序列标注分类器 CRF 基于 token 级别的语义特征向量对单词或短语进行标签预测以获取最优的标签序列，从而实现实体分类。

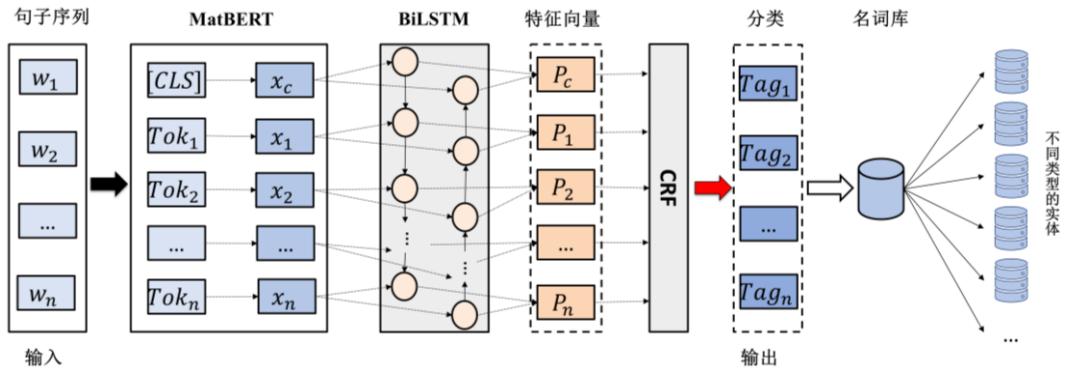


图 3.2 基于多层语义特征融合的材料命名实体识别模型结构图

3.3.1 基于 MatBERT 的多级别语义特征的融合

MatBERT 通过同时编码词嵌入、位置嵌入及句子嵌入信息，来捕获含丰富材料信息的 token 及句子级别的语义特征，并将其融合以形成单词的向量表示。MatBERT 主要采用了 Transformers 的编码器架构，融合了当前单词左侧及右侧的上下文信息。在训练词向量时，编码器不再从左到右或从右到左编码句子来预测单词，而是根据一定的比例随机隐藏或替换部分单词，并根据上下文预测原始单词。此外，MatBERT 模型还增加了句子级别的训练任务来学习句子间的上下文。具体做法是随机替换一些句子，编码器使用前一个句子来预测下一个句子是否为原始句子。联合训练上述两个任务，以抽取含丰富语义信息的 token 级和句子级别的语义特征。MatBERT 动态捕获含丰富材料信息的 token 及句子级别的语义信息，可以有效缓解材料领域词汇的编码问题。

具体地，给定一个句子序列 “[CLS] The space group R3c becomes ... [SEP]” 作为输入。其中，[CLS] 表示一个句子序列的起始位置，[SEP] 表示句子间的间隔符，它们是由 MatBERT 引入的用于句子级别训练任务的特殊标记。每个单词向量的表示由三部分组成：单词嵌入、句子嵌入和位置嵌入。其中，单词的嵌入向

量定义为 $e_w = (e_{[CLS]}^w, e_{w1}^w, e_{w2}^w, \dots, e_{wn}^w, e_{[SEP]}^w)$, 句子的嵌入向量定义为 $e_s = (e_A^s, e_A^s, e_A^s, \dots, e_A^s, e_A^s)$, 位置的嵌入向量定义为 $e_p = (e_0^p, e_1^p, e_2^p, \dots, e_n^p, e_{n+1}^p)$ 。需要注意的是, 单词嵌入向量由 MatBERT 提供的词汇表决定。由于训练样本是一个单句, 所以句子嵌入向量被设置为 0。将上述三个嵌入向量累加得到的单词特征作为 MatBERT 模型的输入, 如图 3.3 所示。通过训练, MatBERT 最后输出的单词向量表示为式 (3.1), 作为 BiLSTM 的输入。

$$x = [x_1, x_2, \dots, x_n] \quad (3.1)$$

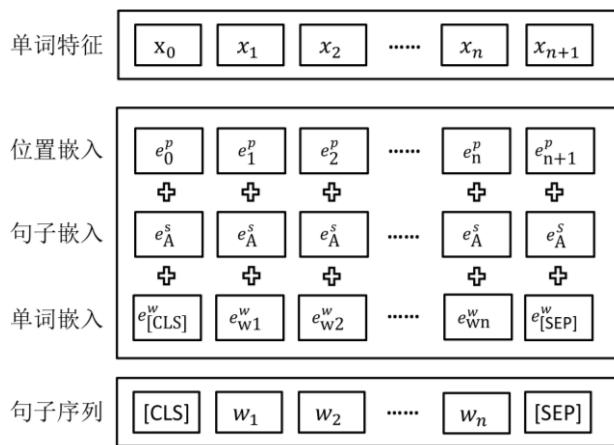


图 3.3 基于 MatBERT 模型的单词向量表示输入

3.3.2 基于 BiLSTM 的局部上下文语义特征的融合

NER 任务需要对句子中所有单词序列进行分类, 因此需要模型具有识别时序信息的能力。RNN 拥有捕捉时序信息进行端到端分类的能力。然而, RNN 在时序信息的传播过程中经常会遭受梯度消失和梯度爆炸的问题。针对该问题, 本节引入 RNN 的一个变体 LSTM 来解决。LSTM 有三个门控单元, 即输入门、遗忘门和输出门, 如图 3.4 所示, 它们能够有选择地保存上下文信息。因此, LSTM 能有效缓解梯度消失及梯度爆炸的问题同时比 RNN 更适合捕捉长距离依赖。

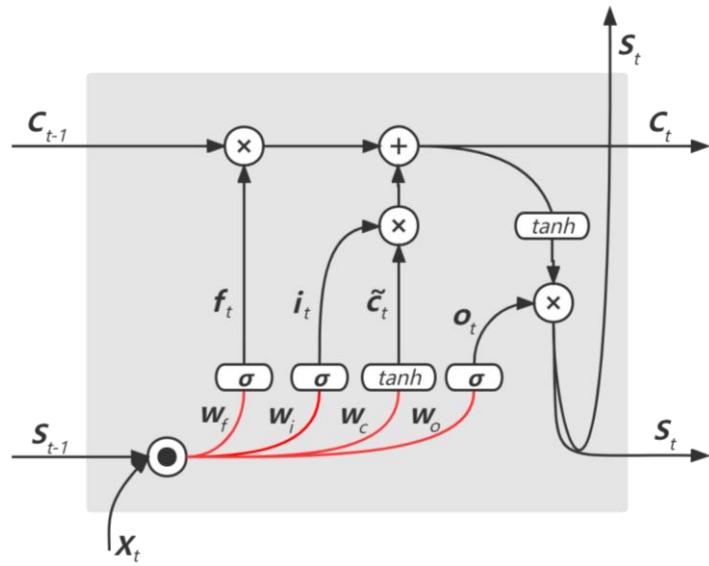


图 3.4 LSTM 单元结构

对于 t 时间，LSTM 单元状态计算如公式（3.2）~（3.7）所示：

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (3.2)$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (3.3)$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (3.4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (3.5)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (3.6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (3.7)$$

其中 σ 和 \tanh 表示不同的激活函数， \odot 表示点积。 W_i, W_f, W_c, W_o 是权重矩阵， b_i, b_f, b_c, b_o 是偏置值。 x_t 是 t 时刻的输入向量， h_t 是 t 时刻的输出向量，由最后一层隐藏层状态获得，其包含了 t 时刻之前的所有有效信息。 i_t, f_t, o_t 分别表示 t 时刻对输入门、遗忘门、输出门的控制。

为了最大程度的捕获当前时刻前后两个方向的局部语义特征，本节采用由前向 LSTM 单元和后向 LSTM 单元组成的 BiLSTM。其中，前向单元的隐藏层表示为 \vec{h}_t ，后向单元的隐藏层表示为 \overleftarrow{h}_t 。通过公式（3.2）~（3.7），可以得到 t 时刻单向隐藏层的输出，如公式（3.8）~（3.9）所示。BiLSTM 最终的隐藏层输出是由前向 LSTM 单元与后向 LSTM 单元的隐藏层输出拼接得到，如公式（3.10）所示。

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}) \quad (3.8)$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t-1}}) \quad (3.9)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (3.10)$$

3.3.3 基于 CRF 的实体分类

机器学习分类任务通常利用 *Softmax* 函数进行分类。然而在面对 NER 序列标记问题时，该方法无法对序列中的每一帧进行分类。此外，相邻的实体标签之间往往存在一定的转移关系，考虑相邻标签之间的关联性能最大程度的为给定的输入句子序列解码出最佳标签链。CRF 作为序列标注问题的分类器能够捕捉到输出标签的强相互依赖关系，从而获得最优的标签序列。因此，本节采用 CRF 模型作为 NER 的分类器。

具体地，假设使用 $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n\}$ 来表示一个通用的输入序列，其中 \mathbf{w}_i 是第 i 个单词的输入向量， $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ 表示对应于 \mathbf{w} 的标签序列。公式(3.11)计算 CRF 模型的评估得分。

$$score(W, y) = \sum_{i=1}^n T_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (3.11)$$

其中， \mathbf{T} 表示转移矩阵， $T_{y_i, y_{i+1}}$ 为 y_i 标签转移到 y_{i+1} 标签的概率分数， P_{i, y_i} 代表第 i 单词被标记为 y_i 的概率分数。公式(3.12)表示句子 S 计算产生标签序列 \mathbf{y} 的概率。

$$p(y|S) = \frac{e^{score(W,y)}}{\sum_{\tilde{y} \in Y_W} e^{score(W,\tilde{y})}} \quad (3.12)$$

其中， \tilde{y} 为真实标签。

在训练过程中，标记序列的似然函数如式(3.13)所示。

$$\log(p(y|S)) = score(W, y) - \log \left(\sum_{\tilde{y} \in Y_W} e^{score(W,\tilde{y})} \right) \quad (3.13)$$

其中， Y_W 表示所有可能的标记集合，通过似然函数可以得到有效的输出序列。最终通过公式(3.14)的计算得出整体概率得分最大的一组序列。

$$y^* = \arg \max_{\tilde{y} \in Y_W} score(W, \tilde{y}) \quad (3.14)$$

通过上述步骤便能为句子序列解码出最优的标签序列从而实现对句子中每

个单词的标签进行精准预测。

3.4 基于材料命名实体的描述符筛选

为了实现多层语义特征融合的材料命名实体识别方法在材料领域的应用,本章进一步提出了基于材料命名实体的描述符筛选方法,其是由基于名词库的材料实体存储及基于重要度计算的描述符筛选策略构成。其中,名词库用于将大规模材料实体信息进行分类存储;在此基础上,设计重要度计算策略以从中筛选出与特定目标材料性能相关的高质量描述符实体,进而通过构建机器学习模型便能实现相关材料性能的预测。

3.4.1 基于名词库的材料实体存储

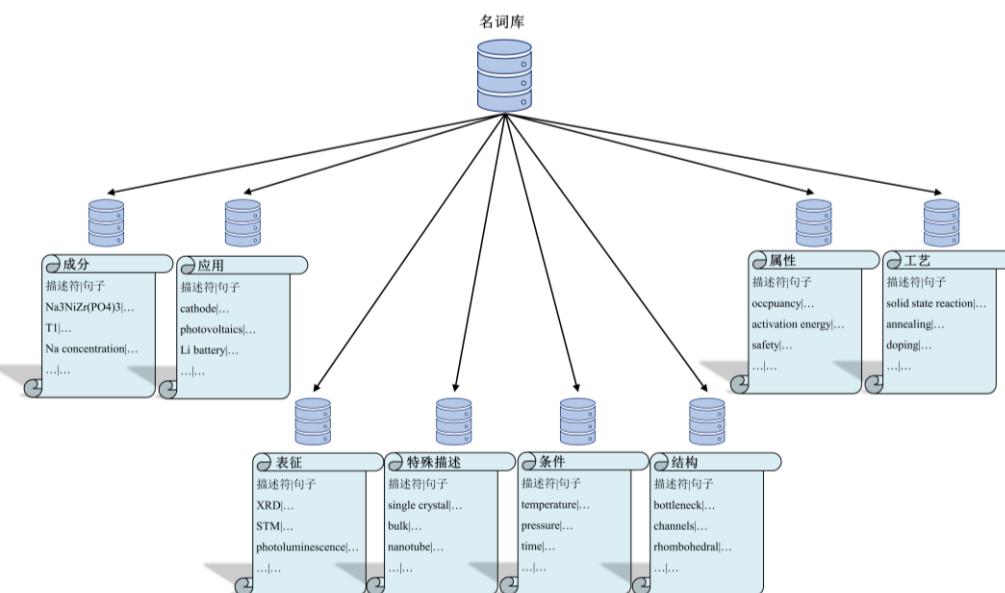


图 3.5 名词库结构

本节建立了一个名词库来存储材料命名实体识别模型抽取的材料实体及其对应的句子信息,如图 3.5 所示。该名词库由八个子库组成,对应八不同类别的材料实体。每个子库中存储的内容是相应类型的材料实体及其出现的句子信息。其中,“activation energy”、“occupancy”和“safety”等属于属性类别的材料实体,而“conduction channels”、“bottleneck”和“rhombohedral symmetry”等属于结构类别的材料实体,每个材料实体后面都有相应的句子列表。此外,本节

还实现了动态添加材料实体及相应的句子到名词库中。

3.4.2 基于重要度计算的描述符筛选

本节提出基于重要度计算的描述符筛选策略，其实现方法和工作流程如图 3.6 所示。首先，通过性能驱动确定所研究的目标材料性能，即特定材料性能，并从名词库中筛选出与其共同出现在同一个句子中的描述符；其次，通过重要度计算策略尽可能地从上述结果中筛选出与特定属性相关的高质量描述符；最后，筛选的描述符可以辅助材料学家以实现材料性能预测或构效关系的研究。

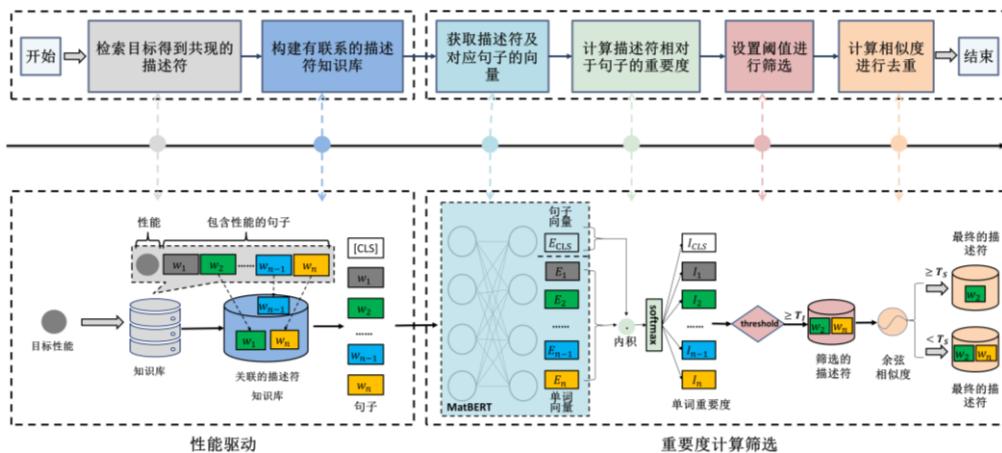


图 3.6 描述符筛选的实现方法和工作流程

为了筛选与特定材料性能预测相关的描述符，本节首先设计了一个性能驱动的策略，如公式 (3.15) 所示。

$$R = \{< D_i, S_i > | (S_i \in KB) \vee ((D_T, D_i) \in S_i) \vee (D_T \neq D_i)\} \quad (3.15)$$

其中， R 表示性能驱动策略筛选的描述符及其对应句子列表的结果集， D_T 表示特定的材料性能， KB 则为存储描述符实体的名词库。 S_i 表示名词库中包含 D_T 的第 i 个句子。

通常情况下，性能驱动的策略需要确定所研究的目标材料的属性，利用这一规则查询名词库，可以筛选出与之共同出现的描述符和相应的句子。然而，通过分析性能驱动的筛选结果发现，共现的描述符并不一定是相关的，因而仅使用共现规则筛选会影响描述符的质量。

为了解决上述问题，本节进一步设计了重要度计算的筛选策略，即通过计算当前描述符相对于句子的重要度进行筛选，如图 3.6 所示。为了计算描述符相对

于其句子中的重要度，我们将 MatBERT 模型最后一层隐藏层输出的单词和句子向量取出，然后将每个单词向量分别与句子向量做内积得到每个词的初始重要度，最后用余弦相似度计算重要度并进行归一化得到最终的重要度。归一化的计算如 (3.16) 所示。需要注意的是，在此使用的 MatBERT 模型与 3.2 节 NER 模型里的相同，只是后者不需要输出单词和句子的向量，而是输入给下游模型进行进一步的特征提取。

$$I_i = \frac{E_i \cdot S_{[CLS]}}{\sum_{i=0}^{n+1} E_i \cdot S_{[CLS]}} \quad (3.16)$$

其中， I_i 表示第 i 个单词的重要性， E_i 为 MatBERT 模型输出的第 i 个单词的嵌入向量， $S_{[CLS]}$ 为对应句子的嵌入向量。之后，设置重要度阈值 T_i 来筛选描述符 w_i ，如图 3.6 所示。此外，为了最大程度的筛除掉前者结果中语义相近的描述符，我们计算任意两个描述符的余弦相似度，如果其小于 T_s （预先设置的相似度阈值），则两个描述符均被保留，否则只保留其中一个。基于上述策略，最终可以筛选出相对高质量与特定材料性能相关的描述符。

3.5 实验

3.5.1 实验数据

为了验证 MatBERT-BiLSTM-CRF 模型的效果和鲁棒性，本章使用 NASICON 型固态电解质和无机材料^[31]实体识别数据集进行实验。

- **NASICON 型固态电解质材料数据集：** 其为第二章构建的数据集，包含 8 种实体类别分别为：“Composition”、“Structure”、“Property”、“Processing”、“Feature”、“Application”、“Characterization”、“Condition”。该数据集含 4868 个句子共计 131380 条数据，其训练集和测试集比例为 8:2。

- **无机材料数据集：** 其是手工注释材料科学文献摘要构建的，其包含 7 种实体类别分别为：“inorganic material (MAT)”、“symmetry/phase label (SPL)”、“sample descriptor (DSC)”、“material property (PRO)”、“material application (APL)”、“synthesis method (SMT)”、“characterization method (CMT)”。该数据集包含 5459 个句子共计 142730 条数据。

3.5.2 实验设置

模型的参数设置如表 3.1 所示。本章采用 AdamW^[109]作为优化器对模型的参数进行调优，批量大小和 Epoch 分别设置为 32 和 100，初始学习率设置为 3e-5。最大句子长度设置为 75，小于该长度的句子将会被填充，大于的则需要截断。单词的词嵌入向量维度设置为 768，LSTM 单元的隐藏层状态维度设置为 128。为了有效防止模型在训练过程中出现过拟合，本章采用了 Dropout^[110]和提前停止^[111]，其中前者设置为 0.1，后者设置为 3。

表 3.1 MatBERT-BiLSTM-CRF 模型实验模型参数

参数名称	值
批量大小 (Batch size)	32
迭代周期 (Epoch)	100
单词向量维度 (Word vector dimension)	768
LSTM 单元维度 (LSTM unit dimension)	128
丢包率 (Dropout rate)	0.1
学习率 (Learning rate)	3e-5
优化器 (Optimizer)	AdamW
提前停止耐力值 (Early stopping patience)	3
最大句子长度 (Max sentence length)	75

3.5.3 评价指标

本章将 F1 作为模型总体的评价指标，该指标是精确率 P (Precision) 和召回率 R (Recall) 的调和平均数，其值越大表明模型的性能越好。精确率 P 、召回率 R 及 F1 值的计算如公式 (3.17) ~ (3.19) 所示。

$$P = \frac{TP}{TP + FP} \quad (3.17)$$

$$R = \frac{TP}{TP + FN} \quad (3.18)$$

$$F1 = \frac{2P * R}{P + R} \quad (3.19)$$

其中， TP 为真正例， FP 为假正例， FN 为假负例。精确率 P 为样本中真实类别为正例且预测类别为正例的数目占所有预测为正例的个数的比例，召回率 R 是样本中真实类别为正例且预测类别为正例的数目占所有真正的正例的个数的比例。

3.5.4 实验结果与分析

(1) NER 模型性能验证

图 3.7 为 NER 模型在测试集上的整体实验表现。模型整体的 F1 值为 0.87，与目前 SOTA 实体识别模型的成绩（F1 为 0.92）^[49]相当。其中，SOTA 实体识别的模型是在只有三个实体标签的手工标记的新闻文章上训练和评估的。然而，我们强调不能将上述两个任务直接进行比较，因为材料文本的建模与新闻类文本有很大的不同，前者具有更特殊、复杂的句子结构。此外，本章 NER 任务拥有更多的实体标签。除了“Application”类，我们的模型在其余每个实体类的 F1 得分都超过了 0.80，表明其在识别不同类型的描述符方面表现良好。模型在“Application”类表现不佳是因为其无法从不充足的训练样本中捕获到丰富的特征来进行判别。

