第 55 卷 第 4 期
2022 年(总 226 期)

西 北 地 质
NORTHWESTERN GEOLOGY

Vol. 55 No. 4
2022(Sum226)

DOI:10.19751/j.cnki.61-1149/p.2022.04.011

基于 BERT 的三维地质建模约束信息抽取方法及意义

邱芹军[1,2],马凯[3,4],朱恒华[5],刘春华[5],谢忠[1,2],谭永健[3,4],陶留锋[1,2]*

(1. 中国地质大学(武汉)计算机学院,湖北 武汉 430074;2. 智能地学信息处理湖北省重点实验室,湖北 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. 山东省地质调查院,山东 济南 250000)

摘 要:地质报告中地质体的几何、拓扑及属性信息是三维地质建模过程中重要约束性信息。但传统的属性信息抽取方法存在覆盖率有限、局限于人工设计特征及模型泛化能力差等问题。面向三维建模任务,总结了地质报告中地质体的几何、拓扑及属性文本的特点,提出了一种基于 BERT-BiLSTM-CRF 的三维地质建模信息抽取方法;基于 BERT 预训练模型,构建融合 BiLSTM 和 CRF 的深度学习模型,通过 BERT 模型获取动态字符深层次语义信息,弥补静态词向量无法解决一词多义的问题,提高地质体复杂建模信息的抽取能力。以 43 篇地质报告为数据源进行模型性能评估,实验结果表明所提出的方法对于地质体三类属性信息抽取准确率达到 90%以上,对于三维地质建模具有重要支撑作用。

关键词:三维地质建模;属性抽取;BERT;约束信息

中图分类号:P628+.3 **文献标志码**:A **文章编号**:1009-6248(2022)04-0124-09

BERT-based Method and Significance of Constraint Information Extraction for 3D Geological Modelling

QIU Qinjun[1,2], MA Kai[3,4], ZHU Henghua[5], LIU Chunhua[5], XIE Zhong[1,2], TAN Yongjian[3,4], TAO Liufeng[1,2,*]

(1. School of Computer Sciences, China University of Geosciences, Wuhan 430074, Hubei, China; 2. Hubei Key Laboratory of Intelligent Geo-Information Processing, Wuhan 430074, Hubei, China; 3. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, Hubei, China; 4. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, Hubei, China; 5. Shandong Institute of Geological Survey, Jinan 250000, Shandong, China)

Abstract: The geometry, topology and attribute information of geological bodies in geological reports are important constraint information in the 3D geological modeling process. However, the traditional attribute information extraction methods have problems such as limited coverage, lim-

收稿日期:2022-01-26;**修回日期**:2022-06-20;**网络发表日期**:2022-11-15;**责任编辑**:贾晓丹

基金项目:国家自然科学基金项目"地球科学知识图谱表示模式与群智协同构建"(42050101)、"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"(41871311)、济南城区四维地质环境可视化信息系统平台建设项目(2018GDCG01Z0301)、山东省重点研发计划(重大科技创新工程)项目"数字孪生城市四维可视化信息系统及其在济南城区的应用"(2019JZZY020105)、中国博士后科学基金(2021M702991)联合资助。

作者简介:邱芹军(1988-),男,博士,副研究员,主要研究领域为地质大数据挖掘与信息抽取研究。E-mail: qiuqinjun@cug.edu.cn。

* **通讯作者**:陶留锋(1984-),男,博士,副研究员,主要研究领域为地质大数据挖掘与信息抽取研究。E-mail: taoliufeng@cug.edu.cn。

ited to artificial design features and poor model generalization ability. Facing the 3D modeling task, the geometry, topology and attribute text characteristics of geological bodies in geological reports are summarized, and a 3D geological modeling information extraction method based on BERT-BiLSTM-CRF is proposed; based on the BERT pre-training model, a deep learning model integrating BiLSTM and CRF is constructed to obtain deep semantic information of dynamic characters through the BERT model to make up for the static word vector cannot solve the problem of multiple meanings of a word, and improve the extraction ability of complex modeling information of geological bodies. The model performance is evaluated with 43 geological reports as the data source, and the experimental results show that the proposed method has an accuracy rate of over 90% for extracting three types of attribute information of geological bodies, which is an important support for 3D geological modeling.

