文章编号：

**基于循环自学习BERT\_BiLSTM\_CRF模型的中文地质报告分词方法**

**摘要：**中文分词是地质报告文本挖掘的基础工作，对于后续的命名实体识别等任务有重要影响。当前通用领域分词模型的精度与语料库的标注规模和质量密切正相关，需要标注大量的语料库才能满足深度学习的训练，而传统的语料库需要耗时费力的人工进行标注。针对这一问题，下文提出一种基于循环自学习的升读学习序列标注模型来对语料库进行自动标注，该模型首先利用BERT生成蕴含有词级特征、语法结构特征的词向量，然后将其输入到BiLSTM-CRF模型中进行训练。实验结果表明，在地质领域测试集上，该模型F1值为95.9%。与已有的BiLSTM-CRF及BiLSTM-CNN-CRF基准模型相比，性能有着较大提升，且提出的循环自学习策略具有很好的泛化能力。

**关键词：**地质报告分词；循环自学习；BERT语言模型；条件随机场CRF

**中图分类号：**TP391  **文献标识码：**A

#### Chinese geological report segmentation based on cyclic self-learning BERT\_BiLSTM\_CRF model

**Abstract:** Chinese word segmentation (CWS) is the foundational work of geological report text mining and has important influence on the following tasks such as named entity recognition. At present, the accuracy of the non-professional field word segmentation model is closely related to the marking scale and quality of corpus, so a large corpus should be marked to meet the deep learning training, while the traditional corpus requires time-consuming and laboring manual marking. In order to solve this problem, a deep learning model based on cyclic self-learning is proposed for automatic labeling the words in corpus. In this model, BERT is first used to generate word vectors with word-level features and grammatical structure features. And then input them into the BiLSTM-CRF model for training. The experimental results show that compared with the existing BiLSTM-CRF and BiLSTM-CNN-CRF models, this model has better performance in domain segmentation, and the proposed cyclic self-learning strategy can be further extended to other domains.

**Key words:** Geological report segmentation; Cyclic self-learning; BERT Bi-LSTM; CRF

**0 引言**

大数据时代背景下，全国地质资料馆中已经积累了海量的地学数据，面对这些海量的地质数据资源，特别是非结构化地质数据，需要进一步建立数据思维、定量思维及努力获取“数据资源”和形成核心“数据知识”的新思维方式。由于地质数据的混合性、变异性、稳健性、相关性等随时间、空间及地质体的不同而各异，如何充分挖掘利用蕴含在地质资料中丰富的知识信息成为研究地质大数据的关键问题[1,2]。为了更好地理解与挖掘地质资料文本中蕴含的地理、地质语义知识，迫切需要一套强有效的地质资料分词方法，为后续的词性标注、命名实体识别、地质语义理解等奠定基础[3,4]。近年来随着学者们对深度学习（Deep Learning）和自然语言处理（Natural Language Processing, NLP）相关研究的深入，通过神经网络进行语言模型建立、文本特征学习和短文本分析等都有广泛应用[5,6]。

在面对地质报告文本分词时，受限于地质报告资料中存在大量关于地名、空间方位、地貌、地层分布、岩性、构造、不良地质、地史等领域专业词汇，传统中文分词方法及通用领域深度学习模型应用到地学领域不可避免会导致存在大量歧义问题，特别是未登录词（Out of Vocabulary，简称OOV）召回率也较低，造成性能急剧下降分词准确率不佳等问题。如“**被查**/**干楚鲁**/粗粒/黑云母/花岗岩/侵入，该套/地层/主要/分布/在/**格日**/**吐**/**防火**/**站**/和/敦德哈布/其勒/南山。”这里传统方法会直接将“被查”、“格日”、“防火”分割出来，直接导致了错误的划分。如表1所示。传统的语料库中需要人工标注，会浪费大量的时间及精力，而且需要标注大量的语料库才能满足深度学习的训练。

表1 地质资料中进行中文分词主要难点

|  |  |  |
| --- | --- | --- |
| 类别 | 示例 | |
| Jieba+外部词典的分词结果 | 正确的分词结果 |
| 地名 | 专题/工作区/位于/**冈底斯/山脉/**中段/南缘/、/**雅鲁藏布江/**之/南 | 专题/工作区/位于/**冈底斯山脉/**中段/南缘/、/**雅鲁藏布江/**之/南 |
| 空间方位 | 整体/呈/**近/东西向/**狭长/带状/展布 | 整体/呈/**近东西向/**狭长/带状/展布 |
| 地貌 | **碎屑/喀斯特**/、/黄土/和/**粘土/喀斯特**/、/**热融/喀斯特**/和/火山岩/区/的/**熔岩/喀斯特**/等 | **碎屑喀斯特**/、/黄土/和/**粘土喀斯特**/、/**热融喀斯特**/和/火山岩/区/的/**熔岩喀斯特**/等 |
| 地质年代 | **晚/侏罗世/**—/**早/白垩世/**和/**中/白垩世/**新特提斯洋/演化 | **晚侏罗世/**—**/早白垩世/**和/**中白垩世/**新特提斯洋/演化 |
| 地质构造 | 大/两会/背斜/，/位于/**汉王山/复式/向斜**/南侧 | 大两会背斜/，/位于/**汉王山/复式向斜**/南侧 |

