www.geojournals.cn/georev

区域地质调查文本中文命名实体识别

邱芹军[1,2)],田苗[3,4)],马凯[3,4)],谢忠[1,2)],金相国[5)],段雨希[5)],陶留锋[1,2)]

1)中国地质大学(武汉)计算机学院,武汉,430074;
2)中国地质大学(武汉)地理信息系统国家地方联合工程实验室,武汉,430074;
3)湖北省水电工程智能视觉监测重点实验室,湖北宜昌,443002;
4)三峡大学计算机与信息学院,湖北宜昌,443002;
5)中国地质大学(武汉)国家地理信息系统工程技术研究中心,武汉,430074

内容提要:作为我国地质调查领域最重要的数据源之一,地质调查报告中蕴含着丰富的地学知识及地质体描述等关键信息,准确高质量地抽取地质命名实体为地学知识图谱构建、知识推理及知识演化提供基础。笔者等在阐述地质命名实体识别任务基础上,分析地质实体不仅包含大量专业术语,还存在实体嵌套、大量长实体等领域特性,进一步增加了地质命名实体识别难度。笔者等提出一种基于轻量级预训练模型(ALBERT)—双向长短时记忆网络(BiLSTM)—条件随机场(CRF)模型的地质命名实体识别方法。首先利用ALBERT对输入字符上下文特征进行建模,并采用BiLSTM对其进行进一步上下文特征表征,最后采用CRF实现标注序列预测。实验结果表明,在构建的地质命名实体识别数据集上,相比于主流的命名实体识别模型算法,本文所提出的方法具有更好的抽取性能,提出的命名实体识别模型能为领域实体识别提供借鉴,同时为地学领域实体关系抽取和地学知识图谱构建提供有力方法支撑。

关键词:地质命名实体识别;轻量级预训练模型;ALBERT;知识图谱;地质报告

地球科学研究已经进入了一个全球覆盖、全天候监测、全要素观测的大数据时代。由于地质对象演化时间长、空间范围大,涉及的因素和过程也非常复杂,关于地质对象的参数、结构、关系、演化等方面的信息存在高度的不完全和不确定性特征,采用自动化的手段从地质大数据中进行数据挖掘有助于发现隐含的知识(Qiu Qinjun et al., 2018; 吴冲龙等, 2020)。在以大数据和数据密集型计算为基础的第四范式主导下,从地质大数据中挖掘有价值的信息并将其应用于地学领域已是必然趋势(Ma Xiaogang et al., 2020;余先川等,2022;雷传扬等,2022;陈忠良等,2022)。当前,地球科学研究已经进入了以大数据驱动为手段构建计算机可理解的知识模式的新阶段(周成虎等,2021;诸云强等,2022)。

2019年中国启动了"深时数字地球国际大科学计划"(Deep-time Digital Earth, DDE),将构建最大的地学基础数据库,形成具有地球演化大数据支撑的"深时数字地球",其中深时地球知识图谱构建正是该项目需要做的重要工作之一(周成虎等,2021; Wang Chengshan et al., 2021; Ma Xiaogang, 2022)。知识图谱是实现语义搜索、智能问答、内容推荐等人工智能任务的知识基础,知识图谱以图的形式存储,以方便检索和推理。命名实体识别作为文本信息抽取、知识图谱构建等任务的基础性工作之一,一直以来都得到了广泛的关注和长足的发展(张雪英等, 2018;Ma Chao et al., 2022;Wang Bin et al., 2022)。

地质报告是地质学家野外考核及科研管理工作所需要的重要优质信息资源,海量高质量的地质报告数据蕴含着丰富可靠的地学知识,如能对地质报告中的这些信息进行自动化抽取和识别,将对自动高效地构建地学知识图谱具有重要意义(Qiu Qinjun et al., 2019a, 2019b)。在地质领域,目前缺少公开

注:本文为国家重点研发计划(编号:2022YFF0711601)、国家自然科学基金资助项目(编号:42050101)和中国博士后科学基金资助项目(编号:2021M702991)的成果。
收稿日期:2022-11-23;改回日期:2023-01-10;网络首发:2023-01-20;责任编辑:章雨旭。Doi: 10.16509/j.georeview.2023.01.085
作者简介:邱芹军,男,1988年生,副研究员;Email: qiuqinjun@cug.edu.cn。通讯作者:马凯,男,1980年生,副教授;Email: makai@ctgu.edu.cn。

标注的数据集,相关研究仍处于起步阶段(储德平等,2021)。地质领域实体识别属于特定领域的实体识别问题,相比通用领域实体识别,它具有如下特点:

(1)地质报告中的地名大多较长且复杂,可读性和可理解性相对偏低。例如:“念青唐古拉主脊”、“昂仁县”、“岗龙乡麻勒果-嘎干拉”等。

(2)地质报告中较多信息都是分条进行陈述,上下文信息不是很连贯。在命名实体识别任务中,实体识别很大程度上依赖于上下文信息,如缺乏上下文信息,会对实体识别产生较大影响。

(3)地质报告中的实体存在大量名称嵌套,即一个实体中包含另一个实体。例如:“薄层细粒长石石英砂岩”,这一单一的岩石实体中包含有“石英”,“石英”属于矿物实体,该情况在命名实体识别中往往会导致识别不全。

(4)地质报告中的实体边界比较模糊。例如“前震旦系念青唐古拉岩群”,在这个词中包含两个实体,分别为“前震旦系”和“念青唐古拉岩群”,它们分别属于地层。综上所述,针对地质领域的文本特性以及实体特点,亟需一种适合该领域的命名实体识别方法。

笔者等主要针对地质报告文本特性以及实体描述特点,通过构建地质领域命名实体数据集,提出了一种基于 ALBERT—BiLSTM—CRF 模型的地质领域命名实体识别方法。为了验证模型性能,使用同一语料库在当前其他主流的命名实体识别模型上进行了对比实验,结果展示笔者等所提出的模型在地质命名实体数据集下具有很好的识别效果。

1 相关工作

地质领域知识图谱构建的基础是命名实体识别,即从未加工的地质领域文本中识别出特定类别的专有名词实体。其准确率直接影响地质领域多种自然语言处理技术的结果(Qiu Qinjun et al., 2019a;谢雪景等,2023;王权于等,2023)。由于基于规则或统计机器学习的传统命名实体识别方法依赖于人工构建文本特征,基于深度神经网络模型的“端到端”识别方法成为当前主要研究方向(焦凯楠等,2021)。

