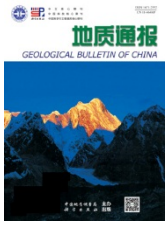

地质通报
Geological Bulletin of China
ISSN 1671-2552,CN 11-4648/P

《地质通报》网络首发论文

题目：结合 BERT 与 BiGRU-Attention-CRF 模型的地质命名实体识别

作者：谢雪景，谢忠，马凯，陈建国，邱芹军，李虎，潘声勇，陶留锋

网络首发日期：2021-09-13

引用格式：谢雪景，谢忠，马凯，陈建国，邱芹军，李虎，潘声勇，陶留锋．结合 BERT 与 BiGRU-Attention-CRF 模型的地质命名实体识别．地质通报.
https://kns.cnki.net/kcms/detail/11.4648.p.20210913.1040.002.html

网络首发时间：2021-09-13 13:35:08
网络首发地址：https://kns.cnki.net/kcms/detail/11.4648.p.20210913.1040.002.html

结合 BERT 与 BiGRU-Attention-CRF 模型的地质命名实体识别

谢雪景[1]，谢忠[1,2]，马凯[3]，陈建国[4]，邱芹军[1,2,*]，李虎[5]，潘声勇[6]，陶留锋[1,2]
XIE Xuejing[1], XIE Zhong[1,2], MA Kai[3], CHEN Jianguo[4], QIU Qinjun[1,2], LI Hu[5], PAN Shengyong[6] and TAO Liufeng[1,2]

1. 国家地理信息系统工程技术研究中心，湖北 武汉 430074；
2. 中国地质大学（武汉）地理与信息工程学院，湖北 武汉 430074；
3. 三峡大学计算机与信息学院，湖北 宜昌 443002；
4. 中国地质大学（武汉）资源学院，湖北 武汉 430074；
5. 济南轨道交通集团有限公司，山东 济南 250000；
6. 武汉中地数码科技有限公司，湖北 武汉 430074；

1. National Engineering Research Center of Geographic Information System, Wuhan 430074, China;
2. School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China;
3. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China;
4. Faculty of Earth Resources, China University of Geosciences, Wuhan 430074, China;
5. Jinan Rail Transit Group Co., Ltd, Jinan 250000;
6. Wuhan Zondy Cyber Science & Technology Co., Ltd. Wuhan 430074, China;

摘要：从地质文本中提取地质命名实体对地质大数据的深度挖掘与应用具有重要意义。本文定义了地质命名实体的概念并制订了标注规范，设计了地质实体对象化表达模型。地质文本存在大量长实体、复杂嵌套实体，这都增加了地质命名实体识别任务的挑战性。针对上述问题，(1) 引入 BERT 模型生成顾及上下文信息的高质量词向量表征；(2) 采用双向门控循环单元-注意力机制-条件随机场（BiGRU-Attention-CRF）对前一层输出的语义编码进行序列标注与解码。通过与主流深度学习模型进行对比，该模型的 F1 值为 84.02%，均比其他模型表现出更为优异的性能，能在小规模地质语料库上有较好的识别效果。

关键词： 命名实体识别；地质命名实体；BERT；注意力机制；BiGRU

Xie X J, Xie Z, Ma K, Chen J G, Qiu Q J, Li H, Pan S Y, Tao L F. Geological Named Entity Recognition based on BERT and BiGRU-Attention-CRF Model

Abstract： Extracting geological named entities from geological texts is of great significance for information extraction from geological texts. In this paper, the concept of geological named entities is defined and annotation specification is developed, and the objectified representation model of geological entities is designed. The presence of a large number of long entities and complex nested entities in geological texts and the small size of the available corpus increase the challenge of the geologically named entity recognition task. To address these problems, (1) the BERT model is introduced to generate high-quality word vector representations that take into account contextual information;(2) BiGRU-Attention-Conditional Random Field (BiGRU-Attention-CRF) is used to sequentially annotate and decode the semantic encoding output from the previous layer. By comparing

基金项目：国家自然科学基金《地球科学知识图谱表示模式与群智协同构建》（批准号：42050101）、国家自然科学基金《基于多模态数据理解及融合的三维地质模型构建方法研究》（批准号：41871311）、国家自然科学基金《城市地质环境时空透视与大数据融合关键技术》（批准号：U1711267）和山东省重大科技创新工程项目（批准号：2019JZZY020105）
第一作者简介：谢雪景（1991-），女，博士研究生，主要从事地质大数据挖掘与信息抽取研究。E-mail：xiexuejing@cug.edu.cn
通讯作者：邱芹军（1988-），男，副研究员，主要从事地质大数据挖掘与知识图谱研究。E-mail：qiuqinjun@cug.edu.cn

with the mainstream deep learning models, the F1 value of this model is 84.02%, both of which show better performance than other models and can have better recognition effect on small-scale geological corpus.

