中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：22721609

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	基于语义感知的材料文献挖掘方法研究

作　　者　张一琳
学科专业　软件工程
导　　师　张瑞
完成日期　二〇二五年五月

姓　　名：张一琳　　学号：22721609

论文题目：基于语义感知的材料文献挖掘方法研究

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：2015年 6月 10日

姓　　名：张一琳　　学号：22721609

论文题目：基于语义感知的材料文献挖掘方法研究

上海大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，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研究者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。

学位论文作者签名: 张一琳

日期:　2025 年　6 月　10 日

上海大学学位论文使用授权说明

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)

学位论文作者签名: 张一琳　　导师签名: 张瑞

日期:　2025 年　6 月　10 日　　日期:　2025 年　6 月　10 日

上海大学工程硕士学位论文

基于语义感知的材料文献挖掘方法研究

作　　者: 张一琳
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
2025 年 5 月

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

Research on Material Literature Mining Methods Based on Semantic Awareness

Candidate: Yilin Zhang
Supervisor: Rui Zhang
Major: Software Engineering

School of Computer Engineering and Science
Shanghai University
May, 2025

摘　要

随着材料科学研究的持续深入，海量文献中蕴含的材料实体及其性能信息成为推动材料研发的重要知识源。然而，由于材料术语表述形式多样、标准不统一，相关信息通常以非结构化文本形式存在，难以获取和利用，同时庞大的文献数量也增加了信息检索的复杂性。如何从材料文献中高效、准确地抽取关键实体信息，已成为当前智能材料研究中的核心问题。针对这一挑战，本文围绕材料文献挖掘展开研究，提出多种面向命名实体识别的模型方法，并构建了性能预测系统。主要工作如下：

（1）针对材料文献中存在的长序列依赖、实体关系复杂的问题，提出语义增强图网络模型，应用到复合材料的文献挖掘领域。该模型通过构建异构图强化语义关联建模，并引入分块注意力机制高效处理长序列问题，有效克服传统模型的局限性。在此基础上，利用深度可分离卷积融合全局与局部语义特征，并结合动态边权重机制与深度评分网络提升节点表示与识别精度，更有效地捕捉材料术语在复杂上下文中的语义关系。

（2）针对通用材料文本中实体边界模糊、长实体识别效果不佳的问题，提出多粒度融合图网络模型，应用于材料科学文献领域的命名实体识别任务。该模型设计了融合多粒度语义特征与边界优化策略的新模块：首先，通过门控融合与跨粒度交互注意力，增强不同尺度语义特征的表示能力；其次，结合条件随机场与对比学习进行联合训练，利用它们各自的优势，协同提升边界识别的准确性和长实体识别的性能。

（3）将所提出的文献挖掘方法应用于碳纤维复合材料性能预测与应用设计。通过挖掘并筛选材料实验文献，提取出九类与力学性能密切相关的关键特征，通过实验验证了文献挖掘成果在性能建模中的应用潜力。此外，设计并实现了一个材料性能预测系统，支持用户上传数据文件并完成模型选择、训练与结果可视化，为材料研究人员提供了一个高效、易用的性能预测工具。

本文的工作通过多项实验验证了其有效性，在提升材料文献挖掘准确性的基础上，为材料信息学的发展提供了理论支撑与实践参考。

关键词：文献挖掘；命名实体识别；异构图网络；对比学习；材料性能预测

ABSTRACT

With the continuous deepening of materials science research, the information on material entities and their properties contained in the vast amount of literature has become an important source of knowledge to promote materials research and development. However, due to the diverse forms of material terminology and the lack of standardized expressions, relevant information is often embedded in unstructured texts, making it difficult to access and utilize directly. Furthermore, the huge amount of literature further complicates the process of information retrieval. How to efficiently and accurately extract key entity information from materials literature has become a core issue in current smart materials research. Aiming at this challenge, this paper focuses on materials literature mining, proposes a variety of modeling methods for named entity recognition and performance prediction, and constructs a performance prediction system. The main contributions are as follows:

(1) To address the challenges of long-range dependencies and complex entity relationships in material science literatures, this paper proposes a Semantic-Rich Graph Network (SRGN) for named entity recognition in composite material literatures. The model strengthens semantic association modeling by constructing a heterogeneous graph and introduces a chunked attention mechanism to efficiently deal with long sequences, effectively overcoming the limitations of traditional models. Building upon this foundation, the model integrates the depthwise separable convolution to fuse global and local semantic features. Coupled with a dynamic edge weighting mechanism and a deep scoring network, it enhances node representation and recognition accuracy, enabling more effective capture of the semantic relationships of material science terminology within complex contexts.

(2) To address the challenges of ambiguous entity boundaries and poor recognition performance for long entities in general material literatures, this paper proposes a Heterogeneous Cross-grained Graph Network (HCG), a multi-granularity fusion model designed for named entity recognition tasks in the domain of material science literatures. The model is designed with a new module that fuses multi-granularity semantic features with boundary op-

timization strategies. Firstly, it employs gated fusion and cross-granularity interactive attention to enhance the representation of semantic features across different scales. Subsequently, HCG combines conditional random fields (CRF) and contrastive learning in joint training, leveraging their complementary strengths to synergistically improve boundary recognition accuracy and long-entity identification performance.

(3) The proposed literature mining method is applied to performance prediction of carbon fiber reinforced polymers. By mining and screening 380 material literatures between 2019 and 2022, nine types of key features that are closely related to mechanical properties are extracted, and the mechanical properties are further modeled and predicted using machine learning methods, which validates the potential of applying the literature mining results in performance modeling. Furthermore, a machine learning-based material performance prediction system is designed and implemented, supporting users in uploading data files and completing model selection, training, and result visualization. This system provides material researchers with an efficient tool for performance prediction.

The methods proposed in this paper have been validated through extensive experiments. Building upon the improvements in the accuracy of material literature mining, they further provide both theoretical foundations and practical references for the advancement of material informatics.

Keywords: Literature Mining; Named Entity Recognition; Heterogeneous Graph Network; Contrastive Learning; Material Performance Prediction

目 录

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：2022YFB3707800）、国家自然科学基金（编号：52273228）、云南省科技攻关计划（编号：202302AB080022）、教育部硅酸盐文物保护重点实验室（上海大学）项目（编号：SCRC2023ZZ07TS）、上海市科技青年人才扬帆计划（编号：23YF1412900）的资助。

1.2　课题背景概述

材料科学作为现代工业技术的基石，直接影响着航空航天、能源装备、交通运输等关键领域的科技发展与进步[1]。工程材料在服役过程中需要承受复杂载荷与环境作用，其力学性能（如强度、韧性、疲劳特性）与物理特性（如导电性、热稳定性）的精准表征与优化设计始终是领域核心课题之一。例如，弯曲强度是评估复合材料力学性能的关键指标之一，它直接影响材料在复杂载荷条件下的承载能力、抗变形性能以及整体结构的稳定性[2]。这对材料成分设计与制备工艺提出了严苛要求。

传统的材料开发严重依赖实验方法探索材料性能[3]，通过进行大量的实验验证和数据测试来确定适用于新技术的材料[4]。为了研究高性能的材料需要进行数千次实验，每次实验涉及各种参数的调整和测试，例如成分比例调整、工艺参数优化等，需要消耗大量时间和经济成本。以碳纤维复合材料为例，它的性能不仅与作为填料的碳纤维以及基体类型相关[5]，还与材料组成、制备工艺参数、样品厚度和测试方法等有关[6]，仅考虑材料组成的调整就可能需要进行数百组的实验，若涉及其他参数或者表面改性等方法，实验规模可能呈指数级增长，难以满足现代工业对材料快速迭代的需求。

随着人工智能技术的突破性发展，材料研究范式正在经历从传统实验和理论方法向数据驱动方法的转变。机器学习技术通过挖掘材料大数据中的隐藏规律，为性能预测和成分设计提供了新途径[7]。然而，实验数据往往难以获取，机器学习模型需要大量数据，且文献中通常包含着关键的信息，例如成分、工艺、测试方法等，因此从科学文献中提取有价值的数据成为深入探索和优化材料性能的关键途径之一。

文献挖掘技术为突破这一瓶颈提供了关键解决方案，旨在自动化提取文献中包含的材料成分、工艺参数、性能指标等结构化信息。它能够帮助材料基因工程发现和预测性能与成分、工艺之间的关系，降低材料开发的成本，为新材料的研发与设计提供数据支撑。因此，本文聚焦于材料领域文献的文本特性，利用自然语言处理技术从非结构化文本中挖掘材料成分、工艺、性能等关键信息，为材料科学的研究提供数据支持。

1.3 课题研究目的和意义

材料科学文献作为该领域知识的核心载体，包含了大量关于成分设计、合成工艺及性能数据等信息，这些信息对于设计和研究新材料而言是至关重要的。目前已有统计，材料科学领域文献发表数量呈指数级增长[8]，从海量文献中自动化提取关键信息已成为知识重用的主要难题。文献挖掘技术结合自然语言处理与机器学习方法，能够从非结构化文本中识别材料实体、解析属性关系，并构建结构化知识库，从而为材料科学的研究和应用提供有效的支持。

目前，许多研究者已经通过自然语言处理技术，从材料科学文献中进行信息挖掘[9-11]。而在合金、不锈钢等材料领域，也有研究者提出了专用的文献挖掘方法，用于从文献中提取成分和性能数据，进行材料性能预测[12]。尽管现有的文献挖掘方法已取得一定进展，但针对材料科学的文献挖掘方法仍存在局限性，特别是在处理复杂信息和多样化数据时。首先，当前文献挖掘研究主要依赖自然语言处理技术，聚焦于关系抽取、命名实体识别等任务。然而，这些任务的方法可能缺乏对特定领域知识的融合，导致在处理材料科学文献时，难以准确识别和解析涉及专业知识的实体类别。其次，许多术语具有强烈的上下文依赖性，这可能会导致传统的命名实体识别、关系抽取等方法在文献中挖掘关键信息时出错。例如，“屈服”一词在材料文献中通常指屈服强度，而在其他工程领域则可能与土木结构的承载力屈服相关。由于缺乏领域知识，传统方法在应对这些语境依赖性较强的术语时可能出现理解偏差。因此，现有文献挖掘方法在材料领域的应用仍存在局限，仍需开发更为专门的技术来弥补这些不足，以提高文献挖掘的准确性和实用性。

现有的文献挖掘工作通常将挖掘得到的结果与机器学习结合，进行性能预测，分析从文献中挖掘的数据，揭示材料性能与成分之间的复杂相互作用，成为加速材料

性能预测和结构设计的重要手段[13]。目前在合金、复合材料等领域，有研究者利用机器学习，对材料进行性能预测，并研究了成分、工艺等参数对性能的影响[14-15]。然而，现有的性能预测工作大多依赖手动收集数据，这通常需要大量的时间和人力成本，且数据的来源有限。此外，手动整理的数据集通常更新周期较长，难以及时整合领域内最新研究成果。因此，通过文献挖掘自动化收集与材料性能相关的数据，并对这些数据进行适当的处理和特征选择，可以为材料性能预测模型提供更为准确、全面的数据支持。

综上所述，针对目前文献挖掘与性能预测工作存在的不足与挑战，本文采用自然语言处理技术，结合异构图、注意力机制以及对比学习等方法，有效提升了文献挖掘的准确率，为材料性能预测和新材料的发现提供支持。

1.4 国内外研究现状

随着材料科学与工程领域快速发展，传统基于实验和理论计算的研究模式越来越难以满足高效、快速的研发需求。近年来，以大数据、自然语言处理和机器学习技术为核心的数据驱动方法逐渐兴起，其中从海量的科学文献中自动抽取关键材料信息以支撑材料设计与开发成为研究热点[16]。研究者们围绕如何从海量材料文献中自动提取有价值的信息展开了大量工作，目前主要的方法可以归纳为以下两类：基于深度学习的方法，以及基于大语言模型的方法。

1.4.1 基于深度学习的方法

早期的材料文献挖掘主要依赖人工设计的规则和特征工程方法。例如，利用正则表达式匹配和专业词典来识别材料名称、化学式等实体，配合预定义的模板提取实验条件或性能数据[17]。这类基于规则和字典的方法针对特定格式或术语具有较高的精度，但需要投入大量人力，难以覆盖文献中措辞多样的表述。

随着深度学习方法在自然语言处理中的崛起，研究者开始将其引入材料文献挖掘，以突破传统方法在人工规则构建和特征工程方面的局限性。深度学习模型能自动从数据中学习特征表示，因而在材料文本中的命名实体识别、关系抽取等任务上展现了优势[18]。例如，Shetty 等人[9]采用 Word2Vec 词向量模型，从 50 万篇聚合物相关论文中提取信息，并利用文献中的领域知识预测聚合物的潜在应用，为材料发现开辟了新的范式；Weston 等人[10]使用文本挖掘技术，从大约 327 万篇材料科学领

域的文献摘要中，成功挖掘了包括材料、性质、表征方法、相位描述符、合成方法及其应用等信息。相比基于规则的传统方法，这些神经网络模型可以更好地捕捉上下文语义，从而提高文献挖掘的准确率。

随着 Transformer 架构和预训练语言模型的出现，材料文献挖掘进入了新阶段。预训练模型在海量通用语料上学习语言表示，然后通过微调适应特定任务，大幅提升了信息抽取的性能。这一范式在材料领域同样取得成功，例如，Gupta 等人[11]提出了一个专门针对材料科学领域的语言模型——MatSciBERT，用于进行科学文本的信息提取，该模型在三个下游任务，即命名实体识别、关系分类和摘要分类上表现出色，显示了其有效性；Wang 等人[19]开发了一个自然语言处理管道，用于捕获和分析化学成分及其性能数据，通过分析 14425 篇文献中的 2531 条数据，预测了未被勘探的 Co 基高温合金，同时还提供了一个开源在线平台，旨在为从文献数据中搜索目标材料提供通用方法；Shetty 等人[20]使用了 240 万份材料科学文献摘要来训练 MaterialsBERT，主要应用于燃料电池、超级电容器和聚合物太阳能电池等电池材料的分析，这一研究验证了从已发表文献到自动提取材料属性信息的全过程的可行性。此外，Zhang 等人[21]提出了一种基于卷积神经网络和 MatBERT 的方法，对 2389 篇钙钛矿材料文献的摘要进行实体提取，获取钙钛矿材料相关知识，分析钙钛矿领域的发展趋势。

尽管基于深度学习的材料文献挖掘研究取得了显著成果，但如何有效解决材料领域特殊术语、实体边界不明晰以及长序列实体识别问题，仍是当前亟待解决的重要挑战，这也为本文后续研究提供了明确的研究方向和基础。

1.4.2 基于大语言模型的方法

随着 GPT-3、GPT-4 等超大规模预训练语言模型的出现，材料文献挖掘进入了一个新的探索阶段。大语言模型（LLM）具有强大的通用语言理解和生成能力，能够在零样本或少样本学习范式下执行复杂的文本推理任务，这为材料领域的文献分析提供了新的思路。目前的研究一方面在评估直接利用大语言模型抽取材料信息的效果，另一方面在探索如何结合提示工程或轻量微调来充分发挥 LLM 的潜力。

Foppiano 等人[22]对比了 GPT-3.5、GPT-4 等 LLM 在材料信息抽取任务中的表现，与传统的 BERT 模型和基于规则的方法进行了对比。实验结果表明，在不进行任何微调的零样本设置下，GPT 系模型并未能超越既有的领域模型基线，尤其在材料实

体命名识别任务上表现平平，这表明超大模型在细粒度提取专业领域实体时仍存在挑战。也就是说，虽然 LLM 具备优秀的通用语言理解和推理能力，但面对材料科学这类专业术语密集、知识体系复杂的领域，若要充分发挥作用，仍需要结合一定的调优策略或配合领域数据。

为进一步提升大模型在材料文献挖掘中的效果，近期一些工作开始尝试对 LLM 进行专门的微调或采用对话式提示工程策略。例如，Dagdelen 等人[23]提出了一种联合命名实体识别和关系抽取的方法，采用 GPT-3 和 Llama-2 等大模型，通过在输出中设计 JSON 格式，使模型从材料化学文献中同时抽取实体及其关系。这种方法避免了传统管道需要先后执行 NER 和关系链接的繁琐步骤，证明了大模型可以通过适当设计输出格式来“一步到位”地完成结构化信息抽取。

近年来，国内学者已逐步开展大语言模型在材料文献挖掘领域的应用探索。例如，时宗彬等人[24]采用本地部署的中文预训练大模型并结合精心设计的提示模板，实现了对有机光伏电池材料文献中材料体系信息的抽取。他们没有对模型进行微调，而是通过在提示模板中添加少量示例并允许大语言模型给出否定回答的方式，识别相应的实例信息。这一研究表明，通过科学设计提示工程方案，大语言模型能够有效实现专业文献信息的精准提取。

尽管 LLM 在材料文献挖掘领域展现出显著的应用潜力，但其在实际应用过程中仍面临一些挑战。首先，LLM 通常是在通用领域的语料库上进行训练的，对材料领域内的专业术语、复杂的化学表达式以及特殊的实验数据表述缺乏深入理解，这导致模型在处理材料科学中特有的语言模式和术语时表现不足，容易出现专业术语识别错误或术语混淆问题。其次，材料科学文献往往包含大量结构化或半结构化的数据，例如实验条件、材料表征数据等。大语言模型主要处理的是非结构化的自然语言文本，无法直接有效地处理文献中经常出现的表格、图表等结构化数据，而这些结构化数据对于材料研究却至关重要。针对上述挑战，仍需研究者在该领域开展大量实验探索。

1.5 论文主要工作

为了解决目前在材料文献挖掘中存在的挑战，从文献文本中有效挖掘出关键信息，本文针对材料文献的特点，结合自然语言处理技术，实现了对材料领域文献内

容的挖掘，并将挖掘的结果应用于性能预测，主要的研究内容和工作如下：

（1）针对复合材料文献中存在的长序列依赖、实体关系复杂的问题，提出一种基于异构图与分块感知的命名实体识别模型——语义增强异构图网络（SRGN）。该模型引入了分块注意力机制，将长序列划分为多个块，在每个块内计算注意力时仅关注局部上下文，从而降低计算复杂度并保持局部敏感度。在处理异构图时，采用深度可分离卷积融合全局-局部特征，增强节点更新的语义信息。此外，采用可学习的动态边权重机制自适应调整节点间的连接权重。为了增强网络的非线性能力，引入了深度评分网络，用于计算预测概率。通过在复合材料文献数据集和材料公共数据集上进行实验，证明了 SRGN 在材料文献挖掘任务中的适用性。

（2）针对通用材料文献中文本实体边界模糊、长实体识别效果不佳的问题，提出一种基于多粒度融合的材料命名实体识别模型——多粒度融合图网络（HCG）。材料文献文本中的实体通常较长，针对这个问题，该模型引入了门控融合机制和跨粒度注意力，增强该方法对信息的表征能力。另外，为了优化实体边界识别，该模型将 CRF 损失与对比学习损失进行联合训练，增强了对实体边界的预测和识别能力。实验表明，HCG 在多个数据集下均表现较好，验证了该模型的有效性。

（3）本文将提出的文献挖掘方法应用于碳纤维复合材料的性能预测。通过挖掘 2019 至 2022 年间的 380 篇碳纤维复合材料实验文献，结合专家知识对挖掘结果进行筛选与分类，提取出九种与材料力学性能相关的关键特征，利用机器学习方法对碳纤维复合材料的弯曲强度与拉伸强度进行了性能预测。此外，本文实现了一个基于机器学习方法的材料性能预测系统，旨在为用户提供一个高效、自动化的工具，进行材料性能预测。该系统支持用户上传材料数据文件，并通过机器学习模型对数据进行处理与分析，从而自动预测材料在不同条件下的性能表现，为材料研究和开发提供科学支持。

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，针对材料科学文献研究文献挖掘方法，实现了从材料文献进行关键信息提取的功能，并通过实验验证了提出方法的有效性。此外，将提出的文献挖掘方法应用于复合材料文献，对弯曲强度与拉伸强度进行了性能预测，并设计了一个性能预测系统。本文总共由六个章节组成，

组织结构如图1.1所示，具体结构如下：

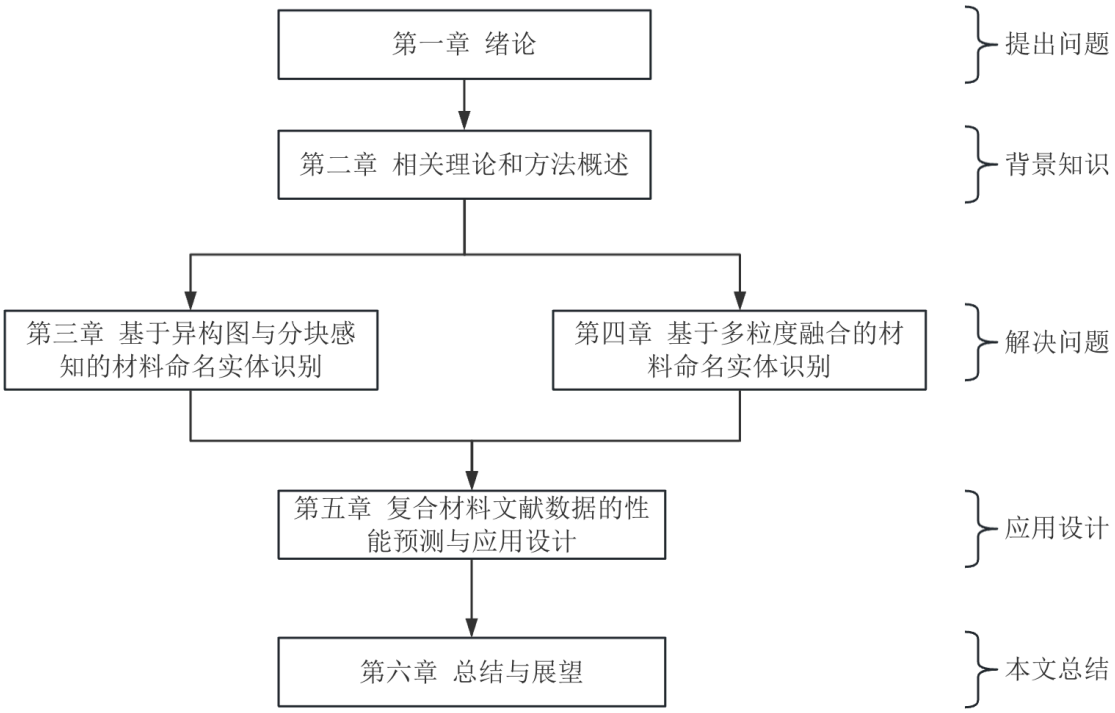


图 1.1　论文组织结构图

第一章为绪论，介绍了课题来源、课题研究目的和意义，分析了材料文献挖掘与性能预测的研究现状及挑战，并提阐述了本文的主要工作与论文组织结构。

第二章为相关理论，详细介绍了与材料文献挖掘相关的理论基础和技术，包括命名实体识别、异构图、注意力机制、对比学习以及深度可分离卷积，为后续章节的算法设计与实现提供了理论支撑。

第三章提出了一种基于异构图与分块感知的材料命名实体识别模型。该模型针对材料文献文本，引入分块注意力机制以降低计算复杂度并保持局部敏感度。在异构图结构中，使用深度可分离卷积融合全局和局部特征，采用动态边权重机制调整节点间连接权重。此外，解码器中引入深度评分网络以增强非线性能力。通过在复合材料和材料公共数据集上的实验，验证了该模型在材料文献挖掘任务中的有效性。

第四章提出了一种基于多粒度融合的材料命名实体识别模型。该模型引入门控融合机制和跨粒度注意力，增强模型对信息的表征能力；其次，将 CRF 损失与对比学习损失进行联合训练，优化对实体边界的识别。通过在多个数据集上的实验，证明了该模型的有效性。

第五章为复合材料文献数据的性能预测与应用设计，通过对复合材料文献的挖

掘，结合机器学习方法进行材料性能的预测与应用设计。该章介绍了如何利用文献中提取的数据进行性能预测，并展示了该方法在复合材料领域的实际应用。

第六章为总结与展望，总结了论文的主要研究成果，讨论了当前研究的不足和未来的研究方向。

第二章　相关理论和方法概述

近年来，随着信息技术的飞速发展，材料科学领域的信息挖掘研究逐渐引起了学术界和工业界的广泛关注。材料文献中蕴含着大量的实验数据、理论分析和技术应用信息，这些信息对于推动新材料的发现和创新具有重要的意义。然而，如何有效地从海量的材料文献中提取有价值的知识和规律，依然是一个挑战。本文的研究目标是针对材料文献中的信息进行深度挖掘，解决数据冗余、信息提取不准确以及文本理解能力有限等问题，从而提升材料研究的效率与准确性。本章节将着重介绍与本文研究密切相关的技术与应用，为后续的研究工作奠定理论基础。

2.1　文献挖掘技术

文献挖掘技术是自然语言处理和信息检索领域的一个重要分支，它涉及从大量文献中自动提取、分析和挖掘有价值的信息。在材料科学领域，随着研究出版物的数量迅速增长，文献挖掘已成为发现新材料、理解材料性质的关键工具。有效的文献挖掘不仅可以帮助科研人员节省查找和分析文献的时间，还能揭示潜在的研究趋势和未被充分探索的科学问题。本节将重点介绍文献挖掘中的三项关键技术：命名实体识别、预训练模型和条件随机场。

2.1.1　命名实体识别

命名实体识别（Named Entity recognition, NER）是自然语言处理领域的研究热点，其核心任务是通过对文本序列的标注，定位并分类具有特定意义的实体。早期研究主要面向通用领域，如新闻语料中的人名、地名和机构名识别。随着应用场景的扩展，NER 的定义逐渐泛化，涵盖生物医学领域的基因蛋白质命名、材料科学中的化学式与工艺参数提取等细分需求。对于命名实体识别方法，其发展如图2.1所示。

传统 NER 方法可分为基于规则与词典的方法、机器学习方法两类。前者依赖领域专家手工编写正则表达式或构建术语库，例如材料领域通过正则匹配化学式或材料名称。后者则以隐马尔可夫模型、支持向量机（SVM）和条件随机场（CRF）为代表，通过特征工程结合概率模型实现序列标注。