目前主流的中文分词包括有监督方法和无监督方法。有监督学习方法能够很好的处理静态文本信息，但是很难识别地质文本中蕴含的丰富的语义关系[7]。主要的原因包括以下几点：①需要构建大量针对专业领域的语料库，语料库的构建代价很高，而且学习模型训练的时间很长，迁移到其他领域分词效果不理想。②专业领域具备很强的异质性，训练好的模型可移植性差；③海量的地质资料中出现新的词汇，不能够及时被训练好的模型或模式识别。相反无监督学习方法因为具备诸多优势，已经在自然语言处理领域得到了广泛的关注。该方法不需要构建大规模的语料库，成本较低，使用统计方法就能够获取文本实时特征[8]。更重要的一点，这类方法能都获取动态的文本特征，更加适合地质领域分词处理。

针对地质报告中存在大量专业词汇问题，且地质资料文本描述多样性、复杂性及灵活性，顾及标注地质领域语料库需要耗费大量的人力及精力等问题，提出一种领域本体辅助下的联合**循环自学习策略**的专业领域训练方法获取地质领域语料库，结合BERT预训练语言模型及Bi-LSTM神经网络，采用四词位标注，在训练过程中加入预先训练的词嵌入分布式向量方法进行中文地质报告分词，实验表明联合自学习策略的引入，并结合加入预训练模型BERT，有效改善了地质领域中文分词效果，具有一定的领域通用性。

1. **地学领域分词相关研究**

中文分词是自然语言处理领域中非常重要的一项基础性研究。很多国内外的学者都致力于这个领域的研究。通用领域的分词方法可以分为基于词典分词和基于统计的分词方法两种。前者主要是根据给定的已构词典进行基于词的规则匹配，如正向最大匹配规则、逆向最大匹配规则[9]以及双向匹配规则[10]等。后者则是通过统计机器学习模型对已标注的中文文本训练得到不同的隐马尔科夫模型（Hidden Markov model，HMM）、条件随机场模型（Conditional Random Field，CRF）[11-14]、深度学习[15-18]等不同的模型，并基于训练得到的模型对未知标签的文本进行分词处理。这些经过预训练得到的统计学模型大多是基于字标注的分词方法（Character-based tagging approach）[16-18, 20]，即将中文分词看作基于句子的序列标注问题，从而能够平等地看待词典词和未登录词的识别问题。目前，双向长短期记忆网络（Long Short-Term Memory，LSTM）在中文文本分词中的应用较为广泛，但在学习较长的句子时，可能因为模型容量难以获取一些重要信息，因此通常会在进行模型训练前增加CRF提取当前词的局部特征[21,22]，或者增加预训练语言模型ELMo[23]、GPT[24]、BERT[25]等学习语言表示。BERT能够产生一系列跨层的词嵌入向量，这些词嵌入向量考虑到了单词出现的上下文相关的双向特征，在很多现有的NLP任务中都具有不错的表现[26]，因此本文选择BERT-Bi-LSTM-CRF模型的完成领域中文分词任务。

但该模型的训练需要大量的标注语料库，当直接采用人民日报的语料库和Bakeoff中的语料库训练模型对地质报告进行分词时，通常会出现切分不一致的情况。例如训练后的模型对“上侏罗统满克头鄂博组该套地层在测区内分布广泛”进行分词时，结果为“**上**/**侏罗**/**统满**/**克头鄂博组**/该/套/地层/在/测区/内/分布/广泛”，但实际上“上侏罗统满克头鄂博组”为一个客观存在的地层对象。这主要是由于训练好的模型中领域知识比较匮乏，计算机仅依赖人民日报语料库和Bakeoff语料库无法学习到地层实体的命名规则。领域词典集合规则和有监督的方法目前还无法有效解决领域分词中的OOV及分词歧义问题。为了克服领域熟语料缺乏的问题，文献[27]使用人工标注后的地质矿产语料和人民日报标注语料两种熟语料库共同作为实验语料，解决了部分未登录词的识别问题。文献[28]利用地质词典和地质矿产资源词典中的术语对知网中的地球科学中文文献进行术语匹配和基于字的标注，以获得用于分词模型训练的熟语料库。文献[3,29]将循环自学习策略引入到领域语料库的标注工作中，进一步改变传统的分词模型在专业领域中的适应性。但在使用循环自学习策略迭代收敛模型性能时，均以人工的方式从标注结果中选择较好的结果加入到训练语料中以扩展训练语料，该策略一定程度上能够改进不同分词模型对领域的适应性，但该人工选择过程存在一定的主观因素，地质专业中未登录词、专业词汇、地名、机构等复杂在人工选择的过程中容易造成疏漏，且相比[29]文献中的新闻、文学、电脑三个领域较为耗时。

为了能够充分发挥地质报告中未标注语料库的作用，改善人工标注的耗时性和不完整性，本文提出一种领域本体辅助下基于循环自学习的中文分词方法，以改进分词模型在地质领域的分词性能和效果。