Keywords: 3D geological modelling; attribute extraction; BERT; constraint information

地质现象在本质上是三维的,通过三维地质模型可以开展流体模拟,进行成矿靶区预测,并帮助研究人员理解区域地质结构,更加准确地分析地质问题。在复杂地质体建模过程中,始终面临着数据稀缺的问题,因此模型构建时都尽量融合地质、物探、遥感、钻孔、化探等多源信息,以提升模型的精度和可靠性(王亚辉等,2019;魏东琦等,2021)。在地质报告中蕴含了丰富的地质建模约束性数据,对于地质模型构建有重要意义。陈麒玉等(2016)指出三维地质模型亟需能够顾及地质语义约束的结构-属性集成三维模型自动构建方法,并提出了基于知识驱动的三维模型构建方法(陈麒玉等,2020),Zhong 等(2019)提出将地质趋势线、约束面、约束线、各向异性等各种隐式约束规则应用在建模过程中,有效提升了建模效率。

三维地质建模中的基本要素包括几何、拓扑和属性 3 类,也是三维构造建模中的重要组成部分(李兆亮等,2016)。其中,几何要素反映地质要素本身的空间位置及几何形态,拓扑要素代表几何图形在连续变形(拓扑变换)下处于不变的性质,属性要素反映的是地质对象的基本特性。

表示三维结构特征的信息分为几何特征与空间关系特征。其中,几何特征包括:表示长度信息,如总长度…km;表示厚度信息,如厚度…km;表示宽度信息,如南北宽…km。

空间关系特征包括:表示确定空间位置的位置信息,例如经纬度坐标;表示邻近或者模糊的拓扑关系,例如位于…之上,位于…之间等;表示模糊空间位置的方位关系,例如在…以东等;表示确定的方位信息,例如总体方位 165°等;表示 2 个地质体的接触关系,如角度整合、不整合、与…整合接触、断层接触等。

表示属性信息特征的信息包括不限于密度、孔隙度、电阻率、矿化度等。

属性抽取是知识图谱构建中的一项基础环节,可为实体属性数据检索、知识图谱构建、实体对象理解等提供应用支持(Qiu et al.,2019a;Qiu et al.,2019b;Qiu et al.,2020)。

目前,面向文本数据的实体及实体关系抽取相关研究工作较多,但实体属性抽取相对较少。虽然有些方法可以直接将属性抽取作为实体关系的形式进行提取,但地质建模中相关的属性很多是对实体的直接描述,不宜将对应的属性值作为实体。

深度学习模型由于在深层结构上自动学习上下文特征具有显著性优势,为地质实体对象建模信息抽取提供了新途径。鉴于 Bi-directional Long Short-Term Memory (BiLSTM)-Conditional Random Field(CRF)在处理序列标注任务方面的优越性,在本文引入 Bidirectional Encoder Representation from Transformers(BERT)模型增强词向量模型泛化能力,提出一种基于 BERT-BiLSTM-CRF 的三维地质建模信息抽取模型。该模型通过非监督学习方式从大规模文本中抽取中文语句双向语义特征,更好表征不同语境中句法和语义信息,有效减少对人工特征和领域知识的依赖,端到端的实现了三维地质建模信息抽取任务。实验结果表明,笔者提出的模型在三维地质建模信息抽取任务中取得了较好的效果。

1 相关工作

在基于文本的信息抽取研究方面，传统的监督学习方法存在标注工作量大、时间代价高等缺点。苏丰龙等(2016)利用支持向量机的分类优势，以及直推式学习在未标注样本上的泛化特点，提出一种改进的半监督学习模型进行信息抽取。Etzioni 等(2005，2008)提出了面向网络信息的抽取系统 KnowItAll。利用 Bootstrapping 等技术从搜索引擎返回的网页中获得可信的正负样本用来训练分类器，再用该分类器来评估搜索引擎返回的更多网页的抽取结果，获得网页文本中实体的关系及实例。

在空间关系信息抽取研究方面，目前主要开展了空间关系关键词、空间语义、机器学习 3 个方面的研究。空间关系关键词相关研究主要基于统计方法和关键词识别的方法来实现空间关系的提取。余丽等(2016a)针对文本蕴含的地理实体关系分布稀疏的特点，提出一种语境增强的关键词提取方法，实现地理实体关系的关键词识别和地理实体关系的抽取。余丽等(2016b)提出一种开放式地理实体关系的自动抽取方法，通过 Bootstrapping 技术统计词语的词性、位置和距离特征来计算语境中词语权值，据此确定描述地理实体关系的关键词。