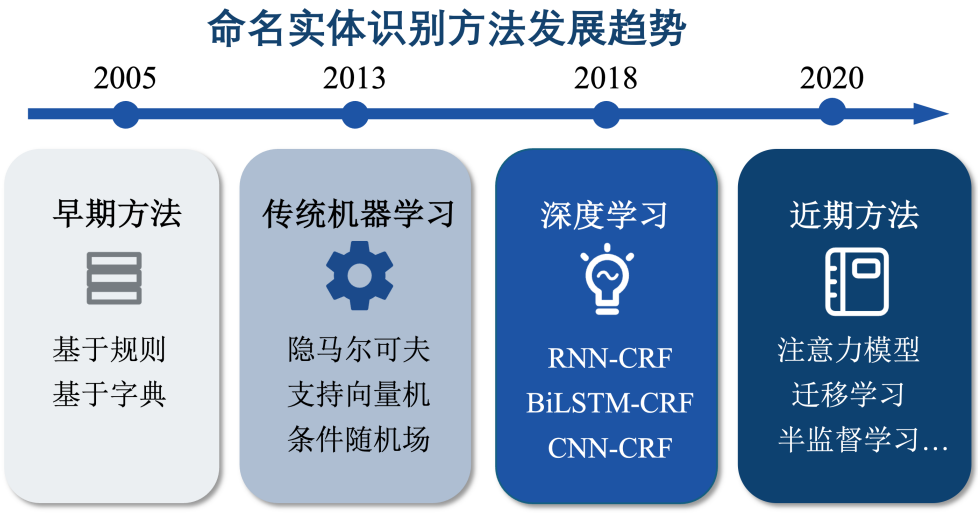


图 2.1 命名实体识别方法发展趋势

深度学习的兴起推动了 NER 技术的范式变革。基于循环神经网络 RNN 的模型，例如 LSTM、BiLSTM 等，通过捕捉上下文语义，显著提升了识别实体的能力。Huang 等人[25]提出的 BiLSTM-CRF 架构将双向长短期记忆网络与 CRF 结合，既利用神经网络提取深层特征，又通过 CRF 约束标签转移概率，成为经典范式。此后，卷积神经网络也被引入，形成 CNN-CRF 模型进行命名实体识别，进一步优化了局部特征提取能力。

目前，NER 技术正朝着高效化、低资源依赖和细粒度理解的方向演进。随着 Transformer 架构的普及，BERT[26]、RoBERTa[27]等基于自注意力机制的预训练模型已经成为主流。预训练模型通过捕捉长距离的上下文依赖关系，显著提升了复杂实体边界判定的准确性，这使得 NER 在处理复杂文本、尤其是长实体时表现尤为突出。以动态稀疏注意力机制[28]为例，它通过动态剪枝无关词元的注意力权重，既降低了计算成本，又保持了实体识别的精度。这类方法尤其适用于处理复杂的、具有多层级结构的实体，例如在材料科学领域中常见的复合材料命名，如“CoFeB/Ta/CoFeB 磁性多层膜”。

尽管通用的 NER 方法取得了显著进展，其在材料科学文献中的应用仍面临诸多挑战，主要包括实体结构复杂、术语边界模糊、长实体识别困难以及领域标注数据稀缺等问题。为缓解小样本学习的局限，元学习与提示学习被广泛应用于材料领域的 NER 任务，通过构建元知识库，使模型能够在仅有少量标注样本的情况下，快速适应新的实体类别，实现跨材料子领域的实体识别。然而，这类方法通常依赖良好

的预训练初始化，且在面对材料文本中的长距离依赖与实体交互建模时，仍存在表达能力受限的情况。

另一方面，半监督与自监督学习策略通过引入自训练机制和一致性正则化，有效利用大量未标注材料文献以提升模型泛化能力。具体做法包括基于高置信度伪标签进行迭代优化，或在原始文本中引入随机扰动以增强模型的稳健性。这些方法在缓解标注数据稀缺问题方面取得了一定进展，但在应对材料文献中高复杂度术语结构、实体间语义关系丰富等特性时，仍表现出识别边界不准、实体类别混淆等不足，需进一步探索更具针对性的建模策略，以提升材料科学文献中实体识别的性能。

2.1.2 预训练模型

近年来，预训练模型在自然语言处理领域取得了革命性的进展，尤其是在 NER、情感分析、机器翻译等各种下游任务中，预训练模型已成为主流技术。这些模型通过在大规模无标签文本数据上进行预训练，能够学习到丰富的语言表示，并通过微调适应不同的任务。预训练模型的成功在于其能够捕捉长距离的上下文关系，并且具备较强的迁移学习能力，极大地提升了在各个文本任务中的性能。

在预训练模型的研究中，BERT 模型[26]是最具代表性和突破性的模型之一，它由谷歌团队于 2018 年提出，并在多个文本任务中设置了新的性能标准。与传统的基于 RNN 或 LSTM 的模型不同，BERT 基于 Transformer[29]架构，通过自注意力机制实现全局上下文交互，能够同时考虑上下文的双向信息，从而显著提升了模型对上下文关系的理解能力。其输入表示由三部分组成：词嵌入（Token Embeddings）、位置嵌入（Position Embeddings）和段嵌入（Segment Embeddings），三者相加后形成输入向量。对于输入序列中的每个词 w_i，其嵌入表示为：

$$h_i^0 = E_{token}(w_i) + E_{pos}(i) + E_{seg}(s_i), \tag{2.1}$$

其中，E_{token} 为词表映射矩阵，E_{pos} 编码词的位置信息，E_{seg} 区分句子归属（对单句任务可忽略）。随后，通过 L 层 Transformer 块进行特征变换，每层 Transformer 包含多头自注意力（Multi-Head Attention）和前馈神经网络（FFN），其中对于多头注意力的介绍，见2.3.2小节的公式 (2.11)。对于特征变换，公式如下：

$$h_i^l = \text{TransformerBlock}(h_i^{l-1}), \tag{2.2}$$

BERT 的预训练任务包含掩码语言建模（Masked Language Modeling, MLM）和下一句预测（Next Sentence Prediction, NSP）。MLM 通过随机遮盖 15% 的输入词，要求模型基于上下文预测被遮盖词。损失函数 $\mathcal{L}_{MLM}$ 通过交叉熵优化，迫使模型学习词汇在双向上下文中的分布式表征，特别擅长捕捉多义词的语境相关语义，公式如下：

$$\mathcal{L}_{MLM} = -\sum_{i \in M} \log P(w_i | w_{\setminus \mathcal{M}}), \tag{2.3}$$

其中，$\mathcal{M}$ 表示被遮盖词的索引集合，$w_{\setminus \mathcal{M}}$ 表示未被遮盖的上下文词序列。NSP 则判断两个句子 S_A 和 S_B 是否连续，损失函数为二分类交叉熵：

$$\mathcal{L}_{NSP} = -\sum \log P(y | S_A, S_B), \tag{2.4}$$

总预训练目标为两者的加权和：

$$\mathcal{L}_{BERT} = \mathcal{L}_{MLM} + \lambda \mathcal{L}_{NSP}, \tag{2.5}$$

BERT 模型的最大优势在于其双向上下文建模能力。在传统的单向语言模型中，模型只能依赖单一方向的上下文信息进行预测，而 BERT 则能够同时考虑前后文信息，这使得它在理解文本的细节和语义方面具有显著优势。BERT 的预训练-微调策略也使得它能够在多个文本任务中取得高效的迁移学习效果，不需要大量的标注数据就能获得良好的性能。

近年来，多个研究者针对 BERT 的不足提出多种变体方法进行改进，例如 RoBERTa[27]、ALBERT[30]以及 ELECTRA[31]等。在材料科学领域，领域适配的预训练模型成为趋势。例如，MatSciBERT[11]在数百万篇材料学论文摘要上继续预训练，增强对化学式、单位及工艺、合成方法等与材料相关的语义理解。

2.1.3 条件随机场

条件随机场[32]（Conditional Random Field, CRF）是一种基于概率图模型的判别式序列标注方法，广泛应用于自然语言处理任务，例如 NER、词性标注和语音识别等。与生成式模型如隐马尔可夫模型不同，CRF 直接建模在给定观测序列条件下的标签序列概率，能够有效捕捉标签间的依赖关系，从而提升在上下文建模方面的表

现。特别是在 NER 任务中，CRF 能有效捕捉实体边界和类别之间的相互关系，从而提升了模型在处理长实体和复杂结构文本时的准确性。

CRF 模型的基本思想是通过引入一个全局的条件概率分布来联合建模序列中各个元素的标签。假设输入序列 $x = x_1, x_2, ..., x_n$ 和对应的标签序列 $y = y_1, y_2, ..., y_n$，CRF 的目标是根据给定的输入序列 x，预测标签序列 y 的条件概率。CRF 模型通过特征函数来表示输入和标签之间的关系，并通过学习这些特征函数的权重来最大化条件概率。公式上，给定输入序列 x，标签序列 y 的条件概率表示为：

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n}\sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, i)\right), \tag{2.6}$$

其中，$f_k(y_{i-1}, y_i, x, i)$ 是特征函数，表示输入序列 x 的第 i 个元素及其标签与前一个标签之间的特征，λ_k 是与特征相关的权重参数，$Z(x)$ 是规范化因子，确保所有标签序列的概率和为 1。特征函数通常通过训练数据学习得到，它反映了标签之间的依赖关系以及与输入序列中某些特征的相关性。

CRF 模型的计算过程依赖于特征函数和规范化因子的计算，后者确保了模型输出的概率是归一化的。为了计算条件概率 $P(y\ x)$，需要计算规范化因子 $Z(x)$，其定义为：

$$Z(x) = \sum_{y'} \exp\left(\sum_{i=1}^{n}\sum_{k} \lambda_k f_k(y'_{i-1}, y'_i, x, i)\right), \tag{2.7}$$

该规范化因子涉及对所有可能标签序列的求和，因此在实际应用中，计算复杂度较高。为了提高计算效率，通常使用动态规划算法，如前向后向算法，来进行优化。

CRF 的优势在于其显式建模标签之间的依赖关系，并通过全局最优解实现整序列的标注预测，这使其在捕捉上下文信息时优于局部优化的传统模型。此外，CRF 具备高度灵活的特征函数设计能力，能够融合词性、上下文线索、实体边界等多种信息，有效增强模型表达能力。然而，CRF 也存在一定局限性：一方面是计算开销较大，特别是在计算规范化因子时，需要对所有可能标签序列进行求和，计算复杂度高。虽然可以通过动态规划算法进行优化，但在处理长文本或大规模数据集时，CRF 仍然存在一定的计算瓶颈。另一方面，CRF 模型对特征的依赖性较强，特征选择和设计需要花费大量的时间和精力，且对于标注数据的需求较大，尤其是在标注数据稀缺的情况下性能可能会受到限制。

2.2 图神经网络

图神经网络（Graph Neural Networks, GNN）是近年来在深度学习领域迅速发展的技术，广泛应用于处理图结构数据。图是由节点（vertices）和边（edges）组成的数学结构，常用于表示实体及其之间的关系。在许多实际问题中，数据不仅仅是以序列或网格的形式存在，还常常以图的形式组织，例如社交网络中的人际关系、分子化学结构、推荐系统中的物品之间的关系等。GNN 的核心任务是通过学习图中节点之间的依赖关系，获取图的嵌入表示，从而解决分类、回归等任务。

图神经网络的一个显著优势是能够通过消息传递机制有效地捕捉节点间的复杂关系，尤其在处理具有结构信息的数据时，展现出了优异的性能。根据图的结构类型，GNN 可以分为同构图和异构图网络两种类型。不同类型的图有不同的建模和处理方法，针对不同的任务，选择合适的图神经网络类型和结构至关重要。图神经网络可以分为同构图与异构图，如图2.2所示，后续将详细介绍它们的不同特点。

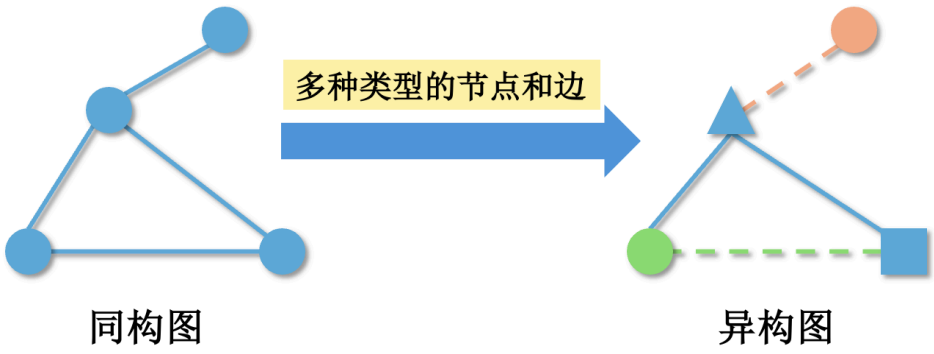


图 2.2　同构图与异构图

2.2.1 同构图

同构图是指图中所有的节点和边都是属于相同类型的图结构。换句话说，在同构图中，所有的节点代表相同的实体类型，所有的边也表示相同类型的关系或连接。最常见的同构图包括社交网络中的人与人之间的关系图，或物品与物品之间的相似性图等。在这些图中，节点和边的类型是统一的，因此可以通过相同的方式进行表示和处理。

在同构图中，图神经网络通过节点之间的信息传递来学习节点的表示。在一个标准的 GNN 框架中，每个节点通过其邻居节点的信息来更新自身的表示。假设存在

一个同构图，节点集合为 $V = \{v_1, v_2, \ldots, v_n\}$，边集合为 $E = \{e_1, e_2, \ldots, e_m\}$，节点的初始表示为 $h_v^{(0)}$。对于每个节点 $v \in V$，图神经网络通过迭代地更新节点的表示来捕捉图的结构信息。每一轮迭代中，节点 v 的新表示 $h_v^{(k+1)}$ 可以通过其邻居节点的表示 $\{h_u^{(k)} \mid u \in \mathcal{N}(v)\}$ 来更新：

$$h_v^{(k+1)} = \text{Update}\left(h_v^{(k)}, \text{Aggregate}\left(\{h_u^{(k)} \mid u \in \mathcal{N}(v)\}\right)\right), \tag{2.8}$$

其中，$\mathcal{N}(v)$ 表示节点 v 的邻居节点集合，Aggregate 是一个聚合操作，通常使用加权求和、平均或最大化等操作来处理邻居节点的信息 Update 是一个非线性映射操作，用于更新节点的表示。在多个迭代后，每个节点会学习到与其邻居关系紧密相关的嵌入表示，进而可以用于分类、回归等任务。

同构图的图神经网络通常具有较为简单的结构，因为所有的节点和边都具有相同的性质。因此，处理同构图时，通常不需要区分不同类型的节点或边，GNN 模型的设计和计算也较为直观和一致。

2.2.2 异构图

与同构图不同，异构图（Heterogeneous Graph）包含多种类型的节点和边。在异构图中，不同类型的节点代表不同的实体类别，边则表示不同类型的关系或交互。异构图的典型例子包括学术论文的引用网络，其中节点包括论文、作者和期刊，边则表示引用、合作等关系。另一例子是电子商务平台中的用户、商品、评论等节点，以及用户购买、评论商品的行为边。

由于异构图包含多种类型的节点和边，因此在处理异构图时，图神经网络需要考虑不同类型的节点和边的特征。为了有效地学习异构图的节点表示，异构图神经网络（Heterogeneous Graph Neural Networks, HGNN）通过设计多种信息聚合策略来分别处理不同类型的节点和边。一种常见的方法是基于消息传递机制，对每种类型的节点和边应用独立的聚合操作，并在每轮迭代时融合不同类型的信息。

在异构图中，节点和边具有不同类型，图神经网络需为每类节点和边设计特定计算规则。给定异构图 G 包含节点类型集合 $\{V_1, ..., V_K\}$ 和边类型集合 $\{E_1, ..., E_L\}$，每个节点 $v \in V_k$ 和边 $e \in E_l$ 使用类型相关的表示。节点更新公式简化为：

$$h_v^{(k+1)} = \text{Update}_k\left(h_v^{(k)}, \sum_{l=1}^{L} \frac{1}{|\mathcal{N}_l(v)|} \sum_{u \in \mathcal{N}_l(v)} W_l h_u^{(k)}\right), \tag{2.9}$$

其中，$\mathcal{N}_l(v)$ 表示通过 E_l 类型边连接的邻居节点；$\mathbf{W}_l$ 为 E_l 边类型专用参数矩阵；Update_k 是 V_k 类型节点的更新函数。通过这种方式，异构图神经网络能够对不同类型的节点和边进行有效的建模，并且能够捕捉节点之间的复杂关系。

异构图神经网络的一个重要挑战是如何有效地处理不同类型的节点和边，以及如何融合来自不同源的信息。近年来，许多方法通过引入多层次的聚合操作、图卷积网络（GCN）和注意力机制等，来增强模型对异构图的建模能力。

2.3 注意力机制

注意力机制通过对输入信息的不同部分赋予不同的权重，使得模型能够聚焦于最相关的信息，从而提升模型的表达能力和处理效率。最早在机器翻译任务中，注意力机制被提出并成功应用，使得神经网络能够在处理长序列时，关注输入序列中的重要部分，从而解决了传统 RNN 模型在长序列建模中的局限性。目前，比较经典的注意力机制包括自注意力机制和多头注意力机制。

2.3.1 自注意力机制

自注意力机制（Self-Attention）最初应用在 Transformer 模型[29]，由谷歌团队在 2017 年提出。该注意力能够在每个位置生成对其他位置的依赖关系，并根据这些关系来加权输入序列中的信息。

自注意力机制允许模型在处理输入时关注序列中每个位置的信息，增强了模型的全局感知能力。它的核心思想是计算每个元素与其他所有元素的相关性，并基于这些关系加权输入信息，如图2.3所示。其计算过程如下：

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.10}$$

其中，Q，K，和 V 分别表示查询（Query）、键（Key）和值（Value）矩阵，$\sqrt{d_k}$ 是一个归一化项，d_k 是键的维度，A 是最终的注意力矩阵，表示各个元素之间的关联强度。通过 Softmax 函数，将每个位置的注意力分数归一化，确保加权后的和为 1。

自注意力机制允许模型在不同的时间步之间直接传递信息，因此在机器翻译和文本生成等序列处理任务中，能够有效捕捉长程依赖关系。

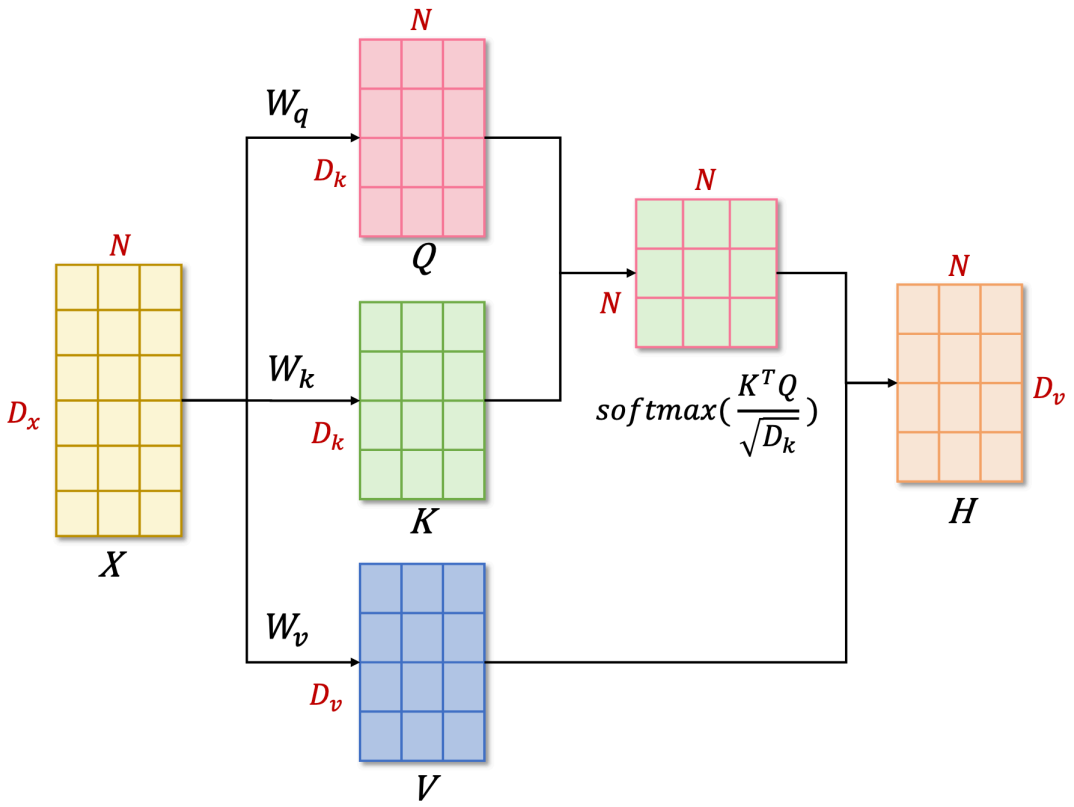


图 2.3　自注意力计算过程

2.3.2　多头注意力机制

多头注意力机制（Multi-Head Attention）是对自注意力机制的扩展，旨在通过并行计算多个注意力头，捕获不同的关联信息，从而进一步提升模型的表达能力。多头注意力机制通过将查询、键和值的矩阵分成多个子空间并分别进行自注意力计算，最终将每个头的输出拼接在一起，再经过线性变换得到最终的输出，如图2.4所示。

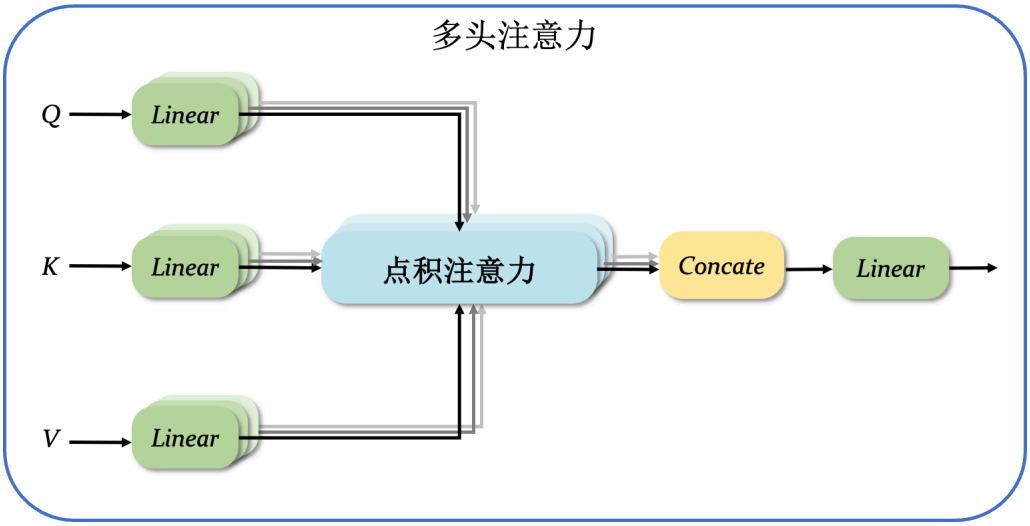


图 2.4　多头注意力结构图

多头注意力的计算过程如以下公式所示：

$$H = \text{concat}(H_1, H_2, \ldots, H_h)W^O, \tag{2.11}$$

其中，H_i 是第 i 个注意力头的输出，W^O 是输出的线性变换矩阵，h 表示头的数量。每个头负责关注输入序列中的不同方面的信息，这样模型能够从不同的角度捕捉序列中的复杂关系。

多头注意力机制极大地提升了自注意力的表示能力，并且在实际应用中，尤其是在自然语言处理任务中，取得了显著的效果。它不仅能帮助模型有效捕捉全局信息，还能通过并行计算提高模型的计算效率。

通过引入自注意力机制和多头注意力机制，模型能够在不同粒度的文本表示中自适应地调整关注重点，从而提升实体识别的准确度。在处理长实体和复杂结构时，注意力机制通过有效的上下文建模，使得模型能够更好地捕捉跨句子、跨段落的实体关系，特别是在长实体边界的准确区分方面，展现出了巨大的优势。

2.4 对比学习

对比学习（Contrastive Learning）是一种无监督表示学习方法，其通过构建正负样本对，驱使模型在特征空间中拉近正样本对的表示距离，同时推离负样本对的表示距离，从而学习具有判别性的数据表征。其核心思想是：通过设计一种损失函数，使得相似的样本在特征空间中聚集在一起，而不相似的样本则被拉开距离。对比学习的一个重要应用是在图像、文本等领域的表示学习中，通过有效的相似度度量，使得模型能够通过简单的计算得到高质量的特征表示。

NER 任务中，对比学习通过对实体的不同文本表述进行特征编码与相似度度量，使模型将同一实体在文本中的各种表述所对应的特征向量，映射至语义空间中更相近的位置，从而提升对实体的精准识别与泛化能力[33]。也就是说，正样本和负样本的特征向量通过对比学习逐步调整，使得相似实体的特征向量更加接近，而不同实体的特征向量相互远离，从而有效提升模型对实体边界的识别能力。在常见的对比学习方法中，使用损失函数来度量样本之间的距离，并通过最小化该损失来进行优化。

最常见的对比学习损失函数是对比损失（Contrastive Loss），其公式如下：

$$L_{\text{contrast}} = \frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot D(z_i, z_i^+) + (1-y_i)\cdot \max(0, m - D(z_i, z_i^-))\right], \tag{2.12}$$

其中，$D(z_i, z_i^+)$ 表示正样本对之间的距离，$D(z_i, z_i^-)$ 表示负样本对之间的距离，y_i 是样本对的标签（1 表示正样本对，0 表示负样本对），m 是一个超参数，用来设置负样本对的边界。

通过这种方式，对比学习在没有明确标签的情况下，利用样本对的相对关系来学习有效的特征表示。在处理 NER 任务中面临的长实体和复杂边界问题时，对比学习通过最大化正样本对之间的相似性来最小化负样本对，使模型具有更精确的边界区分能力。

2.5 深度可分离卷积

深度可分离卷积（Depthwise Separable Convolution, DSC）是一种通过将传统卷积操作分解成两个更简单的操作来降低计算复杂度和参数量的卷积神经网络，最初在 MobileNet 模型[34]中应用。与传统的卷积操作不同，深度可分离卷积将卷积过程分为两个独立的步骤：逐深度卷积（Depthwise Convolution）和逐点卷积（Pointwise Convolution）。深度卷积操作只针对每个输入通道执行卷积，逐点卷积则使用 1x1 卷积对不同通道的特征进行融合，计算过程如图2.5所示。