在中文地质命名实体识别任务方面,由于不准确的中文分词可能会造成误差传递问题,以字向量作为输入特征成为中文语境下的另一解决方案。Fan Runyu 等(2019)提出一种基于深度学习的 NER 模型;即深度多分支 BiGRU—CRF 模型。该模型结合了多分支双向门控递归单元(BiGRU)层和 CRF 模型。在端到端和监督过程中,所提出的模型通过多分支双向 GRU 层自动学习和转换特征,并使用 CRF 层增强输出。实验结果表明,所提出的深度多分支 BiGRU—CRF 模型优于最先进的模型。Enkhsaikhan 等(2021a)提出了一个迭代深度学习 NER 框架,使用远程监督来自动标记特定于域的数据集。实验结果证明了这种方法的有效性,并得到了领域专家的进一步证实。Enkhsaikhan 等(2021b)使用监督深度学习方法提取地质实体进行序列标记,使用领域字典进行远程监督,以及通过使用无监督机器学习算法进行语义分组和聚类来提取实体之间的关系类型。并取得了不错的效果。由此可见,当前基于深度神经网络的命名实体识别方法总体上可以规约为由嵌入层、编码层和解码层构成的三层模型架构,其中适宜的特征表示和深度网络模型是实现识别性能提升的关键。

近年来,为克服传统深度神经网络模型在长程记忆能力等方面存在的不足,国内外学者开始将注意力机制引入深度神经网络模型,并在自然语言处理研究领域取得了较好的应用效果。如 Qiu Qinjun 等(2019a)针对传统的 NER 方法严重依赖于特征方程并且存在大量标注不一致的问题,提出了一种基于注意力机制的条件随机场层双向长短时记忆(Att—BiLSTM—CRF)的地质命名实体识别方法,提取任务中的 F1 平均得分达到 91.47%。Qiu Qinjun 等(2019b)采用基于一元语言模型的随机抽取算法生成由概率标注伪句组成的大规模训练数据集,然后将生成的每个句子作为自我训练和学习算法的输入,结果证明可有效的识别地质命名实体。Devlin 等(2018)提出了 BERT 预训练模型,并在通用领域的 11 项文本分析任务中取得了当时的最佳效果,成为当前国内外自然语言处理领域关注的焦点。在此基础上,储德平等(2021)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。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。Liu Hao 等(2022)提出了一种双阶段微调方法。在第一个微调阶段,使用来自 Transformer 语言模型和地质领域知识(GeoBERT)的双向编码器表示,它将地质领域

知识结合在预先训练的 BERT 模型上,在第二阶段,使用少量样本来完成基于 GeoBERT 的地质报告中的 NER 任务。与构建数据集上的基线模型相比,提出的模型获得了较高的 $F1$ 分数。Yu Yuqing 等(2022)提出一种基于矿物领域深度学习的 NER 模型。与之前的 NER 模型相比,引入了 Transformers 的双向编码器表示。将 BERT 与 LSTM 和 CRF 合并,对 MNER 任务进行比较实验。结果表明,该模型能够有效识别 7 个矿物实体,平均 $F1$ 得分为 0.842。

然而目前现有的 NER 方法无法很好的提取序列信息,且会出现梯度消失和梯度爆炸等现象,其梯度问题并未得到良好的解决。此外,现有的 NER 方法无法解决地质领域实体中存在嵌套及大量长实体等问题。开源的 BERT 模型通常基于通用领域文本进行预训练,对于含有大量专业术语和语法结构的特定领域很难直接适用。适用于中文地质文本描述方式特性的命名实体识别方法仍有待进一步研究。

2　区域地质调查文本特性分析

在领域业务特性方面,地质学是地学中的主要分科。随着科学的发展和人类物质文明的需要,地质学的任务已不仅仅局限于解释地球的过去,研究和反演地球的历史。今天地质学研究的内容十分广泛,包括岩石学、矿石学、古生物学、历史地质学、环境地质学、灾害地质学等(Qiu Qinjun et al., 2018;钟自然, 2018; Wang Chengshan et al.,2021)。深入了解地质领域文本特性是制定语料库标注规则的前提。

地质报告文本隶属于中文科技文的范畴,内容是地质勘察成果的总结及对地质调查研究成果的记录和汇总,是根据研究区域的历史数据及当前的勘察成果整理汇编而成。按专业类型可以分为区域地质调查报告、环境地质报告、水文地质报告、矿产地质报告等,按工作阶段可以分为中间报告、总结报告、补充报告等。这些地质报告虽然内容侧重点不同,但结构都比较规范,且最终都需要通过主管部门的评审。一般来讲地质报告中的图表比较丰富,文本中专业术语较多,用词比较规范,篇幅较长,各章节文本主题性较强,一般章节标题就能代表各部分的主题。

通过对地质调查报告分析,发现不同地区的地质报告文本在其结构方面仍然存在差异,但其编写内容都符合相关行业规范要求。在文本描述特性方面,在每一份地质报告中,都涉及了地质时间、地质构造、地层、矿物、岩石和地点等实体要素,也是报告中知识的重要载体,对这些元素进行高效、准确的识别和抽取是实现地质报告知识挖掘的基础。

地质报告文本是记录一定区域范围内相应的地质条件及地质事件,其中涵盖大量的地质实体,而且实体的类型是多样化的。地质报告文本不管是对地质变化的记录、地质状况的描述还是对地质灾害的统计,其本质上都是对相关的地质实体、其伴随的附属信息及相互之间的语义关系的描述。因此,地质实体是中文地质报告中核心的组成要素,其他的关系及属性的描述都是围绕地质实体这一核心要素展开。地质实体往往是中文地质报告中一系列地质知识的主要体现,对其进行精准有效的识别能够进一步提升地质大数据的深度挖掘。

在文本内容特性方面,本文选取的地质报告文本主要来自尼玛区调查报告❶、杂多县幅区域地质调查报告❷、广东万阳春市幅区域地质调查成果报告❸、金牛镇幅高桥幅区域地质调查报告❹,通过对这些调查报告的分析,发现不同地区的地质报告文本在其结构方面仍然存在较大的差异,这方面并无统一的规范化标准,但其核心仍然是围绕通用的地质本体库等相关行业规范要求来编写的。其中,地质年代、地质构造、地层、矿物、岩石和地点是地质报告中最重要的部分,对其进行高效、准确的识别是实现地质报告信息抽取的基础,具有重要的研究意义和发展前景。

3　语料库构建

3.1　标注体系

实体是知识图谱中的基本单元,是文本中承载信息的重要语言单位。本文在分析中文文本领域命名实体信息描述规则和抽取方法特点的基础上,参考地质领域的行业规范(中国地质调查局地质调查技术标准),结合地质报告中关键信息的描述特点,将地质调查报告和地质科技文献这类文本中出现的关键内容划分为地质时间、地质构造、地层、岩石、矿物、地名六种类型,总结归纳出一套适用于地质领域的命名实体参考体系。地质领域术语识别是构建地质知识图谱的基础性工作,而一个地质领域知识图谱需要包含地质体的属性及层次关系、时空关系,又包括通用类(如地点)的基础要素。地质科学研究具有明显的区域性,而且地质数据具有显著的时空特性,脱离了空间位置的地质数据没有意义。地质时间是地学研究中的基础性的要素也是关键性的要