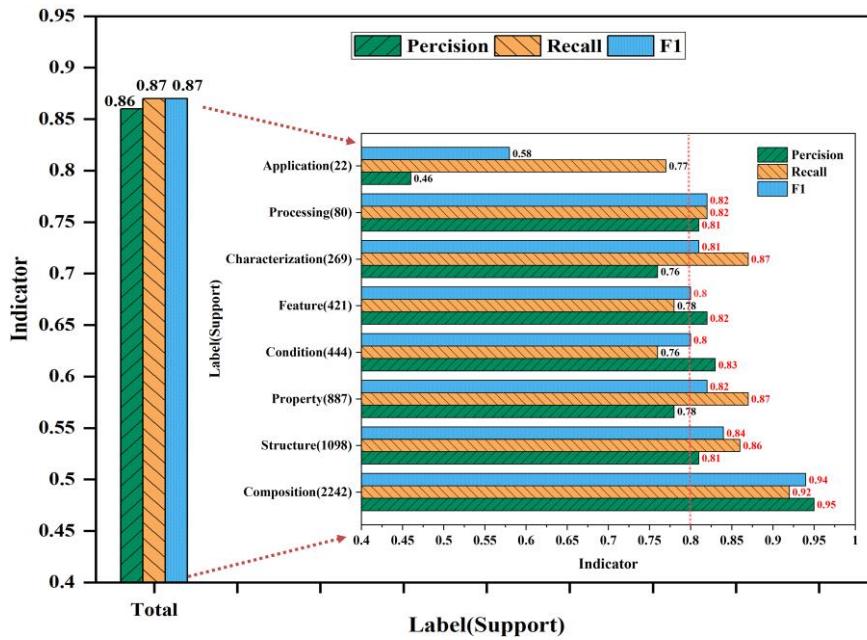


图 3.7 MatBERT-BiLSTM-CRF 模型在测试集上总体精度度量

本节将 MatBERT-BiLSTM-CRF 模型与目前主流的 NER 模型 BiLSTM-CRF^[46]、BiLSTM-CNNs-CRF^[31]和 BERT^[112]进行了比较实验，其结果如图 3.8(a) 所示。从中可以看出，MatBERT-BiLSTM-CRF 模型的 F1 值为 0.87，而上述模型的 F1 值分别为 0.69、0.71 和 0.78。MatBERT-BiLSTM-CRF 模型相较于上述模型分别提高了 18%、16% 和 9%。为了进一步证明 MatBERT 相较于 Word2vec 具

有更能动态捕获复杂材料文本语义信息的能力，同时消除本章模型只能在单一数据集上（2.4 节构建的数据集）表现良好的偏见，我们在无机材料数据集上（800 个手工标记的摘要^[31]）进行了对比实验，结果如表 3.2 所示。其中，MatBERT-BiLSTM-CRF 模型的 F1 值相较于 BiLSTM-CNNs-CRF 提高了 4%。结果表明，MatBERT 的引入可以提取单词的更充分的上下文语义特征，从而能够更好地表示为语义向量；此外，BiLSTM 可以进一步捕获句子序列中单词的局部上下文语义信息。

表 3.2 不同模型在 800 个手工标记的摘要数据集上精度比较

模型	F1 (800 个手工标记的摘要 ^[31])
BiLSTM-CNNs-CRF ^[31]	0.87
MatBERT-BiLSTM-CRF (Ours)	0.91

(2) NER 模型消融实验

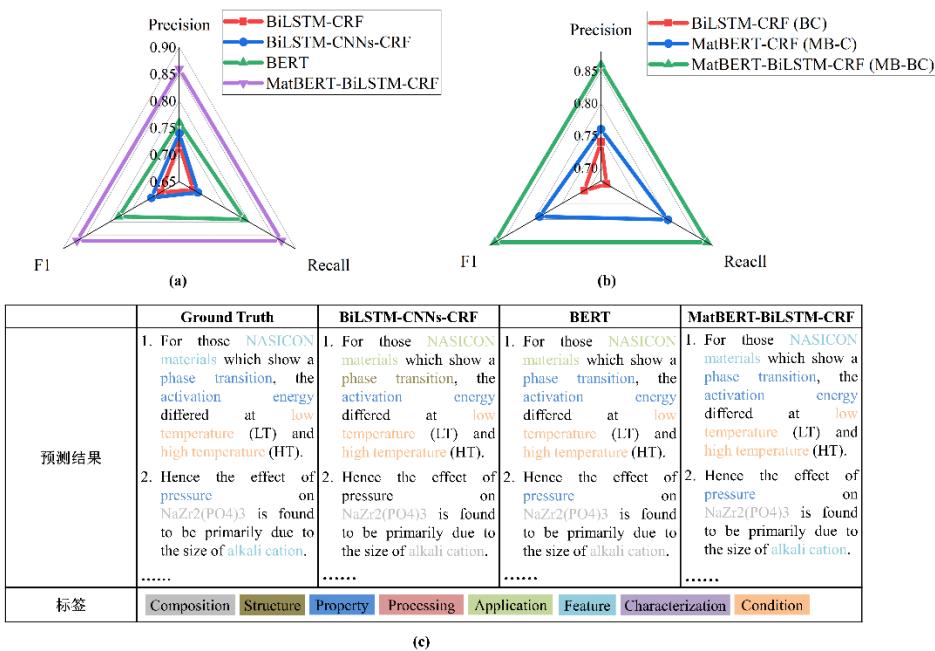


图 3.8 (a) 不同模型的实验结果对比。 (b) 消融实验结果。 (c) 不同模型的实例预测展示，高亮部分表示相应模型与某一描述符实体类型相关的文本区域。

为了验证 MatBERT-BiLSTM-CRF (MB-BC) 模型各组成部分的有效性，本节进一步设计了消融实验。如图 3.8(b)所示，MB-BC 分为两种情况，即 BC (单词的嵌入向量由传统的 Word2vec 训练获得，而不是 MatBERT) 和 MB-C (缺失

BiLSTM 网络) 模型, 分别测试 MatBERT 和 BiLSTM 对 MB-BC 的贡献。基于上述两种情况, MB-BC 的评价指标均受到不同程度的影响。对于 BC 模型, 其 F1 得分降低了 16%, 这表明 MatBERT 可以根据上下文信息动态地生成具有更丰富语义的嵌入向量。而对于 MB-C 模型, 其 F1 得分下降了 8%, 这表明 BiLSTM 可以提高单词的远距离依赖能力, 并捕获更充分的局部上下文语义信息。图 3.8(c) 展示了不同对比模型的实例预测结果。在材料文本 “For those NASICON materials which show a phase transition, the activation energy differed at low temperature (LT) and high temperature (HT)” 中, 可以看出短语 “NASICON materials” 属于 “Feature” 类的描述符实体, 短语 “phase transition” 和 “activation energy” 属于 “Property” 类的描述符实体, 短语 “low temperature” 和 “high temperature” 属于 “Condition” 类的描述符实体。MatBERT-BiLSTM-CRF 模型几乎都能够将其正确的识别出来, 与其它对比模型形成了鲜明对比。例如, 对于 BiLSTM-CNNs-CRF 模型, “NASICON materials” 和 “phase transition” 被分别错误的识别为 “Application” 和 “Structure” 类描述符实体。在材料文本 “Hence the effect of pressure on NaZr₂(PO₄)₃ is found to be primarily due to the size of alkali cation.” 中, 可以看出单词 “pressure” 属于 “Property” 类的描述符实体, “NaZr₂(PO₄)₃” 属于 “Composition” 类的描述符实体, 短语 “alkali cation” 则属于 “Feature” 类的描述符实体。而 “alkali cation” 被比较模型错误的识别为 “Composition” 类, “pressure” 则没有被识别。

(3) MatBERT-BiLSTM-CRF 模型鲁棒性验证

表 3.3 MatBERT-BiLSTM-CRF 模型在原数据集及新数据集上的总体精度度量

标签	F1 (原数据集)	F1 (新数据集)
Application	0.58	0.68
Characterization	0.81	0.82
Composition	0.94	0.93
Condition	0.80	0.83
Feature	0.80	0.78
Processing	0.82	0.80
Property	0.82	0.82
Structure	0.84	0.85
Total	0.87	0.86

为了验证 MatBERT-BiLSTM-CRF 模型的鲁棒性，本节基于额外标注的 35 篇材料文献语料构建测试数据集来进行实验。MatBERT-BiLSTM-CRF 模型在原数据集及新数据集上的实验结果如表 3.3 所示。模型在新数据集上总体 F1 值为 0.86，与原始数据集上的效果相当。此外，新数据集上 F1 值在“Application”和“Condition”类中有了明显的提升，分别为 10% 和 3%。对于其它实体类别，F1 值均在正常范围内波动（ $\pm 3\%$ ），表明模型具有理想的分类性能，即使数据集的分布与训练集不同。综上证明了本章 NER 模型具有良好的鲁棒性。

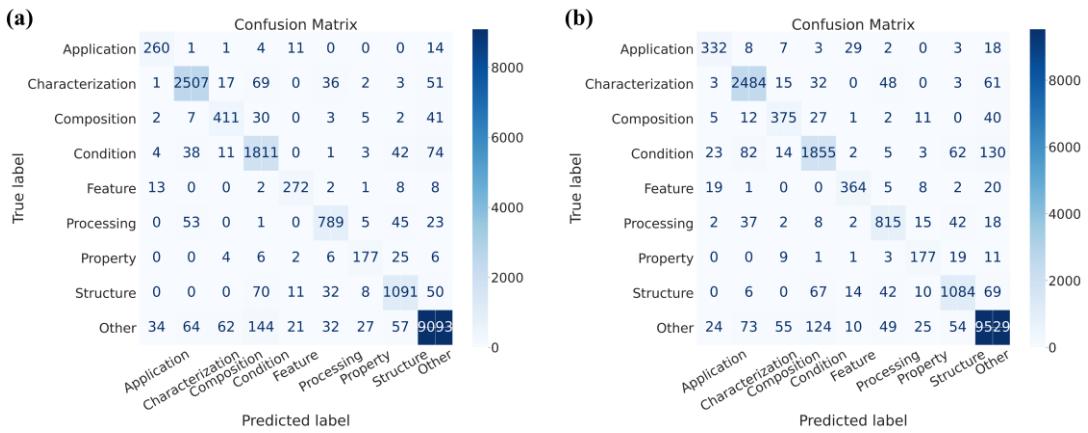


图 3.9 (a) MatBERT-BiLSTM-CRF 模型在初始数据集上的混淆矩阵。(b) MatBERT-BiLSTM-CRF 模型在新数据集上的混淆矩阵。

MatBERT-BiLSTM-CRF 模型在初始和新数据集上的混淆矩阵如图 3.9 所示。对于初始数据集，每个实体类别中所有样本的预测结果几乎都是正确的，其中“Characterization”和“Condition”类预测正确的样本数较多，分别为 2507 和 1811 个；而对于新的数据集，预测正确的样本数较多也是“Characterization”和“Condition”实体类别，分别为 2484 和 1855。值得注意的是，由于本章所关注的是材料实体类别，因此“Other”类预测正确的样本在此不进行分析。然而，初始和新数据集中分别有 70 和 67 个原本属于“Structure”类的样本被模型错误的识别为“Condition”类，这是由于数据集的材料化学式中存在许多元素的键值容易被误判为外部条件值。此外，初始数据集中有 69 个原本属于“Characterization”类的样本被模型错误的识别为“Condition”类，这是由于“Characterization”和“Condition”类在材料文本中具有相似的上下文（即通常作为副词使用），使得在分类过程中混淆了本章的 NER 模型。

3.6 应用

本章将多层语义特征融合的材料命名实体识别模型 MatBERT-BiLSTM-CRF 应用于 NASICON 型固态电解质材料领域。首先，实现了 NASICON 型固态电解质材料实体的抽取存储；在此基础上，从名词库中筛选了激活能相关的描述符，并基于筛选的描述符实现了激活能的预测。

3.6.1 NASICON 型固态电解质材料实体的存储

利用 3.2 节提出的多层语义特征融合的材料命名实体识别模型 MatBERT-BiLSTM-CRFNER 模型，我们从 1808 篇 NASICON 型固态电解质研究相关的文献中抽取了 106896 个不同类别的材料实体信息及其对应的句子，并将其保存在名词库中，不同实体类型的统计信息如图 3.10 所示。

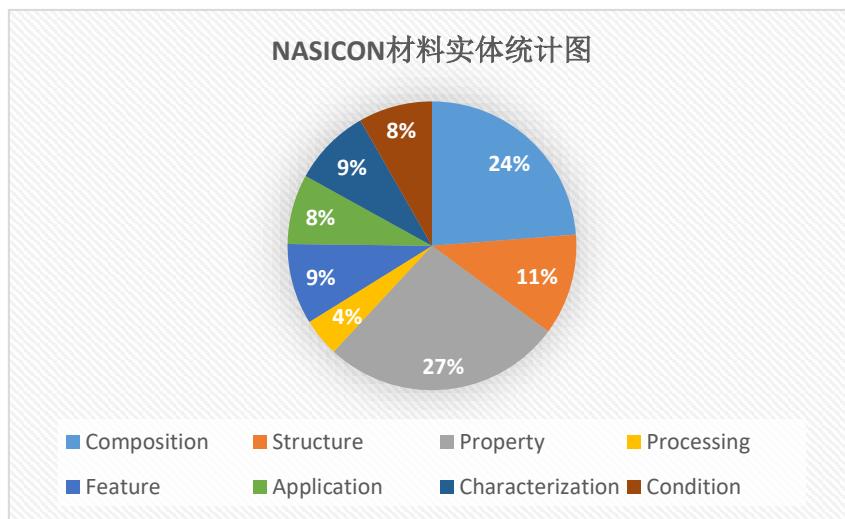


图 3.10 材料命名实体识别模型挖掘的实体信息统计图

3.6.2 NASICON 型固态电解质激活能相关描述符的筛选

我们利用描述符筛选策略从名词库中选取出了 408 个高质量的性能驱动描述符。基于这些筛选出来的描述符，我们利用机器学习模型进行了激活能的预测，详细流程如下：

首先，材料领域专家从不同的角度选择描述符，并基于筛选策略选取的描述

符构建了两份激活能预测的样本数据集。其中，一些专家基于简单分子、结构参数和电子的研究，选取了 31 个与激活能预测相关的描述符特征（如表 3.4 所示）并构建了一份数据集 (*Dataset₃₁*)；另一些则选取包含属性、组成、结构和外部条件等参数的 45 个激活能相关的描述符特征（如表 3.5 所示）并构建了一份数据集 (*Dataset₄₅*)。图 3.11 展示了本文用于激活能预测的部分描述符特征。

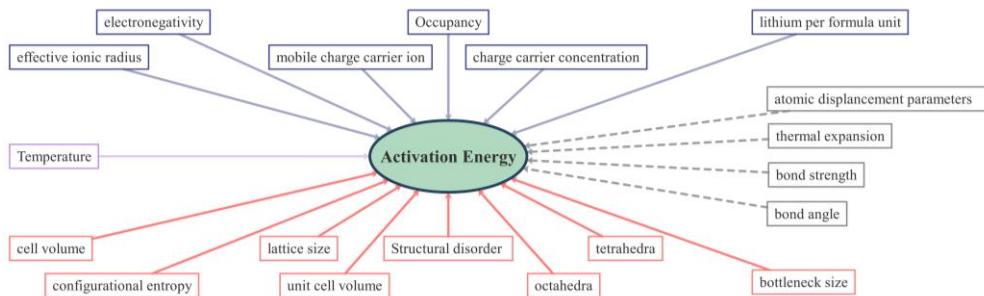


图 3.11 本文选择的用于激活能预测的部分描述符，其中虚线表示仍有待开发的描述符。

从图 3.11 中可以看出，选取的描述符中存在一些可能与预测激活能相关但目前还没有被研究候选的描述符。例如，“bond strength”可能会影响激活能的预测，因为离子输运性能与 M-O 和 X-O 键的成键强度有关。为了实现对激活能的预测，本节从无机晶体结构数据库 (Inorganic Crystal Structure Database, ICSD) 中收集了 85 个描述 NASICON 型固态电解质材料的 CIF 文件。在构建样本数据集的过程中，一些描述符特征值依赖从 CIF 文件中直接读取，而另一些则需通过计算得到。

表 3.4 NASICON 型化合物 31 个描述符的基本含义

序号	名称	描述
1	V_{cell}	晶胞体积
2	a	晶格常数
3	c	晶格常数
4	a/c	晶格常数 a 与 c 之比
5	d	晶胞直径
6	h	晶格 a 平面斜边
7	$D2_{eff}$	M 位上其中一个元素的有效离子半径
8	$D2_{eneg}$	M 位上其中一个元素的电负性
9	$D2_{eneff}$	M 位上其中一个元素的有效电负性
10	$D2_{ionicr}$	M 位上其中一个元素的离子半径
11	$D2_{vol}$	M 位上其中一个元素的有效体积
12	$D2_{volperatom}$	M 位上其中一个元素的体积

13	<i>D2_stoich</i>	<i>M</i> 位上其中一个元素的化学计量数
14	<i>D2_occu</i>	<i>M</i> 位上其中一个元素的占据率
15	<i>D3_eff</i>	<i>M</i> 位剩余元素的平均有效离子半径
16	<i>D3_eneg</i>	<i>M</i> 位剩余元素的电负性
17	<i>D3_eneff</i>	<i>M</i> 位剩余元素的有效电负性
18	<i>D3_ionicr</i>	<i>M</i> 位剩余元素的离子半径
19	<i>D3_vol</i>	<i>M</i> 位剩余元素的有效体积
20	<i>D3_volperatom</i>	<i>M</i> 位剩余元素的体积
21	<i>D3_stoich</i>	<i>M</i> 位剩余元素的化学计量数
22	<i>D3_occu</i>	<i>M</i> 位剩余元素的占据率
23	<i>Na_eff</i>	<i>Na</i> 离子的有效离子半径
24	<i>Na_eneg</i>	<i>Na</i> 离子的电负性
25	<i>Na_eneff</i>	<i>Na</i> 离子的有效电负性
26	<i>Na_ionicr</i>	<i>Na</i> 离子的离子半径
27	<i>Na_vol</i>	<i>Na</i> 离子的有效体积
28	<i>Na_volperatom</i>	<i>Na</i> 离子的体积
29	<i>Na_stoich</i>	<i>Na</i> 离子的化学计量数
30	<i>X1_stoich</i>	<i>X</i> 位上其中一个元素的化学计量数
31	<i>X2_stoich</i>	<i>X</i> 位上剩余元素的化学计量数

表 3.5 NASICON 型化合物 45 个描述符的基本含义

序号	名称	描述
1	O_Na1	Na^+ 在 <i>Na</i> (1)位(Wyckoff 6b)的占据率
2	O_Na2	Na^+ 在 <i>Na</i> (2)位(Wyckoff 18e)的占据率
3	O_Na3	Na^+ 在 <i>Na</i> (3)位(Wyckoff 36f)的占据率
4	C_Na	Na^+ 浓度
5	O_M1	<i>M</i> 位上其中一个元素的占据率(元素周期表序数靠前)
6	O_M2	<i>M</i> 位上剩余元素的占据率(元组周期表序数靠后或 0)
7	EN_M1	<i>M</i> 位上其中一个元素的电负性(元素周期表序数靠前)
8	EN_M2	<i>M</i> 位上剩余元素的电负性(元组周期表序数靠后或 0)
9	Avg_M_EN	<i>M</i> 位的平均有效电负性
10	R_M1	<i>M</i> 位上其中一个元素的离子半径(Å)(元素周期表序数靠前)
11	R_M2	<i>M</i> 位上剩余元素的离子半径(Å)(元组周期表序数靠后或 0)
12	Avg_M_R	<i>M</i> 位的平均有效离子半径(Å)
13	V_M1	<i>M</i> 位上其中一个元素的价态(元素周期表序数靠前)
14	V_M2	<i>M</i> 位上剩余元素的价态(元组周期表序数靠后或 0)
15	Avg_M_V	<i>M</i> 位的平均有效价态
16	O_X1	<i>X</i> 位上其中一个元素的占据率(元素周期表序数靠前)
17	O_X2	<i>X</i> 位上剩余元素的占据率(元组周期表序数靠后或 0)
18	EN_X1	<i>X</i> 位上其中一个元素的电负性(元素周期表序数靠前)
19	EN_X2	<i>X</i> 位上剩余元素的电负性(元组周期表序数靠后或 0)
20	Avg_X_EN	<i>X</i> 位的平均有效电负性
21	R_X1	<i>X</i> 位上其中一个元素的离子半径(Å)(元素周期表序数靠前)

