



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

陈忠良^{1,2)}, 袁峰¹⁾, 李晓晖¹⁾, 张明明¹⁾

1) 合肥工业大学资源与环境工程学院, 合肥, 230009; 2) 安徽省地质调查院, 合肥, 230001

内容提要: 地质调查正在从“数字化”走向“智能化”, 需要在大数据思维的指导下, 面向非结构化数据开展机器阅读和地质知识的自动提取。地学命名实体和关系联合提取是当前研究的难点和核心。本文采用基于大规模预训练中文语言模型的 BERT—BiLSTM—CRF 方法开展岩石描述文本命名实体与关系联合提取。首先, 通过收集数字地质填图工作中的剖面测量和路线地质观测数据, 建立岩石描述语料; 然后, 在岩石学理论指导下分析岩石知识组成, 完成岩石知识图谱命名实体与关系的模式设计, 标注岩石语料; 最后, 开展岩石描述语料知识提取的深度学习训练和消融试验对比。试验结果显示, 大规模预训练中文语言模型 (BERT) 对岩石描述语料知识提取具有较高的适用性。推荐的 BERT—BiLSTM—CRF 模型方法对岩石命名实体与关系联合提取的准确率 ($F1$ 值) 为 91.75%, 对岩石命名实体识别的准确率 ($F1$ 值) 为 97.38%。消融试验证明基于 BERT 的词嵌入层对岩石描述知识提取的性能提升影响显著, 双向长短时记忆网络模型层 (BiLSTM Layer) 能提升实体关系联合提取性能。

关键词: 大数据思维; 深度学习; 预训练中文语言模型; 命名实体识别; 关系提取

随着大数据思维和人工智能技术在地学研究中的不断应用 (周国玉等, 2020; 黄敬军等, 2020; 刘传正和陈春利, 2020; 周永章等, 2021a), 基于自然语言处理技术的机器阅读技术 (Peters et al., 2014; 蒋璟鑫等, 2020) 和地学领域知识提取技术 (Abu-Salih, 2021; 周永章等, 2021b) 也越来越受到地学研究者的关注。

在地学领域, 针对文献的机器阅读技术 (Geodeepdive) 较早应用于以沉积学为主的地质数据库 (Macrostrat; Peters et al., 2014)。在叠层石的时空分布研究中, 机器阅读技术成功从文献中智能提取了叠层石词汇及其所处岩石地层名称 (Peters et al., 2017)。对于非结构化的中文地学文献, 学者开展了基于中文分词和词频统计的文献关键字提取, 采用知识图谱的图方式展示了文献的内蕴信息 (Zhu Yueqin et al., 2017; Wang Chengbin et al., 2018)。基于深度学习的命名实体识别技术也被应用于从文献中提取信息构建地质灾害知识图谱 (Fan Runyu et al., 2019)。刘鹏等 (2020) 引入基于

BiLSTM—CRF 改进的网络结构模型 Lattice—LSTM 进行了煤矿领域知识提取。周永章等 (2021b) 依据斑岩型铜矿床概念模型, 开展华南钦州—杭州成矿带内典型矿床的知识获取、标注和提取。机器阅读技术和地学领域知识提取技术关注的同为如何从非结构化的数据 (如文本) 中提取出事实。在知识图谱中事实既是知识, 是以实体及其关系组成的三元组 (王万良, 2020; Ji Shaoxiong et al., 2021)。目前, 在地学领域知识的自动获取和图谱构建的过程中, 实体识别是其重要内容, 关系提取则是其中的难点和核心 (齐浩等, 2020; 周永章等, 2021b)。

当前, 地质调查正在从“数字化”走向“智能化” (李超岭等, 2015)。建设中的地质调查智能空间以“需求+数据+知识+智能驱动”理念为指导, 研发空间数据自适应感知服务、非结构化数据挖掘服务、地质知识的流程化和智能化应用等 (李丰丹等, 2019)。在非结构化数据挖掘方面, 现有智能空间平台已开展基于位置和关键词的地质报告信息检索技术研究, 基于机器学习的非结构化数据挖掘工作

注: 本文为国家自然科学基金资助项目 (编号: 41820104007, 42072321, 41872247) 的成果。

收稿日期: 2021-07-07; 改回日期: 2022-01-05; 网络首发: 2022-01-20; 责任编辑: 章雨旭。Doi: 10.16509/j.georeview.2022.01.115

作者简介: 陈忠良, 男, 1984 年生, 博士研究生, 高级工程师, 长期从事地质调查与地学信息技术研究; Email: c_mulder@163.com。通讯作者: 袁峰, 男, 1971 年生, 教授, 博导, 矿物学、岩石学、矿床学专业; Email: yf_hfut@163.com。

是需要开展的研究内容(Wu Liang et al., 2017)。在地质调查智能空间平台中,岩石描述文本是除照片、地质报告文档外重要的非结构化数据之一。当前,基于机器学习的岩石和矿物智能识别研究正在逐渐深入(张野等,2018;徐述腾和周永章,2018;任伟等,2021),但针对岩石描述文本的机器阅读和知识提取技术研究却尚未开展。岩石描述文本与地质报告的显著区别是其短文本特点,非常适合开展基于大规模预训练语言模型的深度学习。如 BERT 预训练语言模型的最大支持标记序列为 512 个字词(Devlin et al., 2018)。同时,岩石描述文本中命名实体和关系明确,也有利于人工语料标注并开展实体和关系的联合提取。

