https://doi.org/10.3799/dqkx.2020.309

基于ELMO-CNN-BiLSTM-CRF模型的地质实体识别

储德平[1],万 波[1,2*],李 红[1],方 芳[1,2],王 润[1,2]

1. 中国地质大学地理与信息工程学院,湖北武汉 430078
2. 国家地理信息系统工程技术研究中心,湖北武汉 430078

摘 要: 地质实体是地质文本中的关键和核心信息,对其准确识别是地质信息提取和挖掘的重要前提.设计了ELMO-CNN-BiLSTM-CRF模型,基于预训练字向量构建深层BiLSTM-CRF神经网络模型,通过添加词语动态特征以及词语字符级别的特征,弥补字向量特异性缺失的问题,提高对于地质文本中复杂多词义的识别水平和对地质实体局部特征的提取能力.以《西藏自治区谢通门县雄村铜矿勘探地质报告》为例,对该模型的性能进行了评估,模型的准确率、召回率和$F1$值分别为95.15%、95.26%和95.21%.实验表明相比BiLSTM-CRF和CNN-BiLSTM-CRF模型,该模型在小规模语料地质实体识别方面效果更优,且能够有效识别长地质实体词汇和地质多义词.

关键词: 地质大数据;地质实体;命名实体识别;ELMO-CNN-BiLSTM-CRF;地质文本;数学地质.

中图分类号: P628.4 **文章编号:** 1000-2383(2021)08-3039-10 **收稿日期:** 2020-09-17

Geological Entity Recognition Based on ELMO-CNN-BiLSTM-CRF Model

Chu Deping[1], Wan Bo[1,2*], Li Hong[1], Fang Fang[1,2], Wang Run[1,2]

1. *School of Geography and Information Engineering, China University of Geosciences, Wuhan* 430078, *China*
2. *National Engineering Research Center of Geographic Information System, Wuhan* 430078, *China*

Abstract: Geological entity is the key and core information in geological texts, and its accurate recognition is an important prerequisite for geological information extraction and mining. The ELMO-CNN-BiLSTM-CRF model is designed in this paper. Based on the pre-trained word vector, the deep BiLSTM-CRF neural network model is constructed. By adding dynamic features of words and character-level features of words, it makes up for the lack of specificity of word vectors, improves the recognition level of complex multi-word meanings in geological text and the ability to extract local features of geological entities. Taking the geological survey report of Xiongcun copper mine in Xietongmen County of Xizang Autonomous Region as an example, the performance of the model is evaluated. The accuracy rate, recall rate and $F1$ value of the model are 95.15%, 95.26% and 95.21% respectively. Experiments show that compared with BiLSTM-CRF and CNN-BiLSTM-CRF models, this model is more effective in small-scale corpus geological entity recognition, and can effectively identify long geological entity words and geological polysemants.

Key words: geological big data; geological entity; named entity recognition; ELMO-CNN-BiLSTM-CRF; geological text; mathematical geology.

基金项目: 国家重点研发计划项目(No.2016YFB0502300);中国地质调查局项目(No.12120114074001).

作者简介: 储德平(1997-),男,硕士,研究方向为地质大数据挖掘. ORCID: 0000-0003-3577-4973. E-mail: Chudeping_2019@cug.edu.cn

* **通讯作者:** 万波,ORCID: 0000-0003-2387-5419. E-mail: wanbo@cug.edu.cn

引用格式: 储德平,万波,李红,等,2021.基于ELMO-CNN-BiLSTM-CRF模型的地质实体识别.地球科学,46(8):3039-3048.

随着地质信息技术的发展,海量地质信息得到归纳整合,以科学实验和知识归纳为主要研究手段的地质科学领域,大数据成为新的科学范式(Tolle *et al.*, 2011; Baumann *et al.*, 2016). 在大数据背景下,用数据科学方法对地质学中的大数据进行智能处理,可以从中分析和挖掘有价值的核心信息和关键数据(张广宇等,2020),用于解决地质学和地质工作中的认知、预测、决策、评价等理论和实际问题(赵鹏大,2015). 就地质大数据而言,数字知识包括野外地质观测点数据、空间地图数据、文本数据等诸多数据(谭永杰等,2018;杨宇谦,2018). 对于这些海量的地学数据,通过数据库进行整合、组织、管理并集成数据分析、数据挖掘等工具,具有重要意义(张鸣之等,2013).

自20世纪80年代起,我国在地质调查领域逐步建立了一系列由海量的结构化以及非结构化的数据构成的基础地质数据库,其中多样化、碎片化的非结构化数据在地质成果中相比结构化数据占据更高的比例(李超岭等,2015). 由Word、PDF、图片、图表等非结构化数据构成的海量地质报告蕴含着丰富的地质信息. 面对这些"大数据"、"大资源",如何从中挖掘信息获取知识并建立图谱成为当下研究热点(朱月琴等,2015;蒋秉川等,2018;Wang *et al.*,2018;Fan *et al.*,2019),其中地质实体识别是地质信息提取和挖掘的重要前提.