22	R_X2	X位上剩余元素的离子半径(Å)(元组周期表序数靠后或0)
23	Avg_X_R	X位的平均有效离子半径(Å)
24	V_X1	X位上其中一个元素的价态(元素周期表序数靠前)
25	V_X2	X位上剩余元素的价态(元组周期表序数靠后或0)
26	Avg_X_V	X位的平均有效价态
27	a	晶格常数
28	c	晶格常数
29	V	晶胞体积
30	V_MO ₆	MO ₆ 多面体体积
31	V_XO ₄	XO ₄ 多面体体积
32	V_Na(1)O ₆	Na(1)O ₆ 多面体体积
33	V_Na(2)O ₈	Na(2)O ₈ 多面体体积
34	V_Na(3)O ₅	Na(3)O ₅ 多面体体积
35	BT2	Na ⁺ 从Na(1)位跃迁到Na(3)位的最小瓶颈
36	BT1	Na ⁺ 从Na(1)位跃迁到Na(2)位的最小瓶颈
37	Min_BT	BT1 和 BT2 的最小值
38	RT	可自由通过由骨架离子组成的空隙空间的最大球形探测器 的半径
39	E_Na(1)	Na ⁺ 在Na(1)位的构型熵
40	E_Na(2)	Na ⁺ 在Na(2)位的构型熵
41	E_Na(3)	Na ⁺ 在Na(3)位的构型熵
42	E_Na	Na ⁺ 的构型熵
43	E_M	M位阳离子的构型熵
44	E_X	X位阳离子的构型熵
45	T	测量结构的实验温度

其次，构建机器学习模型进行激活能的预测实验，详细的实验设置如下。85个NASICON 样本分为训练集和测试集，通过 10 折交叉验证来评估 6 个候选预测模型（LASSO、GPR、Ridge、SVR、KNN 和 RF）对样本数据集的泛化能力。候选模型对数据集的总体表现是在所有 10 次迭代中的均方根误差（RMSE）、平均绝对百分比误差（MAPE）及 R 平方（R²）的平均值。RMSE、MAPE 和 R² 的计算公式如（3.20）~（3.22）所示。其中， y_i 表示第 i 条样本激活能的真实值， \hat{y}_i 表示模型对第 i 条样本激活能的预测值， n 表示样本的总数，RMSE 和 MAPE 的值越小表明模型拟合的越好，相反 R² 的值越大表明模型拟合的越好。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3.20)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (3.21)$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \quad (3.22)$$

最后，6个候选机器学习模型的实验结果如表3.6所示。在Dataset₄₅上训练最好的GPR模型相较于在Dataset₃₁上训练最好的GPR模型获得了更好的性能，其中，RMSE、MAPE及R²分别取得了0.035、0.022和0.087的进步。这是因为Dataset₄₅比Dataset₃₁考虑了更加全面的描述符，如“configuration entropy”、“bottleneck”等。此外，在上述两个数据集上，所有机器学习模型的激活能预测准确率(R²)均在80%以上。结果表明，本文提出的多层次语义特征融合的材料命名实体识别方法能够客观有效的识别并筛选出高质量描述符。

表3.6 不同机器学习模型在Dataset₃₁和Dataset₄₅上的实验结果比较

模型	Dataset ₃₁			Dataset ₄₅		
	RMSE	MAPE	R ²	RMSE	MAPE	R ²
LASSO	0.092	0.062	0.856	0.065	0.041	0.928
GPR	0.092	0.061	0.858	0.057	0.039	0.945
Ridge	0.093	0.060	0.851	0.058	0.040	0.943
SVR	0.096	0.069	0.845	0.070	0.060	0.915
KNN	0.109	0.075	0.800	0.100	0.066	0.830
RF	0.101	0.065	0.826	0.060	0.039	0.939

3.7 小结

本章主要研究了基于多层次语义特征融合的材料命名实体识别方法。首先阐述了材料命名实体识别的研究现状及其存在的问题；然后针对材料特殊文本语义特征难以被已有的命名实体识别模型充分融合的问题，提出了多层次语义特征融合的材料命名实体识别方法MatBERT-BiLSTM-CRF以实现实体的抽取。该方法通过构建MatBERT来充分提取词级别的语义特征，引入BiLSTM模型以捕获词的局部上下文语义信息，利用CRF实现实体的精准分类。实验证明，本章提出的方法在不同材料数据集上都达到了理想的识别效果。进一步，本章设计了描述符筛选策略以筛选特定材料性能相关的高质量描述符，并以NASICON型固态电解质激活能为例，证明了所筛选的描述符可以有效构建激活能预测的机器学习样本集。

第四章 基于实体感知的材料关系抽取方法

关系抽取 (Relation Extraction, RE) 可以从文献中自动挖掘出“(主体, 关系, 客体)”形式的实体关系三元组信息, 且在材料领域已取得初步成效。然而, 材料文本中的关系十分复杂, 句子中存在多种重叠关系, 使得材料目标实体及其边界语义信息难以被现有 RE 方法感知, 从而影响其分类准确性。因此, 本章针对上述问题展开材料 RE 方法的研究。首先, 介绍目前材料 RE 任务研究现状及存在的问题; 其次, 提出一种实体感知的材料 RE 模型, 并详细叙述所包含的关键技术; 再次, 在 NASICON 型固态电解质和电池材料关系抽取数据集上进行对比及消融实验来验证模型的有效性; 最后, 以 NASICON 型固态电解质文献为例进行材料实体关系三元组的抽取, 并构建知识图谱对其进行存储。在此基础上, 构建描述符树并对其填充以获取 NASICON 型固态电解质构效关系知识, 进一步利用知识嵌入的机器学习对样本进行特征选择以验证所获知识的有效性。

4.1 问题描述与分析

材料 RE 方法旨在标识和分类材料文本中提及的实体对间的关系, 其不仅能够实现将实体与其属性相关联, 还可以建立实体间的共现关系。在材料科学的研究中, 大多数关系抽取方法都是以依赖解析方式进行的^[52, 96, 97, 113]。这些方法基于远程监督的思想, 通过利用已有句子中一个实体与其它实体的关系来表示句子结构, 进而推广到更多的句子以得到实体间的关系。

预定义规则和 Snowball 算法是材料领域利用依赖解析进行关系抽取的常用方法。例如, Hawizy 等人^[96]将 Snowball 算法引入材料领域进行材料化学实体间关系抽取的研究, 首先为其提供已知的具有正相关关系的种子例子, 其次利用这些种子示例定位句子, 在此基础上, 利用文本相似性聚类来学习典型模式, 并将新句子和学习到的模式进行比较, 最后根据最低相似度的阈值来识别新的关系。Court 等人^[113]通过对 Snowball 算法进行改进来研究磁性材料化合物及其相关居里和尼尔磁相变温度等信息的抽取, 他们将 Snowball 算法的原始二元关系提取功能扩展到四元关系提取, 基于此进行远程监督四元关系的提取, 最终自动生成

了一个磁性材料数据库。Kuniyoshi 等人^[52]利用预定义规则进行材料合成过程中关系信息的抽取，并开发了一个基于深度学习序列标记器和简单的启发式规则关系抽取器的自动机器阅读系统，最终实现从文献中自动提取材料合成过程。Wang 等人^[97]基于远程监督思想进行材料化学成分及其性质间的关系抽取，首先搜集文献构建实体数据集并定义关系规则，然后分别设计基于 BiLSTM-CRF 的 NER 模型和基于 Snowball 的远程监督 RE 模型，最后通过对模型的训练以实现材料文献中化学成分和性质等数据的自动抽取。上述方法在一定程度上均取得了不错的效果，然而基于 Snowball 算法和规则的 RE 方法在进行关系抽取时人工干预较大，且抽取结果中负例较多，因此存在很大的噪声问题。

随着深度学习的不断发展，基于深度学习的 RE 方法逐渐被研究者重视。该类方法不需要太多人工干预且可以有效缓解基于依赖解析方式的材料 RE 方法存在的噪声问题^[114]。然而，材料文本中的关系十分复杂，语句存在大量关系重叠问题，例如，在句子“Increasing the sintering temperature causes the lattice parameters and the unit volume to increase”中，“temperature”作为材料实体既与“lattice parameters”存在影响关系，又与“unit volume”存在影响关系。受限于材料文本的复杂特性，已有的通用领域深度学习 RE 方法难以直接迁移到材料领域进行有效应用。此外，当目标实体具有重叠关系时，已有方法难以充分感知实体及其边界语义信息，因而影响着深度学习模型对材料实体间关系的分类准确性。

综上所述，材料领域仍缺乏有效的深度学习 RE 模型来自动从文献中挖掘实体关系信息。因此，本章提出实体感知的材料关系抽取模型 MatBERT-BiGRU-Softmax。该模型首先用特殊的封闭标记“[]”及“{}”包裹目标实体词，使得 MatBERT 模型能清晰地判别目标实体的边界，从而可以充分感知并提取更丰富的目标实体及句子的语义信息；其次，引入 BiGRU 模型对句子序列进行建模，以捕获目标实体的局部上下文语义信息；再次，利用 Softmax 函数实现材料实体关系的精准分类。

4.2 方法概述

本章研究了面向材料领域的材料关系抽取方法，提出了基于实体感知的材料

关系抽取方法 MatBERT-BiGRU-Softmax。该方法包括目标实体驱动的实体感知、基于实体感知的语义特征提取和基于 *Softmax* 的材料实体关系分类，以快速抽取材料实体关系三元组信息。进一步，提出了基于实体关系的材料知识图谱构建和知识获取方法，包括基于 Neo4j 图数据库的材料知识图谱的构建、基于材料知识图谱的描述符树的构建和基于描述符树的知识获取，以快速获取材料知识并实现材料知识嵌入的性能预测。整个方法以管道的形式进行了材料关系抽取模型构建与应用，具体流程如图 4.1 所示。

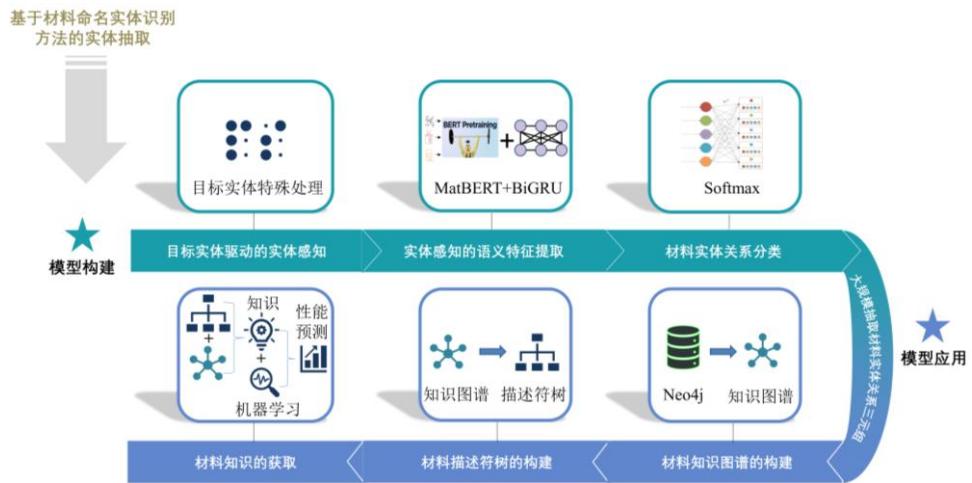


图 4.1 实体感知的材料关系抽取方法及应用流程

4.3 基于实体感知的材料关系抽取

图 4.2 展示了实体感知的材料关系抽取模型 MatBERT-BiGRU-Softmax 的结构，由目标实体驱动的实体感知、基于实体感知的语义特征提取、基于 *Softmax* 的材料实体关系分类三部分组成。在实体感知阶段，通过设计特殊标记 “[]” 和 “{}” 对两个目标实体词进行包裹，使得模型能清晰地感知目标实体及其边界信息，并以此作为输入属性交付给下一阶段；在语义特征提取阶段，首先 MatBERT 用于提取句子级别语义特征和包含句子嵌入、单词嵌入及位置嵌入的单词级别语义特征，然后进一步利用 BiGRU 对句子序列建模以提取句子及目标实体的局部上下文语义特征；在实体关系分类阶段，通过全连接操作拼接句子及目标实体的特征向量，并利用 *Softmax* 函数计算得到候选关系中概率最大的一个来实现关系分类。

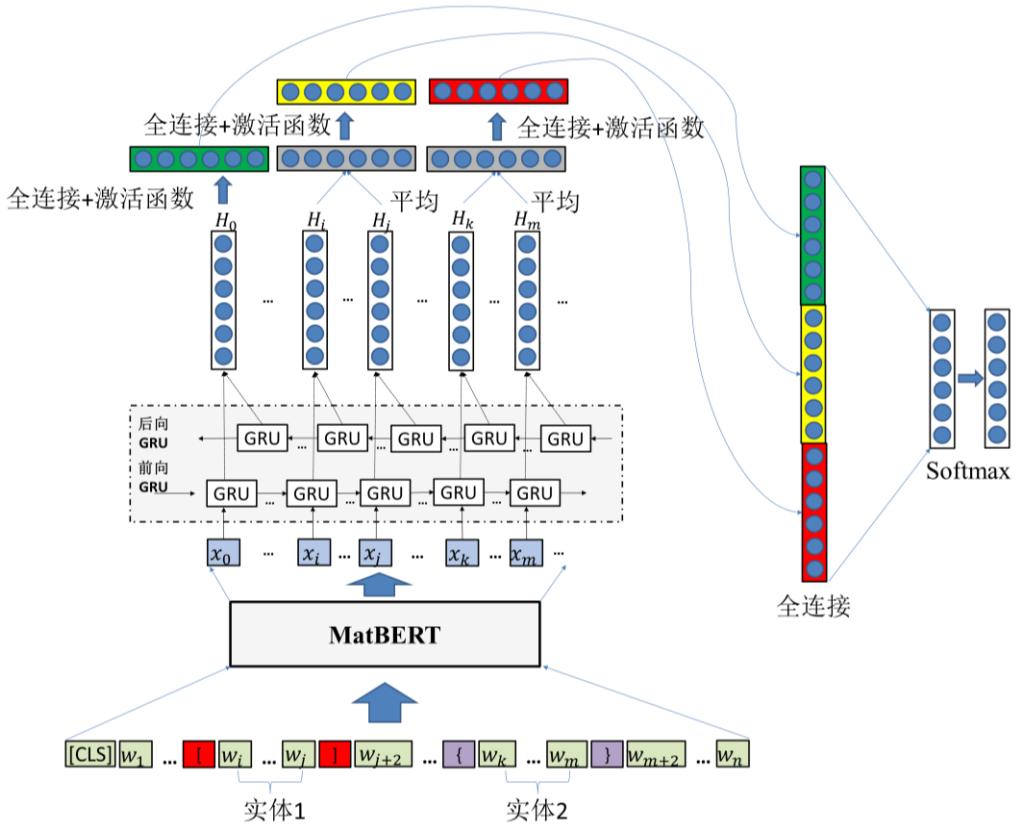


图 4.2 实体感知的材料实体关系抽取模型结构图

4.3.1 目标实体驱动的实体感知

为了使预训练 MatBERT 语言模型能清晰地判别目标实体的边界，从而可以在捕获句子序列中每个单词的上下文语义信息的同时能更加关注目标实体的位置，进而获得更加丰富的目标实体的语义信息，模型的输入层对数据集中的句子进行了如图 4.3 所示的处理。首先，给定关系抽取数据集中包含两个目标实体 e_1 (Subject) 和 e_2 (Object) 的一个句子 (Sentence)，在第一个实体的开始和结束位置分别插入特殊标记 “[” 和 “]”，在第二个实体的开始和结束位置分别插入特殊标记 “{” 和 “}”。其次，为了便于 MatBERT 语言模型对句子信息进行编码，在句子的起始位置添加特殊标记 “[CLS]” 对整个句子的语义信息进行表征，同时在句子的结尾处添加特殊标记 “[SEP]” 以表示当前句子已经结束，以便 MatBERT 语言模型执行句子级别的训练任务。最后，需要将句子序列表示为模型可以接受的向量形式，将句子中的每个单词替换为其在词汇表中出现的序号。

(id)，同时将句子设置为固定的长度格式（对于不满足长度的句子进行切断或者填充处理）。例如，给定数据集中带有目标实体“pressure”和“NaZr₂(PO₄)₃”的一个句子“Hence the effect of pressure on NaZr₂(PO₄)₃ is found to be primarily due to the size of alkali cation.”，两个目标实体“pressure”和“NaZr₂(PO₄)₃”在插入特殊的标记后，句子将变为“Hence the effect of [pressure] on { NaZr₂(PO₄)₃ } is found to be primarily due to the size of alkali cation.”，之后再由输入层对其进行填充或截断以及向量化处理。

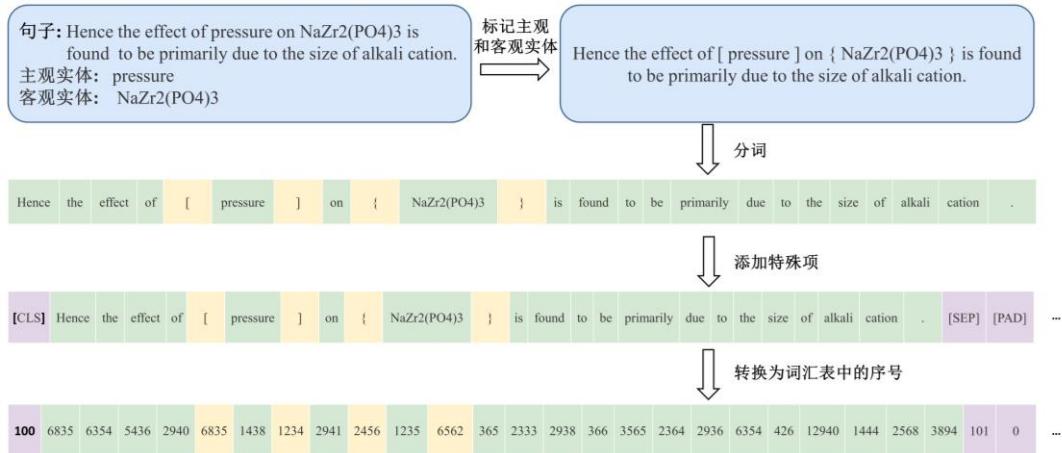


图 4.3 输入数据进行实体感知处理

4.3.2 基于实体感知的语义特征提取

如图 4.2 所示，包含丰富语义信息的句子特征向量 (\mathbf{H}_0) 及目标实体特征向量 ($\mathbf{H}_i \sim \mathbf{H}_j, \mathbf{H}_k \sim \mathbf{H}_m$) 的表示主要由语义特征提取层的 MatBERT 和 BiGRU 模型感知并生成。具体地，给定一个带有目标实体 e_1 和 e_2 的句子 $s = ([CLS], w_1, w_2, \dots, w_n)$ ，其中，[CLS] 表示样本句子序列的开始， w_t 表示句子中的第 t 个单词。在此，假设目标实体 e_1 由单词 $w_i \sim w_j$ 组成， e_2 由单词 $w_k \sim w_m$ 组成。MatBERT 通过同时结合单词嵌入、句子嵌入及位置嵌入信息来编码句子中每个单词的向量并作为其训练的输入向量。之后，通过无监督的训练学习得到其最后隐藏层输出的特征向量表征，为 $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ 。其中 \mathbf{x}_t 表示 MatBERT 模型学到的第 t 个单词的向量表示。

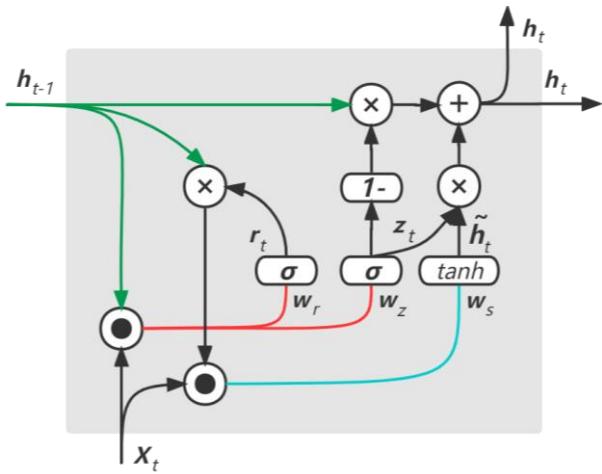


图 4.4 GRU 单元结构

尽管 MatBERT 模型可以捕获句子序列中单词及句子级别的语义特征，但这些语义特征在其各个内部层之间的传递过程中会不可避免地丢失部分特征，尤其是与关系抽取任务密切相关的位置特征信息^[48]。因此，为了弥补上述语义特征信息的丢失，提高模型捕获句子和目标实体局部上下文语义信息的能力，本节使用 BiGRU 来对句子序列建模以捕获局部上下文特征信息。与 RNN 对句子序列建模相比，GRU 结构单元设置了更新门和重置门，门控单元的结构可以选择要保存或者丢掉的上下文语义信息，从而可以缓解 RNN 梯度消失或梯度爆炸的问题。此外，在性能等同于 LSTM 的情况下，GRU 的结构比 LSTM 更简单，训练速度更快。GRU 单元的结构如图 4.4 所示。