在传统卷积中，输入的每个通道和卷积核的每个通道都会进行全连接操作，计算复杂度随着输入通道数 C_{in} 和输出通道数 C_{out} 的增加而迅速增加。对于一个 $C_{in}\times H\times W$ 的输入特征图和一个 $C_{out}\times C_{in}\times K\times K$ 的卷积核，计算量为 $O(C_{in}\times C_{out}\times K^2)$，其中 H 和 W 是特征图的高度和宽度，K 是卷积核的尺寸。深度可分离卷积将这个过程分解为两个部分，首先对每个通道单独进行卷积操作，然后通过逐点卷积进行通道间的信息融合。这种分解减少了计算量和参数量。

在逐深度卷积中，每个输入通道与对应的卷积核进行卷积，输出的特征图与输入的通道数相同。其计算公式为：

$$Y_d = \sum_{i=1}^{C_{in}} W_i * X_i, \tag{2.13}$$

其中，W_i 是针对第 i 个输入通道的卷积核，X_i 是输入特征图的第 i 个通道，$*$ 表示

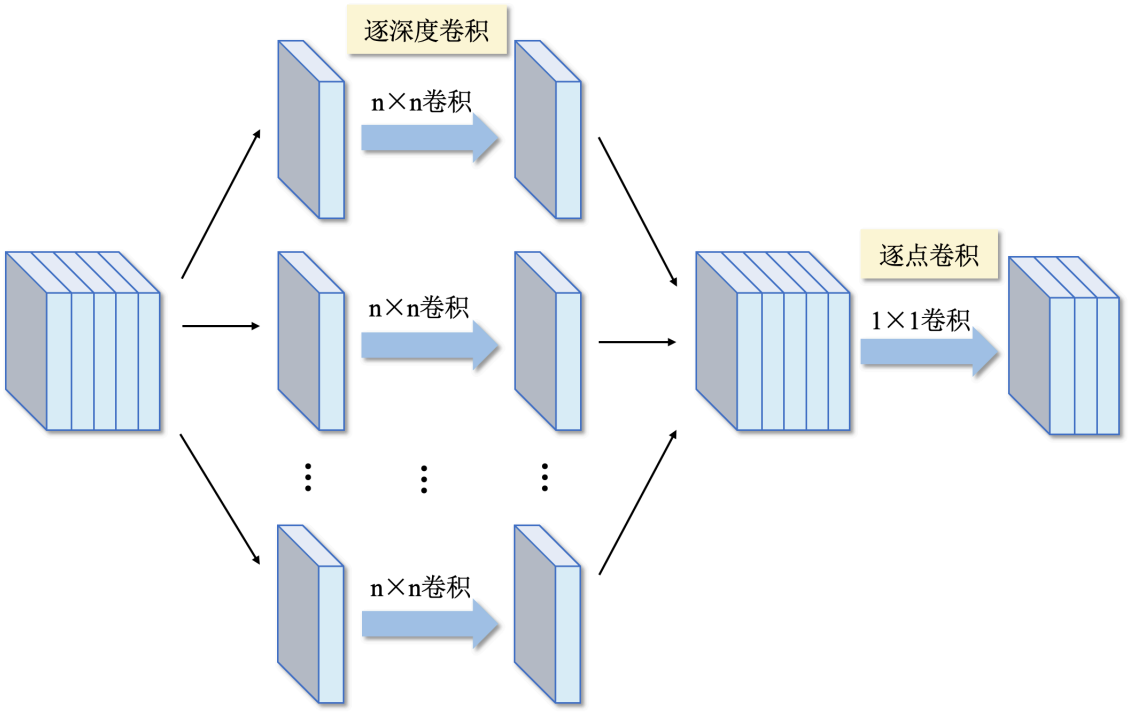


图 2.5 深度可分离卷积计算过程

卷积操作，C_{in} 表示输入通道数。

接下来，逐点卷积使用 1×1 的卷积核对每个通道的输出进行融合，得到最终的输出特征图。逐点卷积的计算公式为：

$$Y_p = \sum_{i=1}^{C_{in}} W_i^{1\times1} * Y_{d_i}, \tag{2.14}$$

其中，$W_i^{1\times1}$ 是 1x1 卷积核，Y_{d_i} 是深度卷积的第 i 个通道输出。通过这种方式，逐点卷积将深度卷积的输出融合为具有更丰富信息的特征图。

通过将卷积操作分解为深度卷积和逐点卷积，深度可分离卷积显著降低了计算复杂度和存储需求。计算复杂度从传统卷积的 $O(C_{in} \times C_{out} \times K^2)$ 降低为 $O(C_{in} \times K^2 + C_{in} \times C_{out})$，其中 C_{in} 是输入通道数，C_{out} 是输出通道数，K 是卷积核的大小。

深度可分离卷积能够有效提升模型的计算效率，尤其在处理长文本或复杂句子时。传统的卷积操作在处理高维文本数据时通常需要大量计算和内存资源，深度可分离卷积通过高效的计算能力，帮助模型更好地捕捉文本中的局部特征，减少计算开销，提高识别精度。

2.6 本章小结

本章系统梳理了论文涉及的核心理论方法及其技术原理。在 NER 方面，基于规则模型、CRF 与预训练模型的递进式框架得到完整阐释，其中 CRF 通过全局概率建模有效解决了序列标注任务中的标签依赖约束问题；针对 GNN，介绍了相关内容，特别是同构图和异构图的处理方法；对于注意力机制，从自注意力单元的上下文关联计算到多头注意力并行特征提取机制，为复杂实体边界判别提供方法论支撑。此外，本章还介绍了对比学习技术与深度可分离卷积网络。通过对这些先进技术的介绍和分析，可以更好地理解它们在材料领域文献挖掘中的潜力，为后续章节的方法提供支持。

第三章　基于异构图与分块感知的材料命名实体识别

近年来，数据驱动的材料研发正逐步成为材料科学领域的重要研究方法。科学文献作为材料研究成果的重要载体，蕴含着大量关于材料成分、制备工艺、性能测试等关键信息，然而这些信息大多以非结构化文本形式存在，传统人工提取方式效率低下且成本高昂。碳纤维复合材料作为典型的先进复合材料，其性能优化涉及多维度参数，但现有文献挖掘方法因标准化数据缺失、实体关系复杂等问题，难以直接应用于该领域。因此，本章提出了一种面向复合材料文献的命名实体识别模型，通过引入分块注意力机制、异构图网络和深度可分离卷积等模块，提升了模型在复杂实体识别和长序列依赖建模方面的能力。

3.1　方法概述

复合材料文献挖掘方法面临的挑战主要体现在两个方面。第一，复合材料文献涉及广泛的实体类别，包括材料组成和加工工艺等专业知识，而传统命名实体识别方法往往缺乏处理这些高度专业实体类别的灵活性。第二，由于领域知识的上下文依赖性，传统 NER 方法在识别和解析文献关键信息时可能受限。例如，复合材料文献中的“屈服”一词可能描述屈服强度等力学性能，而在其他工程领域，它可能与土木结构的承载能力屈服相关。这种上下文依赖性使传统的 NER 方法难以准确识别此类专业术语。此外，复合材料文献通常篇幅较长、语义结构复杂，其文本中实体与上下文之间往往存在跨句甚至长距离的依赖关系，这对实体识别模型的上下文建模能力提出了更高要求。这些局限性表明，迫切需要开发专门针对复合材料的文献挖掘方法，现有方法需进一步优化改进，以应对复合材料领域的复杂性。

为应对复合材料文献中长序列依赖、实体关系复杂等挑战，本章提出一种语义增强图网络模型（Semantic-Rich Graph Network, SRGN），通过引入分块注意力机制以提升对长文本的处理能力，构建异构图结构以融合实体类型与上下文信息，并结合深度可分离卷积实现多尺度特征提取。此外，模型还设计了动态边权重机制和深度评分网络，以增强节点间语义交互与实体类别判别能力。通过这些关键组件的协同作用，SRGN 能够高效地对长序列数据进行语义表达并准确识别文献中的关键信

息。SRGN 模型的整体结构如图3.1所示。

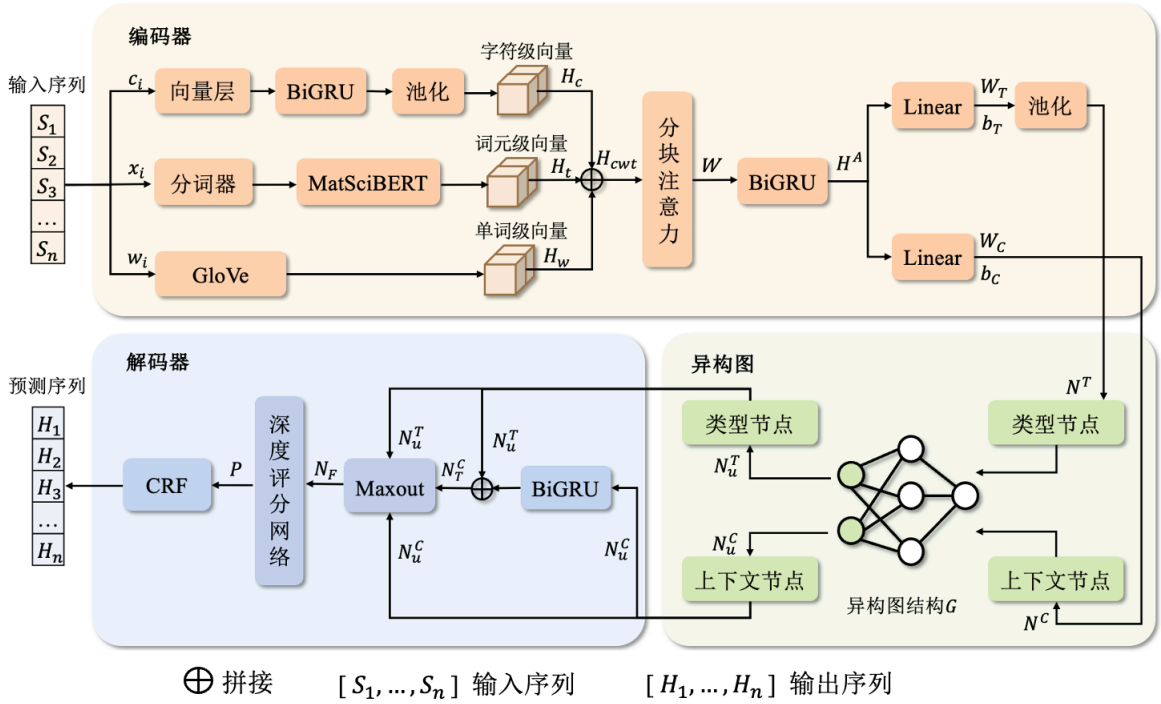


图 3.1　语义增强图网络模型结构图

具体来说，首先编码器部分引入了分块注意力机制，以有效应对长序列建模的挑战。传统 Transformer 的全局自注意力在处理超长序列时计算复杂度和内存开销极高，而分块注意力通过将文本划分为多个块，并对每块分别计算注意力，从而降低了整体计算复杂度，同时保持对关键局部信息的敏感性。其次，SRGN 通过类型节点和上下文节点构建了异构图，融合了文本上下文表示和实体类型表示，并在图网络的迭代过程中不断更新节点信息，增强上下文表征。在异构图的上下文节点局部窗口中，SRGN 应用了深度可分离卷积，旨在提取和融合文本的全局与局部特征。对于边权重，传统的图注意力机制复杂度为 $O(n^2)$，处理长文本时受限。为了进一步提升模型的上下文理解能力和节点间交互，SRGN 引入了自适应的动态边权重机制，通过迭代更新优化节点之间的语义连接。解码器结构中，引入了深度评分网络，用于计算每个实体的预测概率。该网络根据每个实体类型特有的上下文信息进行独立优化，避免了不同类型之间的相互干扰，充分保留了各类型的差异性。通过优化模型设计，进一步提高了 SRGN 实体识别的准确性。

3.2 语义增强图网络

语义增强图网络（SRGN）用于识别和解析材料文献中的实体信息。在该模型中，采用了 GraphNER 模型[35]作为骨干网络。GraphNER 是一种基于异构图的命名实体识别模型，能够比较有效地解决嵌套实体识别问题。但对于材料文献，相同的材料需要上下文来明确具体是哪一种组分，例如环氧树脂在不同的文献或实验中，可能作为基体也可能作为辅助添加剂。因此，本文进一步扩展了模型以进行优化，以保证处理具有复杂的上下文依赖关系，以及包含长序列的数据时具有更好的灵活性。

如图3.1所示，SRGN 的整体结构可划分为编码器、异构图和解码器三个部分。其中，编码器接收输入的文献文本，得到不同粒度的语义信息，包括字符级、词元级以及单词级语义信息，采用分块注意力提升长序列处理效率；随后，通过构建包含类型节点与上下文节点的异构图，在全局范围内建模上下文语义关联；解码器部分将迭代后的两个节点表示结合起来，并采用深度评分网络计算实体标签的预测概率，最后将预测概率送入 CRF[32]中实现命名实体识别。下面对 SRGN 进行详细介绍。

3.2.1 编码器结构

字符级向量表示 H_c 通过向量层 E_c 将字符 c_i 映射为向量，再依次通过双向门控循环单元模型 BiGRU[36]和池化操作获取每个字符的向量，捕捉字符级粒度的语义：

$$H_c = \text{MaxPool}(\text{BiGRU}(E_c(c_i))), \tag{3.1}$$

词元向量表示 H_t 通过预训练模型得到。相比于单词级的静态向量，词元向量能够根据上下文变化捕捉动态信息。输入的文本首先通过预训练模型的分词器进行分词，然后送入到预训练模型中以获得词元向量表示。由于输入的是材料领域的文本，所以预训练模型选择材料领域的 MatSciBERT[11]。具体过程如以下公式所示：

$$H_t = \text{MatSciBERT}(\text{Tokenizer}(x_i)), \tag{3.2}$$

其中，$x_i = \{t_1, t_2, ..., t_n\}$ 表示输入的词元序列。

单词级向量表示 H_w 通过词向量模型 GloVe[37]捕捉单词级别的语义信息，其中 w_i 表示给定句子的词序列：

$$H_w = \text{GloVe}(w_i), \tag{3.3}$$

随后，将三种不同粒度的向量表示通过拼接融合起来，如以下公式所示：

$$H_{cwt} = \text{concat}(H_c; H_t; H_w), \tag{3.4}$$

考虑到处理大量的文本序列时，模型可能不会充分关注与特定任务最相关的信息部分，所以需要引入注意力机制。但对于材料文献，特别是复合材料文献文本而言，序列长度可能相当大，处理效率较低。传统的注意力机制在处理长序列数据时计算成本高昂，主要因为其注意力矩阵的计算复杂度随序列长度呈平方增长。因此，为了应对常见的长序列处理挑战，本章引入了分块注意力机制，结构如图3.2所示。分块注意力通过将序列划分成较小的块，显著降低每次计算的复杂度至 $O(nm)$ $(m \ll n)$。此外，为了缓解分块可能导致的上下文丢失问题，每个块在注意力计算时引入前后块的上下文窗口，使得模型在保持高效计算的同时，仍能捕捉跨块间的语义关联。这种设计在兼顾计算效率的基础上，有效维持了对长距离依赖的建模能力。

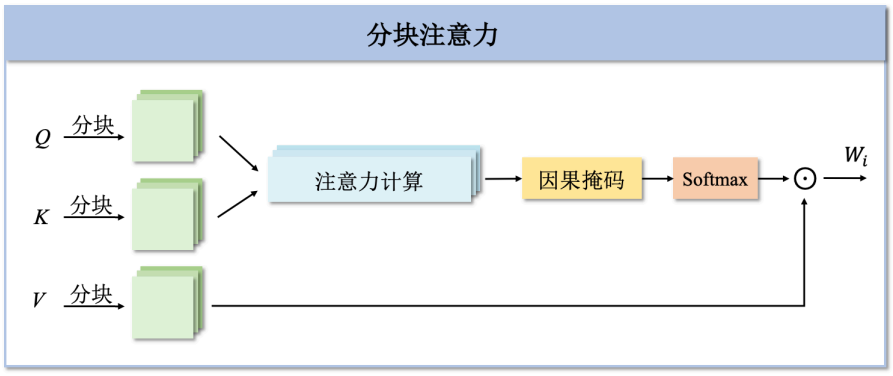


图 3.2 分块注意力机制结构图

在分块注意力机制中，$H_{cwt} \in \mathbb{R}^{B \times L \times H}$ 表示拼接后的特征向量，其中 B 为批量大小，L 为序列长度，H 为特征维度。对于输入的查询（Q）、键（K）、值（V）矩阵，分别将它们划分成大小为 C 的多个块。对于每个块 i，将当前块的查询矩阵 Q_i 与局部上下文窗口的键矩阵 $K_{[start:end]}$ 进行相乘，计算注意力得分 $attn_i$，过程如以下公式所示：

$$attn_i = \frac{Q_i \cdot K^T_{[start:end]}}{\sqrt{H}}, \tag{3.5}$$

其中，H 是特征向量的维度。注意力得分会通过 $\frac{1}{\sqrt{H}}$ 进行缩放，以适应特征空间的维度。

在计算过程中，为了防止信息泄露，在注意力结构中增加了因果掩码 M，确保每个词元仅关注前面的时间步。随后，将通过掩码处理的注意力得分传入 Softmax 中进行归一化处理，得到注意力权重，每个块的最终输出通过将注意力权重与值向量 $V_{[start:end]}$ 进行加权和计算得到单个块的注意力输出 W_i：

$$W_i = \text{Softmax}(attn_i \odot M) \cdot V_{[start:end]}, \tag{3.6}$$

其中，对于因果掩码 $M \in \mathbb{R}^{C\times|K_i|}$，需对每个查询向量位置 p 和键向量位置 q 之间的依赖关系进行约束。设 p 表示当前查询位置，q 表示被关注的键的位置，仅允许 $q \leq p$ 的位置参与注意力计算，掩码矩阵 M 有如下定义：

$$M_{p,q} = \begin{cases} 0 & \text{if } q \leq p \\ -\infty & \text{otherwise} \end{cases}, \tag{3.7}$$

最后，所有块的输出 W_i 被拼接在一起，形成最终的输出 W。在分块注意力机制之后，得到的序列会传入 BiGRU 中得到向量 H^A，捕捉序列中正向和反向的依赖关系。过程如下公式所示：

$$H^A = \text{BiGRU}(W), \tag{3.8}$$

对于向量 H^A, 通过线性变换分别生成类型节点 N^T 和上下文节点 N^C。这两种类型的节点在模型中扮演着重要角色，其中类型节点用于表示实体的类别信息，而上下文节点则捕捉文本或数据中的上下文关系，如以下公式所示：

$$N^T = \text{MaxPool}(W_T H^A + b_T), \tag{3.9}$$

$$N^C = W_C H^A + b_C, \tag{3.10}$$

其中，W_C、b_C 以及 W_T、b_T 表示线性变换中使用的权重和偏置，H^A 是每个 N^T 对应的预定义的实体类别。随后，通过这两类节点构建异构图，融入不同类型的信息，在后续的处理过程中进行交互和信息传递。

3.2.2 异构图结构

在图网络中，节点通常需要捕捉和表示复杂的数据关系。异构图的引入使得模型能够充分利用多种节点类型的特征，从而更好地捕捉和理解不同实体间的复杂关

系。SRGN 的异构图结构中，上下文节点代表材料文献文本中的词或短语，而类型节点代表可能的实体类型。通过构建类型节点和上下文节点的异构图，模型能够有效捕捉文献中的实体关系和上下文信息，提升命名实体识别的精度，同时异构图通过多层迭代更新实现两类节点的交互。每层更新模块包含两个组件：动态边权重机制与深度可分离卷积，分别用于捕捉全局类型关联与局部语义模式。

材料科学文献中的命名实体常呈现局部连续的语句，例如化学式 “LiNiO ”、工艺参数 “1200°C/2h” 等，它们的语义特征在短窗口内高度相关。传统图神经网络依赖全连接注意力机制建模节点关系，但此类方法对局部连续特征的捕捉效率较低，且参数量随序列长度平方增长，难以适应长文本场景。为此，SRGN 在上下文节点的局部窗口中引入深度可分离卷积[34]，在上下文节点更新中显式提取局部语义特征。

原始的深度可分离卷积模块由逐通道卷积与逐点卷积两阶段构成，但为了减少参数量，本章在这部分的设计中仅保留逐通道卷积。对于上下文节点特征 $N^C \in \mathbb{R}^{B\times L\times d}$（$d$ 为隐藏层维度），模块首先通过逐通道卷积提取局部窗口内的空间特征：

$$N^C_{local} = \text{DepthwiseConv1D}(N^C, k = 2w + 1), \tag{3.11}$$

其中，w 为窗口半径，卷积核在通道维度独立作用于每个特征通道，参数量降低至标准卷积的 $\frac{1}{d}$。随后，为进一步增强局部特征的语义指向性，模块将类型节点的全局特征与卷积输出融合。首先对类型节点 N^T 进行全局平均池化：

$$N^T_{global} = \frac{1}{N^T}\sum_{i=1}^{N^T} N^T_i, \tag{3.12}$$

随后，通过线性投影将其维度扩展至序列长度，与卷积特征相加，得到新的上下文节点 N^C_{pre}：

$$N^C_{pre} = \text{GELU}(N^C_{local} + W_g \cdot N^T_{global}), \tag{3.13}$$

其中，W_g 为可学习参数。深度可分离卷积通过解耦空间与通道维度，在保持局部特征提取能力的同时，将参数量降至标准卷积的 $\frac{1}{d}$。

在图结构中，通常采用传统的图注意力机制计算节点间相似度，生成边权重，但其复杂度达到 $O(n^2)$ 难以处理长文本。为了增强节点间的交互和降低复杂度，引入了动态边权重机制，它能更精确地调整节点间的关系，有效提升模型对节点间复杂

语义依赖的建模能力。对于第 t 层更新，将类型节点特征沿序列维度广播，与上下文节点拼接，再基于 MLP 得到动态的权重矩阵：

$$W_{dynamic} = \text{MLP}([N^C; \text{Expand}(N^T_{global})]), \tag{3.14}$$

根据动态权重矩阵，对上下文特征进行加权聚合，更新类型节点，得到 N^T_{pre}：

$$N^T_{pre} = N^T + \sum_{j=1}^{L} W_{dynamic,j} N^C_j, \tag{3.15}$$

动态边权重生成器的应用使得异构图在不使用传统注意力的情况下，以特征拼接和非线性变换生成动态权重矩阵，在减少计算开销的同时增强语义建模的灵活性。

随后，将类型节点 N^T_{pre}、上下文节点 N^C_{pre} 送入异构图的更新层。在这个阶段中，本文参考并应用了 GraphNER[35] 中异构图结构的更新层设置，进行类型节点和上下文节点的迭代与更新，如以下公式所示：

$$N^T_u = \text{UpdateLayer}(N^T_{pre}), \tag{3.16}$$

$$N^C_u = \text{UpdateLayer}(N^C_{pre}), \tag{3.17}$$

在每一层异构图的更新过程中，上下文节点和类型节点的更新存在依赖关系。具体而言，上下文节点的更新基于进入该层前的原始状态，而类型节点的更新则依赖同一层内已更新的上下文节点。这种渐进式更新机制使得模型能够逐步融合异构节点间的语义信息。

最终，将迭代更新后的节点 N^T_u 和 N^C_u 送入解码器部分，实现命名实体识别。

3.2.3 解码器结构

在经过异构图的迭代和更新过程之后，得到更新后的两个节点信息，将它们结合起来通过解码器结构实现命名实体识别。

在解码器模块，N^C_u 通过 BiGRU 处理，用于捕捉上下文依赖。N^T_u 与处理后的 N^C_u 融合，得到 N^C_T，用于融合异构图中得到的类型和上下文信息，增强特定实体类型的识别。过程如以下公式所示：

$$N^C_T = \text{concat}(\text{BiGRU}(N^C_u); N^T_u), \tag{3.18}$$

接下来，N_u^C，N_u^T，N_T^C 进一步与 Maxout 激活函数相结合[38]，得到表示所有实体类型的上下文向量 N_F。过程如以下公式所示：

$$N_F = \text{Maxout}(N_T^C + N_u^C; N_u^T), \tag{3.19}$$

在材料科学文献的 NER 任务中，传统线性网络难以有效捕捉材料实体的复杂语义特征。材料领域实体常呈现复合命名规则及上下文敏感特性，基于线性变换的评分方法存在表征能力不足的问题，无法充分建模材料术语与标签间的非线性关联，容易导致实体边界的误判和类型分类错误。

针对上述问题，在 SRGN 中引入深度评分网络（Deep Scoring Network, DSN）。DSN 采用多层级结构设计，引入了多层非线性变换，通过降维投影生成标签预测概率，增强了模型对文献文本中稀疏实体表达的泛化能力，结构如图3.3所示。

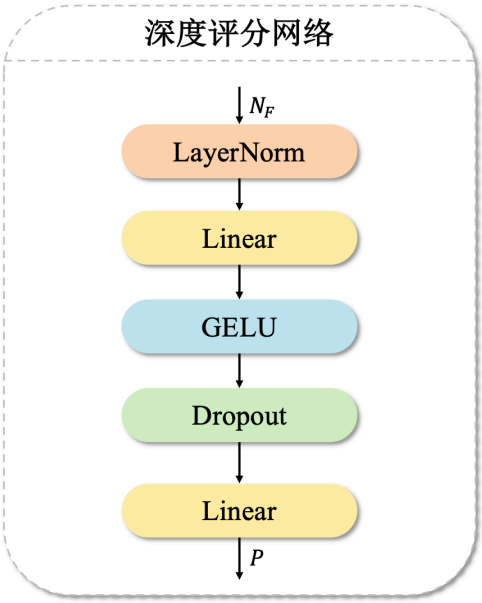


图 3.3　深度评分网络结构图

在 DSN 中，对于包含类型信息的上下文特征 $N_F \in \mathbb{R}^d$，首先进行规范化处理以稳定训练过程。接着，通过执行非线性变换来扩展特征维度，以此增强特征的表达能力。过程如以下公式所示：

$$P = W_2 \cdot \text{Dropout}(\text{GELU}(W_1 \cdot \text{LayerNorm}(N_F))), \tag{3.20}$$

其中，$W_1 \in \mathbb{R}^{2d \times d}$ 为扩展投影矩阵，$W_2 \in \mathbb{R}^{C \times 2d}$ 为标签映射矩阵，$GELU(x) = x \cdot \Phi(x)$ 为高斯误差线性单元。