本文针对地学领域知识的自动获取,特别是关系提取这一难点问题,以智能地质调查空间中岩石描述这一短文本为研究对象,开展基于深度学习的岩石描述文本命名实体和关系联合提取方法研究。分为:中文语料库与词嵌入技术、地学命名实体识别和关系提取的相关研究工作;岩石实体与关系的模式设计、岩石描述语料的收集和标注以及基于 BERT—BiLSTM—CRF 模型的知识提取方法;试验结果和讨论,开展了推荐模型方法的消融试验,比较分析了模型中各部分对岩石知识提取性能的影响。

1 相关工作

开展地学领域知识提取方法研究主要涉及:① 中文地学语料库与地学领域词嵌入技术;② 地学命名实体识别与关系提取。

1.1 中文地学语料库与词嵌入技术

向量是理解和表示文本数据的数学方法。词嵌入技术(Word Embeddings),是将文本的每个字(单词)封装成向量表述的一种技术,是机器学习算法开展自然语言处理的基础。当前,地学领域词嵌入技术主要有 GeoVec,其与通用词嵌入技术相比,在多项地学英文语言处理任务中取得了更好的效果(Padarian and Fuentes, 2019; Fuentes et al., 2020)。然而,地学领域中文语料库和中文地学领域词嵌入技术却相对匮乏。由于本次以中文地学知识提取作为研究对象,在地学领域词嵌入技术缺乏的情况,通用中文词嵌入技术成为可选方案,如 Word2Vec(Mikolov et al., 2013)、BERT(Devlin et al., 2018)等。特别是后者作为一个 Word2Vec 的替代者,在自然语言处理领域的多个方向大幅刷新了精度。通用中文词嵌入技术在地学知识提取中的

应用效果成为本次研究的内容之一。

1.2 地学命名实体识别与关系提取

目前地学领域命名实体识别主要有基于词典及规则的方法和基于机器学习的方法。对于非结构化的中文地学文献,学者开展了基于扩展地质词典及规则的地质知识提取(Zhu Yueqin et al., 2017)。而传统的机器学习方法应用于地学命名实体提取的主要为条件随机场模型(CRF)。Wang Chengbin 等(2018)基于地质词典采用 CRF 模型开展了地学文献的中文分词和词频统计。条件随机场模型对灾害领域命名实体提取实现 $F1$ 值 72.55% 的识别结果(杜志强等,2020)。近年,深度学习方法提取特征逐渐成为主流(周永章等,2018),如 DBN、BiLSTM—CRF、Lattice—LSTM、BiGRU—CRF、ELMO—CNN—BiLSTM—CRF 模型等。DBN 模型在小规模矿产资源地质调查报告语料的地质实体识别评估中,各项评估指标($P, R, F1$)均取得了 90% 以上(张雪英等,2018)。BiLSTM—CRF 模型及其改进的 Lattice—LSTM 模型对煤矿领域命名实体分别取得了 $F1$ 值 91.94% 和 94.04% 的识别结果(刘鹏等,2020)。BiGRU—CRF 模型对地质灾害命名实体识别也取得了 $F1$ 值 94.19 的识别结果(Fan Runyu et al., 2019)。

基于深度学习的方法对地学命名实体识别展现了较好的识别效果,但地学命名实体与关系联合提取目前工作开展较少。随着大规模预训练语言模型 BERT 在自然语言处理领域众多任务中取得最优结果(Devlin et al., 2018),其对地学数据集的命名实体识别特别是实体与关系的联合提取的适应性研究尚未开展。

2 材料与方法

2.1 岩石命名实体与关系的模式设计

岩石观察和描述的内容一般包括颜色、构造、结构、矿物成分的种类和含量,以及依据岩石的分类命名原则对岩石命名。岩石知识图谱是节点和边(有向边)组成的知识三元组表现形式。节点,即实体,内容包含岩石、地层代号、颜色、结构、构造、矿物、第四纪沉积物、接触关系等。边,即关系,是各节点之间广泛的知识关联,如主要矿物、次要矿物、新鲜色、风化色等。图 1 为岩石知识命名实体与关系的模式设计元图。接触关系为岩石之间的接触关系,多为岩石描述语句之外单独语句描述。本次考虑描述语句字数限制,暂对接触关系描述语句只做命名实体