对于时间 t，GRU 单元状态计算如公式 (4.1) ~ (4.4) 所示：

$$r_t = \sigma(w_r \cdot [h_{t-1}, x_t] + b_r) \quad (4.1)$$

$$z_t = \sigma(w_z \cdot [h_{t-1}, x_t] + b_z) \quad (4.2)$$

$$\tilde{h}_t = \tanh(w_h \cdot [r_t * h_{t-1}, x_t] + b_h) \quad (4.3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (4.4)$$

其中， σ 和 \tanh 表示不同的激活函数， \cdot 表示点积。 w_r , w_z , w_h 是权重矩阵， b_r , b_z , b_h 是参数偏置值。 x_t 是 t 时刻的输入向量， h_t 是隐藏层状态，也是输出向量，其包含了 t 时刻之前的全部有效信息。 z_t 是更新门，用于控制前一个单元的隐藏层输出对当前单元状态的影响，更新门的值越大，前一个单元的隐藏层输出对当前单元状态的影响就越大。 r_t 是重置门，用于控制 h_{t-1} 对 \tilde{h}_t 的重要性，重置门的

值越小，上一单元的隐藏层信息被忽略的程度就越大。 \tilde{h}_t 表示在当前单元中需要更新的信息。上述两个门控单元共同作用使得 GRU 能够捕获句子序列长度的依赖性。

为了使模型能最大程度地捕获当前时刻前后两个方向的语义特征，本节采用由前向 GRU 单元和后向 GRU 单元组成的 BiGRU 网络，前向单元的隐藏层表示为 \vec{h}_t ，后向单元的隐藏层表示为 \overleftarrow{h}_t 。通过公式 (4.1) ~ (4.3)，得到 t 时刻单向隐藏层的输出，如公式 (4.5) ~ (4.6) 所示。BiGRU 的隐藏层输出通过前向 GRU 单元与后向 GRU 单元的隐藏层输出拼接得到，如公式 (4.7) 所示。

$$\vec{h}_t = GRU(x_t, \vec{h}_{t-1}) \quad (4.5)$$

$$\overleftarrow{h}_t = GRU(x_t, \overleftarrow{h}_{t-1}) \quad (4.6)$$

$$h_t = [\vec{h}_t, \overleftarrow{h}_t] \quad (4.7)$$

4.3.3 基于Softmax的材料实体关系分类

通过语义特征提取层的 MatBERT 和 BiGRU 模型对句子及单词语义特征进行提取，得到了包含丰富语义信息的句子及单词的向量表示 $\mathbf{h} = (\mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n)$ 。其中， h_t 表示 t 时刻 BiGRU 单元隐藏层的输出。为了对给定句子中的目标实体间的关系进行分类，本节首先取出两个目标实体 (e_1 和 e_2) 及句子的 BiGRU 隐藏层输出向量。其中， $\mathbf{H}_i \sim \mathbf{H}_j$ 和 $\mathbf{H}_k \sim \mathbf{H}_m$ 分别是实体 e_1 及 e_2 的 BiGRU 最终隐藏状态向量， \mathbf{h}_0 为整个句子的 BiGRU 最终隐藏状态向量。其次分别对 $\mathbf{H}_i \sim \mathbf{H}_j$ 和 $\mathbf{H}_k \sim \mathbf{H}_m$ 取平均来得到两个目标实体的向量表示，在对其进行激活函数操作后，在两个向量上再各加一个全连接层，从而得到两个目标实体最后的输出向量 \mathbf{H}'_1 和 \mathbf{H}'_2 。该计算过程如公式 (4.8) ~ (4.9) 所示。

$$H'_1 = W_1 \left[\tanh \left(\frac{1}{j-i+1} \sum_{t=i}^n H_t \right) \right] + b_1 \quad (4.8)$$

$$H'_2 = W_2 \left[\tanh \left(\frac{1}{m-k+1} \sum_{t=k}^m H_t \right) \right] + b_2 \quad (4.9)$$

其中， W_1 、 W_2 和 b_1 、 b_2 共享相同的参数，也即， $W_1=W_2$ ， $b_1=b_2$ 。对于第一项

的最终隐藏状态向量（即“[CLS]”），本节同样添加了一个激活函数及全连接的操作，如公式（4.10）所示。

$$H'_0 = W_0[\tanh(H_0)] + b_0 \quad (4.10)$$

其中，矩阵 W_0 、 W_1 、 W_2 有相同的维度，即 $W_0 \in R^{d \times d}$ 、 $W_1 \in R^{d \times d}$ 、 $W_2 \in R^{d \times d}$ ，其中 d 为 MatBERT 隐藏状态的大小。

最后，将 H'_0 、 H'_1 、 H'_2 拼接，通过一个全连接层和一个 *Softmax* 层，实现最终的关系分类，如公式（4.11）~（4.12）所示。

$$h'' = W_3[\text{concat}(H'_0, H'_1, H'_2)] + b_3 \quad (4.11)$$

$$p = \text{Softmax}(h'') \quad (4.12)$$

其中， $W_3 \in R^{L \times 3d}$ (L 是关系类型的数量)， p 为最终的概率输出，在公式（4.8）~（4.12）中 b_0 、 b_1 、 b_2 、 b_3 均是偏置向量。

4.4 基于实体关系的材料知识图谱构建与知识获取

为了实现实体感知的材料关系抽取方法在材料领域的应用，本章进一步提出了基于实体关系的材料知识图谱构建和知识获取方法，包括基于 Neo4j 图数据库的材料知识图谱的构建、基于材料知识图谱的描述符树的构建和基于描述符树的知识获取。其中，Neo4j 图数据库用于存储模型抽取的材料实体关系三元组信息和构建材料知识图谱；在此基础上，构建材料描述符树并对其进行填充；最后，通过描述符树与知识图谱结合以推理获得材料知识，同时对其进行表示，并通过材料知识嵌入特征选择方法实现对知识的验证。

4.4.1 基于 Neo4j 图数据库的材料知识图谱构建

本节基于 Neo4j 图数据库^[115]实现材料实体关系三元组信息的存储。Neo4j 是一个基于 Java 语言实现的开源 NoSQL 图数据库，其架构旨在优化节点和关系的快速管理、存储和遍历过程。在 Neo4j 中，关系是图数据库中最重要的元素，它表示节点之间的互连，即由一个节点指向另一个节点。Neo4j 中只有两种数据类型：节点和边。节点用于保存材料实体，边来用于连接节点以表示材料实体间的关系。本节通过 Python 的 py2neo 工具包对 Neo4j 图数据库进行读写操作以实现

材料实体关系三元组的存储。

在此基础上，本节基于 Neo4j 图数据库构建材料知识图谱以便于材料描述符树的建立。知识图谱旨在描述真实世界中的各种实体或概念及其关系，并将上述事实构成一张巨大的语义关系网络图，其节点可由实体或概念填充，边则由关系或属性填充。知识图谱在逻辑结构上由模式层和数据层构成，其中数据层主要存储由一系列事实组成的数据，而知识则以事实为单位进行表达，如用（主体，关系，客体）或（实体，属性，属性值）的三元组来表示事实；模式层则构建在数据层之上，主要通过本体库来规范数据层的一系列事实表达。考虑到材料领域知识的特点，本节提出了材料领域知识图谱的特殊性原则（Material-Domain Knowledge Graph of Particularity, MatKGPtcl），如公式（4.13）所示。

$$MatKGPtcl = \langle F_{kno}, O_{kno}, U_{kno} \rangle \quad (4.13)$$

其中， F_{kno} 为知识的表现形式，即材料描述符实体及其间的关系（即描述符实体关系三元组），主要由实体识别到关系抽取的流程获取； O_{kno} 为知识在图谱中的组织方式，即自上而下的自动化构建方式，用于设定材料知识图谱的逻辑结构； U_{kno} 为知识的使用需求，即期望借助描述符逻辑关系图谱的推理能力，发现更多显性及隐性知识，为材料领域研究提供可视化知识显示，进而为材料知识学习方法和模型提供形式化知识支撑。

基于 MatKGPtcl 原则，并在 Neo4j 图数据库的驱动下，最终可以实现目标材料性能驱动的描述符逻辑关系图谱的构建。具体地，首先采取自上（模式层）而下（数据层）的方式对其逻辑结构进行设定；其次，设置目标材料性能并填充至模式层，同时将基于命名实体识别模型抽取的成分、结构、工艺、性能、描述、应用、条件和表征等抽象的描述符实体也填充至模式层；最后，将关系抽取模型挖掘出的描述符实体关系三元组进行填充至数据层。需要注意的是，数据层存储的描述符实体不仅受到模式层的约束，而且还必须和目标材料性能有一定的关系。

4.4.2 基于材料知识图谱的描述符树建立

为了获取材料知识图谱概念层中描述符间的层级关系，从而形式化表示特定材料性能相关的描述符，使其易读、易维护、易复用，本节基于材料图谱构建了

描述符树。描述符树的形式化表示如公式（4.14）所示：

$$\text{Descriptors Tree}(N, E, H, S) \quad (4.14)$$

其中， N 代表节点（Node）， $N = \{N_1, N_2, \dots, N_n\}, n \in [1, m]$ ，用于放置描述符； E 代表边（Edge）， $E = \{E_1, E_2, \dots, E_n\}, n \in [1, m]$ ，边映射了节点之间的关系； H 代表层级（Hierarchy）， $H = \{H_1, H_2, \dots, H_n\}, n \in [1, m]$ ，用于表示描述符的抽象程度，层级越高，概念越抽象； S 代表属性（Slot）， $S = \{S_1, S_2, \dots, S_n\}, n \in [1, m]$ ，用于保存节点或边的其他信息，属性中存放的对象可以根据研究人员的需求进行自定义。最终构建的描述符树的通用形式如图 4.5 所示。

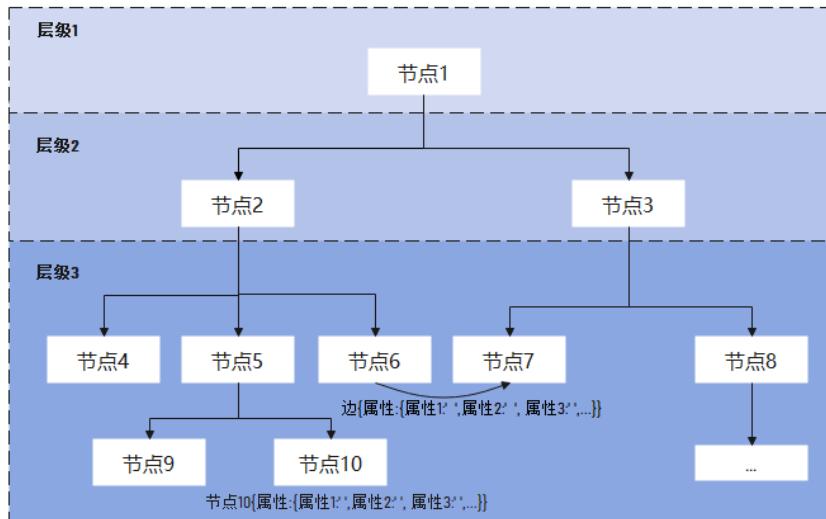


图 4.5 描述符树的通用形式

在此基础上，本节设计了材料描述符树的填充算法。具体步骤如下：

步骤 1：将描述符树整体设计为 3 层。其中，粗粒度层（层级 1）由目标材料性能及其影响因素的抽象类别概念组成，细粒度层（层级 2）由不同类别的参数信息组成，而概念层（层级 3）则由具体的描述符信息组成；

步骤 2：设置不同层的约束关系。概念层、细粒度层和粗粒度层依次受到上层信息的约束，即粗粒度层能够划分出不同类别细粒度层，细粒度层又可衍生出不同的概念层；

步骤 3：设置不同层填充信息的方式。粗粒度层和细粒度层是在领域专家经验知识的指导下人为设定填充的，概念层则由知识图谱检索与推理得到的描述符信息填充。

材料描述符树填充算法的伪代码如算法 4.1 所示。其中， $R =$

$\{R_{ce}, R_{cw}, R_{fo}, R_{lo}, R_{io}, R_{co}, R_{mo}\}$ 表示关系集合，相应内容分别为“Cause-Effect”、“Component-Whole”、“Feature-Of”、“Located-Of”、“Instance-Of”、“Condition-On” 和 “Method-Of” 关系类型。 $Prop_{obj}$ 表示粗粒度层填充的目标材料性能， $Class_1, \dots, Class_i$ 表示粗粒度层填充的影响 $Prop_{obj}$ 的抽象类别概念， $P_{class_1}, \dots, P_{class_i}$ 表示细粒度层填充的不同抽象类别概念下的参数信息。 $r(a, b)$ 表示 a 和 b 之间的关系。

算法 4.1: 材料描述符树填充

输入: 未填充的描述符树 $Unfiled - D_{tree}$; 材料知识图谱 KG_{mat} ;

关系集 $R = \{R_{ce}, R_{cw}, R_{fo}, R_{lo}, R_{io}, R_{co}, R_{mo}\}$; 填充的描述符树 $Filed - D_{tree} = \{\}$ 。

输出: $Filed - D_{tree}$

1: 开始

2: 初始化 $Filed - D_{tree}$

3: $Filed - D_{tree} \leftarrow \{H_1: \{\}, H_2: \{\}, H_3: []\}$ // 对描述符树的层级进行设定

4: $H_1 \leftarrow \{"Prop_{obj}": [Class_1, \dots, Class_i]\}$ // 对 H_1 进行填充

5: $H_2 \leftarrow \{P_{class_1}: [], \dots, P_{class_i}: []\}$ // 对 H_2 进行填充

6: **foreach** $D_m \in KG_{mat}$ **do**

7: **if** $r(D_m, Prop_{obj}) = R_{ce} \wedge (D_m \in P_{class_i})$:

8: $H_{3i} \leftarrow [D_m]$ // 对 H_3 的第 i 个元素进行填充

9: $P_{class_i} = P_{class_i} \cup H_{3i}$ // 建立 D_i 与 P_{class_i} 的层级关系

10: **foreach** $D_n \in KG_{mat}$ **do**

11: **if** $r(D_n, D_m) \in R \wedge (D_n \in P_{class_j})$:

12: $H_{3j} \leftarrow [D_n]$ // 对 H_3 第 j 个元素进行填充

13: $P_{class_j} = P_{class_j} \cup H_{3j}$ // 建立 D_n 与 P_{class_j} 的层级关系

14: 结束

4.4.3 基于描述符树的知识获取

机器学习不仅要从历史数据中挖掘潜在模式，领域知识对它的指导也是至关

重要的。目前，在利用机器学习研究材料构效关系时，领域知识的应用仅局限于数据集的构建与预处理阶段。本课题组提出材料领域知识嵌入的机器学习^[116]，期望通过对材料领域知识进行符号化表示并嵌入机器学习三要素（模型、策略和算法）中，建立不同机器学习阶段的领域知识嵌入方式，实现材料领域知识在机器学习全流程的有机融入，从而构建高精度且具有一定可解释性的机器学习新模型。其中，材料领域知识的获取与表示是基础和难点。

本文所构建的描述符树中涵盖了大量的领域知识，可以为材料领域知识嵌入的机器学习方法自动提供材料领域知识。然而，描述符树中节点的抽象程度是不同的，越上层节点的概念越抽象，所涵盖的描述符也就越多，越下层节点越具体，所涵盖的描述符也就越可能直接作为描述符特征。为了从描述符树得到材料性能及其影响因素间的构效关系知识，本节设计了材料构效关系知识获取算法。该算法是在描述符树（本文 4.4.2 节建立的）和可溯源处理模型（本文 2.2.2 节提出的）的驱动下完成。具体步骤如下：

步骤 1：获取描述符及其间的关系类型信息。从描述符树中遍历概念层的描述符，同时得到描述符间较粗粒度级别的关系（“Cause-Effect”、“Component-Whole”、“Feature-Of”、“Located-Of”、“Instance-Of”、“Condition-On”和“Method-Of”）；

步骤 2：对描述符执行回溯。通过可溯源处理模型对描述符实体对进行溯源，得到同时出现二者的句子，并将其加入候选材料知识库中；

步骤 3：获取材料构效关系知识。遍历候选知识库，并对当前句子进行推理，若当前句子中存在相关性词汇（如“positive”、“negative”、“increase”、“decrease”和“A 升高，B 减小”、“A 降低，B 增大”等，A、B 表示句子中的单词或短语）或影响性规则，表明描述符间存在材料构效关系知识，则将其联合并加入最终的知识库。

材料构效关系知识获取算法的伪代码如算法 4.2 所示。其中， $Filed - D_{tree}$ 为填充后的描述符树，PMTra 为本文 2.2 节的可溯源处理模型， $R = \{R_{ce}, R_{cw}, R_{fo}, R_{lo}, R_{io}, R_{co}, R_{mo}\}$ ，分别表示本文定义的“Cause-Effect”、“Component-Whole”、“Feature-Of”、“Located-Of”、“Instance-Of”、“Condition-

On” 和 “Method-Of” 关系类型。 $r(a, b)$ 表示 a 和 b 之间的关系， S_i 为回溯得到文献中句子， $Corr_{know}$ 为相关性信息。

算法 4.2: 材料构效关系知识获取

输入: 描述符树 $Filed - D_{tree} = \{H_1: \{"Prop_{obj}\": [Class_1, \dots, Class_i]\},$

$$H_2: \{P_{class_1}: [], \dots, P_{class_i}: []\},$$

$$H_3: \{[D_{11}, \dots, D_{1j}], \dots, [D_{i1}, \dots, D_{ij}]\}\};$$

可溯源处理模型 PMTra；

关系集合 $R = \{R_{ce}, R_{cw}, R_{fo}, R_{lo}, R_{io}, R_{co}, R_{mo}\}$ ；

知识集 $K_{now} = \{\}$ ；候选知识库 $Cand_{know} = \{\}$ 。

输出: K_{now}

1: 开始

2: 初始化 $K_{now}, Cand_{know}$

3: **foreach** $D_m, D_n \in Filed - D_{tree} \wedge (D_m, D_n \text{ in } H_3)$:

4: **if** $r(D_m, D_n) \in R$:

5: $S_i = \{D_m \cup D_n \text{ 输入到 PMTra}\} // \text{ 通过 PMTra 回溯得到句子 } S_i$

6: $Cand_{know} = Cand_{know} \cup \{S_i: [D_m, D_n]\} // S_i, D_m, D_n \text{ 加入 } Cand_{know}$

7: **foreach** $S_j \in Cand_{know}$ **do**

8: **if** $\exists Corr_{know} \text{ in } S_j$:

9: $K_{now} = K_{now} \cup \{Corr_{know}: [D_{jm}, D_{jn}]\} // Corr_{know} \text{ 和 } S_j \text{ 的 } D_{jm}, D_{jn} \text{ 加入 } K_{now}$

10: 结束

4.5 实验

4.5.1 实验数据

为了验证 MatBERT-BiGRU-Softmax 模型的有效性和鲁棒性，本章使用 NASICON 型固态电解质和 MatSciRE^[117]材料关系抽取数据集进行实验。

- **NASICON 型固态电解质材料关系抽取数据集:** 其为第二章构建的数据集，包含 7 种语义关系类型和一种人工关系类型，其中，7 种语义关系类型分别

为“Cause-Effect”、“Component-Whole”、“Instance-Of”、“Located-Of”、“Method-Of”、“Condition-On”及“Feature-Of”，人工关系类型为“Other”，即 7 种语义关系之外其余的均属于“Other”类。该数据集包含 2434 个句子中共 2297 个关系，每个句子包含两个特殊符号 e_1 和 e_2 及句子中实体对对应的关系类型。需要注意的是，本文构建的数据集中关系具有方向性，即“Cause-Effect(e_1, e_2)” \neq “Cause-Effect(e_2, e_1)”。

- **MatSciRE 材料关系抽取数据集：**其是通过手工标注材料科学文献所得的，包含五种材料关系，分别为：“Conductivity”、“Coulombic Efficiency”、“Capacity”、“Voltage”及“Energy”。该数据集包含 1255 个句子共 1793 个关系。其中，一个句子可能包含一个或多个关系标签。

4.5.2 实验设置

表 4.1 参数设置

参数名称	值
批量大小 (Batch size)	8
最大句子长度 (Max sequence length)	128
迭代周期 (Number of epochs)	5
丢包率 (Dropout rate)	0.2
线性层大小 (Linear size)	1024
学习率 (Adam learning rate)	2e-5
权值衰减 (Weight decay)	0.01
GRU 隐藏层大小 (GRU hidden size)	64
GRU 层数 (GRU number of layers)	1
CRU 丢包率 (GRU dropout rate)	0.2

模型的参数设置如表 4.1 所示。本章使用 AdamW 优化器^[109]进行参数调优，批的大小设置为 8，每次训练迭代过程中使用一个批次进行参数的更新，每个批次的数据随机训练 5 次，即 epoch 设置为 5，并设置初始学习率为 2e-5。在模型训练之前，需要对句子进行预处理。句子的最大长度设置为 128，小于该长度需要进行填充 (padding) 处理，大于该长度则需要进行截断 (truncation) 处理。GRU 模型包含一层且其隐藏层状态维度设置为 64。此外，我们使用了 Dropout^[110]和 L_2 正则化项 (权值衰减)^[118]以防止模型过拟合，其中 GRU 网络及模型其余部分的 Dropout 均设置为 0.2，权值衰减设置为 0.01。

