

中图分类号: TP391

单位代号: 10280

密 级: 公开

学 号: 20721615

上海大学



专业学位硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题 目	基于上下文感知的材料文献文本 与表格信息挖掘及应用方法研究
--------	----------------------------------

作 者 张家旺

学科专业 软件工程

导 师 张瑞

完成日期 二〇二三年五月

姓 名：张家旺


学号：20721615

论文题目：基于上下文感知的材料文献文本与表格信息挖掘及应用方法研究

上海大学

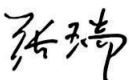
本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主席：

委员：



导 师：

答辩日期：2023.6.7

姓 名：张家旺

学号：20721615

论文题目：基于上下文感知的材料文献文本与表格信息挖掘及应用方法研究

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：张家旺 日期：2023.6.7

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：张家旺 导师签名：张琦 日期：2023.6.7

上海大学工程硕士学位论文

基于上下文感知的材料文献文本与表格信息挖掘及应用方法研究

作者: 张家旺
导师: 张瑞
学科专业: 软件工程

计算机工程与科学学院

上海大学

2023年5月

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

**Research on Text and Table
Information Mining and Application
Methods of Material Literature Based
on Context Awareness**

Candidate: Jiawang Zhang

Supervisor: Rui Zhang

Major: Software Engineering

**School of Computer Engineering and Science
Shanghai University
May, 2023**

摘要

金属材料在现代工业和科技发展中扮演着重要的角色。随着材料基因组计划的深入实施，数据驱动的材料研究方法成为目前的研究热点之一，被认为是材料研究的第四范式。科学文献作为展示材料研究成果的重要方式，其中蕴含着极具价值的材料数据，迫切需要数据挖掘及应用方法从非结构化文献中获取关键信息，为数据驱动的材料研究提供支撑，并从挖掘结果中提取有效特征，构建出最相关的特征集，为材料研究提供更准确的模型。本文对材料文献进行深入分析，并将挖掘结果通过数据驱动的方式应用在材料研究上，提出了基于上下文感知的文献提取方法以及交叉特征选择的材料性能预测方法。本文主要工作和创新点包括：

首先，针对材料文献文本的表述特点和成分表格的结构特征，本文提出一种基于上下文感知的文献信息提取方法，分别对文本和表格信息进行挖掘。使用命名实体识别技术对材料文本进行挖掘，将动态词向量与材料领域静态词向量相融合，使得每个词向量中都包含上下文语境信息和材料领域知识，显著提高材料文本的命名实体识别效果。在不锈钢材料和无机材料的命名实体识别数据集上实验，F1 得分分别为 80.08% 和 88.16%，相较于基线 BERT-BiLSTM-CRF 方法分别提高 2.46% 和 2.19%。针对材料文献中成分表格的结构特点，结合形态学、目标轮廓检测、文本相似度等方法，提出基于传统图像技术的表格识别方法，将成分表格的结构拆解为标题、表头和表体，分别从不同区域中提取出材料名称、元素、元素含量和单位信息。经过实验验证，成分表格识别方法的提取准确率为 85.37%，达到较好的效果。

然后，针对从材料文献上下文中挖掘得到的抗拉强度和材料成分数据，本文提出一种基于文献信息提取的材料性能预测方法。该方法利用 XenonPy 材料信息学库对成分数据进行特征扩充，根据扩充的计算原理，设计一种交叉特征压缩及特征选择方法，筛选得到元素级统计特征和抗拉强度数据，并使用机器学习在这些数据上训练预测模型。实验采用日本国立材料科学研究所公布的数据，结果显示 R^2 得分提高 11.42%，证明所提出的成分特征处理方法能够显著提升模型的预测性能。

最后，以不锈钢为示范材料，将本文提出的文献挖掘和性能预测方法应用在 11,058 篇不锈钢科学文献上。从文献文本中挖掘得到 236 万个材料实体，从文献

表格中提取得到 7970 组材料成分信息，从中筛选出相关数据，对抗拉强度进行数值预测，对抗腐蚀性、延展性、强度和硬度进行变化趋势预测。

本文在多个材料命名实体识别数据集和成分表格数据集上进行实验，并在不锈钢文献上进行应用，证明了所提出的文献挖掘方法与材料性能预测方法的有效性和可行性，为材料性能内禀关系探究、材料数据库建设提供数据源，为数据驱动的材料研究提供一种新的方案。

关键词：文献挖掘；材料性能预测；命名实体识别；表格提取；特征选择

ABSTRACT

Metal materials play an important role in modern industrial and technological development. With the deep implementation of the Materials Genome Initiative, data-driven material research methods have become one of the current research hotspots and are considered the fourth paradigm of materials research. Scientific literature, as an important way to showcase material research results, contains valuable material data that urgently needs data mining and application methods to obtain key information from unstructured literature. This provides support for data-driven materials research and extracts effective features from the mining results to construct the most relevant feature set, providing more accurate models for material research. This paper conducts an in-depth analysis of literature and applies the mining results to material research in a data-driven way. It proposes a literature extraction method based on context awareness and a material property prediction method based on crossed feature selection. The main contributions and innovations of this paper include:

First, based on the characteristics of material literature texts and composition tables, this paper proposes a context-aware method for extracting literature information to mine both text and table information. Named entity recognition technology is used to mine material texts. Dynamic word vectors are fused with material domain static word vectors, making each word vector contain contextual information and material domain knowledge, significantly improving the named entity recognition effect of material texts. In experiments on named entity recognition datasets of stainless steel materials and inorganic materials, the F1 scores are 80.08% and 88.16%, respectively, which are 2.46% and 2.19% higher than the baseline BERT-BiLSTM-CRF method. Based on the structural characteristics of composition tables in material literature, a table recognition method based on traditional image technology is proposed, which combines morphology, target contour detection, text similarity, and other methods. This table recognition method breaks down the composition table structure into title, header, and body, and extracts material names, elements, element content, and unit information from different regions. Experimental results show that the extraction accuracy

of the composition table recognition method is 85.37%, achieving good results.

Second, a material property prediction method based on literature information extraction is proposed for the tensile strength and material composition data obtained from the context of material literature. This method uses the XenonPy materials informatics library to expand the feature of the composition data. Based on the expanded calculation principle, a crossed feature compression and feature selection method is designed to screen the statistically significant elemental features and tensile strength data. Machine learning is then used to train prediction models on these data. The experiment uses data published by the National Institute for Materials Science in Japan. The results show an increase of 11.42% in the R^2 score, demonstrating that the proposed feature processing method for composition can significantly improve the predictive performance of the model.

Finally, using stainless steel as a demonstration material, the literature mining and property prediction methods proposed in this paper are applied to 11,058 scientific literature on stainless steel. A total of 2.36 million material entities are mined from the literature texts, and 7,970 sets of material composition information are extracted from the literature tables. The relevant data are selected for numerical prediction of tensile strength and trend prediction of corrosion resistance, ductility, strength, and hardness.

This paper conducts experiments on multiple datasets for materials named entity recognition and composition tables. It also applies the proposed literature mining method and material property prediction method to stainless steel literature, demonstrating their effectiveness and feasibility. It provides a data source for exploring intrinsic relationships of material properties and constructing material databases, and offers a new approach for data-driven material research.

Keywords: Literature Mining; Material Property Prediction; Named Entity Recognition; Table Extraction; Feature Selection

目 录

第一章 绪论	1
1.1 课题来源	1
1.2 课题背景概述	1
1.3 课题研究的目的和意义	2
1.4 国内外研究现状.....	3
1.4.1 文献挖掘研究概况	3
1.4.2 材料性能预测研究概况	5
1.5 论文主要工作	6
1.6 论文组织结构	7
第二章 相关理论和方法概述	9
2.1 命名实体识别技术.....	9
2.1.1 自然语言处理	9
2.1.2 命名实体识别	10
2.1.3 长短期记忆网络	12
2.1.4 条件随机场	13
2.2 预训练语言模型.....	14
2.2.1 Word2Vec 模型	15
2.2.2 BERT 模型	16
2.3 传统形态学方法.....	17
2.4 材料性能预测	19
2.5 本章小结	21
第三章 基于上下文感知的材料文献信息提取	22
3.1 基于词向量融合的材料命名实体识别.....	23
3.1.1 命名实体识别方法框架	23
3.1.2 动静态词嵌入向量融合	24

3.1.3	材料命名实体提取	26
3.1.4	数据集及标注策略	27
3.1.5	实验环境与评价指标	29
3.1.6	验证动静态词向量融合的有效性	30
3.1.7	融合词向量与单一词向量的对比实验	34
3.1.8	小结	37
3.2	基于传统图像处理的材料成分表格识别	38
3.2.1	成分表格识别方法框架	39
3.2.2	表格文本区域识别	39
3.2.3	材料成分信息提取	42
3.2.4	实验及方法测试	44
3.2.5	小结	46
3.3	本章小结	46
第四章	基于文献信息提取的材料性能预测	48
4.1	性能预测方法	49
4.1.1	性能预测方法框架	49
4.1.2	成分特征扩充方法	50
4.1.3	十字交叉特征压缩及选择方法	51
4.1.4	XGBoost 算法预测抗拉强度	55
4.2	实验分析	56
4.2.1	实验数据准备	56
4.2.2	实验环境与评价指标	57
4.2.3	不同特征预测抗拉强度的对比实验	58
4.2.4	实验结果及分析	60
4.3	本章小结	61
第五章	基于不锈钢文献提取结果的统计分析和性能预测	62
5.1	文献收集与信息提取	62
5.2	提取结果统计分析	62
5.2.1	单实体类别统计分析	62

5.2.2 多实体类别联合分析	66
5.3 不锈钢性能预测.....	67
5.3.1 抗拉强度性能值预测	67
5.3.2 四种性能变化趋势预测	68
5.4 本章小结	70
第六章 总结与展望	71
6.1 总结	71
6.2 展望	72
插图索引	73
表格索引	75
参考文献	76
作者在攻读硕士学位期间发表的论文与研究成果	85
作者在攻读硕士学位期间所作的项目	86
致 谢	87

第一章 绪论

1.1 课题来源

本课题得到国家重点研发计划（编号：2018YFB0704400，2018YFB0704402，2020YFB0704503），国家自然科学基金（面上，编号：52273228），上海市自然科学基金项目（编号：20ZR1419000），上海市“科技创新行动计划”启明星项目（扬帆专项）（编号：23YF1412900），之江实验室科研攻关项目（编号：2021PE0AC02）资助。

1.2 课题背景概述

随着社会文明的发展和科技的进步，材料作为现代社会中不可或缺的基础资源，在航空航天、电子工业、生物医疗等领域都扮演着重要角色，而当今社会需求的变化也对新材料的研发提出更高的要求。传统上新材料的发现依赖于经验法和实验试错法 [1]，需要不断地实验和试验，浪费大量时间和资源。当新材料的发现受到高昂成本和研发周期限制，新材料的制备受到合成技术限制时 [2]，材料研究的第四范式 [3] 和数据驱动研究方法 [4] 的出现，为材料学的发展提供了新的思路和方法。第四范式在实验、理论和计算模拟方面统一前三种材料科学范式 [5]，为材料研究提供了数据分析和计算方法的理论基础，而数据驱动的方法则提供了实现这一理论的技术手段。因此，如何获取海量的优质材料数据是数据驱动方法当前面临的严峻挑战。科学文献作为展现材料研究成果的主要载体，其中包含了研究者的实验设计、研究方法、实验数据等重要信息，因此科学文献可以作为数据驱动方法的重要数据源。然而文献数量庞大，手工提取的文献信息已经无法满足数据驱动的需求，因此材料文献挖掘技术应运而生。材料文献挖掘是以计算机技术和自然语言处理技术为基础，可以从海量材料文献中自动化地提取出重要的信息和知识，以支撑数据驱动方法在材料研究上的应用，对推动材料科学的发展具有重要的意义。

因此，如何对材料文献文本和非文本内容进行挖掘，从中获取有用信息应用于材料研究，已经成为当下研究的重要问题。本文围绕材料文献展开研究，针对文献文本和表格进行信息提取，并将提取结果进行特征处理后应用在材料性能预测上。

1.3 课题研究的目的是和意义

随着网络和数字化技术的发展，数据共享比以往任何时候都更加地便利和快捷 [6]，极大地促进了文献信息的获取和传播。在全球各大科学文献数据库中，存储着大量便携式文档 (PDF) 格式科学文献，以材料科学文献为例，在 Elsevier ScienceDirect 全文数据库中，以“Metal Material”为检索关键词，“2017-2022”为时间区间，可以检索到约 80 万篇科学文献，这些文献都是以非结构化形式予以存储发表，其中包含着大量文本内容（如摘要、正文）和非文本内容（如表格），这些内容中蕴含着大量有价值的信息，包括材料名称、成分、工艺、性能、实验等信息。将文献挖掘与数据驱动的材料研究方法相结合，对大量材料科学文献进行挖掘，使用机器学习处理挖掘得到的数据，可以达到整合借鉴已发表科学文献信息的目的，为新材料低成本改进和设计提供一种可行的方式。

但是文献挖掘涉及多种形式的内容提取，如文字、表格等，这些内容分布在文献的上下文中，这是文献挖掘面临的主要困难和挑战，为了实现更好的文献挖掘，需要采用多种技术和算法。文本是文献中最常见的形式，在现有研究工作中，常利用自然语言处理技术对文献文本进行分类、关键词提取、命名实体识别等任务 [7-10]。这些文献挖掘工作尚未将文献上下文中的文本内容与非文本内容共同挖掘，而非文本也是科学文献内容的重要组成部分，其中包含的重要信息能对文本内容进行有力支撑与补充，因此现有文献挖掘工作存在一定的局限性。此外，现有材料文献挖掘工作常聚焦于关键信息提取、实体标准化或材料数据库构建上 [11-15]，并没有深入探究挖掘结果在材料研究上的应用，而将文献挖掘得到的数据通过数据驱动的方法应用于材料性能预测，可以为材料研究提供新的思路和方案。

对于现有材料性能预测工作而言，其存在的重要问题是过度依赖人工数据集。性能预测需要大量的数据进行模型的训练和验证，但是这些数据十分依赖于专业人员的采集和标注，因此往往需要耗费大量的时间和人力成本，导致人工数据集存在数据来源面狭窄、数据收集困难等缺点。此外，相较于科学文献中最新的材料研究成果数据，人工数据集还存在更新不及时的问题。因此，可以通过文献挖掘的方式，从海量材料文献中收集并整理与性能相关数据，并对数据进行适当处理和特征选择，以构建更为准确、可靠的材料性能预测模型。

综上所述，材料文献挖掘与应用方法尚不完善，材料性能预测依赖人工数据集。

针对上述方法存在的问题和不足，本论文进行了相应的研究和解决，为材料文献挖掘和数据驱动的材料研究提供可行方案。

1.4 国内外研究现状

本论文主要围绕材料文献信息挖掘进行研究，并探究了将挖掘结果应用在材料性能预测上，因此本节将分别介绍文献挖掘和材料性能预测工作的国内外研究现状，这些研究为本工作提供了启发和借鉴。

1.4.1 文献挖掘研究概况

文献挖掘是利用计算机技术对海量文献进行信息提取及分析，是机器学习和自然语言处理等多个领域的交叉研究，目的是发掘文献中的知识和潜在信息，揭示不同文献间的规律和内在联系。近年来，国内外学者在文献挖掘领域取得了不少成果。在文献的自动分类方面 [16-18]，研究者主要采用机器学习和数据挖掘技术，如朴素贝叶斯算法、支持向量机等，根据文献文本的特征将大量文献按照一定的规则或特征划分为若干类别，便于文献检索和管理，提高文献利用价值。在文献文本的情感分析方面 [19-21]，研究者主要采用情感词典、支持向量机、最大熵模型、深度学习等方法，对文献中的情感色彩进行分析，帮助研究者了解文献文本内容的情感倾向和态度，进而深入理解文献内容。在文献的主题挖掘方面 [22-23]，研究者主要采用潜在狄利克雷分配主题模型、词频-逆文档频率、PageRank 图模型等方法，将文本数据转化为主题空间，并根据文本中的词汇共现关系自动发现主题信息。

随着自然语言处理和学科交叉融合的发展，自然语言处理技术被广泛应用于生物医学、材料科学等领域的文献挖掘工作中，研究人员提出许多大规模从文献文本中提取信息并分析的工具或方法 [24]。针对科学文献文本挖掘结果价值性的探究，Westergaard 等人 [25] 对 1500 万篇各类科学文献的正文内容进行文本挖掘，并通过量化基准，评估文献正文的挖掘结果和摘要之间的差异，证明了文献正文挖掘在知识管理和发现上的有效性。Nazemi 等人 [26] 为了探究文献挖掘中摘要与正文是否有相同价值时，采用潜在语义索引和潜在狄利克雷分配方法对计算机和信息科学文献进行挖掘，发现正文中研究主题的一致性值均显著高于摘要内容。

在生物医学文献挖掘领域，Gorrell 等人 [27] 提出生物医学命名实体链接系统

Bio-YODIE, 利用自然语言文本挖掘方法提取医学文本中实体信息, 并将实体与 UMLS 医学语言知识库关联, 以提高对复杂医学词汇的理解度, 为医生的临床决策提供理论支持和病例参考。Dreisbach 等人 [28] 使用自然语言处理和文本挖掘技术提取电子病历文本中症状和治疗方法信息, 并将提取到的患者自述症状映射到标准化医学语言系统中, 构建健康沟通和症状实时评估系统。Weber 等人 [29] 针对化学品和蛋白质实体, 引入大量手动注释的语料库和专用的文本挖掘方法来分析生物医学和临床文本。

在材料科学文献挖掘领域, Weston 等人 [24] 利用文献挖掘技术从 327 万篇材料文献的摘要中挖掘材料名称、对称性或材料相、样品描述符、材料性能、应用、合成方法和表征方法, 将已发表文献的非结构化原始文本映射到允许程式化查询的结构化数据库中, 生成的数据库包含 8,000 万个规范化材料实体, 并基于这个材料实体数据库构建出一个全面的材料科学搜索和知识发现引擎。Guha 等人 [30] 以文献挖掘为基础, 开发自动化工具材料科学信息提取器 MatSciE, 用以从密度泛函理论或第一性原理计算的论文中挖掘材料名、代码、参数、方法和结构信息, 并建立结构化数据库, 易于材料模拟。Kuniyoshi 等人 [12] 针对无机材料文献中包含的材料名称和属性, 提出集成文本序列标注模型和数字标准化模块的文献挖掘框架, 用来分析和预测无机材料发展趋势, 为研究人员提供参考和指导。Shetty 等人 [31] 使用自然语言处理方法进行文献文本挖掘, 从 50 万篇聚合物文献中挖掘文本信息, 并将材料知识嵌入到词向量空间中, 用于材料知识推理和新型聚合物合成预测。Nandy 等人 [32] 为探究有机金属框架 MOFs 的设计潜力, 从 4000 份科学文献中挖掘得到 2000 多种溶剂去除稳定性措施和 3000 个热降解温度, 用以分析稳定性与化学成分、几何结构之间的关系。Cruse 等人 [33] 使用自然语言处理和文本挖掘技术从 500 万篇材料文献中提取金纳米颗粒的合成配方和形态信息, 使用数据驱动的材料研究方法探究影响金纳米颗粒大小及形状的合成参数和潜在机制。

现有针对材料文献的数据挖掘工作尚未将文献上下文中的文本与表格信息共同挖掘。表格作为一种嵌入在文献中的非文本组件, 具有半结构化的特性 [34], 是传达关键信息的重要媒介, 其中包含的重要信息能对文本内容进行有力支撑与补充, 因此现有文献文本挖掘工作存在一定的局限性。

1.4.2 材料性能预测研究概况

数据驱动的材料性能预测是基于现有数据的研究方法，即从大量的数据中挖掘规律和特征，通过建立模型预测新材料的性能。机器学习作为数据驱动方法的一种重要实现手段，它利用已有的数据和统计学方法来构建预测模型，对材料的性能进行预测。已有大量研究人员使用机器学习方法探究材料的性能，Mauro 等人 [35]、Bhasker 等人 [36]、Alcobaca 等人 [37] 利用机器学习方法研究化学成分、转变温度和工艺对玻璃性能的影响，生成预测模型协助功能性玻璃和生物玻璃的研发。Xiong 等人 [38] 在 NIMS 钢数据集上利用随机森林算法在特征集上预测疲劳强度、抗拉强度、断裂强度和硬度，并通过符号回归得到性能值计算公式。Si 等人 [39] 采用高通量粉末冶金技术和机器学习方法，建立了一套完整的多主元素合金的成分设计和基础研究策略，从材料相组成、基体强度、位错运动活化能等方面揭示了不同因素对强度和塑性的影响。Zhang 等人 [40] 基于机器学习构建钢铁材料相位识别方案，可以从显微组织图像中识别出对钢力学性能有重要影响的马氏体相，并从多相微组分中估计其体积分数，估算的马氏体分数是增材制造中预测材料力学性能的基本特征。Geng 等人 [41] 利用化学成分等材料描述符，建立了一种新的数据驱动机器学习模型来预测硼钢的淬透性曲线，并采用最优模型与正交设计相结合的方法，成功地设计出一种淬透性较好的压硬钢。Roy 等人 [42] 采用实验和机器学习结合的方法，预测了基于合金成分和氧化条件的 Fe-Cr-Al 的抗氧化性，发现 Fe-Cr-Al 中的 Mo 元素形成加厚无保护性氧化垢，容易因热膨胀而剥落。但是，现有通过机器学习算法研究材料性能的工作，数据来源面狭窄，数据收集困难，甚至使用人工的方式，并且相较于科学文献中最新材料研究成果数据，这类数据集往往还存在数据陈旧的问题。

在机器学习中，特征处理是一个非常重要的环节。通过对原始数据进行特征提取，可以帮助消除无关特征，构建出强相关特征集，使数据更易于处理，是解决模型效果不佳的有效方法之一。特征选择可以提高模型的效率和准确性，对于实现精确预测和科学发现具有重要的意义。Han 等人 [43] 提出一种数据驱动的方法来估算锂离子电池容量，该方法对电池的电压-放电容量曲线进行分析，以探究电池循环老化的演变模式，并提取具有电化学意义的特征；在特征选择和高斯过程回归的帮助下，建立了一个数据驱动模型来预测锂离子电池的容量。Hu 等人 [44] 运用三种特征选择方法（递归特征消除、过滤法特征选择、套索回归特征选择）和三种机器学习

模型（支持向量回归、线性回归、岭回归），选出包括熔点、晶体结构、门捷列夫序号等材料特性的特征子集来预测无机化合物的弹性性能，为获得无机化合物弹性性能描述符和进一步开发更有效的材料性能预测方法提供参考。Xiong 等人 [38] 利用随机森林和符号回归进行特征选择，使用五种机器学习算法预测疲劳强度、抗拉强度、断裂强度和硬度。对于高维度数据，直接使用机器学习算法进行特征选择会导致计算复杂度的大幅增加，从而降低算法的效率和可行性。此外，特征选择结果可能受到数据分布和样本选取的影响，导致结果不够稳定可靠。因此，为了避免这些问题，常常需要在使用机器学习算法之前，根据高维特征生成的方式进行特征工程，从而提高模型的准确性和可靠性。

综上所述，虽然近年来关于文献挖掘和数据驱动的材料性能预测的研究越来越多，但是只对文献文本挖掘只能获取有限信息，人工收集数据的方式无法为数据驱动的方法快速地提供足够数据，常见的特征选择方法不一定适合所有机器学习任务。因此，如何对材料文献上下文中的文本和非文本内容进行挖掘，从中提取出与材料性能相关的数据，经过特征处理后应用在材料性能预测上是数据挖掘和材料信息学领域急需解决的问题。

1.5 论文主要工作

为了从材料科学文献上下文中挖掘出有价值信息，本文针对材料文本的表述特点和成分表格的结构特征，结合深度学习与传统方法，实现了材料文献上下文内容的挖掘，并将挖掘数据关联地应用在材料性能预测上。本文的主要工作和创新如下：

(1) 利用提出的基于词向量融合的命名实体识别方法对材料文献文本进行实体信息提取。本方法针对材料文本的表述特点，将动态词向量与材料领域静态词向量相融合，使得每个词向量中都包含上下文语境信息和材料领域知识，显著提高了材料文本的命名实体识别效果。在不锈钢和无机材料命名实体识别数据集上实验， $F1$ 得分分别为 80.08% 和 88.16%，与其他方法相比，本方法在材料文本命名实体识别任务中取得更佳的提取效果。

(2) 使用设计的基于传统图像技术的材料成分表格识别方法从文献中提取材料成分信息。针对材料文献中成分表格的结构特点，本方法结合形态学处理、目标轮廓检测、文本相似度计算等方法，对成分表格进行结构拆解，从不同区域中识别并

提取出材料名称、元素、元素含量和单位信息。经过实验验证，本方法成分表格提取准确率为 85.37%，单张平均耗时 4.59 秒，取得较好的准确度和较高的识别速度。

(3) 采用基于文献信息提取的材料性能预测方法对材料抗拉强度值进行预测。本方法将文献上下文中提取结果相结合，以文本中挖掘到的抗拉强度和表格中提取到的材料成分为基础，利用材料信息学库对材料成分进行特征扩充。根据扩充的计算方式，设计出一种交叉特征压缩及选择方法，筛选得到元素级统计特征和抗拉强度数据，并使用机器学习在这些数据上训练预测模型。实验表明该特征处理方法将抗拉强度预测的 R^2 得分提高了 11.42%。

(4) 以不锈钢文献为例，将本文提出的文献提取和性能预测方法应用在 11,058 篇不锈钢科学文献上。从文献中提取得到 236 万个实体信息和 7970 组成分信息，从中筛选出相关数据，对抗拉强度值进行预测，对抗腐蚀性、延展性、强度和硬度的变化趋势进行预测。

1.6 论文组织结构

本文以作者攻读硕士研究生期间参与的课题为基础，针对材料科学文献的特点，研究深度学习与传统方法相结合的文献挖掘方法，以及基于文献提取结果进行材料性能预测的方法。本文总共由六个章节组成，组织结构如图1.1所示，具体如下：

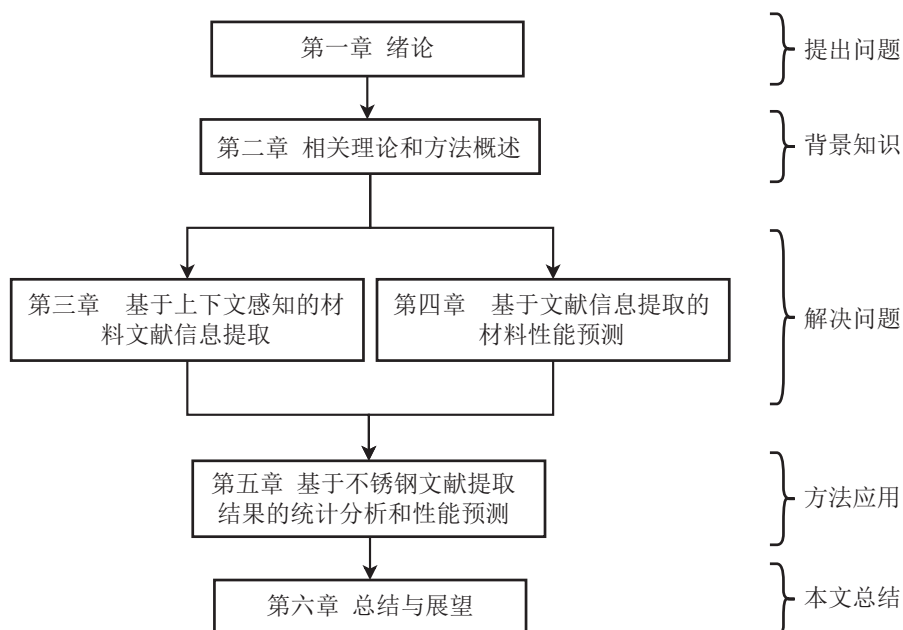


图 1.1 本文组织结构图

第一章介绍课题来源和背景，阐述本课题的研究目的和意义，概述文献挖掘以及性能预测的相关工作，并指明本文的主要工作和创新点。此外还对论文的组织结构进行介绍。

第二章介绍与本职工作相关的背景知识和理论方法。首先，介绍了命名实体识别任务以及涉及的网络结构；其次，对自然语言处理领域经典的预训练语言模型进行介绍；然后，对传统图像处理中的部分方法进行原理解析；最后，介绍材料性能预测的相关理论和所涉及的常用机器学习算法。

第三章提出基于上下文感知的材料文献信息提取方法，分别对文献上下文中的文本和表格进行提取。针对材料文本提出一种 SF-BiLSTM-CRF 的命名实体识别方法，首先对模型结构和原理进行概述，然后详细介绍了动静态词向量融合的过程，最后在不锈钢和无机材料命名实体识别数据集上进行一系列消融和对比实验，并对实验结果进行讨论分析。针对文献表格提出一种基于传统图像处理的材料成分表格识别方法，首先介绍方法的整体框架和各个模块的作用，然后详细叙述表格文本区域识别和成分提取的实现方法，最后对该方法进行实验验证和测试，并分析实验结果。