1. **领域本体辅助下基于循环自学习的中文分词模型**

**2. 1 基于领域本体和循环自学习策略的地质语料库生成**

在地质领域中文分词中，由于地质领域术语的多边性和报告领域的不同，带来了训练语料库中很多未出现的词汇或短语，使得未登录词的识别成为地质领域分词中的关键问题。同时，由于地质报告领域改变导致上下文及上下文语义变化也会引起已登录词（In-Vacabulary，IV）处理的准确性及性能。这就迫切需要领域内大规模的标注语料库发挥作用，因此本文采用一种领域本体辅助下的循环自学习训练策略（图1）自动生成领域熟语料，结合字标注分词方法的深度学习模型，改进分词模型在地质领域的分词性能和效果。

图1 领域本体辅助下的循环自学习流程

在通用领域进行中文分词处理时使用的数据集一般是某些特点的语料库，如人民日报语料库和SIGHAN Bakeoff，语料库中涉及到专业领域的数量并不确定，如地质领域中的空间方位、地貌、地层分布、岩性、构造、不良地质、地史、分析、评价等覆盖率较低，这些语料库针对领域词汇的针对性不强，需要增加大量的领域语料库一起进行模型的训练。因此本文通过对《地质大辞典》等领域专业词典的处理获取到地质专业词汇5870个（图2(a)），地质本体中领域词条2175个（图2(b)）。在图1的初始化模型中将这些词汇作为语料库的一部分加入到深度学习模型中，利用训练得到的模型对未标注的地质语料库进行标注，其中未标注的语料库包括43篇地质报告，共计460037个词。

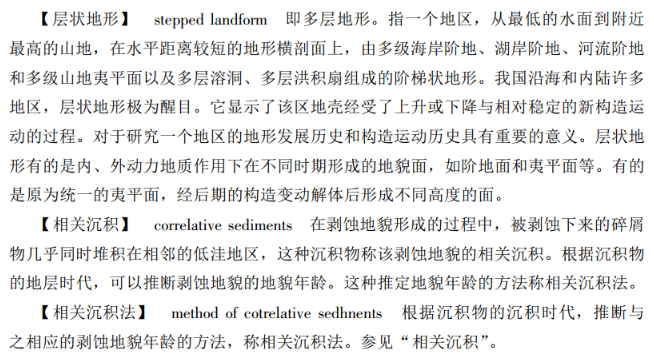


图2 (a) 地质专业词汇

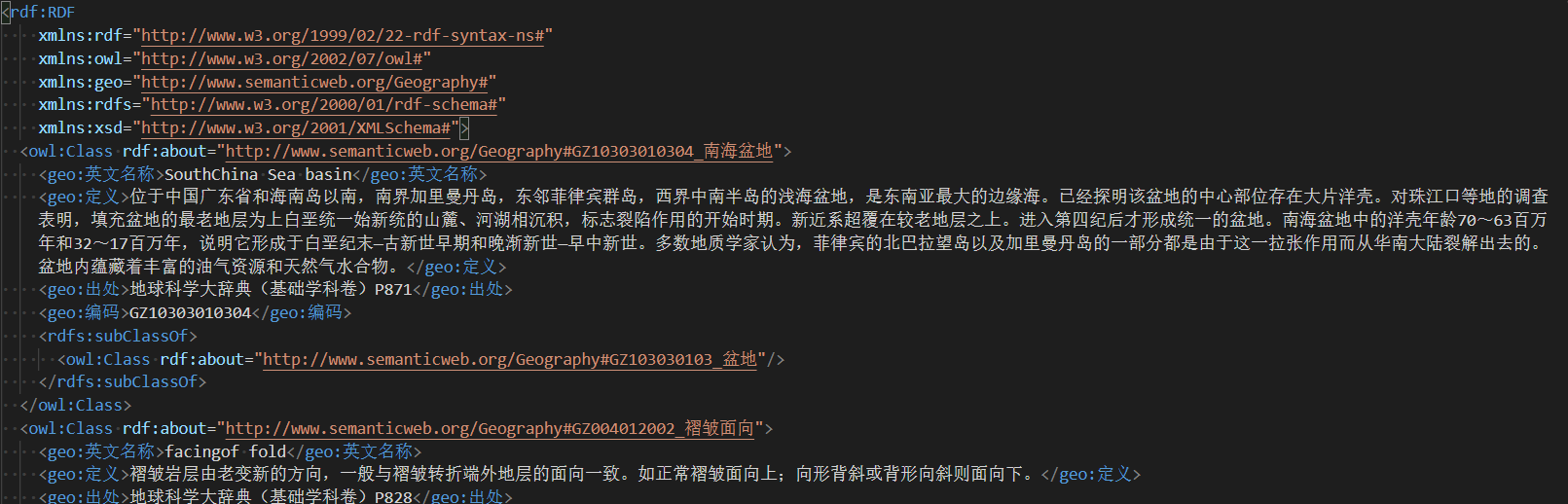


图2 (b) 地质本体

在进行循环自学习的过程，本文将以往的人工“删除”过程，即对语料库中标注错误的进行删除处理，调整为基于本体的领域约束条件判断为“False”和相似度匹配较低的 “删除”过程，即让计算机在现有知识的辅助下自主地从模型中选择较好的结果加入到训练语料库中重新训练，不断执行该过程直至模型收敛。当图1中的第一个判断“本体中的领域词汇的约束条件”结果为Yes时，将模型标注后的领域命名实体返回作为初始模型的输入继续训练模型。当判断结果为No时，计算机将进行第二次判断，计算模型标注得到的领域命名实体与现有实体/词汇之间的相似度，仅当相似度大于给定阈值时认为该标注正确，并将该标注结果返回作为初始模型的输入继续训练模型。模型中的两次判断内容如下：