4.5.3 评价指标

本章使用通用领域关系抽取 SemEval-2010 数据集^[119]官方评分器脚本来评估材料关系抽取方法，它计算 7 个实际关系（不包括“Other”）的宏观平均 F1 值，并考虑方向性。

4.5.4 实验结果与分析

(1) 关系抽取模型性能验证

为了验证 MatBERT-BiGRU-Softmax 模型的有效性及鲁棒性，本节在 NASICON 型固态电解质和 MatSciRE 材料关系抽取数据集上进行了一系列的实验，并与经典的关系抽取方法进行了比较。实验结果如表 4.2 所示，其中 WV，ATT 分别表示利用 Word2vec 模型获得词向量特征及利用注意力机制提取目标实体特征。

表 4.2 实验结果对比

模型	F1	
	NASICON 数据集	MatSciRE 数据集
WV+CNN+ATT ^[120]	0.52	0.58
WV+BiLSTM+ATT ^[58]	0.59	0.65
R-BERT ^[62]	0.66	0.72
D-BERT ^[63]	0.67	0.74
MatBERT-BiGRU-Softmax (Ours)	0.68	0.76

从表中可以看出，MatBERT-BiGRU-Softmax 模型在 NASICON 型固态电解质和 MatSciRE 材料关系抽取数据集上效果最好，F1 相较于 WV+CNN+ATT、WV+BiLSTM+ATT、R-BERT 和 D-BERT 模型分别提高了 16%、9%、2%、1% 和 18%、11%、4%、2%，这是由于 MatBERT-BiGRU-Softmax 模型对输入的数据进行了实体感知操作，即设计特殊标记“[]”和“{}”对目标实体词进行包裹，使得模型清晰地感知到材料目标实体及其边界信息，并由 MatBERT 和 BiGRU 模型充分提取了句子和单词级别的语义特征及单词的上下文语义特征信息。

为了更直观全面地评估 MatBERT-BiGRU-Softmax 模型在测试集上的分类性能，本节给出了模型训练的 P-R 曲线（精确率-召回率曲线）及混淆矩阵，如图

4.6 所示。从图 4.6(a)中可以看出，精确率和召回率大体呈负相关趋势，表明本章构建的 NASICON 型固态电解质材料关系抽取正负样本比例均衡。此外，P-R 曲线与坐标轴围成的阴影部分面积占比较大，表明本章的关系抽取模型在 NASICON 型固态电解质材料关系抽取数据集上有着较好的训练效果。从图 4.6(b) 中可以看出，对角线上的数字在每个真实关系标签上均是最大的，表明测试集中大部分样本预测标签与真实标签都相同，证明了 MatBERT-BiGRU-Softmax 模型对 NASICON 型固态电解质材料实体间的关系具有较好的分类准确性。

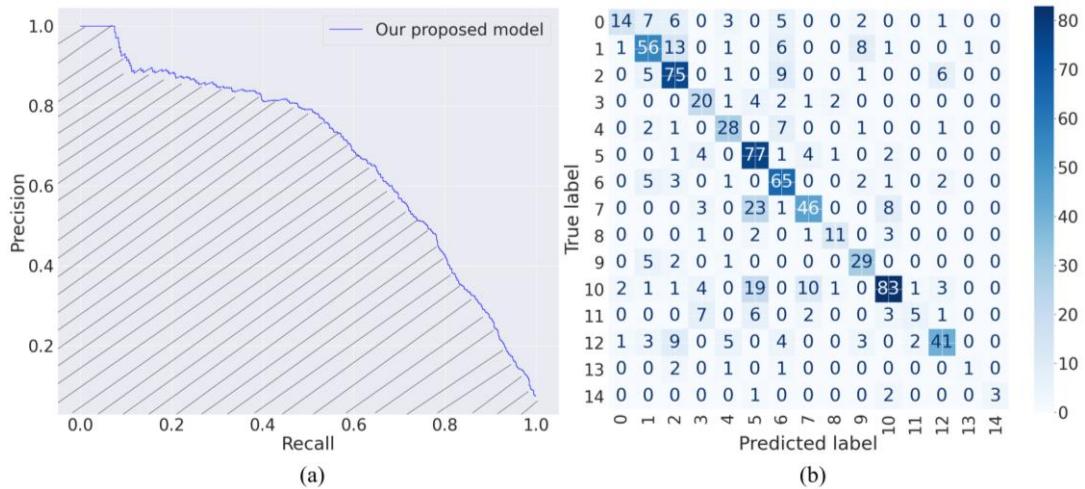


图 4.6 (a) MatBERT-BiGRU-Softmax 模型在 NASICON 型固态电解质材料关系抽取数据集上训练的 P-R 曲线。(b) MatBERT-BiGRU-Softmax 模型在 NASICON 型固态电解质材料关系抽取数据集上训练的混淆矩阵，其中 0-14 依次代表“Cause-Effect”、“Component-Whole”、“Instance-Of”、“Located-Of”、“Method-Of”、“Condition-On” 和 “Feature-Of” 不同方向的关系类别及 “Other” 类别。

(2) 消融实验

为了评估 MatBERT-BiGRU-Softmax 关系抽取模型的组成部分，本节在 NASICON 型固态电解质关系抽取数据集上设计了消融实验。在这组实验中，MatBERT-BiGRU-Softmax 的衍生模型包括：

(a) R-WV-BiGRU：没有通过 MatBERT 模型提取 token 及句子级别的语义特征，而是由 Word2vec 模型得到词嵌入向量作为模型的输入，其目的在于测试 MatBERT 对 MatBERT-BiGRU-Softmax 进行关系分类的影响。

(b) MatBERT-BiGRU-Softmax-NO-SEP-NO-ENT：未对数据集中的目标实

体用特殊封闭标记包裹并且未融合目标实体的语义向量进行关系的分类，其目的在于测试对数据集中句子的目标实体进行特殊标记包裹并融合其语义特征对 MatBERT-BiGRU-Softmax 进行关系分类的影响。

(c) MatBERT-BiGRU-Softmax-NO-ENT：仅通过抽取的句子语义特征对关系进行分类，未融合目标实体词的语义特征，其目的在于测试将目标实体词语义特征融入句子特征对 MatBERT-BiGRU-Softmax 进行关系分类的影响。

(d) R-MatBERT：缺失 BiGRU 网络，其目的在于测试 BiGRU 网络进一步对句子序列建模提取的局部上下文语义信息对 MatBERT-BiGRU-Softmax 进行关系分类的影响。

表 4.3 显示了上述四种结构的消融实验结果，从中可以看出这四种情况的表现都比 MatBERT-BiGRU-Softmax 模型差。

表 4.3 模型在 NASICON 型固态电解质材料关系抽取数据集上消融实验结果

模型	F1
R-WV-BiGRU	0.60
MatBERT-BiGRU-Softmax-NO-SEP-NO-ENT	0.63
MatBERT-BiGRU-Softmax-NO-ENT	0.65
R-MatBERT	0.66
MatBERT-BiGRU-Softmax	0.68

对于 (a)，R-WV-BiGRU 的 F1 值下降最多，为 8%。表明 MatBERT-BiGRU-Softmax 模型中由 MatBERT 语言模型生成材料词嵌入向量可以动态捕获含丰富语义信息的 token 级别及句子级别的语义特征信息，这是因为 MatBERT 模型在对材料文本进行编码的时候可以结合词嵌入、句子嵌入及位置嵌入信息进行学习。

对于 (b)，MatBERT-BiGRU-Softmax 模型对材料关系抽取数据集未加特殊封闭标记包裹及其在进行关系分类时未融合目标实体词的语义特征向量也会造成 F1 值下降，为 5%。表明 MatBERT-BiGRU-Softmax-NO-SEP-NO-ENT 无法定位目标实体并丢失此关键信息。而这些特殊的封闭标记可以提高准确性的原因是它们识别了两个目标实体的位置，并将信息转移到 MatBERT 模型中，这使得 MatBERT 输出包含了两个实体的位置信息。

对于 (c)，MatBERT-BiGRU-Softmax 模型在进行关系分类时未融合目标实体词的语义向量也会造成 F1 值下降，为 3%。表明 MatBERT-BiGRU-Softmax-

NO-ENT 模型在进行关系分类时丢失了部分目标实体的语义信息。而其可以提高准确性的原因是 MatBERT 模型结合目标实体向量的输出进一步丰富了信息，有助于做出更准确的关系预测。

对于 (d)，MatBERT-BiGRU-Softmax 模型未进一步提取 token 及句子级别的局部上下文语义特征也会造成 F1 值下降，为 2%。表明 R-MatBERT 模型在学习动态捕获 token 及句子级别的语义信息时丢失了部分上下文信息。而 BiGRU 可以进一步提升关系分类的准确性的原因是其通过对句子序列进行建模，可以弥补 R-MatBERT 模型丢失的局部上下文语义信息。

上述实验结果表明，使用 MatBERT 语言模型获得 token 及句子级别的向量表征、对数据集中的实体对添加特殊的标记、对关系进行分类时融入目标实体的语义信息及由 BiGRU 模型对句子序列建模提取 token 级别的局部上下文语义特征都对 MatBERT-BiGRU-Softmax 模型有重要贡献。

4.6 应用

本节以 NASICON 型固态电解质材料为例探索所提方法在材料领域的有效应用。首先，构建了 NASICON 型固态电解质材料知识图谱以实现实体关系三元组的存储；其次，建立了 NASICON 型固态电解质激活能描述符树，结合描述符树与知识图谱获取了 NASICON 型固态电解质构效关系知识；最后，将所获知识进行表示后嵌入材料机器学习模型中，并通过知识嵌入的特征选择实验验证材料知识的有效性。

4.6.1 NASICON 型固态电解质材料的知识图谱构建

本节基于 3.2 节的材料实体识别模型和 4.2 节的材料关系抽取模型，以管道方式从 1808 篇 NASICON 型固态电解质材料研究相关的文献中抽取出 260475 个材料实体关系三元组，不同关系类型的统计信息如图 4.7 所示。

在此基础上，利用 Neo4j 图数据库，本节实现了对 MatBERT-BiGRU-Softmax 模型挖掘的材料实体关系三元组信息的存储，如图 4.8 所示。

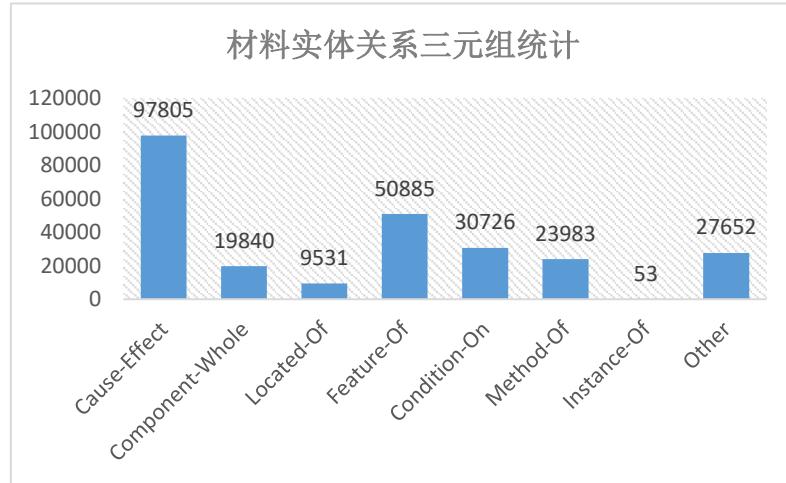


图 4.7 材料关系抽取模型挖掘的实体关系三元组统计图

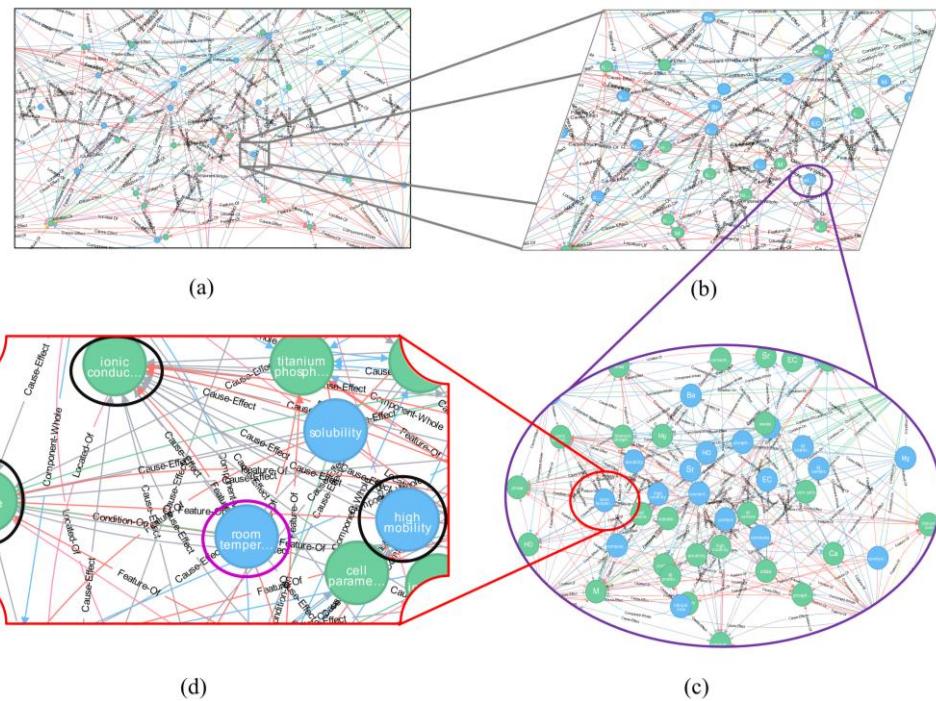


图 4.8 材料实体关系三元组存储的可视化展示。其中，不同颜色的边代表不同的材料关系类型，“Cause-Effect”、“Component-Whole”、“Condition-On”、“Feature-Of”、“Instance-Of”、“Located-Of”、“Method-Of” 对应边的颜色分别为灰、蓝、绿、红、紫、粉和黄色。

图 4.8(a)展示了存储材料实体关系三元组的 Neo4j 图数据库总体概览。从中可以看出，Neo4j 图数据库中不仅蕴含着大量的材料实体信息（节点），而且还蕴含着材料实体间的关系信息（边），即具有关系的两个材料实体会有关系边将其连接起来，由此便可快速定位到材料实体及其关系信息从而便于推理得到知识。在此，本节以“room temperature”为例来展示 Neo4j 图数据库中存储的材料实体

关系三元组信息，图 4.8(b)、图 4.8(c)及图 4.8(d)则为图 4.8(a)的放大版本。从中可以清晰地看出，材料实体“room temperature”和“high mobility”、“phase”及“ionic conductivity”被灰色的关系边连接，因此可得知温度会影响迁移率、相和离子电导率。此外，在图 4.8(c)中，我们还观察到“cell parameter”与“volume”及“M”也被灰色的关系边连接，因此可得知晶胞参数会影响体积且会受“M”位置元素的影响。

在通用领域知识图谱的概念驱动下，结合提出的 MatKGPtcl 准则，本节最终构建了 NASICON 型固态电解质材料激活能驱动的描述符逻辑关系图谱，如图 4.9 所示，并基于 Neo4j 图数据库快速检索功能对其进行填充。

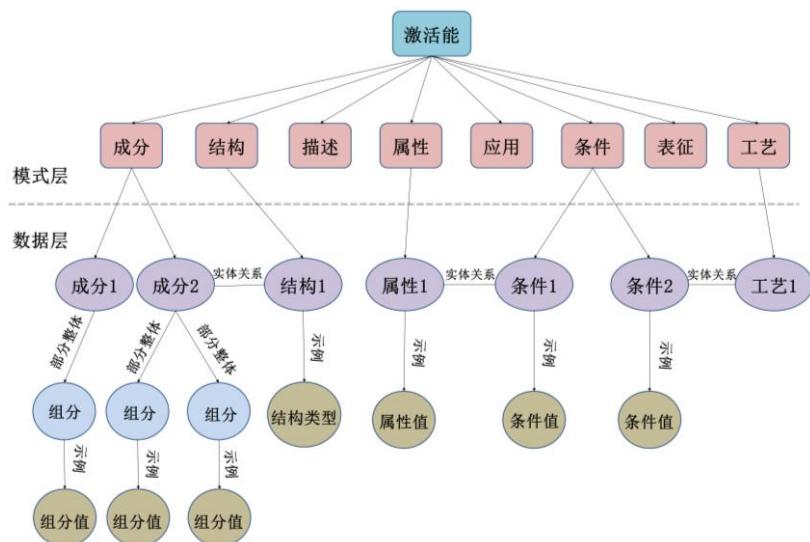


图 4.9 NASICON 型固态电解质材料激活能驱动的描述符逻辑关系图谱

从图 4.9 中可以看出，数据层存储了大量与激活能具有一定关系的、不同类型的描述符信息，且不同类型描述符之间的关系在图谱中也能够显示出来。此外，本节构建的 NASICON 型固态电解质材料激活能驱动的描述符逻辑关系图谱继承了 Neo4j 的检索功能，因而便具备了快速捕获到相应材料信息的能力。

4.6.2 NASICON 型固态电解质材料的描述符树建立

本文 4.4.2 节构建了描述符树的通用形式，本节基于算法 4.1 对描述符树进行填充，以得到 NASICON 型固态电解质激活能描述符树。具体地，首先将粗粒度层的目标性能定义为激活能，同时衍生出命名实体识别设定的材料实体类别中

的“成分”、“结构”、“性能”、“工艺”和“条件”等影响激活能的粗粒度描述符概念，以实现对粗粒度层的填充；其次，细粒度层在每个实体类别下设置细粒度的参数以实现其填充；最后，在材料知识图谱的驱动下实现概念层描述符的填充。以对粗粒度层“性能”下的概念层进行填充为例，由于描述符实体间的关系具有方向性，因此本节分别以头实体和尾实体驱动寻找与激活能的关系类型为“Cause-Effect”描述符实体关系三元组，如图 4.10 所示。将上述两种情况获得的描述符取交集以完成填充操作。

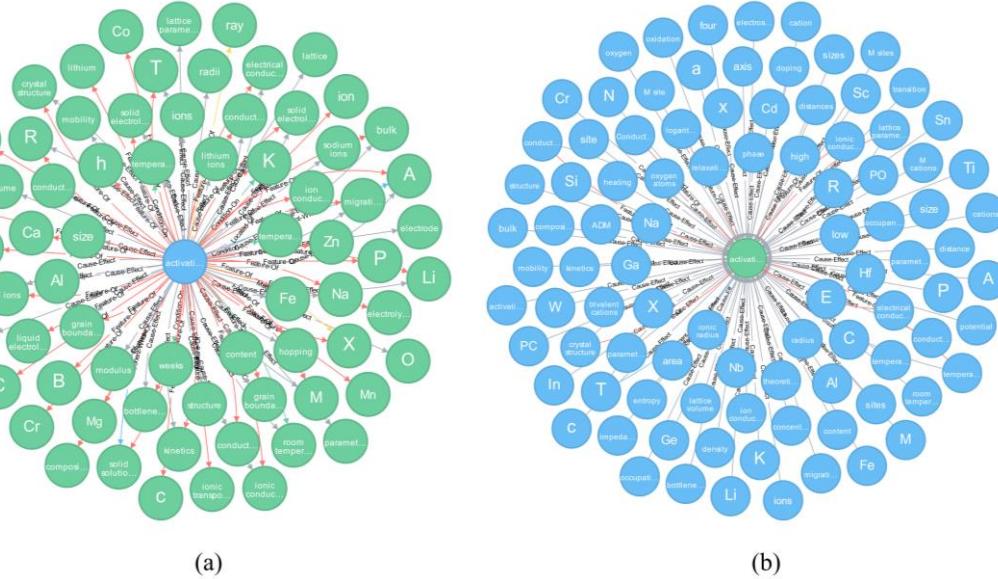


图 4.10 激活能相关描述符实体关系三元组可视化展示，红色箭头代表“Cause-Effect”关系类型。（a）以头实体驱动的激活能相关描述符实体查找。（b）以尾实体驱动的激活能相关描述符实体查找。

此外，在每个概念下，本节基于知识图谱进行了更深层次的检索，如图 4.11 所示。例如，在将“sodium ions”填充至从属于“成分”下的概念层之后，寻找与该描述符（“sodium ions”）具有“Component-Whole”关系的其它描述符进行再次填充。同理，其余概念层描述符树也是基于类似的操作完成填充。最终构建的 NASICON 型固态电解质激活能描述符树如图 4.12 所示。其中，不同层级下详细的描述符信息如表 4.4 所示。