Key words： Named entity recognition; Geological named entities; Bidirectional Encoder Representations for Transformers; Attention mechanism; Bi-Gated Recurrent Unit

2019 年 2 月由国际学术组织共同发起的深时数字地球大科学计划（DDE）提出在地质大数据和人工智能技术共同驱动下重构地球科学知识结构体系，推动地球知识发现与创新[1]。大数据研究已成为一种数据密集型科学范式的代表[2]，通过将大数据及人工智能相关技术方法应用到地质领域，可为地质大数据信息深度挖掘及知识服务提供解决途径。长期积累的庞大地质资料集不仅包括结构化的地质数据，还包括文本、图表等半结构化或非结构化的数据[3]。地质报告作为地质调查工作成果的重要载体，其涵盖了大量有意义的信息及丰富知识，迫切需要对其进行深度挖掘与知识抽取[4–6]。地质报告文本中记录了某些区域范围内的地质环境及其事件，包含了各类地质命名实体[4]。地质命名实体作为地质文本中的核心要素，地质属性与关系的描述均以实体为基础，是实现地质信息抽取和挖掘、构建地学知识图谱的重要前提[7]。文本挖掘技术能够突破地质数据量及空间认知模式等多方面的限制，对地质资料中隐含的语义、时空及其他相关关系等地学信息进行有效抽取[8]。

命名实体识别(named entity recognition, NER)作为自然语言处理中的一项基础任务，对信息抽取、知识图谱构建等起着关键作用。通用领域与一些特定领域命名实体识别的研究已经得到了广泛而成熟的应用[9–11]。目前命名实体识别的方法主要包括：基于字典和规则的方法[12]、基于统计学的机器学习方法[13–16]以及深度学习方法[5,17–20]。基于字典和规则方法需要根据短语搭配模式及词汇特征人工构建实体抽取规则，虽然能在特定领域取得较好的效果，但需要大量专家知识且召回率低；基于统计机器学习方法有隐马尔可夫模型[16]、支持向量机[14]、条件随机场[15]、最大熵模型[13]等。主要依据标注好的训练集定义特征集，应用传统机器学习算法训练统计模型，其识别性能与设计的特征密切相关；基于深度学习方法近年来得到广泛的应用和突破性进展，包括循环神经网络模型（RNN）[9]、卷积神经网络（CNN）[18]、门控神经网络（GRU）[21]等。深度学习方法与机器学习模型相比，能学习到高维度与深层次的特征表示，有利于提高实体识别的泛化能力。

地质命名实体识别属于特定领域的命名实体识别，旨在识别地质中的一些重要概念，包括地质年代、地质构造、地层、岩石、矿物和地点等[5]。目前已有一些学者对地质命名实体识别进行研究。张雪英等[7] 根据地质文本特征制定了地质实体信息的要素分类体系及标注规范，并将深度信念网络模型应用到地质实体信息识别中。马凯[22]结合地质领域本体对地质文本经过分词、去停用词预处理操作后，使用成熟的 BiLSTM-CRF 模型开展命名实体识别任务。Qiu 等[4]参考 word2vec 模型用大量未标注的数据训练地质领域的词嵌入，并使用基于注意机制的 BiLSTM-CRF 模型方法进行句子的语义编码与地质实体识别。储德平等[6]融合 ELMO（Embeddings from Language Models）、CNN、Bi-LSTM-CRF 多种方法提取地质实体，经过字向量化表达后使用 CNN 添加字符特征、ELMO 提取词动态特征来获取输入分布式表示。以上研究的共性是利用深度学习模型能学习到词间深层次非线性特征这一优势来实施地质命名实体识别任务，取得了较好的效果。

但地质命名实体的识别仍面临着一些困难与挑战，地质文本相比于通用领域文本，地质命名实体存在（1）字符长度大；（2）生僻词多；（3）命名实体间相互嵌套等情况。如“石榴子石透辉石矽卡岩”表示一种主要矿物成分为石榴子石和透辉石的矽卡岩，该地质命名实体包含 10 个字符，“石榴子石透辉石矽卡岩”实体中又包括石榴子石、透辉石、矽卡岩、石榴子石透辉石矽卡岩四个地质命名实体，其中包含岩石和矿物两种不同类型的实体。不

同概念层级的地质命名实体之间存在着嵌套关系，具有复杂的层次化结构。因此地质命名实体识别成为一项具有挑战性的任务。

为解决上述问题，提出地质实体的对象化表达模型，制定了地质实体标注规范，并融入 BERT 模型结合基础模型 BiGRU-Attention-CRF 进行地质命名实体识别。BERT 在自行标注的语料库上学习地质文本的语义语法特征，并在 BiGRU-CRF 基础模型中引入注意力机制用于序列标注与解码，能充分考虑与当前实体相关的局部信息，有助于提升地质命名实体识别性能。

1 地质命名实体、对象概念定义及实体标注规范

1.1 概念定义

地质文本数据中蕴含着复杂的地质语义信息，地质调查研究人员可以从中提取感兴趣的地质信息，获取对地质对象的认知。正确地认识地质体是地质空间认知中的核心问题。下面对地质体、地质命名实体及地质对象进行明确的定义。

a. 地质体：指包含矿物、岩石、地层、地质构造等客观存在的泛指任何体积的天然岩石体。
b. 地质命名实体：地质文本中的核心要素和基本信息单位，是描述时空属性及其他附属属性、地质实体关系的基础。
c. 地质对象：指包含地质命名实体、地质命名实体的属性以及地质实体之间关系的统称，代表的是聚合地质实体、时空及关系等特征信息的对象，是对地质文本的实体、属性及时空信息的整体性描述与提取。

地质体与地质命名实体是一个事物的两个方面，地质体是现实世界客观存在的事物，而地质命名实体是在地质文本中所表达的基本信息单元。地质对象则包含了地质命名实体、时空附属属性、实体相互关系的表达，是比“地质命名实体”更为广义的概念。本文中的目标任务主要是对地质命名实体进行识别。

1.2 文本中地质实体对象化表达

对于地质文本中蕴含的语义信息，在表达尺度上可用地质命名实体、地质对象、组团地质对象三种承载方式为核心进行描述，而在空间尺度结构上包含微结构（地质命名实体之间的关系）、对象结构（地质对象之间的关系）与格局结构（组团地质对象之间的关系）等层次。地质实体信息对象化表达是指通过以某一地质实体为核心，将其特定属性、实体关系、空间、时间等相关内容进行关联。该地质命名实体与这些相关内容整体抽象化成一个地质对象，从而实现相互统一的抽象化模型，如图 1 所示。