命名实体识别技术是自然语言处理领域的重要工作之一. 近年来随着神经网络的发展,基于深度学习的命名实体识别技术在医疗、生物、新闻等领域取得了很好的成果(李丽双和郭元凯,2018;刘宇鹏和栗冬冬,2020). 传统的机器学习方法,如条件随机场模型(conditional random fields,CRF)、隐式马尔可夫模型(hidden markov model,HMM)等方法需要添加大量人工特征,特别是对于全新的领域,相比之下基于深度学习的命名实体方法不需要添加任何特征,具有更广泛的通用性(陈曙东和欧阳小叶,2020). 在地学领域,地质实体识别是从地质文本中提取地质信息以支持数据分析和地质解释的关键任务(Qiu *et al.*, 2019). 地质实体识别是识别和提取重要的地质概念的过程,如岩石、地质构造和地质时代等,这些概念统称为地质实体,是其他属性和关系描述的基础. 地质实体识别不仅可以识别提取出地质文本中的关键和核心信息,帮助快速理解文本内容,还可以为地质信息组织管理、关系提取、图谱构建提供数据支持(张雪英等,2020). 然而在地质信息表达过程中,依赖结构化数据的地质命名实体识别无法直接访问锁定在地质文本中的实体信息,这使得地质实体识别成为一项具有挑战性的任务. 相比于一般领域的命名实体识别,地质文本报告内容冗余,文字量巨大,文本特征、模式方面高度多样,常见不同的地质实体表达方式,例如表1中1、2句分别以直述和解释两种方式表达实体;且地质实体词汇组成结构复杂,如表1中"夹薄层泥灰岩和钙质角岩"、"含砾泥质长石石英粉砂岩"等复杂的长地质实体词汇;并且存在大量复杂地质多义词,表1中3、4句两个"断裂"根据句意可以推断出含义不同,第一个断裂属于实体名词,第二个则是区域性描述.

针对地质实体识别研究,张雪英等(2018)提出了基于深度信念网络的地质实体识别,将地质实体组成关系划分了对象、特征和关系三个层次;马凯(2018)针对铜矿床的地质报告文本对铜矿床的概念特征进行了详细的划分,采用BiLSTM-CRF方法开展了与铜矿床相关地质实体的识别任务;Qiu *et al.*(2019)提出了一种基于注意力机制的双向长短时记忆条件随机场层的命名实体识别方法,提取地质实体信息. 虽然上述研究均采用的是深度学习的方式,避免了使用领域知识和人工添加特征工程,但是均基于单一词向量或者字向量开展实验;单纯的词向量或者字向量对于文本所蕴涵的信息并不能充分地表达,存在一定的不足. 本研究利用了卷积神经网络(convolutional neural network,CNN)可以很好的提取局部特征以及基于语言模型

表1 部分地质文本内容

Table 1 Part of geological text content

序号	例句
1	麻木下组主要岩性为灰—灰白色中—厚层状结晶灰岩,含燧石结核大理岩,夹薄层泥灰岩和钙质角岩…
2	拉嘎组(C_2P_1l):该组以含砾碎屑岩为特征. 下部为灰白、灰黄色中层细粒岩屑石英砂岩、长石石英砂岩、含砾泥质长石石英粉砂岩…
3	F1、F2断裂都具有早期韧性剪切、晚期脆性变形的特征,而且F1断裂韧性剪切特征更为明显…
4	如钙碱性岩多见于褶皱区,碱性岩多见于断裂区等…

的词嵌入(Embedding from language models,ELMO)可以很好的提取词语动态特征的特点,设计了一种融合ELMO、CNN、双向长短时记忆网络(Bidirection short and long term memory networks,BiLSTM)和CRF等方法的综合深度学习模型(ELMO-CNN-BiLSTM-CRF)用于提取地质文本中的地质实体.本研究首次提出以多特征融合的深度学习方式提取地质实体,以预训练的外部字向量为基础,该模型首先利用ELMO和CNN提取词语动态特征和词语基于字的特征,并与预训练字向量拼接,得到具有更丰富语义信息的新向量,然后将拼接完成的新向量输入到BiLSTM进行训练;最后应用线性条件随机场模型进行解码,对输出标注间的关系进行约束;从而得到了一个同时利用词语基于字的特征、词语动态特征和字向量的地质实体识别模型.在矿产地质文本报告构成的数据集上验证了该模型的有效性.

1　基于ELMO-CNN-BiLSTM-CRF模型的地质实体识别

深度学习方法作为一种具有多个隐含层的机器学习算法,可以通过更深层次的网络模型来学习和提取更高维度的样本特征(左仁广等,2020),被广泛地应用于文本分类、情感分析、语义分析、命名实体识别等各项研究中,在文本数据挖掘中具有显著优势(Collobert *et al.*, 2011; Ma and Hovy, 2016).本文采用的深度学习模型(ELMO-CNN-BiLSTM-CRF模型)进行地质实体识别的流程主要分为文本字符向量化、特征提取和网络参数训练3个部分,如图1所示.

1.1　文本字符向量化表达

为利用计算机进行自动提取和学习文本中所蕴涵的特征及信息,首先需要将文本转化成计算机能够识别的语言,这个过程就是文本向量化(Turian *et al.*, 2010).将经过预处理的文本序列中每个字符转化成向量形式,向量的维度代表字符所蕴含的语义信息的丰富程度.以向量的形式作为模型的输入,不仅避免了传统机器学习方法的诟病,而且能够很好地表达文本中的语义信息和语法关系(Wang *et al.*, 2020).本研究采用预训练的外部字向量,来完成文本向量化的工作.