张雪英等(2012)通过分析中文文本中描述地理空间关系的语言特点，构建了中文文本的地理空间关系标注体系，实现了中文文本中地理空间关系描述的结构化表达。吕鹏飞等(2017)建立了基于文献的地质实体关系抽取模型，采用统计语言模型作为关系抽取方式，采用 Bootstrapping 算法作为关系扩展方式，进行了关联关系发现和关系扩展发现研究。Elia A 等(2013)通过构建复杂的空间关系表达语法，提出了一种基于词典-语法的方法来实现从非结构化文本中自动提取空间关系。宋卿等(2017)提出 Bootstrapping 的关系种子集自动生成方法，并在迭代过程中加入扩展和过滤规则，最终得到准确度和复用性较高的实体关系提取模式。现有的关键词及语义抽取方法大多基于规则或基于模板匹配方式，这类方法构建的规则与模板迁移性较差，无法有效抽取规则之外的属性信息；基于传统的机器学习方法受限于语料库及特征表征局限性，其严重依赖人工设计特征，无法有效捕获长距离上下文信息，很难直接应用到复杂的非结构化地质建模约束信息抽取中。

2 三维地质建模属性信息描述文本特点

笔者的研究对象主要是来源于全国地质资料馆中的公开地质报告。典型三维地质建模属性见表 1。报告中的三维地质建模信息文本描述方式及形式比较丰富，具备如下的一些特点。

(1)地质实体属性信息类型多样，包括可度量属性和不可度量属性。如“卓戈洞组(D_3z)仅见于妥坝以北、出露宽度在 200 m 左右，厚度大于 149 m”和“永珠组以黑色泥板岩发育为特征”，前者代表可度量属性，后者代表不可度量属性信息。当然，面对不同的地质实体对象，属性信息中的度量值与非度量值涵盖情况不同。

(2)不同地质实体对象属性信息完整表达一般包括“地质实体”、“属性名称”和“属性值”3 类。在某些情况下，地质报告中文本描述时往往只出现属性值信息，一般情况下可以依据属性值类型推断具体属性类型。

(3)中文地质报告文本属性信息描述具备一定的分散性，很多情况下会出现属性信息分散在多个句子、多个段落、多个篇章中。然而，很多情况下属性名称及对应的属性值都聚集在单个的句子描述与表达中。一般一段地质报告文本通常是围绕某一个地质对象主体来进行描述与表达，因此地质对象的实体与属性信息之间常常会保持唯一的对应关系。如：“该地层沿北西—南东构造线呈构造块体分布在石炭—二叠纪多彩蛇绿混杂带中，出露面积很小，不足 5 km^2，控制最大厚度大于 1 039.52 m。地层分区属西金乌兰-金沙江地层分区”。

(4)中文地质报告中地质实体的属性信息描述语句往往比较长，很多情况下在句子层级上，属性名称与具体的属性值之间在位置上是紧密相邻的。如“叠置厚度 1 039.52 m”、“斜长石：28%～47%”、“年龄有(2156±61)Ma”等。也存在一个句子中包含多个属性名称及对应的属性值的情况，根据文本距离最近原则能够形成“属性名称-属性值”这种键值对。

(5)中文自然语言描述具有一定的灵活性与随意性。中文地质报告文本描述中存在部分地质对象

或地质事件的属性信息描述时存在指代或者省略的情况。比如可能会存在直接以省略或者指代的形式来描述上文语境中地质对象或地质事件的相关属性值信息。

(6)中文地质报告文本中关于地质对象或地质事件的描述往往会呈现一定的规律性与特殊性,通常情况下会伴随着大量的触发词或者指示符号,这些触发词或者指示符号对于地质对象或者地质事件的非时空要素信息的描述与表现具有非常显著性的作用。如"地理坐标:起点东经:95°07′03″,北纬:33°15′30″,海拔高程:4 608 m;终点东经:95°06′07″,北纬:33°14′15″,高程:4 463 m"中的":"符号。