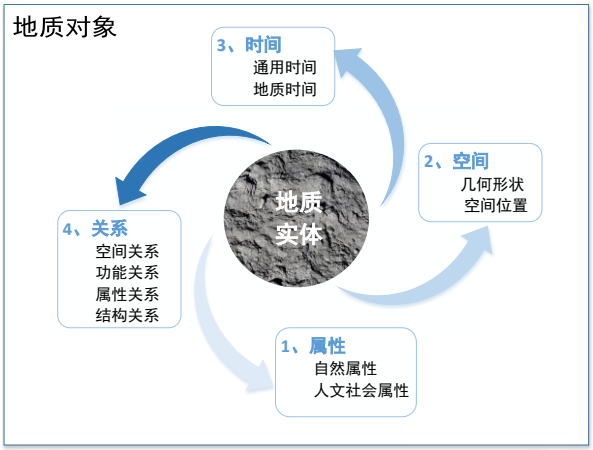

图 1 地质实体对象化表达

Fig. 1 Object oriented expression of geological entity

表 1 按地质实体标识与地质实体要素对地质对象进行详细划分。通过将地质实体与地质实体的空间、属性、时间、关系等实体要素相结合，共同构成一个地质对象，实现基于地质实体的对象化表达。

表 1 地质实体对象化信息

Table 1 Object information of geological entity

大类	二级类	三级类	四级类	五级类
地质对象	地质实体标识			采用名称、编码或符号进行标识
	地质实体相关要素	属性	自然要素	物理
				化学
				生物
			人文要素	社会
				经济
				人文
		空间	几何形状	点
				线
				面
				体
			空间位置	绝对位置
				相对位置
		时间	地质时间	按地质年代区分
			通用时间	按纪年法区分
		关系	空间	拓扑
				距离
				方向
				顺序
			功能	矿物伴生/共生关系
				地层接触关系
			属性	等级关系
				相关关系
				…
			结构	组成关系
				…

1.3 实体标注规范

地质命名实体识别研究属于序列标注问题。序列标注的目标是对给定序列中每个元素标注恰当标签。通常情况下，一个序列指一个句子，而一个元素指句子中的一个字或一个词语。本文结合多种地质命名实体类型，采用 BIOES 标注法，并去除英文单词、特殊符号图表等其他无关信息。标签类型的定义如表 2 所示。

表 2 标签类型定义

Table 2 Label type definition

定义	全称	备注
B	Begin	实体片段的开始

I	Inside	实体片段的中间
E	End	实体片段的结束
S	Single	单个字的实体
O	Other	其他不属于任何实体的字符（包括标点等）

另外，本文将地质命名实体分为六类，分别为地质年代、地质构造、地层、岩石、矿物与地点，其对应的标注为：GTM、GST、STR、ROC、MIN、PLA。

2 BERT-BiGRU-Attention-CRF 模型

本模型框架由 BERT 预训练语言模型、BiGRU 网络、注意力机制与 CRF 层组成，模型架构图如图 2 所示。首先将输入序列输入 BERT 层进行预训练获得上下文相关的表征，用于解决生僻字多、实体嵌套的实体识别关键问题；然后将 BERT 层获取的向量输入 BiGRU 层与 Attention 注意力机制层解决文本长期记忆和长文本依赖问题，可用于解决地质实体字符长度大的关键问题；最后通过 CRF 层进行解码，获得输出标签序列。

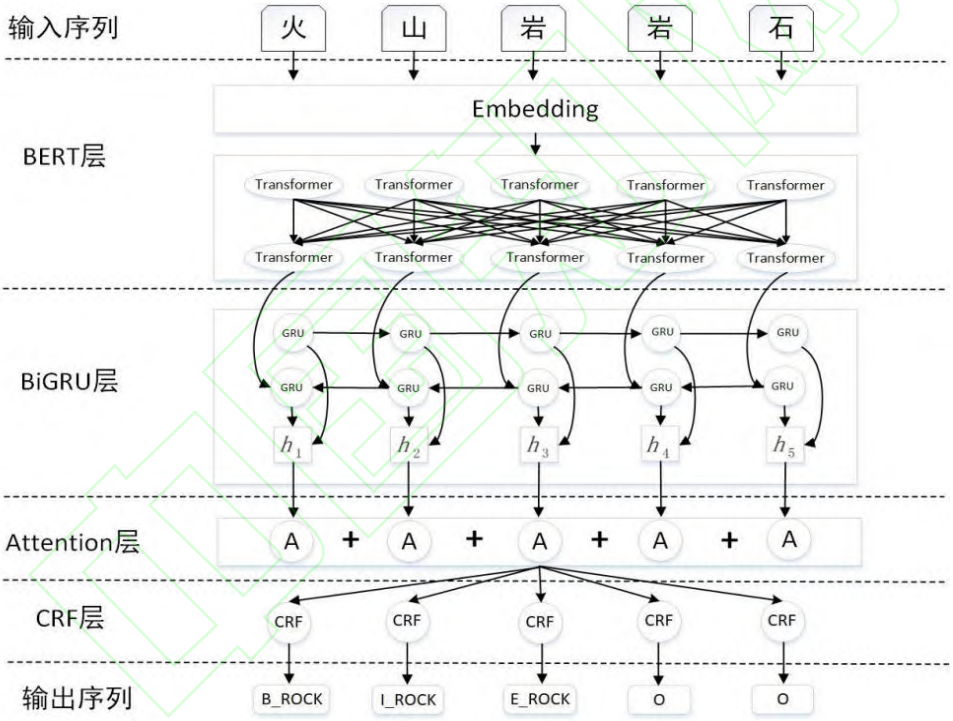

图 2 模型流程图

Fig. 2 Model flow chart

2.1 BERT 预训练模型

本文针对传统语言模型无法解决地质命名实体存在的生僻词多、实体嵌套、语义复杂等问题，特引入 BERT[23]预训练语言模型。相对于 ELMo[24]与 OpenAI-GPT[25]两种预训练模型， BERT 能同时从前后两个方向上提取上下文信息获得词向量的表示。BERT 较前两者所做的改进在于：