1.2　特征提取

在中文命名实体识别过程中,由于中文词语之间不像英文单词之间有空格隔开,因此在正式训练之前需要进行分词,但是直接分词可能会导致很多分词词库中不存在的词语无法被正确切分,从而导致最终识别结果变差.常见的做法是用字向量代替词向量进行训练,避免分词产生的误差.而同一个字在不同词语中含义是存在差异的,直接使用字向量会导致词语的特异性缺失.例如"自"在"自然"一词中指大自然环境,在"自…起"中表示开始的含义.如果单纯使用字向量,这两种情况下"自"所表示的字向量是一样的.为了弥补字向量的缺陷,获得更多语义信息,就需要挖掘更多的特征来丰富字向量.特征提取的过程就是获取更多的语义信息、丰富字向量信息的过程,是在字符向量化的基础上,通过神经网络挖掘文本中隐含的语义特征,包括词语中字的特征以及句子中词语的特征等,并将挖掘出来的特征拼接到预训练字向量上,来获取具有更加丰富信息的训练向量的过程.特征提取主要由CNN提取词语字符级别特征和ELMO提取词语动态特征两部分组成.

1.2.1　CNN提取词语字符级别特征　以往的研究表明,CNN可以很好地提取数据的局部特征(Kim, 2014).故引入CNN来提取实体词语字符级别的特征,如提取出以"…体"、"…带"、"…块"等结尾的实体中最后一个字的特征.如图1中所示,对于所有的字向量构成的矩阵均通过占位符(padding)以最长的词语长度为标准补充到同一大小,然后通过卷积从每个词语字向量构成的矩阵中提取当前词语字符级别的特征,再通过池化进一步提取特征中的关键信息,最后将提取的特征拼接到当前词语对应的每个字的字向量上;既可以保留字向量的优点,也可以充分利用当前字对应词语的语义信息.

1.2.2　ELMO提取词语动态特征　中文文本中存在大量的多义词,在不同的语句中同一个词语存在不同的含义.神经网络在学习相关词语的特征时就会产生影响.同样在地质文本中也存在类似的词语.如"品位"在地质文本中用于描述矿石中有用矿物的含量,也可以描述官吏品级或者人和事物的品质水平;"断裂"既可以指名词裂缝、裂隙,属于地质实体,也可以指动词裂开、分裂,不属于地质实体.诸如Word2vec或者GloVe等工具获取的词向量对于这类多义词汇只能表示成同一个词向量,针对这种情况本文引入了ELMO模型,用于学习不同文本序列中同义词的不同含义,提取词语的动态特征,