第四章基于第三章提取得到的材料文献上下文数据，提出一种材料性能预测方法。首先，对提出的性能预测方法进行总体描述；其次，详细介绍材料成分特征的扩充方法；接着，重点介绍交叉特征压缩及选择方法；最后，利用文献挖掘得到的材料成分和抗拉强度数据，以及获取到的日本国立材料科学研究所数据进行实验，验证本方法对成分特征处理以及性能预测的有效性，并对四组实验的结果进行分析和讨论。

第五章以不锈钢科学文献为例，对第三章和第四章方法进行应用。首先，从文献数据库中收集可开源获取的 11,058 篇英文不锈钢文献，并对文献进行预处理；其次，利用第三章中科学文献提取方法对不锈钢文献进行挖掘；然后，对文本中提取到的材料实体进行不同类别的统计分析；最后，利用从文献中提取得到的数据，对不锈钢抗拉强度进行性能值预测，对不锈钢抗腐蚀性、延展性、强度和硬度进行性能变化趋势预测。

第六章对全文工作和成果进行总结回顾，并展望了未来的研究方向。

第二章 相关理论和方法概述

2.1 命名实体识别技术

2.1.1 自然语言处理

自然语言处理 (NLP) 是人工智能和语言学交叉领域的一个重要分支,旨在使计算机理解、分析和生成自然语言,解决与语言相关的复杂且具有挑战性的任务。按照研究目的可以将 NLP 分为两个主要的子领域,即基础研究和应用研究 [45],如图2.1所示。基础研究主要构建模拟人类语言系统所需的基础模块,包括语言分词、词法分析、句法解析、语义分析、语言建模等;应用研究主要包括信息提取(例如命名实体识别、关系抽取)、语言翻译、文档摘要、问答系统、语音识别等,这些研究在实际生活中具有广泛的应用场景,使得人们可以更加方便地利用自然语言进行交流和处理信息。应用研究往往依赖于基础研究的成果,例如分词和词法分析是许多 NLP 任务的先决条件,而句法分析和语义分析则是很多高级任务的基础。在探索这些研究时,NLP 通常涉及到许多机器学习算法,如朴素贝叶斯、支持向量机、最大熵模型、隐马尔可夫模型等。随着计算机技术的快速发展,深度学习的神经网络结构可以自动地提取和学习语言的各种特征和规律,在处理自然语言时具有更强的表达和泛化能力,因此相比与传统机器学习方法,深度学习在 NLP 领域更受欢迎。例如,在机器翻译中基于短语的统计方法已经被神经机器翻译 [46]、编码器-解码器模型、注意力机制 (Attention) 等所取代,并获得更好的性能;早期基于字典和语法规则的命名实体识别方法已经被循环神经网络 (RNN)、Transformer、条件随机场 (CRF) 等所取代 [47]。此外,NLP 中常涉及的深度学习方法还包括长短时记忆网络 (LSTM)、门控循环单元 (GRU)、卷积神经网络 (CNN) 以及预训练语言模型 (如 Word2Vec、GPT、BERT) 等,这些算法和网络都有各自的特点和适用场景,需要根据具体的任务和数据进行调整,以达到更好的效果。

此外,语料库在 NLP 领域是非常重要的数据资源和基础,对 NLP 相关算法和模型的发展起到了至关重要的作用,常见的语料库包括 Wikipedia、Freebase、SNLI、COCO 等,还有一些特定领域的语料库,用于支持该领域的自然语言处理任务,例如 PubMed 生物医学文献库、MaSciP 材料文本语料库等。随着 NLP 技术在不同领域

的广泛应用，新的模型和大规模语料库也在不断涌现，以实现计算机对自然语言的理解和应用。

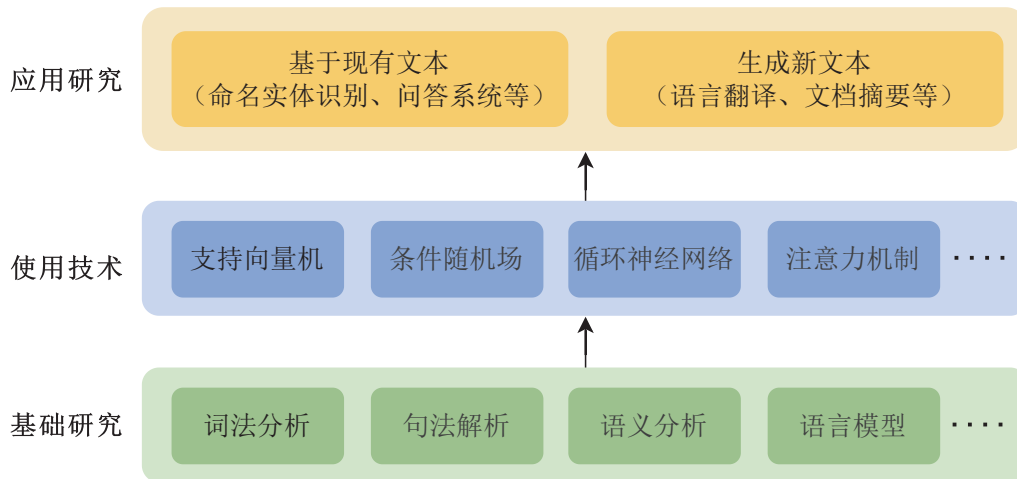


图 2.1 自然语言处理基础研究和应用研究的分类

2.1.2 命名实体识别

命名实体识别（Named Entity Recognition, NER）是 NLP 领域中的一个重要任务，是实现信息抽取、机器翻译、问答系统等任务的基础，其概念最早在第六届消息理解会议（MUC-6）上被提出 [48]，当时计算机科学家开始研究如何从英文新闻的文本中自动识别和提取中具有特定意义的实体名称，例如人名、地名、组织机构名等，一种标准的 NER 总体流程如图 2.2 所示。但最近 NER 也被用于生物医学和材料科学等领域中 [49]。例如，在生物医学领域，NER 被广泛用于从医学文献或临床记录中提取实体信息，包括疾病、药物、治疗方案等，这些实体对于医学诊断、新药物发现、医学知识图谱构建具有重要意义；在材料科学领域，NER 可以从材料文本中提取材料名称、组成、制备方法、性能参数等信息，为材料数据库构建、材料设计和优化等提供支撑，加快研究进度。



图 2.2 一种标准 NER 任务的总体流程

常见的 NER 方法主要包括：基于规则和词典的方法、基于统计的方法和基于深度学习的方法等。基于规则和词典的方法主要依赖于人工设计的规则或定义的字典来识别命名实体，该方法缺点在于需要领域专家和语言专家来编写规则和整理字典，耗费大量人力成本，并且不具备普适性，但在特定领域中这种方法可能会表现得非常有效。基于统计的方法需要根据领域知识和数据特点，设计出适合该 NER 任务的特征，在特征工程的基础上使用传统机器学习算法（如 CRF、SVM 等）在标注好的语料库中训练模型，预测每个单词是否属于命名实体。该方法优点是不依赖手工定义的规则即可对多种类型实体进行识别，并且具有较强的可解释性，其缺点在于需要大量的训练数据，语料库的质量直接关系到模型性能。基于深度学习的方法已经成为 NER 任务的主流方法之一，常用的神经网络包括 RNN、CNN 和 Transformer 等，这些方法能够自动地学习文本中的特征，并且对长文本序列有较好的提取效果。相较于传统的统计学习方法，深度学习提取命名实体更加灵活和准确，但是也需要大量标注数据和计算资源。基于深度学习的 NER 方法通常需要标注好的数据集，并使用预训练语言模型对文本进行词嵌入，然后使用神经网络模型学习命名实体特征，即可使模型拥有识别命名实体的能力。

除了上述方法外，研究人员还提出基于远程监督的方法和基于无监督的方法。基于远程监督的方法利用大规模未标注的语料库和部分标注数据，自动构建训练数据集，从而训练 NER 模型，其主要思想是利用外部的结构化语料库（如 Wikipedia、Freebase 等）和一些启发式规则来处理未标注数据，将其转化为有监督的训练数据。使用远程监督的方法可以节省标注数据的成本，同时可以利用大规模未标注的文本数据来增强 NER 模型的训练效果。但是，远程监督 NER 方法也存在一些挑战，比如怎样解决数据噪声和未知实体类型、如何利用多个知识库等问题，这些都需要进一步的研究和改进。基于无监督的 NER 方法是一种不需要标注数据的 NER 方法，通过聚类、分布式表征学习等无监督方法，自动从未标注的文本中识别命名实体，相比于传统的有监督方法，无监督 NER 方法不需要标注数据，因此能够节省大量的标注成本。但是，由于其使用未标注数据，识别准确率通常较低，需要依赖于后处理和实体消歧等方式来提高其准确性。另外，无监督 NER 方法通常也会受到语料库质量和领域限制等问题的影响，其在实际应用中的效果需要进一步验证和改进。

2.1.3 长短期记忆网络

长短期记忆网络 (LSTM) [50] 是一种递归神经网络, 它具有处理变长序列信息的能力, 并通过门控机制来控制信息的流动, 从而解决了传统递归神经网络中存在的梯度消失问题。LSTM 网络中的每个单元都包含一个单元状态和三个门, 分别为输入门、遗忘门和输出门, 如图2.3所示。

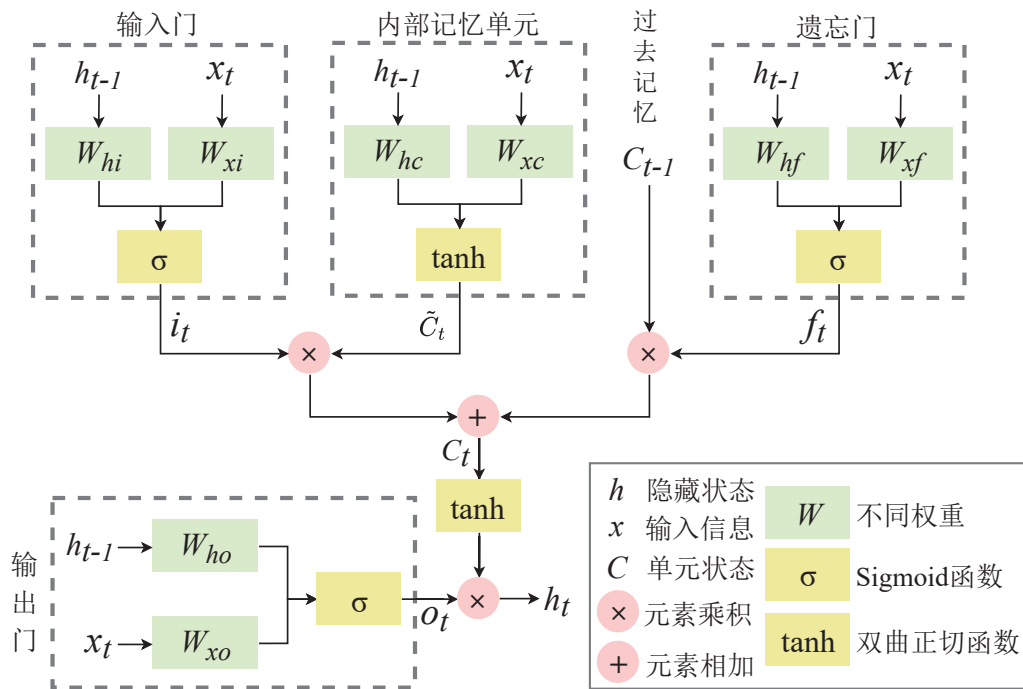


图 2.3 LSTM 神经元细胞结构图

通过控制这些门的开关, LSTM 网络能够根据输入序列中的不同模式来记忆和提取信息, 从而实现对序列数据的建模和预测。LSTM 网络已经被广泛应用于自然语言处理、语音识别、时间序列预测等领域, 成为当前深度学习领域中的重要技术之一, 可以被认为由三个关键阶段组成:

遗忘阶段: 通过遗忘门 f_t 控制过去的单元状态中信息的保留和遗弃, 使用门控机制来控制信息的流动。如公式2.1所示, 其中 x_t 表示当前时间步的输入, h_{t-1} 表示上一个时间步的输出 (隐藏状态), W_{xf} 和 W_{hf} 是遗忘门的权重矩阵, b_f 是遗忘门的偏置向量。

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2.1)$$

输入更新阶段：通过输入门 i_t 控制新输入的信息的添加，如公式2.2所示，其中 W_{xi} 和 W_{hi} 是输入和隐藏状态的权重矩阵， b_i 是输入门的偏置向量， σ 是 sigmoid 激活函数，用于将输入门的值限制在 0 和 1 之间。此外，根据前时间步的输入 x_t 和上一个时间步的输出 h_{t-1} 可以学习到当前单元状态 \tilde{C}_t ，如公式2.3所示，其中 W_{xc} 和 W_{hc} 是当前单元状态的权重矩阵， b_c 是当前单元状态的偏置向量， \tanh 是双曲正切函数，用于将输入的信息映射到-1 和 1 之间。最后更新单元状态 C_t 时，需要考虑短期单元状态 \tilde{C}_t 和获得的长期单元状态 C_{t-1} ，通过遗忘门 f_t 和输入门 i_t 控制不同状态保留的比例，如公式2.4所示，其中 \odot 表示按元素相乘。

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (2.2)$$

$$\tilde{C}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (2.3)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (2.4)$$

输出阶段：通过输出门 o_t 控制当前的状态信息 C_t 哪些需要被输出，如公式2.5所示，其中 W_{xo} 和 W_{ho} 是输出门的权重矩阵， b_o 是输出门的偏置向量，最终得到隐藏状态 h_t 用于后续任务的处理和分析，如公式2.6。

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (2.5)$$

$$h_t = o_t \odot \tanh(C_t) \quad (2.6)$$

LSTM 能够处理文本序列中的长期依赖关系，捕捉到文本中上下文语义信息；通过学习 NER 语料库中的标注数据，预测输入序列中每个单词是否属于预定义的命名实体。在实际应用中，LSTM 已经被广泛被用于 NER 任务，并取得很好的效果，同时也有很多改进的 LSTM 网络结构被提出，例如双向 LSTM 和多层 LSTM，这些结构可以进一步提高 LSTM 网络在 NER 任务中的性能。

2.1.4 条件随机场

条件随机场 (CRF) [51] 是一种基于最大熵模型和隐马尔可夫理论的概率图模型，在 NER 任务中能够捕捉序列中不同实体标签之间的依赖关系，输出可能

性最优的预测序列。假设输入序列为 $\mathbf{x} = (x_1, x_2, \dots, x_m)$ ，对应的标签序列为 $\mathbf{y} = (y_1, y_2, \dots, y_m)$ ，其中 y_i 是 x_i 对应的标签，需要满足的条件概率如下所示：

$$P(y_i | x, y_1, y_2, \dots, y_n) = P(y_i | x, y_{i-1}, y_{i+1}) \quad (2.7)$$

使用 CRF 对输入序列 \mathbf{x} 处理，可以得到状态序列 \mathbf{y} 的概率公式为：

$$P(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{i=1}^m \sum_{k=1}^K \lambda_k t_k(y_i, y_{i-1}, x, i) + \sum_{i=1}^m \sum_{l=1}^L \mu_l s_l(y_i, x, i) \right) \quad (2.8)$$

$$Z(x) = \sum_y \exp \left(\sum_{i=1}^m \sum_{k=1}^K \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i=1}^m \sum_{l=1}^L \mu_l s_l(y_i, x, i) \right) \quad (2.9)$$

其中， $Z(x)$ 是归一化因子， t_k 是转移特征函数， s_l 是状态特征函数，转移特征表示相邻两个标记之间的关系，状态特征表示当前标记和当前输入之间的关系， λ_k 和 μ_l 是特征权重， K 和 L 分别表示转移特征函数和状态特征函数的个数。特征函数的定义是由任务和数据决定的，需要根据任务和数据的特点进行设计。

CRF 的目标是学习一组最优的特征权重 λ ，使得训练数据中的输入序列到输出序列的条件概率最大化，这可以通过最大似然估计来实现。

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \log p(y^{(i)} | x^{(i)}) - \frac{\lambda^2}{2\sigma^2} \quad (2.10)$$

其中， N 是训练集中的样本数量， $y^{(i)}$ 和 $x^{(i)}$ 分别是第 i 个样本的标记序列和观察序列。 λ^2 是正则化系数，用于控制过拟合。CRF 的预测过程是寻找一个最优的输出序列，使得条件概率最大，这可以通过维特比算法 [52] 来实现。

2.2 预训练语言模型

预训练语言模型是使用神经网络模型在大量的无标注文本语料库中学习人类语言的表示方法，包括句法、语义等多层次的语言知识。在 NLP 领域的特定下游任务中，例如 NER、问答任务等，使用规模较小的任务相关语料库对预训练语言模型进行微调，可以提高模型的性能和泛化能力。目前为止，预训练语言模型的发展经历了三个主要阶段，从最初缺乏上下文信息的静态词向量（如 Word2Vec），到基于深

度神经网络的上下文相关词向量（如 ELMo），再到基于 Transformer 和自注意力机制的语言编码器（如 BERT）。

2.2.1 Word2Vec 模型

Word2Vec[53] 是一种无监督学习算法，用于将单词映射到向量空间中，使用固定长度的向量表示单词的语义信息，具有相似语义的单词向量会在向量空间中聚集。Word2Vec 有两种不同的模型：连续词袋模型（CBOW）和 Skip-Gram，两种模型都依赖于独热编码（One-Hot 编码）。

(1) One-Hot 编码

One-Hot 编码是一种将离散数据表示为连续向量的技术，每个离散数据都被表示为具有唯一性的高维向量，该向量的维度等于离散数据集合的大小。在对语料库中的文本内容进行 One-Hot 编码时，使用 0 和 1 进行编码，使得每个字符都有唯一向量表示，例如对语料库中“自然语言”这句话进行编码，“自”对应的向量为 [1, 0, 0, 0]，“然”对应的向量为 [0, 1, 0, 0]，“语”对应的向量为 [0, 0, 1, 0]，“言”对应的向量为 [0, 0, 0, 1]。

(2) CBOW 模型

CBOW 模型的基本思想是利用前后一定范围内的单词预测中间单词，这是一种基于前后文单词的模型，其输入为一组范围内单词的独热编码，输出是中间目标单词的独热编码。例如，如果中心词是“语”，上下文是“自然_言”，则 CBOW 模型会试图预测“语”，具体是通过将上下文中的所有编码向量相加并取平均值得到输入向量，然后将其输入到隐藏层中进行转换，通过全连接层最终输出一个向量，表示预测得到中心词。CBOW 模型结构如图2.4所示，其中 $[x_{1k}, \dots, x_{Ck}]$ 表示 k 位置中心词的前后 C 个单词的 One-Hot 编码，将编码输入矩阵 $W_{V \times N}$ 进行查表， V 为词表大小， N 为隐藏层维度，将查表结果累加求和，并通过 $N \times V$ 维的矩阵 W' 映射到输出层。矩阵 W 和 W' 是由模型训练得到。CBOW 模型的优点包括高效的训练和推理速度，能够处理未知单词并生成高质量的单词嵌入表示。这些嵌入表示可用于各种自然语言处理任务，例如命名实体识别、推荐系统、文本相似度计算等。

(3) Skip-Gram 模型

Skip-Gram 对一个给定的单词前后一定范围内的单词进行建模，来训练这些单

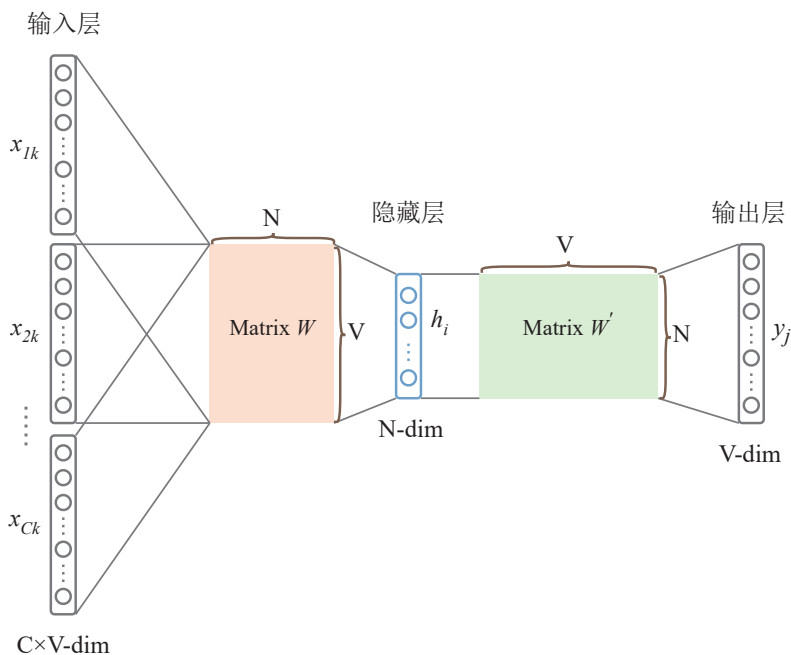


图 2.4 CBOW 模型结构图

词的分布式表示，从而捕捉单词之间的语义关系，生成高质量的词向量。Skip-Gram 模型的结构如图2.5所示，包括：输入层使用一个 One-Hot 编码的词向量，表示当前的中心词；隐藏层利用一个线性变换层，将输入层的词向量映射到一个低维的稠密向量，表示中心词的语义信息；输出层本质是 Softmax 层，将隐藏层的向量作为输入，输出每个词作为上下文词的概率。Skip-Gram 模型的训练过程使用最大化对数似然函数，即最大化给定中心词时，上下文词出现的概率。

综上所述，CBOW 是通过周围单词预测当前单词，而 Skip-Gram 通过当前单词来预测周围单词。在实际应用中，CBOW 模型在大型语料库或者词汇较常见时更加适用，因为它可以更快地训练出词向量；如果语料库较小，或者生僻词较多，则 Skip-Gram 模型能够更准确地训练出词向量。

2.2.2 BERT 模型

BERT (Bidirectional Encoder Representations from Transformers) 模型 [54] 是一种以双向 Transformer 编码器为核心的深度学习语言模型，该编码器具有多个堆叠的自注意力层和前馈神经网络层组成，每个层之间还包括残差连接和层归一化。BERT 可以通过预训练和微调两个阶段来完成 NLP 任务。在预训练阶段，BERT 在大量的未标记的文本语料库上使用两种任务训练模型，语料库主要由英文 BookCorpus[55] 和

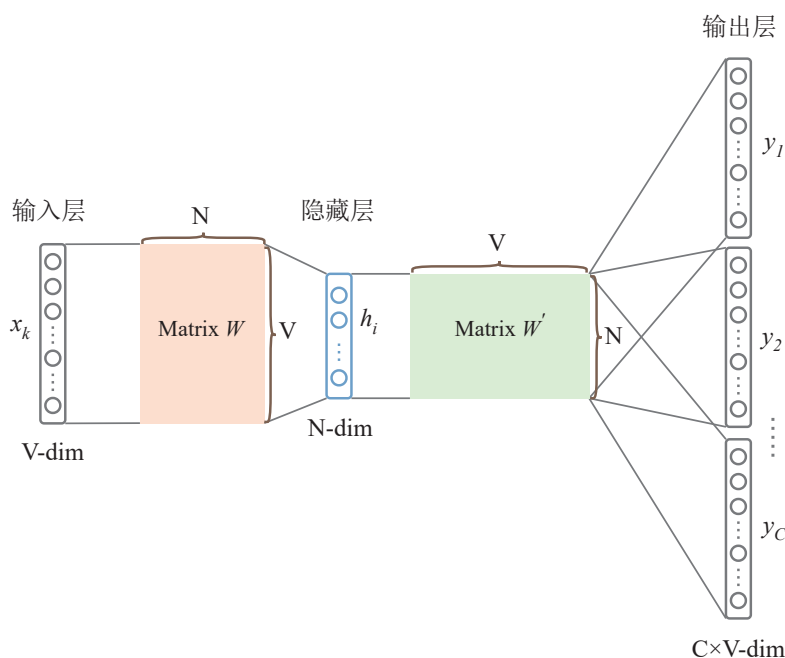


图 2.5 Skip-Gram 模型结构图

Wikipedia 组成，两种任务分别是遮蔽语言建模 (MLM) 和下一句预测 (NSP)，MLM 任务是从输入序列中随机挑选 15% 的单词使用掩码标记替换，训练模型根据上下文信息预测掩码位置处的单词，NSP 任务以两个句子作为输入，并判断这两个句子是否相邻，目的在于帮助 BERT 模型学习句子级别的语义表示。

在预训练完成后，BERT 模型可以根据不同的 NLP 任务进行微调，通常只需要在最后一层编码器上加上额外的输出层，并使用少量任务相关的标签数据进行训练即可。例如，对于文本分类任务，可以在最后一层编码器输出的首位（对应 [CLS] 符号位）加上一个全连接层和一个 Softmax 层；对于问答任务，可以在最后一层编码器输出上加上两个全连接层和一个 Softmax 层，分别预测答案的起始位置和结束位置。BERT 模型在多个自然语言处理任务上都取得了显著地提升，证明了预训练模型在自然语言处理任务中的巨大潜力，并为自然语言处理领域的研究和应用提供了重要的思路。

2.3 传统形态学方法

传统图像处理是数字图像处理的基础，主要使用基于数学模型和信号处理的方法对数字图像进行分析、处理和增强，常用的方法包括图像分割、滤波、增强、边缘检测、目标识别等。在这些方法中，图像通常被视为二维数组，每个元素代表图像

的像素值。传统图像处理方法易于理解和实现，且具有较高的效率和可解释性，因此得到广泛的应用。形态学操作是最常用的传统图像处理方法之一，其基于图像形态学理论，主要用于分析和处理图像中的形状和结构。形态学操作包括膨胀、腐蚀、开运算、闭运算等，可以用于图像增强、图像分割等。

形态学膨胀操作是一种基于集合论的图像处理方法，它可以用来扩大图像中的前景区域，缩小背景区域。设 A 和 B 是二值图像中的两个集合，其中 A 表示图像中的前景像素， B 表示一个结构元素，则 A 对 B 的膨胀操作的数学定义如下：

$$A \oplus B = \{x | (B)_x \cap A \neq \emptyset\} \quad (2.11)$$

其中， $(B)_x$ 表示将结构元素 B 的中心点在 A 上移动后得到的元素集合，上式的意义是，将结构元素 B 在图像的前景像素中滑动，只要 B 与 A 有任何重叠部分，就将 B 加入到膨胀结果中，如图2.6所示。因此，膨胀操作相当于将结构元素 B 映射到图像中的每个前景像素上，从而扩大了前景区域。

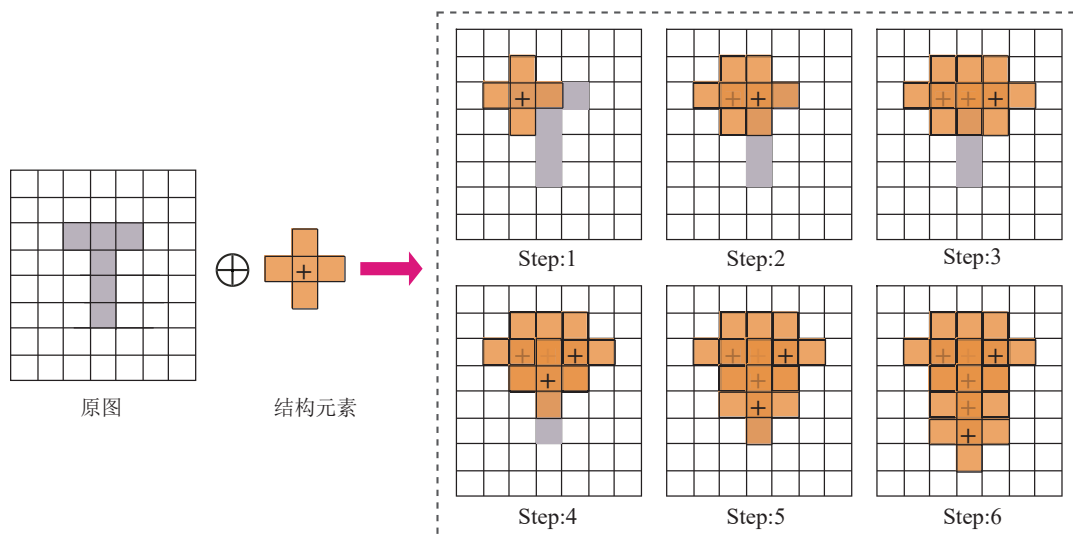


图 2.6 形态学膨胀操作示意图

形态学腐蚀操作也是一种基于集合论的图像处理方法，与膨胀操作效果相反，它可以用来缩小图像中的前景区域，扩大背景区域。设 A 表示图像中的前景像素， B 表示一个结构元素，则 A 对 B 的腐蚀操作表示为：

$$A \ominus B = \{x | (B)_x \subseteq A\} \quad (2.12)$$

其中， $(B)_x$ 表示将结构元素 B 的中心点在 A 上移动后得到的元素集合，上述公式的

含义是将结构元素 B 在 A 中滑动，只有当 B 完全包含在 A 中时，才将 B 的原点（即锚点）加入到腐蚀结果中，如图2.7所示。

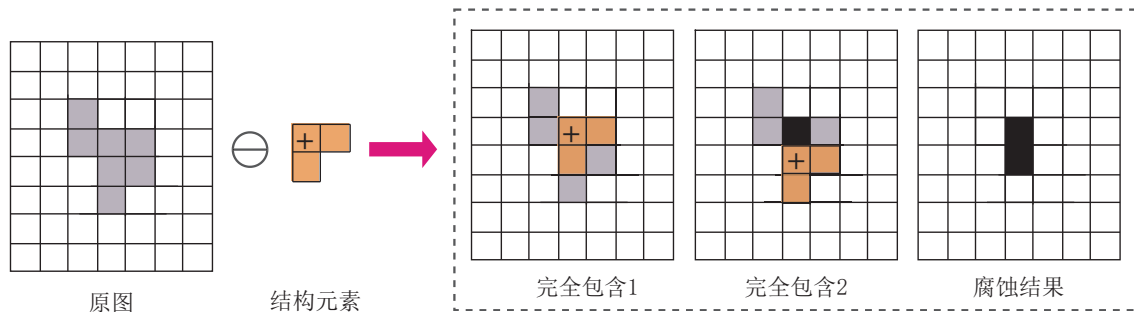


图 2.7 形态学腐蚀操作示意图

此外，开运算是腐蚀和膨胀操作的组合，处理顺序为先腐蚀后膨胀，可以去除物体细节部分或小噪声，同时保持较大物体的形状，其公式见2.13；闭运算本质是先膨胀后腐，可以填补小型空洞或裂缝，同时保留图像中的整体结构，其公式见2.14。

$$A \circ B = (A \ominus B) \oplus B \quad (2.13)$$

$$A \bullet B = (A \oplus B) \ominus B \quad (2.14)$$

2.4 材料性能预测

材料性能预测是指利用计算机模拟或机器学习等方法，根据材料的组成、结构等因素，预测材料在不同条件下的物理、化学或力学等性质，从而为材料的设计和 optimization 提供指导 [56]。相较于传统的实验方法，材料性能预测可以在现有数据基础上，通过计算模拟和机器学习构建数据驱动模型，预测材料的性能和特性，无需多次试验，节约了大量时间和成本。此外，材料性能预测可以探索未知的材料空间，具有发现新的稳定化合物或新的功能材料的可行性，还可以结合高通量筛选技术，快速寻找满足特定需求的候选材料 [57]。