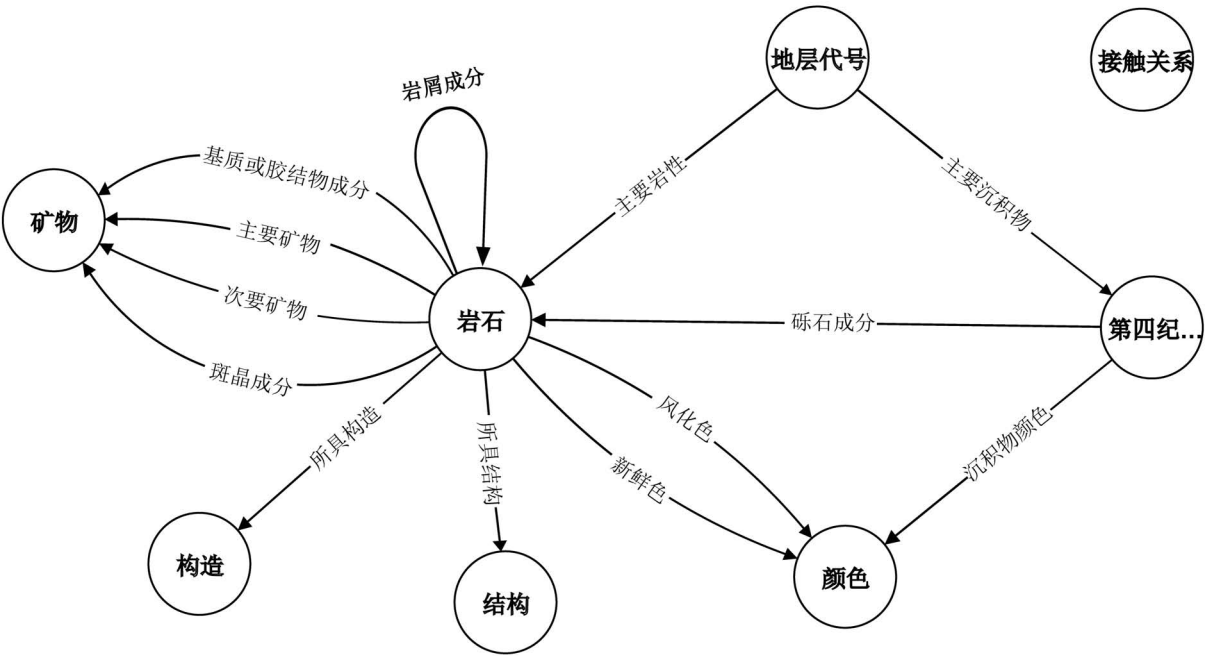


图 1 岩石知识图谱命名实体与关系的模式设计元图
Fig. 1 Meta-graph for named entities and relations of the domain—specific knowledge graph of petrology

设计。

2.2 数据来源及语料标注

岩石知识来源众多,可分为结构化数据、半结构化数据和非结构化数据。其中,非结构化数据以岩石学文献和岩石描述文本为主,是岩石知识提取的主要研究对象之一。特别是岩石描述文本,其是数字填图系统和研发中的智能地质调查系统(李超岭等,2015;李丰丹等,2019)野外采集人员输入的主要数据,成为了本次岩石知识提取的主要语料资源。本次岩石描述文本所描述的岩石类型涵盖了岩浆岩、沉积岩、变质岩和第四纪沉积物。

岩石描述语料标注为“**BIO+命名实体**”方式。其中,B代表命名实体片段的开始;I代表实体片段的中间;O代表字符不为任何实体。由于岩石描述多为围绕某种岩石或第四纪沉积物展开,本次岩石描述语料中的实体关系同样采用“**BIO+关系**”的标注方式。标注工具采用开源的 BRAT(Stenetorp et al., 2012)。采用这种方式,实体与关系联合提取任务将转换成序列标注任务。图 2 为岩石描述语料命名实体和关系标注示例。对于“风化色”这一“岩石”与“颜色”命名实体间的关系,一般岩石描述中均在颜色实体前有“风化”字词,具有一定的前后文语义特征。其它实体关系,如主要矿物、次要矿物

等,均具有类似的前后文语义特征。

最终选取了 300 个地质点的岩石描述开展了标注工作。BIO 标注语料 771 个句子。为了测试命名实体识别和实体与关系联合提取的差异,本次对语料分为两个版本,V1 版为只标注岩石命名实体,V2 版则是岩石命名实体与关系同步标注。两版语料均以 8 : 1 : 1 的比例划分为训练集、验证集和测试集。V1 版岩石命名实体提取语料共标记 771 个句子。其中,629 个句子训练集,66 个句子验证集,76 个句子测试集。V2 版岩石命名实体与关系联合提取语料同样标记 771 个句子。其中,620 个句子训练集,73 个句子验证集,78 个句子测试集。

2.3 基于 BERT—BiLSTM—CRF 模型的岩石命名实体识别与关系提取

本次采用基于大规模预训练中文语言模型的 BERT—BiLSTM—CRF 模型开展岩石描述文本命名实体与关系联合提取。模型结构如图 3 所示。主要包含基于 BERT 的词嵌入层 (BERT—Embedding Layer),双向长短时记忆网络模型层 (BiLSTM Layer)和条件随机场模型层 (CRF Layer)。

基于 BERT 的词嵌入层,首先利用基于大规模中文语料预训练 BERT 中文语言模型输出的字典文件,将输入的岩石描述语句逐字映射转换为字符编

原句	灰 白 色 黑 云 斜 长 片 麻								
序列标注	B-新鲜色 I-新鲜色 I-新鲜色 B-岩石 I-岩石 I-岩石 I-岩石 I-岩石 I-岩石								
原句	岩 , 岩 石 风 化 呈 灰 黄								
序列标注	I-岩石 O O O O O O B-风化色 I-风化色								
原句	色 , 鳞 片 变 晶 结 构 ,								
序列标注	I-岩石 O B-所具结构 I-所具结构 I-所具结构 I-所具结构 I-所具结构 I-所具结构 O								
原句	片 麻 状 构 造 。 矿 物 成								
序列标注	B-所具构造 I-所具构造 I-所具构造 I-所具构造 I-所具构造 O O O O								
原句	分 有 斜 长 石 5 0 % ,								
序列标注	O O B-主要矿物 I-主要矿物 I-主要矿物 O O O O								
原句	石 英 3 0 % , 钾 长 石								
序列标注	B-主要矿物 I-主要矿物 O O O O B-主要矿物 I-主要矿物 I-主要矿物								
原句	1 5 % , 少 量 黑 云 母 。								
序列标注	O O O O O O B-次要矿物 I-次要矿物 I-主要矿物 O								