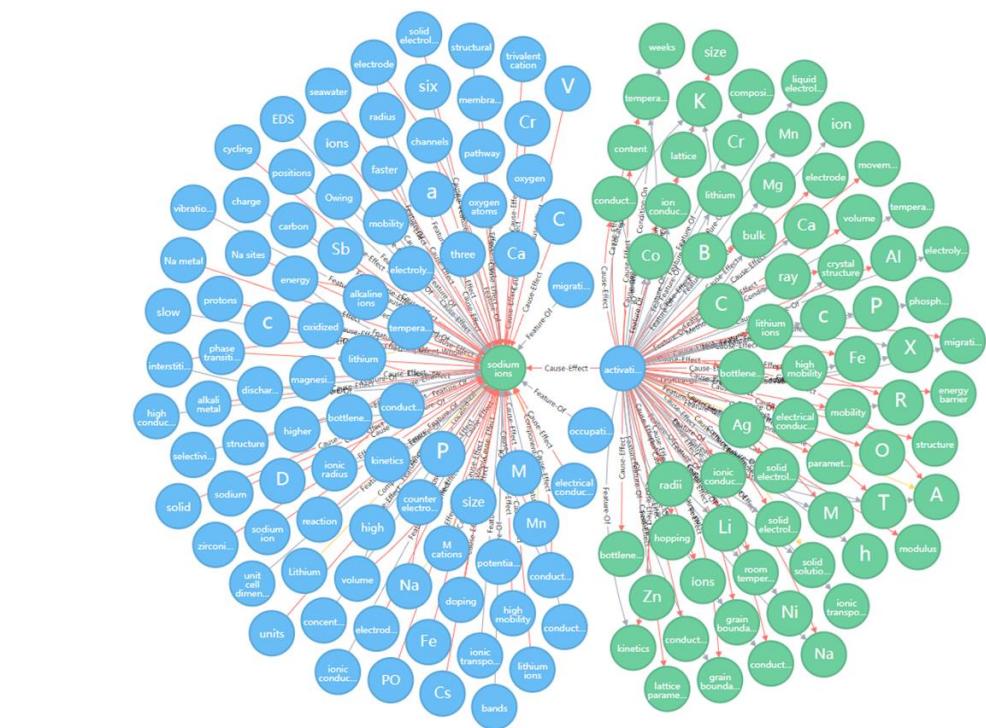


图 4.11 影响激活能成分相关实体关系三元组可视化展示，黄色箭头代表“Component-Whole”关系类型。

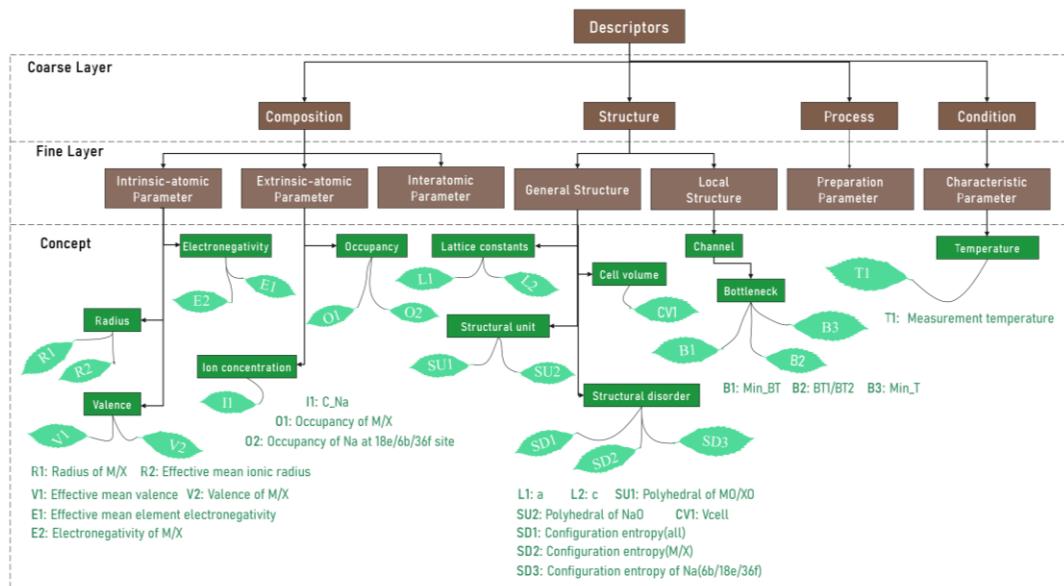


图 4.12 描述符树（仅展示部分描述符）

表 4.4 NASICON 型固态电解质激活能描述符树中不同层级下的描述符

粗粒度层	细粒度层	概念层
Composition	Intrinsic-atomic	Electronegativity of M1; Electronegativity of M2; Electronegativity of X1; Electronegativity of X2; Effective

	parameter	mean element electronegativity of M; Effective mean element electronegativity of X; Polarizability of M1; Polarizability of M2; Polarizability of X1; Polarizability of X2; Effective mean element polarizability of M; Effective mean element polarizability of X; Radius of M1; Radius of M2; Radius of X1; Radius of X2; Effective mean radius of M; Effective mean radius of X; Valence of M1; Valence of M2; Valence of X1; Valence of X2; Effective mean valence of M; Effective mean valence of X; First ionization energy of M1; First ionization energy of M2; First ionization energy of X1; First ionization energy of X2; Second ionization energy of M1; Second ionization energy of M2; Second ionization energy of X1; Second ionization energy of X2; Volume of M1; Volume of M2; Volume of X1; Volume of X2; Effective mean valence of M; Effective mean valence of X; Mass of M1; Mass of M2; Mass of X1; Mass of X2; Effective mean mass of M; Effective mean mass of X; Density of M1; Density of M2; Density of X1; Density of X2; Effective mean density of M; Effective mean density of X; Stoichiometric number of species Na according to formula; Occupancy of Na (6b); Occupancy of Na (18e); Occupancy of Na (36f); Occupancy of M1; Occupancy of M2; Occupancy of X1; Occupancy of X2; Vacancy of M; Vacancy of X; Vacancy of 6b; Vacancy of 18e; Vacancy of 36f;
	Extrinsic-atomic parameter	Bond length; Bond strength; Coulomb interaction
	Interatomic parameter	Formation energies of materials; Vacancy formation energy; Energy landscapes; Entropy of migration
	Overall parameters	a; c; V; Configuration entropy of Na (6b); Configuration entropy of Na (18e); Configuration entropy of Na (36f); Configuration entropy of Na (all); Configuration entropy of M; Configuration entropy of X; Haven ratio; Jump frequency; Attempt frequency; The percentage of the correlated jumps; All jump events; Lattice softness; Lattice symmetry; Debye frequency; Average phonon band center; Dielectric constant; Ion oscillation; Geometry factor; Packing mode; Packing fraction; Volume effect; Compressive strain
Structure	General structure	Polyhedral of MO ₆ ; Polyhedral of XO ₄ ; Polyhedral of Na ₁ O ₆ ; Polyhedral of Na ₂ O ₈ ; Polyhedral of Na ₃ O ₅ ; BT1; BT2; Min_BT; Conduction threshold; The dimensionality of diffusion pathways; The volume of diffusion pathways; Mechanical strength
	Local structure	

Processing	Preparation Parameter	/
Condition	Characteristic Parameter	Temperature; Pressure

注：“/”表示为空。

4.6.3 NASICON 型固态电解质材料的知识获取

数据驱动的机器学习因其能够快速拟合历史数据中的潜在模式并实现材料性能的预测，已被成功应用于 NASICON 型固态电解质材料激活能构效关系的研究。然而，由于缺乏描述符间关联关系、材料性能驱动机制等材料领域知识的指导，数据驱动的机器学习在实际应用中常常出现与材料基础理论认知或原理不一致的结果。因此，本节基于 4.4.3 节设计的算法 4.2，实现从 4.6.2 建立的描述符树中检索出描述符间关系并推理出 NASICON 型固态电解质材料构效关系知识，同时对其进行表示以嵌入到数据驱动特征选择方法中，以验证知识的有效性。

(1) 知识获取的结果

本文将 3.6.2 节构建的包含 31 和 45 个特征的 NASICON 型固态电解质激活能预测数据集 $Dataset_{31}$ 和 $Dataset_{45}$ 作为研究对象以获取特征间的材料构效关系知识。

表 4.5 $Dataset_{31}$ 和 $Dataset_{45}$ 数据集中特征的并集

序号	并集特征名称	描述
1	a 、 c 、 a/c 、 d 、 h	晶格常数
2	V	晶胞体积
3	R_M1 、 R_M2 、 Avg_M_R 、 R_X1 、 R_X2 、 Avg_X_R 、 RT	离子半径
4	EN_M1 、 EN_M2 、 Avg_M_EN 、 EN_X1 、 Avg_X_EN	电负性
5	V_{MO_6} 、 V_{XO_4} 、 $V_{Na(1)O_6}$ 、 $V_{Na(2)O_8}$ 、 $V_{Na(3)O_5}$	多面体体积
6	$D2_stoich$ 、 $D3_stoich$ 、 Na_stoich 、 $X1_stoich$ 、 $X2_stoich$	化学计量数
7	O_{Na1} 、 O_{Na2} 、 O_{Na3} 、 O_M1 、 O_M2 、 O_X1 、 O_X2	占据率
8	C_{Na}	浓度
9	V_M1 、 V_M2 、 Avg_M_V 、 V_X1 、 V_X2 、 Avg_X_V	价态
10	$BT1$ 、 $BT2$ 、 Min_BT	瓶颈
11	$E_{Na(1)}$ 、 $E_{Na(2)}$ 、 $E_{Na(3)}$ 、 E_{Na} 、 E_M 、 E_X	构型熵
12	T	温度

首先取上述两个数据集中描述符特征的并集作为描述符树的检索依据,结果如表 4.5 所示;

在此基础上,基于算法 4.2 的流程,从 NASICON 型固态电解质材料描述符树概念层选择与并集中具有特征关联的描述符,结合回溯模型得到 NASICON 型固态电解质激活能及其影响因素间的关系和相关句子,并将其联合加入候选知识库。候选知识库的结果如表 4.6 所示。

表 4.6 候选知识库

描述符 1	描述符 2	关系	句子
lattice parameters	cell volume	Cause-effect	<p>1. As the Mo⁺ doping content increasing, less Na ions are introduced into the crystal lattice that the lattice parameters of <i>a</i> and <i>c</i> and cell volume increase, as seen in Fig b.</p> <p>2 There was a very slight increase in lattice parameter, <i>a</i>, and hence an increase in lattice volume after immersion for all stability tests.</p>
lattice parameters	bottleneck	Cause-effect	<p>1. The lattice parameters <i>a</i> and <i>c</i>, and the unit cell volume increased with as the <i>X</i> value increased for this system, which likely increased the bottleneck size in the ionic pathway.</p> <p>2. We interpret this conductivity increase as a consequence of an enhanced lithium mobility due to the release of the bottleneck M positions of the NASICON structure promoted by Cr insertion, resulting in a reduction of the activation energy.</p> <p>3. The aim of this work is to show that a relationship exists between the size of the bottleneck between the M and M sites and the activation energy involved in the motion of Li⁺ ions along the conduction channels.</p>
bottleneck	activation energy	Cause-effect	<p>3. The aim of this work is to show that a relationship exists between the size of the bottleneck between the M and M sites and the activation energy involved in the motion of Li⁺ ions along the conduction</p>

ionic radius	occupancy ratio	Cause-effect	channels. 1. This is due to de-mixing of the substituents from the NASICON structure and subsequently unfavorable site occupancy, especially for Al ⁺ ions due to their small ionic radius in octahedral coordination.
channels	activation energy	Feature-Of	1. Taking into account that the activation energy gives a measure of the hindrance in the movement of Li ⁺ ions along the conduction channels, we interpret the two observed regimes as follows: the size of the bottleneck is less than that of Li ⁺ ion in the first regime and larger in the second one. 2. The activation energy involved in the movement of Li ⁺ ions along the conduction channels of the NASICON framework is a parameter that includes at least two effects: one strongly dependent on the size of bottleneck between M and M sites, and the other related to lithium-lattice and/or lithium-lithium interactions.
temperature	ionic radius	Condition-On	1. The relationship between bulk conductivity of the doped NASICON and dopant ionic radius can be seen at a range of temperatures.
valence	occupy	Cause-effect	1. Typically, dopant ions are expected to occupy the cation sites with similar effective ionic radius and valence. 1. On increasing the level of aluminum doping, the unit cell volume slightly decreases, which is a direct consequence of the lower effective ionic radius of Al ⁺ . 2. But in terms of ionic radius, Na is larger than Li, which tends to cause greater expansion in volume during insertion process, further resulting in unsatisfactory cycling stability and inferior specific capacity. 3. In particular, the aforementioned limitation of larger ionic radius leads to
ionic radius	volume	Cause-effect	

ionic radius	activation energy	Feature-Of	huge volume changes, especially when Na is inserted into/extracted from the host materials, thus causing severe structural degradation as well as the consequence of exhibiting poor cycleability and rate capability.
configuration entropy	activation energy	Cause-effect	<p>4. However, the larger ionic radius of K⁺ is likely to induce dramatic volume expansion and sluggish kinetic properties.</p> <p>5. When the ionic radius of the substituent element is smaller than existing element, this cause the cell parameter and the cell volume decrease.</p>
temperature	activation energy	Condition-On	<p>1. With increasing ionic radius of R⁺ in RO, the activation energy decreases and the conductivity increases.</p> <p>2. The activation energy of crystal growth of the NaYRPSi glass decreases as the ionic radius of R is increased.</p>
temperature	occupancy	Condition-On	<p>1. According to this rule, motions with important activation energy improves considerably the entropic term, but motions with lower activation energy only improves slightly configuration entropy associated with Li motion.</p> <p>1. The activation energy is higher at low temperature in the derived garnet - type oxides, indicating the introduction of defects for Li⁺ trapping.</p> <p>1. An increase in temperature leads to a significant decrease in the occupancy of the Na positions.</p> <p>2. The authors also found that the occupancy rate is highly dependent on the temperature.</p>
occupancy	activation energy	Cause-effect	<p>1. At still higher lithium contents, the increase in M site occupancy is accompanied by a gradual rise in activation energy, up to kJ / mol.</p> <p>1. Increasing the sintering temperature causes the lattice parameters and the unit volume to increase.</p>
temperature	volume	Condition-On	

最后，对候选知识库中的句子进行进一步推理。若句子中存相关性词汇或影响性规则，则将其加入最终的知识库中，以此便得到了激活能及其影响因素间的构效关系知识。描述符树中的实体关系知识与 NASICON 型固态电解质材料构效关系知识间的对应关系如表 4.7 所示。

表 4.7 NASICON 型固态电解质构效关系知识

序号	描述符树中的实体关系知识	材料构效关系知识
1	(“lattice parameter”, “Cause-effect”, “cell volume”)	晶格常数与晶胞体积具有正相关关系
2	(“lattice parameter”, “Cause-effect”, “bottleneck”)	晶胞参数与瓶颈具有负相关关系
3	(“lattice parameter”, “Cause-effect”, “occupancy ratio”)	- (晶格常数影响占据率)
4	(“bottleneck”, “Cause-effect”, “activation energy”)	瓶颈与激活能具有负相关关系
5	(“ionic radius”, “Cause-effect”, “occupancy ratio”)	- (离子半径影响占据率)
6	(“temperature”, “Condition-On”, “activation energy”)	温度与激活能具有负相关关系
7	(“ionic radius”, “Cause-effect”, “volume”)	离子半径与体积具有正相关关系
8	(“ionic radius”, “Feature-Of”, “activation energy”)	M 位置元素离子半径越大，激活能越小
9	(“bottleneck”, “Cause-effect”, “occupancy ratio”)	- (瓶颈影响占据率)
10	(“ionic conductivity”, “Cause-effect”, “activation energy”)	激活能与离子电导率具有较明显的负相关关系
11	(“ionic concentration”, “Cause-effect”, “bottleneck”)	- (离子浓度影响瓶颈)
12	(“channel size”, “Cause-effect”, “activation energy”)	通道尺寸与激活能具有负相关关系
13	(“occupancy ratio”, “Cause-effect”, “electronegativity”)	- (占据率影响电负性)
14	(“temperature”, “Condition-On”, “volume”)	降温可以减小晶胞体积
15	(“ionic radius”, “Cause-effect”, “bottleneck”)	- (离子半径影响瓶颈)
16	(“Na concentration”, “Cause-effect”, “activation energy”)	Na 离子浓度与激活能具有负相关关系
17	(“temperature”, “Condition-On”, “ionic radius”)	- (温度影响离子半径)

18	(“occupancy ratio”, “Cause-effect”, “activation energy”)	X位置元素占据率越大，激活能越小 - (构型熵影响占据率)
19	(“configuration entropy”, “Cause-effect”, “occupancy ratio”)	温度与占据率具有正相关关系
20	(“temperature”, “Condition-On”, “occupancy ratio”)	- (价态影响占据率)
21	(“valance”, “Cause-effect”, “occupancy ratio”)	晶胞体积与激活能具有负相关关系
22	(“volume”, “Feature-Of”, “activation energy”)	构型熵越大，激活能越小
23	(“configuration entropy”, “Cause-effect”, “activation energy”)	- (电负性影响激活能)
24	(“electronegativity”, “Cause-effect”, “activation energy”)	

注：“-”未从句子中推理到描述符间明显的正负相关性信息。

(2) 知识的有效性验证

在材料领域知识嵌入的机器学习方法思想的指导下,本课题组研发了材料领域知识嵌入的特征选择方法(Feature Selection method embedded with NCOR, NCOR-FS)^[121]。为了对所获得材料知识的有效性进行验证,本节将该方法作为知识的应用对象。基于所获取的关于描述符与激活能间或描述符间的相关性知识,将其转换成机器学习模型可以利用的表示形式,并嵌入到NCOR-FS方法中,以提高样本的特征选择效果。

● NCOR-FS 方法简介

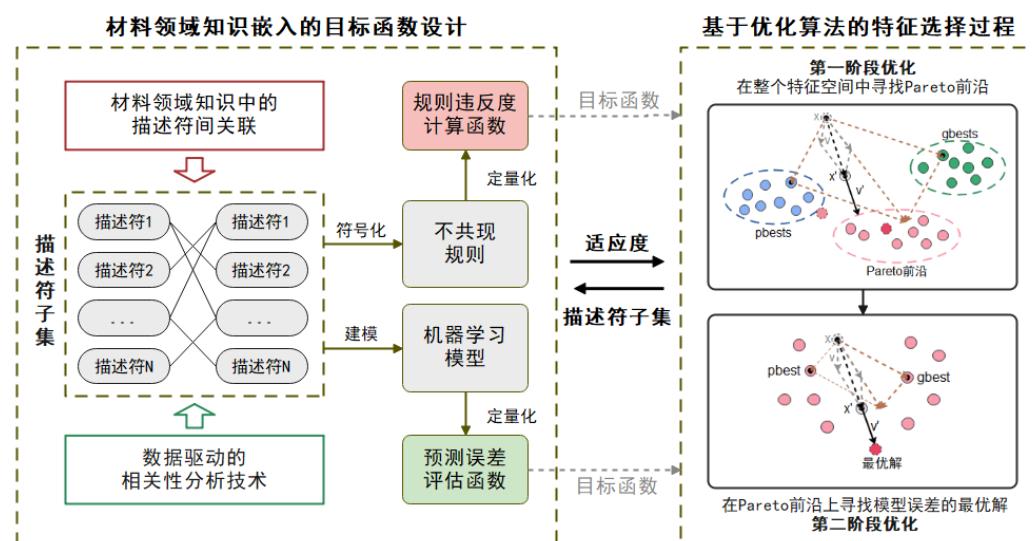


图 4.13 材料领域知识嵌入的特征选择方法流程^[121]

NCOR-FS 方法如图 4.13 所示, 将材料领域知识中描述符间关系与数据驱动的相关性分析技术结合来获取描述符间的不共现关系, 并将其符号化为不共现规则, 其中, 不共现规则的定义如公式 (4.15) 所示。在此基础上, 通过建立任意描述符集合的不共现规则违反度计算函数, 并联合机器学习模型的预测误差评估函数共同作为基于优化算法的特征选择方法的目标函数, 以评估描述符子集的适应度, 最后由两阶段的进化过程来优化特征选择方法的寻优过程。

$$\forall F_i, F_j \in R: (\forall p \in F_i, \forall q \in F_j): \neg(p \in X \wedge q \in X) \quad (4.15)$$

其中, $R = \{F_1, F_2, \dots, F_n\}$ 表示集合, F_i 和 F_j 是 R 中由一个或多个描述符组成的集合, 每个 F 在统计或材料意义上具有高度相关性, 且 $F_i \cap F_j = \emptyset$ ($i, j \in \{1, \dots, n\}, i \neq j$)。 X 为任意一个描述符子集, 当其不满足公式 (4.15) 时称其违反不共现规则 R 。

● 知识到不共现规则的转换

本节基于表 4.5 中的材料构效关系知识 (知识 1、4、5、7、13、14、19、21、23) 来获取如表 4.8 所示的特征间的不共现规则 (规则 1~27), 并将这些知识作为 NCOR-FS 方法的不共现规则。

