基于 SedERNIE-BiLSTM-Attention-CRF 模型的沉积学文本三元组提取

刘晓恩

长江大学地球科学学院，湖北武汉，430100；

摘要：构建沉积学知识图谱对整合海量数据、揭示沉积环境、沉积物特征与地质过程的复杂关系至关重要。针对沉积学专业术语多、句子结构复杂导致实体及属性关系联合抽取困难的问题，本文提出基于 SedERNIE-BiLSTM-Attention-CRF 模型的三元组提取方法。该方法利用 SedERNIE 增强领域语义表示，结合 BiLSTM 捕捉文本双向依赖，引入 Attention 聚焦关键信息，并通过 CRF 全局优化实体及其属性以确保准确性；识别实体后，按潜在关系拼接相关实体对文本输入模型分类形成三元组。实验显示，该方法准确率达 94.8%、召回率 94.0%、F1 值 94.3%，优于对比模型，为高质量沉积学知识图谱构建提供了可靠技术支持。

关键词：知识图谱；沉积学领域；三元组提取；SedERNIE

DOI：10.69979/3041-0673.26.02.109

随着沉积学研究的深入，海量文献和数据涌现，如何从中有效提取、整合知识成为关键问题。沉积学涉及复杂的沉积环境、岩石类型、地质过程等信息，这些信息多以自然语言文本形式存在。构建高效、准确的自动化信息提取方法，对于从这些非结构化数据中获取知识、构建沉积学知识图谱至关重要。

在沉积学领域，机器阅读技术（Geodeepdive）较早应用于地质数据库构建，成功实现了叠层石相关词汇及岩石地层名称的自动提取。针对中文沉积学文献，研究者通过中文分词和词频统计提取关键字，并用知识图谱可视化呈现。随着技术进步，基于深度学习的命名实体识别技术逐步应用于地质领域信息提取。刘鹏等人提出了基于 BiLSTM-CRF 改进的 Lattice-LSTM 模型用于煤矿领域知识提取；周永章等基于斑岩型铜矿床概念模型研究了矿床知识获取方法。这些工作的核心都是从非结构化文本中准确提取事实，在知识图谱中以三元组形式表示。

目前，沉积学实体识别主要有基于词典规则和机器学习两类方法。传统机器学习方法以条件随机场（CRF）为主，Wang Chengbin 等基于地质词典采用 CRF 开展沉积学文献分词和词频统计，在灾害领域实现 F1 值 72.55%。近年来，深度学习方法成为主流，DBN、BiLSTM-CRF、Lattice-LSTM、BiGRU-CRF 等模型在地质实体识别中均取得 90%以上的 F1 值。大规模预训练语言模型 BERT 虽在 NLP 任务中表现良好，但其字符级掩码机制缺乏全局建模能力，难以准确识别符号分隔的连续实体。为此，Sun 等提出 ERNIE 模型，增加短语和实体级掩码机制，更适合中文命名实体识别。但 ERNIE 泛化能力不足，对专业领域适应性有待提高。实体识别后，关系和属性的准确提取仍是难点，而沉积学领域的实体与属性联合提取工作开展较少。

针对上述问题，本文提出基于 SedERNIE-BiLSTM-Attention-CRF 模型的沉积学文本三元组提取方法。首先，SedERNIE 通过领域预训练深度建模沉积学语义特征，提升领域词汇理解能力；其次，结合 BiLSTM 捕捉文本双向语义依赖，引入注意力机制突出关键信息；最后，使用 CRF 进行序列标注优化。在识别实体后，依据实体对间潜在关系进行拼接分类，形成新输入样本以识别关系类型，从而有效捕捉文本中的关系信息，增强知识图谱构建和语义理解深度，确保三元组提取的准确性和一致性。

1 SedERNIE-BiLSTM-Attention-CRF 模型

SedERNIE-BiLSTM-Attention-CRF 模型框架由 SedERNIE 沉积学领域预训练模型、BiLSTM 网络、Attention 层与 CRF 层组成。

该模型利用 SedERNIE 进行沉积学领域语义建模，结合 BiLSTM 捕捉文本双向长期依赖，引入 Attention 机制动态关注关键信息，最后通过 CRF 优化标签序列全局一致性，提升实体识别准确性率。

