上海大学

计算机工程与科学学院

# 2022-2023 学年夏季学期
# 智能应用联合大作业报告

课 题 名 称 ： 面向 NASICON 型固态电解质的描述符自动获取方法研究

学 生 签 名 ： 徐陆骏，李昀哲，倪思远，邱逸辰

指导教师签名：

评语： 得分：

# 一、课题成员

| 姓　名 | 文档撰写工作分工 | 开发工作分工 | 对课题的贡献百分比 |
|---|---|---|---|
| 徐陆骏 | 代码补充 | 数据标注、模型部署、前端开发 | 40 |
| 李昀哲 | 框架确定及细节补充 | 前端优化、后端实现、数据库设计及实现、模型部署、溯源机制实现、知识图谱构建 | 40 |
| 倪思远 | 文献查找 | 数据标注 | 10 |
| 邱逸辰 | 文献查找 | 数据标注 | 10 |

# 二、课题目标

## 1. 简介

目前，数据驱动的机器学习已广泛地应用到 NASICON 型固态电解质材料领域来进行激活能的预测与构效关系的研究。其中，描述符的选择尤为重要，它会影响数据的质量从而影响机器学习的预测性能。然而，已发表的科研文献中存在着大量描述符相关的知识待开发。

本项目利用文本挖掘方法，从小批量 NASICON 型固态电解质文献中，抽取描述符并以此构建模型进行训练，实现自动、高效地获取 NASICON 型固态电解质描述符。

## 2. 国内外研究现状述评及研究意义

## 2.1 项目研究意义

NASICON 型(Na Super-Ion Conductors)复合物在十九世纪六十年代首次被发现，在 1976 年 $Na_1+XZr_2SiXP_3-XO_{12}$ 类材料被开发后，由 Hong 和 Goodenough 命名为“NASICON”。NASICON 型材料目前被认为是适合高压固态电池的固体电解质。

NASICON 型固态电解质因其原料成本低，安全性高，化学性质稳定、电导率高、电化学窗口宽等特点，成为了最具潜力的钠离子无机固态电解质[1]。鉴于

开放的 3D 结构和高离子导电率，NASICON 结构的化合物已经广泛开发应用于二次电池的电极材料和固体电解质材料。固态电池的离子导电性能取决于离子在固态电解质中的扩散，迁移离子的激活能是衡量 NASICON 型固态电解质材料离子输运性能的关键指标之一。因此，通过对 NASICON 型固态电解质材料激活能的准确预测，能够在一定程度上加速新型高性能固态电解质材料的发现过程。

在预测激活能的过程中，描述符是一种有利的特征。大量有关预测激活能的描述符存在于 NASICON 型固态电解质相关的文献中。目前，绝大部分领域专家只能依靠亲自阅读相关文献，结合自身领域知识，手动提取描述符，然后再进行预测工作。这样的过程存在以下问题：其一，NASICON 型固态电解质相关文献规模大、来源广泛，手动提取会消耗极大的时间、人力成本；其二，个体知识存在差别，考虑到描述符的选择至关重要，所以只是依靠人力难以及时获取最新的知识，导致提取描述符的准确度降低。

本项目针对目前 NASICON 型固态电解质描述符的获取方式进行创新，提出面向 NASICON 型固态电解质的描述符自动获取的方法。

主要流程是先对 NASICON 型固态电解质的文献预处理，再进行文献命名实体识别和关系抽取，最后基于预处理文献中提取的描述符，构建机器学习模型，实现 NASICON 型固态电解质描述符的自动获取。此方法目前在这个领域还未有团队深入研究，具有良好的发展前景和现实意义。

## 2.2 实体识别的研究现状

实体识别[2][3]旨在将现实生活中具有实体类信息的名称识别提取出来，可以是具体的客观存在的事物，也可以是某种抽象概念，具体到不同的领域也有不同的定义，在通用领域有人名、地名、机构名；在医疗领域可以将疾病名称，症状，治疗方式，药物名称等定义为实体类别；在军事领域有武器名、组织名、地区名等。这些实体非常清楚的表达了这句话的意思，准确识别出这些实体，对一些下游任务有很大帮助，可以应用在提升信息检索的质量，构建知识图谱等下游任务。

国外在对于命名实体识别任务的研究起步较早，研究经历主要经历了基于规则的方法，基于统计学的机器方法学，基于深度学习的方法。英文只需关注单词本身含义[4]，不需考虑词语边界问题，实现难度相对较低，准确率也相对较高。早在 1991 年 L.Rau 就在金融新闻领域提取公司名称，并随后在 100 万多字的财经新闻中做了测试，准确率超过 95%，表现效果比人为标注索引的公司实体名多出 25%，并在 1996 年 MUC-6 会议上被正式列为信息抽取的一项子任务。Nadeau，David[5]等学者在 2007 年总结了近 15 年内命名实体识别与分类任务的研究进展

和方法，调查结果显示，当时的 NERC 系统实现了从最初的手工规则提取到机器学习技术的广泛使用。近年来，深度学习通过将文本向量化表示和一些非线性处理的语义组合，在命名实体任务上有了非常可观的进展。

## 2.3 关系抽取的研究现状

关系抽取[2]是在实体识别的基础上，将其中有关系的两个实体进行分类以帮助人们得到文本中更为重要的信息，是知识库构建，只能问答，语义分析的重要技术支撑。一个完整的关系抽取系统包含实体识别，实体对齐，关系分类三部分，第一部分是基础，第三部分是核心，不管是国内还是国外，其实现方式不外乎两种，一种是以管道的当时抽取，先做一部分提取出实体，然后在识别出的实体上做抽取任务。另一种方式是做联合抽取，输入文本，直接输出实体关系对。大量实验表明同等数据集和算法下，联合抽取的性能要比管道方式性能好。

实体关系抽取任务在 1998 年 MUC-7 会议上被首次正式提出，当时主要用模板匹配的方式在英文纽约时代新翁服务语料库完成关系分类。随后在机器学习和深度学习方面都有相应的发展，通过分析句法语法结构，得到句子的某些特征，利用这些特征结合机器学习的相关算法去做关系抽取任务。随着 SVM 的发展，核函数设计的多样化也促进了关系抽取任务的发展。像 Bunescu R 和 Mooney R 提出的依赖树核和子序列核[6]。Miwa M， Bansal M[7]在序列和树结构上用长短时记忆网络做到端到端的关系抽取。近些年来，神经网络模型被应用到更多的关系抽取任务，有递归神经网络，卷积神经网络，循环神经网络，图神经网络，transformer 等。

与国外相比，中文实体标注的研究进展有限。无词界，无形态变化的特点使抽取困难大大增加。Wenjie Li, Peng Zhang[8]等人提出一种新型基于语法特征的中文关系抽取模型，不仅利用了实体本身语义信息，还定义了实体之间的九种位置结构，在 ACE2005 数据集上表现良好。Zheng, Suncong[9]等人提出了一种混合神经网络模型，不需要任何手工特征，实体检测部分用来双向 LSTM 模块，关系分类用 CNN 模块并在 ACE2005 数据集上有了最新成果。Tseng, Yuen-Hsien[10]等人开发了一种新型的开放式信息抽取系统，该系统不需要任何特定于关系的人工输入和预先指定的关系合计就可以提取出一组不同的关系，适合目标关系未知的大规模文本语料库。