图 2 岩石描述语料命名实体和关系标注示例

Fig. 2 A example for the annotations of the named entities and relations on lithological description corpus

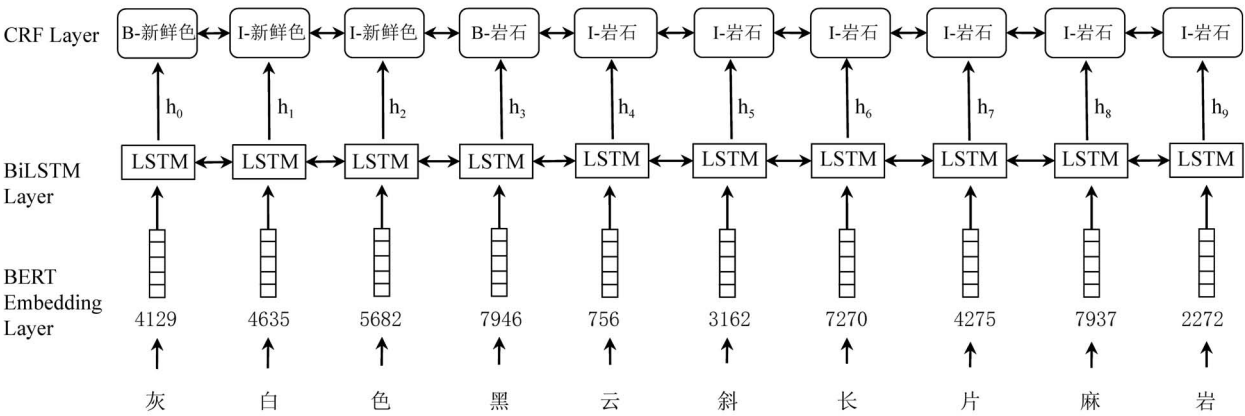


图 3 BERT—BiLSTM—CRF 模型结构示意图

Fig. 3 The schematic diagram of the BERT—BiLSTM—CRF model architecture

码。然后字符编码进入预训练参数初始化的 BERT 模型层转换为词向量输出。本层主要学习输入的岩石描述语句中每个字和符号到对应的岩石命名实体和关系标签的规律(Devlin et al. , 2018)。

双向长短时记忆网络模型层由一个正向和一个反向长短时记忆网络(LSTM)组成。该层主要学习岩石描述语句的上下文信息(刘鹏等,2020)。本层输出为岩石语句中每一个字属于不同岩石命名实体

和关系的概率。

条件随机场模型层则是学习岩石描述句子中相邻岩石命名实体和关系标签之间的转移规则(Wang Chengbin et al. , 2018),如“B—岩石”为岩石实体的开头,一般后面会是“I—岩石”。句子的开头应为“B—”或“O—”。“ I—”只出现在句中和句尾。转移规律的学习能够提高预测的准确度。

模型训练实验均在配置有 Quadro P3200 显卡的移动工作站上完成。机器学习平台采用 Tensorflow-gpu 1.13.1。模型训练参数见表 1 所示。批处理尺寸为 2。输入语句最大长度设定为 500(实际语料句子最大长度为 424)。LSTM 模型的隐藏层神经元数量设置为 128。词向量大小设置为 768。模型训练参数中,编译优化器选择 Ruder(2016)推荐的 Adam 自适应优化器,初始学习率为 0.00002,Dropout 设置为 0.5。

表 1 BERT—BiLSTM—CRF 模型训练参数
Table 1 The experiment settings for the training of the BERT—BiLSTM—CRF model

批处理尺寸	输入最大长度	LSTM 隐藏层节点数	词向量维度	初始学习率	Dropout
2	500	128	768	0.00002	0.5

3 结果和讨论

3.1 评估指标

试验结果评估指标采用(Goutte and Gaussier,

2005)定义三个测试指标:查准率(*precision*, *P*)、召回率(*recall*, *R*)和 *F1* 值。根据模型在测试集上的预测结果,其中 *TP* 为能正确识别岩石命名实体和关系标签的个数、*FP* 为能识别出岩石命名实体和关系但标签类别判定出现错误的个数、*FN* 为应该但没被识别的岩石命名实体和关系个数。按照公式(1~3)可以得到 *P*、*R* 和 *F1*($\beta=1$)值。

$$precision = \frac{TP}{TP + FP}$$
 (1)

$$recall = \frac{TP}{TP + FN}$$
 (2)

$$F_{\beta} = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$
 (3)

3.2 测试结果

BERT—BiLSTM—CRF 模型在 V2 版语料上,训练集损失函数 *loss* 值和验证集损失函数 *loss* 值均逐渐减小,趋于收敛,未见明显过拟合(图 4a)。训练后的模型在测试集上查准率 *P* 为 91.83%,召回率 *R* 为 91.67%,*F1* 值为 91.75%(表 4),表现出较好的实体和关系联合提取效果。表 2 为 BERT—BiLSTM—CRF 模型在 V2 版语料上对主要岩石关系的分项提取准确率。在“沉积物颜色”、“新鲜色”、“所具结构”、“所具构造”、“主要矿物”等关系提取方面均实现了超过 95%的 *F1* 值。在“基质与胶结物成分”、“斑晶成分”、“生物碎屑成分”等关系提取方面 *F1* 值均低于 80%,明显效果不理想。这些岩石与矿物之间的关系多存在于斑状结构、陆源碎屑结构和粒屑结构的岩石描述之中。