与单层线性变换相比，DSN 能更好地捕捉复杂的上下文依赖关系，提高实体识别的准确性。在 NER 任务中，实体之间的关系通常是非线性的，尤其是在长距离上下文依赖关系中，DSN 可以学习到更丰富的模式和特征。其次，DSN 为每种实体类型保留了一个独立的网络，因此每种类型的标签概率计算都可以根据其独特的上下文信息进行独立优化，避免了不同类型之间的干扰，充分保留了类型之间的差异。

后续则将命名实体识别建模为以 BIO 为标注方式的序列标注任务，并通过 CRF[32] 来实现命名实体识别，提取材料文献文本中的目标实体。

3.2.4 损失函数

本章采用 CRF 的条件对数似然损失作为模型的核心损失函数，通过最大化标注序列的条件概率，有效捕捉文本中实体间的依赖关系，提升 NER 的预测准确性。CRF 通过构建输入特征与输出标签序列之间的概率模型，有效地捕捉和利用了标签之间的依赖关系。

CRF 在预测序列标签时，不仅考虑到单个标签的发射概率，还结合了标签之间的转移概率。这种方法能够全局考虑整个标签序列的合理性，从而避免了局部最优解带来的偏差。损失函数为条件对数似然损失，公式如下：

$$\mathcal{L}_{CRF} = \log\left(\sum_{y\in Y} e^{s(x,y)}\right) - s(x, y^*), \tag{3.21}$$

其中，y^* 是真实的标签序列，Y 是所有可能的标签序列集合，$s(x, y)$ 是给定输入 x 下序列 y 的得分。

CRF 损失函数的引入主要解决了两大问题。首先，通过全局归一化处理，CRF 帮助模型克服了标签偏置问题，即模型不会因为某些标签在训练集中出现频率较高而倾向于预测这些标签。其次，CRF 能够显著提升模型对标签序列内部结构依赖关系的学习能力，保证了输出标签在逻辑和结构上的一致性。

3.3 实验与讨论

本章对 SRGN 模型进行了实验验证与结果分析。首先介绍了实验数据集，并说明实验环境、参数设置及评价指标。随后，通过对比实验和消融实验评估模型性能，并探讨各模块对整体效果的影响。

3.3.1 数据集介绍

为了验证 SRGN 的有效性，本章选择了两个材料领域的命名实体识别数据集：碳纤维复合材料数据集 CompMatLitDS 和材料科学公共数据集 MatScholar[10]，分别对方法进行实验评估。其中，CompMatLitDS 是手动进行文献收集与标注的数据集，MatScholar 为材料领域的公共数据集。

（1）复合材料数据集 CompMatLitDS

为了训练 SRGN，同时考虑到复合材料领域缺乏专门的数据集，本章对碳纤维复合材料相关文献进行了手动标注，并构建了一个复合材料文献数据集，命名为 CompMatLitDS。该数据集由 380 篇复合材料相关的 PDF 文献标注而成，这些文献来源于 Elsevier、Wiley、MDPI 等公开文献数据库，时间范围为 2019 至 2022 年。这些文献被划分为 1282 段语料文本，并通过 Doccano 工具[39]进行序列标注，涵盖 13 个实体类别。标注任务完成后，1282 个文本段被转换为 BIO 格式 [17]（B-Begin, I-Inside, O-Outside）。最终，数据集按照 6:2:2 的比例划分为训练集、验证集和测试集，得到 CompMatLitDS 数据集。数据集划分如图3.4（a）所示。SRGN 将在该数据集上进行训练、验证和测试。

针对命名实体识别任务，定义了 13 种复合材料的实体类别以进行序列标注，主要包括：基体（Matrix）、填料（Filler）、复材（Composite）、辅助添加剂（Auxiliary Additives）、材料数值（Material Values）、加工类型（Machining Types）、工艺数值（Process Parameters）、制备方法（Preparation Methods）、成型类型（Forming Types）、性能名称（Property Names）、性能数值（Property Values）、测试方法（Test Methods）、测试标准（Test Standards）。每个实体类别对应的详细解释如表3.1所示，这些类别涵盖了复合材料领域中的核心概念，为命名实体识别任务提供了一个统一的标注标准，以提高标注过程的一致性和准确性。

（2）材料公共数据集 MatScholor

MatScholar 是 Weston 等人[10]对材料文献摘要进行文献挖掘后得到的数据集，涵盖了 1900 年至 2018 年间发表的大部分材料文章的英文摘要。经过预处理和标注，该语料库包含超过 327 万条摘要，其中包括七类实体：无机材料名称（Material，MAT）、样品描述符（Sample Descriptors，DSC）、对称/相标签（Symmetry/Phase label ，SPL）、材料性能（Material Properties，PRO）、应用（Applications，APL）、合成方法（Synthesis

表 3.1　CompMatLitDS 数据集中 13 个实体定义和实体类别的说明

序号	实体类别定义	实体类别说明
1	基体（Matrix）	复合材料的主要成分，包括聚丙烯、环氧树脂、酚醛树脂等
2	填料（Filler）	在基体中加入的颗粒或纤维物质，如碳纤维、玻璃纤维等，改善材料特性
3	复材（Composite）	由基体和填料组成的材料，如碳纤维复合材料
4	辅助添加剂（Auxiliary Additives）	辅助添加剂主要包括固化剂、增溶剂、增塑剂、催化剂和上浆剂等
5	材料数值（Material Values）	基体或填料的含量，通常以质量分数（wt%）或体积分数（vol%）表示
6	加工类型（Machining Types）	复合材料加工过程中使用的加工工艺，如固化、热压、压缩成型等
7	工艺数值（Process Parameters）	工艺条件：温度、时间、压力等；尺寸：长度、宽度、高度等
8	制备方法（Preparation Methods）	熔融、共混和 3D 打印
9	成型类型（Forming Types）	复合材料经过加工后，最终形成的膜状、块状或层状结构等
10	性能名称（Property Names）	文献中提到的性能名称，包括机械性能（拉伸强度、弯曲强度等）和玻璃化转变温度
11	性能数值（Property Values）	与性能名称相关的性能数值
12	测试方法（Test Methods）	用于测试力学性能的方法，如 DSC 测试，三点弯曲测试等
13	测试标准（Test Standards）	测试方法对应的标注，通常为 ASTM 标准

Methods，SMT）和表征方法（Characterization Methods，CMT）。数据集划分如图3.4（b）所示，实体定义和实体类型的说明如表3.2所示。

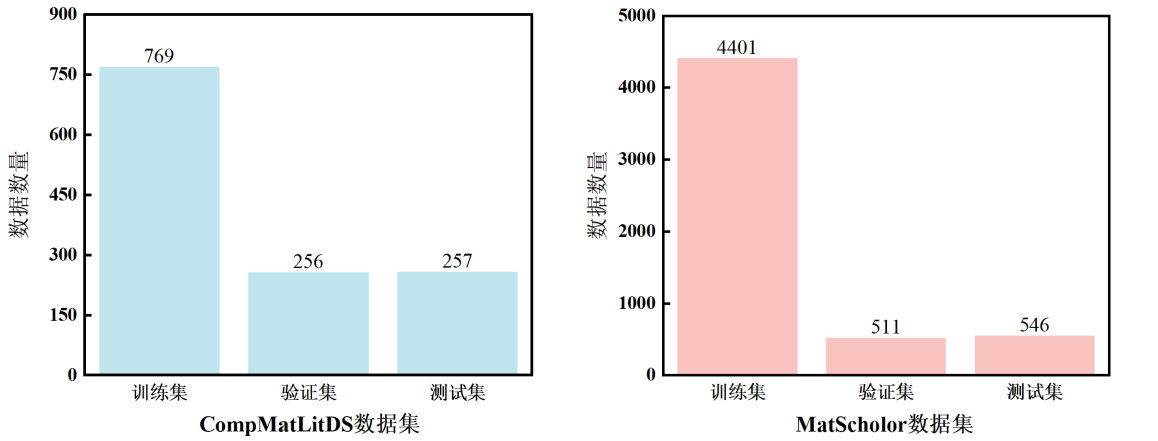
3.3.2　实验环境及模型参数

本章的所有实验都是在运行 Python 3.9.13 和 PyTorch 1.12.1 的环境下进行的，使用的 GPU 型号为 RTX 3090 24GB，CPU 型号为 Intel(R)@2.40GHz。

对于模型的参数设置，默认使用 MatSciBERT[11] 作为预训练模型获得词元级信息; 设置句子的最大长度为 256，字符级信息维度为 50；默认采用 GloVe 模型获得单词级信息，并且维度为 50。另外，设置异构图的迭代次数为 6。

表 3.2　MatScholor 数据集中 7 个实体定义和实体类型的说明

序号	实体类别定义	实体类别说明
1	无机材料名称（MAT）	任何无机固体或合金，以及任何在室温下为非气态的元素
2	样品描述符（DSC）	描述样品类型或形状的特殊描述
3	对称/相标签（SPL）	晶体结构或相的名称，或任何对称性标签
4	材料性能（PRO）	任何可以测量、具有单位和数值的性质，或任何由材料展示的定性属性或现象。
5	应用（APL）	任何高级应用或具体的设备
6	合成方法（SMT）	合成材料的任何技术，或样品生产中的任何步骤
7	表征方法（CMT）	用于表征材料的任何方法，无论是实验还是理论；或用于命名方程或模型的名称



（a）CompMatLitDS 数据集划分　　（b）MatScholor 数据集划分

图 3.4　两个材料数据集的数据划分。(a) CompMatLitDS 数据集划分，(b) MatScholor 数据集划分。

3.3.3　评价指标

通过评估测试集的准确率，可以从更加量化和直观的角度评估模型的性能。因此，本章采用 F1-Score 作为评价指标，用于评估测试集的准确率，它代表精确度和召回率的调和平均值。同时，在实验中也考察了精确率（Precision）和召回率（Recall）这两个指标，因为单一指标在某些情况下可能无法全面反映模型性能的优劣，特别是在处理不平衡数据集时，精确率和召回率的权衡尤为重要。通过同时评估三个指标，在实验中能更全面理解和评价 SRGN 的表现。三个指标的计算如以下公式所示：

$$Precision = \frac{TP}{TP + FP}, \tag{3.22}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3.23}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{3.24}$$

其中，TP 表示正确预测为正类的样本数，FP 表示错误预测为正类的样本数，FN 表示实际为正类但预测为负类的样本数。

3.3.4 对比实验

为了评估 SRGN 在材料命名实体识别任务中的有效性，本章节在 CompMatLitDS 数据集与 MatScholor 数据集上进行了实验。对于训练的参数设置，设定学习率为 2×10^{-5}，学习率预热(Warmup)设置为 0.1，最大梯度设置为 1，权重衰减率设置为 0.01。另外，在所有的实验中，考虑到数据包含材料领域知识，因此均采用 MatSciBERT[11] 作为预训练模型，获得词元级信息。对于评价指标，采用了3.3.3节所描述的精确率、召回率和 F1 三项评价指标。

（1）CompMatLitDS 数据集实验结果

利用 SRGN，从复合材料科学文献中提取关键信息，提取示例如图3.5所示。从图中可以清晰看到，SRGN 能有效识别文献中的关键信息。

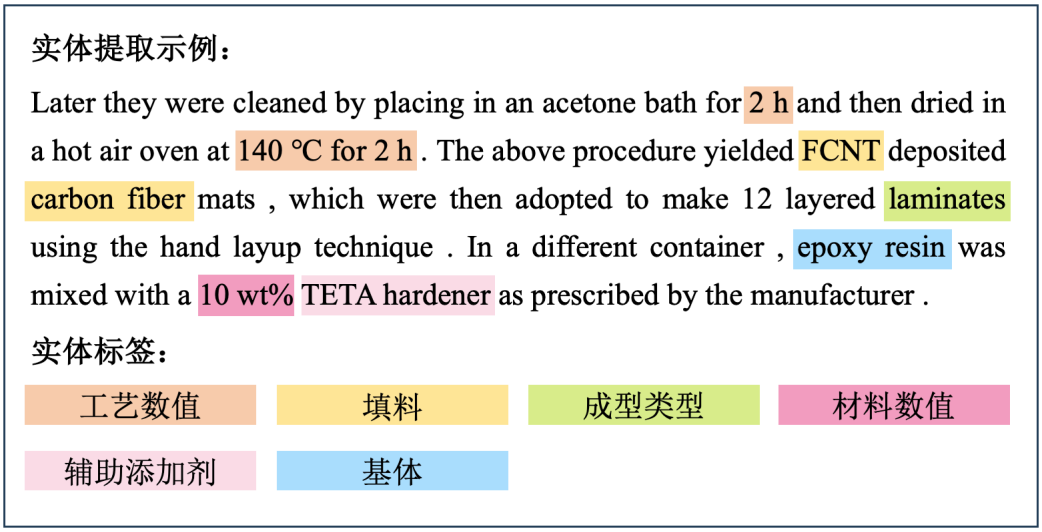


图 3.5 SRGN 的文献挖掘示例图

针对 CompMatLitDS 数据集，选择了近几年在 NER 领域表现比较好的模型进行实验。对比评估了基于 GraphNER[35]、PromptNER[40]、PIQN[41]、W2NER[42]、Binder[43] 以及 SRGN 的性能。GraphNER[35] 是一种基于异构图网络的 NER 模型，通

过构建实体关系图来捕捉上下文中的全局关联；PromptNER[40]采用提示学习范式，将预训练语言模型应用于 NER 任务；PIQN[41]则通过设置全局可学习的实例查询并行抽取实体，每个查询独立预测一个实体，从而避免逐类型或逐实体序列标注；W2NER[42]将命名实体识别建模为词-词关系分类，在二维矩阵上表示实体边界和相邻词关系，并采用多粒度卷积网络提取特征；Binder[43]采用了双编码器架构并结合对比学习，将候选文本片段和预定义的实体类别映射到同一向量空间进行匹配，从而提高识别准确性。这些不同方法的对比实验结果如表3.3所示。实验结果表明，SRGN 在召回率和 F1 上均优于其他对比模型，在精确率上仅次于其他方法，展示出比较明显的优势。具体来说，SRGN 模型达到了 94.56% 的精确率、95.92% 的召回率和 95.24% 的 F1-score，这些结果充分证明了该方法在准确提取实体方面的优越性。

表 3.3　不同模型在 CompMatLitDS 数据集上的对比实验结果

模型	Precision (%)	Recall (%)	F1 (%)
GraphNER[35]	**94.65**	94.85	94.75
PromptNER[40]	76.03	57.53	65.50
PIQN[41]	84.11	78.77	81.35
W2NER[42]	92.35	95.39	93.84
Binder[43]	93.47	93.37	93.42
Ours (SRGN) ✓	94.56	**95.92**	**95.24**

SRGN 能更准确地识别包含领域知识的实体，这体现了分块注意力机制和异构图结构的有效性，提升在处理复杂文本数据时的性能。通过深度可分离卷积优化节点迭代和边权重机制优化边，有效迭代和更新实体间的关系，从而提高了实体识别的准确性和全面性。此外，深度评分网络也在实体关系评分中发挥了关键作用，进一步提升了模型整体的识别效果。

在对比实验中，根据表3.3的结果可以看到，虽然 Binder 和 W2NER 等方法在通用命名实体识别任务中表现出色，但在处理包含特定领域知识的数据时，它们却面临着局限性，因为它们可能无法适应该领域独特的语言模式和复杂的实体结构。GraphNER 作为 SRGN 的骨干网络，通过构建实体关系图来捕捉上下文中的全局关联。虽然这种方法在提高上下文理解能力上有一定的优势，但在复杂的复合材料文献数据集上仍表现出不足，难以捕捉文本中的微妙语义变化，这也证明了 SRGN 的有效性。PromptNER 采用提示学习方法，将预训练语言模型应用于 NER 任务，但由

于复合材料文献常包含特殊术语和复杂的实体结构，该方法在准确抽取这些专业实体方面表现不佳。PIQN 则通过设置全局可学习的实例查询并行抽取实体，每个查询独立预测一个实体，可以提高处理速度和效率，但在实体相互关联和互为上下文的复杂文献中，这种方法可能无法充分理解和利用这些复杂的关系。W2NER 将命名实体识别建模为词-词关系分类，并采用多粒度卷积网络提取特征，清晰识别实体边界，但其局限性在于处理高度嵌套或交叉实体时的复杂性和准确性。Binder 采用了双编码器架构并结合对比学习，这有助于提高模型的泛化能力，通过将候选文本片段和预定义的实体类别映射到同一向量空间进行匹配，但其在复合材料文献中的应用可能受限于对比学习在处理少见或新颖实体类型时的不足，以及在高度专业化领域中模型训练数据的限制。

在 CompMatLitDS 数据集中不同实体类型的对比结果如表3.4所示。实验结果显示 SRGN 模型在多数实体类型上表现较好，达到 90% 以上的准确率，这表明 SRGN 模型在识别和分类复杂的实体关系时具有显著的优势。

表 3.4　CompMatLitDS 数据集中不同实体类型的实验结果

实体类型	Precision (%)	Recall (%)	F1 (%)
基体（Matrix）	97.30	96.26	96.77
填料（Filler）	95.93	98.77	97.33
复材（Composite）	97.06	96.78	96.93
辅助添加剂（Auxiliary Additives）	76.36	77.78	77.06
材料数值（Material Values）	91.23	97.20	94.12
加工类型（Machining Types）	95.51	97.70	96.59
工艺数值（Process Parameters）	88.89	90.64	89.76
制备方法（Preparation Methods）	99.99	96.29	98.11
成型类型（Forming Types）	94.12	92.75	93.43
性能名称（Property Names）	97.60	99.02	98.31
性能数值（Property Values）	94.12	96.97	95.52
测试方法（Test Methods）	91.92	92.86	92.39
测试标准（Test Standards）	97.14	97.14	97.14

然而，在对“辅助添加剂”和“工艺数值”两个实体类别的分析中，本文观察到与其他类别相比，这两个类别的 F1 相对较低。对于“辅助添加剂”，F1 为 77.06%。这一较低的得分主要是由于碳纤维复合材料文献中对此类实体的描述较少，导致在数据集中该类别的样本数量不平衡，如图3.6所示。在复合材料的研究文献中，通常

会更多地关注主要材料的组成（例如基体、填料等）、性能和加工方法，而对辅助添加剂的详细描述较为少见，这直接影响了模型学习到有效特征的机会，进而影响了识别的准确性和全面性。而对于“工艺数值”类别，F1 达到了 89.76%，在整体数据集中也相对较低。这主要是因为该类别涵盖了多种不同的数据类型，如温度、加工时间、尺寸等，这些不同的工艺参数在文本中的表达方式各异，增加了实体识别的复杂性。

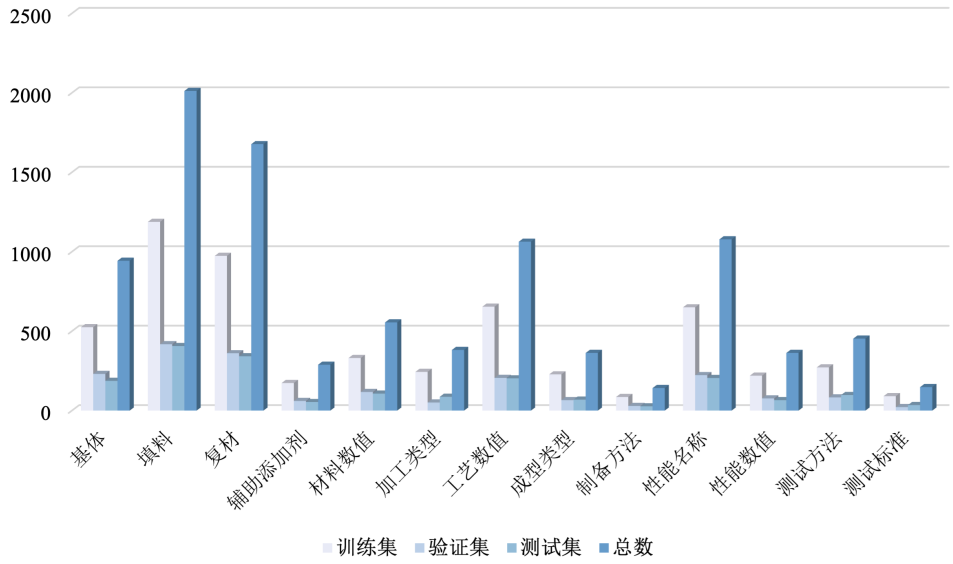


图 3.6　CompMatLitDS 数据集实体分布

（2）MatScholor 数据集实验结果

对于 MatScholor 数据集，考虑到该数据集主要应用于材料领域，本章评估并比较了不同预训练模型在材料科学文本处理任务中的性能，主要包括：Mat2Vec[10]，SciBERT[44]，MatBERT[45]，MatSciBERT[11]，MatTPUSciBERT[46]，以及 SFBC[12]。其中，Mat2Vec[10]方法利用无监督学习，从大规模材料科学文献中进行信息提取，基于词嵌入技术捕捉术语间的语义关系。SciBERT[44]是一种在科学领域的大型语料库上预训练的 BERT 模型。MatBERT[45]方法是一种针对材料科学文献进行优化的 BERT 模型，强调在复杂的材料科学语境中提取有价值的信息；而 MatSciBERT[11]则是专为材料科学领域定制的 BERT 模型，它在材料科学的专业文本上进行了进一步的预训练，以更好地处理行业特定的术语和概念。MatTPUSciBERT[46]是在 TPU 上训练的 SciBERT 版本，专门针对技术和物理更新序列的科学文本，提供更快的处理速度和更好的计算效率。SFBC[12]主要基于动态、静态词向量融合方法，对材料文献中的关

键信息进行提取。这些方法在材料科学领域优化或者在相关数据集上表现出色，通过与这些方法进行比较，可以更好地体现出 SRGN 在特定领域内的优势。

MatScholor 数据集的对比实验结果如表3.5所示。根据实验结果，本文提出的 SRGN 方法达到了 87.91% 的精确率和 89.25% 的召回率，F1 达到了 88.58%。这一结果证明了其在材料科学领域文本处理任务中的有效性。与其它模型相比，Mat2Vec 通过将材料科学文献中的术语嵌入到一个低维空间中，来捕捉词汇之间的语义关系，但由于其本身没有针对上下文关系进行深入学习，导致在复杂文本的处理上精度相对较低，仅达到 80.19%；而 SciBERT 则通过在大规模科学文本上进行预训练，显著提高了对科学文献的理解能力，但它并未专门针对材料科学文献的特殊性进行优化，因此在 MatScholar 数据集上的准确率依然存在提升空间。

表 3.5　不同模型在 MatScholor 数据集上的对比实验结果

模型	Precision (%)	Recall (%)	F1 (%)
Mat2Vec[10]	80.87	79.53	80.19
SciBERT[44]	86.46	86.58	86.52
MatBERT[45]	86.72	88.41	87.56
MatSciBERT[11]	**89.15**	87.63	88.39
MatTPUSciBERT[46]	87.84	88.12	87.98
SFBC[12]	88.47	87.85	88.16
Ours (SRGN) ✓	88.76	**88.40**	**88.58**

相比之下，MatSciBERT 和 MatBERT 都是在材料科学领域专门进行预训练的模型，它们通过在材料文献上进行更多的训练来捕捉专业术语和领域特有的结构。MatSciBERT 的 F1 为 88.39%，比较接近 SRGN 的表现，而 MatBERT 的 F1 为 87.56%，略低于 MatSciBERT。尽管这些模型在材料文本的处理上有所优化，但由于它们没有在捕捉实体之间复杂关系方面进行特殊设计，导致它们在处理较为复杂的实体关系时可能不如 SRGN 模型高效。MatTPUSciBERT 结合了 TPU 加速计算，提高了模型的训练和推理速度，然而这种加速并未显著提高模型在 MatScholar 数据集上的准确率。虽然 MatTPUSciBERT 在准确率上表现良好，但其在召回率上的相对较低值，显示了其在提取一些低频实体或复杂关系时的局限性。SFBC 模型通过引入动、静态词向量融合的方法进行优化，F1 为 88.16%，但未能突破在复杂实体识别和深层次关系建模方面的瓶颈。

与这些模型相比，SRGN 的 F1 达到 88.58%，与 MatSciBERT 相比提高了 0.19%。SRGN 模型不仅优化了实体识别的准确性，还通过引入分块注意力机制和异构图优化了实体之间复杂关系的建模。这使得 SRGN 能够在材料科学文献中，特别是在那些包含多层次实体关系和复杂上下文的文本中，提供更为全面的识别效果。

表 3.6　不同模型在 MatScholor 数据集中不同实体的结果对比

实体类型	Mat2Vec	MatBERT	MSci①	MTPU②	SFBC	**SRGN**
无机材料名称（MAT）	87.65	92.70	92.03	92.20	**92.71**	91.61
样品描述符（DSC）	78.88	88.83	89.71	88.65	**91.32**	90.95
对称/相标签（SPL）	61.38	84.21	81.11	82.52	**84.57**	83.99
材料性能（PRO）	74.47	81.45	81.06	80.63	82.76	**84.89**
应用（APL）	74.89	79.58	84.33	86.22	79.96	**89.02**
合成方法（SMT）	68.85	79.58	83.24	83.49	81.89	**87.91**
表征方法（CMT）	79.08	87.79	**88.95**	86.23	87.81	87.31

① 该方法的完整名称为 MatSciBERT
② 该方法的完整名称为 MatTPUSciBERT

本章对比了 SRGN 模型与其他几种基准模型在 MatScholar 数据集上针对不同实体类型的表现，实验结果如表3.6所示。根据实验结果，SRGN 在大多数实体类别上都展示了出色的性能，并且与其他方模型相比，在不同实体类型上有显著的提升。在“PRO”、“APL”和“SMT”这三个实体类型中，SRGN 分别达到了 84.89%、89.02% 和 87.91% 的 F1，这些结果在所有对比的基准模型中均为最高。而在其他实体类型，如“MAT”和“DSC”等，SRGN 的表现次优，但依然保持了比较好的结果。这表明 SRGN 在识别材料名称和表面处理方法等高度专业性的实体时，能够提供更为精确的识别能力。而与 Mat2Vec 和 MatBERT 等方法相比，SRGN 在处理复杂的材料命名和工艺细节的提取上表现更加突出。Mat2Vec 和 MatBERT 等方法在这些任务中显示出比较低的 F1，尤其在处理领域特定的术语和复杂的实体关系时，SRGN 的上下文理解能力展现了其优势。

3.3.5　消融实验

为了分析不同模块对模型性能的影响，本节对 SRGN 模型进行了消融实验。通过在基模型上逐步增加该方法中的某些组件，在不同数据集上评估了各个模块的贡献。表3.7和表3.8分别展示了在 CompMatLitDS 数据集和 MatScholar 数据集上的消融