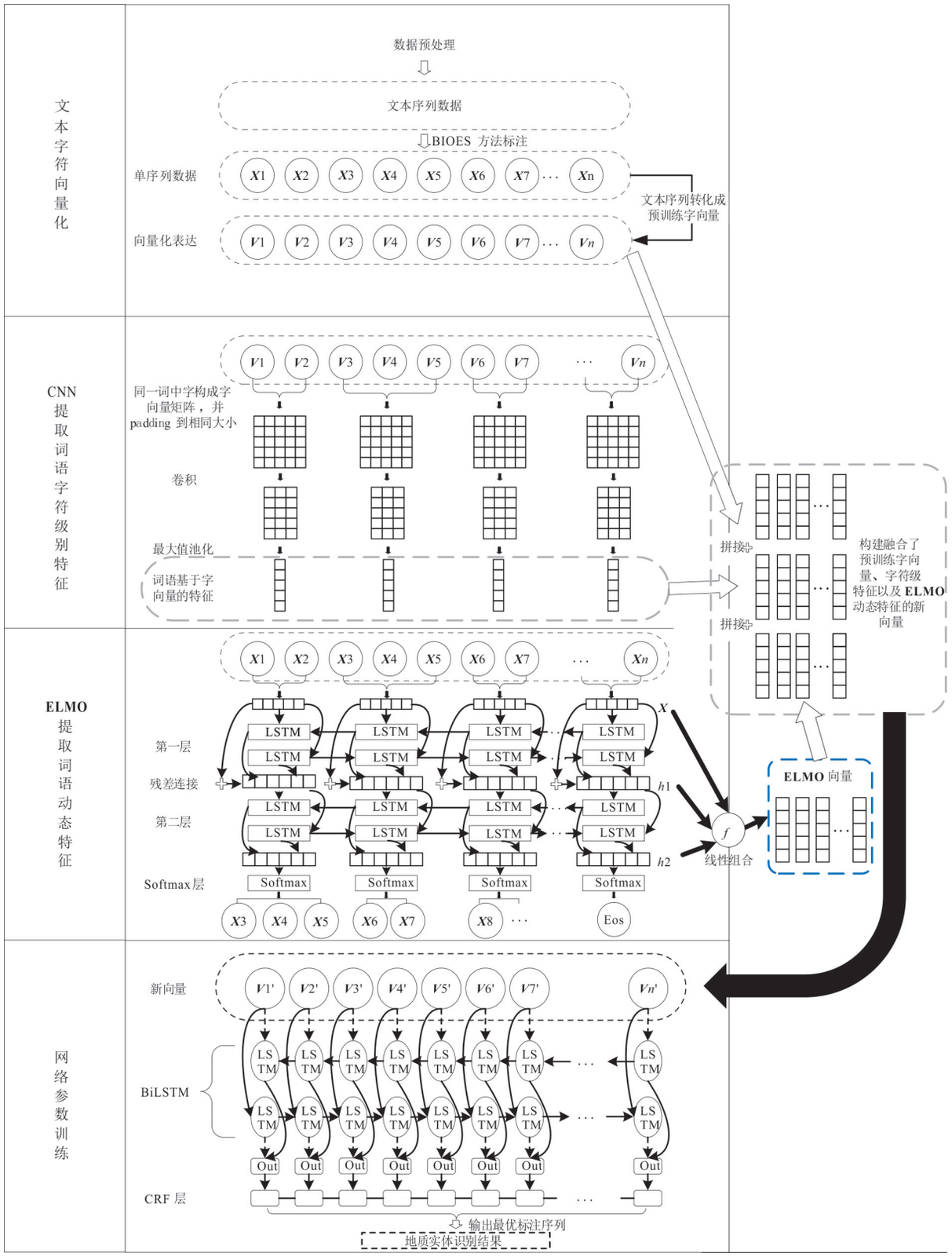

图 1 基于 ELMO-CNN-BiLSTM-CRF 命名实体识别流程

Fig.1 Named entity recognition process based on ELMO-CNN-BiLSTM-CRF

以提高模型的性能.

ELMO 模型最早于 2018 年首次被提出,该模型表示不同于传统的单词类型嵌入,因为每个标记都被分配了一个表示,它是整个输入句子的函数. EL-

MO使用来自双向LSTM的向量,两个LSTM之间通过残差连接,保证网络稳定,顶层softmax根据双向LSTM的输出结果计算上下文的条件概率(赵亚欧等,2020). 该双向LSTM通过耦合语言模型(Language model,LM)目标在大型文本语料库上进行训练,获取模型参数;然后将后续的文本序列输入训练好的双向语言模型中,抽取双向语言模型的输入层和隐含层(如图1中输入层向量为$\boldsymbol{X}$,第一层隐状态为$h1$,第二层隐状态为$h2$)进行加权组合,即可获取文本序列的$\boldsymbol{ELMO}$特征向量. 具体加权组合如下:

$$\boldsymbol{ELMO}_k=\gamma(s_0\boldsymbol{X}_k+\sum_{j=1}^{L}s_jh_k^j), \tag{1}$$

其中,γ为缩放因子;s_j为归一化的系数,表示每个特征的占比;$\boldsymbol{X}_k$表示t_k时刻的输入词向量;s_0表示输入向量对应的特征权重;h_k^j表示t_k时刻第j个隐藏层对应的隐状态,由前向和后向模型的隐状态$\overrightarrow{h_k^j}$和$\overleftarrow{h_k^j}$拼接而成.

1.3　网络参数训练

网络参数训练是利用神经网络学习文本向量中所蕴涵的信息的关键步骤,通过将拼接的词语基于字符级别特征以及词语动态特征的预训练字向量,作为神经网络的输入. 神经网络通过对输入数据进行学习,来迭代更新参数,最终得到训练好的模型. 本文选用的参数训练模型主要结构是双向长短时记忆网络(Hochreiter and Schmidhuber,1997)和条件随机场模型(Lafferty *et al.*,2001).

1.3.1　双向长短时记忆网络　长短时记忆网络(short and long term memory networks, LSTM)由Hochreiter and Schmidhuber(1997)提出,解决了早期循环神经网络经过多层网络传播之后出现的梯度消失或梯度爆炸的现象,是一种特殊的循环神经网络. 由于其能够很好地捕捉时序信息,且能很好地处理具有前后依赖性的信息,被广泛应用于文本信息处理的任务中(Chiu and Nichols,2016).

标准的LSTM只能接受前文信息,只考虑前文信息对当前时刻的影响,忽略了下文信息. 考虑到中文文本上下文联系紧密,本文采用双向LSTM. BiLSTM是LSTM的进一步发展,通过增加了一个后向的LSTM,将前向隐藏层和后向隐藏层结合起来,在递归运算中得到当前时刻输入信息的两种不同向量表示,拼接在一起,作为当前时刻输入信息的向量表示,这样既可以访问前文信息,也可以访问后文信息. BiLSTM详细编码模式见图2.

1.3.2　条件随机场模型　命名实体识别任务一般可以被认为是序列标注的问题,通常BiLSTM的输出结果即可进行序列标注,通过在最顶层添加一个softmax层进行判断,输出概率最大的标签,即可完成输入序列的标注任务. BiLSTM虽然解决了上下文联系的问题,却缺乏对输出标签信息的约束. softmax层的判断是基于当前时刻的信息的判断,没有考虑到整体的联系,输出的结果只是当前时刻信息的最优解,即局部最优解. 相应地就可能会导致输出无效的标签序列. 所谓无效标签序列就是出现诸如{B-GEO,I-TIME,…}这样的输出序列. 在命名实体识别任务中,采用BIOES方法(Lample *et al.*, 2016)进行标注,就不可能出现B-GEO之后为I-TIME的情况.

因此本文引入条件随机场模型,CRF通过对输出整个标签序列进行联合解码,从整体上寻找最优序列,采用全局归一化的方法,避免了标签偏置的问题,解码过程采用的是维特比算法(Viterbi)(Lafferty *et al.*,2001).