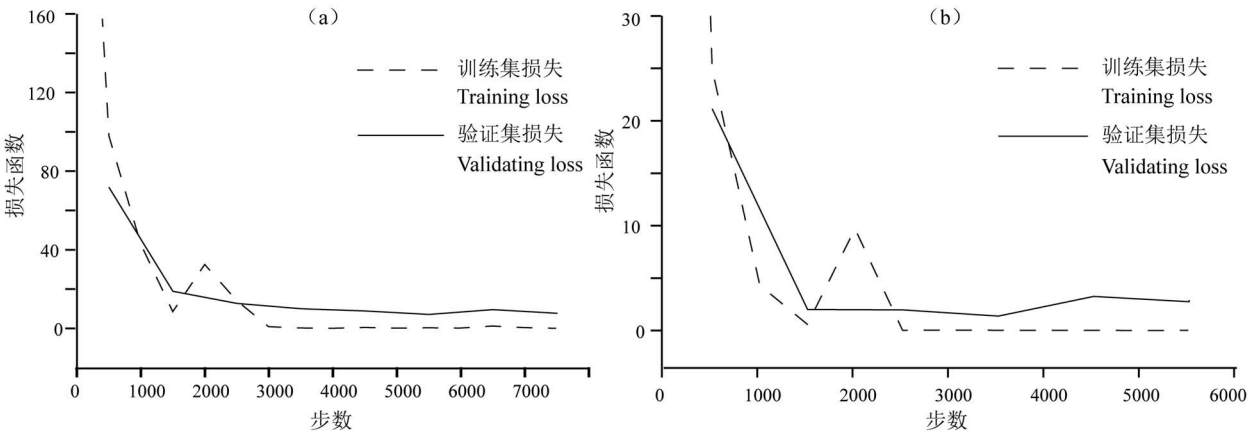


图 4 BERT—BiLSTM—CRF 模型损失函数训练曲线: (a) V2 版语料训练曲线; (b) V1 版语料训练曲线
Fig. 4 Training loss curves of the BERT—BiLSTM—CRF model: (a) training loss curves on the second version corpus; (b) training loss curves on the first version corpus

表 2 BERT—BiLSTM—CRF 模型在 V2 版语料上对主要岩石关系的分项提取准确率

Table 2 Performance of the proposed BERT—BiLSTM—CRF model which was used to extract the major relations from the second version corpus

关系	BERT—BiLSTM—CRF			关系	BERT—BiLSTM—CRF		
	<i>P</i>	<i>R</i>	<i>F1</i>		<i>P</i>	<i>R</i>	<i>F1</i>
主要岩性	75.00	100.00	85.71	所具结构	95.35	97.62	96.47
主要沉积物	100.00	50.00	66.67	所具构造	92.68	100.00	96.20
岩屑成分	100.00	87.50	93.33	主要矿物	97.03	94.23	95.61
斑晶成分	71.43	83.33	76.92	次要矿物	84.00	91.30	87.50
新鲜色	94.52	98.57	96.50	基质或胶结物成分	28.57	66.67	40.00
风化色	86.67	92.86	89.66	砾石成分	100.00	100.00	100.00
沉积物颜色	100.00	100.00	100.00				

表 3 BERT—BiLSTM—CRF 模型在 V1 版语料上对主要岩石命名实体的
分项提取准确率

Table 3 Performance of the BERT—BiLSTM—CRF model which was used to recognize the rock named entities on the first version corpus

命名实体	准确率			命名实体	准确率		
	<i>P</i>	<i>R</i>	<i>F1</i>		<i>P</i>	<i>R</i>	<i>F1</i>
岩石	96.77	95.74	96.26	结构	100.00	100.00	100.00
颜色	98.21	99.10	98.65	构造	100.00	100.00	100.00
第四纪沉积物	100.00	100.00	100.00	矿物	94.59	97.22	95.89
地层代号	93.75	100.00	96.77	接触关系	94.74	100.00	97.30

为了与复杂的关系提取做对比,本次对 BERT—BiLSTM—CRF 模型在 V1 版语料上进行了主要岩石命名实体的识别试验。训练集损失函数 *loss* 值和验证集损失函数 *loss* 值同样逐渐减小,趋于收敛,未见明显过拟合(图 4b)。训练后的模型在测试集上查准率 *P* 为 96.79%,召回率 *R* 为 97.97%,*F1* 值为 97.38%。表 3 为 BERT—BiLSTM—CRF 模型在 V1 版语料上对主要岩石命名实体的分项提取准确率结果。所有命名实体均取得了超过 95%的 *F1* 值,说明 BERT—BiLSTM—CRF 模型对现有语料中岩石命名实体表现出非常好的识别效果。也进一步说明模型是对复杂的岩石物质成分描述知识的提取仍然存在不足。

3.3 消融试验

为了检验 BERT—BiLSTM—CRF 模型中不同部分的作用,本次还开展了针对本文所提模型方法的消融试验。消融试验分别去掉双向长短时记忆网络模型层(BiLSTM Layer)和基于 BERT 的词嵌入层(BERT—Embedding Layer),在 V1 版和 V2 版数据集