素,通过地质时间能够直接反映地质体的时空演化规律。地点反映的是地质体产生的区域范围及空间信息。本文在基于构造地学知识图谱及地学演化机理基础上,结合已有的地质领域本体库,围绕地质体标注与其相关的地质时间、地质构造、地层、矿物、岩石和地点等语言单元。

命名实体识别中的人名、地名、组织机构名等实体,都属于序列标注问题。在序列标注中,目标是对给定序列的每个元素标注一个标签。一般来说,一个序列代表一个句子,而一个元素指的是句子中的一个词语或者一个字。针对地质领域语料库,结合实体类型,在标注语料库的过程中,所采用的标注方法是 BIOES 标注法,分别表示实体片段的开始(Begin)、实体片段的中间(Inside)、实体片段的结束(End)、单个字的实体(Single)、非实体(Other),并去除了英文单词、特殊符号图表等其他无关信息。

3.2 标注结果

地质年代表示地球历史的地质时间单位,是地质历史中一个连续的时间片段。地质构造泛指地壳运动形成的地球构造现象和特征,包括从大地构造到各种构造特征(如断层、褶皱、岩层、岩体等)。地层是具有某种共同特征或属性的岩石体。岩石是天然产出的具有一定结构构造的矿物集合体,它构成地球上层部分(地壳和上地幔),在地壳中具有一定的产状。矿物是具有一定化学组成的天然化合物,它具有稳定的相界面和结晶习性。地名是人们赋予某一特定空间位置上自然或人文地理实体的专有名称。

此次实验将地质报告划分为六大类实体并综合当前语料库构建工作及地质领域文本特性制定标准体系,对地质年代、地质构造、地层、岩石、矿物、地名六大类实体将进行标注。具体标注结果实体类型及标签如表 1 所示,示例如表 2 所示。

4 ALBERT—BiLSTM—CRF 模型

笔者等提出的 ALBERT—BiLSTM—CRF 模型结构图如图 1 所示。该模型图共分为三部分,分别为 ALBERT 层、BiLSTM 层和 CRF 层。首先将原始文本输入 ALBERT 层,通过 ALBERT 层的预训练过程,可以将词表表征为向量的形式,然后将输出的向量输入到 BiLSTM 层进行编码,前向的 LSTM 可以挖掘下文的特征,而反向 LSTM 则可以挖掘上文的信息,最终得到全局特征,即在 t 时刻所得到的隐藏状态 h_t,最终通过 CRF 层,利用 CRF 进行解码,从而输出最好的标签序列。

4.1 ALBERT 预训练语言模型

ALBERT 模型与 BERT 模型的原理相同,均采用了 Transformer 模型的方案,但 ALBERT 模型共享了所有层的参数,使得参数数据大量减少,因此在本次实验研究中,笔者等使用 ALBERT 模型进行预训练。

BERT 模型是一个无监督的预训练模型,基于 Transformer 的双向编码器(Liu Hao et al., 2022),BERT 模型的核心部分就是 Transformer,将其结构进行堆叠,便可以形成 BERT 模型,结构简图如图 2 所示。将原文本进行输入时,每个序列的第一个词

表 1 实体类型及示例

Table 1 Entity type and examples

实体类型	符号	示例	例句
地质年代	GTM	晚白垩世、晚侏罗世、三叠纪、古生代 多数为 xxx 世、xxx 代、xxx 纪…	该亚带侵入**早白垩世**多尼组、化石表明多尼组形成于**阿普特期**,同时发现部分岩脉侵入**晚白垩世**镜柱山组
地质构造	GST	海陆交互相、火山通道相、浅海相等	去申拉组火山岩岩相划分为**爆发相**、**喷溢相**、**喷发沉积相**、**火山通道相**等 4 种基本类型
地层	STR	结扎群、诺日巴尕日保组、巴塘群 多数为 xxx 群、xxx 组、xxx 系	该亚带侵入早白垩世**多尼组**、化石表明**多尼组**形成于阿普特期,同时发现部分岩脉侵入晚白垩世**镜柱山组**
岩石	ROC	英云闪长岩、细粒斜长岩、花岗岩 多数为 xxx 岩	岩石类型以**片麻岩**、**混合岩**、**石英岩**、**大理岩**为主,总厚度大于 5000m,变质程度达角闪岩相
矿物	MIN	斜长石、石英、黑云母、磁铁矿等 多数为 xxx 石、xxx 矿	深解理和颗粒边缘被**纤闪石**交代;基质由**斜长石**微晶和基质组成,**绿泥石**、**玉髓**充填形成杏仁体
地名	PLA	治多县、通天河、阿文俄育、肖恰错等	晚白垩世**折勒—得木我**碱性斑岩亚带分布于图幅南部**昂子错**附近,西延出图,部分地段被第四系覆盖。

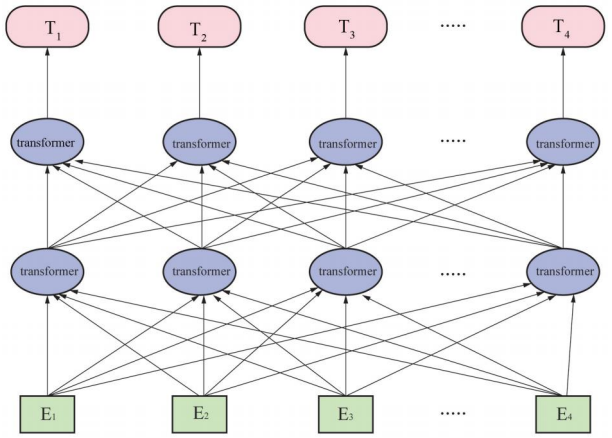

图 1 ALBERT—BiLSTM—CRF 模型图

Fig. 1 ALBERT—BiLSTM—CRF model diagram

始终是特殊分类嵌(【CLS】),而剩下的每一个词则代表一个汉字。BERT 模型的输入分为 3 个部分,分别为词向量、句子向量和位置向量。词向量层可以将每一个词转换为固定维度的向量;句子向量层可以用作分类任务,最后位置向量层是在训练过程中得到的。BERT 模型通过 Transformer 可以有效地捕捉句子之间的上下文关系,使得命名实体识别任务的精度得到了大幅度的提升。