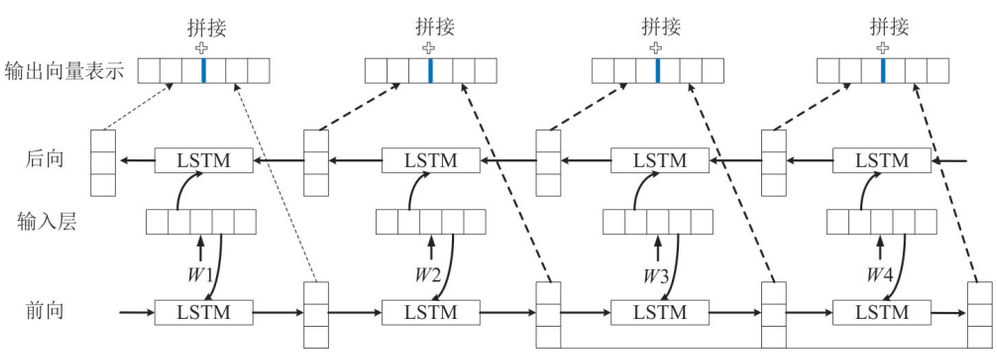

图2　双向长短时记忆网络编码模式

Fig.2　Bidirectional long-time memory network coding mode

2 实验参数及评价标准

2.1 模型参数设置

本文实验全程在PC机环境下完成,64位Windows10操作系统,采用的计算机语言是python3.7.模型的参数采用梯度下降进行优化,详细参数见表2.

参数设置参照Ma and Hovy(2016),CNN层中采用的滤波器窗口大小为3,滤波器个数为30,LSTM状态大小为200,即前、后向的隐状态维度均为100维.ELMO特征向量维度通过多组实验最终设定为40维,详细见结论部分,训练过程中的学习率设为0.001,模型采用Adam优化器进行优化.为防止出现过拟合的情况,在双向LSTM前后添加了dropout层,dropout值为0.5;模型训练过程中采用梯度截断的方法防止梯度爆炸,梯度截断阈值设为5,当梯度大于5或者小于−5时进行梯度截断.

2.2 实验评价标准

实验采用准确率(Precision)、召回率(Recall)、$F1$值3个评价指标来评价模型精度.具体见公式(2)~(4).在本实验中,式中n_correct指原标注是地质实体并被识别为地质实体的词的个数,n_predict指所有被识别为地质实体的词的个数,n_label指所有被标注为地质实体词的实际个数.

$$\text{Precision}=\frac{\text{n\_correct}}{\text{n\_predict}}, \tag{2}$$

$$\text{Recall}=\frac{\text{n\_correct}}{\text{n\_label}}, \tag{3}$$

$$F1-\text{score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{4}$$

单纯考虑Precision和Recall可能会出现矛盾,例如准确率很大而召回率很低的情况或者召回率很大而准确率很低的情况,因此引入$F1$值综合考虑2个指标的影响.

表2 模型参数

Table 2 Model Parameter

层	超参数	数值
CNN	窗口大小	3
	滤波器个数	30
ELMO	映射维度	40
LSTM	状态大小	200
	初始状态	0.0
	孔洞	无
其他	dropout率	0.5
	批量大小	10
	学习率	0.001
	梯度裁剪	5.0

3 实验及结果分析

本研究实验数据来自中国地质调查局全国地质资料馆网站(http://www.ngac.org.cn/)下载的《西藏自治区谢通门县雄村铜矿勘探地质报告》(共计26万字).对报告进行去除图片、表格等一系列非文本的部分,对文本部分去除空格后按句子进行划分,并去除不含地质实体信息的句子,最后共得到1 313条有效句子.对于这些有效数据,本文按照随机抽取的方式,以7:3的比例划分为训练集和测试集.同时为了加快模型的收敛速度,提高模型精度,本实验采用经过预训练的100维字向量作为输入序列的字向量查询表,对于在输入序列中存在而在预训练字向量表中不存在的字采用随机初始化的方式进行随机生成字向量.进行预训练的字向量地质文本由专家挑选出的非标注的地质报告组成,共计520 231字.为增加实验的可靠性,用于预训练的字向量文本只进行去除空格、图片、表格的操作,最大程度保留文本内容.

实验过程参考张雪英等(2018)和马凯(2018)的地质实体划分标准,将地质实体划分为实体对象类(GEO)、地质年代类(TIME)、地质作用类(PROCESS)和其他地质指标类(OTHERS)四类,不对地质实体的组成关系进行提取,同时简化实体类别(表3).

由于实验是基于预训练字向量进行训练的,本文参考《地质大辞典》收录的地质词汇建立标注语料库,采用以上四种地质实体类别对测试集和训练集的所有数据使用计算机进行自动标注,标注完成后进行人工校对,以保证标注的准确性.标注方式为BIOES方法,其中B标注实体词汇的第一个字,I标注实体词汇的所有中间字,E标注实体词汇的最后一个字,S标注单个字的实体词汇,O标注所有非实体词汇.具体标注样例见表4.