上执行 BiLSTM—CRF 模型和 BERT—CRF 模型的实体识别和岩石命名实体与关系联合提取任务。消融试验中的 3 种模型的性能对比见表 4 所示。双向长短时记忆网络模型层(BiLSTM Layer)的去除对 V2 数据集上的实体关系联合提取任务性能稍有影响,*F1* 值从 91.75%降低到 89.45%。但该层的删除对 V1 版数据集上的实体识别任务性能影响不大。去除基于 BERT 的词嵌入层后的 BiLSTM—CRF 模型在两项任务性能上均明显降低,说明此层非常重要。由此可推断,在小规模的地学语料库上,模型中加入大规模预训练中文语言模型 BERT,能够丰富词向语义,更好的学习岩石描述语句中每个字和符号到对应的岩石命名实体和关系标签的规律,可以很好的提升模型的性能。

表 4 不同模型对测试集的预测结果

Table 4 The comparison of the BERT—BiLSTM—CRF model against the BERT—CRF model and BiLSTM—CRF model on rock named entity recognition task and joint entity and relation extraction task

任务	评估指标	BiLSTM—CRF	BERT—CRF	BERT—BiLSTM—CRF
岩石命名实体识别(V1 版语料)	<i>P</i>	69.33	97.57	96.79
	<i>R</i>	75.66	97.57	97.97
	<i>F1</i>	72.36	97.57	97.38
岩石命名实体与关系联合提取(V2 版语料)	<i>P</i>	47.43	89.78	91.83
	<i>R</i>	51.81	89.13	91.67
	<i>F1</i>	49.52	89.45	91.75

4 结论

地质调查正在从“数字化”走向“智能化”,需要在大数据思维的指导下,面向非结构化数据开展机器阅读和地质知识的自动提取。在地质调查智能空间平台中,岩石描述文本是除照片、地质报告文档外重要的非结构化数据之一。岩石描述文本中岩石命名实体的识别,特别是关系提取是本次的研究对象。由于中文地学领域词嵌入技术尚未发布,本文采用通用中文词嵌入技术,针对中文岩石描述语料,开展了基于大规模预训练中文语言模型 BERT—BiLSTM—CRF 方法的岩石领域命名实体识别以及实体与关系联合提取试验研究。试验结果显示,大规模预训练中文语言模型(BERT)对岩石语料知识提取具有较高的适用性,特别是针对岩石描述语料中岩石命名实体提取表现出了较好的识别效果。

由于 BERT—BiLSTM—CRF 模型在斑状结构、陆源碎屑结构和粒屑结构的岩石描述命名实体和关系的联合提取任务上表现出不足,同时模型在单一命名实体识别任务上的良好性能,后续需要继续开展基于流水线模式的实体与关系联合提取研究,并与联合模型方法开展对比。岩石描述标注语料库的数据规模也需要后续进一步提升。大规模中文地学语料库的欠缺制约着地学领域中文词向量技术的开发。共建共享中文地学语料库也是行业内研究者与管理者需要积极推动的工作方向之一。

致谢:感谢审稿专家周永章教授和责任编辑章雨旭研究员提出的宝贵修改意见。感谢安徽省地质调查院王翔高级工程师、吴衡高级工程师、邓佳良高级工程师和黄蒙高级工程师在岩石描述语料收集和标注工作中提供的大力帮助。

参 考 文 献 / References

(The literature whose publishing year followed by a “&” is in Chinese with English abstract; The literature whose publishing year followed by a “#” is in Chinese without English abstract)

杜志强,李钰,张叶廷,谭玉琪,赵文豪. 2020. 自然灾害应急知识图谱构建方法研究. 武汉大学学报(信息科学版), 45(9):1344~1355.

黄敬军,赵增玉,姜素,杨磊,高立,许书刚. 2020. 自然资源管理视角下江苏城市地质调查工作新思考. 地质论评,66(6):1609~1618.

蒋璟鑫,李超,胡修棉. 2020. 沉积学数据库建设与沉积大数据科学研究进展:以 Macrostrat 数据库为例. 高校地质学报,26(1):27~43.

李超岭,李丰丹,吕霞,李健强,刘畅,刘园园. 2015. 地质调查智能空间体系与架构. 测绘学报,44(增刊1):143~151,159.

李丰丹,刘畅,刘园园,吕霞. 2019. 地质调查智能空间框架构建与实践. 地质论评,65(增刊1):317~320.

齐浩,董少春,张丽丽,胡欢,樊隽轩. 2020. 地球科学知识图谱的构建与展望. 高校地质学报,26(1):2~10.

刘传正,陈春利. 2020. 中国地质灾害成因分析. 地质论评,66(5):1334~1348.

刘鹏,叶帅,舒雅,鹿晓龙,刘明明. 2020. 煤矿安全知识图谱构建及智能查询方法研究. 中文信息学报,34(11):49~59.

任伟,张盛,乔计花,黄金明. 2021. 基于深度学习的岩石矿物智能识别. 地质论评,67(增刊1):281~282.

王万良. 2020. 人工智能及其应用(第4版). 北京:高等教育出版社:1~66.

徐述腾,周永章. 2018. 基于深度学习的镜下矿石矿物的智能识别实验研究. 岩石学报,34(11):3244~3252.

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

张野,李明超,韩帅. 2018. 基于岩石图像深度学习的岩性自动识别与分类方法. 岩石学报,34(2):333~342.

周国玉,张明明,沈乐,张淑虹,袁峰,李晓晖,季斌,周宇章. 2020. 铜陵矿集区姚家岭锌金多金属矿床深部地质空间信息相关性数据挖掘. 大地构造与成矿学,44(2):242~250.