## 2.4 描述符的构建与应用研究现状

机器学习方法采用分子描述符来建立模型，由于分子类型不同、相互作用的机理不同、描述符也不同，因而不存在普适的用于建模的分子描述符。尤其在对

作用机理不清楚或机理及其复杂的情况下，常用的方法是首先构建一些列描述符。尤其在对作用机理不清楚或机理极其复杂的情况下，常用的方法是首先构建一系列描述符，再用数学方法筛选描述符用于建模，描述符可以分为表征分子属性、行为、结构、应用、成分等。过去十年来，通过考虑分子的三维特征及其中原子的物理化学性质的方法，对传统描述符进行拓展，已经发展成为一种趋势，产生了 3D-MoRSE、MS-WHIM、自相关、BCUT、RDF、GETAWAY 等描述符。目前，商用的描述符计算软件并不能计算文献中出现的所有描述符，例如，可计算 1600 个描述符的 Dragon 软件不能计算 MS-WHIM 描述符，可计算超过 1200 个描述符的 PreADME 软件不能计算 3D-MoRSE、RDF、GETAWAY 和 MS WHIM 描述符。

综上所述，鉴于 NASICON 存在较高的研究价值和属性预测价值，且材料专家对于预测所需描述符的获取费时费力，以及自身知识存在局限性的前提下，本项目创新性地基于文本挖掘，构建机器学习模型，实现描述符的自动获取，尽可能填充材料专家提供的描述符树，为专家提供更多描述符的选择，减少人力物力的消耗，更好地完成预测激活能的工作。

# 三、课题成果考核指标及其完成情况

本课题的执行过程如图 3-1 所示，开题预计成果为基本实现对描述符的自动获取及知识平台的搭建。目前已基本完成。

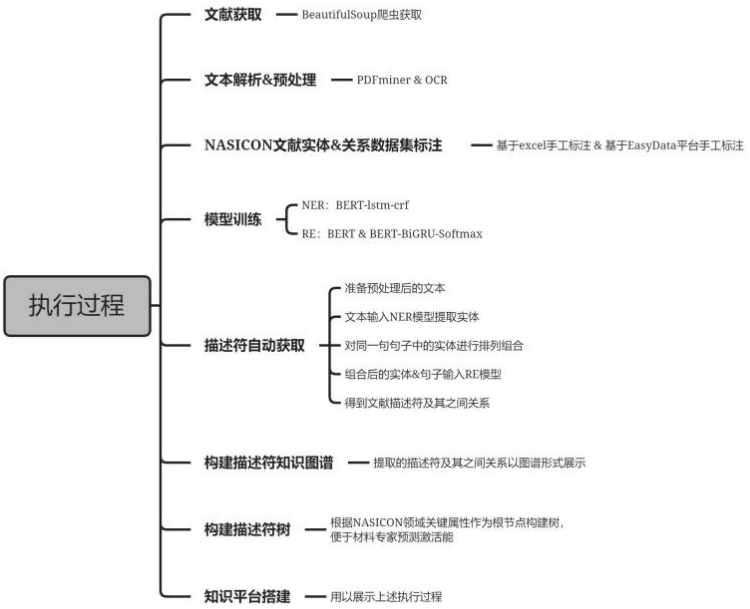


**图 3-1 课题执行过程**

# 四、研究开发内容

## 1. 数据

### 1.1 获取

本研究的核心为对于 NASICON 型固态电解质领域文献的知识获取，故数据来源主要为 Web Of Science 网站获取的文献。利用爬虫、正则等方法，筛选并获取符合要求的论文及元数据，论文以 PDF 格式存储，元数据存储于 excel 表格中，将其定义为

$$D_{Meta}\{title, Volume, Issue, Pages, author, Abstract, Keywords, journal\}$$

爬虫获取的伪代码如算法 4-1 所示：

**算法 1**：爬虫获取文献元数据

**输入**：目标网页前缀

**输出**：文献元数据 csv 文件

```
1：开始;
2：指定文献信息表存放路径;
3：foreach page do
4：      PageToDict 得到每一页文献的所有信息
5：          foreach dict.keys() do
6：              toString（所需的 column 即 D_Meta 所需的列名）;
7：      to_csv()  // 得到元数据的 csv 文件
8：结束
```

### 1.2 处理

文献预处理是有监督材料数据集构建过程中一个必要环节，它可以对下游文本挖掘任务产生重大的影响[11]。对于知识获取而言，无法直接对 PDF 格式的输入数据进行处理。文献预处理的手段根据文本挖掘方法及其每个阶段使用工具的差异而有所不同。其中，预处理初期旨在得到材料纯文本数据。PDF 科研文献是文本数据存储的主要载体之一，具有易获得、更全面的特质。因此，大规模提取

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

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

**基于 ChemDataExtractor 和人工清理的文本处理：**在预处理后期，需要对纯文本数据进行处理以得到可以用于标注的数据。在英文文本中，标点符号是识别句子较为明显的方式之一，然而材料文本具有领域特殊性，即材料科学领域的语言常常因为由多个词、符号和其它类型结构实体组成的术语而变得复杂。例如，$(La_{0.8}Sr_{0.2})_{0.97}MnO_3$、$(1-x)Pb(Zr_{0.52}Ti_{0.48})O_{3-x}BaTiO_3$等。因此，材料领域需要专门的文本处理工具，其对材料科学文本挖掘的成功十分重要。
ChemDataExtractor 的功能在上述三种工具中的功能最完备，且其操作简单，对用户友好，同时可以处理通用领域的材料化学文本。此外其版本迭代快，表明该工具会实时将最新技术融入进来，未来具有较 大的竞争优势。因此，我们将 ChemDataExtractor 作为材料文本符号化处理工具，以实现对材料领域复杂文本进行文本分段、分句及分词等操作以得到干净的能进行标注的半结构化文本数据

## 1.3 标注

**表 4-1 材料实体标签设计的对比**

| 设计人 | 设计目标 | 标签类别数 | 标签类别 | 适用领域 | 可解决问题 |
|---|---|---|---|---|---|
| Swain 等 | 开发能自动从大规模非结构化材料文献中挖掘化学信息的工具 | 3 | 属性、关系和测量 | 通用材料 | 大规模化学数据库快速创建问题 |
| Weston 等 | 将材料发现的新结果与已发表文献 | 7 | 无机材料、相、描述符、性能、应用、 | 无机材料 | 材料查找、指导文献搜索与总结及回 |

| | | | | | |
|---|---|---|---|---|---|
| | 联系起来 | | 合成方法和表征方法 | | 答简单的元问题 |
| Wang 等 | 从文献中自动挖掘出数据驱动材料设计所需的高质量可靠数据 | 4 | 耐热温度、密度、固相及液相温度 | 合金材料 | 具有高耐热温度的钴基单晶高温合金预测问题 |
| Pan 等 | 构建语义表示框架用于锂离子阴极的文献挖掘 | 3 | 无机材料信息、锂离子阴极和描述符 | 电池材料 | 锂离子阴极开发的候选材料寻找问题 |