随着计算材料科学的发展，研究人员提出多种性能预测方法，在预测具体材料的性能时可以根据不同的输入和输出选择合适的方法，主要可以分为以下几类：

(1) 基于理论模型的方法：理论模型是基于物理和化学原理的模型，通常建立材料基本方程和数学模型来预测材料的性能和特性。这种方法适用于对材料的基本结

构和物理化学特性已有深入了解的情况下，具有较高的准确性和预测能力，但需要耗费大量时间和精力来建立模型和进行计算。常见理论包括密度泛函理论（DFT）、分子动力学模拟（MD）、含时密度泛函理论（TDDFT）等。

(2) 基于计算模型的方法：计算模型是通过计算机模拟来预测材料的性能和特性，其基本思想是通过计算机模拟材料的微观结构和宏观行为，来预测材料的性能和特性。这种方法具有高效性和可重复性，能够通过调整模型参数来优化预测结果，但模拟和分析的过程需要大量的计算和时间资源。常见计算方法包括有限元方法（FEM）、分子力场（MFF）等。

(3) 基于数据模型的算法：数据模型是通过统计分析已有数据中的关系来预测材料的性能和特性，其基本思想是通过机器学习等技术，从已有数据中学习模型，来预测新材料的性能和特性，机器学习在材料性能预测研究中的一般工作流程如图2.8所示。这种方法适用于数据量大、维度高的情况，能够有效地发现材料的隐藏关系和规律，但是需要定义合适的材料描述符，如成分，结构，元素周期表特征等。常见算法包括决策树、随机森林、支持向量机等。

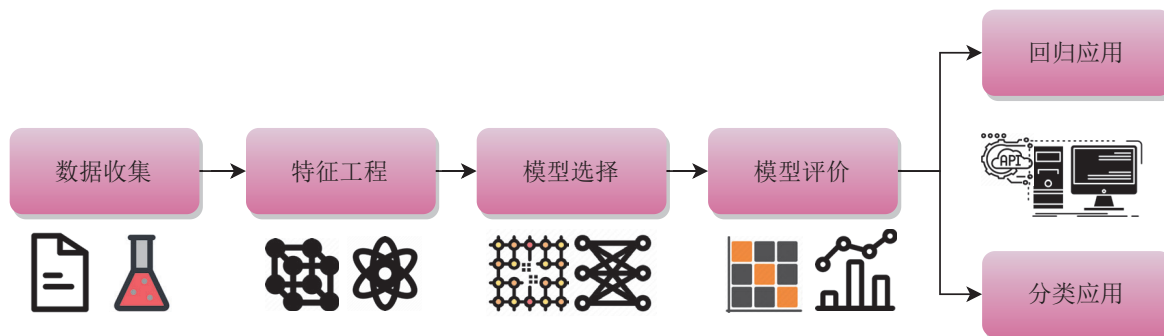


图 2.8 机器学习在材料性能预测研究中的工作流程

此外，也有一些在线的工具和平台提供材料性能预测功能，用户无需干预或调参，输入材料的成分、工艺或结构等元素，即可获得相关性能的预测值，如 NIMS 钢疲劳预测器 [58]、Matbench[59]，MPpredictor[60] 等。

性能预测已经在材料领域得到广泛应用，例如：从大量的候选材料中筛选出具有特定性质的材料，如高熵合金 [61]、超导材料 [57] 等；从材料的图像或谱图中提取特征，识别材料的相或缺陷，已经在纳米材料 [62]，电池材料 [63] 等研究中得到应用；模拟材料在不同条件下原子运动和微观结构演化，如晶体生长、相变 [64] 等。

然而，材料性能预测也面临着一些挑战和局限，数据质量和数量是影响材料性

能预测准确性的关键因素，因为收集到的实验数据可能存在噪声、缺失和系统偏差等问题，从而导致模型的预测结果不够准确；此外材料的性质数据往往是稀疏和不均匀的，对于新型材料或者极端条件下的材料，数据获取成本和难度都很高，限制了模型的构建。另外，描述符的选择、预测模型的选择和优化等都对性能预测工作有重要的影响。

2.5 本章小结

本章首先介绍自然语言处理的基本概念和不同研究类别。接着，阐述命名实体识别在具体领域上的应用以及实现方法，包括基于规则、基于统计和基于深度学习的方法，以及远程监督和无监督方法。随后，详细介绍命名实体识别方法中涉及的长短期记忆网络和条件随机场的结构和原理。本章还介绍了预训练语言模型，例如 Word2Vec 模型、BERT 模型。然后对传统图像处理中形态学的常用方法进行回顾，并对膨胀和腐蚀的原理进行详细介绍。最后，介绍了材料性能预测的定义、方法种类和具体应用。

第三章 基于上下文感知的材料文献信息提取

数据驱动下的新材料研发是目前的研究热点之一，被认为是材料研究的第四范式，可以实现材料的快速研发；而科学文献作为展示材料研究成果的重要方式，蕴含着极具价值的材料研究数据，迫切需要挖掘方法从非结构化文献的上下文中获取关键信息。然而，现有文献挖掘工作常常忽视了包含重要信息的非文本内容如表格。因此，本章针对现有文献挖掘存在的缺点提出一种面向材料科学文献的挖掘方法，分别从材料文献的上下文中挖掘文本信息和表格数据以供分析与应用，如图3.1所示。具体而言，针对材料科学文献中语句表达的特点，提出的命名实体识别模型将通用领域语言模型动态词向量与材料领域语言模型静态词向量相融合，让词向量同时融入上下文信息和材料领域知识，实现材料实体的提取；针对材料文献中材料成分表格的结构特点，提出的基于传统图像处理的成分表格提取方法依靠单元格文本和位置信息，可以提取出材料名称、元素、元素含量和含量单位，并与上下文中提取得到的材料实体相互结合，在材料性能预测上进行应用。

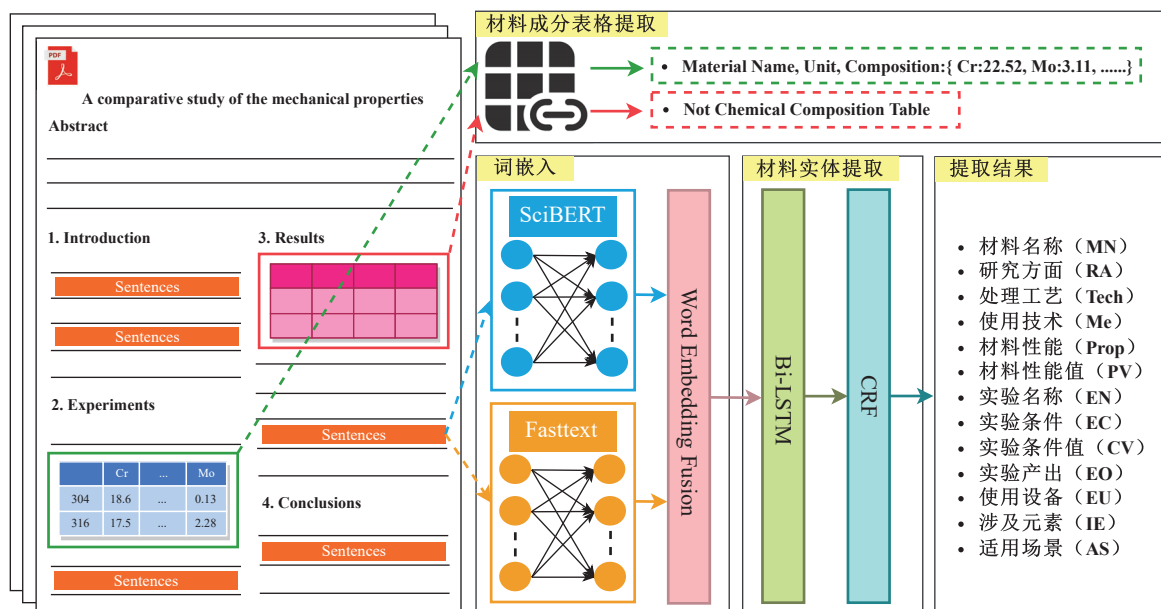


图 3.1 基于上下文感知的材料文献信息提取方法结构图

3.1 基于词向量融合的材料命名实体识别

命名实体识别 (NER) 是自然语言处理中的一个重要序列标注任务, 即将一个句子中的每个词与一个预定义的实体类型进行匹配, 其目的是在文本中识别和提取特定类型的实体。材料文献的文本中, 涉及的实体通常包括材料名称、材料性能、制备方法、实验条件、实验设备等, 通过对这些文本进行命名实体识别, 可以为材料科学领域的研究提供有益的信息支持。

3.1.1 命名实体识别方法框架

本文使用自然语言处理技术对文献文本进行数据挖掘, 提出一种称为 SF-BiLSTM-CRF (SFBC) 的材料文本命名实体识别模型, 能够从材料文本中提取处预定义类别的材料实体。SFBC 模型具体由通用领域动态语言模型 SciBERT[65]、材料领域静态语言模型 Fasttext[66] 和词向量融合模块组成, 模型的下游结构包括双向长短期记忆网络 (Bi-directional Long Short-Term Memory, Bi-LSTM)、全连接层 (Fully Connected layer, FC) 和条件随机场 (Conditional Random Field, CRF), 模型整体结构如图3.2所示。针对材料科学文献中包含众多领域性专业词汇的特点, 现有的大规模预训练语言模型如 BERT[54]、SciBERT[66] 等在对这些句子进行向量表示时缺乏重要的材料领域知识, 而使用材料文本语料库训练得到的语言模型如材料 Word2Vec[67]、Mat2Vec[24]、材料 Fasttext[66] 等对文献句子进行表征时, 对每个词有固定的向量表示, 无法捕捉到词之间的上下文关联性, 不能解决一词多义问题。因此, SFBC 模型将通用领域语言模型 SciBERT[65] 的动态词向量与材料领域语言模型 Fasttext[66] 的静态词向量相融合, 使得每个词向量中都包含上下文信息和材料领域知识, 从而弥补通用领域动态语言模型在表征时材料领域知识的缺失, 解决材料领域静态语言模型在表征时缺少上下文信息的问题。得到的融合词向量被输入到双向 LSTM 层中, 每个词元位置的前向和后向隐藏状态拼接得到特征向量; 特征向量将被输入到 FC 层, 得到每个词元对应各个材料实体类别的可能性得分, 并通过 softmax 函数将得分转换为概率; 将 FC 层输出的概率矩阵和训练得到的转移矩阵作为 CRF 层的输入, 通过维特比 [52] 动态规划算法找到最优的材料实体标签序列作为材料 NER 任务的输出结果。

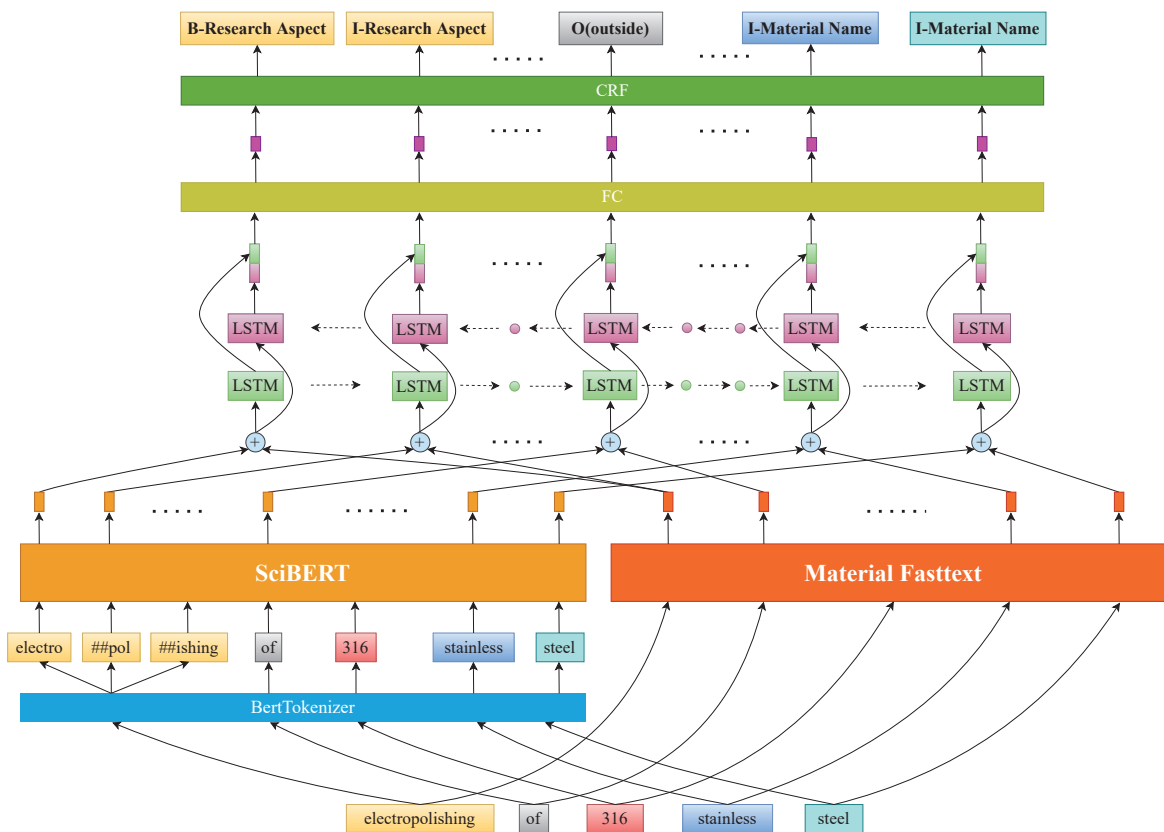


图 3.2 材料命名实体识别模型 SFBC 的结构图

3.1.2 动静态词嵌入向量融合

SFBC 模型使用 SciBERT[65] 预训练语言模型对材料文本执行动态词嵌入，使用材料领域 Fasttext[66] 语言模型执行静态词嵌入。SciBERT[65] 模型以 BERT[54] 模型为基础，使用计算机科学和生物医学两个领域的 114 万篇科学文献语料库进行微调，其基于科学文本的词汇表 SciVocab 大小为 31090。原生 Fasttext[68] 模型是一个文本分类和词向量表示的语言模型，使用固定维度、固定内容的高维向量表示单词信息。Fasttext[68] 模型使用字符级别的 N-gram 表示策略来提高词序列的表达能力，一方面能够优化低频词的表示效果，解决 Out-Of-Vocabulary 问题；另一方面能够捕捉到单词内部的结构信息，不仅考虑单个词，还考虑词的组合。Kim 等人在原生 Fasttext[68] 模型的基础上，使用 250 万篇材料科学文献训练得到用于辅助材料合成研究的材料领域 Fasttext[66] 模型。

在自然语言处理领域中，词元 (token) 代表自然语言的基本构成单元，这些单元可以是单词、短语、词的一部分、符号或者其他任何有意义的单元等。使用词元分析器 tokenizer 将输入句子中的单词词元化，是自然语言处理相关任务的基础步骤，对

下游流程有着重要的影响,其能将自然语言文本分割成可以赋予意义的基本单元,以提取关键信息和特征。使用单词序列 $Seq = [W_1, W_2, \dots, W_n]$ 表示输入的句子,其中 n 为句子中单词的数量, W_i 为句子中的单词, i 代表单词在句子中的序号, $1 \leq i \leq n$ 。将单词序列 Seq 输入到 SciBERT[65] 的词元分析器中,该分析器使用 WordPiece[69] 词元化策略,单词 W_i 将会被切分为若干个词元,表示为:

$$W_i = [T_{i1}, T_{i2}, \dots, T_{im_i}] \quad (3.1)$$

其中 m_i 为单词 W_i 切分得到词元的数量, T_{ij} 表示词元, $m_i \geq 1, 1 \leq j \leq m_i$ 。经过词元化处理,单词序列 Seq 被转化成词元序列,把每个词元添加上句中位置编码信息后,输入到 SciBERT[65] 模型中执行融合上下文信息的词嵌入,通过模型的多层自注意力网络,得到每个词元的隐藏状态向量。因此词元 T_{ij} 被表征为 768 维向量 TV_{ij} ,进而得到单词 W_i 的动态词向量 BV_i ,其本质为 $m_i * 768$ 维向量,表示为:

$$BV_i = [TV_{i1}, TV_{i2}, \dots, TV_{im_i}] \quad (3.2)$$

使用材料领域 Fasttext[66] 模型对句子序列 Seq 中的单词 W_i 进行静态词嵌入,得到 100 维静态词向量 FV_i 。将单词 W_i 的动态词向量 BV_i 与材料领域静态词向量 FV_i 进行词向量融合,具体过程为词元向量 TV_{ij} 逐个与向量 FV_i 叠加,得到 868 维向量 $[TV_{ij}, FV_i]$,最终得到单词 W_i 的 $m_i * 868$ 维融合词向量 WV_i :

$$WV_i = [[TV_{i1}, FV_i], [TV_{i2}, FV_i], \dots, [TV_{im_i}, FV_i]] \quad (3.3)$$

输入句子中每个单词都被表示为融合向量,进而单词序列 Seq 被向量化表示为 $\sum_{i=1}^n m_i * 868$ 维向量 SV_i :

$$SV_i = [WV_1, WV_2, \dots, WV_n] \quad (3.4)$$

序列向量 SV_i 被输入进下游 BiLSTM-CRF 网络中,用以提取材料命名实体。

如图3.2所示,以单词“electropolishing”为例,其被词元分析器切分为三个词元:“electro”、“pol”和“ishing”,每个词元都被 SciBERT[65] 表示为 768 维动态向量。此外,单词“electropolishing”还被材料领域 Fasttext[66] 表示为 100 维静态向量。在三个词元的动态向量中分别拼接上单词“electropolishing”静态向量,即可得到单

词的融合向量，其维度为 3×868 。与此相同，将输入句子中每个单词都表征为融合向量，最终将单词序列向量输入到下游网络结构中提取材料实体。

3.1.3 材料命名实体提取

在 SFBC 模型中，BiLSTM 层的输入维度为 868 维，单向 LSTM 的隐层状态 h_t 和细胞状态 c_t 的特征维度为 434，将正向和反向 LSTM 层的输出拼接成 868 维特征作为 BiLSTM 的输出。将融合得到的 $\sum_{i=1}^n m_i \times 868$ 维词嵌入向量 SV_i 输入到 BiLSTM 结构中，通过双向循环神经网络得到每个词元在文本序列中的上下文特征向量，帮助 SFBC 模型更好地理解句子中的语义知识和句法信息，从而实现更准确地实体识别。经过 BiLSTM 层处理后，以 $\sum_{i=1}^n m_i \times 868$ 维特征向量作为输出，并输入到全连接层 FC 中。

全连接层 FC 能够将每个词元位置的 868 维特征向量映射到 N 维实体类别空间，即 $\mathbb{R}^{868} \mapsto \mathbb{R}^N$ ， N 是每个词元位置候选实体类别的数量，其取决于数据集中预定义的实体类别数量和使用的序列标注策略。例如，数据集中预定义的实体类别数量为 n ，并且使用的标注策略是 BIO[70]，“B”表示实体的开始，“I”表示实体的内部，将“B”和“I”分别拼接在数据集中预定义的实体类别的头部，能够有效地划分实体边界，并区分不同类型的实体，因此 n 类实体转换为 $2 \times n$ 类；“O”表示非实体，数据集中所有未标注的单词被设定为“O”，最终 $N = 2 \times n + 1$ 。经过 FC 层的处理，将 BiLSTM 层得到的特征向量映射为每个词元对应各个标签类别的得分，该得分也称为发射分数，所有词元的发射分数构成发射矩阵，矩阵维度为 $\sum_{i=1}^n m_i \times N$ 。

在 CRF 层中，转移矩阵表示不同实体标签之间转换的概率或得分，为了使转移矩阵更加健壮，加入两个特殊的标签：START 和 END，分别表示句子的开始和结束，因此转移矩阵为 $(N + 2) \times (N + 2)$ 的方阵。转移矩阵可以随机初始化，并在训练过程中更新，从而学习到标签之间的约束条件和依赖关系。以 $\sum_{i=1}^n m_i \times N$ 维的发射矩阵和 $(N + 2) \times (N + 2)$ 维的转移矩阵作为 CRF 层的输入，将各个实体标签类别之间转移概率组合起来计算整个标签序列对应输入文本序列的得分，通过维特比 [52] 动态规划算法求出所有可能路径得分之和，并利用 softmax 函数计算每条路径对应输入文本序列的概率值。在训练时，目标是最大化正确标签序列对应输入文本序列的概率值；在预测时，目标是找出概率值最大的那条路径作为输出结果，路径中的每

个节点为预测得到的每个词元对应的实体标签。最终 SFBC 模型对单词序列 Seq 的命名实体识别结果是长度为 $\sum_{i=1}^n m_i$ 的实体类别标签序列。

以图3.2中的句子为例, 输入的单词序列 [electropolishing, of, 316, stainless, steel] 被切分为词元序列 [electro, pol, ishing, of, 316, stainless, steel], SFBC 模型能够预测每个词元对应的实体标签类别, 输出结果为 [B-Research Aspect, I-Research Aspect, I-Research Aspect, O, B-Material Name, I-Material Name, I-Material Name]。经过后处理, 即可得到“electropolishing”实体类别为研究方面 (Research Aspect, RA), “of”为非实体, “316 stainless steel”实体类别为材料名称 (Material Name, MN)。

3.1.4 数据集及标注策略

为了验证 SFBC 模型在材料文本上提取命名实体的有效性和泛化性, 本文选取两种材料领域 NER 数据集: 不锈钢材料 NER 数据集 (SLSNerData) [71] 和无机材料 NER 数据集 (InorgNerData) [24], 分别对 SFBC 模型进行实验评估。其中, SLSNerData 数据集是在本文中进行收集、标注并整理开源, InorgNerData 数据集为材料领域公开 NER 数据集。

(1) 不锈钢 SLSNerData 数据集

本文从 Elsevier ScienceDirect 数据库中收集得到 250 篇开源不锈钢主题的英文文献, 从文献摘要和正文中获取 2453 条句子, 并使用 Doccano[72] 工具对这些句子进行命名实体序列标注, 以创建 SLSNerData 数据集。其中, 训练集包括 1956 条数据, 测试集包括 467 条数据。数据集中使用 13 种实体标签定义不锈钢实体对象, 包括: 材料名称、研究方面、处理工艺、使用技术、材料性能、性质值、实验名称、实验条件、条件值、实验产出、使用设备、涉及元素和适用场景。每个类别定义如下:

材料名称 (Material Name, MN): 文献中存在若干个材料名称, 这些材料名称以全名或缩写的形式出现, 例如 “after very short aging times in *super duplex stainless steel* and *hyper duplex stainless steel* plates and welds” [73]。

研究方面 (Research Aspect, RA): 研究方面表示科学文献中讨论与研究的重点, 例如 “moreover, from the analysis of *scanning strategy*” [74]。

处理工艺 (Technology, Tech): 处理工艺表示科学文献里提及材料研究过程中所涉及到的工艺, 例如 “some specimens were *cold-rolled* again to investigate the work-

hardening behavior” [75]。

使用技术 (Method, Me): 使用技术表示材料研究实验中使用的研究技术, 例如 “the microstructures were analyzed by *electron backscatter diffraction (EBSD)*” [76]。

材料性能 (Property, Prop): 材料属性表示科学文献中研究的材料性能特征, 例如 “because the two-phase structure of ferrite and austenite possesses good *toughness*, high *strength* and excellent *corrosion resistance*” [77]。

性能值 (Property Value, PV): 性能值表示材料性能的变化趋势或具体性能数值, 例如 “because the two-phase structure of ferrite and austenite possesses *good* toughness, *high* strength and *excellent* corrosion resistance” [77] 或者 “a tensile strength of *627MPa* with a maximum elongation of *50%*” [78]。

实验名称 (Experiment Name, EN): 实验名称表示科学文献中的研究所做的实验, 例如 “the *tensile experiment* was conducted using an MTS testing machine” [79]。

实验条件 (Experiment Condition, EC): 实验条件表示实验中涉及的影响因素、变量等, 例如 “the EIS measurement was performed at 25°C in 3.5wt%-*NaCl solution* over a *frequency range* of 10mHz to 100kHz” [79]。

条件值 (Condition Value, CV): 条件值表示实验条件的具体值, 例如 “the EIS measurement was performed at 25°C in 3.5wt%-NaCl solution over a frequency range of *10Hz to 100kHz*” [80]。

实验产出 (Experiment Output, EO): 实验产出表示从实验中获得的结果, 例如 “the *flow stress curves* under different conditions were depicted in Figure 6” [81]。

使用设备 (Equipment Used, EU): 使用设备表示在各种材料研究实验中使用的实验设备的名称, 例如 “the solidus and liquidus temperatures of each SS-B4C were determined using a *differential scanning calorimeter*” [82]。

涉及元素 (Involved Element, IE): 涉及元素表示研究人员在研究过程中重点探究的元素, 这些元素可能是材料自身成分的一部分, 也可能是研究人员在实验过程中额外添加的, 例如 “*Yttrium*, as a reactive element, has similar features with cerium and laudanum” [83]。

适用场景 (Applicable Scenario, AS): 适用场景是指科学文献中被研究材料的应用场景, 例如 “these results are helpful to promote the application of duplex stainless steel

in the fields of *nuclear power fields*” [77]。

(2) 无机材料 InorgNerData 数据集

Weston 等人 [24] 对 800 篇材料文献的摘要进行手工标注，形成材料领域 NER 数据集，因为这些文献主要对无机材料进行研究，故本文使用 InorgNerData 指代此数据集。该数据集定义了 7 种实体标签类型：无机材料名 (MAT)、对称性或材料相 (SPL)、样品描述符 (DSC)、材料性能 (PRO)、材料应用 (APL)、合成方法 (SMT) 和表征方法 (CMT)，这些实体类别的定义一定程度与材料科学四面体理念（加工、结构、性能和特性）相互关联，不同实体标签的具体含义如表 3.1 所示。InorgNerData 数据集中，训练集包括 4401 条标注数据，测试集包括 546 条标注数据。

表 3.1 InorgNerData 数据集实体标签定义

序号	标签类别	标签含义
1	Inorganic Material (MAT)	无机材料的名称，如：GaN
2	Symmetry/Phase Label (SPL)	材料对称性或相的标签，如：tetragonal
3	Sample Descriptor (DSC)	材料类型或形状的描述符，如：thin films
4	Material Property (PRO)	材料特定的性能，如：band gap
5	Material Application (APL)	材料的应用方面，如：laser diodes
6	Synthesis Method (SMT)	材料合成的方法，如：laser-assisted sol-gel
7	Characterization Method (CMT)	材料表征的方法，如：neutron diffraction