1. 本体中的领域词汇的约束条件

定义本体的类和等级体系的方法可分为自底向上、自顶向下和混合方法三种。自底向上的构建模式从最底层、最细小的类开始定义，然后逐渐向上将这些细化的小类进行综合形成上一层的类，并不断迭代该过程直至本体构建完全。而自顶向下的模式则与之相反，先确定所构建本体的框架和体系再逐层向下逐层细化直至本体构建完全。前者很容易由于开发者对领域知识认识的不够全面而造成上层本体构建的不完整，后者则可能因为收集到的小类不够细化或不够充足同样造成本体构建的不完整，因为在实际工作中多采用混合方法，即先构建部分显而易见的类，再根据已构建的类分别向上、向下进行泛化和细化。图2(b)中的本体即采用混合方法进行构建，涉及地质领域中的大量领域词汇和领域约束规则，部分如表2所示：

表2 领域命名实体约束条件示例

|  |  |  |
| --- | --- | --- |
| **领域命名实体** | **约束规则** | **示例** |
| **组**（区域地层单位） | *R1：*可以在地理名称之后加上岩石性质，但“组”不能省略 | 峨眉山玄武岩组 |
| **段**（区域地层单位） | *R2：*可以用地理名称，也可以不用地理名称用岩性命名 | 湾湾沟段、  石灰岩段 |
| **细碎屑岩类岩石** | *R3：*胶结物+结构+基本名称 | 泥质粉砂岩 |
| **中碎屑岩类岩石** | *R4：*胶结物+结构+碎屑成分+基本名称 | 中粒长石石英砂岩 |
| **泥质岩类岩石** | *R5：*颜色+混入物+粘土矿物成分+基本名称 | 高岭石粘土（岩） |

2. 领域词汇的相似度

词语的相似度计算即以定量化的方式计算多个词语之间的共性信息，其计算方法可分为四类：基于字符串的相似度计算、基于语料库的相似度计算、基于知识的相似度计算和其他方法，如句法分析、混合方法等。其中基于语料库的相似度计算又可以分为基于词袋（Bag of Words Model，BOW）、基于神经网络的方法和基于搜索引擎的方法[30]。基于神经网络的方法中最典型的就是word2vec模型[31]，通过网络训练可以从非结构化文本中为每一个词训练出与之对对应的低维向量，且每一维的值代表一个具有一定的语义和语法上解释的词语特征。这种表示方式能使相似的词语在同一向量空间中的余弦距离上更加接近，因此本文将采用该模型进行词语相似度的计算，采用的公式见式(1)：

式中，表示词和的相似程度，值越大说明两个词语的相似度越高，反之则相似度越低。的取值范围即为的取值范围[-1,1]。

在本文当的值大于给定阈值θ时，我们认为模型标注的实体与已知实体相似，即模型所标注的实体正确。反之则所标注的实体错误，当标注的实体错误时将会被执行“删除”操作，不会被返回到初始模型作为输入。

根据现有本体中的约束规则能够较好地对提取到的领域实体进行标签正确性的判断，但是具有一定的局限性，这主要是由于本体中的具有规则约束的领域实体有限，如“更新统粘性土”作为一种地质实体并不符合表2中的规则R5。因此本体约束条件判断后，本文还进行了领域词汇相似度的计算，以提高计算机对标签正确性的判断。

**2. 2 基于BERT-BiLSTM-CRF模型的分词方法**

目前主流的中文分词任务都被当作是一种字符级别的序列标注问题，即将具体的分词过程转换成每个字在文本序列中进行标注的过程[15]。常用的是四词位标注{B，E，M，S}来获取词语所在边界信息，其中B、M、E分别代表单词所处的位置的开头、中间及结尾，S代表的是单个字。

近些年来，预训练模型对于中文分词的上游任务来说一直都是研究的热点问题。其中BERT作为目前先进的语言预处理表征模型，能够有效获取高质量的词向量表征模式，更加利于下游的任务。本文在进行中文分词研究当中采用的是BERT-BiLSTM-CRF结构，其整体结构图如图4所示，其基础架构采用的是BiLSM-CRF如图3所示。

BERT-BiLSTM-CRF结构主要包括三个模块。首先将标注之后的语料库输入到BERT预训练语言模块中获取对应的词向量表征，然后将生成的词向量再次输入到BiLSTM网络当中进行后续的处理，最后利用CRF层对BiLSTM的输出进行解码与标签预测，最后通过最后得到的预测标签来进行中文分词的整个流程。