对于ElMO特征向量维度的选择,本文进行了多组实验,实验结果见图3,可以看出当ELMO维度为40维时,模型效果最好.分析原因可能是本文使用的小规模语料库,用于训练ELMO特征向量的数据偏少,导致提取的词语特征信息不够充分.当特征维度偏小时,包含的信息量不足,进而影响模型精度;当维度偏大时,ELMO特征信息在组合向量

表3　地质实体类别划分及相关样例

Table 3　Classification of geological entities and related samples

实体类型	样例
实体对象(GEO)	雅鲁藏布江缝合带、班公湖－怒江缝合带、狮泉河－纳木错断裂带、冈底斯－念青唐古拉地体、花岗岩、白朗蛇绿岩带等
地质年代(TIME)	第四纪、震旦纪、前寒武纪、古生代等
地质作用(PROCESS)	大理岩化、绢云母化、铜矿化等
其他地质指标(OTHERS)	品位、倾角、产状等

表4　BIOES标注样例

Table 4　BIOES labeled sample

字	标注	字	标注
有	O	围	O
时	O	绕	O
包	O	黄	B-GEO
裹	O	铜	I-GEO
少	O	矿	E-GEO
量	O	呈	O
石	B-GEO	环	O
英	E-GEO	带	O
.	O	分	O
有	O	布	O
的	O	.	O

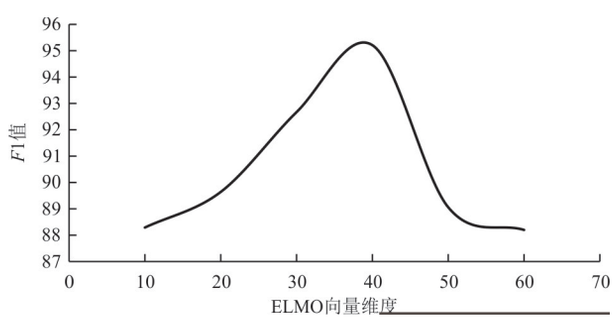

图3　ELMO特征向量维度影响

Fig.3　Influence of ELMO eigenvector dimension

中占比过大,会导致原有的字符向量和CNN特征向量的信息被稀释.因此通过实验,本文最终选取ELMO特征向量维度为40维.

为了进一步比较模型的性能,本文采用以下3种模型在相同的训练集上进行训练,并采用相同的测试集进行测试,3种模型的性能对比见表5.

为保证实验的准确性,本文进行了多组迭代实验,由表5可以看出,基于相同的数据和相同的迭代数,本文提出的模型的准确率、召回率、$F1$值均优于前2个模型,且都在90%以上.由此可知,在小规模语料库上,模型中加入ELMO提取的词语动态特征,丰富字向量语义,可以很好地提升模型的性能.同时可以看出用于比较的模型的主体都是BiLSTM-CRF神经网络,通过在该模型的基础上为输入向量添加相应的特征向量来丰富语义,CNN和ELMO的目的都是为了提取不同情况下的语义信息.通过将两者的提取的语义信息拼接到BiLSTM-CRF模型的输入向量上,就可以让神经网络学习到更多的特征信息.同时模型具有很好的移植性,对于不同类型的地质文本只需要训练出合适的网络模型即可.

表5　不同模型训练结果

Table 5　Training results of different models

模型	Precision	Recall	$F1$
迭代次数为100			
BiLSTM-CRF	86.51%	87.24%	86.87%
CNN-BiLSTM-CRF	92.49%	91.01%	91.74%
ELMO-CNN-BiLSTM-CRF	94.83%	94.39%	94.61%
迭代次数为200			
BiLSTM-CRF	87.36%	88.70%	88.03%
CNN-BiLSTM-CRF	93.24%	90.48%	91.84%
ELMO-CNN-BiLSTM-CRF	94.95%	94.57%	94.76%
迭代次数为500			
BiLSTM-CRF	89.61%	87.76%	88.68%
CNN-BiLSTM-CRF	92.17%	91.13%	91.64%
ELMO-CNN-BiLSTM-CRF	95.15%	95.26%	95.21%

从表6可以看出本文提出的模型对于如“花岗闪长岩岩脉”、“斑状黑云母角闪石花岗岩”等由多个实体名词构成的混合长实体名称能进行很好的识别;同时本文对复杂多义词也进行了很好的处理,如表中“…花岗闪长岩岩脉断裂形成层状地形…”和“…断裂南侧较老地层及局部超基性岩覆于北侧…”中“断裂”一词前者是动词,不属于地质实体,后者是名词,属于实体,模型进行了很好的区分.但是从表中也可以看出本实验存在一些识别问题,总结其中主要问题如下:(1)对于具有相似特征的词汇上的判断存在一些问题,如“班公湖－怒江缝合带”和“中心相带”两个词,前者属于大地构造

表6 ELMO-CNN-BiLSTM-CRF模型部分识别实例

Table 6 ELMO-CNN-BiLSTM-CRF model partial identification instance

原文内容	标注信息	识别结果
…主要对雄村铜矿体进行了较为详细的研究…	铜矿体	铜矿
…花岗闪长岩岩脉断裂形成层状地形…	花岗闪长岩岩脉	花岗闪长岩岩脉
…断裂南侧较老地层及局部超基性岩覆于北侧…	断裂、地层、超基性岩	断裂、地层、超基性岩
…常具有强烈的绢云母化及泥化…	绢云母化、泥化	绢云母化、泥化
…中心相带以石英二长岩为主,边缘相带为花岗岩、斑状黑云母角闪石花岗岩…	石英二长岩、花岗岩、斑状黑云母角闪石花岗岩	中心相带、石英二长岩、花岗岩、斑状黑云母角闪石花岗岩
…磨棱岩化花岗岩基本保留原岩特征…	磨棱岩化花岗岩	岩化、花岗岩

单元,是地质实体,后者是描述方位词,非地质实体;因为两者具有相似的特征,都是“…体”,故模型在识别过程中进行了误分.(2)对于嵌套关系的词有时存在漏字的情况,如“铜矿体”在识别过程中只识别出其中的“铜矿”,漏了“体”;“磨棱岩化花岗岩”由“磨棱岩”、“岩化”、“花岗岩”混合构成,模型在学习过程中只提取了局部信息.