笔者等使用 ALBERT 模型判断句子中的每一个单词是否为实体,微调时将整个句子作为输入,在每一个时间片输出一个概率,从而得到实体类别。

4.2　BiLSTM 层

LSTM 是 RNN 的一种,使用 LSTM 模型可以更好地捕捉到较长距离的依赖关系,这是由于 LSTM 在训练的过程中可以学到记忆哪些信息和遗忘哪些信息(Schmidhuber et al. , 1997)。LSTM 模型是由 t 时刻的输入词 X_t ,细胞状态 C_t ,临时细胞状态 $\widetilde{C_t}$,隐藏层状态 h_t ,遗忘门 f_t ,记忆门 i_t ,和输出状态 o_t 组成。

(1) 计算遗忘门的输出。LSTM 中的遗忘门可以在一定概率上控制是否遗忘上一层的隐藏细胞状态。当输入上一序列的隐藏状态 h_{t-1} 和本次序列输入的数据 x_t 时,可以通过一个激活函数得到遗忘门的输出,即 f_t 。其计算公式为:

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{1}$$

(2)计算输出门。其主要负责处理当前序列位置的输入,主要分为两个阶段,在第一个阶段主要使用 sigmoid 激活函数,输出 i_t ,第二个部分则使用 tanh 激活函数,输出为 $\widetilde{C_t}$ 。其计算公式为:

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{2}$$

$$\widetilde{C_t} = tanh(W_c \cdot [h_{t-1}, X_t] + b_C) \tag{3}$$

(3) 在进行输出门之前,应先计算当前时刻的细胞状态 C_t 。其计算公式如下:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{4}$$

(4) 最后计算当前时刻的隐藏层状态及输出的结果。其计算公式如下:

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{5}$$

$$h_t = o_t \cdot tanh(C_t) \tag{6}$$

前向的 LSTM 和反向的 LSTM 相结合即 BiLSTM,这种构建网络结构可以更方便的提取上下文信息,使之充分的提取文本特征,其结构图如图 3 所示。

4.3　CRF 层

通过 BiLSTM 层后,得到的输出结果表示该单词对应各个类别的分数,将这些分数作为 CRF 层的输入,类别序列中分数最高的类别即预测的最终结果。选择加入 CRF 层,是因为在 BiLSTM 层可对序列中的上下文信息进行分析,并输出最终的得分并选取最高的得分,但 BiLSTM 模型却无法考虑到序列之间的限制关系。通过 CRF 层,可以为最后预测

表 2 数据标注示例

Table 2 Data set annotation examples

语料	黄	铜	矿	呈	星	散	状	细	脉	
标注	B-MIN	I-MIN	E-MIN	O	O	O	O	O	O	O
语料	状	,	团	块	状	,	不	均	匀	分
标注	O	O	O	O	O	O	O	O	O	O
语料	布	于	砂	岩	中	,	砂	岩	表	面
标注	O	O	B-ROC	E-ROC	O	O	B-ROC	E-ROC	O	O
语料	孔	雀	石	、	铜	蓝	等	内	部	为
标注	B-MIN	I-MIN	E-MIN	O	B-MIN	E-MIN	O	O	O	O
语料	黄	铜	矿	沉	积	性	矿	化	点	。
标注	B-MIN	I-MIN	E-MIN	O	O	O	O	O	O	O

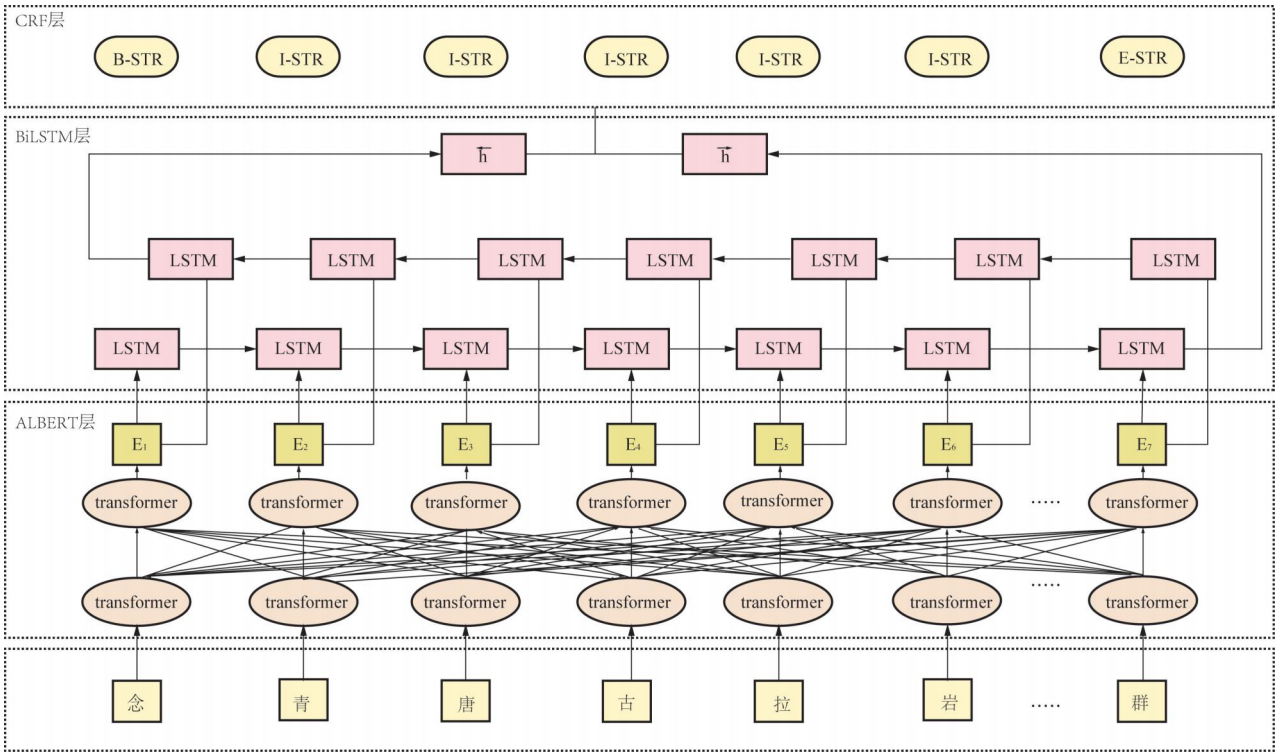

图 2 BERT 模型图

Fig. 2 BERT model diagram

的标签添加一些约束条件,并且保证标签的合法性,主要约束为:

(1) 一个实体的开头是以“B-”和“O”开始的,而非“I-”。

(2) 标签“B-label1, I-label2, …, E-label3”中,label1,label2,label3 应该属于同一类实体,否则就判定为非法标签序列。

(3)有效的标签序列只能为“O B-label”,而“O I-label”则是非法序列。

CRF 模型中有两类特征函数,分别为状态特征函数和转移特征函数。状态特征函数是用当前节点,即某个输出位置的状态分数表示;而转移特征是用上一个节点到当前节点的转移分数表示。笔者等考虑线形链 CRF。即当定义好一组特征函数时,给每一个特征函数 f_j 赋予一个权重 λ_j,便可以利用特征函数集来进行评分,数学表达式如下:

$$score(l \mid s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1}) \qquad (7)$$