表 1　典型三维地质建模属性样例表

Tab. 1　Typical three-dimensional geological model attribute sample

类型	样　　例
几何	……贡德勇玛岩体,地表形态近圆形,……
	……呈椭圆形,长 18 km,宽 8～10 km,总体呈北西向延伸
	……恒星错铜多金属矿化点位于昌都县妥坝乡,地理坐标:东经 97°40′32″－97°48′05″,北纬 31°25′36″－31°31′57″,……
拓扑	…背斜轴线多呈北西向,向斜轴线大多为北西西、北西向……
	……轴向近南北,并表现出东翼陡,……
属性	……F91 断层为额艾普断裂带之主断层,走向 330～349 m,延伸约 80 km,断层破碎带宽 30～70 m…
	……上盘倾向 56°,下盘倾向 234°,倾角 79°
	……北东翼产状 40°∠60°,南西翼产状 230°∠25°……
	……导致岩石中 Ba/Sr 值大部分接近 1,Ba/Rb 值大多在 1.66～5.77,Rb/Sr 值在 0.19～0.67 之间

3　基于 BERT-BiLSTM-CRF 模型

当前,随着深度神经网络技术在信息抽取任务中的广泛应用,在实体属性抽取领域不依赖人工特征的端到端模型 BiLSTM-CRF 逐渐成为主流。而经过预训练后的 BERT 模型能够更好处理文本信息,与深度神经网络模型相结合并应用到后续文本识别任务中,能够明显提升准确率。

3.1　模型总体架构

笔者所构建的 BERT-BiLSTM-CRF 模型是在 BiLSTM-CRF 模型基础上增加了预训练模型 BERT,采用端到端的方式实现三维地质建模约束性信息序列标注任务。BERT 模型通过非监督学习方式从大规模文本中抽取中文语句双向语义特征,能够更好地表征不同语境中句法和语义信息,有效地减少对人工特征和领域知识的依赖。BERT-BiLSTM-CRF 模型总体架构(图 1)。整个模型可分为 BERT 层、BiLSTM 层及 CRF 层 3 部分。文本序列输入后,首先通过 BERT 预训练过程将词语表示为向量形式;随后将词向量输入 BiLSTM 层进行语义编码,进一步获取上下文句子特征;最后通过 CRF 层进行解码并得到概率最大的预测标签序列。

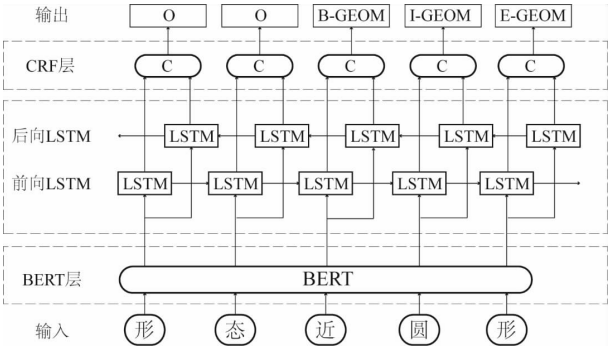

图 1　BERT-BiLSTM-CRF 模型图

Fig. 1　BERT-BILSTM-CRF model diagram

3.2　BERT 预训练模型

BERT 模型采用了双向 Transformer 编码结构(图 2)。Transformer 拥有较强的特征提取能力,其核心 Self-attention 机制能够通过一个句子中词语之间的相关度来调整权重值从而获得每个词的表征,这样每个词语向量值不仅包含词语本身含义,还包含了与其他词语的关联程度。因此,BERT 在对每个词语进行编码时,能够充分考虑词语本身含义及上下文中其他词语权重大小,从而具有强大的编