实验结果。

首先，根据 CompMatLitDS 数据集的消融实验结果（见表3.7），基模型（即 Graph-NER）的 F1 为 94.75%。当增加分块注意力模块时，模型的性能有小幅度提升，达到 94.95%，这表明分块注意力机制在提取局部特征并改善模型的上下文理解方面具有一定作用。加入异构图结构中的自适应边权重机制和深度可分离卷积后，F1 进一步提高至 95.09%，说明它们能够有效优化不同特征的贡献。加入深度评分网络后，模型性能再次得到提升，F1 达到 95.24%。这一结果表明，深度评分网络提高了 SRGN 对复杂实体的准确识别，增强了在处理复杂数据时的鲁棒性。

表 3.7　SRGN 在 CompMatLitDS 数据集的消融实验结果

模块	Precision (%)	Recall (%)	F1 (%)
基模型	94.65	94.85	94.75
+ 分块注意力	95.69	94.22	94.95
+ 自适应边权重	95.16	94.96	95.06
+ 深度可分离卷积	95.11	95.06	95.09
+ 深度评分网络✓	94.56	95.92	**95.24**

在 MatScholar 数据集上的消融实验（见表3.8）中，基模型的 F1 为 87.92%。加入分块注意力后，模型性能提升至 88.12%。自适应边权重以及深度可分离卷积的引入使得 F1 分别达到 88.22%、88.49%，进一步验证了 SRGN 方法在优化特征贡献方面的重要性。而在引入深度评分网络之后，模型的 F1 达到了 88.58%，这一提升表明深度评分网络在材料文献处理中的应用具有一定的优势。

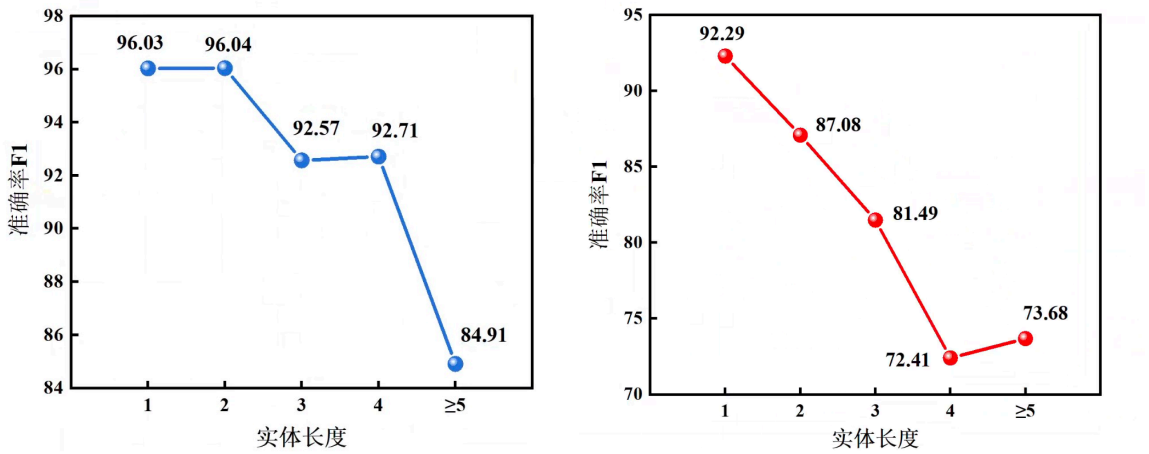
表 3.8　SRGN 在 MatScholor 数据集的消融实验结果

模块	Precision (%)	Recall (%)	F1 (%)
基模型	88.08	87.76	87.92
+ 分块注意力	88.31	87.92	88.12
+ 自适应边权重	88.43	88.01	88.22
+ 深度可分离卷积	88.58	88.40	88.49
+ 深度评分网络✓	88.76	88.40	**88.58**

总体来看，消融实验结果表明，每个模块的引入都能不同程度地提高模型的整体性能，尤其是在复杂特征和多样化数据集的处理上，分块注意力、自适应边权重、

深度可分离卷积和深度评分网络的结合，使得 SRGN 能够更好地捕捉文本中的复杂关系，提升对材料文献的理解和处理能力。消融实验的结果验证了 SRGN 模型的不同模块对其最终性能的贡献，并为未来的研究提供了进一步优化模型架构的依据。

此外，为了评估 SRGN 在处理材料文献文本中的不同类型实体时的表现差异，需要对不同实体长度的实验结果进行分析。因为在材料文献中通常包含很多较长的实体，而长实体则包含更多的上下文信息和复杂的关系，通过对不同实体长度的分析可以发现模型在识别短实体与长实体时的优势和不足，从而为进一步优化模型提供参考。实验结果如图3.7所示。



（a）CompMatLitDS 上实体长度结果　　（b）MatScholor 上实体长度结果

图 3.7　两个材料数据集上实体长度的实验结果。（a）CompMatLitDS 上实体长度结果，（b）MatScholor 上实体长度结果。

CompMatLitDS 数据集上的实验结果如图3.7（a）所示。当实体长度为 1 时，SRGN 模型的 F1 为 96.03%，这表明对于短小实体的识别，模型能够实现非常高的准确度。对于长度为 2 的实体，F1 稍有提升，达到 96.04%。然而，对于数据集中长度为 3 和 4 的实体，模型的 F1 分别降至 92.57% 和 92.71%，显示出随着实体的复杂度增加，模型的识别能力有所下降。对于长度 ≥5 的实体，F1 进一步下降至 84.91%。这表明 SRGN 模型在处理较短实体时表现优异，但在处理更长、更复杂的实体时，相对于处理长度 ≤2 的短实体而言，该模型可能面临更多的上下文依赖和结构复杂性，从而影响了其识别性能。

在 MatScholar 数据集上，不同实体长度的实验结果如图3.7（b）所示。SRGN 模型在长度为 1 的短实体上的表现也最为突出，F1 为 92.06%。当实体长度增加时，模

型的 F1 逐渐下降。具体而言，长度为 2 的实体的 F1 为 86.12%，长度为 3 的实体为 83.37%，长度为 4 的实体为 77.35%。当实体长度为 ≥5 时，F1 降至 75.98%。这些结果表明，SRGN 在短实体的识别上表现较为优秀，而在处理较长实体时，模型的表现受到一定影响，可能是由于长实体涉及到更多的上下文关系和实体间的依赖，增加了模型学习和推理的复杂性。

3.4 本章小结

针对复合材料文献以及长序列处理问题，本章提出了 SRGN，用于从复合材料文献中挖掘材料组成、加工工艺及性能测试等关键信息。该模型通过分块注意力机制降低长文本计算复杂度，引入融合了上下文与实体类型信息的异构图结构，结合深度可分离卷积和自适应动态边权重机制优化语义连接，显著提升了实体识别与复杂关系建模能力。实验表明，SRGN 在 CompMatLitDS 数据集和 MatScholar 数据集上分别达到了 95.24% 和 88.58% 的 F1，验证了所提模型的有效性。在消融实验中，通过对不同模块的增加，体现了引入分块注意力机制、深度可分离卷积、动态边权重机制以及深度评分网络的重要性。在不同实体长度的实验分析中，SRGN 表现出了在短实体识别上的强大能力，但随着实体长度的增加，尤其是在处理更长的实体时，模型的表现有所下降。这表明在处理复合材料文献中复杂的长实体时，方法的性能受到实体内部和实体间复杂关系的影响。在第四章中将进一步优化模型的上下文建模能力，提高模型在处理长实体时的表现。这将有助于提升 SRGN 在多样化实体类型上的广泛适应性和准确性。

第四章　基于多粒度融合的材料命名实体识别

第三章提出的 SRGN 模型成功实现了针对复合材料领域文献的命名实体识别，该模型通过精确地提取材料文本数据中的关键信息，显著提高了信息处理的效率和准确性，能够从材料科学文献中有效捕捉到组成元素、性能数据等关键特征，为复合材料的设计提供了数据基础。然而，实验表明 SRGN 在长实体识别方面存在不足，这也是材料文献挖掘中普遍面临的挑战。通用材料文献中同样可能包含由多个词组成的长实体，如具体的化学配方或复杂的材料处理工艺描述。如何正确划定长实体的边界、准确识别较长的实体是一项挑战。因此，本章针对长实体识别以及准确区分实体边界的问题，对 SRGN 进行优化，使其能有效识别各种材料科学文献中的多样化实体，应用于通用材料领域的文献挖掘任务。

4.1　方法概述

尽管 SRGN 在复合材料文献挖掘中表现良好，但在长实体识别与跨领域迁移方面仍存在局限。为解决上述问题，本章提出了多粒度融合图网络（Heterogeneous Cross-grained Graph Network, HCG），应用于通用材料文献中的命名实体识别任务。该模型引入可学习的门控机制、跨粒度交互注意力机制，以及结合对比学习与 CRF 损失的联合训练策略，对 SRGN 进行了有效改进。通过门控机制和交互注意力机制，增强了模型对多粒度语义信息的建模能力；通过引入对比学习与 CRF 损失进行联合训练，优化了对实体边界的识别，从而提升了模型在长实体识别与复杂上下文理解中的性能表现。HCG 的整体结构如图 4.1 所示。

HCG 模型在多个方面进行了改进。首先，在编码器部分，针对输入的三种不同粒度的向量，本章引入了一种可学习的门控机制。由于直接拼接这些粒度可能会导致特征冗余，因此通过门控机制，每个粒度的信息能够根据上下文进行加权，以确保各粒度特征在网络中贡献的权重是动态调整的。其次，SRGN 中的分块注意力机制在解决长序列问题上虽然有效，但未能显著提高对长实体边界区分和识别准确率。为此，本章提出了一种基于传统多头注意力机制改进的交互注意力（Talking Attention），作为跨粒度的注意力机制，使得不同粒度之间的信息能够更加有效地交互与融合，从

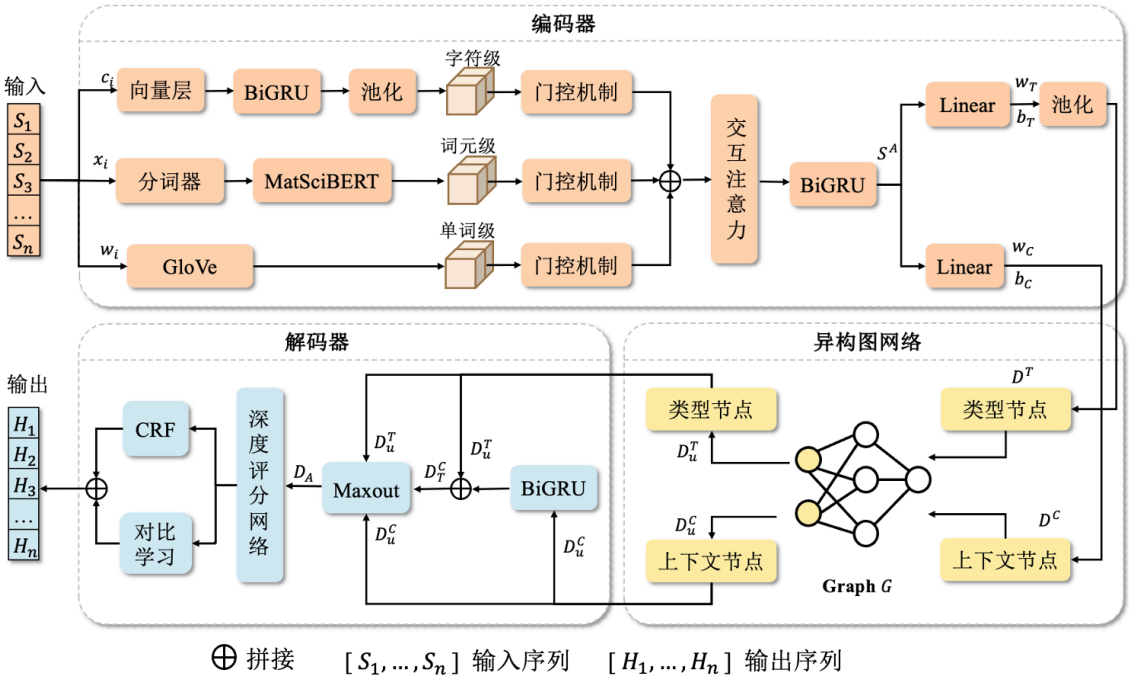


图 4.1　多粒度融合图网络（HCG）模型结构图

而提高长实体的识别精度。最后，在解码器部分，本章引入了对比学习的策略，并在训练过程中采用了对比损失与 CRF 损失的联合训练方法。这种联合训练方式不仅弥补了 CRF 在捕捉细粒度语义区分方面的不足，还有效解决了通用材料领域下长实体识别的难题。通过对比学习和 CRF 损失的结合，模型能够更好地识别和区分不同类型的实体，尤其是在面对复杂实体时表现得更精准。

HCG 通过以上改进，在通用材料文献挖掘中展现出了更强的性能和适应性。接下来，本章将详细介绍各个模块的设计细节，并通过实验验证所提出方法的有效性和优势。

4.1.1　多粒度融合模块

在材料科学文献分析中，文本特征通常包含字符级的化学符号模式、词级的专业术语语义以及上下文级的材料属性关联。大多数的方法都是通过直接拼接融合多粒度特征，即对字符级、单词级和词元级向量进行简单拼接。然而，这种处理方式可能在材料文本处理中面临严峻挑战。不同粒度特征的分布空间存在显著差异：字符级向量关注元素符号的组合模式，例如”Fe-Cr” 中的短横连接符号；单词级向量反映材料学术语定义，例如”austenite” 的晶体结构语义；而词元向量主要是通过 BERT 系列的语言模型，捕捉材料属性间的长程依赖。采用直接拼接的方法可能会带来特

征冗余问题，因为不同粒度的信息具有不同的语义重要性和信息丰富度，简单的拼接无法有效区分这些差异。此外，这种方法在处理文本数据时也很难精确地控制信息流的重要性，导致模型在长实体的边界识别和精确信息捕捉方面表现不佳。此外，固定融合策略无法适应材料文本的多样性，例如在解析化学方程式时需要强化字符特征，而在理解材料制备工艺时需依赖上下文关联。

为了解决上述问题，本章引入了一个基于可学习的门控机制的改进方案。该方法的核心思想是通过轻量级参数化门控网络，自适应调节不同粒度特征的融合权重，使模型能够根据内容的语义重要性自适应地调整各粒度信息的贡献。每种粒度的输入向量都通过一个独立的门控单元处理。门控单元包括一个线性变换和一个 Sigmoid 激活函数。相较于固定拼接方式，本方案具有三个优势：首先，通过门控实现特征空间校准，缓解分布差异；其次，动态权重分配可捕捉不同文本片段的语义侧重；最后，门控网络仅引入少量参数，可以避免维度爆炸问题。

具体来说，H_c、H_w、H_t 分别表示字符级、单词级、词元级向量，对于每个特征 $H_m \in \{H_c; H_w; H_t\}$，可学习的门控 g_m 计算重要性权重，如以下公式所示：

$$g_m = \sigma(W_m \cdot H_m + b_m), \tag{4.1}$$

其中，g_m 表示可学习的门控机制针对不同粒度特征的权重；σ 是 Sigmoid 激活函数，用于将加权结果映射到（0,1）区间，确保输出的权重具有合理的范围；W_m 和 b_m 分别表示线性变换中可学习的参数，即权重矩阵和偏置。

随后，计算得到的门控权重 g_m 与对应粒度的输入向量 H_m 进行元素乘法，将原始输入向量根据其重要性进行缩放，从而使得重要的特征得到强调，不重要的特征被抑制，如以下公式所示：

$$H'_m = g_m \otimes H_m, \tag{4.2}$$

其中，$\otimes$ 表示元素相乘。所有经过门控处理的粒度特征向量 H'_m 被拼接或相加，形成最终的特征表示。

该门控机制的可学习性主要体现在两个方面。一方面，门控模块中的线性变换参数在训练过程中是可优化的，模型能够通过数据驱动的方式自动学习每种粒度特征的最优权重分配，从而实现更精确的信息融合。另一方面，该机制具备良好的上下文自适应能力，能够根据不同输入样本的语义特征动态调整各粒度信息的贡献大

小，使模型在面对不同类型、不同结构的材料文献时具备更强的灵活性和表达能力。这样的方法可以基于上下文语义进行动态加权，增强了模型对长实体的识别能力。

4.1.2 交互注意力机制

在通用材料领域的文本分析任务中，长实体识别（如复杂化学物质名称、复合材料结构描述等）面临显著挑战。传统序列标注模型依赖局部上下文窗口，难以捕捉长距离依赖；而标准 Transformer 模型虽能建模全局关系，但其多头注意力机制中各个注意力头独立计算，导致头间信息割裂，限制了模型对跨多语义单元的长实体的联合推理能力。针对这一问题，本研究提出一种基于多头注意力改进的交互注意力机制，通过动态融合不同注意力头的语义聚焦模式，增强模型对长实体边界及内部结构的感知能力。

交互注意力的引入来源于对话交互注意力[47]（Talking Heads Attention），其核心思想在于引入可学习的跨头投影层，动态调整注意力头之间的交互模式，从而突破传统机制中头的严格独立性限制。因此，本章节结合了 Talking Heads Attention 的设计思想，并在传统的多头注意力上，针对训练稳定性与计算效率进行了改进。交互注意力机制的结构如图4.2所示。

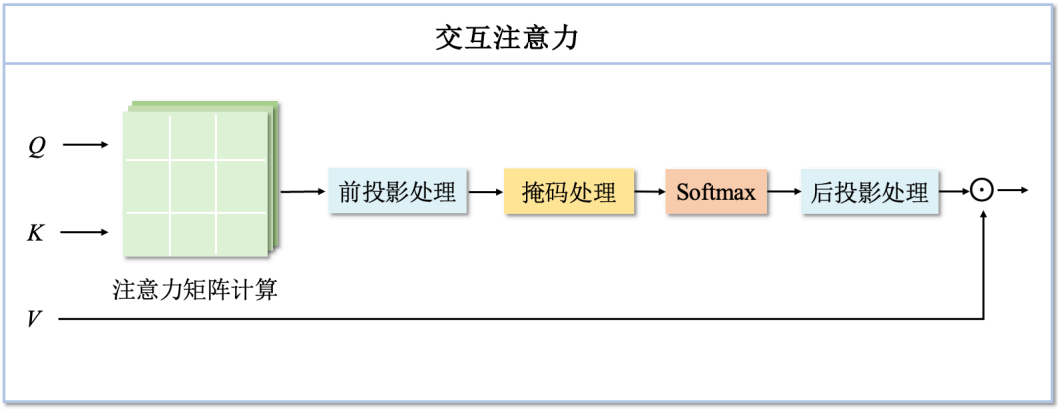


图 4.2 交互注意力模型结构图

对于本节的交互注意力机制，首先，在投影层初始化策略上，采用 Dirac 初始化方法将卷积核权重近似为单位矩阵，确保训练初期投影层近似恒等变换，从而保留传统多头注意力的初始行为模式，有利于梯度传播的稳定性。其次，采用双向投影设计，模型可以动态混合不同注意力头的分数矩阵。这打破了传统多头注意力中头的独立性，允许多个头之间共享信息。另外，在投影层实现方式上，将原始的全连接

层替换为 1×1 卷积操作，利用卷积核的权重共享特性降低参数复杂度。此外，为进一步增强注意力权重的泛化能力，在投影层后引入 Dropout 操作，缓解过拟合风险。

具体来说，在4.1.1的多粒度融合之后，融合得到的向量首先通过线性变换生成查询（Query）、键（Key）、值（Value）向量，并拆分为 H 个注意力头。与传统多头注意力直接计算头间独立注意力分数不同，交互注意力在 Softmax 归一化前后分别插入可学习的跨头投影层。在计算初始注意力分数矩阵 A 后，首先通过预投影矩阵 $W_{pre} \in \mathbb{R}^{H\times H}$ 对头维度进行线性组合，生成跨头增强的注意力分数。该操作可视为对不同注意力头关注模式的动态加权融合，使模型能够自动学习材料文本中化学成分、物理属性等不同语义维度间的关联强度，如以下公式所示：

$$A = \frac{Q \cdot K^T}{\sqrt{d}}, \tag{4.3}$$

$$A' = W_{pre} \cdot A, \tag{4.4}$$

随后，对 A' 进行 Softmax 归一化得到注意力权重 S，再通过后投影矩阵 W_{post} 进行二次校正，最终生成融合多粒度语义信息的注意力分布 S'：

$$S' = W_{post} \cdot \text{Softmax}(A'), \tag{4.5}$$

整个交互注意力的过程如以下公式所示：

$$Attention(Q,K,V) = \text{Proj}\left(W_{post} \cdot \underbrace{\text{Softmax}\left(W_{pre} \cdot \frac{QK^T}{\sqrt{d}}\right)}_{\text{跨头交互注意力权重}} \cdot V\right), \tag{4.6}$$

其中，$\text{Proj} : \mathbb{R}^{d_{model}} \to \mathbb{R}^{d_{model}}$ 表示投影层，将多头输出重映射至统一语义空间，确保与后续模块的兼容性。

相较于原始 Talking Heads Attention 机制的实现，本研究针对材料领域文本特性进行了两项关键改进：

（1）采用 Dirac 初始化策略，将 W_{pre} 和 W_{post} 初始化为单位矩阵的近似形式，确保训练初期近似等价于标准注意力机制，缓解冷启动阶段参数敏感性问题；

（2）设计维度置换投影架构，将原始实现中的二维卷积操作替换为维度置换 + 线性投影组合，通过将注意力分数矩阵从 $[B,H,L,L]$ 调整为 $[B,L,L,H]$ 后应用全连接层，提升 GPU 内存访问效率。

交互注意力机制通通过引入轻量级的跨头投影层，在保持多头注意力并行计算优势的同时，显著提升了模型对复杂依赖关系的建模能力。这样的方式优化了各粒度之间的信息交互，提升内存访问效率，但在训练时可能增加少量的计算开销。

4.1.3 对比学习

在解码器结构中，尽管 CRF 在序列标注任务中广泛应用，并具备良好的结构建模能力，但其对细粒度语义的建模能力仍存在一定局限。尤其在面对复杂结构文本如通用材料文献时，由于序列本身为 BIO 标注，实体内部标签间的语义关联未被显式建模，容易因局部噪声导致边界判定错误。另外，CRF 在训练过程中仅依赖于标签序列的似然最大化，忽略了标签之间的全局语义距离分布，不利于模型在跨领域或实体多样性较高的任务中保持鲁棒性。因此，为进一步增强模型对实体边界的感知能力，尤其是提升对长实体识别与区分的能力，本章在解码器结构中引入了对比学习，在标签语义空间中拉近同类实体标签的距离，同时推远非实体标签，增强模型对实体内部一致性的感知能力，并通过与 CRF 的联合损失训练提升模型表现。

在模型对每个候选实体位置进行标签预测时，首先计算对应的标签分数，随后通过一个共享的投影层将这些标签分数映射至一个统一的低维对比空间。设给定批次样本的发射分数 $E \in \mathbb{R}^C$，其中 C 为标签数，则投影可以表示为：

$$H = \text{Proj}(E) = E \cdot W_C, \tag{4.7}$$

其中，$W_c \in \mathbb{R}^{C \times d}$ 是一个可学习的对比投影矩阵，d 是对比空间的维度。

在该对比空间中，模型采用以下约束策略：（1）将同一实体内部的起始标签与后续的标签作为正样本对；（2）将标签为 O 的位置作为负样本，通过采样方式与锚点（B 标签）进行区分。这样做的目标是使得锚点与其正样本在表示空间中尽可能靠近，而与负样本尽可能远离。具体对比损失定义如下：

$$\mathcal{L}_{contrast} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(h_i^\top h_i^+/\tau)}{\exp(h_i^\top h_i^+/\tau) + \sum_{j=1}^{M} \exp(h_i^\top h_j^-/\tau)}, \tag{4.8}$$

其中，h_i 为锚点向量，h_i^+ 为正样本，h_j^- 为负样本；τ 为温度参数，用于缩放点积；N 为所有锚点数量，M 为每个锚点的负样本采样数。

训练过程中，最终采用联合损失函数优化模型，综合对比学习损失和 CRF 损失

(即3.2.4小节所提到的对数似然损失):

$$\mathcal{L}_{total} = \mathcal{L}_{CRF} + \lambda \cdot \mathcal{L}_{contrast}, \tag{4.9}$$

其中，λ 为平衡两个损失项的超参数，通常取值较小，用于稳定训练。

另外，该设计使得模型在优化 CRF 标签转移概率的同时，显式约束实体内部标签的语义连续性。为提升计算效率，算法采用负采样策略，每个锚点需计算与随机负样本的对比损失。对于整个基于对比学习与 CRF 联合损失的训练过程，具体实现流程如算法4.1 所示。首先从标注数据中提取 B 标签位置作为锚点，遍历其后续连续 I 标签构建正样本对，随后在 O 标签中随机采样构建负样本对，最终基于交叉熵损失优化对比空间中的样本分布，与 CRF 进行联合损失训练。

算法 4.1: 基于对比学习与 CRF 的联合损失训练过程

输入: 发射得分 $\mathcal{E}$，标签序列 $\mathcal{T}$，掩码 $\mathcal{M}$，温度系数 τ，权重系数 λ

输出: 联合损失 $\mathcal{L}_{\text{total}}$

$\mathcal{H} \leftarrow \text{Proj}(\mathcal{E})$ // 将发射得分映射到对比空间

$\mathcal{A} \leftarrow \text{GetAnchor}(\mathcal{T} = \text{B}, \mathcal{M})$ // 提取 B 标签作为锚点

$\mathcal{P} \leftarrow \text{GetPositive}(\mathcal{T} = \text{I})$ // 匹配同实体下的 I 标签作为正样本

$\mathcal{N} \leftarrow \text{SampleNeg}(\mathcal{T} = \text{O}, k)$ // 从 O 标签中采样负样本

if $\mathcal{A} = \emptyset$ ***or*** $\mathcal{P} = \emptyset$ **then** // 无正负样本则跳过

  $\mathcal{L}_{\text{contrast}} \leftarrow 0$

else

  $S^+ \leftarrow \text{Sim}(\mathcal{A}, \mathcal{P})/\tau$ // 计算锚点与正样本相似度

  $S^- \leftarrow \text{Sim}(\mathcal{A}, \mathcal{N})/\tau$ // 计算锚点与负样本相似度

  $L_{\text{contrast}} \leftarrow \text{CrossEntropy}(\text{Concat}(S^+, S^-), \mathbf{0})$ // 对比损失

end

$\mathcal{L}_{\text{CRF}} \leftarrow \text{CRF}(\mathcal{E}, \mathcal{T}, \mathcal{M})$ // CRF 条件似然损失