(3) 数据集标注策略

SLSNerData 和 InorgNerData 数据集均使用 BIO (Begin, Inside, Outside) [70] 标注策略，其中“B”表示实体的开始，“I”表示实体的内部，“O”表示实体的外部，即非实体。以 SLSNerData 数据集中的“flow stress”为例，被标注的类别是“Research Aspect”（研究方面，RA），转换成 BIO 格式后，“flow”被标记为“B-Research Aspect”，“stress”被标记成“I-Research Aspect”；以“DIM”为例，标记的类别为“Material Name”（材料名称，MN）；转换成 BIO 格式后，“DIM”标记为“B-Material Name”；其他未被标注的单词都为“O”类。

3.1.5 实验环境与评价指标

(1) 实验环境

实验所用操作系统均为 Ubuntu 18.04.5 LTS，采用 Python 3.8 版本进行编码，使用的深度学习框架版本为 Torch 1.7.1，硬件平台为 Intel(R) Xeon(R) Silver 4210R CPU

@2.40GHz, GPU 型号为 GeForce RTX 3090 TURBO 24G。

(2) 评价指标

按照 Guha 等人 [30] 提出的命名实体识别评价指标的计算方式, 将句子的 NER 手工标注记为真实值 (GT), 模型对句子的 NER 预测结果记为预测值 ($Pred$), 使用准确率 ($Precision, P_{ner}$)、召回率 ($Recall, R_{ner}$) 和 F1 得分 ($F1\text{-score}, F1_{ner}$) 作为实验的评价指标。如图3.3所示, 以 SLSNerData 数据集中句子为例, 样本句子的手工注释结果在图中的 GT 表, 模型预测结果在图中的 $Pred$ 表, 正确的预测结果在图中的 $GT \cap Pred$ 表, “开始”和“结束”表示输入句子中实体的开始和结束位置, “类别”是实体类型。评价指标的计算公式如3.5-3.7所示。

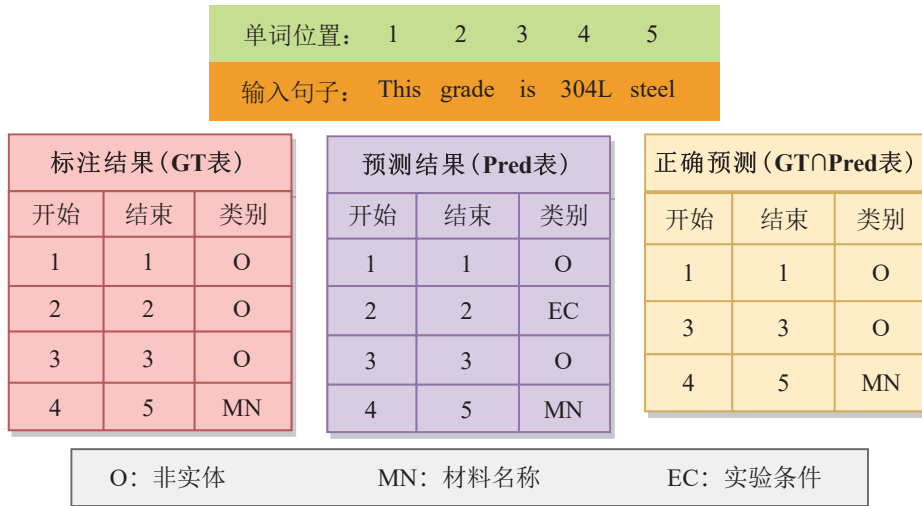


图 3.3 命名实体识别评价指标示意图

$$P_{ner} = \frac{|GT \cap Pred|}{|Pred|} \quad (3.5)$$

$$R_{ner} = \frac{|GT \cap Pred|}{|GT|} \quad (3.6)$$

$$F1_{ner} = \frac{2P_{ner}R_{ner}}{P_{ner} + R_{ner}} \quad (3.7)$$

3.1.6 验证动静态词向量融合的有效性

为了验证通用动态词向量与领域静态词向量相互融合的策略, 在材料领域 NER 任务上的有效性, 本实验选取两种大规模预训练语言模型 BERT[54] 和 SciBERT[66],

与两种材料领域语言模型材料 Word2Vec[67] 和材料 Fasttext[66] 进行词向量融合, 得到四种融合词向量: BERT+ 材料 Word2Vec、BERT+ 材料 Fasttext、SciBERT+ 材料 Word2Vec 和 SciBERT+ 材料 Fasttext。BERT[54] 模型使用 BooksCorpus[55] 以及英语维基百科语料作为训练数据, 单词量分别是 8 亿以及 25 亿 [84]; SciBERT[66] 模型以 BERT[54] 模型为基础, 使用 114 万篇科学文献专门针对科学文本进行训练和微调; 材料 Word2Vec[67] 模型基于原生 Word2Vec 技术 [53], 使用 64 万篇材料合成相关文献训练得到; 材料 Fasttext[66] 模型使用原生 Fasttext[68] 模型的训练方法, 在 250 万篇材料科学文献数据上训练而成。本实验分别在 SLSNerData 和 InorgNerData 数据集上, 使用四种融合词向量模型与四种单一词向量模型进行对比实验, 所有语言模型与下游网络 BiLSTM-CRF 组成命名实体识别模型, 使用 Adam 作为优化器, 学习率为 0.002, LSTM 使用 dropout 正则优化, 参数设置为 0.4, 每个模型训练 500 轮, 批量大小为 20。

SLSNerData 数据集实验结果

各个模型在 SLSNerData 数据集上 NER 总体效果如表3.2所示, 在 13 种不同实体类别上的 F1 得分如表3.3所示。

对不同模型的总体得分情况分析: (1) 使用 BERT+ 材料 Word2Vec 对材料文本进行词嵌入, 生成融合词向量, 并输入到 BiLSTM-CRF 网络中提取材料实体的 F1 得分为 79.32%, 相较于仅使用 BERT 的 F1 得分提高 1.70%, 相较于仅使用材料 Word2Vec 的 F1 得分提高 11.97%; (2) 使用 BERT+ 材料 Fasttext 生成融合词向量提取材料实体的 F1 得分为 77.91%, 相较于仅使用 BERT 的 F1 得分提高 0.29%, 相较于仅使用材料 Fasttext 的 F1 得分提高 9.62%; (3) 使用 SciBERT+ 材料 Word2Vec 生成融合词向量提取材料实体的 F1 得分为 80.01%, 相较于仅使用 SciBERT 的 F1 得分提高 1.03%, 相较于仅使用材料 Word2Vec 的 F1 得分提高 12.66%; (4) 使用 SciBERT+ 材料 Fasttext 生成融合词向量提取材料实体的 F1 得分为 80.08%, 相较于仅使用 SciBERT 的 F1 得分提高 1.10%, 相较于仅使用材料 Fasttext 的 F1 得分提高 11.79%。

对比静态词向量、动态词向量和融合词向量实验, 本文提出在材料 NER 任务中使用融合词向量, 能够充分考虑上下文语境信息和材料知识, 显著提高 NER 的准确度。其中, SciBERT 与材料 Fasttext 词向量相融合效果最佳, 准确率、召回率和 F1 得分分别达到 80.35%、79.80% 和 80.08%, 相较于基线模型 BERT-BiLSTM-CRF 分别

提高 2.10%、2.79% 和 2.46%，500 轮次训练中不同指标的变化见图3.4(a)。

表 3.2 融合与非融合方法在 SLSNerData 上总体得分对比

模型类型	语言模型	P_{ner}	R_{ner}	$F1_{ner}$
静态	材料 Word2Vec (2017)	68.66	66.08	67.35
	材料 Fasttext (2020)	69.04	67.56	68.29
动态	BERT (2018)	78.25	77.01	77.62
	SciBERT (2020)	78.67	79.28	78.98
动静融合	BERT+ 材料 Word2Vec	79.29	79.34	79.32
	BERT+ 材料 Fasttext	76.12	79.80	77.91
	SciBERT+ 材料 Word2Vec	78.09	82.02	80.01
	SciBERT+ 材料 Fasttext	80.35	79.80	80.08

对不同模型在 13 种材料实体上的 F1 得分进行分析：(1) BERT+ 材料 Word2Vec 在 1 种实体上 F1 得分最高，具体为 EU 类别得分 80.99%；(2) BERT+ 材料 Fasttext 在 2 种实体上 F1 得分最高，分别为 RA 类别得分 66.78%，EN 类别得分 91.81%；(3) SciBERT+ 材料 Word2Vec 在 3 种实体上 F1 得分最高，分别为 PV 类别得分 83.16%，EC 类别得分 61.17%，AS 类别得分 73.44%；(4) SciBERT+ 材料 Fasttext 在 4 种实体上 F1 得分最高，分别为 MN 类别得分 87.80%，CV 类别得分 89.31%，EO 类别得分 82.74%，IE 类别得分 82.65%。此外，BERT 在 1 种实体上 F1 得分最高，具体为 Me 类别得分 76.33%；SciBERT 在 2 种实体上 F1 得分最高，分别为 Tech 类别得分 85.19%，Prop 类别得分 80.23%。因此，在 13 个类别中，采用融合策略的模型在 10 类中取得最高得分，SciBERT+ 材料 Fasttext 的融合效果最佳。总体而言，将通用领域动态词向量与材料领域静态词向量融合，能够有效地提高材料文本 NER 准确度。

InorgNerData 数据集实验结果

各个模型在 InorgNerData 数据集上 NER 总体效果如表3.4所示，实验分析如下：(1) 使用 BERT+ 材料 Word2Vec 生成融合词向量，提取 7 类材料实体的 F1 得分为 87.70%，比仅使用 BERT 的 F1 得分提高 1.73%，比仅使用材料 Word2Vec 的 F1 得分提高 11.42%；(2) 使用 BERT+ 材料 Fasttext 生成融合词向量，提取 7 类材料实体的 F1 得分为 87.27%，相较于仅使用 BERT 的 F1 得分提高 1.30%，相较于仅使用材料 Fasttext 的 F1 得分提高 9.21%；(3) 使用 SciBERT+ 材料 Word2Vec 生成融合词向量，提取 7 类材料实体的 F1 得分为 88.02%，相较于仅使用 SciBERT 的 F1 得分提高 1.50%，相较于仅使用材料 Word2Vec 的 F1 得分提高 11.74%；(4) 使用

表 3.3 融合与非融合方法在 SLSNerData 上 13 类实体 F1 得分对比

实体类别	W2V ^①	Fast	BE	Sci	BE+W2V	BE+Fast	Sci+W2V	Sci+Fast
MN	71.97	71.41	83.81	85.17	85.92	82.83	85.80	87.80
RA	53.11	54.20	65.64	61.92	63.56	66.78	65.69	63.34
Tech	70.15	65.22	77.99	85.19	79.82	82.77	82.10	79.47
Me	53.33	60.99	76.33	70.45	68.06	69.56	72.50	70.59
Prop	73.11	74.66	74.18	80.23	78.42	78.85	78.87	78.63
PV	67.58	71.94	75.94	70.88	73.80	73.19	83.16	71.68
EN	81.54	85.48	86.17	88.79	90.42	91.81	90.99	88.97
EC	48.28	54.29	55.95	58.62	57.58	59.09	61.17	56.91
CV	81.61	80.83	85.58	86.46	89.01	86.34	86.06	89.31
EO	71.43	72.00	75.58	79.19	82.19	72.03	82.41	82.74
EU	71.61	64.20	74.09	76.54	80.99	74.89	76.53	74.37
IE	65.95	70.42	77.46	77.03	78.72	77.87	78.32	82.65
AS	60.47	59.52	68.18	70.41	66.67	62.99	73.44	62.30

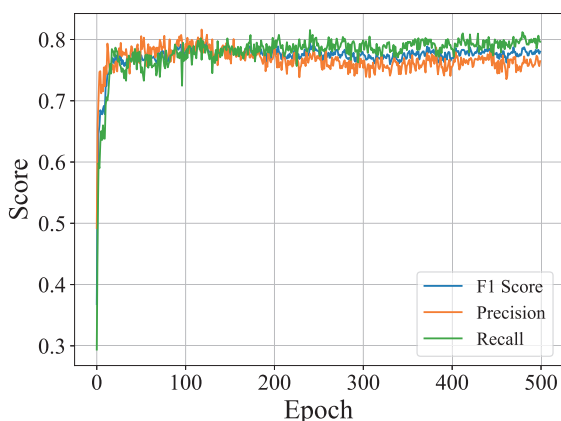
① 本表格中使用“W2V”代替“材料 Word2Vec”，使用“Fast”代替“材料 Fasttext”，使用“BE”代替“BERT”，使用“Sci”代替“SciBERT”

SciBERT+ 材料 Fasttext 生成融合词向量，提取 7 类材料实体的 F1 得分为 88.16%，相较于仅使用 SciBERT 的 F1 得分提高 1.64%，相较于仅使用材料 Fasttext 的 F1 得分提高 10.10%。

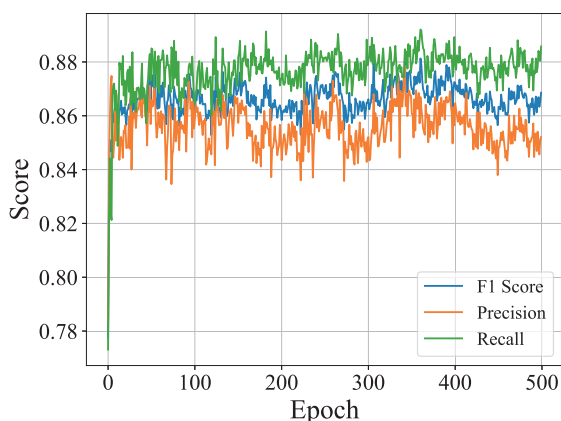
对比三类模型的实验结果可以发现，材料领域静态语言模型受语料库规模和词向量表示时缺乏上下文及语境信息的限制，其性能远不如在大规模语料库上训练得到的预训练语言模型；而 BERT 系列语言模型，虽然使用 transformer 结构和多任务预训练，克服了静态语言模型的缺点，实现基于上下文和句子层面的深度双向语言表征，但是在领域特征明显的自然语言处理任务中，其性能也受到限制。在材料领域 NER 任务中，将通用领域动态词向量与材料领域静态词向量相融合做词嵌入，使得词向量中包含上下文语境信息和领域知识，能够显著提高模型的实体识别效果，其中 SciBERT 与材料 Fasttext 词向量相融合效果最佳，准确率、召回率和 F1 分别达到 88.47%、87.85% 和 88.16%，相较于基线模型 BERT-BiLSTM-CRF 分别提高 2.16%、2.22% 和 2.19%，500 轮次训练中不同指标的变化见图 3.4(b)。

表 3.4 融合与非融合方法在 InorgNerData 上总体得分对比

模型类型	语言模型	P_{ner}	R_{ner}	$F1_{ner}$
静态	材料 Word2Vec (2017)	77.05	75.52	76.28
	材料 Fasttext (2020)	78.21	77.91	78.06
动态	BERT (2018)	86.25	85.69	85.97
	SciBERT (2020)	86.46	86.58	86.52
动静融合	BERT+ 材料 Word2Vec	87.41	87.99	87.70
	BERT+ 材料 Fasttext	86.93	87.62	87.27
	SciBERT+ 材料 Word2Vec	87.05	88.99	88.02
	SciBERT+ 材料 Fasttext	88.47	87.85	88.16



(a) SLSNerData 数据集



(b) InorgNerData 数据集

图 3.4 SFBC 模型在不同数据集上训练曲线。(a) 图为 SLSNerData 数据集, (b) 图为 InorgNerData 数据集。

3.1.7 融合词向量与单一词向量的对比实验

为了验证通用动态词向量与领域静态词向量相互融合的策略,在材料领域NER任务上的有效性,本实验选取 Mat2Vec[24]、ALBERT[85]、ClinicalBERT[86]、BioBERT[87]、MatBERT[88]、MatSciBERT[89]、MatTPUSciBERT[90] 语言模型与 SciBERT+Fasttext 融合模型进行对比,分别在 SLSNerData 和 InorgNerData 数据集上进行材料NER实验。Mat2Vec 材料语言模型是利用 Word2Vec[53] 技术构建的材料领域语言模型,其使用 330 万篇材料、化学、物理等领域的文献摘要进行模型训练;ALBERT 是一种轻量级 BERT 模型,通过嵌入参数因式分解和跨层参数共享等技术,显著降低了模型的参数量,同时保持了与 BERT 相当或更优的性能;ClinicalBERT 是在 MIMIC-III 数据集的临床笔记数据上训练得到的医学领域语言模型;BioBERT 基于 BERT 模型,使用 PubMed 和 PMC 等大规模的生物医学语料库训练得到的模型;

MatBERT 材料语言模型是在 BERT 的基础上，使用包含 5000 万个段落的语料库进行训练与微调的材料语言模型；MatSciBERT 材料语言模型是以 SciBERT 为基础的材料语言模型，其使用 15 万篇材料文献进行训练和微调；MatTPUSciBERT 材料语言模型使用 SciBERT 模型对参数初始化，并在张量处理单元（TPU）上使用 70 万篇材料文献进行训练和微调。实验中，所有语言模型与下游网络 BiLSTM-CRF 组成 NER 方法，使用 Adam 作为优化器，学习率为 0.002，LSTM 使用 dropout 正则优化，参数设置为 0.4，每个模型训练 500 轮，批量大小为 20。

SLSNerData 数据集实验结果

各个模型在 SLSNerData 数据集上 NER 总体效果如表3.5所示，在 13 种不同实体类别上的 F1 得分如表3.6所示。本文提出的 SciBERT 词向量与材料 Fasttext 词向量相融方法的 F1 得分为 80.08%，比 Mat2Vec 得分高 10.35%，比 ALBERT 得分高 3.2%，比 ClinicalBERT 得分高 3.53%，比 BioBERT 得分高 2.7%，比 MatBERT 得分高 0.62%，比 MatSciBERT 得分低 0.14%，比 MatTPUSciBERT 得分高 0.39%。

表 3.5 不同方法在 SLSNerData 上总体得分对比

模型类型	语言模型	P_{ner}	R_{ner}	$F1_{ner}$
静态	Mat2Vec (2019)	69.68	69.76	69.73
	ALBERT (2019)	77.96	75.83	76.88
动态	ClinicalBERT (2019)	77.15	75.96	76.55
	BioBERT (2020)	77.53	77.24	77.38
	MatBERT (2021)	78.83	80.11	79.46
	MatSciBERT (2022)	79.96	80.48	80.22
	MatTPUSciBERT (2022)	78.98	80.42	79.69
动静融合	SciBERT+ 材料 Fasttext	80.35	79.80	80.08

通过对比实验可以发现，融合词向量在材料 NER 任务上的性能普遍高于 BERT 系列的非材料领域语言模型，而相较于材料领域的 BERT 模型（MatBERT、MatSciBERT 和 MatTPUSciBERT），融合词向量策略无需在大量材料文献语料库中进行训练和微调，即可取得近似或者更优的性能。对 13 种材料实体具体的 F1 得分进行分析，SciBERT+ 材料 Fasttext 在 4 种实体上取得 F1 最高分，分别为 MN (87.80%)、EN (88.97%)、CV (89.31%) 和 EO (82.74%)；MatTPUSciBERT 在 3 种实体上取得 F1 最高分，分别为 Tech (84.19%)、Me (71.72%) 和 EU (78.15%)；MatSciBERT 在 2 种实体上取得 F1 最高分，分别为 Prop (82.12%) 和 EC (63.41%)；MatBERT 在

2 种实体上取得 F1 最高分，分别为 PV (72.91%)、IE (83.33%) 和 AS (72.27%)；BioBER 在 1 种实体上取得 F1 最高分，为 RA (66.58%)。由此也可以发现，与材料领域最新的预训练语言模型相比，融合词向量策略无需使用大量语料库对语言模型进行训练，就可取得近似的提取效果，显著减少时间和资源的消耗。

表 3.6 不同方法在 SLSNerData 上 13 类实体 F1 得分对比

实体类别	M2V ^①	ALB	ClinB	BioB	MB	MSB	MTB	SciBERT+ 材料 Fasttext
MN	75.00	82.15	82.17	81.82	86.05	85.16	86.85	87.80
RA	53.58	62.48	66.51	66.58	66.13	66.54	61.92	63.34
Tech	71.75	80.75	82.38	78.96	84.16	83.15	84.19	79.47
Me	61.44	72.13	67.16	66.76	70.14	70.99	71.72	70.59
Prop	77.35	78.21	79.42	79.62	78.81	82.12	79.49	78.63
PV	71.72	67.84	69.52	69.19	72.91	71.29	72.54	71.68
EN	86.61	87.41	88.42	88.32	88.54	88.89	88.77	88.97
EC	55.22	59.31	57.36	59.36	58.13	63.41	60.71	56.91
CV	78.65	84.55	82.18	85.99	85.09	87.26	87.87	89.31
EO	71.70	70.99	71.37	74.34	79.10	81.95	76.61	82.74
EU	65.38	69.10	66.86	71.72	76.00	76.81	78.15	74.37
IE	67.41	80.53	81.42	81.48	83.33	82.19	75.90	82.65
AS	70.83	64.46	57.58	68.57	72.27	69.11	69.70	62.30

① 本表格中使用“M2V”代替“Mat2Vec”，使用“ALB”代替“ALBERT”，使用“ClinB”代替“ClinicalBERT”，使用“BioB”代替“BioBERT”，使用“MB”代替“MatBERT”，使用“MSB”代替“MatSciBERT”，使用“MTB”代替“MatTPUSciBERT”

InorgNerData 数据集实验结果

各个模型在 InorgNerData 数据集上 NER 总体效果如表3.7所示。实验中，将一种材料领域静态语言模型 (Mat2Vec) 和三种材料领域动态语言模型 (MatBERT、MatSciBERT 和 MatTPUSciBERT) 与本文提出的融合方法 (SciBERT+ 材料 Fasttext) 进行对比，使用这五种 NER 方法从 InorgNerData 数据集中提取 7 类材料实体。使用 MatSciBERT 模型做词嵌入的方法提取表征方法 (CMT) 实体的效果最佳，F1 得分为 88.95%；使用 MatTPUSciBERT 的方法提取无机材料名称 (MAT)、材料应用 (APL) 和合成方法 (SMT) 实体的效果最佳，F1 得分分别为 93.2%、86.22% 和 83.49%；使用 SciBERT+ 材料 Fasttext 的方法提取对称性或材料相 (SPL)、样品描述符 (DSC) 和材料性能 (PRO) 实体的效果最佳，F1 得分分别为 84.57%、91.32% 和 82.76%。从不同方法在 InorgNerData 数据集上的总体得分来看，MatSciBERT 模型生成的词向量在

NER 任务上效果最佳, F1 得分为 88.39%, 而本文方法 SciBERT+ 材料 Fasttext 的 F1 得分为 88.16%, 仅比最高分低 0.23% 分。从 7 种实体类别的 F1 得分分析, SciBERT+ 材料 Fasttext 在 3 个类别上得分最高。因此, 本文方法具有显著优势, 无需在材料文献语料库中继续训练语言模型就可以获得与目前最佳方法相当的性能, 避免重新训练模型的开销, 节省大量的时间和计算资源。

表 3.7 不同方法在 InorgNerData 上准确率、召回率和 F1 得分对比

语言模型	评价指标	MAT	SPL	DSC	PRO	APL	SMT	CMT	SUM ^①
Mat2Vec	P_{ner}	85.69	76.07	75.65	81.14	78.04	64.12	80.81	80.87
	R_{ner}	89.71	51.45	82.38	68.81	71.98	74.34	77.42	79.53
	$F1_{ner}$	87.65	61.38	78.88	74.47	74.89	68.85	79.08	80.19
MatBERT	P_{ner}	90.12	82.42	88.91	84.66	77.78	79.15	84.27	86.72
	R_{ner}	95.43	86.71	88.76	78.48	81.46	82.30	91.61	88.41
	$F1_{ner}$	92.70	84.51	88.83	81.45	79.58	79.58	87.79	87.56
MatSciBERT	P_{ner}	90.54	78.07	87.42	83.47	89.85	84.76	86.77	89.15
	R_{ner}	95.66	84.39	92.11	78.77	79.14	81.74	91.22	87.63
	$F1_{ner}$	93.03	81.11	89.71	81.06	84.33	83.24	88.95	88.39
MatTSBERT ^②	P_{ner}	90.66	81.81	86.33	83.58	88.99	84.81	81.2	87.84
	R_{ner}	95.90	83.24	91.10	77.88	83.62	82.19	91.94	88.12
	$F1_{ner}$	93.20	82.52	88.65	80.63	86.22	83.49	86.23	87.98
SciBERT+Fast ^③	P_{ner}	89.51	84.66	92.64	85.92	79.59	80.83	87.02	88.47
	R_{ner}	96.13	84.49	90.02	79.68	80.33	82.96	88.62	87.85
	$F1_{ner}$	92.71	84.57	91.32	82.76	79.96	81.89	87.81	88.16

① 模型在 InorgNerData 数据集上总体准确率、召回率和 F1 得分

② 语言模型的完整名称为 MatTPUSciBERT

③ 语言模型的完整名称为 SciBERT+ 材料 Fasttext

3.1.8 小结

在面向材料文本的命名实体识别方法中, 基于通用动态语言模型和领域静态语言模型的特点和缺点, 提出一种动静态词向量融合的命名实体识别方法 SFBC, 阐述了不同语言模型词向量融合的原理并详细介绍向量融合的构建流程。该方法针对材料科学文献文本的特点, 利用 SciBERT 模型能够更好地捕捉科学领域的词汇、语法、语义和上下文信息的优点, 与材料领域 Fasttext 模型具有材料领域知识的优点相结合, 更好对材料文本进行向量表征, 用以提取材料实体。本节中使用 SLSNerData 和 InorgNerData 两种材料 NER 数据集进行实验, 结果表明 SFBC 模型在两种数据集

上的 P、R 和 F1 得分均高于基线模型 BERT-BiLSTM-CRF，F1 得分提高 2 分以上。此外，相较于材料领域最新的语言模型 MatSciBERT，将 SciBERT 与材料 Fasttext 融合用于材料命名实体识别能取得近似的效果，在两个数据集上 F1 得分仅低 0.14 和 0.23 分，但本文方法具有显著的优势，它无需收集数十万篇甚至百万篇材料文献对语言模型进行训练就可以取得近似效果，节省了大量的时间和计算资源。实验证明了将通用领域语言模型的动态词向量与材料领域语言模型的静态词向量相融合，使得每个词向量中都包含上下文信息和材料领域知识，有助于提高材料文本命名实体识别效果。

3.2 基于传统图像处理的材料成分表格识别

表格识别是指从图像或文档中检测和提取表格的内容和结构的过程，材料成分表格能够详细地记录材料的组成成分，包括元素、化合物、含量、比例等信息，这对于理解材料的性质和特性非常重要。大多数科学文献以便携式文档格式（Portable Document Format, PDF）呈现，因此本文的材料文献挖掘方法以 PDF 格式文献为挖掘对象。PDF 科学文献中的表格无法直接获取，材料成分表格识别方法将 PDF 文献逐页转化为图像格式，使用图像目标检测 YOLOv3[91] 算法训练表格区域检测模型，实现从文献版面图像中识别并截取出表格区域，再使用传统图像处理和形态学方法对图像型材料成分表格进行识别，并从中提取出材料成分信息，见示意图 3.5。

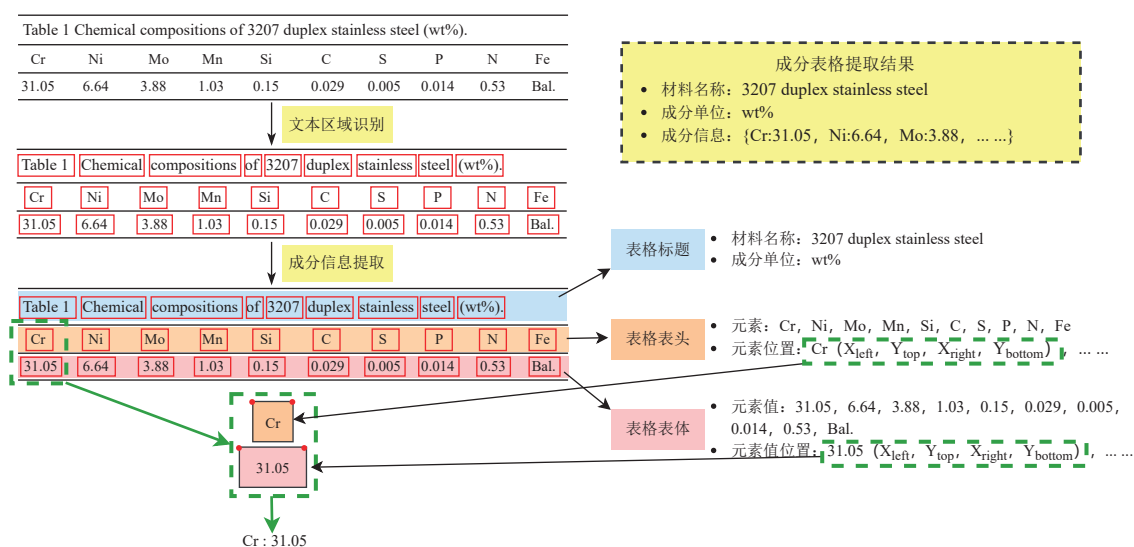


图 3.5 材料成分表格识别方法示意图

3.2.1 成分表格识别方法框架

材料成分表格识别方法的整体流程如图3.6所示，使用识别方法可以从图像型表格中提取材料名称、元素、元素含量和成分单位，并以 JSON 键值对格式存储，便于计算机读取成分信息。识别方法总体可以分为五个步骤：表格图像预处理、文本区域检测、成分表格筛选及分类、表格结构解析和成分信息提取。文本区域检测目的是从表格图像中准确地识别出包含文本信息的区域，以便后续的文本识别和表格结构分析等处理步骤；图像型成分表格是本方法的识别对象，表格筛选目的是从材料文献的表格图片中筛选出成分表格图片；成分表格可以分为单种材料表格和多种材料表格，成分表格分类目的是将两种表格进行区分，不同类型采用不同的提取规则；表格结构解析能够将表格拆解为标题、表头和表体，有助于分离出表格中不同类型的信息，提高识别效率和准确率；成分信息提取是从成分表格的不同区域中提取出材料名称、元素、元素含量和成分单位信息。