抛开模型训练过程中的偶然性,分析问题出现的主要原因有:(1)测试集中的部分实体信息在训练集中未出现过,标注不够完整,导致模型在训练时缺乏相应的知识,因此在测试时出现错分的情况.(2)模型训练的数据量偏少,个别特征出现次数过少,模型对于这类特征的学习存在不足.对于以上出现的问题,应考虑通过增加数据集以及丰富地质实体语料库进行解决,例如搜集各种类型的地质文本报告,获取更加全面的地质实体标注语料,有利于神经网络模型的完善.

4 结论

(1)本文针对地质命名实体问题,提出了通过多特征融合的深度学习方法来识别地质文本实体.利用ELMO来提取动态词向量,获取词语的语义特征,以及利用CNN提取词语字符级别的特征来丰富字向量所包含的信息,进而构建出了ELMO-CNN-BiLSTM-CRF神经网络模型.该模型弥补了单纯使用词向量或者字向量的不足,通过加入特征提高了对地质实体中复杂多义词的区分水平以及对地质实体局部特征的提取能力.

(2)实验表明,通过本文提出的模型可以不用添加任何人工特征,仅通过少量有标注的语料就可以学习到文本所包含的丰富的特征信息,在小规模地质语料上取得了相对于BiLSTM-CRF和CNN-BiLSTM-CRF模型更好的性能.

(3)在后续研究中,将进一步增加数据集以及丰富地质实体语料库,同时将构建地质实体信息关系网络,将地质实体之间的所包含的关系用网络的形式进行表达,如矿物之间的伴生关系、岩石类别的包含关系,通过网络节点进行连接,为进一步充分利用地质信息提供支持.

References

Baumann, P., Mazzetti, P., Ungar, J., et al., 2016. Big Data Analytics for Earth Sciences: The Earth Server Approach. *International Journal of Digital Earth*, 9(1): 3−29. https://doi.org/10.1080/17538947.2014.1003106

Chen, S.D., Ouyang, X.Y., 2020. Overview of Named Entity Recognition Technology. *Radio Communications Technology*, 46(3): 251−260 (in Chinese with English abstract).

Chiu, J. P. C., Nichols, E., 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4: 357−370. https://doi.org/10.1162/tacl_a_00104

Collobert, R., Weston, J., Bottou, L., et al., 2011. Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12(1): 2493−2537.

Fan, R. Y., Wang, L. Z., Yan, J. N., et al., 2019. Deep Learning-Based Named Entity Recognition and Knowledge Graph Construction for Geological Hazards. *ISPRS International Journal of Geo-Information*, 9(1): 15. https://doi.org/10.3390/ijgi9010015

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735−1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jiang, B.C., Wan, G., Xu, J., et al., 2018. Geographic Knowledge Graph Building Extracted from Multi-Sourced Heterogeneous Data. *Acta Geodaetica et Carto-*

graphica Sinica, 47(8): 1051－1061 (in Chinese with English abstract).

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification. Conference on Empirical Methods in Natural Language Processing (EMNLP). The Association for Computational Linguistics, Doha.

Lafferty, J.D., McCallum, A., Pereira, F., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco.

Lample, G., Ballesteros, M., Subramanian, S., et al., 2016. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. The Association for Computational Linguistics, San Diego. https://doi.org/10.18653/v1/n16-1030

Li, C.L., Li, J.Q., Zhang, H.C., et al., 2015. Big Data Application Architecture and Key Technologies of Intelligent Geological Survey. *Geological Bulletin of China*, 34(7): 1288－1299 (in Chinese with English abstract).

Li, L.S., Guo, Y.K., 2018. Biomedical Named Entity Recognition with CNN-BLSTM-CRF. *Journal of Chinese Information Processing*, 32(1): 116－122 (in Chinese with English abstract).

Liu, Y.P., Li, D.D., 2020. Chinese Named Entity Recognition Method Based on Bi-Directional LSTM-CNN-CRF. *Journal of Harbin University of Science and Technology*, 25(1): 115－120 (in Chinese with English abstract).

Ma, K., 2018. Research on the Key Technologies of Geological Big Data Representation and Association (Dissertation). China University of Geosciences, Wuhan (in Chinese with English abstract).

Ma, X. Z., Hovy, E., 2016. End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). The Association for Computational Linguistics, Berlin. https://doi.org/10.18653/v1/p16-1101

Qiu, Q. J., Xie, Z., Wu, L., et al., 2019a. GNER: A Generative Model for Geological Named Entity Recognition without Labeled Data Using Deep Learning. *Earth and Space Science*, 6(6): 931－946. https://doi. org/10.1029/2019ea000610

Qiu, Q. J., Xie, Z., Wu, L., et al., 2019b. BiLSTM-CRF for Geological Named Entity Recognition from the Geoscience Literature. *Earth Science Informatics*, 12(4): 565－579. https://doi. org/10.1007/s12145-019-00390-3

Tan, Y.J., Qu, H.G., Wen, M., 2018. On Big Data of Geological Survey. *Geomatics World*, 25(2): 7－11 (in Chinese with English abstract).

Tolle, K. M., Tansley, D. S. W., Hey, A. J. G., 2011. The Fourth Paradigm: Data-Intensive Scientific Discovery. *Proceedings of the IEEE*, 99(8): 1334－1337. https://doi.org/10.1109/jproc.2011.2155130

Turian, J.P., Ratinov, L., Bengio, Y., 2010. Word Representations: A Simple and General Method for Semi-Supervised Learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. The Association for Computational Linguistics, Upsala.