1.1 SedERNIE 预训练模型

ERNIE 预训练模型依托 Transformer 架构与自注意力机制，通过多层结构堆叠搭建深度语义表征体系，核心优势在于能够融合文本知识，强化模型的语义表达能力。该模型采用单元整体字符掩码的预训练策略，可自主学习实体固有语义信息与长距离语义关联，相较于 BERT 模型的掩码机制，更适配中文命名实体识别任务。不同于仅依托掩码语言模型完成预训练的 BERT，ERNIE 新增了关系填空等专项预训练任务，有效挖掘文本内部隐藏的语义关联与逻辑关系，进一步提升模型的上下文语义理解能力。但原生 ERNIE 模型面向通用自然语言场景训练，未吸纳专业领域语义特征与先验知识，难以适配沉积学专项研究场景，领域适用性存在明显短板。

针对 ERNIE 模型在沉积学领域适配性不足的缺陷，本文对该模型进行领域化优化升级。研究收集整理沉积学相关学术期刊论文、会议文献、地质勘查报告、专业数据库及线上行业资源，经过清洗、筛选、标准化等一系列数据预处理操作，构建专属的沉积学领域预训练语料库，并基于该领域语料库完成二次预训练，迭代得到沉积学专用预训练模型 SedERNIE。

SedERNIE 模型通过领域专项预训练，充分习得沉积学专业先验语义知识，有效提升了对全新地质命名实体以及各类常规地质实体的识别精准度。在此基础上，本文在 SedERNIE 分词环节引入自定义地质领域核心词表，解决了符号分隔、组合式地质命名实体分词错乱的问题，保障模型能够完整捕捉复合型地质实体的语义特征与知识信息，全方位优化模型对地质命名实体的识别性能。

1.2 BiLSTM 网络结构

RNN（循环神经网络）广泛应用于许多 NLP（自然语言处理）任务中。LSTM（长短期记忆）是 RNN 的一种变体，它可以有效克服基于 RNN 对长期依赖关系造成的梯度爆炸和梯度损失。BiLSTM 的基本思想是通过 LSTM 的两个隐藏层从输入序列中获取上下文信息。最后，将 LSTM 的两个隐藏向量连接起来创建上下文向量$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$。在本文中，使用 BiLSTM 来提取上下文中的特征。

1.3 注意力机制层

BiLSTM 网络能够缓解传统模型的长期记忆缺陷，完成文本全局特征的提取工作，但面对沉积学文本时存在明显局限性，无法有效捕捉文本内的长距离语义依赖，且在长文本场景下容易丢失关键的局部细节特征。为弥补 BiLSTM 在局部特征提取层面的短板，本文引入注意力机制，量化文本字符与上下文之间的关联度，针对性解决沉积学实体字符跨度大带来的长距离依赖难题。该机制可自主提升沉积学相关实体语义特征的权重，显著优化模型对文本局部特征的捕捉能力。与此同时，注意力机制能够挖掘自然语言问句词汇与属性词汇间的潜在语义关联，引导模型聚焦文本核心有效信息，过滤冗余信息。在实验过程中，模型将问句词向量作为输入生成注意力矩阵，完成问句与属性的特征表征后，和原始词向量矩阵进行特征拼接，最终将融合后的向量矩阵输入卷积神经网络。整体结构中，注意力层会对 BiLSTM 输出的特征向量完成动态权重分配，融合两层网络的特征优势，输出兼具全局信息与局部细节的复合特征向量。

1.4 CRF 层

CRF 常用于序列标注任务。他可以在 BiLSTM-Attention 的基础上加入一些约束，确保输出标签之间的序列顺序正确。因此，CRF 层作为最终的输出解码层，用于获取沉积学预测标签序列。给定一组随机变量 X 为观测序列与输出序列 Y，利用条件概率 P(X/Y) 描述 CRF 模型。对于一句文本，$X = \{x_1, x_2, \cdots, x_n\}$ 表示其观测序列，对于输出序列标签 $Y = \{y_1, y_2, \cdots, y_n\}$

2 实验与结果分析

2.1 实验数据与预处理

由于沉积学领域公共数据集匮乏，我们从学术期刊、地质调查报告及在线资源中收集整理了 277260 字沉积学文本，经筛选得到 1500 条样本，包含 6 种预定义属性和 9000 个地质三元组。数据集按 8:2 划分为训练集和验证集，并进行了噪声去除、格式统一等预处理。实体及属性关系设计参考了陈忠良等整理的岩石关系体系。