对于有监督材料科学文本挖掘性能而言，数据标注起到至关重要的作用，本节也是我认为整个项目的核心关键。类别标签设计是影响有监督材料科学文本挖掘模型性能优劣的先决条件之一，在一定程度上决定了数据标注的质量以及文本挖掘的结果（即期望从文本中挖掘出何种类型的材料信息）。材料命名实体识别文本挖掘工作发展至今，不同领域专家从不同的下游任务出发，设计了适用于其独特下游任务的实体标签。本项目使用的标签基于刘悦老师课题组的过往项目，对比了 Swain 等人[12]、Weston 等人[13]、Wang 等人[14]和 Pan 等人[15]设计的实体标签，如表 4-1 所示。从表中可以看出，标签的设定需要从材料的应用点出发，同时要确定标签的可适用领域及其可以解决的问题。本节以通用领域描述符的自动识别为研究目标，期望通过挖掘出不同类型的描述符信息来实现对材料性能预测及构效关系的研究。然而，特定材料的属性会受到多种类型描述符的影响。例如，NASICON 型固态电解质激活能会受到如成分、结构、工艺及性能等描述符的影响。材料四面体即材料学四要素，旨在研究材料的成分、结构、制造、性能以及它们之间的关系，体现了对材料间构效关系的研究。故在处理工艺-结构-属性-性能四面体准则的驱动下，对需要挖掘的内容进行了总结，如表 4-2 所示。

**表 4-2 文本挖掘内容分析**

| 挖掘内容 | 描述 |
|---|---|
| 成分信息 | 材料文献中经常会对所研究的材料成分进行概述，包括组成的元素以及元素的含量。 |
| 结构信息 | 材料的结构可能会受到其成分的影响，且其可能会影响材料的属性。 |

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

为了实现从材料文献中自动获取描述符信息，通过对上述信息进行高度的抽象，本节设计了 8 个描述符类别实体标签，分别为：成分（Composition）、结构（Structure）、属性（Property）、工艺（Processing）、表征（Characterization）、应用（Application）、特别描述（Feature）及外界环境（Condition），其可以概括绝大多数的材料文献的描述符信息。表 4-3 展示了每个标签的定义及示例。同时，为了实现自动从文献中抽取出描述符间的关联关系，本节设计了 8 种实体类型之间的 8 种关系类型，分别为：原因-影响（Cause-Effect）、部分-整体（Component-Whole）、特征（Feature-Of）、位置（Located-Of）、实例（Instance-Of）、条件（Condition-On）、方法（Method-Of）及其它（Other）类，其具体定义如表 4-4 所示。

**表 4-3 材料领域 8 种实体类型定义**

| 实体标签 | 定义 | 例子 |
|---|---|---|
| Composition | 任何与化学式有关；描述材料内部与含量相关的内容等。 | NaCl, CaCl2; Na concentration, Electrons charge carriers |
| Structure | 晶体结构、相的名称；用于刻画晶体结构的名称等。 | Fcc, Phase; Bottleneck, Channel, Path |
| Property | 带单位的可度量值；材料表现出来定性的性质或现象；描述材料产生物理、化学过程行为，或者物理、 | Conductivity, Activation, Radius; Ferroelectric, Metallic; Phase |

| | | |
|---|---|---|
| | 化学机制的名词等。 | transition, Ionic reaction |
| Processing | 任何合成材料的技术；任何合成材料的技术；材料改性的手段等。 | Solid state reaction, Annealing; Doping |
| Characterization | 用来表征材料,实验或理论的任何方法；也可以是一个模型或者是公式等。 | XRD, STM, Photoluminescence, DFT; Bethe-Salpeter equation |
| Application | 任何高级的应用；任何特定的器件、系统等。 | Cathode, Photovoltaics; Battery Management System |
| Feature | 样品类型、形状的特殊说明等。 | Single crystal, Bulk, nanotube, Quantum dot |
| Condition | 描述材料所处的环境（材料的外部条件）。 | Temperature, Pressure |

**表 4-4 材料领域 8 种关系类型定义**

| 关系标签 | 定义（A 和 B 为描述的实体） | 可能存在关系的实体类型 |
|---|---|---|
| Cause-Effect | A 对 B 有影响 | Property-Property 、Composition-Structure、Structure-Property |
| Component-Whole | A 是 B 的部分 | Composition-Composition |
| Feature-Of | A 是 B 的特征 | Property-Property 、Composition-Property |
| Located-Of | A 占据了 B 位置 | Composition-Structure |
| Instance-Of | A 是 B 的实例 | Composition-Composition、Structure-Structure 、Property-Property |
| Condition-On | A 的条件是 B | Processing-Condition |
| Method-Of | A 的表征方法是 B | Property-Characterization |
| Other | A 与 B 无明显关系 | - |

为了获得用于模型训练学习的有监督材料实体识别及关系抽取数据集，需要研究者手工标注部分样本。选择合适的标注工具有利于提高标注效率。

**实体和关系数据集标注：**不同的文本挖掘方法有着不同格式的数据需求，导致相应的标注流程也有所差异。对于材料命名实体识别任务，本节选择 EasyData 工具进行材料实体和关系识别数据集的标注。通过对定义的描述符实体及关系标签分析可知，材料文本不同实体类型间可能会有重叠关系，而目前已有的标注工具大都只能针对单一实体类型间的关系进行标注。

由此实现了对于 NASICON 型固态电解质文献较高质量文本数据的构建。

# 2. 模型/算法

## 2.1 来源

模型基于 BERT、BiLSTM、CRF，进行融合改进，即：通过 MatBERT 来同时编码材料词嵌入、位置嵌入及句子嵌入信息从而动态捕获材料复杂文本间的深层语义特征，以缓解材料复杂句子的一词多义及代词指代 的编码问题，并利用 BiLSTM 对句子序列进行建模，以抽取材料局部上下文语义特征，最后采用 CRF 对句子最优的标签序列进行预测，以实现材料实体的精准分类。

数据存储基于 MySQL 数据库和 Neo4j 数据库进行进一步开发。

## 2.2 模型/算法的描述

### 2.2.1 命名实体识别

材料 NER 方法能够通过识别和分类文本中提及的概念来挖掘具有语义价值的实体对象。这些实体对象不仅可以映射到材料性能上，还能为研究人员找到相似的化合物或纳入注释标记提供巨大的帮助。然而，材料领域 NER 的研究依然处于起步阶段，使用的技术以传统基于字典、规则、机器学习的单一或组合方法偏多。例如，Lowe 等人[16]将字典和基于模式的技术组合进行 NER 的研究，使用命名约定规则研发了生物材料 NER 工具 LeadMine。Lezan 等人[18]通过解析化学文本实验合成部分的化学标记，基于模式规则和字典的组合技术进行材料化学 NER 的研究并开发了一个材料化学 NER 工具 ChemicalTagger。Swain 等人[17]提出了基于字典和机器学习（CRF）组合技术的化学 NER 方法，实现了化学实体及其相关属性、测量和关系标签自动抽取的 ChemDataExtractor 工具的研

发。上述材料 NER 方法可以在小数据集上达到一个很高的识别准确率，但是在面对大规模的数据集或其它领域时便不适用，往往需要重新设置新的规则或者收集新的字典。此外，手工设计规则及收集字典的步骤往往需要花费大量的人力、物力及财力。