$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CRF}} + \lambda \cdot \mathcal{L}_{\text{contrast}}$ // 联合损失

return $\mathcal{L}_{total}$

本小节在解码器的结构中引入对比学习，由投影层将标签分数映射至对比空间

后，与 CRF 损失函数共同进行联合损失计算。该设计既保留了 CRF 的序列建模优势，通过对比学习增强对长距离依赖的建模能力，能有效区分相邻实体的边界。

4.2 实验与讨论

为验证所提出的 HCG 在通用材料领域任务中的有效性，本章在多个公开数据集上开展了系统的实验研究。本章将对实验设置与结果进行详细说明和分析。首先，介绍所使用的实验数据集以及实验环境，并说明模型训练中的关键参数设置及评价指标；随后，分别开展与当前主流方法的对比实验，以验证 HCG 的综合性能；最后，通过消融实验深入探究模型各个组成模块对整体性能的影响，进一步印证本文方法的设计合理性与有效性。

4.2.1 数据集介绍

在本章的实验中，HCG 主要面向通用材料文献领域的命名实体识别任务，因此，实验选用了 Matscholor 数据集[10]，该数据集专门针对无机材料文献的文本数据进行设计，涵盖了包括材料名、化学元素、物理属性等多个实体类别。具体而言，Matscholor 数据集包含了 7 种实体类型，详细的实体类别及其数据集划分可参考第3.3.1节的介绍。该数据集不仅能够很好地展示我们方法在材料文献领域的应用效果，还能够为后续的研究提供一个标准化的测试平台。然而，考虑到本章方法的通用性和更广泛的适用性，本研究还选择了两个具有代表性的命名实体识别数据集——CoNLL2003 数据集[48]和 GENIA 数据集[49]进行实验，以验证所提出方法在其他领域的表现。

GENIA 数据集是通用于生物医学领域中一个成熟且广泛使用的语料库，它的创建旨在支持分子生物学领域信息提取和文本挖掘系统的开发和评估。GENIA 包含了来自 PubMed 的 2000 篇摘要，并介绍了四种主要的实体类型：蛋白质（protein）、DNA、RNA 和细胞类型（cell type）。该数据集具有较高的标注质量和领域特性，是生物医学命名实体识别任务中一个标准的评测集。该数据集关于训练集、验证集、测试集的统计信息如表4.1所示。

CoNLL2003 数据集则是一个涵盖了多个领域的广泛使用的命名实体识别基准数据集。该数据集由新闻文章组成，标注了四种实体类别：人名（per）、组织（org）、地点（loc）和其它项（misc）。CoNLL2003 数据集因其语料丰富、标注明确而成为命名实体识别研究中的经典数据集，广泛用于验证各种 NER 模型的效果。该数据集关

表 4.1　GENIA 数据集的统计信息

信息	训练集	验证集	测试集
句子数量	14041	3453	3250
实体数量	23499	5648	5942

于训练集、验证集、测试集的统计信息如表4.2所示。

表 4.2　CoNLL2003 数据集的统计信息

信息	训练集	验证集	测试集
句子数量	15203	1854	1669
实体数量	46142	5506	4367

在实验中，对于每个数据集，都按照原始数据集的划分方法进行训练集、验证集和测试集的分配，保证实验结果的可靠性，同时避免过拟合的风险。

4.2.2　实验环境

本章的所有实验均在以下环境中进行。首先，实验使用的操作系统为 Ubuntu 18.04.5, 所有代码均在 Python 3.9.13 环境下运行。深度学习框架选择了 PyTorch 1.12.1 作为主要开发和训练工具。在硬件方面，所有实验均在一台搭载 NVIDIA RTX3090 24GB 的图形处理单元上运行。此外，实验还使用了配置为 Intel(R) Core(TM) i9-10900K @2.40GHz 的中央处理器，确保在数据预处理、模型训练等任务中具有足够的计算资源。

4.2.3　参数设置以及评价指标

在本章的实验中，对于专门针对通用材料文献领域的 Matscholor 数据集，采用了 MatSciBERT 作为预训练模型，以准确获得材料领域的词元级信息，并设置句子的最大长度为 256，确保有效处理材料文献中较长的文本信息，捕捉到更多的上下文特征。而在处理通用领域的 CoNLL2003 和 GENIA 数据集时，分别选择了 BERT 和 BioBERT 作为预训练模型，以获得领域特定的词元表示。对于这两个数据集，句子的最大长度设置为 128，以适应数据集的文本长度和模型的计算需求。方法中对比学习的部分，对于三个数据集，均设置温度系数为 1.0，损失函数中的 λ 均设置为 0.5。

在所有实验中，迭代轮数设置为 30。训练使用了 AdamW 优化器，用于减小模型训练过程中的过拟合现象。学习率设置为 2e-5，最大梯度归一化设置为 1e0，这有助于加速模型的收敛，并保持梯度的稳定性。为了进一步提升训练效果，采用了线性预热衰减学习率调度来逐步调整学习率，从而提高模型在训练初期的稳定性并避免梯度爆炸。

在模型参数方面，训练时异构图的迭代次数设置为 6，这一设置帮助模型在多个粒度层次间进行有效的知识融合，进一步提高模型的表达能力。同时，交互注意力所使用的注意力头数量被设置为 6，保证注意力机制能够充分捕捉到不同实体间的细粒度关系。

在评价指标方面，实验时采用了 NER 中常用的指标，包括精确率（Precision）、召回率（Recall）和 F1。这些指标能够全面衡量模型在不同实体类别上的识别能力，并为模型的性能比较提供了可靠的依据。这三项评价指标的详细信息可以参考第3.3.3小节的介绍。

4.2.4 对比实验

为了验证 HCG 的有效性与通用性，本节在三个具有代表性的数据集上进行了广泛的对比实验，分别是针对通用材料领域的 MatScholor 以及通用领域的 CoNLL2003 和 GENIA 数据集。为确保评价的全面性和客观性，实验采用了精确率（Precision）、召回率（Recall）和 F1 作为评价指标。具体实验过程和结果将在以下部分详细介绍。

（1）MatScholor 数据集对比实验

在本节的对比实验中，针对 MatScholor 数据集，评估了多种现有的命名实体识别模型，包括 Mat2Vec[10]、MatBERT[45]、MatSciBERT[11]、MatTPUSciBERT[46]、SFBC[12]、GraphNER[35] 以及 PIQN[41]。这些模型大多数是针对材料领域开发的，能够从材料文献文本中提取与工艺、性能等相关的关键信息。另外，为了进一步验证 HCG 的优势，还与第三章提出的 SRGN 进行了对比。所有这些模型的实验结果将与 HCG 进行详细对比，以验证在 NER 任务中的有效性和通用性。

在 MatScholor 数据集上进行的对比实验结果如表4.3所示。结果表明，HCG 的表现均优于对比的几种模型。具体来说，HCG 的精确率、召回率和 F1 分别达到了 88.90%、88.60% 和 88.70%。相较之下，表现最好的对比模型 MatTPUSciBERT 在 F1 上为 87.98%，而 Mat2Vec 和 MatBERT 分别为 80.87% 和 86.72%。这些结果表明，

HCG 在 NER 中表现出色，能够在材料文献领域中更好地捕捉和识别关键信息。

表 4.3　不同模型在 MatScholor 数据集上的实验结果

模型	Precision (%)	Recall (%)	F1 (%)
Mat2Vec[10]	80.87	79.53	80.19
MatBERT[45]	86.72	<u>88.41</u>	87.56
MatSciBERT[11]	**89.15**	87.63	88.39
MatTPUSciBERT[46]	87.84	88.12	87.98
SFBC[12]	88.47	87.85	88.16
GraphNER[35]	88.08	87.76	87.92
PIQN[41]	85.72	83.97	84.84
SRGN	87.91	**89.25**	<u>88.58</u>
Ours (HCG) ✓	<u>88.05</u>	**89.25**	**88.65**

首先，与传统的基于嵌入方法的模型相比，HCG 方法在所有指标上均有显著提升。Mat2Vec 使用的是传统的词嵌入方法，缺乏深层次的上下文理解，而 HCG 通过引入可学习的门控机制来融合多粒度特征，从而能够更有效地捕捉长实体的边界信息，这一点尤其在材料文献中至关重要。Mat2Vec 的结果显示其召回率较低，为 79.53%，说明该方法未能充分识别长实体或复杂实体之间的关系。与 MatBERT、MatSciBERT 等基于 BERT 的方法相比，HCG 仍然展现了较为明显的优势。MatBERT 和 MatSciBERT 都是针对材料文献领域的预训练模型，尽管它们在理论上能够更好地理解材料领域的文本，但它们仍然受到传统 BERT 模型局限性的影响，尤其在长实体识别和细粒度的边界区分上存在不足。HCG 在这方面的优势得益于其引入的多头注意力机制，这一机制能够在实体识别过程中处理更为复杂的上下文信息，并且在解码器部分联合优化了对比学习和 CRF 损失，从而提升了模型的边界区分能力和整体识别性能。

GraphNER 作为 HCG 与第三章 SRGN 的基模型，虽然通过图结构捕捉实体间关系，但由于缺乏足够的特征融合机制和高效的注意力机制，实验结果相对较弱，尤其是在处理长实体或复杂结构时，未能充分发挥图网络的潜力。相比之下，HCG 在这方面通过引入改进的交互注意力机制有效地提升了模型对不同粒度特征的融合能力，使得长实体和复杂边界的识别能力得到了显著增强。SFBC 通过动静态词向量融合的方式进行 NER，但是根据实验结果可以看到对于一些特有的长实体识别方面表

现不足，仅依赖标签转移概率，采用 CRF 进行实体识别，易受局部噪声影响。PIQN 方法虽在通用领域的命名实体识别任务中表现出一定优势，但对于通用的材料领域特有的任务上未能充分优化。

（2）CoNLL2003 数据集对比实验

为验证本章提出的 HCG 模型在通用领域的适应性，进一步在基准 CoNLL2003 数据集开展对比实验。该数据集涵盖新闻语料中的人名、地名、组织机构名等实体，虽未明确标注长实体比例，但其细粒度标注特性可有效检验模型的泛化能力。在实验中，本节选择了多个近年来在该数据集上表现优异的模型进行对比，包括基于阅读理解的 BERT-MRC[50]、双仿射标注框架 BiaffineNER[51]、基于预训练生成的 Template-BART[52] 与 BARTNER[53]、异构图网络方法 GraphNER[35]、并行标注框架 PIQN[41]、扩散模型 DiffusionNER[54] 以及提示学习框架 PromptNER[40]。这些模型分别从特征交互、解码机制或生成式架构等不同角度优化实体识别，近年来在 NER 上都取得了较好的成绩，但很少有方法显式关注长实体边界与识别的问题。

对比实验结果如表4.4根据实验结果，HCG 在 F1 值上表现优异，达到了 92.90%，超越了其他的模型。F1 值是衡量模型精确度和召回率平衡的综合指标，从实验结果可以看到，HCG 能在这两者之间找到较好的平衡，从而使得其 F1 值表现出色。

表 4.4　不同模型在 CoNLL2003 数据集上的实验结果

模型	Precision (%)	Recall (%)	F1 (%)
BERT-MRC[50]	92.47	<u>93.27</u>	<u>92.87</u>
BiaffineNER[51]	92.85	92.15	92.50
Template-BART[52]	91.72	**93.40**	92.55
BARTNER[53]	92.31	93.45	92.88
GraphNER[35]	92.55	92.43	92.49
PIQN[41]	**93.29**	92.46	<u>92.87</u>
DiffusionNER[54]	92.99	92.56	92.78
PromptNER[BERT-large][40]	92.48	92.33	92.41
Ours (HCG) ✓	<u>93.19</u>	92.60	**92.90**

与 BERT-MRC、BiaffineNER 等传统 BERT 基础的模型相比，HCG 不仅在精确度上有所提升，而且在召回率和 F1 上的提升更为显著。BERT-MRC 在 F1 上为 92.87%，但其模型对于长实体的识别和边界的精准区分仍存在不足，而这正是 HCG 的优势所

在。HCG 引入了可学习的门控机制来融合多粒度特征，从而使得模型在处理长实体边界和复杂实体关系时，能够获得更精确的定位与识别。

相比于基于异构图网络的 GraphNER，HCG 在图结构的基础上进一步引入了改进的交互注意力机制，通过多头注意力机制优化了模型对实体间关系的捕捉能力。这使得 HCG 在 F1 值上较 GraphNER 的 92.49% 有了明显提升（+0.41%），尤其是在复杂关系和长实体的边界处理方面，HCG 能够有效克服 GraphNER 未能处理好的问题。GraphNER 在处理长实体时可能未能充分捕捉到长实体之间的细微关系，而 HCG 通过多粒度特征融合、交互注意力机制和对比学习优化，使得其 F1 在同类方法中处于领先地位。对于 DiffusionNER 和 PIQN，它们虽然能够提高精确度和召回率，但它们可能偏重于某一方面，导致在 F1 的综合表现上无法达到 HCG 的平衡。尤其是对于长实体的识别和细粒度实体边界的处理，HCG 的交互注意力和对比学习的结合使得其在整体表现上更为均衡，在 CoNLL2003 数据集上表现出了较其他对比模型更为优秀的性能，证明了 HCG 在通用领域中的高效性与通用性。

（3）GENIA 数据集对比实验

本小节也在 GENIA 数据集上进行了广泛的对比实验，以验证 HCG 的通用性与有效性。实验时，选择了近年来在该数据集上取得较好成绩的对比模型，包括 BARTNER[53]、LogSumExpDecoder[55]、PO-TreeCRFs[56]、GraphNER[35]、Latent Lexicalized Constituency Parsing[57]、Span-level Graph[58]、Pointer Networks[59]、SpanPred+SEQ[60] 和 UniversalNER[61]。这些模型都在 GENIA 数据集上展示了不同的优势，采用了各自特有的架构和策略，例如 BARTNER 采用了 BART 模型进行编码，GraphNER 和 Span-level Graph 方法则利用了图网络结构进行实体识别；Pointer Networks 指针网络是基于选区解析的嵌套实体识别框架；SpanPred+SEQ 方法融合跨度预测与序列标注；Hashing 则采用结构化对比哈希方法进行命名实体识别；UniversalNER 是基于指令调优用于命名实体识别的大模型。

实验结果如表4.5所示，HCG 的表现比较突出，达到了 79.85%，相比于其他对比模型展现了明显的优势。同时，结合精确率与召回率的实验结果，也表明 HCG 在精确度和召回率之间找到了良好的平衡，并且能够有效提升实体识别的整体性能。

首先，BARTNER 在 GENIA 数据集上取得了 79.23% 的结果，它使用 BERT 的变体模型 BART 进行文本编码，能够对输入文本进行有效的建模，但它未能充分应对长实体和复杂边界的识别问题。相比之下，HCG 通过引入可学习的门控机制来融

表 4.5 不同模型在 GENIA 数据集上的实验结果

模型	Precision (%)	Recall (%)	F1 (%)
BARTNER[53]	78.87	<u>79.60</u>	79.23
LogSumExpDecoder[55]	79.20	78.67	78.93
PO-TreeCRFs[56]	78.20	78.20	78.20
GraphNER[35]	<u>80.81</u>	78.71	<u>79.74</u>
LLCP[57] ①	78.39	78.50	78.44
Span-level Graph[58]	77.92	**80.74**	79.30
Pointer Networks[59]	78.08	78.26	78.16
SpanPred+SEQ[60] *②	-	-	79.20
UniversalNER[61] *	-	-	77.54
Hashing[62] *	-	-	78.79
Ours (HCG) ✓	**81.67**	78.11	**79.85**

① 该模型的全称是 Latent Lexicalized Constituency Parsing
② * 表示来自原文献的实验结果

合多粒度特征，能更好地处理长实体的边界问题，并且能够在实体间的细粒度关系上取得更好的表现，显著优于 BARTNER（+0.62%）。

对于采用图网络的 GraphNER 与 Span-level Graph 而言，这两个方法的实验结果分别达到了 79.74% 和 79.30%。GraphNER 通过图结构建模实体间的关系，尽管在某些情况下能够有效捕捉实体间的相互作用，但它在长实体边界的精准定位和复杂关系的建模上表现较弱。Span-level Graph 虽然能够识别更多的实体，但它在避免错误识别方面的能力较弱。HCG 通过引入基于多头注意力改进的交互注意力，可以在多个粒度层次之间有效地融合特征，并且通过对比学习优化了模型的解码器部分，使得 HCG 能够更精确地捕捉长实体的边界以及平衡精确率与召回率的结果。Pointer Networks 的实验结果达到 78.16%，而 HCG 与之相比提升了 1.69%。这可能是由于在处理生物医学长实体时，Pointer Networks 采用了结构一致性约束，虽然能够避免实体跨度的交叉重叠，却限制了模型对部分重叠实体的灵活建模能力。

HCG 对比 SpanPred+SEQ 模型（F1=79.20%），提升了 0.65%。该方法通过融合序列标注（SEQ）、条件随机场（SeqCRF）与跨度预测（SpanPred）三类基础方法，利用多数投票或学习组合策略提升性能。在生物医学场景中，该方法的混合设计虽能覆盖不同粒度的实体线索，但其本质仍是静态特征组合，缺乏对多粒度交互的动态建模。并且当 SpanPred 与 SEQ 对同一跨度的实体类型预测不一致时，组合机制无

法有效消解歧义。

Hashing 提出结构化对比哈希方法，引入位级 CKY 算法优化结构化预测。该方法在通用结构化任务中展现出潜力，但根据实验结果可能存在计算复杂度随实体长度呈指数增长的问题，这些会导致 Hashing 在面对长实体时准确率较低。而 UniversalNER 是一种利用大模型进行实体识别的方法，通过指令调优从 ChatGPT 蒸馏得到，旨在实现开放域任意实体类型的识别，但指令驱动的泛化能力高度依赖预设的实体类型描述，而 GENIA 中包含生物医学的名词，缺乏通用语义解释，导致模型混淆功能描述与实体边界。相比于 UniversalNER，HCG 提升了 2.31%。

对于 LogSumExpDecoder、PO-TreeCRFs 和 Latent Lexicalized Constituency Parsing 这三个方法，虽然它们在 F1 值上取得了一定成绩，但它们的表现均未超过 HCG。具体来说，HCG 相比 LogSumExpDecoder 提升了 0.92%，较 PO-TreeCRFs 提升了 1.65%，而与 Latent Lexicalized Constituency Parsing 相比，HCG 的 F1 分数提高了 1.41%。这些提升结果也表明了 HCG 中的门控机制、交互注意力以及对比学习优化的有效性。

4.2.5 消融实验

为了进一步验证 HCG 的有效性，本节通过消融实验对 HCG 的各个关键组件进行逐一剖析。通过移除或替换模型中的不同模块，分析各个改进的设计对模型性能的具体影响，深入理解每个组件对最终性能的贡献。

在 MatScholor、CoNLL2003 以及 GENIA 三个数据集上的消融实验结果分别如表4.6、4.7以及4.8所示。实验首先从基线模型 SRGN 开始，通过逐步加入门控机制、交互注意力机制和对比学习模块，观察每个部分对模型性能的影响。每个模块的加入都有效提升了模型的整体性能，特别是在长实体的边界识别和复杂实体之间关系的建模上，表现出显著的改善。具体来说，门控机制增强了模型在多粒度特征融合方面的能力，交互注意力机制进一步提高了模型对实体间细粒度关系的捕捉能力。加入对比学习后，模型在所有三个数据集上的 F1 值均达到了最优，这一结果验证了对比学习对 HCG 模型的显著贡献，尤其是在优化对比损失和 CRF 损失的联合训练方面，提升了模型的整体鲁棒性。

此外，本章在 MatScholor、CoNLL2003 和 GENIA 三个数据集上分析了各个实体长度的识别效果，尤其关注 HCG 在长实体处理方面的表现。首先，对于 MatScholor 数据集上，考虑到本章主要针对通用领域的长实体识别，所以将 HCG 的结果与 SRGN

表 4.6　HCG 在 MatScholor 数据集上的消融实验结果

模块	Precision (%)	Recall (%)	F1 (%)
SRGN（基模型）	92.73	92.33	92.53
+ 门控机制	92.68	**92.60**	92.64
+ 交互注意力	92.98	92.56	92.77
+ 对比学习✓	**88.05**	**89.25**	**88.65**

表 4.7　HCG 在 CoNLL2003 数据集上的消融实验结果

模块	Precision (%)	Recall (%)	F1 (%)
SRGN（基模型）	92.73	92.33	92.53
+ 门控机制	92.68	**92.60**	92.64
+ 交互注意力	92.98	92.56	92.77
+ 对比学习✓	**93.19**	**92.60**	**92.90**

做了对比，以验证有效性，具体如表4.9所示。根据实验结果，SRGN 方法对于长度为 1 的短实体表现良好，F1 值为 92.29%。然而，随着实体长度的增加，SRGN 的性能逐渐下降，在长度大于 5 的实体上，F1 仅为 73.68%。相比之下，HCG 在处理长实体方面展示了显著优势。在长度为 1 时，HCG 的 F1 值为 92.06%，接近 SRGN 的表现，但对于其他长度的实体，HCG 的 F1 值始终高于 SRGN，尤其是在长度大于 5 的实体上取得了 75.98% 的 F1 值，对比 SRGN 提升了 2.3%。这一结果表明，HCG 在长实体边界的区分和识别方面都取得了显著提升。

根据实验结果可以发现，HCG 更适用于长实体识别。门控机制通过融合多粒度特征，使得模型在处理长实体时，能够更精确地捕捉到不同粒度的信息，进而有效提升了长实体的边界区分能力。此外，HCG 引入的基于多头注意力改进的交互注意力机制，能够更好地建模实体之间复杂的关系，尤其是在长实体的情境下，通过不同注意力头的多角度分析，加强了模型对长实体结构的理解。最后，HCG 在解码器

表 4.8　HCG 在 GENIA 数据集上的消融实验结果

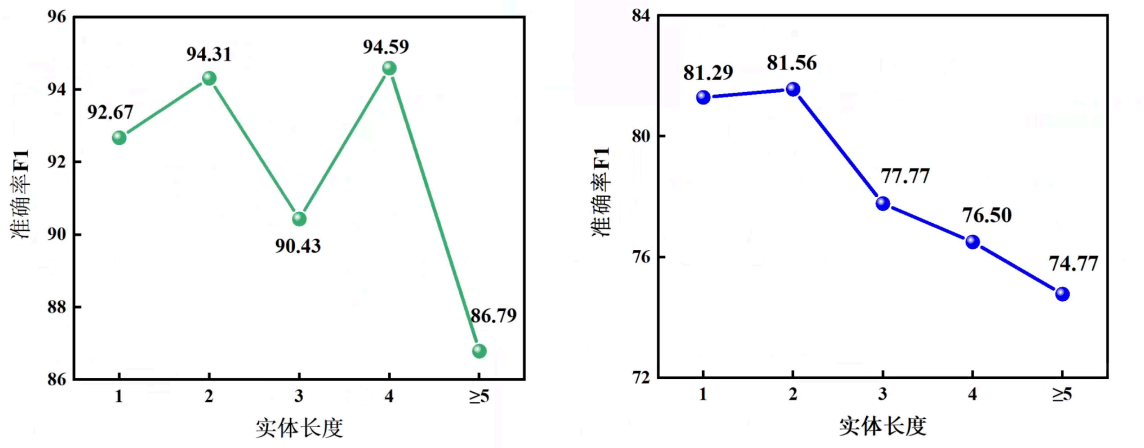
模块	Precision (%)	Recall (%)	F1 (%)
SRGN（基模型）	81.26	77.95	79.57
+ 门控机制	81.25	76.59	79.65
+ 交互注意力	81.15	**78.44**	79.77
+ 对比学习✓	**81.67**	78.11	**79.85**

表 4.9　HCG 与 SRGN 在 MatScholor 数据集上的实体长度分析

实体长度	SRGN	HCG
1	**92.29**	92.06
2	87.08	**86.12**
3	81.49	**83.37**
4	72.41	**77.35**
≥ 5	73.68	**75.98**

中引入的对比学习机制，通过优化对比损失和 CRF 损失的联合训练，进一步增强了模型在长实体边界和复杂结构上的鲁棒性，从而显著提升了模型的整体性能。

对于 CoNLL2003 和 GENIA 数据集, 实验结果如图4.3所示。虽然这些数据集的重点领域不同，但从 HCG 的表现来看，其对长实体的识别能力具有良好的通用性。在 CoNLL2003 数据集上，HCG 在实体长度为 1 至 4 的范围内均表现较好，特别是在长度为 4 的实体上，F1 值达到 94.59%。在 GENIA 数据集上，HCG 同样表现出了对长实体的良好识别能力。尽管 F1 略低于 CoNLL2003 数据集，但总体来看，HCG 在 GENIA 数据集的长实体识别上保持了一定优势。

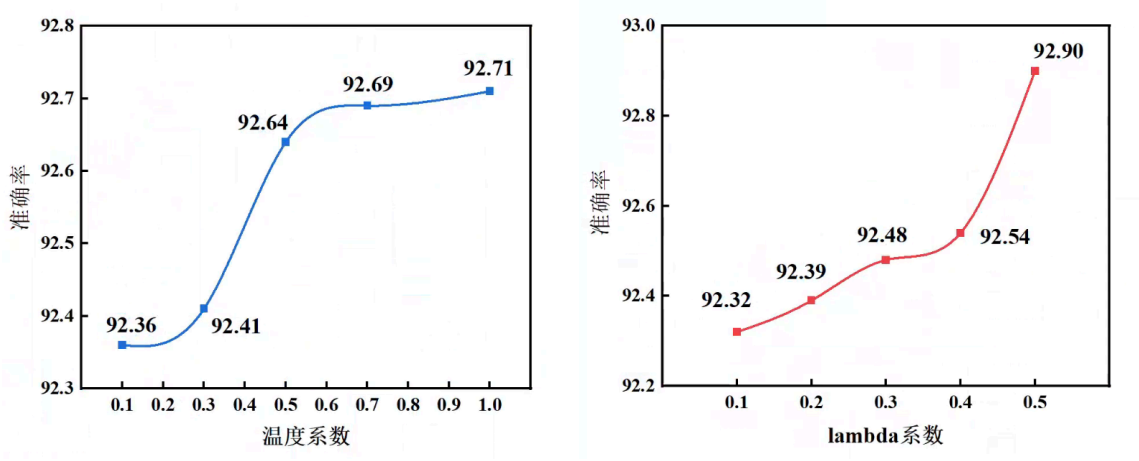


（a）CoNLL2003 数据集上实体分析　　（b）GENIA 数据集上实体分析

图 4.3　两个材料数据集上实体长度的实验结果。（a）CompMatLitDS 上实体长度结果，（b）MatScholor 上实体长度结果。

不同数据集上的实体分析实验表明，与上一章的 SRGN 方法相比，HCG 在长实体处理上的能力得到显著提升，表明其在通用材料领域文献文本中处理长实体的有效性。此外，HCG 在 CoNLL2003 和 GENIA 数据集上的表现也表明该方法具备较强的通用性，能够有效适用于多个领域的长实体识别任务。