其中 s 代表一个句子,l 为所标注的序列,表达式中有两个求和,外层求和用来求每一个特征函数 f_j 评分值的和,里面的求和用来求句子中每个位置单词的特征值的和。对这个分数进行指数化和标准化,即可得到标注序列 l 的概率值 $p(l \mid s)$,数学表达式如下:

$$p(l \mid s) = \frac{\exp[score(l \mid s)]}{\sum_{l'} \exp[score(l' \mid s)]} \qquad (8)$$

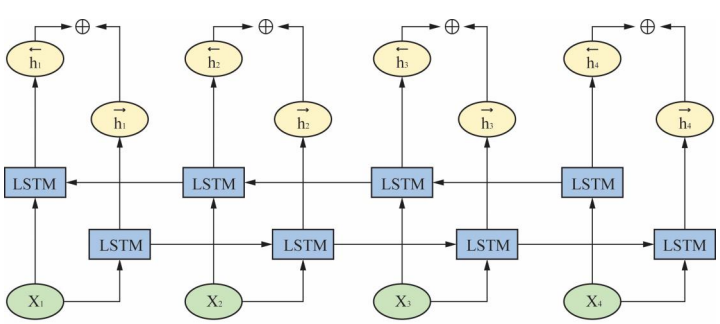

图 3 BiLSTM 模型图

Fig. 3 BiLSTM model diagram

实验采用 ALBERT—BiLSTM—CRF 对地质领域进行命名实体识别,其主要优势在于利用 ALBERT 模型中的 Transformer 可以有效地捕捉句子之间的上下文关系,使得命名实体识别任务的精度得到了大幅度的提升;其次采用 BiLSTM 这一模型可以更好地捕捉较长距离之间的依赖关系;最终通过使用 CRF 层,可以使

得为模型提供一个标签约束关系,最终选择得分最高的序列作为最终的输出结果。

5　实验分析

5.1　语料库设置

实验所采用的语料库为自行标注的面向地质领域命名实体识别语料库。该语料库包含四份地质报告文本,分别为治多县福地质报告、广东万阳春市地质报告、尼玛区调查报告以及金牛镇福、高桥区域地质报告。语料库中共含有 6 类实体,包括地质年代、地质构造、地层、岩石、矿物、地名。语料库共包含标记句子 10803 句,已标注实体字数 100106 字,非实体字数 598406,6 类实体具体数目如表 3 所示。

实验将按照 8∶1∶1 的比例将语料库划分为训练集、验证集、测试集,并采用传统的准确率(Precision,P)、召回率(Recall,R)和 F1 值($F1$)作为评价指标。

5.2　实验及参数设置

实验对模型 ALBERT—BiLSTM—CRF 参数设置如下表 4 所示。

由于地质语料库文本的特殊性,经常出现长句描述,故设置最大输入句子序列长度为 256;对于 ALBERT—BiLSTM—CRF 模型,优化器设置为 Adam,设置 Batch size 为 16,迭代次数为 20 次;为了防止过拟合现象,加入 Dropout 并设置参数为 0.1。

为了验证 ALBERT—BiLSTM—CRF 模型的有效性,在使用该地质语料库的基础上,设置了 6 组对比实验,使用的模型包括了 ALBERT—BiLSTM、BiLSTM—CRF、BiLSTM—Attention—CRF、BERT—BiLSTM—CRF、IDCNN、IDCNN_v2,均为命名实体识别领域常见模型。

5.3　实验结果

5.3.1　不同实体实验结果比较

ALBERT—BiLSTM—CRF 与 ALBERT—BiLSTM 模型实验结果如表 5 所示。

从表 5 可以看出,在地质语料库中,所构建的 ALBERT—BiLSTM—CRF 模型 $F1$ 值达到了 76.27%,明显高于 ALBERT—BiLSTM 模型的 65.17%,这是由于 CRF 层可以增加一些约束到最后的预测标签,使得非法序列减少,从而增强最后模型预测结果的准确性。对于各个实体的预测结果来说,矿物类型实体(MIN)预测准确度最高,F1 值达到了 80.41%,这是因为矿物类型实体数量较多,而且词汇本身多为固有词汇,例如“黑云母”、“石英”等,歧义较少;地名实体(PLA)预测结果最低,且两模型 $F1$ 值差异较大,说明 CRF 层对语料库中相对复杂的地名名词有着较好的识别效果;其他实体,包括地质年代(GTM)、地质构造(GST)、地层(STR)、岩石(ROC)$F1$ 值均衡,总体来看识别效果较为理想。

5.3.2　不同模型的性能比较

表 6 和表 7 分别描述了无预训练模型和 BERT

表 3 实体数量统计

Table 3 Statistics of the number of entities

实体类型	数量	占比
地质年代(GTM)	1864	7.99%
地质构造(GST)	1359	5.82%
地层(STR)	3016	12.92%
岩石(ROC)	9827	42.09%
矿物(MIN)	4924	21.09%
地点(PLA)	2355	10.09%

表 4 模型参数设置

Table 4 Model parameters setting

参数名	参数值
MAX_SEQ_LEN	256
Batch_size	16
Epoch	20
Dropout	0.1
Optimizer	Adam
BiLSTM_Units	128

表 5 ALBERT—BiLSTM—CRF 模型实验结果

Table 5 Experimental results of ALBERT—BiLSTM—CRF model

实体类型	ALBERT—BiLSTM—CRF			ALBERT—BiLSTM		
	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$ (%)
GTM	81.82	68.94	74.83	59.68	64.26	61.89
GST	88.57	56.36	68.89	55.63	47.88	51.47
STR	80.00	69.27	74.25	64.44	64.80	64.62
ROC	76.68	72.79	74.68	62.12	65.27	63.66
MIN	79.22	81.64	80.41	68.41	76.87	72.39
PLA	65.88	69.14	67.47	36.36	54.32	43.56
Average	**78.64**	**74.03**	**76.27**	**62.78**	**67.76**	**65.17**

表 6 无预训练模型条件下不同模型的对比实验结果

Table 6 Comparative experimental results of different models without pre-training model

模型	P(%)	R(%)	$F1$(%)
IDCNN	63.30	63.98	62.84
IDCNN_v2	68.30	61.33	64.33
BiLSTM—Attention—CRF	60.50	62.08	60.11
BiLSTM—CRF	59.38	58.15	56.97
ALBERT—BiLSTM—CRF	**78.64**	**74.03**	**76.27**

表 7 预训练模型条件下不同模型的对比实验结果

Table 7 Comparative experimental results of different models based on pre-training model conditions