1)掩码语言模型（MLM, Masked Language Model）。为获得一个上下文相关的双向特征表示，BERT 在预训练阶段随机屏蔽掉 15%的标记，并根据上下文预测这些标记，可以更好地根据全文理解单词的语义。并以一定的概率保留单词的语义信息展示给模型，使信息

不至于完全被遮掩。因此，若出现生僻词，可以根据上下文进行预测，从而有效解决地质命名实体生僻词多、语义复杂的问题。

2)下一句预测模型（Next Sentence Prediction）。对于实体嵌套的问题，同一个字符可能同时具有两个或以上的标签。问题较为复杂，因此传统的基于字符级别的序列标注任务 Flat NER 由于嵌套实体问题变成了 Nested NER。解决该问题的方法有多种，其中基于阅读理解的方法[26]近年来展现出较好的效果，它为 Flat NER 与 Nested NER 提供统一的处理框架。阅读理解的任务就是查询句子中是否存在指定问题的答案，实体通过在给定上下文中回答问题来提取。由于 BERT 具有“下一句预测”模型，它允许在模型中输入同时两个不同的句子，擅长处理句子的匹配任务，可作为基础模型完成阅读理解的任务。本文嵌套实体最终输出的结果为最外层的实体，因此在输出序列中每个字符只有一个标签。

为了明确地表达地质文本中的一个句子，BERT 的输入为字符级 Embedding 的序列，对于每一个字符，其表征由字符 embedding、句子级 embedding、以及位置 embedding 求和获得。其中句子级 embedding 对应句子的唯一的向量表示。位置 embedding 表示字符在句中的位置，字符或词语在句中的位置不同可能会导致完全不同的语义。

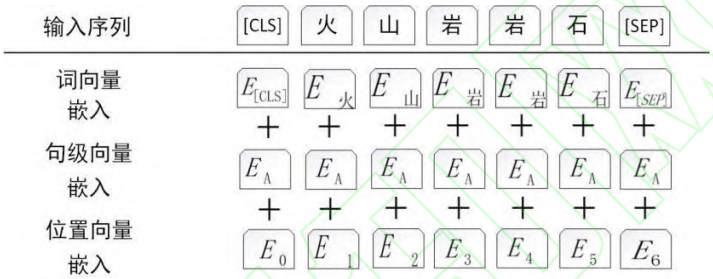

图 3 BERT 预训练模型词向量组成

Fig. 3 Word vector composition of BERT pre-training model

2.2 BiGRU 网络结构

门控循环单元（Gated Recurrent Unit, GRU）[21]与 LSTM 一样，是为了解决 RNN 长期记忆与反向传播梯度消失的问题所设计。其性能效果与 LSTM 类似，优势体现在参数少、硬件和时间成本较低，在小样本数据集上泛化能力效果较好。GRU 内部结构如图 4 所示。

GRU 结合当前的节点输入x^t与上个节点传输下来的状态h^{t-1}得出当前节点的输出y^t和传递给下个节点的隐状态h^t。网络内部参数传递与更新公式如（1）-（4）所示。

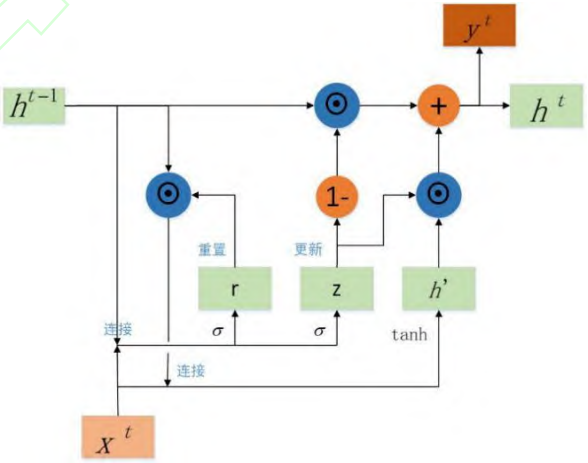

图 4 GRU 内部结构图

Fig. 4 Internal structure of GRU

$$r^t = \sigma(w^{rx}x^t + w^{rh}h^{t-1} + b^r) \quad (1)$$

$$z^t = \sigma(w^{zx}x^t + w^{zh}h^{t-1} + b^z) \quad (2)$$

$$h' = \tanh(w^{xh}x^t + r^t \odot w^{hh}h^{t-1}) \quad (3)$$

$$h^t = (1 - z^t) \odot h^{t-1} + h' \odot z^t \quad (4)$$

其中σ为$sigmoid$ 函数，该函数用来充当门控信号，可将数值控制在[0,1]范围内。门控信号越接近 1，表示记忆下来的数据越多。反之，遗忘的越多。r为控制重置的门控，z为控制更新的门控。h' 指候选隐藏状态。w^{rx}、w^{rh}等为权重矩阵，b^r、b^z等为偏置量。$\odot$是 Hadamard 积，即将矩阵中对应的元素相乘。