**基于 MatBERT 的多级别语义特征的融合**。图 4-1 展示了基于多层语义特征融合的材料命名实体识别模型 MatBERT-BiLSTM-CRF 的结构，由基于 MatBERT 的多级别语义特征的融合、基于 BiLSTM 的局部上下文语义特征的融合和基于 CRF 的材料实体分类三部分组成。其中，MatBERT 同时编码词嵌入、位置嵌入及句子嵌入信息，以捕获含丰富材料信息的 token 及句子级别的语义特征并将其融合形成单词的向量表示；在此基础上，BiLSTM 对句子序列建模，以进一步捕获单词的局部上下文语义特征，从而最终获得 token 级别的语义特征向量；序列标注分类器 CRF 基于 token 级别的语义特征向量对单词或短语进行标签预测以获取最优的标签序列，从而实现实体分类。

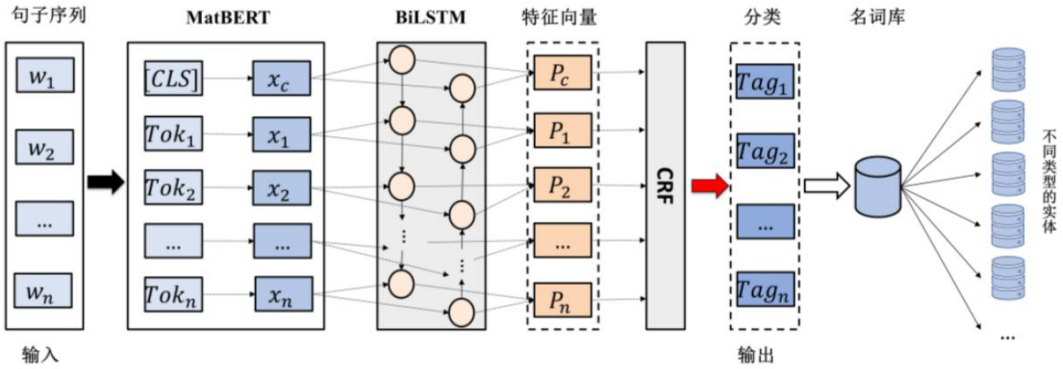


**图 4-1 基于多层语义特征融合的材料命名实体识别模型结构图**

**基于 BiLSTM 的局部上下文语义特征的融合**。NER 任务需要对句子中所有单词序列进行分类，因此需要模型具有识别时序信息的能力。RNN 拥有捕捉时序信息进行端到端分类的能力。然而，RNN 在时序信息的传播过程中经常会遭受梯度消失和梯度爆炸的问题。针对该问题，本节引入 RNN 的一个变体 LSTM 来解决。LSTM 有三个门控单元，即输入门、遗忘门和输出门，如图 4-2 所示，它们能够有选择地保存上下文信息。因此，LSTM 能有效缓解梯度消失及梯度爆炸的问题同时比 RNN 更适合捕捉长距离依赖。

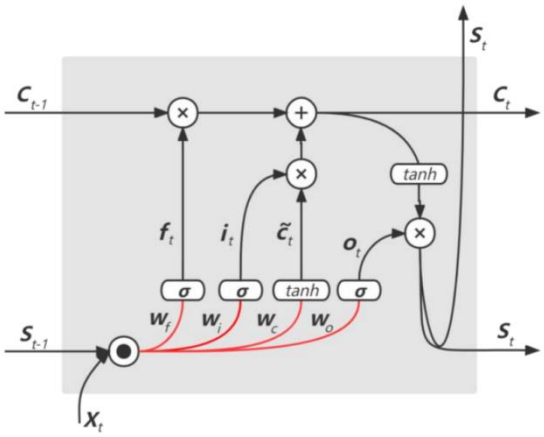


**图 4-2** LSTM 单元结构

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

具体地，假设使用 $w = \{w_1, w_2, \ldots, w_n\}$，来表示一个通用的输入序列，其中 $\mathbf{w_i}$ 是第 i 个单词的输入向量，$y = \{y_1, y_2, \ldots, y_n\}$表示对应于 w 的标签序列。公式（1）计算 CRF 模型的评估得分。

$$score(W, y) = \sum_{i=1}^{n} T_{yi,yi+1} + \sum_{i=1}^{n} P_{i,yi} \tag{1}$$

其中，**T** 表示转移矩阵，$T_{yi,yi+1}$ 为 $y_i$ 标签转移到 $y_{i+1}$ 标签的概率分数。

由此，对于输入的语句，进行标签的预测，建立实体库。

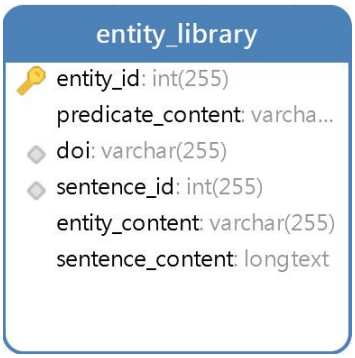


**图 4-3** 实体库结构

### 2.2.2 关系抽取

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

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

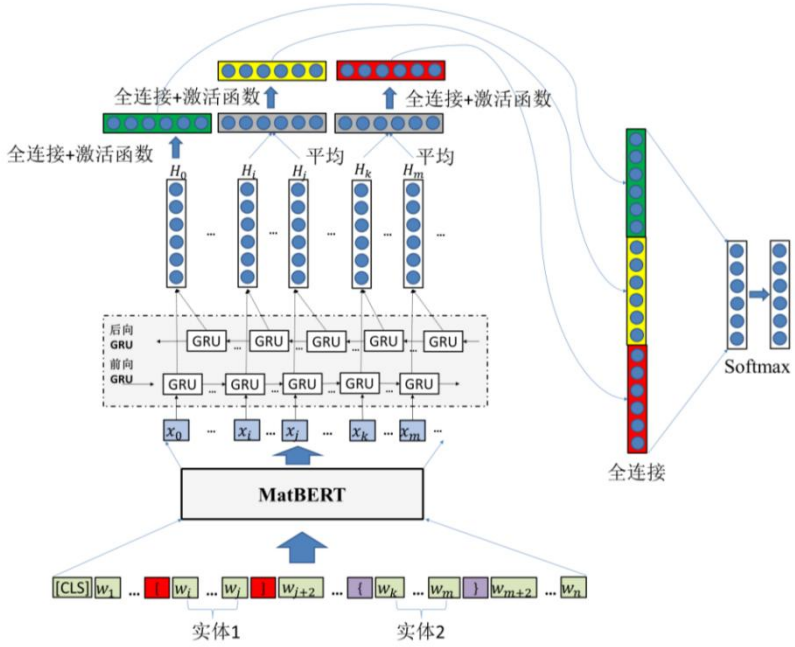


**图 4-4 实体感知的材料实体关系抽取模型结构图**

由此，得到关系库。

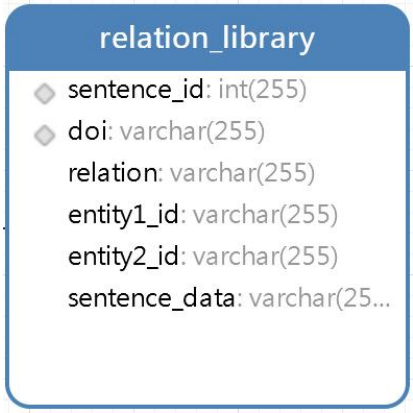


**图 4-5 关系库结构**

### 2.2.3 基于实体关系的材料知识图谱构建

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

基于 Neo4j 图数据库[19]实现材料实体关系三元组信息的存储。Neo4j 是一个基于 Java 语言实现的开源 NoSQL 图数据库，其架构旨在优化节点和关系的快速管理、存储和遍历过程。在 Neo4j 中，关系是图数据库中最重要的元素，它表示节点之间的互连，即由一个节点指向另一个节点。Neo4j 中只有两种数据类型：节点和边。节点用于保存材料实体，边来用于连接节点以表示材料实体间的关系。本节通过 Python 的 py2neo 工具包对 Neo4j 图数据库进行读写操作以实现材料实体关系三元组的存储。