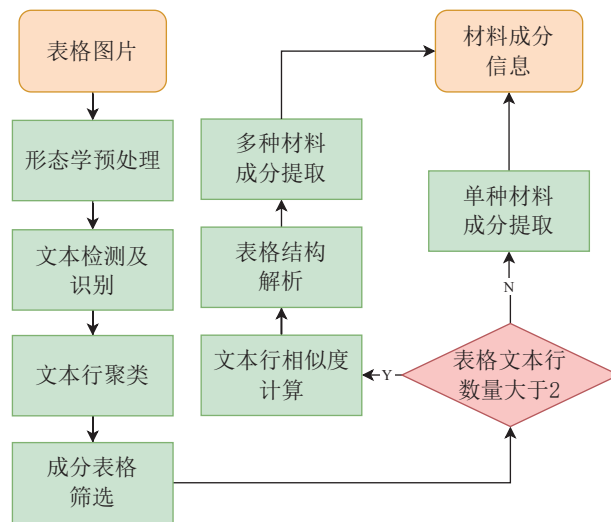


图 3.6 材料成分表格提取整体流程图

3.2.2 表格文本区域识别

电子文档中表格结构形态多样，表格线作为表结构的组成部分，虽然可以作为拆解表格结构、识别文本区域的重要依据，但是在科学文献中，学术性风格的表格通常没有完整的框线结构，因此表格线对文献表格识别意义较小。此外，表格线会在图像中会形成大量的直线和交点，影响表格中的文本检测和识别，加剧表格识别任务的难度。因此对科学文献中表格进行识别时不能依赖于表格线信息，需要对表

格图像进行预处理，以去除表格线，预处理流程和效果如图3.7所示。

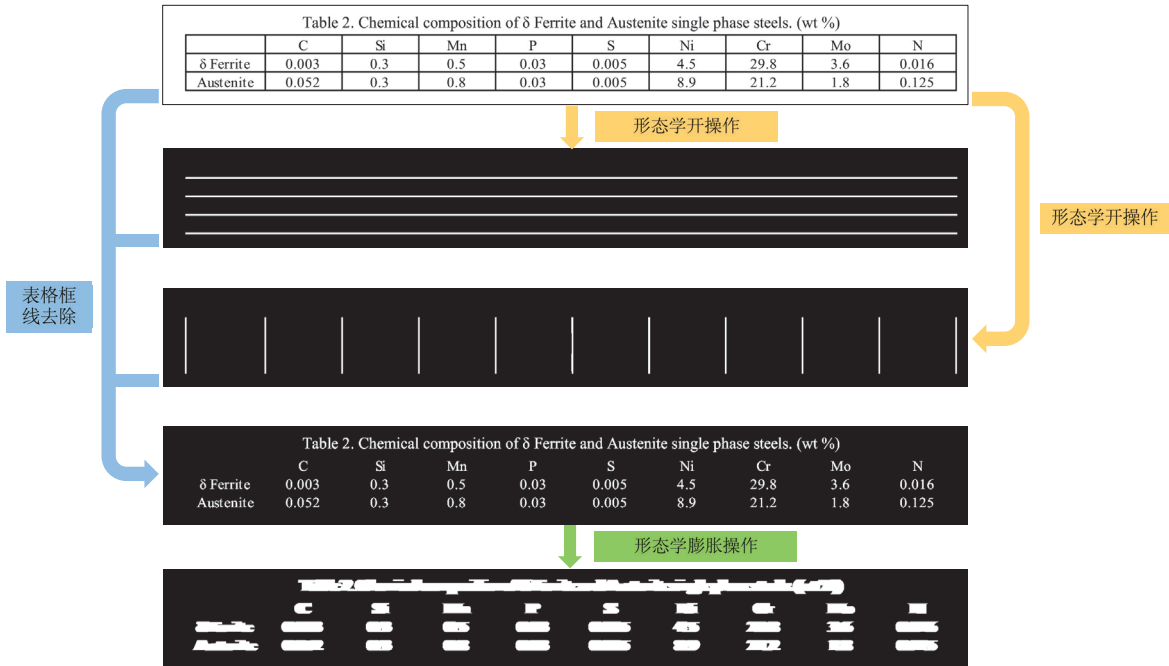


图 3.7 成分表格形态学预处理的过程

成分表格识别方法首先使用 OpenCV 将 RGB 格式的表格图像转化为表格灰度图，去除 RGB 彩色图像中可能存在的干扰，提供更简洁的图像特征信息，灰度化计算公式如3.8所示：

$$Gray(x, y) = 0.299 \times R(x, y) + 0.587 \times G(x, y) + 0.114 \times B(x, y) \quad (3.8)$$

其中， $R(x, y)$ 、 $G(x, y)$ 、 $B(x, y)$ 分别表示位置 (x, y) 处像素的红、绿、蓝三个通道的值， $Gray(x, y)$ 表示转换后位置 (x, y) 的像素灰度值。为了进一步凸显表格图像中的线条轮廓、文字边缘等特征信息，对表格灰度图执行二值化操作得到二值化表格图。二值化是将灰度图像中的像素灰度值转换为只有两个值（通常是黑和白），基本原理是根据设定的阈值 ($thresh$) 对灰度图像进行分割，将灰度值小于阈值的像素设为 0（黑色），将灰度值大于等于阈值的像素设为 255（白色），其数学表达式为：

$$Binary(x, y) = \begin{cases} 255, & \text{if } Gray(x, y) \geq thresh \\ 0, & \text{if } Gray(x, y) < thresh \end{cases} \quad (3.9)$$

其中， $Gray(x, y)$ 表示灰度图像在位置 (x, y) 处的像素灰度值， $thresh$ 表示设定的阈值， $Binary(x, y)$ 表示二值化后在位置 (x, y) 的像素值，本方法中 $thresh$ 设定为 200。

接着，使用形态学方法检测并去除表格线。对于二值化表格图中的水平表格线，使用一个高度较小、宽度较大的矩形结构元素对二值化表格图像进行开操作，其本质为先进进行腐蚀操作，将垂直线前景像素区域完全融入背景像素区域，水平线的边缘像素也会被融入背景像素区域，导致水平线被细化；再进行膨胀操作，增强水平线特征，将水平线恢复成近似原来的宽度，由此可以得到二值化的水平线图。对于二值化表格图中的垂直表格线，使用一个高度较大、宽度较小的矩形结构元素对二值化表格图像进行开操作，其本质也是先进进行腐蚀操作，将垂直线前景像素区域完全融入背景像素区域，垂直线的边缘像素也会被融入背景像素区域，导致垂直线被细化；再进行膨胀操作，增强垂直线特征，将垂直线恢复成近似原来的宽度，由此可以得到二值化的垂直线图。将水平和垂直线图相加得到二值化表格线图，前景为白色线条，背景为黑色像素，得到最终的表格线检测结果。对二值化表格线图进行取反操作，使得前景为黑色线条，背景为白色像素，其数学表达式为：

$$Binary_{inv}(x, y) = 255 - Binary(x, y) \quad (3.10)$$

其中 $Binary_{inv}(x, y)$ 表示取反后在位置 (x, y) 的像素值。将取反后的二值化表格线图（黑色线条、白色文字、白色背景）与二值化表格图（白色线条、白色文字、黑色背景）进行逐个像素与操作，白色与白色像素与操作结果为白色像素，白色与黑色像素与操作结果为黑色像素，由此可以得到无线二值化表格图（黑色线条、白色文字、黑色背景），并对此图进行形态学膨胀操作，将单个字符连通成字符块，增强文本区域的特征，完成形态学预处理操作。

准确地从表格图片中识别出文本内容及坐标信息是表格识别的关键，本方法的处理过程如图3.8所示。针对预处理得到的文本区域特征增强图采用基于二值图像的轮廓检测方法，从二值图像中找到文本连通区域，并将它们表示为一系列的点集合，对区域最外层轮廓进行检测，并采用只保留轮廓端点的近似方法，从而检测出文本区域，得到文本区域轮廓图。根据文本区域轮廓在图中的坐标位置，可以从表格原图中截取出文本区域，并使用开源光学字符识别（OCR）模型 [92] 进行文本识别，最终可以得到文本内容及区域坐标。根据文本区域坐标信息垂直方向上的数值，对文本区域行分类，根据水平方向的数值，对文本区域进行列分类，把文本区域文字识别结果记为 $T_{i,j}$ ， i 为文本区域的行号和 j 为文本区域在本行中的序号，由第 i 行文

本区域组成的文本行内容记为 TL_i ,

$$TL_i = [T_{i,0}, T_{i,1}, \dots, T_{i,n_i}] \quad (3.11)$$

其中, n_i 表示第 i 行中文本区域的数量, 坐标信息记为 $L_{i,j} = (X_l, Y_t, X_r, Y_b)$, 其中 (X_l, Y_t) 为文本区域左上角坐标, (X_r, Y_b) 为文本区域右下角坐标。

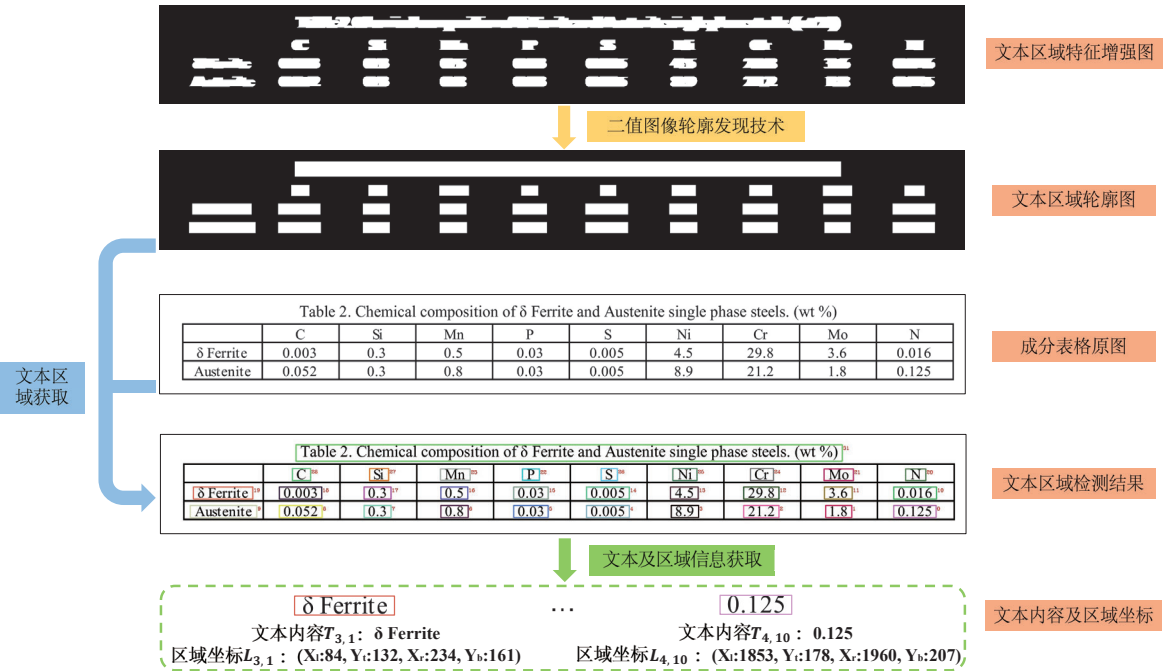


图 3.8 成分表格文本内容及区域坐标获取

3.2.3 材料成分信息提取

经过表格文本区域识别, 表格中文本区域的文本信息和坐标信息已经被获取, 识别方法对表格结构进一步解析, 以获取材料成分信息。提取成分表格信息, 一方面需要从科学文献表格中筛选出成分表格, 另一方面需要将表格拆解为标题、表头和表体三个组成部分, 因为每个部分都有不同的结构和语义信息, 这些信息对于成分数据提取至关重要。科学文献中表格标题通常在表格框线之外, 因此可以依据表格顶部横线划分出表格标题区域和表格内容区域 (包括表头和表体), 即横线上方为表格标题, 下方为表格内容。对表格标题进行正则匹配, $pattern = "\b[C|c]ompositions?\b"$, 判断该表格是否为成分表格, 本方法只对成分表格进行信息提取。科学文献中常见的材料成分表格结构如图3.9所示。

Table 1 Chemical compositions of 3207 duplex stainless steel (wt%).									
Cr	Ni	Mo	Mn	Si	C	S	P	N	
31.05	6.64	3.88	1.03	0.15	0.029	0.005	0.014	0.53	

Table 2 Chemical compositions of 409 FSS and 309L (wt%).									
	C	Si	Mn	P	S	Cr	Ni	Mo	
409FSS	0.03	0.57	0.77	0.03	0.01	11.1	0.25	0.01	
309L	0.02	0.42	1.88	0.02	0.001	23.03	13.74	0.08	

材料名称	单位	元素	元素含量
------	----	----	------

图 3.9 材料成分表格结构示意图

针对成分表格的表头和表体区域，如果文本行总数等于 2 时，则第一行为表头，第二行为表体，该表格为单种材料成分表格。表头和表体部分文本行的数量大于 2 时，使用文本行间的余弦相似度来区分表头和表体。具体是利用材料领域 Fasttext[66] 语言模型，将表头和表体部分的文本区域内容 $T_{i,j}$ 表征为 100 维文本向量 $V_{i,j}$ ，进而将文本行内容 TL_i 转化为文本行向量 VL_i ：

$$VL_i = V_{i,0} + V_{i,1} + \dots + V_{i,n_i} \tag{3.12}$$

其中， n_i 表示第 i 行中文本区域的数量。将当前文本行向量 VL_i 和下一行向量 VL_{i+1} 的余弦相似度记为 S_i ，计算公式如下：

$$S_i = \frac{VL_i \cdot VL_{i+1}}{\|VL_i\| \|VL_{i+1}\|} \tag{3.13}$$

依次计算表头和表体区域文本行向量与下一行向量之间的余弦相似度，文本内容相似度越高，余弦相似度得分越高。在表体中通常展示的是同一类数据，因此表体文本行之间存在极高的相似性，实验中表体文本行之间的相似度均大于 0.6。而表头末行和表体首行相似度低，多行表头内的文本行之间相似度也较低，若有两个及以上 $S_i < 0.6$ ，则表明表头区域有多行文本，不对该类型表格进行提取。若计算发现仅有一个 $S_i < 0.6$ ，可以推断出表头区域有单行文本，表体区域包含多行文本，该表格为多种材料成分表格。

针对单种材料成分表格，表格标题中含有材料名称和含量单位，表头中包括组成元素，表体中为元素含量；针对多种材料成分表格，表格标题中包含单位，表头中包括元素，表体部分含有材料名称和元素含量。将单种材料表格标题输入到上一节

的材料文本命名实体识别 SFBC 模型和单位提取正则表达式中, 分别提取出材料名称和单位; 将多种材料表格标题输入到单位提取正则表达式中, 可以提取出单位, 再从表体部分每行首个文本区域获得材料名称。对于从单种和多种材料成分表格中提取出元素及其含量, 成分表格识别方法依次取出表头中文本区域文字信息 $T_{i,j}$ 和坐标信息 $L_{i,j}$, 根据坐标信息 $L_{i,j}$ 中 X_l 和 X_r 的数值跨度, 寻找位于序号 (i, j) 文本区域下方序号为 $(i+t, j)$ 的文本区域, 得到的 $T_{i,j}$ 为元素文本, $T_{i+t,j}$ 为元素对应的含量数值, 其中 $1 \leq t \leq n$, n 为表体文本行数量, 即表格中包含材料成分信息的数量。将表头中所有文本区域和表体中每一本文行处理完毕, 即可得到由材料名称、元素、元素含量和单位组成的材料成分信息。

3.2.4 实验及方法测试

为了验证基于传统图像处理的材料成分表格识别方法的准确率和效率, 本文分别做了两组实验和一组测试, 分别为表格检测实验、表格文本识别实验和成分表格提取测试。依靠表格区域检测实验, 验证 YOLOv3 算法从文献版面图像中检测表格的精确性; 使用表格文本识别实验, 验证成分表格识别方法获取文本区域和字符 OCR 的准确性; 通过成分表格提取测试, 检验成分表格识别方法获取成分信息的准确率和时间效率。

(1) 表格检测实验

以“Metal Material”为检索关键词, 从 Elsevier ScienceDirect 全文数据库中获取 100 篇文献用于表格区域检测实验。成分表格识别方法中基于 YOLOv3 的表格检测模型是在 800 张文献版面图像数据集中训练, 该数据集使用 labeling 工具对表格区域进行标注。将训练得到的 YOLOv3 模型、BCL 工具 [93] 和 TableSeer[94] 在 100 篇科学文献上进行表格检测实验。人工对 100 篇科学文献中的表格数量进行统计, 总计 265 个, 本文的 YOLOv3 模型提取到 262 个表格, 其中表格区域完整且正确的 244 个; 通过 BCL 转化得到 492 个表格, 远超出人工统计的 265 个, 该方法引入大量错误的表格; 使用 TableSeer 方法定位得到 219 表格, 其中正确的数量为 194 个。

对 YOLOv3 模型未检测到的 3 个表格进行分析, 其中一个表格横向占满整个页面, 一个表格纵向占满整个页面, 一个表格单元格内展示不是文本而是图片, 以上三种类型并未包含在训练数据中, 导致模型无法检测这几类表格区域。对表格区域不

完整或边界溢出的 18 个表格进行分析，区域不完整的主要表现在多行表格标题情况下截取不完整，主要由于数据集中这类数据太少而导致检测结果不佳；边界溢出是因为表格区域和文献文本区域之间间距太小，将部分文本识别成表格的一部分。总体分析，使用 YOLOv3 对文献版面图像中表格进行检测，能够达到较好的效果。

(2) 表格文本识别准确率实验

以单元格（表格标题也视为一个单元格）为基本单位，对检测得到的 244 个图像型表格进行人工数据标注，表格标注结果记为集合 $\mathbf{A}_i = \{t_j\}_{j=1}^{n_i}$ ，其中 i 代表表格序号， $i = 1, 2, \dots, 244$ ， t_j 表示标注数据中单元格文本内容， n_i 代表标注数据中第 i 个表格单元格数量。使用成分表格识别方法对 244 个图像型表格进行表格文本区域识别，识别结果记为集合 $\mathbf{R}_i = \{c_j\}_{j=1}^{m_i}$ ， c_j 表示识别结果中单元格文本内容， m_i 代表识别结果中第 i 个表格单元格数量，按照公式 3.14 计算单张表格识别结果得分。

$$Score_i = |\mathbf{A}_i \cap \mathbf{R}_i| \quad (3.14)$$

对所有表格得分累加求和并除以单元格总数，即可得到表格文本识别准确率得分，如公式 3.15 所示。经过实验，本方法文本识别准确率为 89.04%，错误的识别结果主要由带上下标的文字内容所导致。

$$Acc = \frac{\sum_{i=1}^{244} Score_i}{\sum_{i=1}^{244} |A_i|} \quad (3.15)$$

(3) 成分表格提取测试

从 244 个表格图像中筛选得到 107 个材料成分表格，单种材料成分表格和多种材料成分表格数量分别为 58 和 49 个，使用人工的方式对成分表格进行标注，标注内容包括材料名称、元素及对应含量、元素单位，总计得到 205 条成分信息。使用本文提出的成分表格识别方法对 107 个成分表格进行识别，并与标注数据对比，从 58 个单种材料表格中成功提取得到 54 条正确的成分信息，从 49 个多种材料成分表格中正确提取到 121 条成分信息。因此，单种材料提取准确率为 93.10%，多种材料表格提取准确率为 82.31%，综合准确率为 85.37%。对错误结果或无结果的情况进行分析，发现错误的结果主要是文字识别存在问题，无结果的情况主要出现在多行表头的成分表格中。此外，经过测试每张成分表格的平均提取耗时为 4.59 秒，效率较高。

3.2.5 小结

在成分表格识别方法小节中，基于科学文献中学术表格的特点和成分表格的结构，提出一种基于传统图像处理的材料成分表格识别方法。在使用目标检测算法从文献版面图像中获取表格区域图像后，该方法分为五个步骤对成分表格进行识别：第一步，对表格图像预处理，去除表格线，并增强文本区域的特征；第二步，利用二值图像轮廓发现方法获取文本区域，使用 OCR 模型对文本区域的文字内容进行识别；第三步，根据表格标题从文献表格中筛选出成分表格，依据表头和表体区域文本行数量将成分表格分为单材料和多材料两类；第四步，计算表头和表体区域相邻文本行的余弦相似度，对多材料成分表格的结构进行解析；第五步，基于成分表格的结构特征，从标题、表头和表体中提取出材料名称、元素、元素含量和单位。经过实验发现，成分表格识别方法的提取准确率为 85.37%，达到较好的效果。

3.3 本章小结

在本章中，针对材料文献上下文中的文本和非文本内容，本文提出一种基于上下文感知的材料文献信息提取方法，分别对文献文本和成分表格进行挖掘，为材料研究提供数据支持。面向材料文本的命名实体识别模型 SFBC 利用不同类别语言模型能够提供不同特征的特点，将通用领域动态词向量与材料领域静态词向量相融合，使得每个词向量中都包含上下文信息和材料领域知识，将融合词向量输入 BiLSTM-CRF 网络中，能够准确地从材料文本中提取出材料实体。在 SLSNerData 和 InorgNerData 数据集上进行实验，结果表明词向量融合策略能够显著提高 NER 模型的性能。SFBC 模型的融合词向量与材料领域最新的预训练语言模型 MatSciBERT 的词向量相比，无需在大规模材料语料库上训练语言模型，就能在材料 NER 任务中达到与 MatSciBERT 近似的效果，极大地节约时间和计算资源。

基于传统图像处理的材料成分表格识别方法对材料文献中的表格进行筛选，并对成分表格进行识别，从中提取得到材料名称、元素、元素含量和单位信息。该方法根据表格线的形态特征，利用不同形状的矩形结构元素执行膨胀腐蚀操作，将表格线去除；使用二值图像轮廓检测方法获取文本区域的坐标信息，并利用 OCR 模型识别文本区域的文字内容；随后基于规则将成分表格结构拆解为标题、表头和表体，最终从不同区域中提取出材料名称、元素、元素含量和成分单位信息。相较于深度

学习表格识别方法，本方法通过传统方法就能准确地提取出材料成分信息，无需消耗时间和计算资源训练模型。从材料文献上下文中提取得到的文本信息和成分信息将会在材料性能研究的工作中得到应用。

但是，本章提出的方法也存在不足。第一，在文本挖掘方法中，动态语言模型在词嵌入时融合上下文信息，能够解决一词多义问题，而静态语言模型使用固定内容的向量进行词嵌入，无法对一词多义进行表征。因此，在动态词向量中融入材料领域静态词向量，虽然能够补充材料领域知识特征，但是也削弱了动态语言模型对一词多义的表征能力。第二，在基于传统图像处理的表格识别方法中，只能对材料成分表格进行结构化提取，并且传统方法依赖于手工设计的特征和表格结构规则定义，难以应对复杂的表格结构。本工作会在后续研究中探索并解决上述问题。

第四章 基于文献信息提取的材料性能预测

上一章中提出的基于上下文感知的材料文献提取方法对文献上下文中的文本和表格信息进行挖掘，可以得到材料名称、成分、性能等数据。本章中对文献提取结果进行进一步材料学研究，提出一种基于文献信息提取的材料性能预测方法，方法整体架构如图4.1所示。

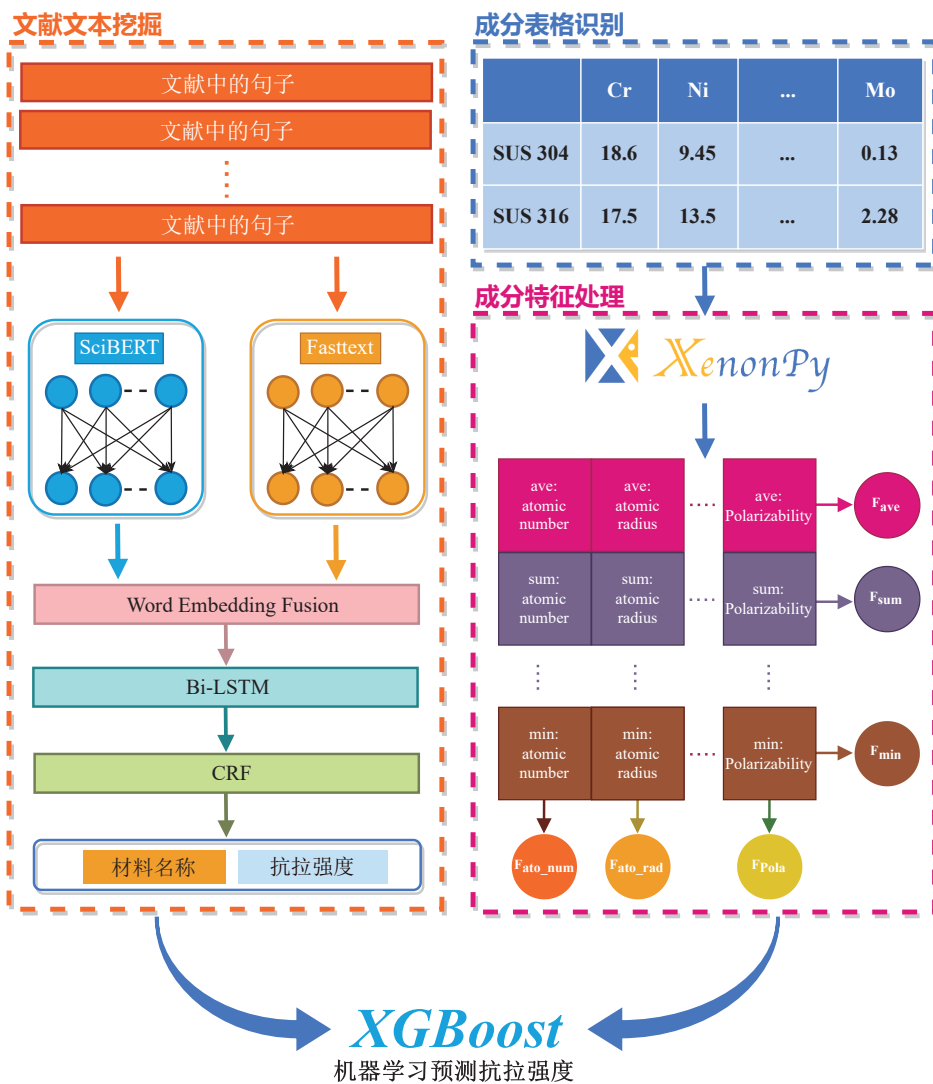


图 4.1 基于文献信息提取的材料性能预测方法整体架构

材料性能预测是通过计算机模拟和预测方法，快速准确地预测材料的性能和行为，可以为材料设计、制造和应用提供重要的指导和帮助。机器学习作为一种新的

材料研究手段，在数据驱动的基础上，为材料科学领域带来了新的机遇。机器学习方法使用已知数据训练模型，从而使模型具有预测材料性能的能力。训练数据的多样性和广泛性对于机器学习预测材料性能至关重要，而材料文献中包含大量与性能相关的实验或结论，因此可以作为性能研究重要的数据源。此外，特征选择在机器学习中具有重要的作用，可以帮助消除无关特征，构建出强相关特征集，是解决高维数据问题、提高模型效果的有效方法之一，对于实现精确预测和科学发现具有重要的意义。

4.1 性能预测方法

4.1.1 性能预测方法框架

对于从文献上下文中挖掘得到的材料成分及抗拉强度数据，本文提出一种材料成分特征扩充及压缩选择的方法，利用筛选后的成分特征与抗拉强度数据共同训练性能预测模型。使用材料信息学库扩充材料成分的特征，根据特征扩充的计算方式，将扩充后的高维特征重塑成矩阵结构，分别对特征矩阵中所有行和列进行特征压缩，并对压缩结果进行重要性评价，从中筛选出重要特征所在的行和列，最终组合为特征筛选结果，用于抗拉强度预测。性能预测方法具体处理流程包括：(1) 使用 XenonPy 材料信息学库 [95] 将文献表格中的材料成分数据扩充为 406 维成分特征；(2) 使用提出的交叉交叉特征压缩及选择方法从扩充后的 406 维材料成分特征中选择重要特征用于抗拉强度预测；(3) 借助机器学习 XGBoost 算法 [96]，使用筛选得到的重要成分特征以及从文献文本中挖掘得到的抗拉强度数据，训练抗拉强度预测模型。训练得到的预测模型，能够依据材料成分预测抗拉强度值。抗拉强度性能预测方法的过程如图4.2所示。