重置门控得到重置之后的数据$r^t \odot h^{t-1}$，再将该值与x^t进行拼接，经过一个$tanh$激活函数将输出值控制在[-1,1]，得到隐藏状态h' 。更新门控同时进行遗忘和选择记忆操作，其中$(1 - z^t) \odot h^{t-1}$对之前节点的状态进行选择性遗忘，$h' \odot z^t$对隐藏状态进行选择性记忆。

双向循环网络整体在序列标注任务中一直优于前馈循环网络，从 GRU 单元中只能获得上文的信息，不能获得未来的信息，因此，本文使用双向 GRU 即 BiGRU 获得上下文信息。

2.3 注意力机制层

GRU 可以在一定程度上解决长期记忆的问题，提取全局特征。但难以解决地质文本中长距离依赖问题，在长文本中难以保留局部细节信息。为了弥补 BiGRU 提取局部特征所存在的缺陷，本文引入 Attention 机制[27]提取句子中不同的字符与上下文的关联程度，有利于解决地质命名实体字符长度大导致的长距离依赖问题。Attention 机制对与地质命名实体相关的语义增加特征权重，提升局部特征提取效果。

注意力机制层对 BiGRU 层输出的特征向量h^t 进行权重分配，计算得到t时刻 BiGRU 层和注意力层的共同输出特征向量c_t。

$$c^t = \sum_{i=1}^{n} a^{t,i} h^i \quad (5)$$

$$a^{t,i} = \frac{\exp(\text{score}(s^{t-1}, h^i))}{\sum_{i'=1}^{n} \exp(\text{score}(s^{t-1}, h^{i'}))} \quad (6)$$

$$score(s^t, h^i) = v \tanh(w[s^t, h^i]) \quad (7)$$

其中$a^{t,i}$为注意力函数。$score$函数为对齐模型，它基于t时刻的输入和输出的匹配程度分配分数，定义每个输出给每个输入隐藏状态多大的权重。

2.4 CRF 层

CRF [15]常用于序列标注任务中。它可以在 BiGRU-Attention 的基础上加入一些约束，确保输出标签之间的顺序正确。因此，CRF 层作为最终的输出解码层，用于获取地质命名实体预测标签序列。

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$P(Y/X)$来描述CRF模型。

对于一句文本，$X=\{x_1,x_2,...,x_n\}$ 表示其观测序列，对于输出标签序列$Y=\{y_1,y_2,...,y_n\}$，其分数为：

$$sore(\mathrm{X},\mathrm{y})=\sum_{1}^{m}Q_{i,\mathrm{y}_i}+\sum_{0}^{m}A_{\mathrm{y}_i,\mathrm{y}_{i+1}} \tag{8}$$

其中Q为注意力机制输出的分数矩阵$m*k$，其中m为句子的长度，k为实体类型不同的标签数量。$Q_{i,j}$表示第 i 个词第 j 个标签的分数。A 是一个大小为$k+2$的转移分数矩阵，其中$A_{y_i,y_{i+1}}$表示由标签i转移到标签$i+1$的分数。

Y_X是对于句子X所有可能的标注序列，最终解码时通过维特比(Viterbi)算法得到得分最大的预测标签序列。

$$y^*=\arg\max(\mathrm{score}(\mathrm{X},\mathrm{y})) \tag{9}$$

3 实验结果与分析

3.1 数据集

本文数据来源于中国地质调查局全国地质资料馆网站（NGAC），以矿产资源地质调查报告为主，共计 12 万余字。对收集的地质报告采用人工标注的方式将文本中六种地质命名实体类别 GTM、GST、STR、ROC、MIN 与 PLA 进行标注，数据标注示例如表 3 所示。实验共标注了 15891 个地质命名实体，标注完成后每类实体在训练集、开发集、测试集的数量比例约为 3:1:1，三者的数量分布如图 5 所示。

表 3 数据标注示例

Table 3 Data set annotation examples

语料	英	安	岩	亚	类	出	露	于	去
标注	B-ROC	I-ROC	E-ROC	O	O	O	O	O	B-STR
语料	申	拉	组	的	上	部	层	位	，
标注	I-STR	I-STR	E-STR	O	O	O	O	O	O
语料	主	要	分	布	于	重	昌	、	玉
标注	O	O	O	O	O	B-PLA	E-PLA	O	B-PLA
语料	雄	等	地	。					
标注	E-PLA	O	O	O					