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

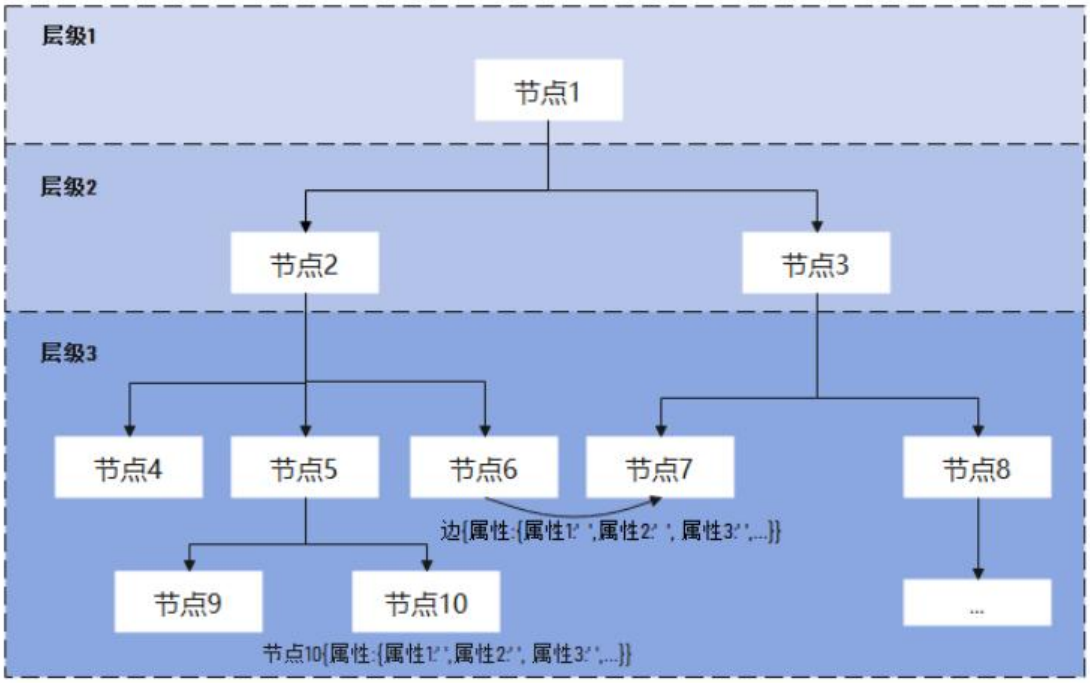


**图 4-6 描述符树的通用形式**

在此基础上，本节设计了材料描述符树的填充算法。具体步骤如下：

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

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

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

## 3. 系统架构

本项目主要利用 Vue+SpringBoot+MySQL 的技术路线进行平台开发。Vue 主要用于前端页面的渲染和与用户的交互，向后端发出服务请求；SpringBoot 主要用于服务的搭建和处理，同时和数据库进行交互。

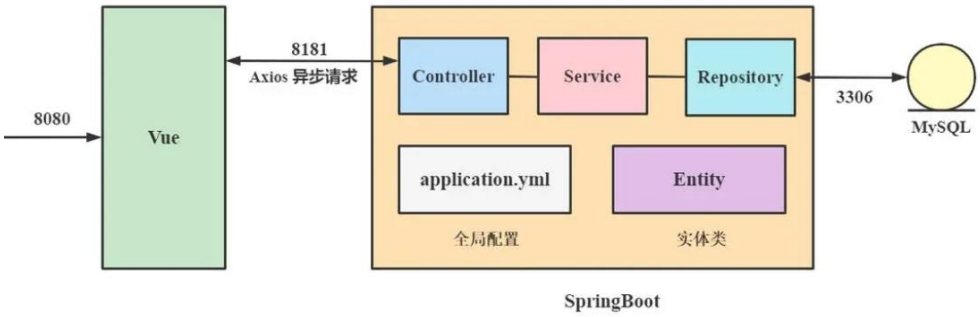


图 4-7 系统架构

## 4. 模块介绍

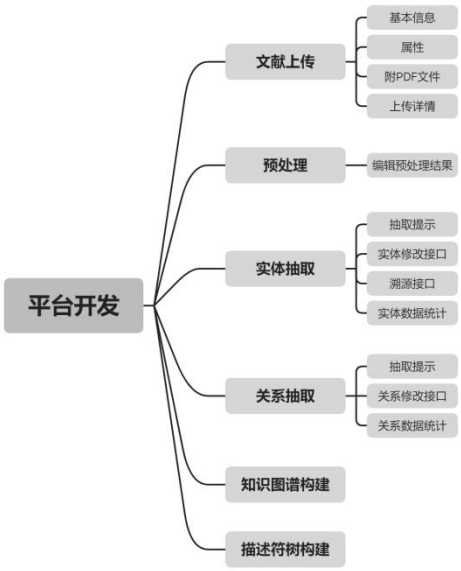


图 4-8 平台开发模块

### 4.1 文献上传

对于整个知识平台而言，主要用以处理前期已经获取的 PDF 格式的文献，故在最开始，需要将文献上传从本地上传至平台。本模块主要包括两种上传方式：单文献上传和多文献上传，分别用以处理不同用户的不同需求。图 4-9 展示了用户可以对不同的上传方式进行选择。

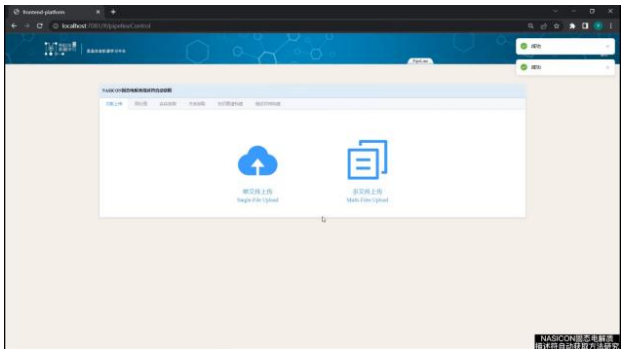

**图 4-9 上传方式选择**

无论是单文献上传还是多文献上传，都需要对文献的基本元信息进行填写。包括标题、PDF 格式的文件命名、作者、摘要、关键字、发表日期、DOI 等，其中对于必填项如不进行填写将会无法提交。注意，必填项通常为能够唯一表示某文献的信息，如 DOI。