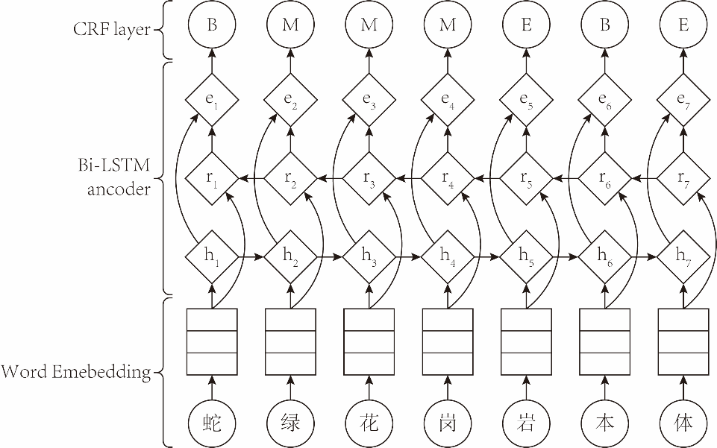


图3 基础的BiLSTM-CRF架构

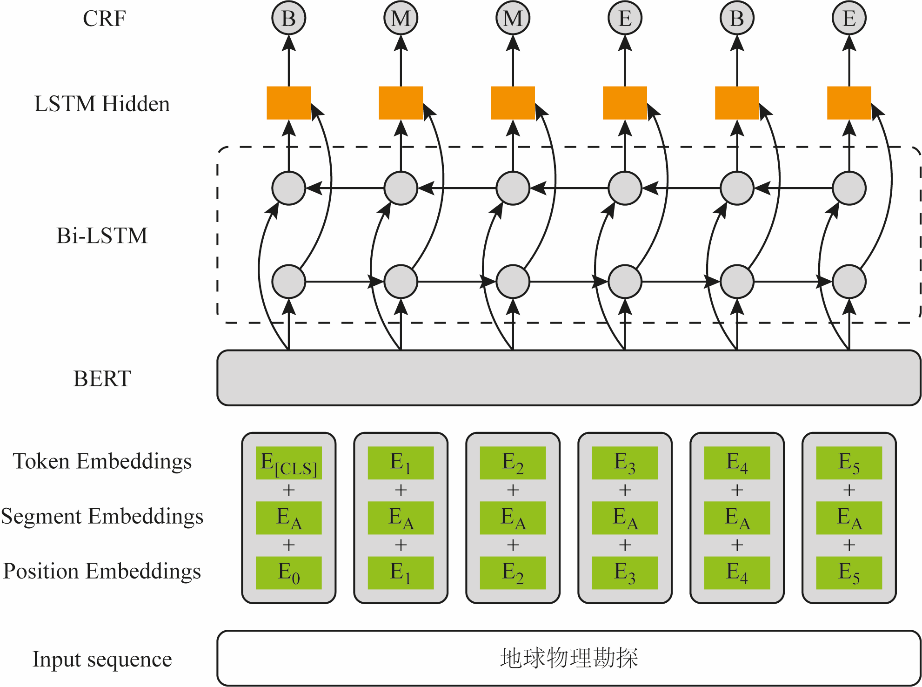


图4 本文中所采用的BERT-BiLSTM-CRF架构

本文所呈现的架构当中最大的优势在于利用了BERT语言预处理模型，这一模型区域别传统的事先训练一系列的字向量与词向量，只需要将序列句子通过BERT语言预处理模型，就能自动化地抽取序列句子中语言丰富的词级特征、语法结构特征及其丰富的语义特征。对于BERT语言预处理模型来说，每个层次上上获取及学习的特征是递进式的，其底层上主要是学习短语这一级别的表征信息，中层这一层次上主要是学习序列当中的句法内部结构表征信息，最后层次上是抓取整个序列中丰富的语义信息[32]。

对于传统的词向量来说主要是获取字符级别及其短语级别的表征信息，对于句法内部结构级别及其句子级别上的表征很少能够捕获，而BERT语言预处理模型能够很好的捕捉语义化的词向量，对于长距离依赖的序列具有较好的效果。

1. **领域本体辅助下基于循环自学习的中文分词模型**

**3. 1 实验环境、数据集和评测指标**

3.1.1 实验基础环境配置

本文所提出的BERT-BiLSTM-CRF架构在进行实验时，采用的环境配置如表3所示：

表3 BERT-BiLSTM-CRF模型训练环境及配置

|  |  |
| --- | --- |
| 操作系统 | Ubuntu 14.04 64bit |
| CPU配置 | 2 \* Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz |
| GPU配置 | 2 \* Nvidia Tesla K20 |
| 内存 | 96GB |
| Python | 2.7 |
| 深度学习框架 | Keras |

3.1.2实验参数配置

本文在实验的过程中选择的是Adam优化器，学习率设置为0.001，其中网络架构当中LSTM维度设置为200，batch\_size设置为64，max\_seq\_len设置为128，实验中为了防止过拟合采用Dropout，设置为0.05。