表 4.8 $Dataset_{31}$ 和 $Dataset_{45}$ 上知识驱动的不共现规则

数据集	编号	知识来源	不共现规则	
			F_1	F_2
$Dataset_{45}$	规则 1	知识 19	{O_Na(1)}	{E_Na(1)}
	规则 2		{O_Na(2)}	{E_Na(2)}
	规则 3		{O_Na(3)}	{E_Na(3)}
	规则 4		{O_M1, O_M2}	{E_M}
	规则 5		{O_M1, O_M2, EN_M1, EN_M2}	{avg_EN_M}
	规则 6		{O_X1, O_X2, EN_X1, EN_X2}	{avg_EN_X}
	规则 7		{O_M1, O_M2, R_M1, R_M2}	{avg_R_M}
	规则 8		{O_X1, O_X2, R_X1, R_X2}	{avg_R_X}
	规则 9		{O_M1, O_M2, V_M1, V_M2}	{avg_V_M}
	规则 10		{O_X1, O_X2, V_X1, V_X2}	{avg_R_V}
	规则 11		{a, c}	{V}

<i>Dataset</i> ₃₁	规则 12	知识 23	{E_Na(1), E_Na(2), E_Na(3)}	{E_Na}
	规则 13	知识 4	{BT1, BT2}	{Min_BT}
	规则 14	知识 1	{a, c}	{a/c}
	规则 15		{a, c}	{V_cell})}
	规则 16		{Na_ionicr}	{Na_volperatom}
	规则 17		{D2_ionicr}	{D2_volperatom}
	规则 18		{D3_ionicr}	{D3_volperatom}
	规则 19		{Na_eff}	{Na_ionicr, Na_Stoich}
	规则 20	知识 7	{D2_eff}	{D2_ionicr, D2_Stoich}
	规则 21		{D3_eff}	{D3_ionicr, D3_Stoich}
	规则 22		{Na_vol}	{Na_ionicr, Na_Stoich}
	规则 23		{D2_vol}	{D2_ionicr, D2_Stoich}
	规则 24		{D3_vol}	{D3_ionicr, D3_Stoich}
	规则 25		{D3_vol}	{D3_ionicr, D3_Stoich}
	规则 26	知识 14	{Na_eneff}	{Na_eneg, Na_Stoich}
	规则 27		{D3_eneff}	{D3_eneg, D3_Stoich}

注: F_1 和 F_2 表示描述符集合。

由知识 1 推出表 4.8 中规则 11 和规则 14~15: 从知识 1 “晶格常数与晶胞体积具有正相关关系” 可以看出, “晶格常数” 与 “晶胞体积” 具有正相关关系, 当它们一起作为候选描述符时, 由 “晶格常数” 构成的集合就与 “晶胞体积” 构成的集合形成了一个不共现规则, 即 $R = \{F_1, F_2\}$, 且 $F_1 = \{"lattice parameter"\}$, $F_2 = \{"cell volume"\}$, 对应到 $Dataset_{45}$ 和 $Dataset_{31}$ 中的相关描述符特征便可得到规则 11 和规则 14~16。

由知识 5 推出表 4.8 中规则 7~8: 从知识 5 “离子半径与占据率存在相关关系” 可以看出, “离子半径” 与 “占据率” 具有相关关系, 当它们一起作为候选描述符时, 由 “离子半径” 构成的集合就与 “占据率” 构成的集合形成了一个不共现规则, 即 $R = \{F_1, F_2\}$, 且 $F_1 = \{"occupancy ratio"\}$, $F_2 = \{"ionic radius"\}$, 对应到 $Dataset_{45}$ 和 $Dataset_{31}$ 中的相关描述符特征便可得到规则 7~8。

由知识 7 推出表 4.8 中规则 16~21: 从知识 5 “离子半径与体积存在相关关系” 可以看出, “离子半径” 与 “体积” 具有相关关系, 当它们一起作为候选描述符时, 由 “离子半径” 构成的集合就与 “体积” 构成的集合形成了一个不共现规则, 即 $R = \{F_1, F_2\}$, 且 $F_1 = \{"ionic radius"\}$, $F_2 = \{"volume"\}$, 对应到 $Dataset_{45}$ 和 $Dataset_{31}$ 中的相关描述符特征便可得到规则 16~25。

由知识 13 推出表 4.8 中规则 5~6: 从知识 1“占据率与电负性存在相关关系”可以看出，“占据率”与“电负性”具有正相关关系，当它们一起作为候选描述符时，由“占据率”构成的集合就与“电负性”构成的集合形成了一个不共现规则，即 $R = \{F_1, F_2\}$ ，且 $F_1 = \{"occupancy ratio"\}$, $F_2 = \{"electronegativity"\}$ ，对应到 $Dataset_{45}$ 和 $Dataset_{31}$ 中的相关描述符特征便可得到规则 5~6。

由知识 19 推出表 4.8 中规则 1~4: 从知识 1“晶格常数与晶胞体积具有正相关关系”可以看出，“占据率”与“构型熵”具有相关性，当它们一起作为候选描述符时，由“占据率”构成的集合就与“构型熵”构成的集合形成了一个不共现规则，即 $R = \{F_1, F_2\}$ ，且 $F_1 = \{"occupancy ratio"\}$ ， $F_2 = \{"configuration entropy"\}$ ，对应到 $Dataset_{45}$ 和 $Dataset_{31}$ 中的相关描述符特征便可得到规则 1~4。

由知识 21 推出表 4.8 中规则 9~10: 从知识 1“占据率与价态存在相关关系”可以看出，“占据率”与“价态”具有相关性，当它们一起作为候选描述符时，由“占据率”构成的集合就与“价态”构成的集合形成了一个不共现规则，即 $R = \{F_1, F_2\}$ ，且 $F_1 = \{"occupancy ratio"\}$, $F_2 = \{"valance"\}$ ，对应到 $Dataset_{45}$ 和 $Dataset_{31}$ 中的相关描述符特征便可得到规则 9~10。

类似地，由知识 4、14 和 23 分别可以得到表 4.8 中的规则 13、规则 26~27 和规则 14。

- 基于 NCOR-FS 方法的 NASICON 型固态电解质激活能预测

表 4.9 $Dataset_{45}$ 上未嵌入知识和嵌入知识的特征选择实验结果^[121]

模型	未嵌入知识的特征选择			嵌入知识的特征选择		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
LASSO	0.058	0.035	0.943	0.051	0.034	0.958
GPR	0.052	0.037	0.954	0.043	0.026	0.970
Ridge	0.051	0.033	0.956	0.049	0.027	0.963
SVR	0.071	0.057	0.916	0.060	0.046	0.942
KNN	0.079	0.051	0.894	0.053	0.056	0.948
RF	0.051	0.035	0.953	0.521	0.035	0.955

将表 4.8 中获取的不共现规则融入 NCOR-FS 建模中，实现 $Dataset_{45}$ 和 $Dataset_{31}$ 数据集上的特征选择和激活能预测，以验证我们获取知识的应用价值。

首先，通过对比我们发现，表 4.8 中所获得 $Dataset_{45}$ 特征间的不共现规则和

本课题组基于专家经验知识获得的不共现规则^[121]完全一致。因此，我们给出了本课题组 NCOR-FS 方法通过将上述不共现规则嵌入到机器学习模型中进行特征选择的实验结果，如表 4.9 所示。从中可以看出，6 个机器学习模型在 *Dataset₄₅* 数据集上知识嵌入特征选择的实验结果（*RMSE*、*MAPE* 和 *R²*）均要优于未嵌入知识的结果。其中，知识嵌入的特征选择 GPR 模型的实验结果最好，*RMSE*、*MAPE* 和 *R²* 分别为 0.043、0.026、0.970，相较于未嵌入知识的特征选择最优模型取得了 0.9%、1.1% 和 1.6% 的进步；6 个模型中泛化性能较差的（SVR）*R²* 也达到 0.9423。上述结果表明了知识嵌入机器学习特征选择的有效性，进而证明了本节获取的相关材料构效关系知识的真实性与可用性。

进一步，本节将表 4.8 中所获得 *Dataset₃₁* 特征间的不共现规则形式化表示成不共现规则违反度后嵌入 NCOR-FS 方法的特征选择机器学习模型中，并在 *Dataset₃₁* 数据集上进行了大量的特征选择实验，结果如表 4.10 所示。

表 4.10 *Dataset₃₁* 上未嵌入知识和嵌入知识的特征选择实验结果

模型	未嵌入知识的特征选择			嵌入知识的特征选择		
	<i>RMSE</i>	<i>MAPE</i>	<i>R²</i>	<i>RMSE</i>	<i>MAPE</i>	<i>R²</i>
LASSO	0.079	0.071	0.910	0.074	0.058	0.925
GPR	0.097	0.070	0.842	0.102	0.094	0.851
Ridge	0.080	0.073	0.894	0.076	0.067	0.922
SVR	0.085	0.076	0.889	0.084	0.072	0.903
KNN	0.079	0.070	0.913	0.068	0.052	0.928
RF	0.081	0.072	0.906	0.078	0.069	0.920

从中可以看出，6 个机器学习模型在 *Dataset₃₁* 数据集上知识嵌入特征选择的实验结果（*RMSE*、*MAPE* 和 *R²*）均要优于未嵌入知识的结果。其中，知识嵌入的特征选择 KNN 模型的实验结果最好，*RMSE*、*MAPE* 和 *R²* 分别为 0.068、0.052、0.928，相较于未嵌入知识的特征选择最优模型取得了 0.9%、1.8% 和 1.5% 的进步；相比之下，GPR 模型的效果最差，*RMSE*、*MAPE* 和 *R²* 分别为 0.076、0.067、0.851，这是由于 GPR 模型对当前材料数据不够敏感造成的，除此之外其余模型在 *Dataset₃₁* 数据集上知识嵌入的特征选择实验结果中 *R²* 均超过 90%。上述结果表明了知识嵌入特征选择的有效性。此外，我们检查 KNN 模型十折交叉实验过程发现，表现最好的一次预测选择了 8 个描述符的组合（序号分别为：0、3、4、19、24、28、29、30），表明知识嵌入的特征选择可以实现筛选相关性较低的描

述符组合进行 NASICON 型固态电解质化合物激活能的预测，进而证明了基于表 4.8 中的材料构效关系知识对数据驱动机器学习模型的指导意义。

4.7 小结

本章主要研究了基于实体感知的材料关系抽取方法。首先阐述了材料实体关系抽取的研究现状及其存在的问题；然后针对材料目标实体及其边界语义信息难以被已有的关系抽取模型感知的问题，提出了基于实体感知的材料关系抽取方法 MatBERT-BiGRU-Softmax 以实现关系的抽取。该方法首先通过对目标实体词分别用特殊的封闭标记“[]”及“{}”包裹，使得 MatBERT 模型能充分感知目标实体以提取更丰富的目标实体及句子的语义信息；其次，引入 BiGRU 模型来感知目标实体及其周围的局部上下文语义特征；最后，由 *Softmax* 函数计算得到候选关系中概率最大的一个来实现材料关系的分类。实验证明，本章提出实体关系抽取方法在不同的材料数据集上都达到了理想的分类效果。进一步，本章以 NASICON 型固态电解质为例，构建了材料知识图谱实现对获取的实体关系三元组的存储；在此基础上，建立了描述符树以推理并获取材料构效关系知识，并将获得的知识表示后嵌入数据驱动的特征选择方法中进行激活能的预测，结果证明了所获取知识的有效性。

第五章 结论与展望

大量的材料领域知识以非结构化文本的形式存储在已发表科研文献中。如何自动从中抽取出数据与知识以指导材料性能优化仍是亟待解决的问题。其关键在于快速获得有监督材料科学文本挖掘数据集及构建高性能的材料科学文本挖掘模型。本文针对材料领域文本挖掘所面临的高质量有监督文本挖掘数据集标注难、材料特殊文本语义特征难以被已有的命名实体识别模型充分融合、材料目标实体及其边界语义信息难以被已有的关系抽取模型感知等问题，分别提出了基于数据增强的有监督材料科学文本挖掘数据集构建方法、基于多层语义特征融合的材料命名实体识别方法和基于实体感知的材料关系抽取方法，并通过实验证明了所提出的方法能快速有效地构建材料科学文本挖掘数据集、识别材料文本中的实体、分类材料文本实体间的关系。进一步，将上述方法应用到 NASICON 型固态电解质材料领域的文本挖掘，可以识别出材料实体以获得描述符，可以挖掘出实体关系以得到材料构效关系知识，从而构建领域知识嵌入的机器学习模型来实现激活能的更精准预测。

5.1 本文主要工作

具体来说，本文的主要工作与贡献如下：

(1) 针对材料有监督文本挖掘数据集标注难的问题，提出了基于数据增强的有监督材料科学文本挖掘数据集构建方法。该方法包括可溯源的文献自动获取、下游任务驱动的文献预处理、材料实体/关系数据标注以及融合材料领域知识的有条件文本数据增强 (cDA-DK 模型)。其中，可溯源的文献自动获取将溯源机制嵌入网络爬虫程序中，以实现可溯源目标文献数据的自动获取；下游任务驱动的文献预处理以材料文本特性为约束选择合适的文本预处理工具，以实现干净、预标注材料文本数据的获取；材料实体/关系数据标注分析了已有的材料科学文本挖掘数据标注场景、在材料专家的指导下设计了有监督数据标签并选择了合适的工具进行标注，以获得部分高质量文本挖掘样本数据；cDA-DK 模型通过对融合材料领域知识的预训练 DistilRoBERTa 模型进行微调，使其能感知材料领域的

特殊性并学习到复杂的上下文语义信息，从而实现在有限手工标注可溯源且高质量样本的基础上自动生成大规模的文本数据。在 NASICON 型固态电解质和无机材料实体识别数据集的对比实验，证明了 cDA-DK 模型能有效生成高质量的文本数据。最后，在增强前后的 NASICON 型固态电解质实体识别数据集上分别训练实体识别模型，增强后的模型精确率、召回率和 F1 分别提高了 5%、3% 和 4%。

(2) 针对材料特殊文本语义特征难以被已有的命名实体识别模型充分融合的问题，提出了多层语义特征融合的材料命名实体识别方法 MatBERT-BiLSTM-CRF。该方法首先通过构建 MatBERT 模型编码词、位置以及句子嵌入信息来充分提取词级别的语义特征；其次，引入 BiLSTM 模型对句子序列进行建模以捕获词的局部上下文语义特征；再次，利用序列标注分类器 CRF 对单词进行标签预测以获取最优的标签序列来实现材料实体的识别。最后，在 NASICON 型固态电解质和无机材料实体识别数据集的对比实验结果表明，我们的模型相较于 BiLSTM-CRF、BiLSTM-CNNs-CRF、BERT 模型，F1 性能指标分别提升了 18%、16% 和 9%。此外，应用 MatBERT-BiLSTM-CRF 模型，从 1808 篇 NASICON 型固态电解质文献中抽取了 106896 个材料实体；进而提出基于重要度计算的描述符筛选策略，成功筛选出 408 个激活能相关的候选描述符；在此基础上，利用数据驱动的机器学习进行了激活能预测，模型的 R^2 性能指标达到了 95%。

(3) 针对材料目标实体及其边界语义信息难以被已有的关系抽取模型感知的问题，本文提出了基于实体感知的材料关系抽取方法 MatBERT-BiGRU-Softmax。该方法首先通过将数据集中目标实体词分别用特殊的封闭标记包裹，使得 MatBERT 模型能充分感知目标实体以提取更丰富的目标实体及句子的语义信息；其次，引入 BiGRU 模型对句子序列进行建模来捕获目标实体周围的局部上下文语义信息；再次，利用 Softmax 函数计算得到候选关系中概率最大的一个来实现材料关系的分类。最后，在 NASICON 型固态电解质和 MatSciRE 材料关系抽取数据集上的对比实验结果表明，我们的模型相较于 WV+CNN+ATT、WV+BiLSTM+ATT、R-BERT 模型，F1 性能指标分别提升了 16%、9% 和 2%。此外，应用 MatBERT-BiGRU-Softmax 模型抽取了 260475 个 NASICON 型固态电解质材料实体关系三元组；进而提出了基于 Neo4j 图数据库的材料知识图谱构建

和基于材料知识图谱的描述符树建立，成功获取了 24 条 NASICON 型固态电解质构效关系知识；藉此，为领域知识嵌入的特征选择方法提供知识，进行了 NASICON 型固态电解质激活能的预测，在两份数据集上模型的 R^2 性能指标较未嵌入知识的模型提高了 1.4% 和 1.5%。

5.2 展望

开发基于自然语言处理的文本挖掘新模型，并将其成功应用于材料领域是一项艰巨的任务，要想针对不同材料问题提出通用完备的方法绝非易事。本文已对所做工作进行了详细的总结和分析，并进一步在阅读大量文献的基础上，提出了几个值得继续探索和研究的方向：

(1) 本文研究了基于数据增强的有监督文本挖掘数据集构建方法。通过将材料知识融入到预训练 DistilRoBERTa 语言模型并对其进行微调来学习材料文本的复杂特性，以实现高质量材料科学文本挖掘数据集的扩充。然而，由于所收集的材料知识有限，因此模型学习到的材料文本特性也有一定的局限性。如在 MatSciBERT（在大量材料语料库上训练的语言模型）上改进，可以扩展到更多材料领域高质量数据的增强。

(2) 本文研究了基于多层语义特征融合的材料实体识别方法。通过设置多层的材料语义特征提取模型，进而准确捕捉材料文本的语义特征并将其融合进行材料实体的分类。由于中文的句子结构比英文的复杂，目前对于中文材料命名实体识别的研究尚未见报道。因此，未来工作可以研究中文材料命名实体识别，从中文文本里发现更多的材料信息，以推动材料命名实体识别的发展。

(3) 本文研究了基于实体感知的材料关系抽取方法。在实体识别的基础上，通过对材料关系抽取数据集中的目标实体进行特殊标记，使模型能够清晰感知并捕获实体感知的单词及句子级别的语义特征，实现了材料实体关系三元组的抽取。然而，管道式的实体关系抽取存在误差传播且不利于重叠关系的提取。实体关系联合抽取可以实现实体和关系的交互抽取，能有效的减少误差传播。因此，未来工作可以研究基于联合抽取的材料知识发现方法，从而更直接地抽取材料文本中蕴含的知识。

参考文献

- [1] Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data[J]. IEEE Intelligent Systems, Vol.24, No.2, 2009, pp.8-12.
- [2] Gebhardt R S, Du P, Wodo O, et al. A data-driven identification of morphological features influencing the fill factor and efficiency of organic photovoltaic devices[J]. Computational Materials Science, Vol.129, 2017, pp.220-225.
- [3] Ghadbeigi L, Sparks T D, Harada J K, et al. Data-mining approach for battery materials[C]. //Proceedings of the 2015 IEEE Conference on Technologies for Sustainability. IEEE, 2015, pp. 239-244.
- [4] Agrawal A, Choudhary A. An online tool for predicting fatigue strength of steel alloys based on ensemble data mining[J]. International Journal of Fatigue, Vol.113, No.8, 2018, pp.389-400.
- [5] Crews J H, Smith R C, Pender K M, et al. Data-driven techniques to estimate parameters in the homogenized energy model for shape memory alloys[J]. Journal of Intelligent Material Systems and Structures, Vol.23, No.17, 2012, pp.1897-1920.
- [6] Srinivasan S, Broderick S R, Zhang R, et al. Mapping chemical selection pathways for designing multicomponent alloys: An informatics framework for materials design[J]. Scientific Reports, Vol.5, No.1, 2015, pp.1-8.
- [7] Ras E J, Rothenberg G. Heterogeneous catalyst discovery using 21st century tools: A tutorial[J]. RSC Advances, Vol.4, No.12, 2014, pp.5963-5974.
- [8] Michopoulos J G, Hermanson J C, Iliopoulos A, et al. Data-driven design optimization for composite material characterization[J]. Journal of Computing & Information Science in Engineering, Vol.11, No.2, 2011, pp.255-267.
- [9] de Oca Zapiain D M, Popova E, Kalidindi S R. Prediction of microscale plastic strain rate fields in two-phase composites subjected to an arbitrary macroscale strain rate using the materials knowledge system framework[J]. Acta Materialia, Vol.141, 2017, pp.230-240.