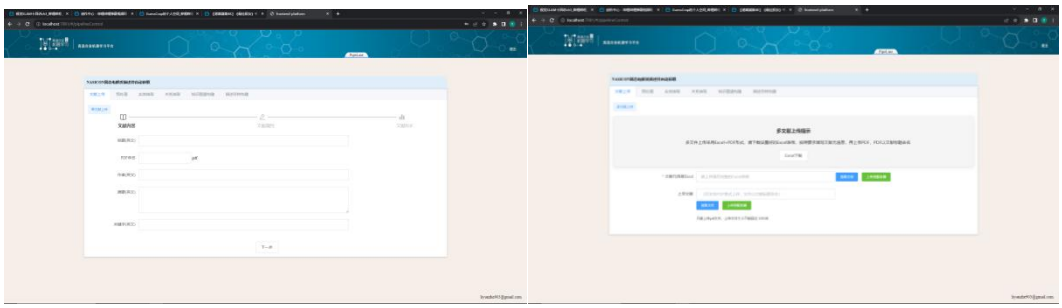

**图 4-10 单文献&多文献上传**

**单文献上传。**单文献会在页面直接填写该文献的详细信息。

**多文献上传。**页面提供一个 excel 模板，用以为用户提供填写多行文件数据的接口，同时，如在 excel 中用户仅填写一行，最后的实际效果和单文献上传是相同的。


**图 4-11 文献上传结果展示**

完成上传即可对上传后的文献信息进行纵览。其中所有的数据全部通过数据库存储和后端进行访问。

## 4.2 预处理

预处理模块主要对已经完成上传的 PDF 模块进行 PDF2TXT 的转换，并对转换后的结果进行分句。


**图 4-12 选择需要预处理的文献**

预处理完成后，往往会出现需要修改或完全无法进一步处理的句子，会在转换完成后，提供修改和删除的接口。

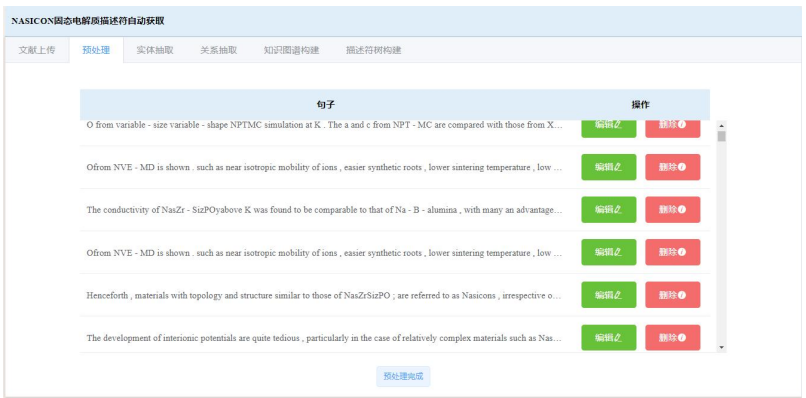


**图 4-13 预处理结果编辑和删除**

## 4.3 实体抽取

完成预处理后可以通过命名实体识别获取实体，如图 4-14 所示，会在页面中展示抽取方法概述和开始按钮。

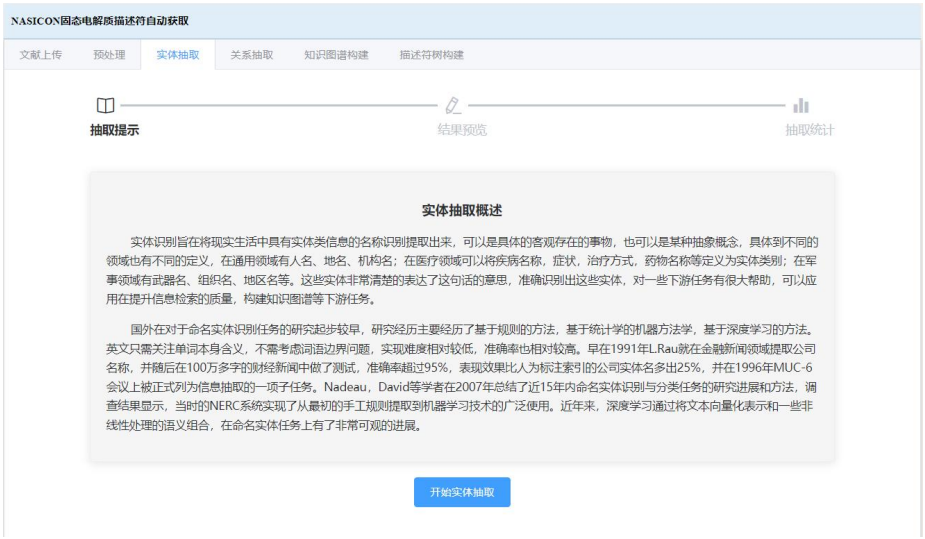


图 4-14 实体抽取页面

**编辑**。由于抽取的过程完全自动，难以一次完美得到想要的结果，因此提供了预览的接口，支持编辑操作。同时，由于数据已经在预处理阶段经过清洗和筛选，故不在此提供删除接口，确保数据的完整性。

**溯源**。对于溯源接口而言，初期并未考虑该功能。在不断地研发和测试过程中，我们发现，对于数据的可靠性，即某个实体、某个关系、某个句子在产生巨大价值时，究竟源自哪一篇文献。实现方式十分简单，但对于这个过程实现的思路却显得尤为重要，如图 4-15 所示。