在进行分词性能评估上，采用中文分词常用的准确率P、召回率R以及综合指标F1=2(P\*R) / (P+R)作为评测指标。

**3. 2 实验设计及结果分析**

通过不断的调整网络的超参数来训练BERT-BiLSTM-CRF网络模型，通过对语料库采取多次迭代的方式得到较好的训练结果。

表4展示了使用循环自学习方法在地质报告中分词实验结果。从结果中可以明显看出，通过循环自学习的方法可以有效地将地质报告中的专业词汇、地名机构、岩石、岩性等分割开，Bakeoff模型由于未加入训练语料库，对于地质专业的词汇“岛弧”、“罕山鹿场”等进行了错误的分割。Bi-LSTM相比较LSTM+CRF模型，分词效果更好，特别是对“毛希嘎达巴”、“阿拉坦达巴道班幅”地名能够很好进行分割。

表4 三种方法在地质领域分词结果对比

|  |  |
| --- | --- |
| **模型** | **实验结果** |
| LSTM-CRF | 形成/规模/较大/的/消减/活动/板块/边缘/（/岛/弧/或/活动/大陆/边缘/）/  该/隆起/带/位于/罕山/鹿场/－/北沙拉/大队/一带/  上/二叠/统/林西/组/组成/林西/组/的/地层/倾向/  沿/该/带/零星/分布/晚/侏罗世/侵入岩/  在/本幅/中/仅/在/北部/包含/一部分/  该/火山口/位于/阿拉坦达/巴道班/幅/的/毛希嘎/达巴/一带/ |
| BiLSTM+CRF | 形成/规模/较大/的/消减/活动/板块/边缘/（/岛弧/或/活动/大陆/边缘/）/  该/隆起/带/位于/罕山鹿场/－/北沙拉大队/一带/  上/二叠统/林西组/组成/林西组/的/地层/倾向/  沿/该/带/零星/分布/晚/侏罗世/侵入岩/  在/本幅/中/仅/在/北部/包含/一部分/  该/火山口/位于/阿拉坦达/巴道班幅/的/毛希/嘎达巴/一带/ |
| BERT-BiLSTM-CRF | 形成/规模/较大/的/消减/活动/板块/边缘/（/岛弧/或/活动/大陆/边缘/）/  该/隆起带/位于/罕山鹿场/－/北沙拉大队/一带/  上/二叠统/林西组/组成/林西组/的/地层/倾向/  沿/该/带/零星/分布/晚/侏罗世/侵入岩/  在/本幅/中/仅/在/北部/包含/一部分/  该/火山口/位于/阿拉坦达巴道班幅/的/毛希嘎达巴/一带/ |

表5展示的是循环自学习语料库在地质报告中的分词效果。从循环自学习的结果来看，随着迭代次数的增加，准确率、召回率和未登录词的召回率都在不断增加，最后趋近于稳定，说明随着语料不断迭代，循环自学习开始发挥作用，迭代到一定次数后，F值发生了波动，说明循环自学习作用开始变小。从结果中看出，循环自学习方法在地质领域分词有了较大幅度提升，F值达到96.7%，特别是在未登录词召回率方面，提升效果更加明显，达到了88.1%。

循环自学习发生作用，是由于深度学习模型能学习专业语料库中的构词规则，同时“删除”操作能够一方面指导网络进行迭代训练，多次网络的迭代训练，深度学习通过自学习功能提高分词性能。

表5领域本体辅助下基于循环自学习的地质报告分词结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **迭代次数** | **P** | **R** | **F** | **IV Recall** | **OOV Recall** |
| 1 | 0.871639 | 0.869937 | 0.8707872 | 0.8846 | 0.7845 |
| 2 | 0.887936 | 0.875641 | 0.8817456 | 0.8941 | 0.7963 |
| 3 | 0.885698 | 0.871546 | 0.878565 | 0.8946 | 0.8021 |
| 4 | 0.89784 | 0.887963 | 0.8928742 | 0.8996 | 0.8325 |
| 5 | 0.901452 | 0.897412 | 0.8994275 | 0.9025 | 0.8411 |
| 6 | 0.914589 | 0.904589 | 0.9095615 | 0.9165 | 0.8463 |
| 7 | 0.934896 | 0.92369 | 0.9292592 | 0.9356 | 0.8512 |
| 8 | 0.948963 | 0.940326 | 0.9446248 | 0.9496 | 0.8601 |
| 9 | 0.953649 | 0.951123 | 0.9523843 | 0.9687 | 0.8732 |
| 10 | **0.963214** | 0.961145 | **0.9621784** | **0.9678** | 0.8812 |
| 20 | 0.96156 | 0.9568 | 0.9591741 | 0.9645 | 0.8803 |

为了验证所使用的BERT-BiLSTM-CRF模型在中文地质分词中较其他以长短期记忆网络及其混合模型具有更好的表现，本文还进行了基于不同网络结构的地质报告分词结果对比实验，如表6所示。双向长短期记忆网络对中文分词的序列标注结果整体都优于长短期记忆网络。BERT预训练语言模型的关键部分是Transformer结构, 能够联合上下文语境进行预训练，学习到短语级别的信息表征以及丰富的语言学特征、语义信息特征。BERT-BiLSTM-CRF明显较其他模型有更好的表现，较模型BiLSTM-CRF在准确率、召回率上分别提升了0.037、0.04。