周永章,王俊,左仁广,肖凡,沈文杰,王树功. 2018. 地质领域机器学习、深度学习及实现语言. 岩石学报,34(11):3173~3178.

周永章,左仁广,刘刚,袁峰,毛先成,郭艳军,肖凡,廖杰,刘艳鹏. 2021a. 数学地球科学跨越发展的十年:大数据、人工智能算法正在改变地质学. 矿物岩石地球化学通报,40(3):556~573,777.

周永章,张前龙,黄永健,杨威,肖凡,吉俊杰,韩枫,唐磊,欧阳冲,沈文杰. 2021b. 钦杭成矿带斑岩铜矿知识图谱构建及应用展望. 地学前缘,28(3):67~75.

Abu-Salih B. 2021. Domain-specific knowledge graphs: A survey[J][OL]. Journal of Network and Computer Applications, 185: 103076; <https://doi.org/https://doi.org/10.1016/j.jnca.2021.103076>

Devlin J, Chang Mingwei, Lee K, Toutanova K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding [OL]. Retrieved from <http://arxiv.org/abs/1810.04805>.

Du Zhiqiang, Li Yu Zhang Yeting, Tan Yuqi, Zhao Wenhao. 2020&. Knowledge graph construction method on natural disaster emergency. Geomatics and Information Science of Wuhan University, 45(9):1344~1355.

Fan Runyu, Wang Lizhe, Yan Jining, Song Weijing, Zhu Yingqian, Chen Xiaodao. 2019. Deep learning-based named entity recognition and knowledge graph construction for geological hazards[J][OL]. ISPRS international Journal of Geo-Information, 9(1); <https://doi.org/10.3390/ijgi9010015>.

Fuentes I, Padarian J, Iwanaga T, Willem Vervoort R. 2020. 3D lithological mapping of borehole descriptions using word embeddings [J][OL]. Computers & Geosciences, 141: 104516; <https://doi.org/https://doi.org/10.1016/j.cageo.2020.104516>.

Goutte C, Gaussier E. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Losada D E, Fernández-Luna J M. eds. Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, vol 3408. Berlin: Springer; https://doi.org/10.1007/978-3-540-31865-1_25.

Huang Jingjun, Zhao Zengyu, Jiang Su, Yang Lei, Gao Li, Xu Shugang. 2020&. New thoughts on urban geological survey concerning natural resource management in Jiangsu Province. Geological Review, 66(6):1609~1618.

Ji Shaoxiong, Pan Shirui, Cambria E, Martinen P, Yu P S. 2021. A survey on knowledge graphs: Representation, acquisition, and applications[J][OL]. IEEE Transactions on Neural Networks and Learning Systems; 1 ~ 21; <https://doi.org/10.1109/TNNLS.2021.3070843>.

Jiang Jingxin, Li Chao, Hu Xiumian. 2020. Advances on sedimentary database building and related research: Macrostrat as an example. Geological Journal of China Universities, 26(1): 27~43.

Li Chaoling, Li Fengdan, Lü Xia, Li Jianqiang, Liu Chang, Liu Yuanyuan. 2015. The architecture of geological survey intelligent space. Acta Geodaetica et Cartographica Sinica, 44(S1): 143 ~ 151; 159.

Li Fengdan, Liu Chang, Liu Yuanyuan, Lü Xia. 2019. Construction and application of the geological survey intelligent space. Geological Review, 65(S1): 317~320.

Liu Chuanyang, Chen Chunli. 2020. Research on the origins of geological disasters in China. Geological Review, 66(5): 1334 ~ 1348.

Liu Peng, Ye Shuai, Shu Ya, Lu Xiaolong, Liu Mingming. 2020. Coalmine safety: Knowledge graph construction and its QA approach. Journal of Chinese Information Processing, 34(11): 49~59.

Mikolov T, Chen Kai, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. In ICLR.

Padarian J, Fuentes I. 2019. Word embeddings for application in geosciences: Development, evaluation, and examples of soil-related concepts[J][OL]. Soil, 5(2): 177~187; <https://doi.org/10.5194/soil-5-177-2019>.

Peters S E, Zhang Ce, Livny M, Ré C. 2014. A machine reading system for assembling synthetic paleontological databases[J][OL]. Plos One, 9(12): e113523; <https://doi.org/10.1371/journal.pone.0113523>.

Peters S E, Husson J M, Wilcots J. 2017. The rise and fall of stromatolites in shallow marine environments[J][OL]. Geology, 45(6): 487~490; <https://doi.org/10.1130/G38931.1>.

Qi Hao, Dong Shaochun, Zhang Lili, Hu Huan, Fan Junxuan. 2020. Construction of earth science knowledge graph and its future perspectives. Geological Journal of China Universities, 26(1): 2~10.

Ren Wei, Zhang Sheng, Qiao Jihua, Huang Jinming. 2021. The rock and mineral intelligence identification method based on deep learning. Geological Review, 67(S1): 281~282.

Ruder S. 2016. An overview of gradient descent optimization algorithms [OL]. Retrieved from <http://arxiv.org/abs/1609.04747>.

Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In: EACL 2012 - Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; 102~107.

Wang Chengbin, Ma Xiaogang, Chen Jianguo, Chen Jingwen. 2018. Information extraction and knowledge graph construction from geoscience literature[J][OL]. Computers and Geosciences, 112: 112~120; <https://doi.org/10.1016/j.cageo.2017.12.007>.