接着对数据进行了序列标注。沉积岩文本数据采用“BIO+实体命名”方式。B 表示实体起始位置，I 表示实体中间位置，O 表示非实体字符。在本次地质语料中，实体属性采用了“BIO+属性”的标注方式。标注工具选用了开源的 dcooano。使用这种标注方式将联合提取任务转化为序列标注任务。

2.2 实验环境和参数配置

本文采用 Python3.10+Pytorch2.2.1 的环境进行模型的训练与测试。训练过程中使用 earlystop 和 dropout 技术来减少过拟合，Adam 优化器来计算和优化模型训练中的网络参数。实验中引入的 SedERNIE 模型架构，它是一个含有 12 个 Transformer 层、768 维隐层和 12 头多头注意力机制的模型。具体的参数设置如表 2。根据实验，模型大约在 50 个 epoch 内收敛。

2.3 评估指标

在实验中，采用准确率（Precision,P）、召回率（Recall,R）和 F1 分数作为评估指标。它们的定义如下：

$$P = \frac{TP}{TP + FP} \times 100\%$$

$$R = \frac{TP}{TP + FN} \times 100\%$$

$$F1 = \frac{2PR}{P + R} \times 100\%$$

其中，TP 表示预测为正例实际上也是正例，FN 表示预测为负例实际上是正例，FP 表示预测是正例实际上是负例。

2.4 结果与分析

实验结果表明，本文模型相较于 CRF、BiLSTM-CRF、BiLSTM-Attention-CRF 及 ERNIE-BiLSTM-Attention-CRF，F1 值分别提升 22.5%、12.9%、5%和 4.1%。SedERNIE 结合上下文语境自动提取特征，有效解决了未登录词识别问题；BiLSTM 通过捕捉双向语义依赖，提升了易混淆实体的区分能力；SedERNIE 预训练模型学习到多粒度特征，有效解决了实体嵌套等复杂语义问题，展现出更强的沉积学领域适应能力。

三元组提取本质上是两个阶段任务。第一阶段的实体和属性抽取自然会影响第二阶段的三元组抽取，因此有必要先分析这两个子模块的性能。在实体识别方面，本文在自建数据集上的 F1 值达到了 93.1%。这表明使用BERT大参数量的预训练模型在实体提取中的表现出色。在属性提取方面，模型在自建数据集上的 F1 值达到了 94.1%。这表明基于注意力机制的属性提取子模块性能良好。

3 结论

本文提出基于 SedERNIE-BiLSTM-Attention-CRF 模型的沉积学文本三元组提取方法。SedERNIE 增强领域语义表示，BiLSTM 捕捉双向依赖，Attention 聚焦关键信息，CRF 实现全局优化。实验表明该方法在三元组提取的精度和召回率上均优于传统方法，为沉积学知识图谱构建提供了技术支撑。未来研究将着力于引入更多领域知识、实现跨学科共享，并提升模型对动态数据的自适应更新能力。

参考文献

[1]刘鹏,叶帅,舒雅,等.煤矿安全知识图谱构建及智能查询方法研究[J].中文信息学报,2020,34(11):49-59.

[2]周永章,张前龙,黄永健,等.钦杭成矿带斑岩铜矿知识图谱构建及应用展望[J].地学前缘,2021,28(3):67-75.

[3]王万良.人工智能及其应用[M].高等教育出版社,2016.

[4]杜志强,李钰,张叶廷,等.自然灾害应急知识图谱构建方法研究[J].武汉大学学报(信息科学版),2020,45(9).

[5]张雪英,叶鹏,王曙,等.基于深度信念网络的地质实体识别方法[J].岩石学报,2018,34(2):343-351.

[6]杨森淇,段旭良,肖展,等.基于 ERNIE+DPCNN+BiGRU 的农业新闻文本分类[J].计算机应用,2023,43(5):1461-1466.

[7]齐浩,董少春,张丽丽,等.地球科学知识图谱的构建与展望[J].高校地质学报,2020,26(1):2-10.

[8]陈忠良,袁峰,李晓晖,等.基于 BERT—BiLSTM—CRF 模型的中文岩石描述文本命名实体与关系联合提取[J].地质论评,2022,68(2):742-750.