Wang, C. B., Ma, X. G., Chen, J. G., et al., 2018. Information Extraction and Knowledge Graph Construction from Geoscience Literature. *Computers & Geosciences*, 112: 112－120. https://doi. org/10.1016/j. cageo.2017.12.007

Wang, J. M., Hu, Y. J., Joseph, K., 2020. NeuroTPR: A Neuro-Net Toponym Recognition Model for Extracting Locations from Social Media Messages. *Transactions in GIS*, 24(3): 719－735. https://doi. org/10.1111/tgis.12627

Yang, Y.Q., 2018. Current Situation, Problems and Countermeasures of Geological Prospecting Units Participate in the "Big Data" Project Construction. *Natural Resource Economics of China*, 31(7): 31－34 (in Chinese with English abstract).

Zhang, G.Y., Fu, J.Y., Ouyang, Z. Z., et al., 2020. The Importance of Space Database Establishment Based on DGSS in Big Data Environment. *Earth Science*, 45(9): 3451－3460 (in Chinese with English abstract).

Zhang, M.Z., Yu, M.L., Wang, Y., et al., 2013. Designing and Building the National Geo-Environment Monitoring Data Warehouse. *Earth Science*, 38(6): 1347－1355 (in Chinese with English abstract).

Zhang, X.Y., Ye, P., Wang, S., et al., 2018. Geological Entity Recognition Method Based on Deep Belief Networks. *Acta Petrologica Sinica*, 34(2): 343－351 (in Chinese with English abstract).

Zhang, X.Y., Zhang, C.J., Wu, M.G., et al., 2020. Spatio-Temporal Features Based Geographical Knowledge Graph Construction. *Scientia Sinica Informationis*, 50(7): 1019－1032 (in Chinese with English abstract).

Zhao, P.D., 2015. Digital Mineral Exploration and Quantita-

tive Evaluation in the Big Data Age. *Geological Bulletin of China*, 34(7): 1255－1259 (in Chinese with English abstract).

Zhao, Y.O., Zhang, J.Z., Li, Y.B., et al., 2020.Sentiment Analysis Using Embedding from Language Model and Multi-Scale Convolutional Neural Network. *Journal of Computer Application*, 40(3): 651－657 (in Chinese with English abstract).

Zhu, Y.Q., Tan, Y.J., Zhang, J.T., et al., 2015. A Framework of Hadoop Based Geology Big Data Fusion and Mining Technologies. *Acta Geodaetica et Cartographica Sinica*, 44(S1): 152－159 (in Chinese with English abstract).

Zuo, R.G., Peng, Y., Li, T., et al., 2020. Challenges of Geological Prospecting Big Data Mining and Integration Using Deep Learning Algorithms. *Earth Science*, 46(1): 350－358 (in Chinese with English abstract).

附中文参考文献

陈曙东, 欧阳小叶, 2020. 命名实体识别技术综述. 无线电通信技术, 46(3): 251－260.

蒋秉川, 万刚, 许剑, 等, 2018. 多源异构数据的大规模地理知识图谱构建. 测绘学报, 47(8): 1051－1061.

李超岭, 李健强, 张宏春, 等, 2015. 智能地质调查大数据应用体系架构与关键技术. 地质通报, 34(7): 1288－1299.

李丽双, 郭元凯, 2018. 基于CNN-BLSTM-CRF模型的生物医学命名实体识别. 中文信息学报, 32(1): 116－122.

刘宇鹏, 栗冬冬, 2020. 基于BLSTM-CNN-CRF的中文命名实体识别方法. 哈尔滨理工大学学报, 25(1): 115－120.

马凯, 2018. 地质大数据表示与关联关键技术研究(博士学位论文). 武汉: 中国地质大学.

谭永杰, 屈红刚, 文敏, 2018. 论地质调查工作大数据. 地理信息世界, 25(2): 7－11.

杨宇谦, 2018. 地勘单位参与"大数据"项目建设的现状、问题及对策. 中国国土资源经济, 31(7): 31－34.

张广宇, 付俊彧, 欧阳兆灼, 等, 2020. 大数据时代下基于DGSS系统下空间数据库建立的重要性. 地球科学, 45(9): 3451－3460.

张鸣之, 喻孟良, 王勇, 等, 2013. 国家级地质环境数据仓库的设计与实现. 地球科学, 38(6): 1347－1355.

张雪英, 叶鹏, 王曙, 等, 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.

张雪英, 张春菊, 吴明光, 等, 2020. 顾及时空特征的地理知识图谱构建方法. 中国科学: 信息科学, 50(7): 1019－1032.

赵鹏大, 2015. 大数据时代数字找矿与定量评价. 地质通报, 34(7): 1255－1259.

赵亚欧, 张家重, 李贻斌, 等, 2020. 融合基于语言模型的词嵌入和多尺度卷积神经网络的情感分析. 计算机应用, 40(3): 651－657.

朱月琴, 谭永杰, 张建通, 等, 2015. 基于Hadoop的地质大数据融合与挖掘技术框架. 测绘学报, 44(S1): 152－159.

左仁广, 彭勇, 李童, 等, 2020. 基于深度学习的地质找矿大数据挖掘与集成的挑战. 地球科学, 46(1): 350－358.