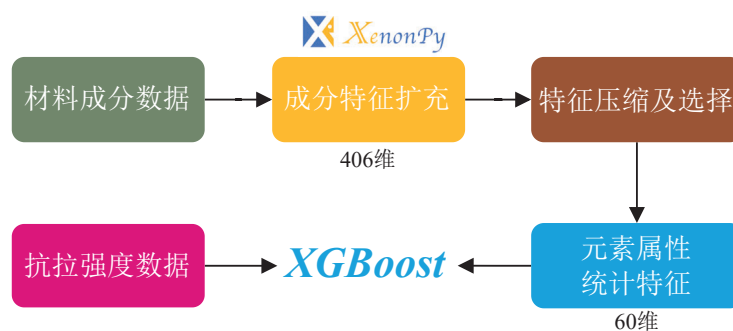


图 4.2 基于文献信息提取的材料性能预测方法流程

4.1.2 成分特征扩充方法

XenonPy 是一个应用于材料信息学领域的机器学习工具，并封装成 Python 库以供使用，旨在提供高效的数据处理和建模工具，以加速材料设计和发现的过程。本文利用该库对文献挖掘得到的材料成分进行特征扩充。现版本 XenonPy 材料信息学库包括材料描述符库、预训练模型库 XenonPy.MDL（支持迁移学习）和相应的机器学习工具，这些算法可以用于材料性质预测、材料分类和聚类任务。此外，XenonPy 还提供了模型解释和可视化的功能，以帮助用户理解模型的预测结果和特征重要性。XenonPy 还提供与公共材料数据库的接口，包括 Materials Project、OQMD、ICSD、CIF 等数据库，用户可以将数据库导入到 XenonPy，以便进行材料分析和建模。

XenonPy 材料描述符库中提供元素级属性数据，具体为元素周期表中从氢 (H) 到钷 (Pu) 这 94 种元素的 58 种元素级属性，如原子核中质子数 (atomic_number)、原子半径 (atomic_radius)、晶面上占有率的赫芬达尔-希尔斯曼指数 (hhi_p)、密度泛函理论带隙能量 (gs_bandgap)、范德华力系数 (c6_gb)、瞬时偶极子形成能力 (Polarizability) 等。此外，XenonPy 材料描述符库还提供 7 种统计特征计算器：加权均值、加权和、加权方差、几何平均数、调和平均数、最大值和最小值计算器，用于对 58 种元素级属性进行统计学特征运算。使用统计特征计算器对元素级属性进行特征运算可以对材料成分数据进行特征扩充。以一种由两种元素组成的二元化合物 $A_{w_A}B_{w_B}$ 为例，设其元素级属性分别为 $f_{A,i}$ 和 $f_{B,i}$ ($i = 1, 2, \dots, 58$)，7 种统计特征数学计算公式见表 4.1。其中 w_A^* 和 w_B^* 表示归一化处理后的对应元素含量，即 $w_A^* + w_B^* = 1$ 。使用 7 种统计特征计算器对 58 种元素级属性进行特征运算，即可将材料成分扩充为 406 (7*58) 维元素属性统计特征。

表 4.1 XenonPy 库中 7 种统计特征计算器及数学计算公式

序号	统计计算器	数学计算公式
1	加权均值 (ave)	$f_{ave,i} = w_A^* f_{A,i} + w_B^* f_{B,i}$
2	加权和 (sum)	$f_{sum,i} = w_A f_{A,i} + w_B f_{B,i}$
3	加权方差 (var)	$f_{var,i} = w_A^* (f_{A,i} - f_{ave,i})^2 + w_B^* (f_{B,i} - f_{ave,i})^2$
4	几何平均数 (gmean)	$f_{gmean,i} = \sqrt[w_A + w_B]{f_{A,i}^{w_A} * f_{B,i}^{w_B}}$
5	调和平均数 (hmean)	$f_{hmean,i} = \frac{w_A + w_B}{\frac{1}{f_{A,i}^{w_A}} + \frac{1}{f_{B,i}^{w_B}}}$
6	最大值 (max)	$f_{max,i} = \max(f_{A,i}, f_{B,i})$
7	最小值 (min)	$f_{min,i} = \min(f_{A,i}, f_{B,i})$

以 $\text{Cr}_{14.4}\text{Ni}_{3.2}\text{Mo}_{0.08}\text{N}_{0.1}\text{Mn}_{0.5}\text{Si}_{0.7}\text{Cu}_{3.4}\text{C}_{0.18}$ 为例，该种材料是由 8 种元素组成的八元化合物，对于原子核中质子数 (atomic_number) 这种元素级属性而言，根据表 4.1 中二元化合物的 7 种统计特征计算方式，可以计算得到该八元化合物原子核中质子数这种元素级属性的 7 个统计特征值，如表 4.2 所示。

表 4.2 样例材料原子核中质子数属性的统计特征计算结果

序号	元素级属性统计特征名称	元素级属性统计特征值
1	原子核中质子数加权均值	24.964084
2	原子核中质子数加权和	564.388
3	原子核中质子数加权方差	11.395915
4	原子核中质子数几何平均数	24.665568
5	原子核中质子数调和平均数	24.214897
6	原子核中质子数最大值	42.0
7	原子核中质子数最小值	6.0

4.1.3 十字交叉特征压缩及选择方法

使用 XenonPy 材料信息学库对材料成分数据进行特征转换，将成分特征扩充至 406 维，特征维度的上升为数据分析和模型学习提供依据，但同时也引入大量低重要性或冗余特征，影响机器学习模型对数据的处理和学习能力。特征选择在机器学习中至关重要 [38]，其能够从高维特征中消除不相关的特征，选择出与结果密切相关的特征，进而降低数据维度，以便提高机器学习任务的性能和准确性 [97]。根据 XenonPy 材料信息学库成分特征扩充的计算方式，本文提出一种十字交叉特征压缩及选择方法，分别对 7 种统计特征计算器的输出结果进行水平压缩，如图 4.3(a) 所示；对 58 种元素属性特征（如原子核中质子数 atomic_number、原子半径 atomic_radius、原子体积 atomic_volume、Rahmy 原子半径 atomic_radius_rahm 等）进行垂直压缩，如图 4.3(b) 所示。根据水平和垂直方向的压缩结果对特征进行筛选和组合。将 7 种统计特征计算器对 58 种元素级属性的计算结果分别记为 $f_{ave,i}$ 、 $f_{sum,i}$ 、 $f_{var,i}$ 、 $f_{gmean,i}$ 、 $f_{hmean,i}$ 、 $f_{max,i}$ 和 $f_{min,i}$ ($i = 1, 2, \dots, 58$)，以加权均值统计特征计算器 (ave) 为例，将其水平压缩后的特征记为 F_{ave} ，计算公式如下：

$$F_{ave} = \frac{\sum_{i=1}^{58} |f_{ave,i}| * f_{ave,i}}{\sum_{i=1}^{58} f_{ave,i}} \quad (4.1)$$



(a) 特征水平压缩



(b) 特征垂直压缩

图 4.3 406 维元素属性统计特征交叉压缩

如是分别计算其余 6 种统计特征计算器水平压缩后的特征，分别记为 F_{sum} 、 F_{var} 、 F_{gmean} 、 F_{hmean} 、 F_{max} 和 F_{min} ，最终将 406 维特征水平压缩为 7 维统计计算器特征。

以元素级属性特征原子核中质子数 (atomic_number) 为例，其在 58 种元素级属性中序号为 1，将其垂直压缩后的特征记为 F_{ato_mum} ，计算公式如下：

$$F_{ato_mum} = \frac{\sum_{cal \in \{ave, sum, var, gmena, hmean, max, min\}} |f_{cal,1}| * f_{cal,1}}{\sum_{cal \in \{ave, sum, var, gmena, hmean, max, min\}} f_{cal,1}} \quad (4.2)$$

如是分别计算其余 57 种元素级属性垂直压缩后的特征，分别记为 F_{ato_rad} 、 F_{ato_vol} 、 F_{ato_radh} 等，最终将 406 维特征垂直压缩为 58 维元素属性特征。

本文使用 lightGBM[98] 分别对水平和垂直压缩得到的 7 维统计计算器特征和 58 维元素属性特征进行特征重要性评估，其基本原理基于特征分裂的次数 (Split)。LightGBM 是一种基于决策树的梯度提升框架，它采用了 Leaf-wise 的增长策略，该策略每次从当前所有叶子中，找到分裂增益最大的一个叶子进行分裂，如此循环。基于分裂次数的特征重要性评估方法是计算每个特征在所有树中被用作分裂点的次数，并将次数作为特征重要性的度量，次数越多，说明该特征越重要，本质上可以反映出特征在模型中出现的频率。LightGBM 计算基于分裂次数的特征重要性的数学公式如下：

$$FI_j = \sum_{t=1}^T I(split_t = j) \quad (4.3)$$

其中， FI_j 表示特征 j 的重要性， $I(split_t = j)$ 为特征 j 在第 t 次分裂中是否被选为最佳分裂特征的指示函数， T 为总的分裂次数。特征重要性评估结果可以识别出对模型预测结果贡献较大的特征，从而优化特征选择的结果，构建相关度更高的特征集，提高模型性能。使用 LightGBM 基于分裂次数的特征重要性评估方法对 7 维统计计算器特征进行重要性评分，从中选出 5 维重要特征，结果见表 4.3；对 58 维元素属性特征进行重要性评分，从中选出 12 维重要特征，结果见表 4.4。使用筛选得到的 5 种特征统计计算器对 12 种元素级属性进行计算，得到 60 维元素属性统计特征用于抗拉强度预测。

在化学信息学中，不同化合物的性能和性质通常受分子结构、原子体积、共价半径等多种因素的影响。在对材料进行性能预测时，需要统筹考虑这些因素之间的相互作用和权重，加权方差 (var) 的特点是可以考虑每个因素的权重和变异程度，因

表 4.3 406 维元素属性统计特征水平压缩结果重要性评分

序号	统计方式特征	得分	序号	统计方式特征	得分
1	加权方差 (var)	311	4	几何平均数 (gmean)	199
2	加权和 (sum)	255	5	加权均值 (ave)	191
3	调和平均数 (hmean)	216			

表 4.4 406 维元素属性统计特征垂直压缩结果重要性评分

序号	元素级属性特征	得分	序号	元素级属性特征	得分
1	hhi_p	113	7	num_s_unfilled	60
2	gs_bandgap	93	8	electron_affinity	59
3	sound_velocity	85	9	num_p_valence	58
4	gs_mag_moment	73	10	c6_gb	48
5	hhi_r	66	11	boiling_point	46
6	num_p_unfilled	60	12	melting_point	39

此可以更好地描述化合物之间的差异和相似性；加权和 (sum) 的特点是可以更好地描述数据的权重分布情况，并将数据值相加，但不能反映数据之间的比例关系；调和平均数 (hmean) 的特点是对异常值具有较强的鲁棒性，可以更好地处理原始数据中存在的异常值；几何平均数 (gmean) 的特点是可以很好地描述原始数据中的比例关系，特别是在数据中存在多个维度时，对于不同维度之间的比例关系具有很好的鲁棒性；加权均值 (ave) 的特点是可以考虑每个数据点的权重，不仅反映了原始数据中每个数据点的贡献大小，还考虑了每个数据点的个数或者样本量。因此使用上述特征计算器扩充成分特征，能够更全面地反映原始材料成分数据的信息，提高模型的预测准确度。

材料元素级属性特征中，hhi_p 特征表示材料的赫芬达尔—赫希曼指数 (HHI)，即组成元素的摩尔分数的平方和，这个指标可以反映材料的组成复杂度。hhi_p 特征对抗拉强度有很大影响。抗拉强度衡量材料在受到拉伸力时能够承受的最大应力，会受到材料的微观结构和缺陷的影响，而这些因素又与材料中元素的分布和相互作用有关 [99]。一般来说，如果材料中元素的多样性和均匀性较高，那么材料的微观结构和缺陷就会较少，从而提高抗拉强度。gs_bandgap 特征表示材料的密度泛函理论 (DFT) 带隙能，即在零温度下基态的电子能带之间的能量差，特征值反映材料的导电性，值越小表示材料导电性越佳。DFT 带隙能对抗拉强度的影响与材料的结构、相

变等因素有关，有些研究表明，使用非融合结构的宽带隙受体可以提高材料的抗拉强度和稳定性 [100]。sound_velocity 特征表示声速，具体指的是声波在介质中的传播速度，它与介质的密度和弹性有关。sound_velocity 特征对抗拉强度有很大影响，因为声速反映了材料的刚度和韧性，这些性质与材料的抗拉能力密切相关 [101]，声速越高，表明材料刚性越好，抗拉强度越高；声速越低，表明材料越偏柔性，抗拉强度越低。因此元素级属性可以反映材料的基本组成和特性，构建出高维的特征空间，能够提高机器学习模型的预测能力。

4.1.4 XGBoost 算法预测抗拉强度

上一章中基于上下文感知的文献提取方法能够从材料科学文献中提取出材料成分及抗拉强度数据，材料成分包括铬 (Cr)、镍 (Ni)、钼 (Mo)、氮 (N)、锰 (Mn)、硅 (Si)、铜 (Cu) 和碳 (C) 这八种元素，抗拉强度为材料在拉伸载荷作用下所能承受的最大拉伸应力。使用 XenonPy 材料信息学库将材料成分数据扩充至 406 维元素属性统计特征，再将 406 维特征进行交叉交叉特征压缩及筛选获得 60 维与抗拉强度密切相关的特征用于抗拉强度预测。使用机器学习 XGBoost (Extreme Gradient Boosting) 算法 [96]，利用经过特征扩充、交叉压缩及特征选择得到的 60 维特征和抗拉强度数据训练得到预测模型。

XGBoost 是一种基于梯度提升决策树的机器学习算法，它通过对多个弱学习器 (决策树) 组合来构建一个强学习器。在每次迭代时，使用梯度下降来拟合一个新的弱学习器，并对前面模型的误差进行加权，使得模型能够集中于更难预测的样本。在回归预测中，XGBoost 使用平方误差损失函数来衡量预测值与真实值之间的误差。在每个决策树上，XGBoost 通过最小化平方误差损失函数来确定最佳分裂点，并使用梯度提升算法来逐步优化每个决策树的预测效果。具体来说，XGBoost 会计算每个样本的损失函数梯度和二阶导数，然后使用这些信息来训练每个决策树，最终将多个决策树的预测结果进行加权平均得到最终的预测结果。与传统的支持向量机、人工神经网络等方法相比较，XGBoost 能获得更高的准确率，并且对奇异值体现出较强的鲁棒性，因此对原始数据的标准化要求低，是目前相关领域中处理回归预测问题的最常用算法之一。

4.2 实验分析

使用从材料文献中提取得到的材料成分和抗拉强度训练预测模型，元素包括 Cr、Ni、Mo、N、Mn、Si、Cu 和 C，元素含量单位为质量百分含量 (mass%)，抗拉强度单位为兆帕 (MPa)。利用机器学习回归任务相关算法，使用挖掘得到的数据训练抗拉强度预测模型，依据材料成分对抗拉强度进行预测，并使用 NIMS 材料数据库 (MatNavi) [102] 中 Kinzoku 库的认证数据进行验证，证明文献提取方法的可行性、材料成分特征处理方法的有效性和回归预测的准确性。

4.2.1 实验数据准备

(1) 用于训练模型的文献提取数据

在 Elsevier ScienceDirect 全文数据库中，以 “Stainless Steel” 为检索词，“2012-2021” 为时间区间进行检索，从结果中筛选并收集得到 11,058 篇可以开源获取的英文文献，将文献提取方法应用于这些文献，并对提取结果进行筛选，最终从中得到符合要求的 321 条不锈钢材料 “材料成分-抗拉强度” 数据。提取结果数据特征统计信息见表4.5。

表 4.5 从文献中提取到的不锈钢成分-抗拉强度数据特征统计信息

序号	数据特征项	特征描述	最大值	最小值	均值	标准差
1	C	铬含量	8.60	27.57	18.07	2.75
2	Ni	镍含量	0.03	25.24	8.78	4.29
3	Mo	钼含量	0.005	9.00	1.90	1.40
4	N	氮含量	0.0047	0.75	0.13	0.12
5	Mn	锰含量	0.013	18.20	2.22	2.92
6	Si	硅含量	0.02	6.90	0.52	0.46
7	Cu	铜含量	0.0032	4.66	0.74	1.17
8	C	碳含量	0.002	1.30	0.06	0.14
9	Tensile	抗拉强度 (MPa)	413	670	609	34.87

(2) 用于验证模型的权威数据

从 NIMS 材料数据库 (MatNavi) 中 Kinzoku 库获取 69 条不锈钢材料的 “材料成分-抗拉强度” 数据，用这些认证数据对本文方法进行验证。MatNavi 由日本国立材料科学研究所 (NIMS) 组建，拥有 6 个基础性能数据库 (聚合物 PoLyInfo、无机材料 AtomWork、计算相图 CPDDB、计算电子结构 CompES-X、扩散 Kakusan 和热物理特

性)、2个工程数据库(金属材料 Kinzoku 和 CCT 图)、4个结构材料数据库(蠕变 CDS、疲劳 FDS、腐蚀 CoDS 和强度 SDS)和3个数据应用系统(复合材料设计和性能预测 CompoTherm、金属偏析预测 SurfSeg 和金属氧化物预测 InerChemBond)。其中,金属材料数据库(Kinzoku)包括金属材料的密度、弹性模量、泊松比、硬度、韧性、拉伸、蠕变、蠕变断裂特性和各种疲劳特性等数据,可按 JIS 标准号、ASTM 标准号、材料名称、材料形式等进行检索。本文使用的 Kinzoku 库部分不锈钢认证数据的特征统计信息见表4.6。

表 4.6 Kinzoku 库中不锈钢成分-抗拉强度数据特征统计信息

序号	数据特征项	特征描述	最大值	最小值	均值	标准差
1	C	铬含量	16.39	25.25	17.99	1.17
2	Ni	镍含量	8.20	33.92	11.60	3.27
3	Mo	钼含量	0.02	2.56	0.82	1.02
4	N	氮含量	0.0074	0.102	0.03	0.02
5	Mn	锰含量	0.54	1.85	1.43	0.34
6	Si	硅含量	0.28	0.82	0.57	0.12
7	Cu	铜含量	0.02	2.92	0.20	0.35
8	C	碳含量	0.07	0.09	0.06	0.02
9	Tensile	抗拉强度 (MPa)	540	676	596	29.97

4.2.2 实验环境与评价指标

(1) 实验环境

实验所用操作系统均为 Windows 10, 代码采用 Python 3.8 版本进行编码, 使用的机器学习框架为 Scikit-learn 和 xgboost 库, 硬件平台为 Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz, 内存为 32GB, 未使用 GPU。

(2) 评价指标

实验采用均方根误差 ($RMSE$)、平均绝对误差 (MAE) 和决定系数 (R^2) 作为评价指标。 $RMSE$ 是预测值与真实值偏差的平方与观测次数 n 比值的平方根, 用来衡量预测值同真值之间的偏差, 对异常值敏感, 定义如公式4.4所示; MAE 是所有单个预测值与算术平均值的偏差的绝对值的平均, 可以避免误差相互抵消的问题, 因而可以准确反映实际预测误差的大小, 定义如公式4.5所示; R^2 反映因变量的全部变动能通过回归关系被自变量解释的比例, 使用均值作为误差基准, 衡量预测误差

同均值基准误差之间的关系，定义如公式4.6所示。 $RMSE$ 和 MAE 的取值范围与实际回归任务相关，无固定范围，数字越小模型效果越好； R^2 是回归任务的综合性评价指标，具有任务无关性，取值范围通常为 0 到 1 之间，越接近于 1 则可认为回归模型拟合效果越好。当 $R^2 < 0$ 时，模型预测能力差，且最小值没有下限。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |(y_i - \hat{y}_i)| \quad (4.5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y}_i - y_i)^2} \quad (4.6)$$

其中， n 表示样本数量， y_i 表示真实值， \hat{y}_i 表示预测值。

4.2.3 不同特征预测抗拉强度的对比实验

为验证本章提出的基于文献信息提取的材料性能预测方法的有效性，分别设计如下对比实验：

实验 1：使用文献提取方法得到的“材料成分-抗拉强度”数据，其中成分包括 8 种元素，再分别使用随机森林 (RF)、梯度提升树 (GBDT)、极端梯度增强 (XGBoost) 和 K 近邻 (KNN) 算法训练抗拉强度预测模型，并使用 Kinzoku 认证数据进行验证，根据抗拉强度预测模型的预测结果值和 Kinzoku 中真实值计算 3 种评价指标。

实验 2：使用 XenonPy 材料信息学库将原始 8 维成分数据扩充为 406 维元素属性统计特征，在“元素属性统计特征-抗拉强度”数据上，使用实验 1 中相同的 4 种机器学习算法训练抗拉强度预测模型，并使用 Kinzoku 库认证数据进行验证并计算 3 种评价指标。

实验 3：在实验 2 中“元素属性统计特征-抗拉强度”数据的基础上，直接使用 LightGBM 算法对 406 维特征进行重要性评分，筛选出重要特征（分别包括 40 维、50 维、60 维和 70 维特征），得到“LG 元素属性统计特征-抗拉强度”数据。再使用实验 1 中相同的 4 种机器学习算法训练抗拉强度预测模型，并使用 Kinzoku 库认证数据进行验证并计算 3 种评价指标。

实验 4: 在实验 2 中“元素属性统计特征-抗拉强度”数据的基础上, 使用本文提出的交叉交叉特征压缩及选择方法, 对 406 维元素属性统计特征进行特征筛选, 获取重要特征 (分别包括 40 维、50 维、60 维和 70 维特征), 得到“重要元素属性统计特征-抗拉强度”数据。再使用实验 1 中相同的 4 种机器学习算法训练抗拉强度预测模型, 并使用 Kinzoku 库认证数据进行验证并计算 3 种评价指标。

实验中, 使用的 RF、GBDT 和 KNN 算法均源自 Scikit-learn 库, 使用的 XGBoost 算法来自 xgboost 库, 所有算法均使用库中默认参数构建模型。以 $RMSE$ 、 MAE 和 R^2 作为评价指标, 4 种机器学习算法在不同特征上的实验结果如表 4.7-表 4.9 所示。

表 4.7 不同实验中 $RMSE$ 评价指标的对比结果

	维度	RF	GBDT	KNN	XGBoost
实验 1	8 维	20.74	23.01	20.57	18.41
实验 2	406 维	20.15	20.69	20.44	23.75
实验 3	40 维	19.63	19.02	30.79	25.83
	50 维	19.22	20.78	30.79	24.32
	60 维	19.53	17.69	30.79	27.45
	70 维	19.87	19.81	30.79	29.01
实验 4	40 维	20.00	20.12	19.25	18.61
	50 维	19.75	18.38	19.25	19.74
	60 维	19.15	17.16	20.44	16.76
	70 维	19.11	16.76	20.44	25.92

表 4.8 不同实验中 MAE 评价指标的对比结果

	维度	RF	GBDT	KNN	XGBoost
实验 1	8 维	15.97	17.63	17.53	14.34
实验 2	406 维	16.41	16.38	16.81	17.37
实验 3	40 维	16.17	16.33	24.12	19.95
	50 维	15.96	16.12	24.12	19.71
	60 维	16.13	14.79	24.12	21.01
	70 维	16.38	16.11	24.12	22.60
实验 4	40 维	15.98	16.24	16.06	15.54
	50 维	15.51	14.95	16.06	15.94
	60 维	15.27	14.20	16.81	13.90
	70 维	15.19	13.96	16.81	21.50

表 4.9 不同实验中 R^2 评价指标的对比结果

	维度	RF	GBDT	KNN	XGBoost
实验 1	8 维	0.4899	0.3719	0.4988	0.5987
实验 2	406 维	0.5198	0.4927	0.5051	0.3322
实验 3	40 维	0.5432	0.5714	-0.123	0.2098
	50 维	0.5621	0.4889	-0.123	0.2997
	60 维	0.5478	0.6291	-0.123	0.1076
	70 维	0.5319	0.5345	-0.123	0.0034
实验 4	40 维	0.5251	0.5202	0.5611	0.5897
	50 维	0.5375	0.5999	0.5611	0.5384
	60 维	0.5652	0.6512	0.5051	0.6671
	70 维	0.5666	0.6463	0.5052	0.2043

4.2.4 实验结果及分析

4 种机器学习模型在 8 维“材料成分-抗拉强度”数据上,预测效果一般。相较于实验 1,实验 2 中 RF、GBDT、KNN 在 406 维数据上,预测效果有所提升,但 XGBoost 预测效果较差。实验 3 中,使用 LightGBM 对 406 元素属性统计特征进行筛选,在 4 种维度的筛选结果上,RF 和 GBDT 预测效果有明显改善,但 KNN 和 XGBoost 模型的预测误差大于实验 2 中的误差,因此直接使用 LightGBM 对 406 维数据进行筛选效果不佳。在实验 4 中,使用本文提出的交叉交叉特征压缩及选择方法对 406 维特征进行筛选,在 4 种机器学习模型上预测效果整体优于其余实验。所有实验中,RF 最优结果出现在实验 4 的 70 维数据中, $R^2=0.5666$;GBDT 最优结果出现在实验 4 的 60 维数据中, $R^2=0.6512$;KNN 最优结果出现在实验 4 的 50 维数据中, $R^2=0.5611$;XGBoost 最优结果出现在实验 4 的 60 维数据中, $R^2=0.6671$,相较于实验 1 中的 R^2 得分提高 11.42%。XGBoost 最优模型的真实值与预测值散点图如图 4.4 所示。本方法通过特征水平压缩,将每种计算器对 58 种元素级属性的计算结果压缩为统计特征,使得每个特征值中都涵盖所有元素级属性信息,在特征筛选时能够选出更重要的统计特征计算器;通过特征垂直压缩,将每种元素级属性被 7 种特征计算机处理后的结果压缩为元素属性特征,使得每个特征值都考虑所有的统计信息,在特征选择时能够选出影响力更大的元素级属性。因此,使用本文提出的交叉交叉特征压缩对 406 维特征进行处理后,再使用 LightGBM 对水平和垂直压缩结果进行筛选,得到的元素属性统计特征在抗拉强度预测上能够取得更好的效果。

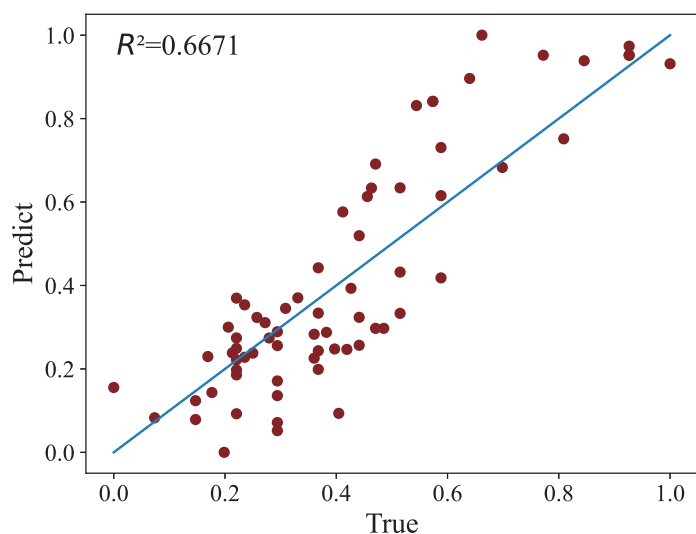


图 4.4 实验 4 中 XGBoost 在 60 维特征上真实值与预测值散点图

4.3 本章小结

本章提出一种基于文献信息提取的材料性能预测方法，旨在协助材料性能改善工作，加快数据驱动的新材料研发步伐。通过上一章节的文献提取方法从科学文献上下文的文本中挖掘得到材料名称和抗拉强度，从图像型表格中提取得到材料成分，由此组合得到材料成分和抗拉强度数据。依靠 XenonPy 材料信息学库将材料成分扩充为 406 维元素属性统计特征，再使用本文提出的交叉交叉特征压缩及选择方法对 406 维特征进行筛选，最终使用机器学习回归算法，在筛选得到的特征和抗拉强度数据上训练得到抗拉强度预测模型。从 11,058 篇不锈钢科学文献中挖掘得到 321 条满足要求的材料成分和抗拉强度数据，在 RF、GBDT、KNN 和 XGBoost 算法上训练模型，并使用 MatNavi 中 Kinzoku 库 69 条不锈钢数据进行测试和评价指标计算。通过对比实验可以发现，本文提出的交叉交叉特征压缩及选择方法对 406 维元素属性统计特征处理效果更好，显著提高模型对抗拉强度预测的准确性。其中，XGBoost 算法在 60 维特征上效果最佳， $R^2=0.6671$ 。