模型	*P*(%)	*R*(%)	*F*1(%)
BERT—Trans—CRF	70.11	70.09	70.10
BERT—CNN—CRF	71.19	71.21	71.20
BERT—BiLSTM—CRF	73.21	73.05	73.13
BERT—CNN—BiLSTM—CRF	72.99	72.18	72.58
Lattice—LSTM—CRF	71.98	71.77	71.87
ALBERT—BiLSTM—CRF	**78.64**	**74.03**	**76.27**

预训练模型两组实验条件下不同模型的综合对比结果。总体而言,本文模型在两组实验中的精确率、召回率和 *F*1 值三方面均优于其他对比实验模型,具有更佳的综合识别效果。

在第一组无预训练模型实验条件下,所有对比模型均基于随机初始化的字向量。BiLSTM—CRF 模型的实验结果较差,*F*1 指标仅有 60.11%。其原因主要在于该模型仅使用了 BiLSTM 作为字符特征提取单元,对于字符级命名实体的前后顺序方向性特征提取能力不足,导致识别效果较差。类似地,IDCNN 模型的 *F*1 值也仅有 62.84%,综合识别效果也有待提升。与 BiLSTM—CRF 模型相比,BiLSTM—Attention—CRF 模型的正确率提升 1.12%,召回率提升 3.93%,*F*1 值也提升 3.14%。

正如本文第 3 节所述,除了蕴含丰富的专业术语以外,地质领域命名实体还具有较强的字符多义性、位置相关性以及方向敏感性等特点,结合 ALBERT 和 BiLSTM 特征提取能力的本文方法能够同时捕获长距离和方向性特征,相较于 BiLSTM—Attention—CRF 模型,对比实验结果的正确率提升了 19.26%,召回率提升了 15.88%,*F*1 值提升了 19.3%,取得了在第一组实验测试集下的综合最优效果。

第二组实验采用 BERT 中文预训练模型将输入序列字符映射为字向量,并固定 BERT 模块参数,对下游对比模型进行参数微调。如表 7 所示的实验结果表明,BERT 预训练模型的引入对所有实验对比模型的识别性能均有较大提升作用,第一组实验中效果较差的 Trans—CRF 模型识别 *F*1 值达到 70.1%。CNN—CRF 模型的 F1 值提升至 71.2%。笔者等提出的 ALBERT—BiLSTM—CRF 模型的识别正确率提升至 78.64%,召回率提升至 74.03%,*F*1 值为 76.27%,优于 Lattice—LSTM—CRF 模型 *F*1 值 4.4%,仍然具有实验测试集下最优的综合识别效果。

5.3.3 案例分析

与通用领域命名实体识别语料库相比,笔者等所构建的面向地学领域命名实体识别语料库具有一些特殊的特点,使得实验结果相比于通用领域命名实体识别准确率还存在一定差距(如表 8 所示)。具体特点如下:

(1)在地学领域特别是地质报告文本中,地名大多比较长而且复杂,如“尼玛县岗龙乡麻勒果-嘎干拉”、“麻勒果”等;而通用领域语料库(例如《人民日报》语料库)中主要包括人名、地名、机构名和其他专有名词,其往往出现的是比较常见且熟知的地名,如“北京、台湾等”。因此采用的 BiLSTM 能够更

表 8 地质实体识别抽取结果

Table 8 Extraction Results of Geological NER

句子	正确抽取类型	错误抽取类型
以昂杰组中上部为代表,主要岩性为中粒石英杂砂岩、中细粒长石石英砂岩,表现为色浅、层厚,粒粗等特征	(1) 昂杰组(STR) (2) 中粒石英杂砂岩(ROC) (3) 细粒长石石英砂岩(ROC)	(1)石英(MIN) (2)砂岩(ROC)
通过以上资料分析,结合区域构造背景可以看出,第一期的北向逆冲期参与变形的地层主要为中晚三叠世巴颜喀拉山群	(1) 中晚三叠世(GTM) (2) 巴颜喀拉山群(STR)	(1) 中晚三叠世巴颜喀拉山群(GTM)
出露于尼玛县卓尼乡—申扎县夏谷北,呈东西向带状展布,出露面积约 1 200 km^2	(1) 尼玛县卓尼乡—申扎县夏谷北(PLA)	(1) 尼玛县卓尼乡(PLA) (2) 申扎县夏谷北(PLA)
该亚带闪长岩形成于俯冲构造阶段,其侵入早白垩世则弄群火山岩,与其空间关系密切	(1) 闪长岩(ROC) (2) 早白垩世(GTM) (3) 则弄群火山岩(ROC)	(1) 亚带闪长岩(ROC) (2) 早白垩世则弄群火山岩(ROC)
私荣藏布以西下部为灰白、灰黄色中层细粒长石石英砂岩	(1) 灰黄色中层细粒长石石英砂岩(ROC)	(1) 灰白、灰黄色中层细粒长石石英砂岩(ROC) (2)石英(MIN)

好利用句子级的语义特征,而特征模板只能在窗口内提取,无法利用上下文内容。而《人民日报》中出现的实体大多数比较灵活且长度较短,所以同一模型下《人民日报》的训练效果可能更好一些。

(2)相对于地质报告,《人民日报》中的实体边界更加明确,更有利于模型理解上下文信息。而地质报告相对复杂晦涩,且存在大量干扰性公式符号等使得整体较为复杂,因而模型可能对边界信息抽取不完整,界限不明确从而影响抽取的结果。

(3)地质报告中的实体存在实体名称嵌套,即在一个实体中包含另一个实体。例如:“薄层细粒长石石英砂岩”,这一单一的岩石实体中包含有“石英”,“石英”属于矿物实体,这种情况在命名实体识别中可能会发生错误。

(4)地质报告中很多信息都是分条进行陈述,上下文信息不是很连贯。在 NER 任务中,实体识别很大程度上依赖于上下文信息,如果缺乏上下文信息,会对实体识别产生影响。其次从可用性角度上看,本次实验研究是针对地质报告文本中的六大类实体,并且最终的实验结果基本上达到了 75%以上。

6　总结

地球科学研究已经进入了大数据时代,在以大数据驱动的第四科学范式下,地球科学大数据相关的分析理论与方法还亟需完善与拓展,建立全域地学知识图谱模型、探究地学知识演化规律等,是目前地学知识研究中的重要发展战略与前沿性课题。而地质命名实体的识别是地学知识图谱构建的基础性工作。

在上述目标任务驱动下,针对地学领域文本描述特性,笔者等提出了一种基于 ALBERT—BiLSTM—CRF 模型的地质领域命名实体识别方法。该方法利用 ALBERT 对输入字符上下文特征进行建模,并采用 BiLSTM 对其进行进一步上下文特征表征,最后采用 CRF 实现标注序列预测。实验结果表明,在构建的数据集上,该方法能够有效识别地质年代、地质构造、地层、岩石等领域实体,同时与现有方法相比,具有更好的性能。

在未来的研究工作中,笔者等还需要构建更加完善及丰富的语料库,研究顾及不同粒度下的地质命名实体识别,以及考虑引入更多的领域先验知识及多源数据源,进一步提升领域命名实体识别性能。

附　录　/　appendix

ALBERT—A lightweight pre-training model / 轻量级预训练模型

BERT—Bidirectional Encoder Representations for Transformers / 预训练模型

BiLSTM—Bi-directional long and short-term memory network / 双向长短时记忆网络

CRF—Conditional random field / 条件随机场

NER—Named entity recognition / 命名实体识别

注　释　/　Notes

❶ 河南省地质调查院. 2002. 尼玛区幅 H45C001003 1/25 万区域地质调查报告. 河南省地质调查院.