图 4-15 实体抽取预览、编辑、溯源

**数据统计**。抽取完成后，会对抽取结果每个标签类别的数量进行统计，同时，提供下载接口，便于用户对各类延申研究提供数据，如图 4-16 所示。

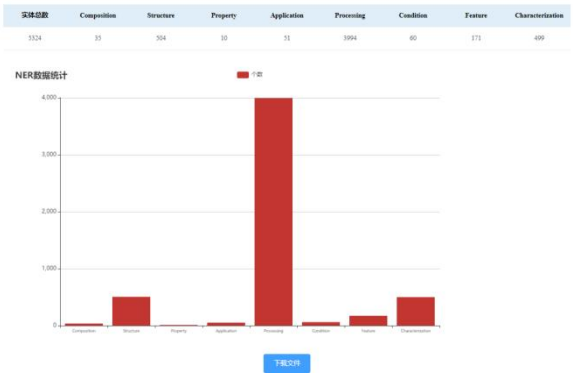


图 4-16 抽取结果数据统计

## 4.4 关系抽取

实体抽取完成后，对于已经完成抽取的实体进行关系抽取。同实体抽取页面类似，会首先展示抽取方法概述，和开始抽取按钮，如图 4-17 所示。

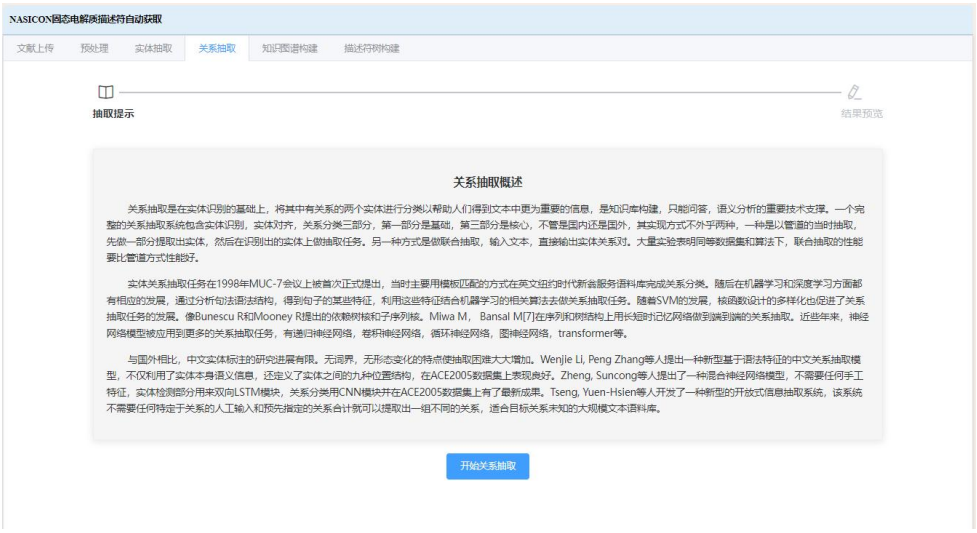


图 4-17 关系抽取概述

**编辑**。由于抽取的过程完全自动，难以一次完美得到想要的结果，因此提供了预览的接口，支持编辑操作。

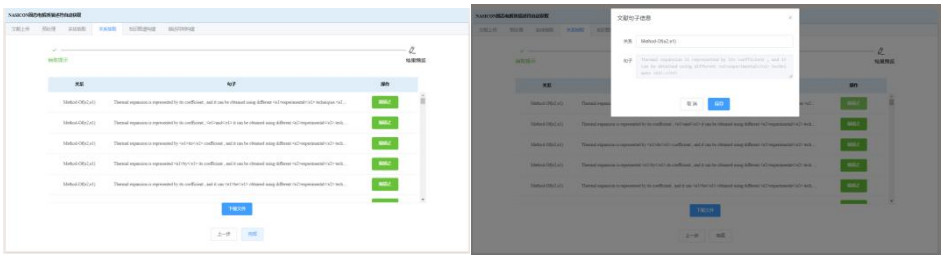

图 4-18 关系抽取结果预览、编辑

## 4.5 知识图谱构建

利用 neo4j 图数据库及 python 相关图数据库操作工具包，实现根据已有实体、关系数据在程序中对数据库的操作，在前端页面展示数据库中图谱，如图 4-19 所示。为后续根据核心属性构建描述符树打下基础。


图 4-19 抽取结果知识图谱展示

## 4.6 描述符树构建

材料专家在多年来对材料属性预测及构效关系的描述符选择研究中构造了一个分层多粒度的描述符逻辑树，该描述符的每个节点在材料上都有自己的含义且它们之间关系从上到下、从粗到细都映射在逻辑树中。通过该树可以为研究人员在材料构效关系的研究中描述符选择问题上提供先验知识，从而大大减少研究人员在描述符选择所耗费的时间。利用 Graphivz 库对知识图谱中的数据进行二次筛选，得到描述符树，如图 4-20 所示。

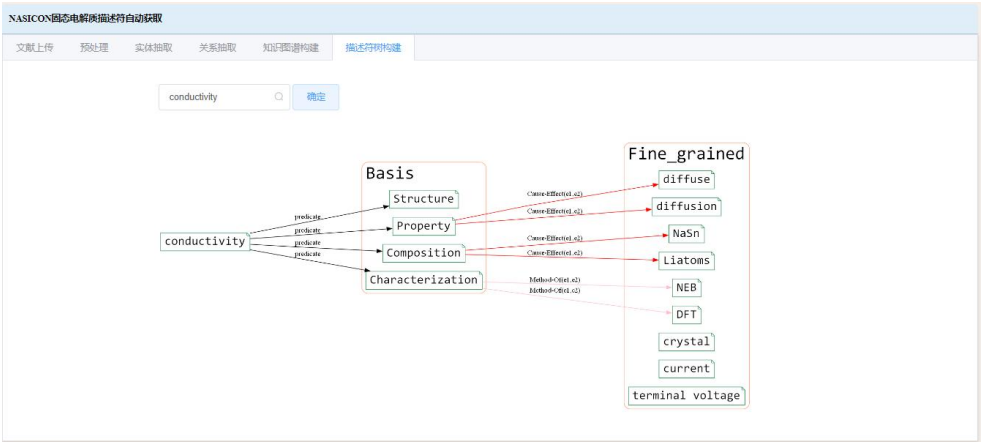


图 4-20 描述符树构建

## 5. 技术难点

**文献预处理方面：**对于 PDF2TXT 的处理始终无法通过程序手段得到完美、无需修改的预处理结果。仍然需要手工清洗。在众多方法中，OCR 能实现相对较好的效果。

**抽取算法方面：**本课题的目前仅对命名实体识别和关系抽取两个任务进行了基本的实现和使用，对于实际场景的真实部署，还需要考虑如并发访问、推理加速等问题，以达到更好的实际使用效率。

**溯源方案方面：**目前的溯源方案已基本实现对于知识图谱中某一节点属于哪一个句子、哪一篇文献的定位，以及可以快速得到该文献的所有元数据。对于未来而言，数据的改动、添加、删除是频繁的，如何做到记录和溯源完整改动、添加、删除的日志信息和版本是一大难点。

**知识图谱展示方面：**目前已经实现了前端页面展示知识图谱，但效果较为单一，仅能完成拖拽等操作，对于节点和节点延申信息的可交互性是未来值得改进的地方。

## 6. 创新点

在预测激活能的过程中，描述符是一种有利的特征。大量有关预测激活能的描述符存在于 NASICON 型固态电解质相关的文献中。目前，绝大部分领域专家只能依靠亲自阅读相关文献，结合自身领域知识，手动提取描述符，然后再进行预测工作。这样的过程存在以下问题：其一，NASICON 型固态电解质相关文献规模大、来源广泛，手动提取会消耗极大的时间、人力成本；其二，个体知识存在差别，考虑到描述符的选择至关重要，所以只是依靠人力难以及时获取最新的知识，导致提取描述符的准确度降低；其三，在海量数据、大数据的时代背景下，数据的伪造、篡改变得愈发容易，而高质量数据又是文本挖掘的前提，同时，提取的描述符无法定位其来源及其来源关联的信息，数据可靠度难以保证，拓展研究难以开展。

针对上述三个问题，本项目针对目前 NASICON 型固态电解质描述符的获取方式及溯源机制进行创新，提出面向 NASICON 型固态电解质的描述符自动获取的方法。

自动获取描述符的主要流程是：先对 NASICON 型固态电解质的文献预处理，拆分为以句子为单位的输入对象；将句子输入经过本下游任务数据训练和微调的命名实体识别（NER）模型，得到对句子中实体识别输出；将识别的实体输入经

过本下游任务数据训练和微调的关系抽取（RE）模型，得到对应实体之间的关系输出；将实体和关系通过<实体名,实体类型,句子>和<实体 1,实体 2,关系名,句子>的形式存入实体库和关系库，实现 NASICON 型固态电解质描述符的 pipeline 流水线化自动获取，解决手动提取费时费力和由于知识背景不足导致的人为选择描述符困难的问题。

定位描述符来源和保证可靠度的主要方式是定义了可靠溯源链 RTC(Reliable Traceability Chain)模型，将整个任务看作链条以表征该任务下每一个对象之间的关联，其中的每一个溯源对象称为链板（link plate），可以用如下四元组描述：

$$< O\text{bject}, P\text{in}, I\text{dentifier}, M\text{echanism}(P\text{in}) > \quad (2)$$