此外，本节也在 CoNLL2003 数据集上对损失函数中的重要参数进行了实验，重点分析了对比学习中的温度系数（temp）和损失函数中的 lambda 系数对模型性能的影响。考虑到对比学习的灵活性与方法的通用性，实验通过系统地调整这两个参数，旨在找到最适合本实验设置的参数组合。



（a）温度系数的实验结果　　（b）lambda 系数的实验结果

图 4.4　CoNLL2003 数据集上对于不同参数的实验结果。（a）温度系数的实验结果，（b）lambda 系数的实验结果。

首先，对于温度系数的实验，采用控制变量法，控制 lambda 系数为 0.1，实验结果如图4.4（a）所示。当温度系数设置为 0.1 时，模型的 F1 值达到了 92.36%。随着温度系数逐步增大，F1 值增加，分别为 92.41%（temp=0.3）、92.64%（temp=0.5）、92.69%（temp=0.7）和 92.71%（temp=1.0）。这一趋势表明，较大的温度系数有助于模型在训练过程中保持较为平滑的特征分布，进而提升准确率。即温度系数为 1.0 时的表现最为优越，并且在本任务中较高的温度系数能够更好地促进对比学习的效果。

针对 lambda 系数的实验，保证温度系数 temp 为 1.0，实验结果如图4.4（b）所示。结果表明，当 lambda 设置为 0.5 时，模型的 F1 值达到 92.90%，为所有实验设置中表现最佳的值。在 lambda 为 0.1 时，F1 值为 92.32%。随着 lambda 系数不断增大，准确率 F1 逐步提高至 92.39%（lambda=0.2），92.48%（lambda=0.3）和 92.54%（lambda=0.4）。这一趋势表明，lambda 系数对模型的影响较为显著。较低的 lambda 值（如 0.1）时，模型的 F1 值较低，这可能是因为对比损失的权重过小，导致模型未能充分优化。而 lambda 值逐步增大时，F1 值逐渐提升，表明对比损失和 CRF 损失的联合优化更能有效地提高模型的性能，尤其是在 lambda 为 0.5 时，达到最优效果。

实验结果表明，在 CoNLL2003 数据集上的最佳参数设置为温度系数为 1.0 和 lambda 为 0.5。此时，模型能够在对比学习和 CRF 损失的联合优化下，发挥出最佳的性能。温度系数较大时能够增强特征之间的关系，而较高的 lambda 值则确保了对比损失和 CRF 损失的平衡，最终优化了模型的长实体边界区分能力和识别准确性。

4.3 本章小结

针对第三章 SRGN 在长实体识别与边界区分上的不足，提出了多粒度融合图网络（HCG）。HCG 通过引入可学习门控机制和交互注意力机制，增强了跨粒度信息交互；通过对比学习结合 CRF 损失联合训练策略，优化细粒度语义捕捉，显著提升了通用材料文献中长实体的识别精度。实验表明，HCG 在 MatScholar、CoNLL2003 和 GENIA 数据集上的 F1 值分别达 88.65%、92.90% 和 79.85%。通过消融实验验证了各个组件的有效性，实体长度分析进一步证明了 HCG 能够有效解决长实体识别和边界准确区分问题，尤其是在实体长度大于等于 5 时，表现有了明显提升。最后，通过进行对比学习和损失函数参数的实验，找到了最佳的参数组合，为 HCG 的优化提供了可靠依据。虽然该方法在实验中取得了良好的效果，但也存在一些问题，比如由于增加了交互注意力，可能在训练时会产生比较高的内存，对设备有一定要求。

第五章　复合材料文献数据的性能预测与应用设计

复合材料作为一种具有广泛应用前景的先进材料，其性能预测一直是材料科学研究中的核心问题之一。由于复合材料的性能受多种因素的影响，包括组成成分、加工工艺、环境条件等，因此从大量的文献中挖掘出有价值的关键信息并进行精准的性能预测，对于加速新材料的研发具有重要意义。因此，本章针对碳纤维复合材料，通过文献挖掘和机器学习，实现对力学性能的高效预测和实际应用设计的支持。首先，采用第三章提出的 SRGN 进行信息提取，从大量复合材料文献中提取出与材料组成、工艺、性能相关的关键信息，根据专家知识进行数据整理与筛选，确定可能对力学性能影响较大的特征，利用机器学习方法提供准确的性能预测。其次，针对碳纤维复合材料的性能预测需求，开发了一个自动化工具。该工具以提取得到的材料组成、加工工艺和性能数据为基础，采用机器学习方法进行性能预测。通过该系统，研究人员可以在现有数据的基础上高效开展材料性能预测与分析，从而为复合材料的设计优化和应用提供支持。

5.1　复合材料性能预测

在这一小节中，首先基于 SRGN 提取得到的实体信息，进一步筛选整理出影响碳纤维复合材料力学性能的关键特征，随后使用不同机器学习模型对弯曲强度、拉伸强进行性能预测。下面将进行详细介绍。

5.1.1　文献收集与挖掘

本章从 Elsevier、Wiley、MDPI 等公开文献数据库，根据“碳纤维复合材料”与“力学性能”关键字，筛选出从 2019 年到 2022 年相关的 380 篇 PDF 文献。利用第三章提出的 SRGN 模型，从复合材料文献的非结构化文本中提取 13 种实体，包括基体、填料、复材、辅助添加剂、材料数值、加工类型、工艺数值、制备方法、成型类型、性能名称、性能数值、测试方法、测试标准。随后，对提取的结果进行数据筛选与整理，最终收集到 105 条与弯曲强度相关的数据、112 条与拉伸强度相关的数据，并确定了 9 个共同影响弯曲强度、拉伸强度的关键特征。这 9 种特征包括基体、填

料、填料量、加工类型、加工温度、加工时间、厚度、几何形状和湿热条件。具体如表5.1所示，其中输入特征“填料类型”表示碳纤维复合材料中的第二填料的类型，第一填料类型默认为碳纤维。

表 5.1　9 个关键特征的详细信息

特征	特征类型	特征名称	特征范围
输入特征	分类特征	基体类型	环氧树脂，尼龙 6，聚丙烯等
		填料类型	石墨烯，碳纤维，碳纳米管等
		加工类型	固化，热压，压缩成型
		湿热条件	有，无
		几何形状	弯曲，扁平
	数值特征	填料含量（wt%）	0-76.18
		温度（℃）	25-415
		时间（minutes）	5-1440
		厚度（mm）	0.2-6.2
输出特征	数值特征	弯曲强度（MPa）	14.5-1378.0
		拉伸强度（MPa）	2.4-1893.0

5.1.2　实验细节

为了确保模型的泛化能力和评估的准确性，采用十折交叉验证来划分数据集。鉴于数据规模相对较小，在模型训练过程中引入了正则化项，以避免过拟合，增强模型的泛化能力。另外，为了评估性能，采用决定系数 (R^2) 和平均绝对误差 (MAE) 作为性能预测的评估指标，具体计算方法如以下公式所示：

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2}, \tag{5.1}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{5.2}$$

其中，n 为样本数，$\hat{y}_i$ 为预测值，y_i 为真实值，$\bar{y}_i$ 为均值。

5.1.3　性能预测实验分析

对于性能预测对比，选择了 7 种机器学习模型，包括 XGBoost、随机森林（RF）、梯度提升决策树（GBDT）、梯度树（DT）、AdaBoost、K-近邻（KNN）和 LightGBM，

并将这些模型的预测结果进行了对比。对于十折交叉验证条件下的弯曲强度和拉伸强度的性能预测结果，分别如表5.2，表5.3所示。

表 5.2　弯曲强度性能预测结果

机器学习模型	R^2	MAE
XGBoost✓	**0.87**	**77.40**
RF	0.86	81.63
GBDT	0.85	80.17
DT	0.84	84.67
AdaBoost	0.81	96.30
KNN	0.73	121.60
LightGBM	0.60	164.78

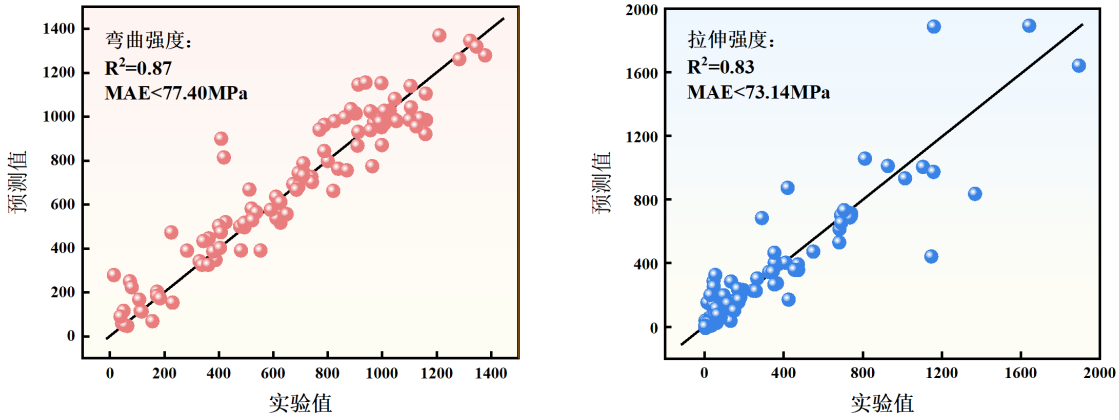
表 5.3　拉伸强度性能预测结果

机器学习模型	R^2	MAE
XGBoost✓	**0.83**	**73.14**
RF	0.79	79.86
GBDT	0.71	88.72
DT	0.59	94.53
AdaBoost	0.41	164.64
KNN	0.47	97.27
LightGBM	0.36	150.58

实验结果表明，使用机器学习方法构建的性能预测模型与数据拟合良好，其中XGBoost 在弯曲强度与拉伸强度的性能预测方面表现都最佳。这一结果也间接验证了 SRGN 从文献中提取关键信息的准确性和可靠性。

为了观察数据的拟合情况，基于表5.2和表5.3的预测结果，绘制了拟合曲线图。图5.1（a）、（b）分别显示了复合材料力学性能预测值与实验值之间的关系。弯曲强度的 R^2 值为 0.87，MAE 为 77.40MPa；拉伸强度的 R^2 值为 0.83，MAE 为 73.14MPa。这些结果反映出该模型具有良好的拟合能力和较强的泛化性能，同时也证实了其在处理未知数据时的可靠性。

在来自不同文献、有限的数据条件下，性能预测仍然获得了较高的预测精度，主要原因如下：



(a) 弯曲强度性能预测结果　　(b) 拉伸强度性能预测结果

图 5.1　复合材料力学性能的预测结果。（a）弯曲强度性能预测结果，（b）拉伸强度性能预测结果。

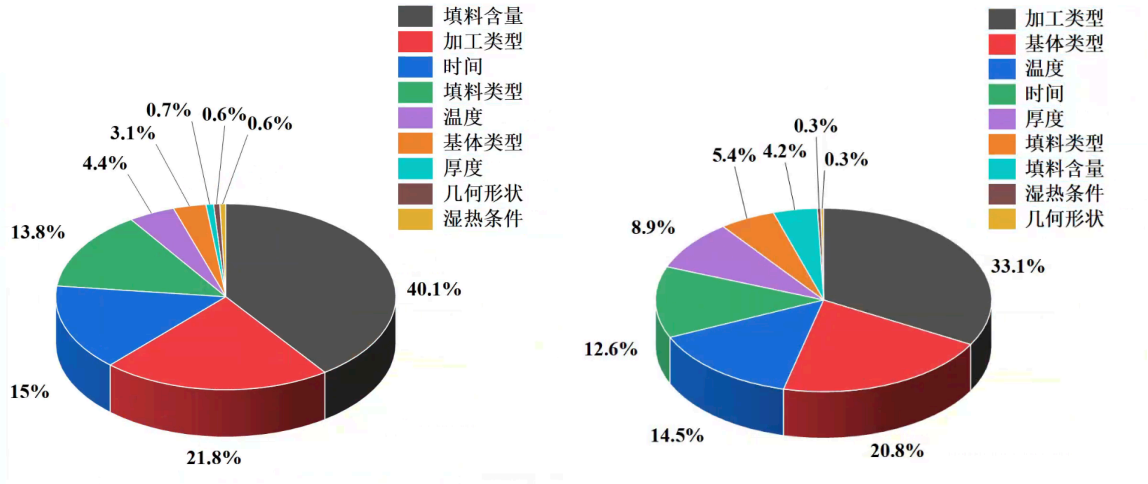
(1) 各输入特征之间的相关性。数据集中，弯曲强度的范围是 14.5-1378.0MPa，拉伸强度的范围是 2.4-1893.0MPa。在进行性能预测时，综合考虑了复合材料组分、工艺等因素对弯曲强度的影响，因此，模型对材料性质的预测更为准确。

(2) 输入特征的确定主要是基于文献挖掘方法中实体识别的结果，并结合专家知识进行整理和确定。根据性能预测实验，预测结果与实验结果的一致性表明，所选特征和研究方法具有科学性和有效性。

为了评估弯曲强度与拉伸强度预测模型中各输入特征的重要性，引入了基于 SHAP[63]（Shapley Additive Explanations）的特征贡献分析方法。在进行重要性分析的方法中，SHAP 是用于解释机器学习模型的常用方法，正 SHAP 值、负 SHAP 值分别反映了某一特征对模型预测过程中起到的正面或者负面贡献[14]。因此，SHAP 方法能够定量地评估每个特征对预测结果的影响。

复合材料力学性能的特征重要性结果如图5.2所示。对于弯曲强度，图5.2（a）中显示了各特征对弯曲强度的影响情况。首先，“填料含量”特征在贡献度上占据了最大比重，贡献达 40.1%，这表明该特征对弯曲强度的预测结果有着最显著的正面影响。其次，“加工类型”和“时间”特征的贡献较为突出，分别占比 21.8% 和 15%。这表明在弯曲强度的预测中，加工类型和时间同样是重要的影响因素之一。相对较小的贡献来自“温度”和“几何形状”，这表明它们对弯曲强度的影响较为有限。

对于拉伸强度，如图5.2（b）所示，结果显示，“加工类型”和“基体类型”同样占据了较大比重，分别为 33.1% 和 20.8%，反映出在拉伸强度的预测中，这两项



(a) 弯曲强度特征重要性　　(b) 拉伸强度特征重要性

图 5.2　复合材料力学性能的特征重要性。(a) 弯曲强度特征重要性，(b) 拉伸强度特征重要性。

特征对模型输出的重要性。相比之下，“温度”对拉伸强度的贡献为 14.5%，相对较小，但仍具有一定的影响。而其他例如时间、厚度等特征在重要性比例中占比更少，说明在拉伸强度的影响因素中，这些特征较为次要。

通过 SHAP 值的特征贡献分析，能够明确识别出在不同类型强度预测中，哪些特征对模型输出的影响最大，从而为进一步的复合材料优化和性能改进提供数据驱动的决策支持。

5.2　性能预测系统应用设计

复合材料性能预测系统的设计旨在为用户提供一个便捷的工具，帮助通过上传材料相关文件进行性能预测与分析。接下来将详细介绍该系统的构建以及工作流程。

5.2.1　开发环境与相关工具

本系统基于 Python 3.9 开发，为了保证系统的图形用户界面能流畅运行，PyQt 5.15.7 框架被选用，它是 Python 中最常用的 GUI 开发工具之一，支持丰富的控件和布局管理，能够实现多种功能的交互设计。开发环境配置为 Windows 11 操作系统，搭配 PyCharm 2023.1 专业版作为集成开发工具。机器学习模块依赖于 scikit-learn 1.2.2 库实现数据预处理与交叉验证功能。对于数据可视化部分，采用 matplotlib 3.7.1 与 seaborn 0.12.2 协同工作，便于对预测结果进行直观分析。

此外，本系统在配备 16GB 显存的 NVIDIA RTX 3060 显卡及 Intel Core i7-12700H 处理器的硬件环境下运行。在该配置下，系统能够稳定完成模型加载与预测任务，具备良好的界面响应性能，可有效支持数据的处理与分析。

5.2.2 界面设计

系统界面的整体设计如图5.3所示，主要划分为四个功能区域，分别承担数据输入、日志展示、模型配置与训练以及结果输出等任务。

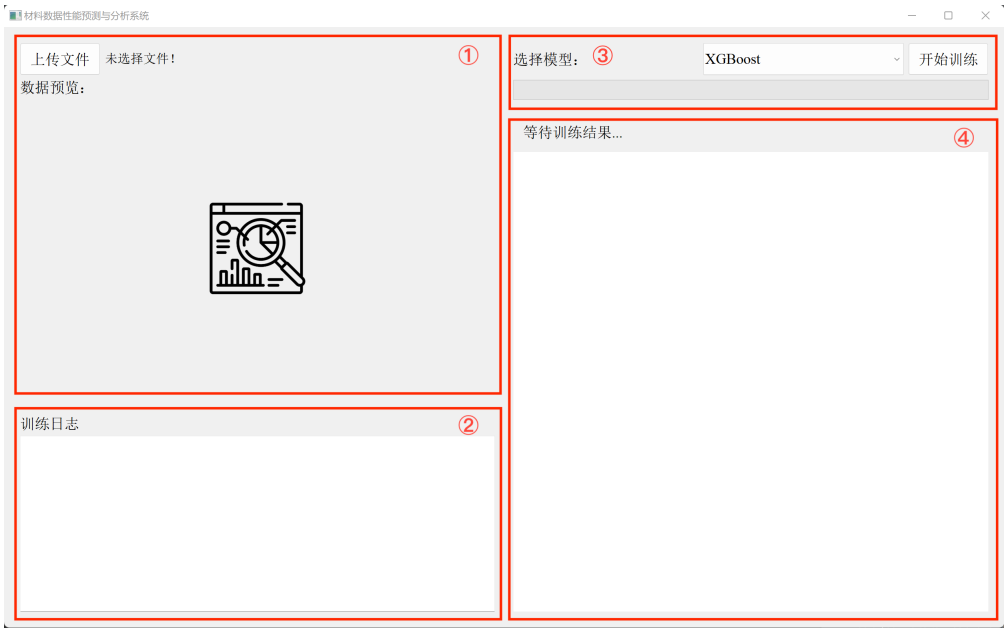


图 5.3　系统界面的整体设计

区域 1 用于上传用户的数据文件，系统支持多种格式的输入文件，包括 Excel 文件（.xlsx 与.xls）和文本文件（.csv），以保证良好的兼容性和操作灵活性。文件成功上传后，系统将在数据预览区域展示相应的表格内容。数据预览采用分页方式，每页展示 10 行数据，用户可通过界面提供的翻页按钮灵活查看全部数据内容，从而确保数据输入的准确性和完整性。

区域 2 用于展示系统运行过程中的日志信息，涵盖数据加载、数据维度提示、预处理进度、训练过程中的性能指标输出等。通过该区域，用户能够清晰掌握操作执行的全过程，有效提升系统的可解释性与交互体验。区域 3 则承担模型选择与训练任务，系统内置了七种主流的机器学习回归模型，包括 XGBoost、AdaBoost、随机森林、GBDT、DT、KNN 以及支持向量回归（SVR），用户可根据实际需求灵活选择合

适模型。模型选择完成后，点击“开始训练”按钮即可启动训练流程，下方的进度条将实时显示训练进度，提升用户对任务执行状态的掌握度。

区域 4 为性能预测结果的展示区域。考虑到模型评估的科学性和稳定性，系统采用十折交叉验证的方式对机器学习模型进行训练和测试。最终的评估结果将以平均性能指标形式展示，主要包括决定系数 R^2、平均绝对误差（MAE）以及均方误差（MSE），以全面衡量模型的拟合能力与误差水平。此外，为进一步增强结果的可视化表达，系统还提供了数据拟合图，直观展示实际值与预测值之间的关系，便于用户对模型预测性能进行深入分析与理解。

5.2.3 整体流程

该系统的整体流程如图5.4所示，用户通过点击区域 1 的“上传文件”按钮选择需要进行性能预测的材料文献数据，在选择文件时仅支持 xlsx、xls 以及 csv 三种格式的文件，如图5.5所示。在选择完成后，区域 1 的界面上会展示该数据文件的内容，用户可以通过点击“上一页”和“下一页”进行数据预览，如图5.6所示。

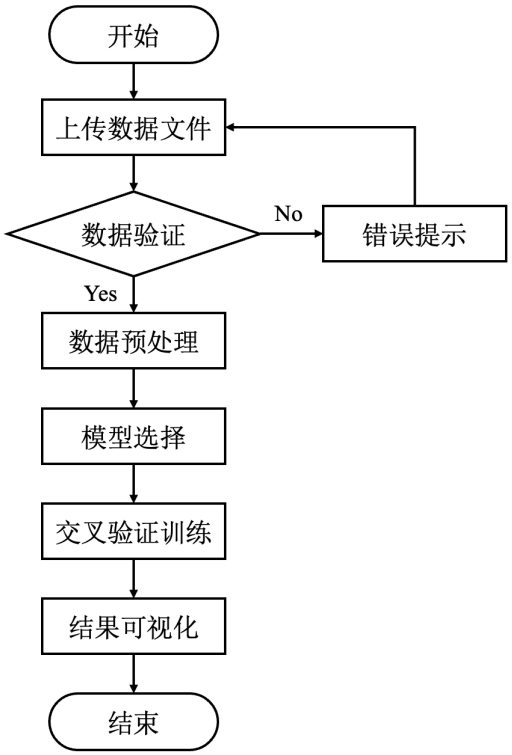


图 5.4　系统的整体流程

在完成数据文件加载操作后，界面下方的区域 2 随即显示日志信息，明确提示

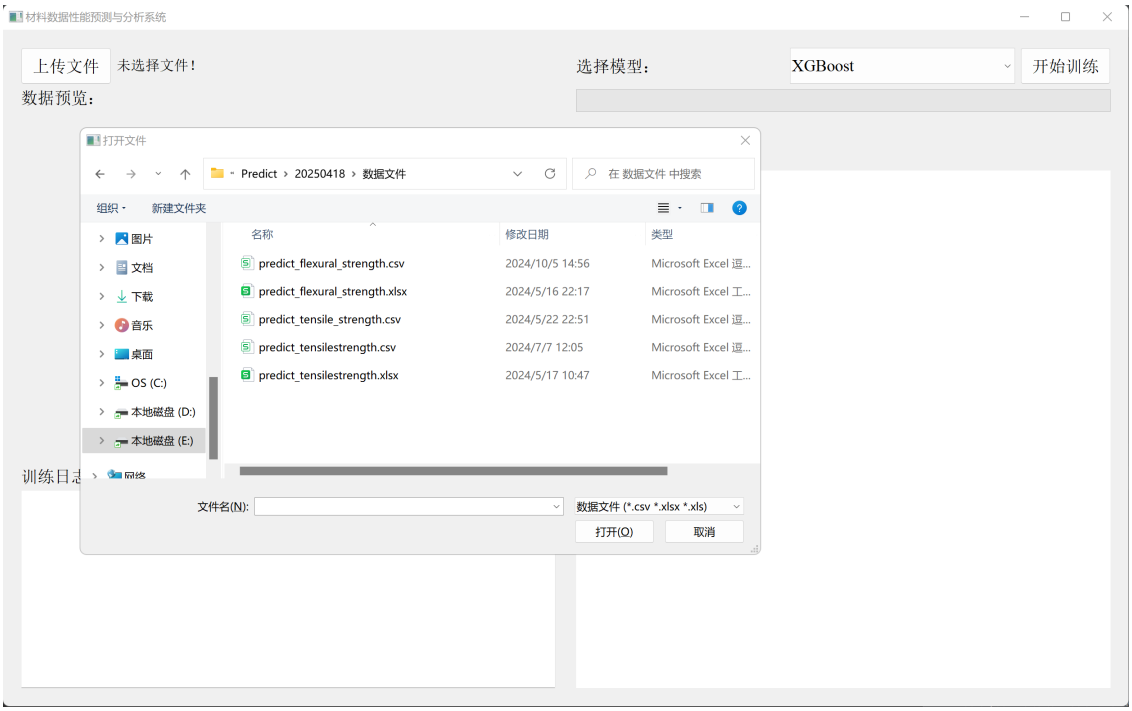


图 5.5　上传数据文件图示

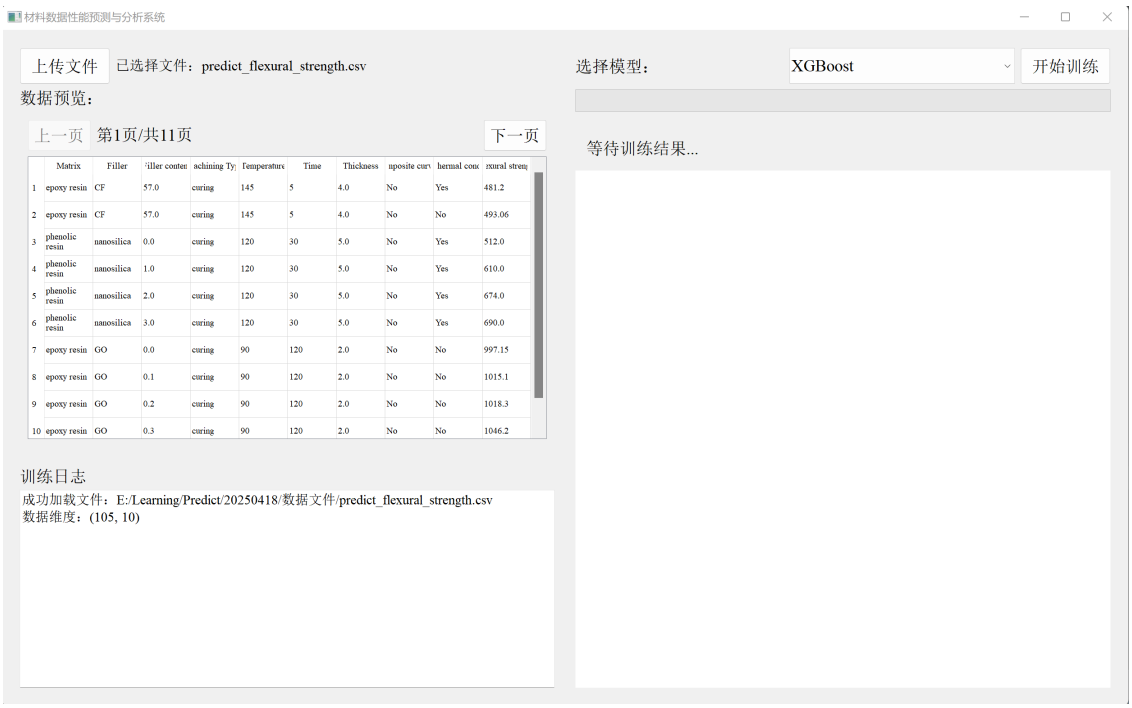


图 5.6　数据预览图示

“已成功加载文件”，同时展示所加载数据的维度。随后，用户可在界面右侧的区域 3 找到“选择模型”下拉框。通过点击该下拉框，用户能够从预设选项中挑选合适的机器学习模型，用以对已上传文件开展性能预测工作。图5.7展示了这一操作流程，在此示例中，用户选择了 XGBoost 模型。当用户完成模型选择后，系统将自动调用 XGBoost 模型进行性能预测。预测过程采用十折交叉验证的训练方法，以确保模型评估的准确性和可靠性。在训练过程中，每一折的训练详情会实时显示在区域 2 的日志中，使用户能够全面了解训练进展。与此同时，界面下方的进度条会动态展示当前的训练进度，直观呈现训练的完成情况。具体的展示效果可参考图5.8。