❷ 青海省地质调查院. 2006. 杂多县幅 I46C004004 1/25 万区域地质调查报告. 青海省地质调查院.

❸ 广东省地质调查院. 2004. 广东阳春市幅 F49C002003 1/25 万区域地质调查成果报告. 广东省地质调查院.

❹ 湖北省地质调查院. 2009. 金牛镇幅(H50E012003)高桥幅(H50E013003)1/5 万区域地质调查报告. 广东省地质调查院.

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)

陈忠良, 袁峰, 李晓晖, 张明明. 2022. 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. 地质论评, 68(2): 742~750.

储德平, 万波, 李红, 方芳, 王润. 2021. 基于 ELMO—CNN—BiLSTM—CRF 模型的地质实体识别. 地球科学, 46(8): 3039~3048.

焦凯楠, 李欣, 朱容辰. 2021. 中文领域命名实体识别综述. 计算机工程与应用, 57(16): 1~15.

雷传扬, 刘兆鑫, 文辉, 范敏, 蒋华标, 王波, 马国玺, 谢海洋, 陶海江, 郝金波. 2022. 基于多源数据和先验知识约束的复杂地质体三维建模研究. 地质论评, 68(4): 1393~1411.

王权于, 李振华, 涂志鹏, 陈冠宇, 胡君, 陈嘉麒, 陈建军, 吕国斌. 2023. 基于 BERT—BiGRU—CRF 模型的岩土工程实体识别[OL]. 地球科学: 1~13; https://doi.org/10.3799/dqkx.2022.462

吴冲龙, 刘刚, 周琦, 张夏林, 徐凯. 2020. 地质科学大数据统合应用的基本问题. 地质科技通报, 39(4): 1~11.

余先川, 王桂安. 2000. 空间推理和空间知识表示研究进展. 地质论评, 46: 384~386.

谢雪景, 谢忠, 马凯, 陈建国, 邱芹军, 李虎, 潘声勇, 陶留锋. 2023. 结合 BERT 与 BiGRU—Attention—CRF 模型的地质命名实体识别[OL]. 地质通报: 1~13.

钟自然. 2018. 做好传统地质、建好绿水青山——在中国地质学会第十二次全国会员代表大会上的讲话. 地质论评, 64(1): 10~14.

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: 343~351.

周成虎, 王华, 王成善, 侯增谦, 郑志明, 沈树忠, 成秋明, 冯志强, 王新兵, 闾海荣, 樊隽轩, 胡修棉, 侯明才, 诸云强. 2021. 大数据时代的地学知识图谱研究. 中国科学: 地球科学, 51(7): 1070~1079.

诸云强, 孙凯, 胡修棉, 闾海荣, 王新兵, 杨杰, 王曙, 李威蓉, 宋佳, 苏娜, 牟兴林. 2022. 大规模地球科学知识图谱构建与共享应用框架研究与实践[OL]. 地球信息科学学报: 1~13. DOI: 10.12082/dqxxkx.2022.210696

Chen Zhongliang, Yuan Feng, Li Xiaohui, Zhang Mingming. 2022&.

Joint extraction of named entities and relationships from Chinese rock description text based on BERT—BILSTM—CRF model. Geological Review, 68 (2): 742~750.

Chu Deping, Wan Bo, Li Hong, Fang Fang, Wang Run. 2021&. Geological entity recognition based on ELMO—CNN—BiLSTM—CRF model. Earth Science, 46 (8): 3039~3048.

Devlin J, Chang M W, Lee K, Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding [OL]. arXiv preprint arXiv: 1810.04805.

Devlin J, Chang Ming-Wei, Lee Kenton, Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding[OL]. arXiv preprint arXiv: 1810.04805

Enkhsaikhan M, Liu Wei, Holden E J, Duuring P. 2021a. Auto-labelling entities in low-resource text: a geological case study. Knowledge and Information Systems, 63(3): 695~715.

Enkhsaikhan M, Holden E J, Duuring P, Liu Wei. 2021b. Understanding ore-forming conditions using machine reading of text [OL]. Ore Geology Reviews, 135: 104200.

Fan Runyu, Wang Lizhe, Yan Jining, Song Weijing, Zhu Yingqian, Chen Xiaotao. 2019. Deep learning-based named entity recognition and knowledge graph construction for geological hazards. ISPRS International Journal of Geo-Information, 9(1): 15.

Jiao Kainan, Li Xin, Zhu Rongchen. 2021&. Overview of Chinese domain named entity recognition. Computer Engineering and Applications, 57 (16): 1~15.

Lei Chuanyang, Liu Zhaoxin, Wen Hui, Fan Min, Jiang Huabiao, Wang Bo, Ma Guoxi, Xie Haiyang, Tao Haijiang, Hao Jinbo. 2022&. Research on 3D modeling of complex geological body based on multi-source data and prior geological knowledge. Geological Review, 68 (4): 1393~1411.

Liu Hao, Qiu Qinjun, Wu Liang, Li Wenjia, Wang Bin, Zhou Yuan. 2022. Few-shot learning for name entity recognition in geological text based on GeoBERT[OL]. Earth Science Informatics, 15: 979~991; https: //doi. org/10. 1007/s12145-022-00775-x

Ma Chao, Kale A S, Zhang Jiyin, Ma X G. 2022. A knowledge graph and service for regional geologic time standards [OL]. Geoscience Frontiers; https: //doi. org/10. 1016/j. gsf. 2022. 101453

Ma X G. 2022. Knowledge graph construction and application in geosciences: A review[OL]. Computers & Geosciences; https: //doi. org/10. 1016/j. cageo. 2022. 105082

Ma X G, Ma Chao, Wang Chengbin. 2020. A new structure for representing and tracking version information in a deep time knowledge graph [OL]. Computers & Geosciences, 145: 104620.

Qiu Qinjun, Xie Zhong, Wu Liang, Li Wenjia. 2018. DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. Computers & geosciences, 121: 1~11.