Wang Wanliang. 2020. Artificial Intelligence (4th edition). Beijing: Higher Education Press; 1~66.

Wu Liang, Xue Lei, Li Chaoling, Lv Xia, Chen Zhanlong, Jiang Baode, Guo Mingqiang, Xie Zhong. 2017. A knowledge-driven geospatially enabled framework for geological big data[J][OL]. ISPRS International Journal of Geo-Information, 6(6); <https://doi.org/10.3390/ijgi6060166>.

Xu Shuteng, Zhou Yongzhang. 2018. Artificial intelligence identification of ore minerals under microscope based on deep learning algorithm. Acta Petrologica Sinica, 34(11): 3244~3252.

Zhang Xueying, Ye Peng, Wang Shu, Du Mi. 2018. Geological entity recognition method based on deep belief networks. Acta Petrologica Sinica, 34(2): 343~351.

Zhang Ye, Li Mingchao, Han Shuai. 2018. Automatic identification and classification in lithology based on deep learning in rock images. Acta Petrologica Sinica, 34(2): 333~342.

Zhou Guoyu, Zhang Mingming, Shen Le, Zhang Shuhong, Yuan Feng, Li Xiaohui, Ji Bin, Zhou Yuzhang. 2020. Data mining of deep geological spatial information of the Yaojialing Zinc-gold polymetallic deposit. Geotectonica et Metallogenia, 44(2): 242~250.

Zhou Yongzhang, Wang Jun, Zuo Renguang, Xiao Fan, Shen Wenjie, Wang Shugong. 2018. Machine learning, deep learning and python language in field of geology. Acta Petrologica Sinica, 34(11): 3173~3178.

Zhou Yongzhang, Zuo Renguang, Liu Gang, Yuan Feng, Mao Xiancheng, Guo Yanjun, Xiao Fan, Liao Jie, Liu Yanpeng. 2021a. The great-leap-forward development of mathematical geoscience during 2010~2019: Big Data and Artificial Intelligence Algorithm are Changing Mathematical Geoscience. Bulletin of Mineralogy, Petrology and Geochemistry, 40(3): 556~573; 777.

Zhou Yongzhang, Zhang Qianlong, Huang Yongjian, Yang Wei, Xiao Fan, Ji Junjie, Han Feng, Tang Lei, Ouyang Chong, Shen Wenjie. 2021b. Construction knowledge graph for the porphyry copper deposit in the Qingzhou-Hangzhou Bay area: Insight into knowledge graph based mineral resource prediction and evaluation. Earth Science Frontiers (China University of Geosciences (Beijing); Peking University), 28(3): 67~75.

Zhu Yueqin, Zhou Wenwen, Xu Yang, Liu Ji, Tan Yongjie. 2017. Intelligent learning for knowledge graph towards geological data[J][OL]. Scientific Programming; <https://doi.org/10.1155/2017/5072427>.

Based on BERT—BiLSTM—CRF model the named entity and relation joint extration of Chinese lithological description corpus

CHEN Zhongliang^{1,2)}, YUAN Feng¹⁾, LI Xiaohui¹⁾, ZHANG Mingming¹⁾

1) School of Resources and Environment Engineering, Hefei University of Technology, Hefei, 230009;

2) Geological Survey of Anhui Province, Hefei, 230001

Abstract: At present, the geological survey is developing from digitization towards the direction of intelligence. According to the big data thinking, the machine reading technique and the auto-extraction of geological knowledge based on the unstructured data deserves academic concern in geosciences. The problem about joint extraction of the geological named entity and relation is the key to this research and yet it is lack of study. This paper proposes the BERT—BiLSTM—CRF model based on the pre-trained Chinese language representation model which was called BERT to conduct the joint task of geological named entity recognition (NER) and relation extraction (RE) on the lithological description corpus. First, the sentence-level corpus was collected from the the profiling and field geological observation data which were produced by the digital geological survey information system designed by China Geological Survey (CGS). Second, based on the theory of petrology, the meta-graph was projected for the rock named entities and relations and the corpus was manual labeled. Third, the comparison experiment of geological knowledge extration task were carried out on the labeled corpus. The experiment results showed that the BERT model does apply to the NER and RE task on the lithological description corpus. The performance (*F1*) achieved by the proposed BERT—BiLSTM—CRF model on the lithological named entity and relation joint extraction task reached 91.75%, and *F1* even reached 97.38% on the task of the named entity recognition. The ablation experiments indicated that the influence of the BERT-embedding layer is prominent on the lithological knowledge extration task and the BiLSTM layer can improvement the performance of the entity and relation joint extraction task.

Keywords: big data thinking; deep learning; pretrained Chinese language model; named entity recognition; relation extraction

Acknowledgements: This study was financially supported by the National Natural Science Foundation of China (Nos. 41820104007, 42072321, 41872247).

First author: CHEN Zhongliang, male, born in 1984, senior engineer, doctor candidate, mainly engaged in the regional geological survey and information technology research on geosciences; Email: c_mulder@163.com

Corresponding author: YUAN Feng, male, born in 1971, professor, doctoral tutor, mainly engaged in mineralogy, petrology and ore deposit; Email: yf_hfut@163.com

Manuscript received on: 2021-07-07; Accepted on: 2022-01-05; Network published on: 2022-01-20

Doi: 10.16509/j.georeview.2022.01.115 **Edited by:** ZHANG Yuxu