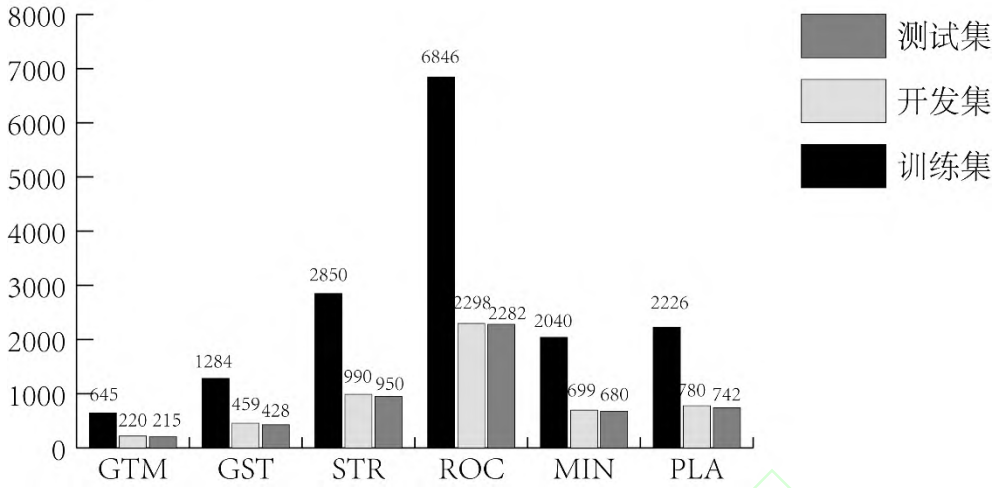

图 5 实体类别数量分布

Fig. 5 Distribution of entity categories number

3.2 实验环境和参数设置

本文采用 Python3.8 + keras 2.3 的实验环境进行模型的训练与测试。实验中引入 BERT-Base 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。另外，GRU 网络层、Attention 机制层等其他参数设置如表 4 所示。

表 4 模型参数设置

Table 4 Model parameter setting

GRU 隐层维数	Attention 隐层维数	最大序列长度	优化函数学习率（Adam）	Dropout
128	50	128	5×10^{-5}	0.2

3.3 评估标准

实验利用精确率（P）、召回率（R）与 F1 值共 3 个评价指标评价 6 大类地质命名实体的命名实体识别效果，3 个评价指标的计算方法如下：

$$P=\frac{TP}{TP+FP} \tag{10}$$

$$R=\frac{TP}{TP+FN} \tag{11}$$

$$F1=\frac{2PR}{P+R} \tag{12}$$

其中TP指样本为正，预测结果为正；FP指样本为负，预测结果为正；FN指样本为正，预测结果负。

3.4 实验结果与分析

图6详细描述了地质年代、地质构造、地层、岩石、矿物、地点六类不同实体的在精确率、召回率、F1值的结果。各类实体识别结果中地名实体识别效果最好，而地质构造和地层的准确率总体上较其他实体的准确率低。这可能是因为地名在较长时间范围内变化的可能性小。另外，地质报告具有区域固定的特点，出现的地名个数较少且频率较高，易于地名识

别与抽取。而地质构造和地层实体专业程度高，存在未登录词占有较大比重的情况，在进行实体识别时较为复杂。

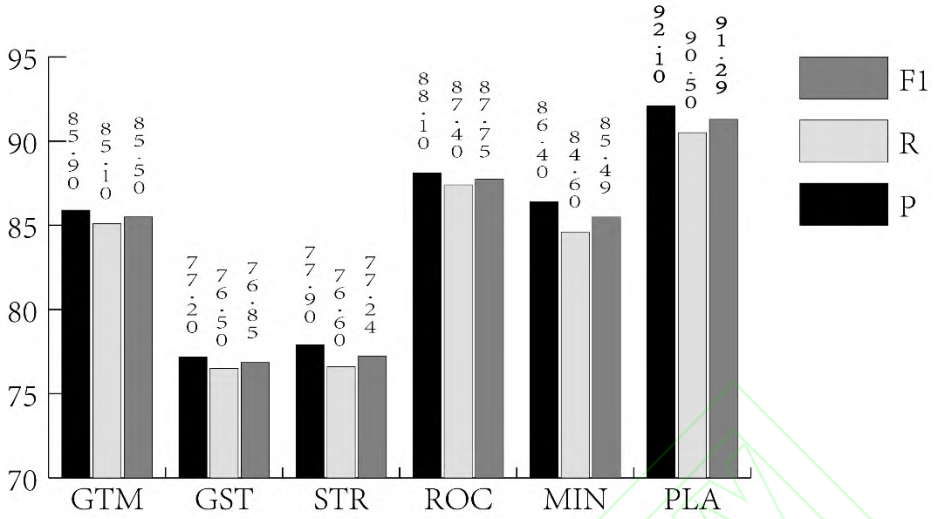

图6 不同实体识别效果

Fig. 6 Different entity recognition effect

为了证明本文框架的性能优越性，使用相同数据集对命名实体识别领域几种主流算法模型的地质命名实体识别效果进行对比实验验证，6 种模型的性能如表 5 所示。

表 5 不同模型识别效果

Table 5 Different model recognition effect

实验	模型	P	R	F1
1	CRF	71.50	69.90	70.69
2	BiLSTM-CRF	73.70	71.50	72.58
3	BiGRU-Attention-CRF	79.10	78.50	78.80
4	ELMO-CNN-BiLSTM-CRF	82.30	80.70	81.49
5	BERT-BiGRU-CRF	83.80	81.20	82.48
6	BERT-BiGRU-Attention-CRF	**84.60**	**83.45**	**84.02**

由表 5 可以看出，本文提出的模型在 P、R、F1 的效果都优于其他模型，具体表现在：

（1)本文模型相对于传统的基于统计的方法 CRF 有比较大的优势，分别提升了 13.1%、13.55%和 13.33%。CRF 在分词基础上通过设置特征模板获取特征，因此对未登录词的实体识别效果较差。而本文模型利用 BERT 结合上下文语境自动提取语料库特征的优势，能有效识别未登录词。

（2）实验 2 相对于实验 1 增加了 BiLSTM 网络，F1 值提高了 1.89%。证明考虑文本前后向信息能提升总体识别效果，如更容易区分“辉长岩”和“辉绿辉长岩”这类容易混淆的词，而 BiGRU 与 BiLSTM 性能类似。

（3）本文模型相对于实验 4，F1 值提高了 2.53%。证明相对于地质命名实体识别领域已有的方法 ELMO-CNN-BiLSTM-CRF 模型在性能上得到一定的提升。