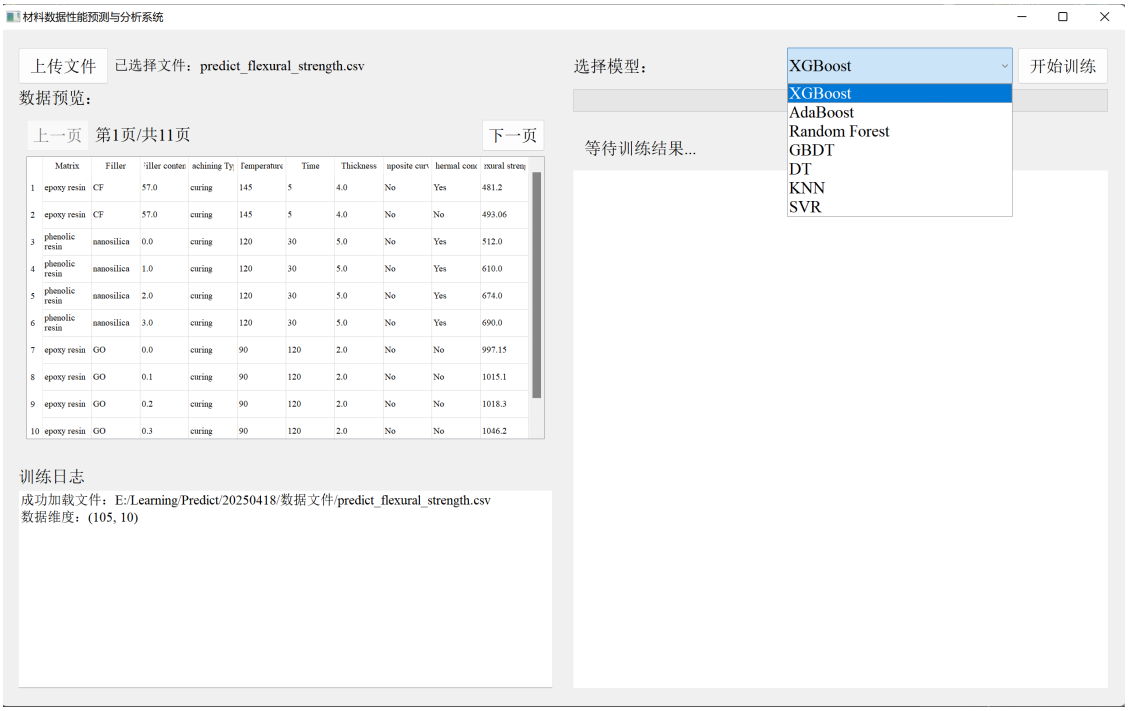


图 5.7　选择模型图示

模型训练完成后，系统将在区域 4 的界面展示训练结果，具体内容如图5.9所示。展示结果为十折交叉验证中每一折结果的平均值，包含决定系数（R^2）、平均绝对误差（MAE）和均方误差（MSE）。此外，界面下方将呈现数据拟合曲线图，图中以红色误差带标注预测值与真实值之间的误差范围，为用户后续分析提供直观参考。

该性能预测系统为复合材料的性能预测与优化提供了一个高效、便捷的工具，系统集成了上传数据文件、模型选择与训练、结果展示等功能，提供了一个直观易用的操作界面。通过结合机器学习模型和数据可视化技术，系统不仅提升了预测精度，还为用户提供了丰富的分析功能，极大地促进了复合材料研究与设计的智能化进程。

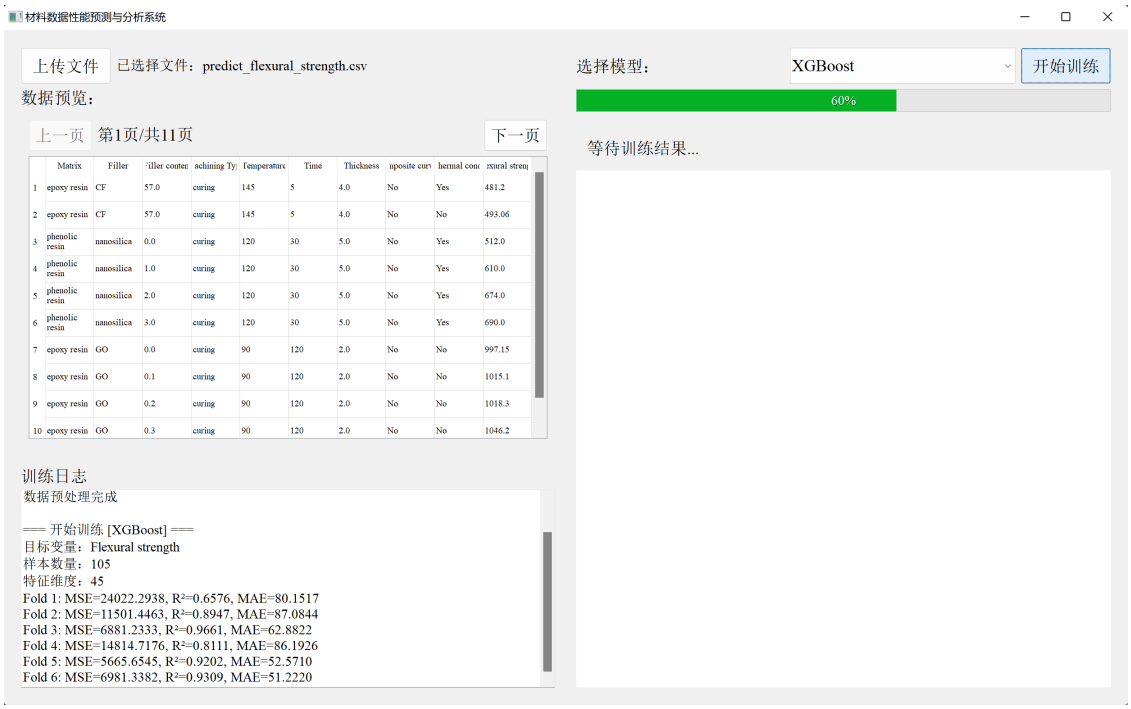


图 5.8　模型训练图示

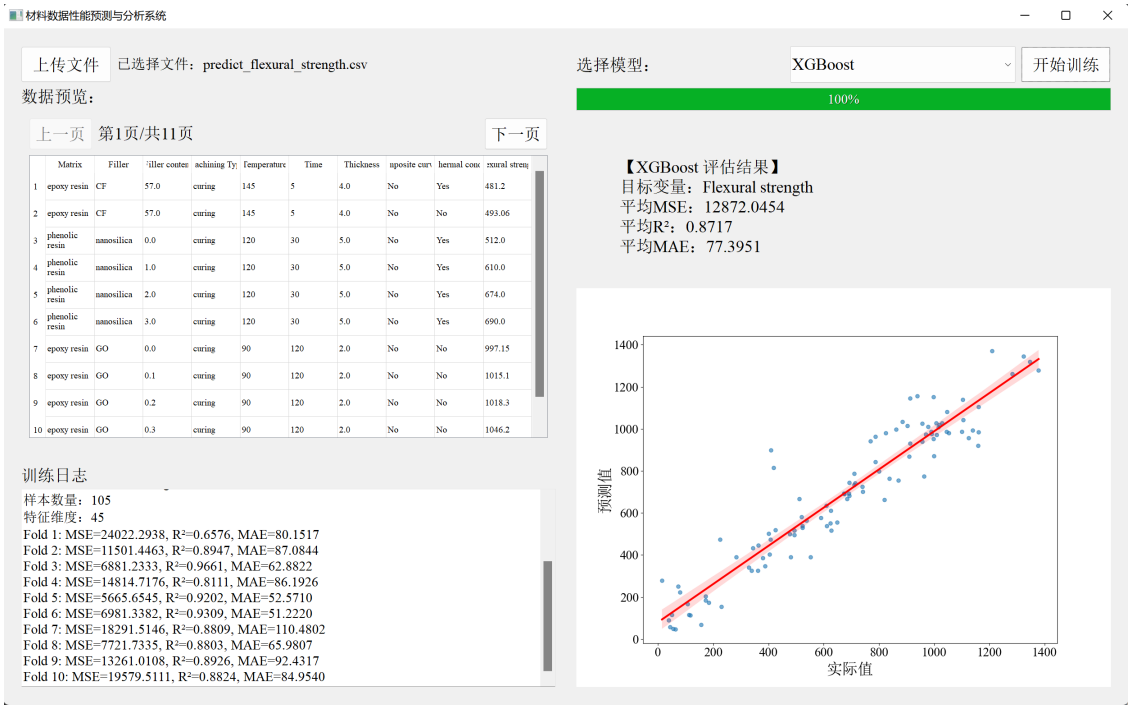


图 5.9　性能预测结果图示

5.3 本章小结

本章主要介绍了复合材料文献数据的性能预测以及性能预测系统的应用设计。对于性能预测，主要从文献收集与挖掘、实验细节，以及性能预测实验分析三个部分对性能预测进行阐述。通过实验数据和机器学习模型的应用，验证了预测方法的有效性，并对弯曲强度与拉伸强度的特征重要性分别进行了分析。最后，设计和实现了复合材料性能预测的系统，用户通过该系统能够高效地进行复合材料性能的预测与分析。系统集成了数据文件上传、模型选择与训练、结果展示等多项功能，提供了一个直观易用的操作界面。模型训练完成后的结果展示清晰明了，通过十折交叉验证的评价指标，用户可以全面了解模型的预测性能。此外，数据拟合曲线图的展示，为进一步的性能分析与优化提供了有力支持。

第六章　总结与展望

6.1　总结

本文围绕材料文献挖掘与性能预测问题，提出并实现了几种改进的方法，并在多个数据集上验证了其有效性。在材料文献挖掘领域，本文的主要贡献体现在如下方面：

（1）针对复合材料文献中长序列依赖、实体关系复杂的问题，提出了语义增强图网络模型，通过异构图结构实现命名实体识别任务。针对材料文献中长文本序列的处理难题，引入了分块注意力机制，降低计算复杂度并保留局部语义敏感度。在异构图特征处理环节，采用深度可分离卷积融合全局-局部特征，强化节点更新过程中的语义信息表征。同时引入可学习的动态边权重机制，实现节点间连接权重的自适应调节。为提升网络非线性映射能力，构建深度评分网络用于预测概率计算。通过在复合材料文献数据集与公开材料数据集上进行实验，证实了该模型在材料文献挖掘任务中的有效性与适用性。

（2）针对材料文献中实体边界模糊以及长实体识别效果不佳的问题，提出了多粒度融合图网络模型，通过门控融合机制与跨粒度注意力机制的协同作用，增强模型对多尺度语义信息的表征能力；同时设计 CRF 损失与对比学习损失的联合训练策略，优化实体边界识别的准确性。实验结果表明，该模型在多个基准数据集上均表现出优异性能，验证了其在复杂实体结构处理以及长实体识别场景中的有效性。

（3）本文将语义增强图网络模型应用于碳纤维复合材料的力学性能预测研究。通过对 2019-2022 年间 380 篇碳纤维复合材料实验文献的深度挖掘，结合领域专家知识对挖掘结果进行筛选与分类，提取出九项与材料力学性能相关的关键特征指标。基于上述特征，采用机器学习算法对碳纤维复合材料的弯曲强度与拉伸强度开展性能预测分析。同时，本文设计了材料性能预测系统，旨在为用户提供高效、自动化的材料性能预测工具。系统支持用户上传材料数据文件，通过内置机器学习模型自动完成数据处理与特征分析，实现多条件下材料性能预测。

本文的工作不仅为材料文献挖掘任务提供了有效的解决方案，还为材料性能预测提供了高效的工具，展示了良好的预测效果和较高的应用价值。这些研究内容为

新材料的发现和材料性能优化提供了数据支持，也为相关领域的研究提供了新的思路和方法。

6.2 展望

尽管本文提出的材料文献挖掘与性能预测方法在多个方面取得了良好的效果，但仍存在一定的局限性，并且在未来的研究中有进一步改进和拓展的空间。

（1）本文中提出的文献挖掘方法主要依赖于结构化数据的处理，对于文献中存在的一些具有复杂语义关系的非结构化数据，如材料文献中的图片、表格等，模型的处理能力尚显不足。未来可以结合自然语言处理和计算机视觉技术，进一步扩展系统的能力，使其能够处理更为复杂的多模态数据，提升文献挖掘的全面性和准确性。

（2）在材料文献挖掘方面，未来可以进一步提升模型的精度与鲁棒性，尤其是针对复杂材料类型和复杂实验环境下的实体识别任务。随着材料领域知识库的不断丰富和数据的增加，如何从大量文献中高效且准确地提取有价值的特征，仍然是一个值得探索的方向。此外，本文所使用的多粒度融合与对比学习方法在处理长实体识别任务时具有较好的性能，但面对更复杂的文本结构和多样化的实体类型，仍然有待进一步优化。

（3）性能预测方面，尽管本文采用的机器学习模型在碳纤维复合材料的弯曲强度与拉伸强度预测中取得了不错的效果，但随着数据量的增加和材料种类的多样化，现有模型可能面临性能下降或过拟合的问题。因此，未来可以尝试结合更多先进的深度学习技术，提升模型的泛化能力与稳定性，以及进一步评估性能预测系统在不同应用场景下的实际可用性，包括响应速度、资源消耗与预测精度等方面的综合表现。此外，如何结合实验数据与专家知识进行多模态学习，进一步提高预测精度，仍是未来研究中的一个重要方向。

上述研究方向仍然需要深入的理论研究和大量的实验工作。

参考文献

[1] ZHOU T, SONG Z, K. S. 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] GNANASEKARAN R K, SHANMUGAM B, RAJA V, et al. Multi-disciplinary optimizations on flexural behavioural effects on various advanced aerospace materials: A validated investigation[J]. MATERIALE PLASTICE, 2022, 59: 214-242.

[3] LIU X, FAN K, HUANG X, et al. Recent advances in artificial intelligence boosting materials design for electrochemical energy storage[J]. Chemical Engineering Journal, 2024, 490: 151625.

[4] 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.

[5] NAYANA V, BALASUBRAMANIAN K. Advanced polymeric composites via commingling for critical engineering applications[J]. Polymer Testing, 2020, 91: 106774.

[6] CEVDET K, CENGIZ C, GUNERI A. Use of silane coupling agents to improve epoxy–rubber interface[J]. European Polymer Journal, 2003, 39(6): 1125-1132.

[7] FU Z, LIU W, HUANG C, et al. A review of performance prediction based on machine learning in materials science[J]. Nanomaterials, 2022, 12(17): 2957.

[8] KONONOVA O, HE T, HUO H, et al. Opportunities and challenges of text mining in materials research[J]. iScience, 2021, 24(3).

[9] SHETTY P, RAMPRASAD R. Automated knowledge extraction from polymer literature using natural language processing[J]. iScience, 2021, 24(1).

[10] WESTON L, TSHITOYAN V, DAGDELEN J, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature.[J]. Journal of Chemical Information and Modeling, 2019.

[11] GUPTA T, ZAKI M, KRISHNAN N M A, et al. Matscibert: A materials domain language model for text mining and information extraction[J]. npj Computational Materials, 2022, 8(1): 102.

[12] ZHANG R, J. Z, Q. C, et al. A literature-mining method of integrating text and table extraction for materials science publications[J]. Computational Materials Science, 2023, 230: 112441.

[13] FU Z, LIU W, HUANG C, et al. A review of performance prediction based on machine learning in materials science[J]. Nanomaterials, 2022, 12(17).

[14] ZHU L, LUO Q, CHEN Q, et al. Prediction of ultimate tensile strength of al-si alloys based on multimodal fusion learning[J]. Materials Genome Engineering Advances, 2024, 2(1): e26.

[15] GUO P, MENG W, XU M, et al. Predicting mechanical properties of high-performance fiber-reinforced cementitious composites by integrating micromechanics and machine learning[J]. Materials, 2021, 14(12).

[16] 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).

[17] 赵海霞, 李磊, 吴信东, 等. 面向知识图谱的信息抽取[J]. 数据挖掘, 2020, 10(4): 282-302.

[18] KONONOVA O, HE T, HUO H, et al. Opportunities and challenges of text mining in materials research[J]. iScience, 2021, 24(3).

[19] WANG W, JIANG X, TIAN S, et al. Automated pipeline for superalloy data by text mining[J]. npj Computational Materials, 2022, 8(1): 9.

[20] SHETTY P, RAJAN A C, KUENNETH C, et al. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing [J]. npj Computational Materials, 2023, 9(1): 52.

[21] ZHANG J, ZHANG L, SUN Y, et al. Named entity recognition in the perovskite field based on convolutional neural networks and matbert[J]. Computational Materials Science, 2024, 240: 113014.

[22] FOPPIANO L, LAMBARD G, AMAGASA T, et al. Mining experimental data from materials science literature with large language models: an evaluation study[J]. Science and Technology of Advanced Materials: Methods, 2024, 4(1): 2356506.

[23] DAGDELEN J, DUNN A, LEE S, et al. Structured information extraction from scientific text with large language models[J]. Nature Communications, 2024, 15(1): 1418.

[24] 时宗彬, 朱丽雅, 乐小虬. 基于本地大语言模型和提示工程的材料信息抽取方法研究[J]. 数据分析与知识发现, 2024, 8(7): 23-31.

[25] HUANG Z, XU W, YU K. Bidirectional lstm-crf models for sequence tagging[A]. 2015. arXiv: 1508.01991.

[26] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[27] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[A]. 2019. arXiv: 1907.11692.

[28] CHILD R, SGRAY S, RADFORD A, et al. Generating long sequences with sparse transformers[A]. 2019. arXiv: 1904.10509.

[29] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[30] LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[A]. 2020. arXiv: 1909.11942.

[31] CLARK K, LUONG M T, LE Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[A]. 2020. arXiv: 2003.10555.

[32] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001: 282–289.

[33] 刘军平, 王润鹏, 胡新荣, 等. 基于对比学习和重排序的实体链接方法研究[J]. 计算机工程, 2025: 1-11.

[34] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[A]. 2017. arXiv: 1704.04861.

[35] WEN X, ZHOU C, TANG H, et al. Type-supervised sequence labeling based on the heterogeneous star graph for named entity recognition[A]. 2022. arXiv: 2210.10240.

[36] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1724-1734.

[37] PENNINGTON J, SOCHER R, MANNING C. GloVe: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1532-1543.

[38] GOODFELLOW I, WARDE-FARLEY D, MIRZA M, et al. Maxout networks[C]// Proceedings of Machine Learning Research: Vol. 28 Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA, 2013: 1319-1327.

[39] NAKAYAMA H, KUBO T, KAMURA J, et al. doccano: Text annotation tool for human[Z]. 2018.

[40] SHEN Y, TAN Z, WU S, et al. PromptNER: Prompt locating and typing for named entity recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 12492-12507.

[41] SHEN Y, WANG X, TAN Z, et al. Parallel instance query network for named entity recognition[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 947-961.

[42] LI J, FEI H, LIU J, et al. Unified named entity recognition as word-word relation classification[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10965-10973. DOI: 10.1609/aaai.v36i10.21344.

[43] ZHANG S, CHENG H, GAO J, et al. Optimizing bi-encoder for named entity recognition via contrastive learning[C]//The Eleventh International Conference on Learning Representations. 2023.

[44] BELTAGY I, LO K, COHAN A. SciBERT: A pretrained language model for scientific text[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 3615-3620.

[45] 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.

[46] FOPPIANO L, ORTIZ SUAREZ P. Material scibert (tpu): Improving language understanding in materials science[Z]. 2022.

[47] SHAZEER N, LAN Z, CHENG Y, et al. Talking-heads attention[A]. 2020. arXiv: 2003.02436.

[48] TJONG KIM SANG E F, DE MEULDER F. Introduction to the conll-2003 shared task: language-independent named entity recognition[C]//CONLL '03: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4. USA: Association for Computational Linguistics, 2003: 142-147.

[49] KIM J D, OHTA T, TATEISI Y, et al. Genia corpus—a semantically annotated corpus for bio-textmining[J]. Bioinformatics, 2003, 19(suppl_1): i180-i182.

[50] LI X, FENG J, MENG Y, et al. A unified MRC framework for named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 5849-5859.

[51] YU J, BOHNET B, POESIO M. Named entity recognition as dependency parsing [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 6470-6476.

[52] CUI L, WU Y, LIU J, et al. Template-based named entity recognition using BART [C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021: 1835-1845.

[53] YAN H, GUI T, DAI J, et al. A unified generative framework for various NER subtasks [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 5808-5822.

[54] SHEN Y, SONG K, TAN X, et al. DiffusionNER: Boundary diffusion for named entity recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 3875-3890.

[55] WANG Y, SHINDO H, MATSUMOTO Y, et al. Nested named entity recognition via explicitly excluding the influence of the best path[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 3547-3557.

[56] FU Y, TAN C, CHEN M, et al. Nested named entity recognition with partially-observed treecrfs[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(14): 12839-12847.

[57] LOU C, YANG S, TU K. Nested named entity recognition as latent lexicalized constituency parsing[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 6183-6198.

[58] WAN J, RU D, ZHANG W, et al. Nested named entity recognition with span-level graphs[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 892-903.

[59] YANG K, S.and Tu. Bottom-up constituency parsing and nested named entity recognition with pointer networks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 2403-2416.

[60] VERMA H, BERGLER S, TAHAEI N. Comparing and combining some popular NER approaches on biomedical tasks[C]//The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Association for Computational Linguistics, 2023: 273-279.

[61] ZHOU W, ZHANG S, GU Y, et al. UniversalNER: Targeted distillation from large language models for open named entity recognition[C]//The Twelfth International Conference on Learning Representations. 2024.

[62] WANG Y, UTIYAMA M. To be continuous, or to be discrete, those are bits of questions[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024: 8036-8049.

[63] LUNDBERG S M, LEE S. A unified approach to interpreting model predictions[C]// NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 4768-4777.

攻读硕士学位期间取得的研究成果

一、论文

[1] Zhang, R., Zhang, Y., Chen, Q., Han, Y. A Method Based on Heterogeneous Graph and Contrastive Learning in Named Entity Recognition. 2025 8th International Conference on Advanced Algorithms and Control Engineering.（EI 会议，已录用，导师一作，本人二作）

[2] Zhang, R., Zhang, Y., Chen, Q., Han, Y. Mechanical Property Prediction of Materials via a Literature Mining Approach. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology.（EI 会议，已录用，导师一作，本人二作）

[3] Zhang, R., Zhang, Y., Chen, Q., Han, Y., et al. Prediction of Flexural Strength of Carbon Fiber Reinforced Polymers Based on Literature Mining.（导师一作，本人二作，submitted: Macromolecular Rapid Communications）

二、软著

[1] 软件名称：基于异构图的文献信息提取平台 V1.0，开发人：张瑞、张一琳、韩越兴。登记号：2024SR1573898，登记日期：2024.10.21，申请人：上海大学。（导师一作，本人二作）

致　谢

时光荏苒，转眼间研究生生活即将结束。回顾求学的这段历程，内心充满了感激与不舍。

首先，我想特别感谢我的导师张瑞老师。张瑞老师严谨务实，对学术研究精益求精，总能敏锐地指出问题所在，为我指明前进的方向；在论文撰写过程中不厌其烦地帮助我理清思路，鼓励我勇于创新。她对于科研细节的认真把控，以及对学术研究的热爱与坚持，深深地影响了我，使我受益终身。除了学术上的精心指导，张瑞老师也经常关心我的生活与成长，给予了我极大的支持与帮助，真正体现了一位导师严谨而温暖的风范。衷心祝愿张瑞老师身体健康、工作顺利、桃李满园。

同时，真诚地感谢陈侨川老师与韩越兴老师在我研究生阶段给予的悉心帮助与指导。两位老师在课题研究中给予我诸多建设性的意见，在关键节点为我指明方向、解决疑惑。他们渊博的学识与和蔼耐心的态度，使我在科研道路上不断进步。每当我遇到困难或瓶颈，两位老师都能及时给予鼓励，让我充满信心地继续前行。

特别感谢我的父母，在这段求学时光中，他们用无声的陪伴、无私的支持和深沉的爱为我遮风挡雨，使我得以安心追求自己的理想。父母的鼓励与关怀，是我坚持下去最温暖的力量。

感谢课题组的每一位同学。研究生阶段，与大家朝夕相处、共同探讨学术、相互帮助，给我带来了许多难忘的回忆。你们的陪伴与支持，让我的科研之路变得更加精彩。

最后，再次向所有曾给予我关怀和帮助的老师、同学与朋友们致以最诚挚的谢意！愿你们一切顺利、幸福美满！