表6 不同网络结构下的地质报告分词结果

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **P** | **R** | **F** |
| LSTM-CRF | 0.91 | 0.891 | 0.900 |
| BiLSTM | 0.918 | 0.905 | 0.911 |
| BiLSTM-CRF | 0.924 | 0.914 | 0.919 |
| BiLSTM-CNN- CRF | 0.939 | 0.927 | 0.933 |
| BERT-BiLSTM-CRF | 0.961 | 0.958 | 0.959 |

实验中还对比分析了前15轮中F1值更新的情况，其结果如图5所示。从图中可以看出，BERT-BiLSTM-CRF在训练的初期就能够快速达到较高水平并进一步提升，最后维持在比较高的水平上；但是传统的深度学习模型来说，在训练初期获取的结果比较低，经过多次迭代才能上升到一定水平，但最终结果还是无法超过BERT-BiLSTM-CRF本文模型

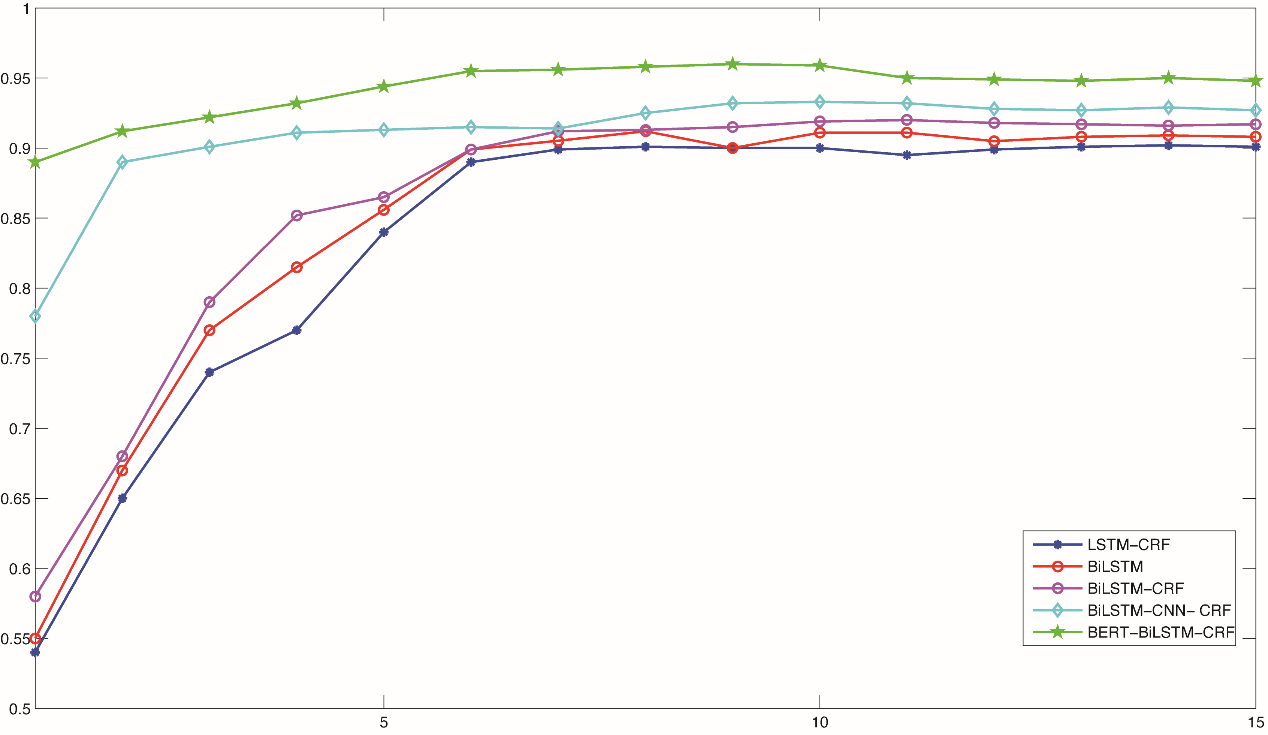


图5 不同网络结构下F1值的更新情况

表7分析了本文的方法无法处理地质报告如下的几种分词：①上下文中分词结果存在不一致情况，比如“成矿带”有时能够正确分割有时却不能，分词全局性不够强，需要增加约束条件对上下文分词进行约束，保证上下文分词一致性；②对语境中词语在特征表现上无显著差异的词语分割不开，比如“有色金属贵金属”。未来的工作中尝试加入全局约束条件，以提高分词的准确性。

表7 地质资料分词实例提取中常见错误分析

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 描述 | 样例 | 错误分析 |
| 1 | 上下文中分词结果不一致，分词全局性不够强 | 测区/成矿/背景/处于/内蒙古大兴安岭成矿省/Ⅱ/级/**成矿**/**带**/ | “成矿带”应该作为一个词，在分词实验中有些地方可以分开，有些地方不能分开 |
| 2 | 语境中词语在特征表现上无显著差异 | 以/**有色金属贵金属**/及/其它/近期/有/经济效益/的/矿产/ | “有色金属”和“贵金属”在该语境中无特征上的显著差异，但应划分为“有色金属**”**和**“**贵金属” |