（4）本文模型相对于实验 5，F1 值提高了 1.54%。证明增加注意力机制可以获得更多局部特征，尤其是与当前输出关联程度较高的信息。如识别句子“蛇绿岩各单元之间及其与周围地质体多为断裂构造接触”中的“蛇绿岩”实体，词语“断裂构造”比“多为”的关联程度更大，所以注意力机制更关注“断裂构造”。

（5）本文模型相对于实验 3，F1 值提高了 5.22%，性能得到显著提升。证明 BERT 能

学习到字级、词级、句级特征，可以更为全面了解句子语义。因此，融入 BERT 模型的地质命名实体识别方法通过基于上下文语境的深度双向语义理解，有效解决实体嵌套等复杂语义问题，从而有效提升地质命名实体识别精度。

从识别性能上，相比较于人民日报语料库来说，地学领域的命名实体有其独特的领域特点，比如地学中的命名实体长度一般比较长、出现的次数也比较少，比如“尼玛县岗龙乡麻勒果－嘎干拉”、“麻勒果”等；而人民日报标注语料库中主要包括人民、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、安徽省、合肥市等”。从语料库的规模上来说，人民日报语料库规模更大，数据集更多，而深度学习的训练是需要大量的数据集学习其中的数据特征从而才能够对新输入的文本进行识别。

表 6 部分识别结果示例

Table 6　Some examples of recognition results

原文信息	参考识别信息	识别信息
上白垩统红色磨拉石建造及新近系山间盆地复陆屑建造	上 B-STR 白 I-STR 垩 I-STR 统 E-STR 红 B-ROC 色 I-ROC 磨 I-ROC 拉 I-ROC 石 E-ROC 建 O 造 O 及 O 新 B-STR 近 I-STR 系 E-STR 山 O 间 O 盆 O 地 O 复 O 陆 O 屑 O 建 O 造 O	上 B-STR 白 I-STR 垩 I-STR 统 E-STR 红 B-ROC 色 I-ROC 磨 I-ROC 拉 I-ROC 石 E-ROC 建 O 造 O 及 O 新 B-STR 近 I-STR 系 E-STR 山 O 间 O 盆 O 地 O 复 O 陆 O 屑 O 建 O 造 O
三节泥盆系出露于尼玛县张恩-申扎县喀郭尔一带	三 O 节 O 泥 B-STR 盆 I-STR 系 E-STR 出 O 露 O 于 O 尼 B-PLA 玛 I-PLA 县 I-PLA 张 I-PLA 恩 I-PLA － I-PLA 申 I-PLA 扎 I-PLA 县 E-PLA 喀 I-PLA 郭 I-PLA 尔 E-PLA 一 O 带 O	三 O 节 O 泥 B-STR 盆 I-STR 系 E-STR 出 O 露 O 于 O 尼 B-PLA 玛 I-PLA 县 I-PLA 张 I-PLA 恩 I-PLA － I-PLA 申 I-PLA 扎 I-PLA 县 I-PLA 喀 I-PLA 郭 I-PLA 尔 E-PLA 一 O 带 O
以拉惹－康如断裂为界进一步划分	以 O 拉 B-GST 惹 I-GST － I-GST 康 I-GST 如 I-GST 断 I-GST 裂 E-GST 为 O 界 O 进 O 一 O 步 O 划 O 分 O	以 O 拉 O 惹 O － O 康 B-GST 如 I-GST 断 I-GST 裂 E-GST 为 O 界 O 进 O 一 O 步 O 划 O 分 O
原岩为碎屑岩－中基性火山岩	原 O 岩 O 为 O 碎 B-ROC 屑 I-ROC 岩 E-ROC － O 中 B-ROC 基 I-ROC 性 I-ROC 火 I-ROC 山 I-ROC 岩 E-ROC	原 O 岩 O 为 O 碎 B-ROC 屑 I-ROC 岩 E-ROC － O 中 O 基 O 性 O 火 I-ROC 山 I-ROC 岩 E-ROC

本文部分实体识别结果如表 6 所示，可以观察到，地层单位实体“上白垩统”、“新近系”和岩石实体“红色磨拉石”等基础的地质命名实体能被有效识别；另外本模型也能精准识别出长实体与生僻的地名词“尼玛县张恩-申扎县喀郭尔”，该实体是由尼玛县、张恩、申扎县、喀郭尔多个独立词组成的嵌套实体，对于这类嵌套实体，本文的做法是识别最外层实体，选择性忽略内层实体，直接将尼玛县张恩-申扎县喀郭尔作为地质命名实体识别结果输出。

仍有部分信息识别存在一定的问题，（1）对部分仅用符号分隔的连续出现的实体字符

无法进行精确识别。如“拉惹-康如断裂”实体中未能将“拉惹”作为地质构造的字符识别出来，而是将“-”字符后的“康如”作为该地质构造实体的开始字符。（2）模型只对局部信息进行了识别。如“中基性火山岩”岩石实体也只识别出“火山岩”，将“中基性”标注为其他字符。出现上述问题的原因在于语料库规模小，训练集中标注数据覆盖度不够，导致有些知识在训练过程中未被学习到，从而影响识别效果。对于这些问题，可以通过增加地质文本语料库规模，完善地质实体标注信息得到有效解决。

4 结 语

本文提出融入BERT的地质命名实体识别方法，首先引入BERT模型获得上下文相关的双向特征表示，更为全面地了解句子语义，有效解决了地质文本语料库规模小、生僻词多、实体嵌套且语义复杂等问题；然后引入注意力机制改进BERT-BiGRU-CRF模型，识别过程中关注关联程度更高的信息，有利于长文本实体的识别。实验结果表明，本文设计的BERT-BiGRU-Attention-CRF模型能有效识别地质命名实体，F1值达到84.02%。本文提出的地质命名实体识别方法对地质命名实体的属性、空间、关系等其他相关信息的抽取具有重要作用，有利于最终实现地学知识图谱构建。

参考文献

[1] 成秋明. 深时数字地球:全球古地理重建与深时大数据[J]. 国际学术动态, 2019(6): 28–29.