其中$O$bject 是链板的具体内容，可以是实体名、句子名、文献名等；$P\text{in} = \{\text{pin}^1, \text{pin}^2, \ldots \text{pin}^n\}$为每一个链板的销，用以关联和溯源其他多个链板的信息，可以是实体所在源句、源文献的$I$dentifier；$I$dentifier 是链板的唯一标识符，以区分不同的链板，也作为$P$in 的关联标识，可以是所属句子的 ID，所属文献的 DOI 等；$M\text{echanism}(P\text{in})$是基于$P$in 的关联或溯源机制，在材料文本挖掘过程中，RTC 模型的溯源对象 Link Plate 中$M\text{echanism}(P\text{in})$被实例化为基础溯源与过程溯源两个部分，基础溯源着重描述文献、句子的基本元信息，包括但不限于文献标题、作者、摘要、关键字、文本内容、收录期刊、发表日期、影响因子、DOI 和存储路径等；过程溯源着重描述在 pipeline 流水线化自动获取过程中每一步的加工信息和改动记录，包括但不限于模型的输入输出，实体库、关系库结果，删改时间戳等，文献的基础溯源和过程溯源如图 4-21 所示。

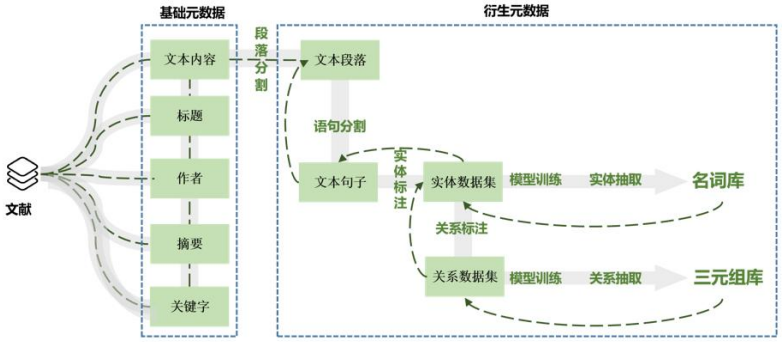


**图 4-21 文献的基础溯源与过程溯源示意图**

RTC 模型使得实验过程中所有的数据获取、中间结果、数据来源、处理过程都有了实时记录，有效地实现了可溯源性治理，较大程度提升了文本挖掘对于 NASICON 型固态电解质描述符处理的可信度。同时，该治理模型具有良好的泛化性，只需更换对应的下游任务数据和定义对应的溯源机制，即可迁移至其他材

料领域进行有效应用。

## 五、应用场景或演示效果

本项目针对目前 NASICON 型固态电解质描述符的获取方式进行创新，提出面向 NASICON 型固态电解质的描述符自动获取的方法。通过搭建知识平台，实现对上传文献进行全流程的处理。可以对文献进行预处理、命名实体识别、关系抽取、抽取结果编辑、抽取结果统计、抽取结果展示、描述符知识图谱构建、描述符树构建。自动、高效地获取描述符，并实现根据所需预测的热点属性构建描述符树。

对于上述每一步的中间结果，都采用了 RTC 模型对数据进行标记，支持包括 MySQL，JSON，csv 等多种形式的存储，在抽取结果展示页面下，对于“溯源”信息有单独的展示弹窗便于查看。根据执行过程全流程，构建体系化、直观化的操作平台，实现上传文献-预处理-实体抽取-关系抽取-知识图谱构建-描述符树构建的全功能及展示平台，便于材料专家对机器学习预测结果进行分析。

## 六、成果分享

GitHub 仓库链接：README 中包含了 Bilibili 演示视频链接

https://github.com/LIYunzhe1408/Research-on-automatic-descriptor-acquisition-method-for-NASICON-solid-electrolyte

## 参考文献

[1] 邓祥宇. NASICON 型固态电解质 Li_（1.3）Al_（0.3）Ti_（1.7）（PO_4）_3 的制备改性及电化学性能研究[D].湖北工业大学,2021.

[2] 缪日健. NASICON 型固态电解质的制备及其电化学性能研究[D].广东工业大学,2021.

[3] 王东.基于深度学习的实体关系抽取方法研究[D].齐鲁工业大学,2021.

[4] 孙镇, 王惠临. 命名实体识别研究进展综述. 现代图书情报技术, 2010, 26(6): 42-47

[5] Roy, Arya. Recent Trends in Named Entity Recognition (NER), 2021.

[6] Razvan Bunescu, Raymond Mooney. A Shortest Path Dependency Kernel for Relation Extraction, 2005

[7] Makoto Miwa, Mohit Bansal. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures,2016

[8] Li, W., Zhang, P., Wei, F., Hou, Y., & Lu, Q. A novel feature-based approach to Chinese entity relation extraction, 2008

[9] Suncong Zheng, Hongyun Bao, J. Xu, Yuexing Hao, Zhenyu Qi and H. Hao, "A Bidirectional Hierarchical Skip-Gram model for text topic embedding," 2016 International Joint Conference on Neural Networks (IJCNN), 2016, pp. 855-862, doi: 10.1109/IJCNN.2016.7727289.

[10] Tseng, Yuen-Hsien & Lee, Lung-Hao & Lin, Shu-Yen & Liao, Bo-Shun & Liu, Mei-Jun & Chen, Hsin-Hsi & Etzioni, Oren & Fader, Anthony. Chinese Open Relation Extraction for Knowledge Acquisition. 10.3115/v1/E14-4003, 2014

[11] Kim E, Huang K, Saunders A, et al. Materials synthesis insights from scientific literature via text extraction and machine learning[J]. Chemistry of Materials, Vol.29, No.21, 2017, pp.9436-9444

[12] Swain M C, Cole J M. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature[J]. Journal of Chemical Information and Modeling, Vol.56, No.10, 2016, pp.1894-1904.

[13] Weston L, Tshitoyan V, Dagdelen J, et al. Named entity recognition and normalizationapplied to large-scale information extraction from the materials scienceliterature[J].Journal of Chemical Information and Modeling, Vol.59, No.9, 2019, pp.3692-3702.

[14] Wang W, Jiang X, Tian S, et al. Automated pipeline for superalloy data by text mining[J]. npj Computational Materials, Vol.8, No.1, 2022, pp.1-12

[15] Nie Z, Zheng S, Liu Y, et al. Automating materials exploration with a semantic knowledge graph for li-ion battery cathodes[J]. Advanced Functional Materials, 2022, pp.2201437

[16] Lowe D M, O'Boyle N M, Sayle R A. Efficient chemical-disease identification and relationship extraction using wikipedia to improve recall[J]. Database, Vol.2016,No.4,2016.

[17] Swain M C, Cole J M. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature[J]. Journal of Chemical Information and Modeling, Vol.56, No.10, 2016, pp.1894-1904.

[18] Hawizy L, Jessop D M, Adams N, et al. Chemicaltagger: A tool for semantic textmining in chemistry[J]. Journal of Cheminformatics, Vol.3, No.1, 2011, pp.1-13

[19] Miller J J. Graph database applications and concepts with neo4j[C]. //Proceedings of the Southern Association for Information Systems Conference. 2013, pp. 1-7.