Qiu Qinjun, Xie Zhong, Wu Liang, Tao Liufeng. 2019a. GNER: A Generative Model for Geological Named Entity Recognition Without Labeled Data Using Deep Learning. Earth and Space Science, 6 (6): 931~946.

Qiu Qinjun, Xie Zhong, Wu Liang, Li Wenjia. 2019b. BiLSTM—CRF for geological named entity recognition from the geoscience literature [OL]. Earth Science Informatics, 12: 565~579. https: //doi. org/10. 1007/s12145-019-00390-3

Schmidhuber J, Hochreiter S. 1997. Long short-term memory. Neural Computation, 9(8): 1735~1780.

Wang Quanyu, Li Zhenhua, Tu Zhipeng, Chen Guanyu, Hu Jun, Chen Jiaqi, Chen Jianjun, Lü Guobin. 2023&. Geotechnical Named Entity Recognition Based on BERT—BiGRU—CRF Model [OL]. Earth Science: 1~13. https: //doi. org/10. 3799/dqkx. 2022. 462

Wang Bin, Wu Liang, Li Wenjia, Qiu Qinjun, Xie Zhong, Liu Hao, Zhou Yuan. 2021. A semi-automatic approach for generating geological profiles by integrating multi-source data [OL]. Ore Geology Reviews, 134: 104190.

Wang Bin, Wu Liang, Xie Zhong, Qiu Qinjun, Zhou Yuan, Ma Kai, Tao Liufeng. 2022. Understanding geological reports based on knowledge graphs using a deep learning approach [OL]. Computers & Geosciences, 168: 105229.

Wang Chengshan, Hazen R M, Cheng Qiuming, Stephenson, M H, Zhou Chenghu, Fox P. 2021. The Deep-Time Digital Earth program: data-driven discovery in geosciences [OL]. National Science Review, 8(9): nwab027.

Wu Chonglong, Liu Gang, Zhou Qi, Zhang Xialin, Xu Kai. 2020&. Fundamental problems of integrated application of big data in geoscience. Bulletin of Geological Science and Technology, 39 (4): 1~11.

Xie Xuejing, Xie Zhong, Ma Kai, Chen Jianguo, Qiu Qinjun, Li Hu, Pan Shengyong, Tao Liufeng. 2023&. Geological named entity recognition based on BERT and BiGRU—Attention—CRF Model [OL]. Geological Bulletin of China: 1~13.

Yu Xianchuan, Wang Gui'an. 2000&. Progress in spatial reasoning and spatial knowledge representation. Geological Review, 46: 384~386.

Yu Yuqing, Wang Yuzhu, Mu Jingqin, Li Wei, Jiao Shoutao, Wang Zhenhua, Zhu Yueqin. 2022. Chinese mineral named entity recognition based on BERT model [OL]. Expert Systems with Applications; https: //doi. org/10. 1016/j. eswa. 2022. 117727

Zhang Xueying, Ye Peng, Wang Shu, Du Mi. 2018&. Geological entity recognition method based on Deep Belief Networks. Acta Petrologica Sinica, 34: 343~351.

Zhong Ziran. 2018#. More efforts to do traditional geology, more beautful to construct blue streams and green Hills. Geological Review, 64 (1): 10~14.

Zhou Chenghu, Wang Hua, Wang Chengshan, Hou Zengqian, Zheng Zhiming, Shen Shuzhong, Cheng Qiuming, Feng Zhiqiang, Wang Xinbing, Lu Hairong, Fan Junxuan, Hu Xiumian, Hou Mingcai, Zhu Yunqiang. 2021&. Prospects for the research on geoscience knowledge graph in the big data era. Science China: Earth Sciences, 51 (7): 1070~1079.

Zhu Yunqiang, Sun Kai, Hu Xiumian, Lü Hairong, Wang Xinbing, Yang Jie, Wang Shu, Li Weirong, Song Jia, Su Na, Mu Xinglin. 2022&. Research and practice on the framework for the construction, sharing, and application of large-scale geoscience knowledge graphs [OL]. Journal of Geo-information Science: 1~13; DOI: 10.12082/dqxxkx. 2022. 210696

Chinese named entity recognition for regional geological survey text

QIU Qinjun[1, 2)], TIAN Miao[3, 4)], MA Kai[3, 4)], XIE Zhong[1, 2)],
JIN Xiangguo[5)], DUAN Yuxi[5)], TAO Liufeng[1, 2)]

1) *School of Computer Science, China University of Geosciences, Wuhan*, 430074;

2) *National Local Joint Engineering Laboratory of Geographic Information System, Wuhan*, 430074;

3) *Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, Hubei*, 443002;

4) *College of Computer and Information Technology, China Three Gorges University, Yichang, Hubei*, 443002;

5) *National Engineering Research Center for Geographic Information System, Wuhan*, 430074

Abstract: As one of the most important data sources in the field of geological survey in China, geological survey texts contain a wealth of geological knowledge and descriptions of geological bodies and other key information, and accurate and effective extraction of geological entities in this field can provide the basis for geological knowledge graph and knowledge inference. In this paper, based on the description of the geological named entity recognition task, it is analysed that geological entities contain a large number of terminologies along with domain characteristics such as entity nesting and a large number of long entities, which further increase the difficulty of geological named entity recognition. A lightweight pre-training model (ALBERT) — bi-directional long and short-term memory network (BiLSTM) — conditional random field (CRF) model is proposed for geological named entity recognition. Firstly, ALBERT is used to model the contextual features of the input characters, and BiLSTM is used to further characterize the contextual features, and finally CRF is used to achieve annotated sequence prediction. The experimental results show that the proposed method has superior extraction performance than the mainstream named entity recognition model algorithms on the constructed geological named entity recognition datasets, and the proposed named entity recognition model can provide reference for domain entity recognition, as well as provide powerful methodological support for entity relationship extraction and geological knowledge graph construction in the geoscience domain.

Keywords: Geological named entity recognition; ALBERT pre-trained models; knowledge graph; regional geological survey

Acknowledgements: The work is supported by the National Key R & D Program of China (No. 2022YFF0711601), National Natural Science Foundation of China (No. 42050101), and China Postdoctoral Science Foundation (No. 2021M702991)

First author: QIU Qinjun, male, born in 1988, associate professor, mainly engaged in the research of geological text mining and knowledge mining; Email: qiuqinjun@cug.edu.cn

Corresponding author: MA Kai, male, born in 1980, associate professor, mainly engaged in the research of geological text mining and knowledge mining; Email: makai@ctug.edu.cn

Manuscript received on: 2022-11-23; Accepted on: 2023-01-10; Published online on: 2023-01-20

Doi: 10.16509/j.georeview.2023.01.085　　**Edited by**: ZHANG Yuxu