1. **结论**

针对非结构化地质报告数据的混合性、变异性、稳健性、相关性等随时间、空间及地质体的不同而各异问题，本文提出了一种循环语料训练的BERT-BiLSTM-CRF的中文分词方法。针对语料库的构建，区别传统的方法采用循环自学习策略来丰富通用领域语料库，结合BERT-BiLSTM-CRF深度学习模型，无需依赖领域性的知识及人工设计特征，相比传统的词向量表征能够更好的表征字符级别的语义及语句特性。实验结果表明，本文提出的方法能够有效增强非结构化地质报告分词能力，且能够识别大量地质专业未登录词。未来的研究方向在于将这一方法扩展到其他专业化领域，用来检测本文所提出的循环自学习策略及方法在专业领域中文分词上的普适性。

**参考文献**

1. Ma X, Carranza E J M, Wu C, et al. Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services[J]. Computers & Geoences, 2012, 40:107-119.
2. Liang W , Lei X , Chaoling L , et al. A Knowledge-Driven Geospatially Enabled Framework for Geological Big Data[J]. Isprs International Journal of Geo Information, 2017, 6(6):166.
3. Qiu, Q., Z. Xie, L. Wu, A cyclic self-learning Chinese word segmentation for the geoscience domain. Geomatica,2018, 72, 16-26.
4. Qiu Q , Xie Z , Wu L , et al. GNER: A Generative Model for Geological Named Entity Recognition Without Labeled Data Using Deep Learning[J]. Earth and Space ence, 2019.
5. Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
6. Dauphin Y N , Fan A , Auli M , et al. Language Modeling with Gated Convolutional Networks[J]. 2016.
7. Graves, Alex. [Studies in Computational Intelligence] Supervised Sequence Labelling with Recurrent Neural Networks Volume 385 [J]. 2012, 10.1007/978-3-642-24797-2.
8. 刘一佳, 车万翔, 刘挺, 等. 基于序列标注的中文分词、词性标注模型比较分析[J]. 中文信息学报, 2013, 27(4).
9. 罗桂琼, 费洪晓, 戴弋. 基于反序词典的中文分词技术研究. 计算机技术与发展, 2018, 18, 80-83.
10. 莫建文, 郑阳, 首照宇, 等. 改进的基于词典的中文分词方法[J]. 计算机工程与设计, 2013, 34(005):1802-1807.
11. 刘泽文, 丁冬, 李春文. 基于条件随机场的中文短文本分词方法[J]. 清华大学学报:自然科学版, 2015(08):906-910.
12. 阚琪. 基于条件随机场的命名实体及实体关系识别的研究与应用[D]. 北京交通大学, 2015.
13. Du L , Xia C , Deng Z , et al. A Machine Learning Based Approach to Identify Protected Health Information in Chinese Clinical Text[J]. International Journal of Medical Informatics, 2018, 116(AUG.):24.
14. 张义, 李治江. 基于高斯词长特征的中文分词方法[J]. 中文信息学报, 2016, 30(5).
15. 任智慧, 徐浩煜, 封松林, 等. 基于LSTM网络的序列标注中文分词法[J]. 计算机应用研究, 2017, 34(005):1321-1324,1341.
16. Xu, J., X. Sun. Dependency-based Gated Recursive Neural Network for Chinese Word Segmentation. In Meeting of the Association for Computational Linguistics, 2016, 567-572.
17. Chen, X., X. Qiu, C. Zhu & X. Huang, Gated Recursive Neural Network for Chinese Word Segmentation, 2015.
18. Shu, X., J. Wang, X. Shen, A. Qu, Word segmentation in Chinese language processing. Statistics and Its Interface, 2017.
19. Peters, M. E., Ammar, W., Bhagavatula, C. & Power, R. Semi-supervised sequence tagging with bidirectional language models, Acl, 2017
20. Li, L., W. Wang, B. He, Y. Zhang, A hybrid method for Chinese address segmentation. International Journal of Geographical Information Science, 2017.
21. Jia Y , X. X. Chinese Named Entity Recognition Based on CNN-BiLSTM-CRF. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). 2018, IEEE.
22. Wang Xuan, Xu Ruifeng, He Yulan et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications, 2017.
23. Peters, M. E. et al. Deep contextualized word representations, 2018.
24. Radford, A. & Salimans, T. Improving Language Understanding by Generative Pre-Training, 2018.
25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
26. Ke Z, Shi L, Meng E, et al. Unified Multi-Criteria Chinese Word Segmentation with BERT[J]. arXiv, 2020.
27. 陈婧汶, 陈建国, 王成彬, 等. 基于条件随机场的地质矿产文本分词研究[J]. 中国矿业, 2018(9):69-74,101.
28. Wang C , Ma X , Chen J , et al. Information extraction and knowledge graph construction from geoscience literature[J]. Computers & Geosciences, 2018, 112(MAR.):112-120.
29. 韩冬煦, 常宝宝. 中文分词模型的领域适应性方法[J]. 计算机学报, 2015, 38(002):272-281.
30. 陈二静, 姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017, 1(6):1-11.
31. Mikolov T , Grave E , Bojanowski P , et al. Advances in Pre-Training Distributed Word Representations[J]. 2017.
32. Jawahar G , Sagot B , Djamé Seddah. What Does BERT Learn about the Structure of Language? [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.