但是，本章提出的方法也存在不足。影响材料性能的因素繁杂，除了材料成分之外，加工工艺、微观结构、晶体结构等都对材料性能有重要影响，这些因素相互作用，共同决定了材料的性能。因此，本章方法存在一定局限性，在未来的研究中，会探究从科学文献中提取更多性能相关数据，并对数据特征进行处理，使得模型能更好地预测各种因素和材料性能之间的内禀关系。此外，使用机器学习 XGBoost 预测抗拉强度其背后的预测原理可解释性较低，需要进一步研究可解释性机器学习。

第五章 基于不锈钢文献提取结果的统计分析和性能预测

本章以不锈钢为示范材料，将第三章中提出的基于上下文感知的材料文献信息提取方法应用在 11,058 篇英文不锈钢科学文献上，将第四章中提出的性能预测方法应用在不锈钢文献提取结果上。首先，使用 SFBC 模型从文献文本中提取得到的 13 类总计 236 万个材料实体，并基于提取结果统计分析 2012-2021 年间不锈钢研究热点的变化和部分实体类别之间潜在关系。其次，利用基于传统图像处理的成分表格识别方法对 11,058 篇科学文献中的成分表格进行识别，提取得到 7970 组材料成分信息。最后，从 236 万个材料实体和 7970 组成分信息中筛选出材料成分和抗拉强度数据，使用交叉特征选择及性能预测方法，对不锈钢材料的抗拉强度进行预测；此外，还筛选出材料成分、工艺、性能和性能变化数据，直接使用机器学习算法对不锈钢材料抗腐蚀性、延展性、强度和硬度进行性能变化趋势预测。

5.1 文献收集与信息提取

本文从 Arxiv 和 Elsevier ScienceDirect 数据库中收集 11,058 篇 PDF 格式的开源英文不锈钢科学文献，为了进行文献提取，需要对 PDF 格式的科学文献进行预处理。预处理步骤包括：(1) 将 PDF 转换为 Word 格式，提取文献的文本内容，并对文本分句以用于文本提取；(2) 使用 Python 中的 fitz 和 PyMuPDF 工具库将 PDF 文献逐页转化为版面图像，用于成分表格识别。预处理后，可以从文献文本中识别 13 类材料实体，从图像型成分表格中提取成分信息。材料实体类别包括材料名称、研究方面、处理工艺、使用技术、材料性能、性质值、实验名称、实验条件、条件值、实验产出、使用设备、涉及元素和适用场景；成分信息包括材料名称、元素、元素含量和单位。

5.2 提取结果统计分析

5.2.1 单实体类别统计分析

在本节中分析了 2012 年至 2021 年间材料名称、处理工艺、使用技术、材料性能、涉及元素和适用场景这六种材料命名实体研究热度的趋势，绘制了不同年份抗

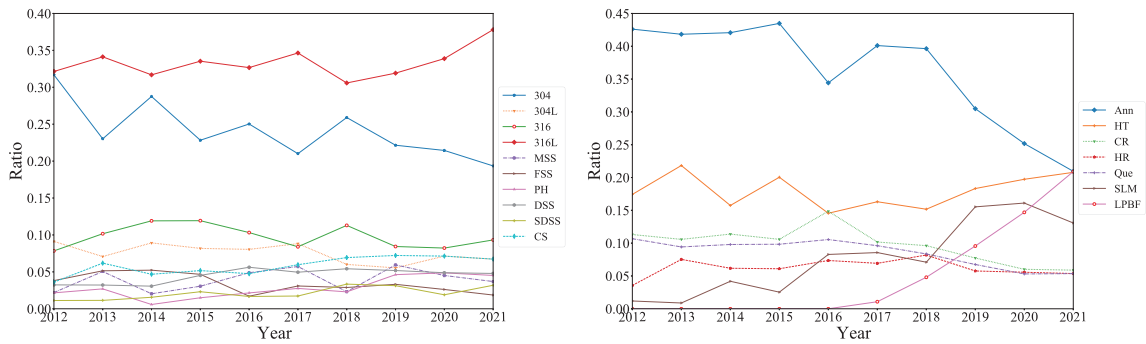
拉强度和屈服强度值的分布情况。具体如下：

材料名称：对奥氏体不锈钢、马氏体不锈钢、铁素体不锈钢、双相不锈钢和沉淀硬化不锈钢出现频率的比例进行统计，不锈钢的具体型号和统计结果在图5.1(a)中。316L 不锈钢是一种低碳不锈钢，近年来受到研究人员的更多关注，其频率比例从 2012 年的 0.33 增加到 2021 年的 0.38。因为添加了钼 (Mo) 元素，使其具有更好的耐腐蚀性和高温性能，特别适用于海水、卤水和高温环境，这与对应适用场景统计结果相一致，适用场景统统计结果显示不锈钢在核电站和航空航天领域中得到了更广泛地使用。然而，304 不锈钢的研究热度逐渐下降，其频率比例从 2012 年的 0.32 降至 2021 年的 0.21。此外，316 不锈钢、碳钢、超级双相不锈钢等材料在不锈钢文献中的受关注度都在增加，统计结果图显示不锈钢领域的研究趋势和热点，可以为相关研究人员提供参考，更好地把握未来研究和发展方向。

处理工艺：收集退火、热处理、选择性激光熔化、激光粉末床融合等 7 种工艺在文献中出现频率的数据，详细信息见图5.1(b)。其中，退火工艺的应用比例持续下降，特别是在 2018 年之后，频率比例从 0.4 连续降至 0.2。退火工艺使不锈钢结构趋向平衡状态，降低了其耐蚀性，同时退火也会使不锈钢的硬度过低，不适合用于高强度和高耐磨场合，因此退火工艺的频率显著下降。激光粉床融合与增材制造技术密切相关，增材制造技术是一种新兴的快速制造技术，受到学术界和工业界广泛关注 [5]，因此越来越多的研究涉及激光粉床融合，其比例在 2021 年达到 0.2。

材料性能：收集硬度、耐蚀性、屈服强度、拉伸强度等 8 种性能的数据，统计结果见图5.2(a)。不锈钢的耐蚀性一直最受关注，但近年来其研究热度却有所下降。相反，越来越多的材料科学文献涉及不锈钢的强度，这种变化的原因在于应用需求的改变。不锈钢的强度与其耐磨性、耐久性和稳定性密切相关，在工业制造（如汽车和飞机制造）中采用轻质、高强度的不锈钢材料，能够提高车身和机身的承载能力和疲劳强度。

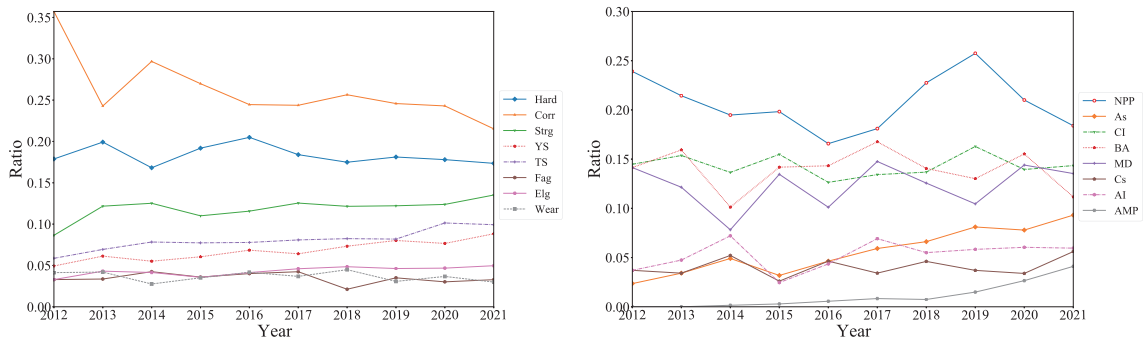
适用场景：对不锈钢的应用场景进行统计分析，主要包括核电站、航空航天、化工行业等领域。从图5.2(b)中可以观察到不同应用领域在材料文献中的出现情况呈现出较大波动，但核电站一直是最重要的应用场景。此外，不锈钢在航空航天和增材制造零件中的应用越来越广泛。统计结果可以指导研究人员确定研究发现，例如开发适用于航空航天应用的轻质高强度不锈钢或使用增材制造方法生产不锈钢零件等。



(a) 材料名称研究热度变化

(b) 处理工艺研究热度变化

图 5.1 材料名称和处理工艺研究热度在不同年份间的变化。(a) 图中 304 是奥氏体 304 不锈钢, 304L 是奥氏体 304L 不锈钢, 316 是奥氏体 316 不锈钢, 316L 是奥氏体 316L 不锈钢, MSS 是马氏体不锈钢 (牌号包括 410&420&440), FSS 是铁氧体不锈钢 (牌号包括 409&430&446), PH 是马氏体沉淀硬化型不锈钢、DSS 是双相不锈钢 (牌号包括 UNS S31500&S31803), SDSS 是超级双相不锈钢 (牌号包括 UNS S32750&S32550), CS 是碳钢。(b) 图中 Ann 是退火, HT 是热处理, CR 是冷轧, HR 是热轧, Que 是淬火, SLM 是选择性激光熔化, LPBF 是激光粉末床融合。



(a) 材料性能研究热度变化

(b) 适用场景研究热度变化

图 5.2 材料性能和适用场景研究热度在不同年份间的变化。(a) 图中 Hard 是硬度, Corr 是耐腐蚀性, Strg 是强度, YS 是屈服强度, TS 是抗拉强度, Fag 是抗疲劳性, Elg 是伸长率, Wear 是耐磨性。(b) 图中 NPP 是核电领域, As 是航空航天, CI 是化学工业, BA 是生物医学, MD 是医疗器械, Cs 是建筑, AI 是汽车工业, AMP 是增材制造零件。

使用技术: 统计背向散射电子衍射技术 (EBSD)、X 射线衍射 (XRD)、能量色散光谱 (EDS) 等技术, 其中 EBSD 最常用并且所占比例逐年增加, 完整的数据显示在图5.3中。EBSD 是一种在扫描电子显微镜中进行毫米到纳米尺度量级的定量微观结构分析的技术, 可以提供材料的晶体学信息, 例如晶界角度、晶粒大小、纹理、应变等。此外, EBSD 还可以与其他技术相结合, 如能量色散光谱、波长色散谱等, 实现多模式分析, 并提高材料表征的效率和准确性。

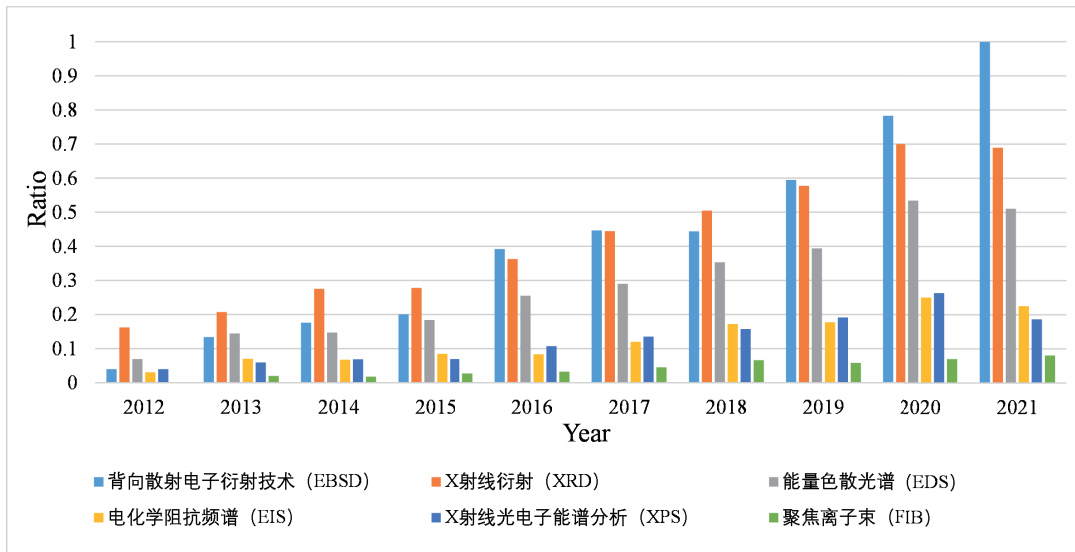


图 5.3 使用技术研究热度在不同年份间的变化

涉及元素：材料成分中的元素对材料的性能有重要影响，本文对 15 种元素数据进行统计分析，例如铬 (Cr)、氮 (N)、镍 (Ni)、钼 (Mo) 等，统计结果显示在图5.4中。对于不锈钢，研究人员最关注的元素是铬，该元素是不锈钢的主要合金元素，只有当铬含量达到一定值时，钢材才具有耐蚀性，原因在于铬元素使得钢材电极电位发生突变，从负电位变成正电位，从而抑制空气中的氧化反应。另一方面，铬元素可以在钢材表面形成一层紧密的氧化铬膜（称为钝化膜），防止水和空气接触钢材，从而保护钢材免受进一步的腐蚀。因此，在不锈钢材料中，铬元素是提高耐腐蚀性和保护金属光泽的关键元素。

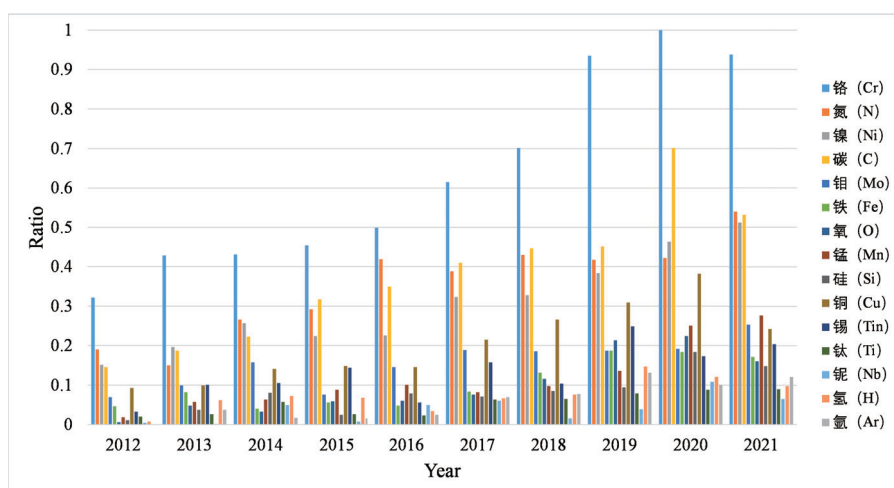


图 5.4 涉及元素研究热度在不同年份间的变化

强度数值：按年份分类，收集得到 2100 组抗拉强度和屈服强度值，使用分类散

点图将数据显示在图5.5中。可以清晰地观察到，文献中提到的强度值普遍越来越高。例如，2021年强度为1000MPa的频率与以前的年份相比显著增加。

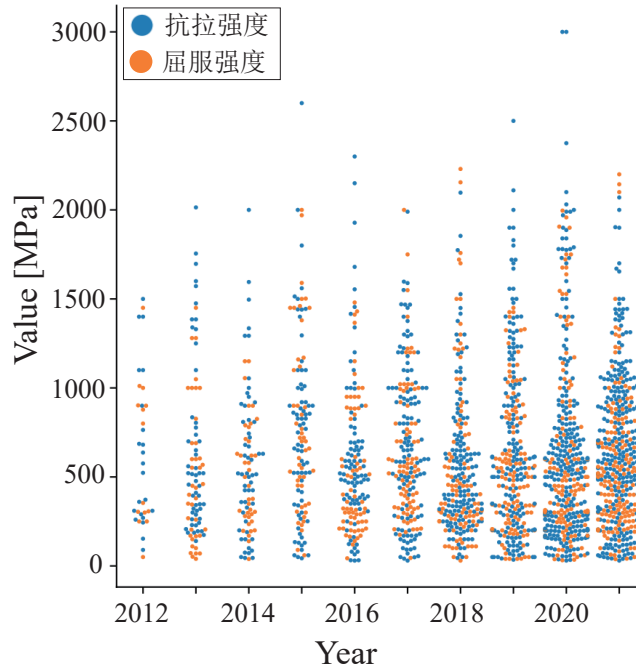
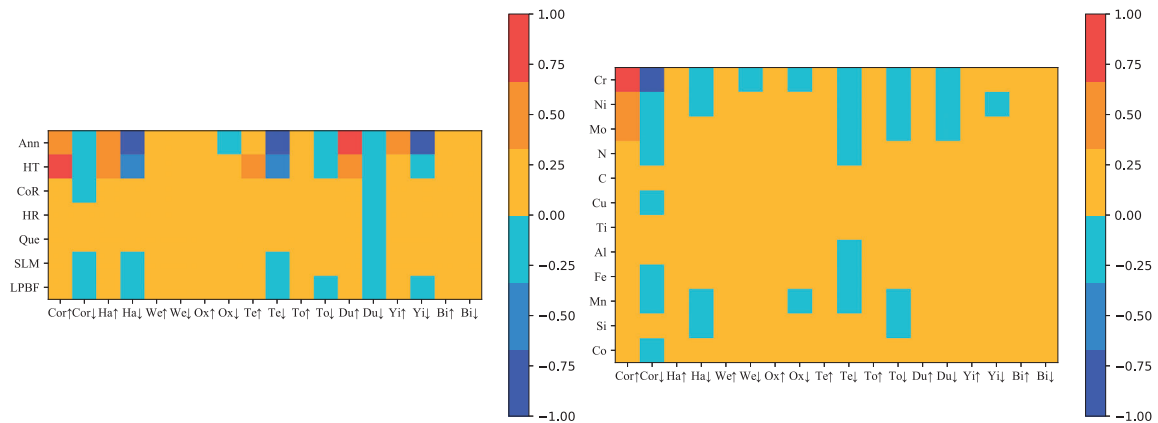


图 5.5 抗拉强度和屈服强度值在不同年份的分布

5.2.2 多实体类别联合分析

从文献提取结果中收集 2012 年至 2021 年期间的 674 条三元组（性能，性能趋势，工艺）数据，并绘制“性能-趋势-工艺”关系热力图，如图5.6(a)所示；整理 1065 条三元组（性能，性能趋势，涉及元素）数据，并绘制“性能-趋势-元素”关系热力图，如图5.6(b)所示。图5.6(a)和图5.6(b)分别展示了统计数据中，7 种工艺对 9 种性能变化的影响，以及 12 种元素对 9 种性能变化的影响，颜色越红表示性能提高出现频次越高，颜色越蓝表示性能降低出现频次越高。性能包括抗腐蚀性、硬度、耐磨性、抗氧化性、抗拉强度、韧性、延展性、屈服强度和生物相容性；性能趋势的描述包括提高（↑）和降低（↓）；工艺包括退火、热处理、冷轧、热轧、淬火、选择性激光熔化和激光粉末床融合；涉及元素包括铬、镍、钼、氮、碳、铜、钛、铝、铁、锰、硅和钴元素。从图5.6(a)中可以观察到，退火和热处理工艺与抗腐蚀性、延展性的提高密切相关，与硬度、抗拉强度屈服强度的衰退也密切相关。从图5.6(b)中可以观察到，铬、镍和钼在提高材料耐腐蚀性方面起着重要作用。



(a) 性能-趋势-工艺

(b) 性能-趋势-元素

图 5.6 工艺、元素与性能变化趋势关系的热力图。图例中颜色越接近红色表示性能提高出现频次越高，颜色越接近蓝色表示性能降低出现频次越高。(a) 和 (b) 图的 X 轴表示不同性能的提高或降低，性能包括抗腐蚀 (Cor)、硬度 (Ha)、耐磨性 (We)、抗氧化性 (Ox)、抗拉强度 (Te)、韧性 (To)、延展性 (Du)、屈服强度 (Yi) 和生物相容性 (Bi)，性能趋势包括提高 (↑) 和降低 (↓)。(a) 图中 Y 轴表示工艺，包括退火 (Ann)、热处理 (HT)、冷轧 (CoR)、热轧 (HR)、淬火 (Que)、选择性激光熔化 (SLM)、激光粉末床融合 (LPBF)。(b) 图中 Y 轴表示元素，包括铬 (Cr)、镍 (Ni)、钼 (Mo)、氮 (N)、碳 (C)、铜 (Cu)、钛 (Ti)、铝 (Al)、铁 (Fe)、锰 (Mn)、硅 (Si) 和钴 (Co)。

5.3 不锈钢性能预测

本文从不锈钢文献提取得到的 236 万个材料实体和 7970 组成分信息中筛选出材料成分和抗拉强度数据，使用第四章中提出的性能预测方法，对不锈钢材料的抗拉强度进行预测。此外还筛选出材料成分、工艺、性能和性能变化数据，使用机器学习算法对不锈钢材料抗腐蚀性、延展性、强度和硬度进行性能变化趋势预测。

5.3.1 抗拉强度性能值预测

从文献文本和表格提取结果中，收集到 321 条不锈钢“材料成分-抗拉强度”数据用于训练预测模型，成分元素包括铬、镍、钼、氮、锰、硅、铜和碳，元素含量单位为质量百分含量 (mass%)，抗拉强度单位为兆帕 (MPa)。第四章中提出的交叉特征选择及性能预测方法，对不锈钢材料的抗拉强度进行预测，具体内容已在第四章实验部分介绍过，最终使用 XGBoost 训练的抗拉强度预测模型 R^2 得分为 0.6671。

5.3.2 四种性能变化趋势预测

从文献文本和表格的提取结果中，收集到 376 条（成分，工艺，耐腐蚀趋势）数据用于训练耐腐蚀预测模型，313 条（成分，工艺，延展性趋势）数据用于训练延展性预测模型，265 条（成分，工艺，硬度趋势）数据用于训练硬度预测模型，756 条（成分，工艺，强度趋势）数据用于训练强度预测模型，训练得到的模型以材料成分和工艺为输入，以性能变化趋势为输出。工艺和性能趋势数据来自文本提取结果，材料成分数据来自表格识别结果，材料成分由铬、镍、钼、氮、锰、铝、硅、钛、铜、碳和钴元素组成，工艺包括退火、热处理、热轧、冷轧、淬火、回火、选择性激光熔化和激光粉末床融合，性能变化趋势分为提高和降低。

在收集得到四组数据上分别进行性能预测实验。使用机器学习决策树 (DT)、随机森林 (RF)、K 近邻 (KNN)、AdaBoost 和 GBDT 算法，在不同数据上训练对应的性能预测模型，并使用十折交叉验证的方式分别验证 DT、RF、KNN、AdaBoost 和 GBDT 的准确性，评价指标为准确率 (Precision, P_{prop})、召回率 (Recall, R_{prop}) 和 F1 得分 (F1-score, $F1_{prop}$)，计算公式如 5.1-5.3 所示，实验结果见表 5.1 所示。

$$P_{prop} = \frac{TP}{TP + FP} \quad (5.1)$$

$$R_{prop} = \frac{TP}{TP + FN} \quad (5.2)$$

$$F1_{prop} = \frac{2P_{prop}R_{prop}}{P_{prop} + R_{prop}} \quad (5.3)$$

其中， TP 表示正确预测性能提高的样本数量， FP 表示将性能降低错误的预测为性能提高的样本数量， FN 表示将性能提高错误的预测为性能降低的样本数量， TN 表示正确预测性能降低的样本数量。其中，GBDT 算法整体预测效果最好，因此使用 GBDT 在四组数据上训练四个性能预测模型，所有模型都使用 GBDT 默认参数进行初始化。抗腐蚀性预测模型 F1 得分为 81.02%，延展性预测模型 F1 得分为 83.51%，强度预测模型 F1 得分为 80.13%，硬度预测模型 F1 得分为 80.23%。每个模型都可以根据输入的材料成分和工艺，预测相应的性能变化趋势。

此外，为了验证文本信息与表信息相结合（工艺 + 成分）的有效性，本文还单独使用文本信息（工艺）和单独使用表格信息（成分）对四种性能的变化趋势进行预

测, 结果如表5.2和表5.3所示。图5.7显示了使用 GBDT 对文本数据和表格数据、仅文本数据和仅表格数据进行性能预测的得分比较, 由此可见将文本和表格提取结果应用于性能预测的效果, 要比仅使用文本或表格提取结果更好。

表 5.1 机器学习在文本和表格数据上对四种性能趋势预测的评价指标得分

算法	抗腐蚀			延展性			强度			硬度		
	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$
DT	76.29	60.42	65.49	71.44	67.18	60.62	55.05	91.42	66.31	79.61	52.41	62.20
RF	57.27	82.46	58.57	60.33	90.20	71.66	57.04	83.51	64.03	69.25	66.97	67.13
KNN	72.85	78.16	74.72	67.79	78.14	71.91	74.43	83.51	78.60	70.28	75.82	72.17
AdaBoost	80.13	84.41	81.93	79.86	84.34	81.66	79.26	81.70	80.10	74.77	74.91	74.52
GBDT	79.93	84.38	81.02	82.45	85.06	83.51	75.07	86.20	80.13	82.92	78.59	80.23

表 5.2 机器学习在文本数据上对四种性能趋势预测的评价指标得分。每个数据都与表5.1中的对应数据比较, “↑”意思是得分提高, “↓”意思是得分下降。

算法	抗腐蚀			延展性			强度			硬度		
	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$
DT	84.29↑	48.98↓	61.43↓	65.34↓	66.79↓	51.86↓	36.44↓	70.00↓	47.89↓	65.16↓	47.99↓	50.42↓
RF	62.37↑	80.87↓	67.83↑	57.04↓	91.52↑	69.69↓	39.33↓	80.00↓	52.49↓	64.49↓	65.66↓	64.46↓
KNN	58.73↓	74.53↓	64.35↓	66.18↓	67.69↓	63.94↓	61.55↓	72.68↓	65.52↓	66.47↓	68.97↓	65.31↓
AdaBoost	58.97↓	87.35↑	69.80↓	71.73↓	67.66↓	68.98↓	60.66↓	79.45↓	68.06↓	66.97↓	50.05↓	54.79↓
GBDT	66.33↓	85.63↑	74.19↓	76.09↓	70.71↓	72.96↓	61.88↓	82.46↓	70.28↓	77.23↓	65.16↓	69.16↓

表 5.3 机器学习在表格数据上对四种性能趋势预测的评价指标得分。每个数据都与表5.1中的对应数据比较, “↑”意思是得分提高, “↓”意思是得分下降。

算法	抗腐蚀			延展性			强度			硬度		
	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$	P_{prop}	R_{prop}	$F1_{prop}$
DT	78.80↑	52.37↓	61.01↓	75.82↑	61.33↓	59.99↓	56.22↑	82.40↓	58.72↓	68.29↓	59.25↑	58.94↓
RF	70.05↑	69.85↓	67.63↑	63.53↑	83.94↓	71.34↓	58.58↑	81.61↓	60.79↓	62.79↓	71.57↑	64.98↓
KNN	73.21↑	69.61↓	70.21↓	70.30↑	75.64↓	71.97↑	69.02↓	74.51↓	70.72↓	68.60↓	60.54↓	63.62↓
AdaBoost	77.81↓	75.54↓	76.27↓	76.40↓	86.42↑	80.46↓	68.12↓	73.92↓	70.30↓	68.79↓	70.04↓	68.94↓
GBDT	80.30↑	79.27↓	79.57↓	83.78↑	83.74↓	82.31↓	75.15↑	76.47↓	75.55↓	75.14↓	77.00↓	75.15↓

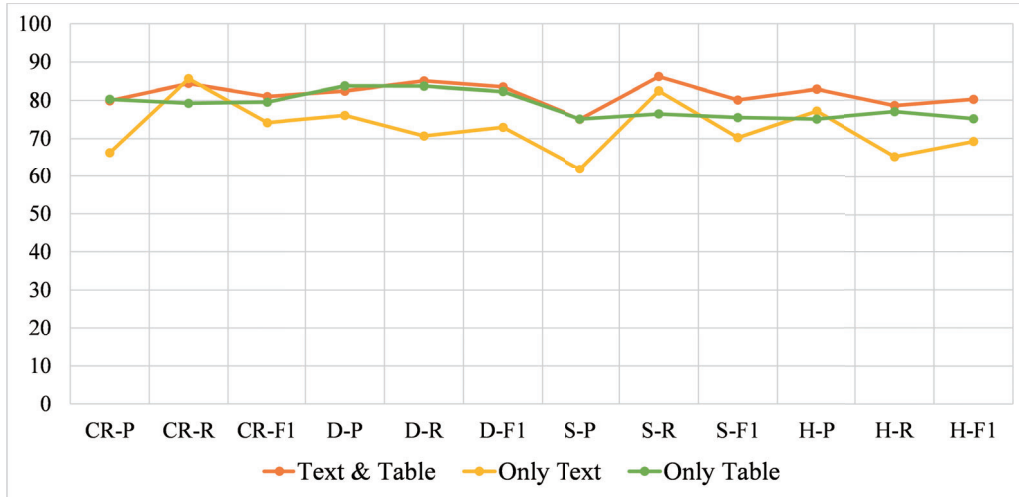


图 5.7 使用 GBDT 在文本表格结合、仅文本和仅表格数据上预测性能趋势的得分比较。X 轴中 CR 代表抗腐蚀性，D 代表延展性，S 代表强度，H 代表硬度，P 为准确率 (P_{prop})，R 为召回率 (R_{prop})，F1 为 F1 得分 ($F1_{prop}$)；Y 轴表示评价指标得分。