[2] 赵鹏大. 地质大数据特点及其合理开发利用[J]. 地学前缘, 2019, 26(4): 1–5.

[3] 王成彬,马小刚,陈建国. 数据预处理技术在地学大数据中应用[J]. 岩石学报, 2018, 34(2): 303–313.

[4] QIU Q, XIE Z, WU L, et al. BiLSTM-CRF for Geological Named Entity Recognition from the Geoscience Literature[J]. Earth Science Informatics, 2019, 12(4): 565–579. DOI:10.1007/s12145-019-00390-3.

[5] 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 Science, 2019, 6(6): 931–946. DOI:10.1029/2019EA000610.

[6] 储德平,万波,李红,等. 基于 ELMO-CNN-BiLSTM-CRF 模型的地质实体识别[J/OL]. 地球科学, 2020. https://kns.cnki.net/kcms/detail/42.1874.P.20201109. 1600.008.html. DOI:10.3799/dqkx.2020.309.

[7] 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.

[8] QIU Q, XIE Z, WU L, et al. Automatic Spatiotemporal and Semantic Information Extraction from Unstructured Geoscience Reports Using Text Mining Techniques[J]. Earth Science Informatics, 2020, 13(4): 1393–1410. DOI:10.1007/s12145-020-00527-9.

[9] 高学攀,杜楚,吴金亮. 基于 BiLSTM-CRF 的军事命名实体识别方法[J]. 无线电工程, 2020, 50(12): 1050–1054.

[10] 张靖宜,贺光辉,代洲,等. 融入 BERT 的企业年报命名实体识别方法[J/OL]. 上海交通大学学报, 2020. https://kns.cnki.net/kcms/detail/31.1466.U.20201016.1557.001.html. DOI:10.16183/j.cnki.2020.009.

[11] 钟原,刘小溶,王杰,等. 基于 NER 的石油非结构化信息抽取研究[J]. 西南石油大学学报(自然科学版), 2020, 42(6): 1–9. DOI:10.11885/j.issn.1674 5086.2020.05.12.01.

[12] MIKHEEV A, MOENS M, GROVER C. Named Entity Recognition without Gazetteers[C/OL]//Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics -. Bergen, Norway: Association for Computational Linguistics, 1999. http://portal.acm.org/citation.cfm?doid=977035. 977037. DOI:10.3115/977035.977037.

[13] BERGERT A L. A Maximum Entropy Approach to Natural Language Processing[J]. Computational

Linguistics, 1996, 22(1): 38–71. DOI:doi:10.1016/0169-7439(95)00072-0.

[14] ZHANG Z. Weakly-Supervised Relation Classification for Information Extraction[C]//Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management - CIKM '04. Washington, D.C., USA: ACM Press, 2004: 581. DOI:10.1145/1031171.1031279.

[15] SARAWAGI S, COHEN W W. Semi-Markov Conditional Random Fields for Information Extraction[J]. NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems, 2004: 1185–1192.

[16] LISON P, HUBIN A, BARNES J, et al. Named Entity Recognition without Labelled Data: A Weak Supervision Approach[J/OL]. ArXiv:2004.14723, 2020[2020–12–12]. http://arxiv.org/abs/2004.14723.

[17] JI J, CHEN B, JIANG H. Fully-Connected LSTM–CRF on Medical Concept Extraction[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(9): 1971–1979. DOI:10.1007/s13042-020-01087-6.

[18] 贺琳,张雨,巴韩飞. 基于 CNN-BiGRU-CRF 模型的外来海洋生物实体识别[J/OL]. 大连海洋大学学报, 2020. https://doi.org/10.16535/j.cnki.dlhyxb.2020-194. DOI:10.16535/j.cnki.dlhyxb.2020-194.

[19] 宋建炜,邓逸川,苏成. 基于预训练语言模型的建筑施工安全事故文本的命名实体识别研究[J/OL]. 图学学报, 2020. https://kns.cnki.net/kcms/detail/10. 1034.T.20201118.1817.052.html.

[20]陈剑,何涛,闻英友,等. 基于 BERT 模型的司法文书实体识别方法[J]. 东北大学学报(自然科学版), 2020, 41(10): 1382–1387.

[21] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation[J/OL]. ArXiv:1406.1078, 2014. http://arxiv.org/abs/1406.1078.

[22] 马凯. 地质大数据表示与关联关键技术研究[D]. 武汉: 中国地质大学（武汉）, 2018.

[23] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[J/OL]. ArXiv:1810.04805, 2019. http://arxiv.org/abs/1810.04805.

[24] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations[J]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018: 2227–2237.

[25] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving Language Understanding by Generative Pre-Training[EB/OL](2018). https://s3-us-west2. amazonawa.com/openai-assets/research-covers/language-unsupervised/language _understanding_ paper.pdf.

[26] LI X Y, FENG J R, MENG Y X, et al. A Unified MRC Framework for Named Entity Recognition[J/OL]. ArXiv:1910.11476, 2020. http://arxiv.org/abs/1910.11476.

[27] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J/OL]. ArXiv:1409.0473, 2016. http://arxiv.org/abs/1409.0473.