- [10] Chen W, Xu Y, Jin R, et al. Text mining-based review of articles published in the journal of professional issues in engineering education and practice[J]. Journal of Professional Issues in Engineering Education and Practice, Vol.145, No.4, 2019, pp.06019002.
- [11] Ananiadou S, Thompson P. Supporting biological pathway curation through text mining[C]. //Proceedings of International Conference on Data Analytics and Management in Data Intensive Domains. Springer, 2016, pp. 59-73.
- [12] 魏小梅. 生物事件抽取联合模型研究 [D]; 武汉大学, 2016.
- [13] Lu Z, Hirschman L. Biocuration workflows and text mining: Overview of the biocreative 2012 workshop track ii[J]. Database, Vol.2012, 2012.
- [14] Li Z, Yang Z, Xiang Y, et al. Exploiting sequence labeling framework to extract document-level relations from biomedical texts[J]. BMC bioinformatics, Vol.21, No.1, 2020, pp.1-14.
- [15] Torayev A, Magusin P C, Grey C P, et al. Text mining assisted review of the literature on li-o₂ batteries[J]. Journal of Physics: Materials, Vol.2, No.4, 2019, pp.044004.
- [16] El - Bousiyydy H, Lombardo T, Primo E N, et al. What can text mining tell us about lithium - ion battery researchers' habits?[J]. Batteries & Supercaps, Vol.4, No.5, 2021, pp.758-766.
- [17] Nie Z, Liu Y, Yang L, et al. Construction and application of materials knowledge graph based on author disambiguation: Revisiting the evolution of lifepo4[J]. Advanced Energy Materials, Vol.11, No.16, 2021, pp.2003580.
- [18] Mahbub R, Huang K, Jensen Z, et al. Text mining for processing conditions of solid-state battery electrolytes[J]. Electrochemistry Communications, Vol.121, No.11, 2020, pp.1388-2481.
- [19] Swain M C, Cole J M. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature[J]. Journal of Chemical Information and Modeling, Vol.56, No.10, 2016, pp.1894-1904.
- [20] Huang S, Cole J M. A database of battery materials auto-generated using

- chemdataextractor[J]. *Scientific Data*, Vol.7, No.1, 2020, pp.1-13.
- [21] 孙镇, 王惠临. 命名实体识别研究进展综述 [J].*现代图书情报技术*, 2010, 26(6): 42-47.
- [22] 徐健, 吴振新, 张智雄. 实体关系抽取的技术方法综述[J]. *现代图书情报技术*, Vol.24, No.8, 2008, pp.18-23.
- [23] Zhao X, Lopez S, Saikin S, et al. Text to insight: Accelerating organic materials knowledge extraction via deep learning[C]. //Proceedings of the Association for Information Science and Technology. 2021, pp. 558-562.
- [24] Rocktäschel T, Weidlich M, Leser U. Chemspot: A hybrid system for chemical named entity recognition[J]. *Bioinformatics*, Vol.28, No.12, 2012, pp.1633-1640.
- [25] Leaman R, Wei C-H, Lu Z. Tmchem: A high performance approach for chemical named entity recognition and normalization[J]. *Journal of Cheminformatics*, Vol.7, No.1, 2015, pp.1-10.
- [26] Kim E, Huang K, Jegelka S, et al. Virtual screening of inorganic materials synthesis parameters with deep learning[J]. *npj Computational Materials*, Vol.3, No.1, 2017, pp.1-9.
- [27] Kim E, Huang K, Saunders A, et al. Materials synthesis insights from scientific literature via text extraction and machine learning[J]. *Chemistry of Materials*, Vol.29, No.21, 2017, pp.9436-9444.
- [28] Kim E, Huang K, Tomala A, et al. Machine-learned and codified synthesis parameters of oxide materials[J]. *Scientific Data*, Vol.4, No.1, 2017, pp.1-9.
- [29] Krallinger M, Rabal O, Lourenco A, et al. Information retrieval and text mining technologies for chemistry[J]. *Chemical Reviews*, Vol.117, No.12, 2017, pp.7673-7761.
- [30] Mysore S, Kim E, Strubell E, et al. Automatically extracting action graphs from materials science synthesis procedures[J]. arXiv preprint arXiv:171106872, 2017.
- [31] Weston L, Tshitoyan V, Dagdelen J, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature[J].

- Journal of Chemical Information and Modeling, Vol.59, No.9, 2019, pp.3692-3702.
- [32] He T, Sun W, Huo H, et al. Similarity of precursors in solid-state synthesis as text-mined from scientific literature[J]. Chemistry of Materials, Vol.32, No.18, 2020, pp.7861-7873.
- [33] Allahyari M, Pouriyeh S, Assefi M, et al. A brief survey of text mining: Classification, clustering and extraction techniques[J]. arXiv preprint arXiv:170702919, 2017.
- [34] Rau L F. Extracting company names from text[C]. //Proceedings of the Seventh IEEE Conference on Artificial Intelligence Application. IEEE Computer Society, 1991, pp. 29-32.
- [35] Grishman R. The nyu system for muc-6 or where's the syntax? [R]: New York University Computer Science Department, 1995.
- [36] Collins M, Singer Y. Unsupervised models for named entity classification[C]. //Proceedings of the 1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora. 1999, pp. 100-110.
- [37] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging[C]. //Proceedings of the Conference on empirical methods in natural language processing. 1996, pp. 133-142.
- [38] McCallum A, Freitag D, Pereira F C. Maximum entropy markov models for information extraction and segmentation[C]. //Proceedings of the International Conference on Machine Learning. 2000, pp. 591-598.
- [39] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. //Proceedings of the 18th International Conference on Machine Learning. 2001, pp. 282-289.
- [40] Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what\''s in a name[J]. Machine Learning, Vol.34, No.1-3, 1999, pp.211-231.
- [41] Chieu H L, Ng H T. Named entity recognition with a maximum entropy approach[C]. //Proceedings of the 7th Conference on Natural language learning. 2003, pp. 160-163.
- [42] Lin Y F, Tsai T H, Chou W C, et al. A maximum entropy approach to biomedical

- named entity recognition[C]. //Proceedings of the 4th International Conference on Data Mining in Bioinformatics. Citeseer, 2004, pp. 56-61.
- [43] Yamada H, Kudo T, Matsumoto Y. Japanese named entity extraction using support vector machine[J]. Transactions of Information Processing Society of Japan, Vol.43, No.1, 2002, pp.44-53.
- [44] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, Vol.12, No.11, 2011, pp.2493–2537.
- [45] Mikolov T, Karafiat M, Burget L, et al. Recurrent neural network based language model[C]. //Proceedings of the Conference of the International Speech Communication Association. Makuhari, 2010, pp. 1045-1048.
- [46] Huang Z, Xu W, Yu K. Bidirectional lstm-crf models for sequence tagging[J]. arXiv preprint arXiv:150801991, 2015.
- [47] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, Vol.30, 2017, pp.1-11.
- [48] Yan H, Deng B, Li X, et al. Tener: Adapting transformer encoder for named entity recognition[J]. arXiv preprint arXiv:191104474, 2019.
- [49] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:181004805, 2018.
- [50] Miller S, Fox H, Ramshaw L, et al. A novel use of statistical parsing to extract information from text[C]. //Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics. 2000, pp. 1-8.
- [51] 邓肇, 樊孝忠, 杨立公. 用语义模式提取实体关系的方法 [J].计算机工程, 2007, 33(10): 212-214.
- [52] Kuniyoshi F, Makino K, Ozawa J, et al. Annotating and extracting synthesis process of all-solid-state batteries from scientific literature[J]. arXiv preprint arXiv:200207339, 2020.
- [53] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum

- entropy models for information extraction[C]. //Proceedings of the ACL Interactive Poster and Demonstration Sessions. 2004, pp. 178-181.
- [54] 甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取 [J]. 计算机研究与发展, 2016, 53(2): 284-302.
- [55] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, Vol.3, No.2, 2003, pp.1083-1106.
- [56] Zhou G, Zhang M, Ji D-H, et al. Tree kernel-based relation extraction with context-sensitive structured parse tree information[C]. //Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Citeseer, 2007, pp. 728-736.
- [57] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]. //Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012, pp. 1201-1211.
- [58] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]. //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, pp. 207-212.
- [59] Liu C, Sun W, Chao W, et al. Convolution neural network for relation extraction[C]. //Proceedings of the International Conference on Advanced Data Mining and Applications. Springer, 2013, pp. 231-242.
- [60] Xu K, Fe Ng Y, Huang S, et al. Semantic relation classification via convolutional neural networks with simple negative sampling[J]. Computer Science, Vol.71, No.7, 2015, pp.941-949.
- [61] Zhang Y, Qi P, Manning C D. Graph convolution over pruned dependency trees improves relation extraction[J]. arXiv preprint arXiv:180910185, 2018.
- [62] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]. //Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019, pp. 2361-2364.

- [63] Huang Y, Li Z, Deng W, et al. D - bert: Incorporating dependency - based attention into bert for relation extraction[J]. CAAI Transactions on Intelligence Technology, Vol.6, No.4, 2021, pp.417-425.
- [64] 魏晓, 王晓鑫, 陈永琪, 等. 基于自然语言处理的材料领域知识图谱构建方法[J]. 上海大学学报(自然科学版), 2022, 28(3): 386-398.
- [65] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]. //Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009, pp. 1003-1011.
- [66] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2010, pp. 148-163.
- [67] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]. //Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: technical papers. 2014, pp. 2335-2344.
- [68] Jat S, Khandelwal S, Talukdar P. Improving distantly supervised relation extraction using word and entity based attention[J]. arXiv preprint arXiv:180406987, 2018.
- [69] Greenberg J, Zhao X, Adair J, et al. Hive-4-mat: Advancing the ontology infrastructure for materials science[C]. Research Conference on Metadata and Semantics Research. Springer, 2020, pp. 297-307.
- [70] Kuniyoshi F, Ozawa J, Miwa M. Analyzing research trends in inorganic materials literature using nlp[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2021, pp. 319-334.
- [71] Nie Z, Zheng S, Liu Y, et al. Automating materials exploration with a semantic knowledge graph for li - ion battery cathodes[J]. Advanced Functional Materials, 2022, pp.2201437.
- [72] Gupta T, Zaki M, Krishnan N. Matscibert: A materials domain language model for text mining and information extraction[J]. npj Computational Materials, Vol.8, No.1,

- 2022, pp.1-11.
- [73] Elton D C, Turakhia D, Reddy N, et al. Using natural language processing techniques to extract information on the properties and functionalities of energetic materials from large text corpora[J]. arXiv preprint arXiv:190300415, 2019.
- [74] Mysore S, Jensen Z, Kim E, et al. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures[J]. arXiv preprint arXiv:190506939, 2019.
- [75] Conneau A, Schwenk H, Barrault L, et al. Very deep convolutional networks for text classification[J]. arXiv preprint arXiv:160601781, 2016.
- [76] Kononova O, He T, Huo H, et al. Opportunities and challenges of text mining in materials research[J]. Iscience, Vol.24, No.3, 2021, pp.102155.
- [77] Jiao X, Yin Y, Shang L, et al. Tinybert: Distilling bert for natural language understanding[C]. //Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020. 2020, pp. 1-12.
- [78] Dai X, Adel H. An analysis of simple data augmentation for named entity recognition[J]. arXiv preprint arXiv:201011683, 2020.
- [79] Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv preprint arXiv:190111196, 2019.
- [80] Wu X, Lv S, Zang L, et al. Conditional bert contextual augmentation[C]. //Proceedings of the International Conference on Computational Science. Springer, 2019, pp. 84-95.
- [81] Chawla N V, Bowyer K W, Hall L O, et al. Smote: Synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, Vol.16, 2002, pp.321-357.
- [82] Zhou Z, He Q, Liu X, et al. Rational design of chemically complex metallic glasses by hybrid modeling guided machine learning[J]. npj Computational Materials, Vol.7, No.1, 2021, pp.1-10.
- [83] Branco P, Torgo L, Ribeiro R P. Smogn: A pre-processing approach for imbalanced regression[C]. //Proceedings of the First international workshop on learning with

- imbalanced domains: Theory and applications. PMLR, 2017, pp. 36-50.
- [84] Xiong J, Zhang T-Y. Data-driven glass-forming ability criterion for bulk amorphous metals with data augmentation[J]. Journal of Materials Science & Technology, Vol.121, 2022, pp.99-104.
- [85] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, Vol.63, No.11, 2020, pp.139-144.
- [86] Naaz F, Herle A, Channegowda J, et al. A generative adversarial network-based synthetic data augmentation technique for battery condition evaluation[J]. International Journal of Energy Research, Vol.45, No.13, 2021, pp.19120-19135.
- [87] Jiao X, Yin Y, Shang L, et al. Tinybert: Distilling bert for natural language understanding[J]. arXiv preprint arXiv:190910351, 2019.
- [88] Morris J X, Lifland E, Yoo J Y, et al. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp[J]. arXiv preprint arXiv:200505909, 2020.
- [89] Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training[J]. Advances in Neural Information Processing Systems, Vol.33, 2020, pp.6256-6268.
- [90] Yu A W, Dohan D, Luong M-T, et al. Qanet: Combining local convolution with global self-attention for reading comprehension[J]. arXiv preprint arXiv:180409541, 2018.
- [91] Tanaka F H K d S, Aranha C. Data augmentation using gans[J]. arXiv preprint arXiv:190409135, 2019.
- [92] Malandrakis N, Shen M, Goyal A, et al. Controlled text generation for data augmentation in intelligent artificial agents[J]. arXiv preprint arXiv:191003487, 2019.
- [93] Kumar V, Choudhary A, Cho E. Data augmentation using pre-trained transformer models[J]. arXiv preprint arXiv:200302245, 2020.
- [94] Murray-Rust P, Townsend J A, Adams S E, et al. The semantics of chemical markup language (cml): Dictionaries and conventions[J]. Journal of Cheminformatics, Vol.3, No.1, 2011, pp.1-12.

- [95] Jessop D M, Adams S E, Willighagen E L, et al. Oscar4: A flexible architecture for chemical text-mining[J]. Journal of Cheminformatics, Vol.3, No.1, 2011, pp.1-12.
- [96] Hawizy L, Jessop D M, Adams N, et al. Chemicaltagger: A tool for semantic text-mining in chemistry[J]. Journal of Cheminformatics, Vol.3, No.1, 2011, pp.1-13.
- [97] Wang W, Jiang X, Tian S, et al. Automated pipeline for superalloy data by text mining[J]. npj Computational Materials, Vol.8, No.1, 2022, pp.1-12.
- [98] Sun C C. Materials science tetrahedron—a useful tool for pharmaceutical research and development[J]. Journal of Pharmaceutical Sciences, Vol.98, No.5, 2009, pp.1671-1687.
- [99] Armstrong S, Church K, Isabelle P, et al. Natural language processing using very large corpora[J]. Inverse Document Frequency A Measure of Deviations from Poisson, Vol.26, No.2, 1999, pp.293-294.
- [100] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:190711692, 2019.
- [101] Gokaslan A, Cohen V. Openwebtext corpus [Z].
<http://Skylion007.github.io/OpenWebTextCorpus>; OpenWebTextCorpus. 2019
- [102] Canepa P, Gautam G S, Hannah D C, et al. Odyssey of multivalent cathode materials: Open questions and future challenges[J]. Chemical Reviews, Vol.117, No.5, 2017, pp.4287-4341.
- [103] Kumar P P, Yashonath S. Ionic conduction in the solid state[J]. WILEY - VCH Verlag, Vol.118, No.1, 2006, pp.135-154.
- [104] Chen S, Wu C, Shen L, et al. Challenges and perspectives for nasicon - type electrode materials for advanced sodium - ion batteries[J]. Advanced Materials, Vol.29, No.48, 2017, pp.1700431.
- [105] Masquelier C, Croguennec L. Polyanionic (phosphates, silicates, sulfates) frameworks as electrode materials for rechargeable li (or na) batteries[J]. Chemical Reviews, Vol.113, No.8, 2013, pp.6552-6591.
- [106] Lowe D M, O'Boyle N M, Sayle R A. Efficient chemical-disease identification and

- relationship extraction using wikipedia to improve recall[J]. Database, Vol.2016, No.4, 2016.
- [107] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:13013781, 2013.
- [108] Yimam S M, Ayele A A, Venkatesh G, et al. Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets[J]. Future Internet, Vol.13, No.11, 2021, pp.275-293.
- [109] Loshchilov I, Hutter F. Fixing weight decay regularization in adam[C]. //Proceedings of the International Conference on Learning Representations. 2018, pp. 1-14.
- [110] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, Vol.15, No.1, 2014, pp.1929-1958.
- [111] Prechelt L. Early stopping-but when? [M]. Neural networks: Tricks of the trade. Springer. 1998: 55-69.
- [112] Devlin J, Chang M-W, 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. Association for Computational Linguistics, 2019, pp. 4171-4186.
- [113] Court C J, Cole J M. Auto-generated materials database of curie and néel temperatures via semi-supervised relationship extraction[J]. Scientific Data, Vol.5, No.1, 2018, pp.1-12.
- [114] Kumar S. A survey of deep learning methods for relation extraction[J]. arXiv preprint arXiv:170503645, 2017.
- [115] Miller J J. Graph database applications and concepts with neo4j[C]. //Proceedings of the Southern Association for Information Systems Conference. 2013, pp. 1-7.
- [116] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习 [J]. 硅酸盐学报, 2022, 50(3): 863-876.

- [117] MatSciRE. Material_science_relation_extraction (matscire) [Z].
https://github.com/MatSciRE/Material_Science_Relation_Extraction; Github. 2022
- [118] Park M Y, Hastie T. L1 - regularization path algorithm for generalized linear models[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol.69, No.4, 2007, pp.659-677.
- [119] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[J]. arXiv preprint arXiv:191110422, 2019.
- [120] Shen Y, Huang X-J. Attention-based convolutional neural network for semantic relation extraction[C]. //Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016, pp. 2526-2536.
- [121] Liu Y, Zou X, Ma S, et al. Feature selection method reducing correlations among features by embedding domain knowledge[J]. Acta Materialia, Vol.238, 2022, pp.118195.

作者在攻读硕士学位期间公开发表的论文

- [1]. Yue Liu, Xianyuan Ge, Zhengwei Yang et al. An automatic descriptors recognizer customized for materials science literature [J]. Journal of Power Sources, 545(2022) 231946. (SCI: <https://doi.org/10.1016/j.jpowsour.2022.231946>, 中科院 1 区 top 期刊, IF=9.794, 第二作者, 导师第一作者)
- [2]. Yue Liu, Lin Ding, Xianyuan Ge et al. Domain Knowledge Discovery from Scientific Abstracts on Nickel-based Single Crystal Superalloys [J]. Science China Technological Sciences (已接收, 第三作者, 导师第一作者)
- [3]. 刘悦, 葛献远, 杨正伟, 等. 文本数据的描述符识别方法、装置及介质. CN114997176A [P]. 2022. (专利, 第二作者, 导师第一作者)

作者在攻读硕士学位期间所作的项目

- [1]. 2021 年 01 月-至今，国家自然基金面上项目“领域知识嵌入的机器学习方法研究镍基单晶高温合金蠕变构效关系”（项目编号：52073169）
- [2]. 2021 年 12 月-至今，国家重点研究发展计划项目“数据驱动的新型高性能功能材料智能化研发与应用”子课题“功能材料数据质量提升及专用数据库建设”（项目编号：2021YFB3802101）

致 谢

光阴荏苒，日月如梭，在上海大学研究生求学生涯已经接近尾声。回首研究生求学过往，我经历过挫折与失败，也曾想过逃避与放弃，是导师、同学以及家人的鼓励与陪伴让我走到现在！因此，借此毕业论文撰写的机会，我谨向这段时间里给予我极大帮助的老师、同学和家人表示最诚挚的感谢。

首先，感谢我的导师刘悦教授。刘老师在学术科研和日常生活上都给予我最大的帮助。在学术上，尤其是大小论文的撰写中，刘老师都以认真严格的态度细心教导我，使我在学术水平、表达能力及逻辑思维能力上都得到了提升。在生活中，刘老师不仅教会了我许多如何为人处事的技巧，而且在我遇到挫折的时候能第一时间来问候并给予我帮助和鼓励。刘老师实事求是的学习态度和严谨细致的科研作风将影响和激励我的一生，对我的关心和教诲我更将永远铭记。

其次，感谢材料科学与工程学院的施思齐教授。施老师在日常的工作学习上对我悉心指导，不仅传递给我面对科研的态度，而且还教会我要提升自己的眼界。同时，感谢张博锋老师、吴绍春老师、邹国兵老师和段圣宇老师。在课题组的学术研讨中，四位老师都给了我很多宝贵的建议，他们的建议对我在从事学术研究的过程中提供了莫大的帮助。

另外，还要感谢研究生期间参与的国家自然基金面上项目“领域知识嵌入的机器学习方法研究镍基单晶高温合金蠕变构效关系”（项目编号：52073169）与国家重点研究发展计划项目“数据驱动的新型高性能功能材料智能化研发与应用”（项目编号：2021YFB3802101）对我个人科研能力的提高。感谢上海智能计算系统工程技术研究中心提供的计算资源和技术支持。

在这里要特别感谢同门孙拾雨，为我论文提供了很多想法以及杨正伟师兄、刘大晖师弟、马舒畅师妹，是他们协助我完成了大小论文的撰写以及修改。同时，还要感谢课题组内的其余同门，他们在学术和生活上也都给予了我很大的帮助。

最后，感谢在背后默默付出的家人，他们的鼓励和支持是我前进的动力。

研究生生涯即将结束，未来我会继续努力，“路漫漫其修远兮，吾将上下而求索”！