5.4 本章小结

本章将第三章文献提取方法和第四章性能预测方法应用在 11,058 篇不锈钢文献上，统计材料名称、处理工艺、使用技术、材料性能、涉及元素和适用场景这六种命名实体在近十年间研究热度的变化，汇总了抗拉强度和屈服强度值的分布情况，还分析了不同实体间的潜在关系。此外，利用文献提取结果中材料成分和抗拉强度数据训练得到抗拉强度值预测模型，还筛选出材料成分、工艺、性能和性能变化数据，对不锈钢材料抗腐蚀性、延展性、强度和硬度的变化趋势进行预测。

第六章 总结与展望

6.1 总结

本研究结合自然语言处理、传统图像处理和计算材料学等多种领域知识，对利用计算机技术实现材料文献数据挖掘及应用进行了重点研究。由于科学文献上下文中包含文本和非文本内容，现有文献挖掘工作常常忽视了包含重要信息的非文本内容如表格，导致文献信息提取不充分。本文针对材料文献内容的特点，结合现有理论知识提出一种基于上下文感知的材料文献提取方法，并将挖掘结果在材料性能预测上进行应用。本文的研究成果如下所示：

(1) 针对材料文献文本包含众多专业词汇的特点，在 3.1 节中提出基于动静态词向量融合的命名实体识别方法。该方法将包含上下文语境信息的动态词向量与具有材料领域知识的静态词向量相融合，使得每个词向量中都嵌入上下文信息和领域知识，在不需要使用大规模材料语料库对语言模型微调的前提下，显著提高了材料文本命名实体识别效果。

(2) 根据材料文献中成分表格的结构特征，在 3.2 节中提出基于传统图像技术的成分表格识别方法。该方法在预定义成分表格结构规则的前提下，使用形态学方法、二值化轮廓检测技术、文本相似度计算等方法，将成分表格分类为单种材料表格和多种材料表格，并进一步把表格拆解为标题、表头和表体，分别从这三种区域中提取出材料名称、元素、元素含量和单位信息。

(3) 基于从材料文献上下文中提取到的文本和表格数据，在第四章中提出一种基于文献信息提取的材料性能预测方法。从文献提取结果中筛选出材料成分和抗拉强度数据，使用 XenonPy 材料信息学库对成分数据进行特征扩充，根据特征扩充的原理分别对元素属性统计特征进行交叉特征压缩和特征选择，使用机器学习算法在选择的特征上预测抗拉强度。该方法显著提高了抗拉强度性能的预测效果，并为数据驱动的材料性能预测提供了一种不依赖于人工数据的可行性方法。

(4) 以不锈钢为示范材料，将文献提取和性能预测方法应用在 11,058 篇科学文献上，对上述方法的可行性和有效性进行验证，应用结果可以为不锈钢材料的研究方向提供指导，为性能研究和优化提供参考。

本文对提出的三种方法都进行了实验探究，实验结果表明（1）命名实体识别方法能更准确地提取材料命名实体；（2）成分表格识别方法能够简单准确地获取文献中成分信息；（3）材料成分特征处理能够提高性能预测的准确率。本工作为材料内幕关系的探究、材料数据库建设提供数据来源，为数据驱动的材料研究方法提供了一种新的可行方案。

6.2 展望

尽管本文所提出的材料文献挖掘及应用方法能够更好地从科学文献中获取相关信息，并辅助材料性能研究工作。但是，仍存在一定局限性，作者认为本工作在以下两个方面值得更进一步的探究：

（1）文献挖掘方面：首先，科学文献中不仅包含文本和表格内容，一些非文本内容比如提供关键信息的数据统计图、示意图、公式等也包含着重要内容，这些信息的提取结果能为文本内容提供支撑和补充。其次，自然语言处理、计算机视觉等领域的发展，多种数据源融合的多模态方法已经被广泛应用，可以利用端到端的多模态深度学习模型对文献中各类信息进行融合提取，以实现计算机对文献的充分理解。此外，现有文献挖掘工作主要针对各种文献库进行挖掘，但未来文献挖掘可能会结合多种数据源，如专利、科技新闻、社交媒体等，以获取更全面的信息。

（2）材料应用方面：首先，数据的质量是影响数据驱动材料研究的关键点之一，需要更完善的数据挖掘和处理方法，修正或去除包含误差和噪声的材料数据，提高模型的准确性和预测能力。其次，数据驱动材料研究需要大量数据对模型进行训练和测试，某些高端材料或实验成本高的材料可能缺乏足够的的数据，可以借助半监督学习、迁移学习等方法充分利用现有数据。此外，数据驱动的材料研究中使用的机器学习模型，其背后的预测原理可能难以解释，导致在实际应用中存在障碍，可以进一步发展可解释性机器学习，以促进材料科学领域的发展与应用。

以上问题仍需要不断的实验和探索，完善文献挖掘和材料研究工作，加快数据驱动的材料研发步伐，推动材料科学的发展和创新，促进材料研发的成果转化，带动经济增长和社会进步。

插图索引

图 1.1	本文组织结构图	7
图 2.1	自然语言处理基础研究和应用研究的分类	10
图 2.2	一种标准 NER 任务的总体流程	10
图 2.3	LSTM 神经元细胞结构图	12
图 2.4	CBOW 模型结构图	16
图 2.5	Skip-Gram 模型结构图	17
图 2.6	形态学膨胀操作示意图	18
图 2.7	形态学腐蚀操作示意图	19
图 2.8	机器学习在材料性能预测研究中的工作流程	20
图 3.1	基于上下文感知的材料文献信息提取方法结构图	22
图 3.2	材料命名实体识别模型 SFBC 的结构图	24
图 3.3	命名实体识别评价指标示意图	30
图 3.4	SFBC 模型在不同数据集上训练曲线	34
图 3.5	材料成分表格识别方法示意图	38
图 3.6	材料成分表格提取整体流程图	39
图 3.7	成分表格形态学预处理的过程	40
图 3.8	成分表格文本内容及区域坐标获取	42
图 3.9	材料成分表格结构示意图	43
图 4.1	基于文献信息提取的材料性能预测方法整体架构	48
图 4.2	基于文献信息提取的材料性能预测方法流程	49

图 4.3	406 维元素属性统计特征十字交叉压缩	52
图 4.4	实验 4 中 XGBoost 在 60 维特征上真实值与预测值散点图	61
图 5.1	材料名称和处理工艺研究热度在不同年份间的变化	64
图 5.2	材料性能和适用场景研究热度在不同年份间的变化	64
图 5.3	使用技术研究热度在不同年份间的变化	65
图 5.4	涉及元素研究热度在不同年份间的变化	65
图 5.5	抗拉强度和屈服强度值在不同年份的分布	66
图 5.6	工艺、元素与性能变化趋势关系的热力图	67
图 5.7	使用 GBDT 在不同数据上进行性能预测的评价指标得分比较	70

表格索引

表 3.1	InorgNerData 数据集实体标签定义.....	29
表 3.2	融合与非融合方法在 SLSNerData 上总体得分对比.....	32
表 3.3	融合与非融合方法在 SLSNerData 上 13 类实体 F1 得分对比	33
表 3.4	融合与非融合方法在 InorgNerData 上总体得分对比	34
表 3.5	不同方法在 SLSNerData 上总体得分对比.....	35
表 3.6	不同方法在 SLSNerData 上 13 类实体 F1 得分对比	36
表 3.7	不同方法在 InorgNerData 上准确率、召回率和 F1 得分对比.....	37
表 4.1	XenonPy 库中 7 种统计特征计算器及数学计算公式.....	50
表 4.2	样例材料原子核中质子数属性的统计特征计算结果	51
表 4.3	406 维元素属性统计特征水平压缩结果重要性评分	54
表 4.4	406 维元素属性统计特征垂直压缩结果重要性评分	54
表 4.5	从文献中提取到的不锈钢成分-抗拉强度数据特征统计信息	56
表 4.6	Kinzoku 库中不锈钢成分-抗拉强度数据特征统计信息.....	57
表 4.7	不同实验中 $RMSE$ 评价指标的对比结果	59
表 4.8	不同实验中 MAE 评价指标的对比结果	59
表 4.9	不同实验中 R^2 评价指标的对比结果	60
表 5.1	机器学习在组合数据上对四种性能趋势预测的评价指标得分	69
表 5.2	机器学习在文本数据上对四种性能趋势预测的评价指标得分	69
表 5.3	机器学习在表格数据上对四种性能趋势预测的评价指标得分	69

参考文献

- [1] ZHOU T, SONG Z, SUNDMACHER K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design[J]. *Engineering*, 2019, 5(6): 1017-1026.
- [2] LOPEZ-BEZANILLA A, LITTLEWOOD P B. Growing field of materials informatics: databases and artificial intelligence[J]. *MRS Communications*, 2020, 10(1): 1-10.
- [3] TOLLE K M, TANSLEY D S W, HEY A J. The fourth paradigm: Data-intensive scientific discovery[J]. *Proceedings of the IEEE*, 2011, 99(8): 1334-1337.
- [4] HIMANEN L, GEURTS A, FOSTER A S, et al. Data-driven materials science: status, challenges, and perspectives[J]. *Advanced Science*, 2019, 6(21): 1900808.
- [5] TIAN C, LI T, BUSTILLOS J, et al. Data-driven approaches toward smarter additive manufacturing[J]. *Advanced Intelligent Systems*, 2021, 3(12): 2100014.
- [6] ZHANG J, LI K, YAO C, et al. Event-based summarization method for scientific literature[J]. *Personal and Ubiquitous Computing*, 2021, 25: 959-968.
- [7] ZEYU Z, HAO W, ZIBO Z, et al. Construction and application of gcn model for text classification with associated information[J]. *Data Analysis and Knowledge Discovery*, 2021, 5(9): 31-41.
- [8] 胡少虎, 张颖怡, 章成志. 关键词提取研究综述 [J]. *数据分析与知识发现*, 2020, 5(3): 45-59.
- [9] KUMAR A, STARLY B. “fabner” : information extraction from manufacturing process science domain literature using named entity recognition[J]. *Journal of Intelligent Manufacturing*, 2022, 33(8): 2393-2407.
- [10] LI J, SONG H, LI J, et al. Comprehensive analysis and classification of natural language questions based on bi-lstm-crf[C]//2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). Piscataway: IEEE, 2022: 1254-1258.
- [11] OLIVETTI E A, COLE J M, KIM E, et al. Data-driven materials research enabled by natural language processing and information extraction[J]. *Applied Physics Reviews*, 2020, 7(4): 041317.

- [12] KUNIYOSHI F, OZAWA J, MIWA M. Analyzing research trends in inorganic materials literature using nlp[C]//Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2021: 319-334.
- [13] HONG Z, WARD L, CHARD K, et al. Challenges and advances in information extraction from scientific literature: a review[J]. JOM, 2021, 73(11): 3383-3400.
- [14] ZHANG L, HE M. Text mining for energy materials[J]. Journal of Research in Science and Engineering, 2022, 4(3): 117-130.
- [15] SMITH A, BHAT V, AI Q, et al. Challenges in information-mining the materials literature: A case study and perspective[J]. Chemistry of Materials, 2022, 34(11): 4821-4827.
- [16] 谢红玲, 奉国和, 何伟林. 基于深度学习的科技文献语义分类研究 [J]. 情报理论与实践, 2018, 41(11): 149.
- [17] 罗鹏程, 王一博, 王继民. 基于深度预训练语言模型的文献学科自动分类研究 [J]. 情报学报, 2020, 39(10): 1046-1059.
- [18] LILI S, PENG J, JING W. A study on the automatic classification of chinese literature in periodicals based on bert model[J]. Libraly Journal, 2022, 41(5): 109.
- [19] MORAES R, VALIATI J F, NETO W P G. Document-level sentiment classification: An empirical comparison between svm and ann[J]. Expert Systems with Applications, 2013, 40(2): 621-633.
- [20] ZHANG L, WANG S, LIU B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1253.
- [21] BIRJALI M, KASRI M, BENI-HSSANE A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends[J]. Knowledge-Based Systems, 2021, 226: 107134.
- [22] CHEN Y, ZHANG H, LIU R, et al. Experimental explorations on short text topic mining between lda and nmf based schemes[J]. Knowledge-Based Systems, 2019, 163: 1-13.
- [23] ALBALAWI R, YEAP T H, BENYOUCEF M. Using topic modeling methods for short-text data: A comparative analysis[J]. Frontiers in Artificial Intelligence, 2020, 3: 42.
- [24] WESTON L, TSHITTOYAN 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, 2019, 59(9): 3692-3702.

- [25] WESTERGAARD D, STAERFELDT H H, TONSBORG C, et al. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts[J]. *PLoS computational biology*, 2018, 14(2): e1005962.
- [26] NAZEMI K, KLEPSCH M J, BURKHARDT D, et al. Comparison of full-text articles and abstracts for visual trend analytics through natural language processing[C]//2020 24th International Conference Information Visualisation (IV). Piscataway: IEEE, 2020: 360-367.
- [27] GORRELL G, SONG X, ROBERTS A. Bio-yodie: A named entity linking system for biomedical text[J]. *arXiv preprint arXiv:1811.04860*, 2018.
- [28] DREISBACH C, KOLECK T A, BOURNE P E, et al. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data[J]. *International journal of medical informatics*, 2019, 125: 37-46.
- [29] WEBER L, MUNCHMEYER J, ROCKTASCHEL T, et al. Huner: improving biomedical ner with pretraining[J]. *Bioinformatics*, 2020, 36(1): 295-302.
- [30] GUHA S, MULLICK A, AGRAWAL J, et al. Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature[J]. *Computational Materials Science*, 2021, 192: 110325.
- [31] SHETTY P, RAMPRASAD R. Automated knowledge extraction from polymer literature using natural language processing[J]. *Iscience*, 2021, 24(1): 101922.
- [32] NANDY A, DUAN C, KULIK H J. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal-organic frameworks[J]. *Journal of the American Chemical Society*, 2021, 143(42): 17535-17547.
- [33] CRUSE K, TREWARTHA A, LEE S, et al. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities[J]. *Scientific Data*, 2022, 9(1): 234.
- [34] 张宇童, 李启元. 表格检测与结构识别综述 [J]. *计算机工程与应用*, 2022, 58(22): 1-11.
- [35] MAURO J C, TANDIA A, VARGHEESE K D, et al. Accelerating the design of functional glasses through modeling[J]. *Chemistry of Materials*, 2016, 28(12): 4267-4277.
- [36] BHASKAR P, KUMAR R, MAURYA Y, et al. Cooling rate effects on the structure of 45s5 bio-glass: Insights from experiments and simulations[J]. *Journal of Non-Crystalline Solids*, 2020, 534: 119952.

- [37] RAVI A, et al. Prediction of reduced glass transition temperature using machine learning[J]. arXiv preprint arXiv:2005.08872, 2020.
- [38] XIONG J, ZHANG T, SHI S. Machine learning of mechanical properties of steels[J]. Science China Technological Sciences, 2020, 63(7): 1247-1255.
- [39] SI S, FAN B, LIU X, et al. Study on strengthening effects of zr-ti-nb-o alloys via high throughput powder metallurgy and data-driven machine learning[J]. Materials & Design, 2021, 206: 109777.
- [40] ZHANG B, SHIN Y C. Data-driven phase recognition of steels for use in mechanical property prediction[J]. Manufacturing Letters, 2021, 30: 27-31.
- [41] GENG X, CHENG Z, WANG S, et al. A data-driven machine learning approach to predict the hardenability curve of boron steels and assist alloy design[J]. Journal of Materials Science, 2022, 57(23): 10755-10768.
- [42] ROY I, FENG B, ROYCHOWDHURY S, et al. Understanding oxidation of fe-cr-al alloys through explainable artificial intelligence[J]. MRS communications, 2023, 13: 1-7.
- [43] 韩云飞, 谢佳, 蔡涛, 等. 结合高斯过程回归与特征选择的锂离子电池容量估计方法 [J]. 储能科学与技术, 2021, 10(4): 1432.
- [44] 胡建军, 曹卓, 但雅波, 等. 基于特征选择和机器学习的材料弹性性能预测 [J]. 华南理工大学学报 (自然科学版), 2019, 47(5): 48-55.
- [45] LAURIOLA I, LAVELLI A, AIOLLI F. An introduction to deep learning in natural language processing: Models, techniques, and tools[J]. Neurocomputing, 2022, 470: 443-456.
- [46] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
- [47] YADAV V, BETHARD S. A survey on recent advances in named entity recognition from deep learning models[J]. arXiv preprint arXiv:1910.11470, 2019.
- [48] GRISHMAN R. Information extraction: Techniques and challenges[C]//Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School. Berlin: Springer, 1997: 10-27.
- [49] TREWARTHA A, WALKER N, HUO H, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science[J]. Patterns, 2022, 3(4): 100488.

- [50] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural computation*, 1997, 9(8): 1735-1780.
- [51] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//*Proceedings of the eighteenth international conference on machine learning*. New York: ACM, 2001: 282–289.
- [52] VITERBI A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm[J]. *IEEE transactions on Information Theory*, 1967, 13(2): 260-269.
- [53] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. *arXiv preprint arXiv:1301.3781*, 2013.
- [54] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. *arXiv preprint arXiv:1810.04805*, 2018.
- [55] ZHU Y, KIROS R, ZEMEL R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]//*Proceedings of the IEEE international conference on computer vision*. Piscataway: IEEE, 2015: 19-27.
- [56] CHIBANI S, COUDERT F X. Machine learning approaches for the prediction of materials properties[J]. *Apl Materials*, 2020, 8(8): 080701.
- [57] GOODALL R E, LEE A A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry[J]. *Nature communications*, 2020, 11(1): 6280.
- [58] AGRAWAL A, CHOUDHARY A. An online tool for predicting fatigue strength of steel alloys based on ensemble data mining[J]. *International Journal of Fatigue*, 2018, 113: 389-400.
- [59] DUNN A, WANG Q, GANOSE A, et al. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm[J]. *npj Computational Materials*, 2020, 6(1): 138.
- [60] ZUO H, JIANG Y, YANG Y, et al. Prediction of properties of metal alloy materials based on machine learning[J]. *arXiv preprint arXiv:2109.09394*, 2021.
- [61] WANG X, TRAN N D, ZENG S, et al. Element-wise representations with ecnet for material property prediction and applications in high-entropy alloys[J]. *npj Computational Materials*, 2022, 8(1): 253.

- [62] GUO K, YANG Z, YU C H, et al. Artificial intelligence and machine learning in design of mechanical materials[J]. *Materials Horizons*, 2021, 8(4): 1153-1172.
- [63] ALLEN A E, TKATCHENKO A. Machine learning of material properties: Predictive and interpretable multilinear models[J]. *Science advances*, 2022, 8(18): eabm7185.
- [64] KAUTZ E J. Predicting material microstructure evolution via data-driven machine learning[J]. *Patterns*, 2021, 2(7): 100285.
- [65] BELTAGY I, LO K, COHAN A. Scibert: A pretrained language model for scientific text[J]. *arXiv preprint arXiv:1903.10676*, 2019.
- [66] KIM E, JENSEN Z, VAN GROOTEL A, et al. Inorganic materials synthesis planning with literature-trained neural networks[J]. *Journal of chemical information and modeling*, 2020, 60(3): 1194-1201.
- [67] KIM E, HUANG K, TOMALA A, et al. Machine-learned and codified synthesis parameters of oxide materials[J]. *Scientific data*, 2017, 4(1): 1-9.
- [68] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. *arXiv preprint arXiv:1607.01759*, 2016.
- [69] SCHUSTER M, NAKAJIMA K. Japanese and korean voice search[C]//2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). Piscataway: IEEE, 2012: 5149-5152.
- [70] RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning[J]. *Natural language processing using very large corpora*, 1999, 11: 157-176.
- [71] ZHANG R, ZHANG J, HAN Y, et al. Ner-scibert-fasttext-bilstm-crf[EB/OL]. 2023. <https://github.com/han-yuexing/NER-SciBERT-Fasttext-BiLSTM-CRF>.
- [72] NAKAYAMA H, KUBO T, KAMURA J, et al. doccano: Text annotation tool for human[EB/OL]. 2018. <https://github.com/doccano/doccano>.
- [73] HOSSEINI V A, THUVANDER M, LINDGREN K, et al. Fe and cr phase separation in super and hyper duplex stainless steel plates and welds after very short aging times[J]. *Materials & Design*, 2021, 210: 110055.

- [74] ZHAO C, BAI Y, ZHANG Y, et al. Influence of scanning strategy and building direction on microstructure and corrosion behaviour of selective laser melted 316L stainless steel[J]. *Materials & Design*, 2021, 209: 109999.
- [75] MASUMURA T, TSUCHIYAMA T. Effect of carbon and nitrogen on work-hardening behavior in metastable austenitic stainless steel[J]. *ISIJ International*, 2021, 61(2): 617-624.
- [76] TABRIZI T R, SABZI M, ANIJAN S M, et al. Comparing the effect of continuous and pulsed current in the gtaw process of aisi 316l stainless steel welded joint: microstructural evolution, phase equilibrium, mechanical properties and fracture mode[J]. *Journal of Materials Research and Technology*, 2021, 15: 199-212.
- [77] MA Q, LUO C, LIU S, et al. Investigation of arc stability, microstructure evolution and corrosion resistance in underwater wet fcaw of duplex stainless steel[J]. *Journal of Materials Research and Technology*, 2021, 15: 5482-5495.
- [78] ZHANG C, ZHU J, JI C, et al. Laser powder bed fusion of high-entropy alloy particle-reinforced stainless steel with enhanced strength, ductility, and corrosion resistance[J]. *Materials & Design*, 2021, 209: 109950.
- [79] SALAHI S, KAZEMIPOUR M, NASIRI A. Effects of microstructural evolution on the corrosion properties of aisi 420 martensitic stainless steel during cold rolling process[J]. *Materials Chemistry and Physics*, 2021, 258: 123916.
- [80] NIE J, WEI L, JIANG Y, et al. Corrosion mechanism of additively manufactured 316 L stainless steel in 3.5 wt.% nacl solution[J]. *Materials Today Communications*, 2021, 26: 101648.
- [81] LEE S Y, TAKUSHIMA C, HAMADA J I, et al. Macroscopic and microscopic characterizations of portevin-lechatelier effect in austenitic stainless steel using high-temperature digital image correlation analysis[J]. *Acta Materialia*, 2021, 205: 116560.
- [82] TAKAI T, FURUKAWA T, YAMANO H. Thermophysical properties of austenitic stainless steel containing boron carbide in a solid state[J]. *Mechanical Engineering Journal*, 2021, 8(4): 2000540.
- [83] ZHANG X, YANG S, LI J, et al. Evolution of oxide inclusions in stainless steel containing yttrium during thermo-mechanical treatment[J]. *Journal of Materials Research and Technology*, 2020, 9 (3): 5982-5990.

- [84] ANNAMORADNEJAD I, ZOGHI G. Colbert: Using bert sentence embedding for humor detection[J]. arXiv preprint arXiv:2004.12765, 2020.
- [85] LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
- [86] ALSENTZER E, MURPHY J R, BOAG W, et al. Publicly available clinical bert embeddings[J]. arXiv preprint arXiv:1904.03323, 2019.
- [87] LEE J, YOON W, KIM S, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining[J]. *Bioinformatics*, 2020, 36(4): 1234-1240.
- [88] TREWARTHA A, WALKER N, HUO H, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science[J]. *Patterns*, 2022, 3(4): 100488.
- [89] GUPTA T, ZAKI M, KRISHNAN N A. Matscibert: A materials domain language model for text mining and information extraction[J]. *npj Computational Materials*, 2022, 8(1): 102.
- [90] SUAREZ L F P O. Material scibert (tpu): Improving language understanding in materials science[EB/OL]. 2022. <https://huggingface.co/lfoppiano/MatTPUSciBERT>.
- [91] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
- [92] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2016, 39(11): 2298-2304.
- [93] PENG J, SHI X, SUN Y, et al. Qtlminer: Qtl database curation by mining tables in literature[J]. *Bioinformatics*, 2015, 31(10): 1689-1691.
- [94] LIU Y, BAI K, MITRA P, et al. Tableseer: automatic table metadata extraction and searching in digital libraries[C]//*Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. New York: ACM, 2007: 91-100.
- [95] WU S, LAMBARD G, LIU C, et al. iqspr in xenonpy: a bayesian molecular design algorithm[J]. *Molecular informatics*, 2020, 39(1-2): 1900107.
- [96] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. New York: ACM, 2016: 785-794.

- [97] DHAL P, AZAD C. A comprehensive survey on feature selection in the various fields of machine learning[J]. *Applied Intelligence*, 2022, 52: 1-39.
- [98] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]// *Advances in Neural Information Processing Systems*. New York: Curran Associates, 2017: 3146–3154.
- [99] SACKS M D. Effect of composition and heat treatment conditions on the tensile strength and creep resistance of sic-based fibers[J]. *Journal of the European Ceramic Society*, 1999, 19(13): 2305-2315.
- [100] BI P, ZHANG S, REN J, et al. A high-performance nonfused wide-bandgap acceptor for versatile photovoltaic applications[J]. *Advanced Materials*, 2022, 34(5): 2108090.
- [101] NOWACKI K, KASPRZYK W. The sound velocity in an alloy steel at high-temperature conditions[J]. *International Journal of Thermophysics*, 2010, 31: 103-112.
- [102] TANIFUJI M, MATSUDA A, YOSHIKAWA H. Materials data platform-a fair system for data-driven materials science[C]//2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI). Piscataway: IEEE, 2019: 1021-1022.

作者在攻读硕士学位期间发表的论文与研究成果

学术论文发表情况

1. Rui Zhang, Jiawang Zhang, Yuexing Han, Qiaochuan Chen. Material Property Prediction by Integrating Literature-mining and Crossed Feature Selection. 2023 4th International Conference on Computer Information and Big Data Applications Engineering. EI 会议, 导师一作, 作者二作, 已录用。

专利授权情况

1. 专利名称: 基于图像处理的文献表格内容识别与信息提取方法, 发明人: 韩越兴、张家旺、张瑞、陈侨川、钱权、夏锦桦、王迎港, 专利号: ZL202110185627.9, 授权日: 2022 年 08 月 09 日, 授权公告号: CN112861736B。

软著授权情况

1. 软件名称: 表格文字识别与复原软件 V1.0, 开发人: 韩越兴、张家旺、张瑞, 登记号: 2021SR0492854, 申请人: 上海大学, 开发完成日期: 2020 年 12 月 10 日, 登记日期: 2021 年 04 月 02 日。
2. 软件名称: PDF 科学文献图表公式检测软件 V1.0, 开发人: 张瑞、张家旺、韩越兴, 登记号: 2022SR1446824, 申请人: 上海大学, 开发完成日期: 2022 年 05 月 10 日, 登记日期: 2022 年 11 月 01 日。

作者在攻读硕士学位期间所作的项目

1. 项目来源：国家重点研发计划项目
项目名称：材料基因组工程专用数据库平台建设与示范应用
项目编号：2018YFB0704400
执行期限：2018.07-2022.06
2. 项目来源：国家重点研发计划项目
项目名称：材料基因工程关键技术与支撑平台
项目编号：2020YFB0704500
执行期限：2020.09-2022.08
3. 项目来源：上海市自然科学基金项目
项目名称：小样本环境下物体自适应识别方法研究
项目编号：20ZR1419000
执行期限：2020.07-2023.06
4. 项目来源：之江实验室科研攻关项目
项目名称：智能计算材料平台建设与示范应用
项目编号：2021PE0AC02
执行期限：2021.11-2024.11

致 谢

光阴似箭，岁月如梭，在上海大学三年的研究生求学生活即将画上句号。回顾这段不断进取的时光，我所取得的每一个进步和成果，都凝聚着老师的指导、同学朋友的支持以及家人的鼓励，在此向你们表达最真挚的感激之情。

首先，感谢我的导师张瑞老师，老师对待学术严谨治学、精益求精，对待工作踏实细致、言传身教，是我永远学习的榜样。在张老师的悉心指导下，我能够逐个击破所遇到的难题，从论文选题、方法构思、实验设计、数据处理到论文撰写与修改，每个阶段都离不开张老师细致入微的帮助。值此论文完成之际，我向张老师表达诚挚的谢意，云深雾茫，师恩难忘！

同时，感谢课题组韩越兴老师和陈侨川老师对我研究的细致指导和支持。在每一次的沟通和交流中，不断地提高我的学术水平，启发我对问题有更深层的思考和分析；难忘在研究工作中，提供解决方案和指导，为我论文的完成提供了关键性的支持。再次感谢所有给予我支持和帮助的老师，祝愿各位老师家庭幸福美满，学术蓬勃发展，桃李满天下！

此外，感谢实验室和课题组同学的帮助和陪伴。讨论往往能消除难题，协作常常能激发灵感，在科研道路上，我们一起互帮互助，勇往直前，共同度过了三年的难忘时光。感谢我的家人和朋友，他们一直是我的坚实后盾，给予我无尽的支持和鼓励。

最后，衷心感谢百忙之中参与论文评审和答辩的各位专家教授，感谢您们的宝贵建议和专业指正。