码能力。

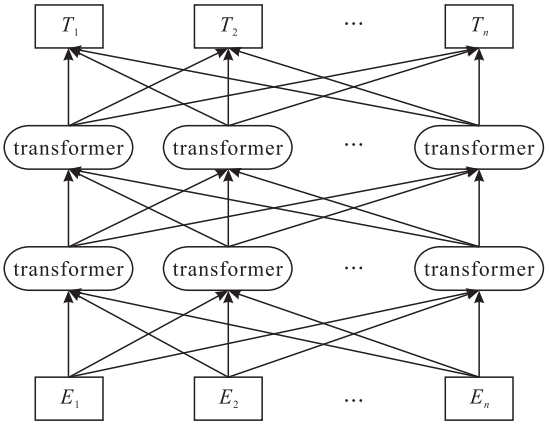

图 2 BERT 模型图

Fig. 2 Model structure diagram

BERT 模型的输入序列见图 3。其中的每个字符向量是由词嵌入、句子嵌入和位置嵌入 3 部分求和得到，通过深层双向编码最终生成词向量。其中，词嵌入对应每个词唯一向量表示；句子嵌入对应句子的向量表示；位置嵌入表示字符在句中位置。另外，BERT 模型还通过掩码语言模型(Masked language model)和下一句预测(Next sentence prediction)来获取字符级和句子级的语义关系。掩码语言模型通过随机隐藏句子中某些词语，预测被隐藏的词语，从而训练出深度双向语言表示，下一句预测是为了理解句子间的关联关系，获得句子级语义关系。

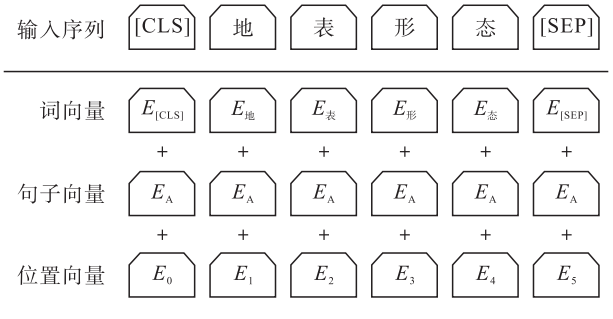

图 3 BERT 预训练模型词向量组成图

Fig. 3 BERT pre-training model word volume composition

3.3 BiLSTM 模型

为了解决循环神经网络(RNN，recurrent neural networks)的梯度消失和梯度爆炸问题，出现了长短期记忆网络(LSTM，directional long short-term memory)。相较于传统 RNN 模型，LSTM 引入了记忆单元与门控机制，用来控制信息的记忆、更新与遗忘，从而可以更好地捕捉句子间长距离的依赖关系。相比于 GRU 单元，LSTM 更加高效，LSTM 单元见图 4。

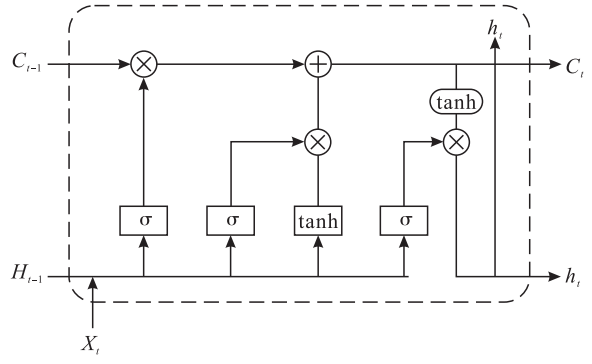

图 4 LSTM 内部结构图

Fig. 4 LSTM internal structure diagram

每个单元通过输入门、输出门与遗忘门来控制单元状态，输入门决定输入到单元的信息；输出门决定单元可输出的信息；遗忘门决定遗忘单元中有哪些信息。具体计算公式如下(1)～(6)所示。

(1)计算输入门：

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

(2)计算遗忘门：

$$f_t=\sigma(W_f\cdot[h_{t-1},X_t]+b_f) \tag{2}$$

(3)计算输出门：

$$o_t=\sigma(W_o\cdot[h_{t-1},X_t]+b_o) \tag{3}$$

$$h_t=o_t*\tanh(C_t) \tag{4}$$

(4)细胞状态更新：

$$\widetilde{C_t}=\tanh(W_c\cdot[h_{t-1},X_t]+b_C) \tag{5}$$

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C_t} \tag{6}$$

BiLSTM 网络由前向 LSTM 与后向 LSTM 两层网络相连构成，该架构能够更好地理解上下文信息，捕捉双向语义依赖关系，从而提升实体识别性能。

3.4 CRF 模型

条件随机场 CRF 是一种条件概率分布模型，经常用于词性标注、实体识别等任务中。在构建模型中，序列输入通过 BiLSTM 层后，输出结果即为词语对应各个类别的权重分数，将其作为 CRF 层输入，序列中分数最高的类别即认定为最终预测结果。在 BiLSTM 层之后加入 CRF 作为输出解码层，可以为

最终的预测标签加入约束条件,并保证标签合法性,确保标签之间顺序正确。所添加的约束条件通过 CRF 层自动学习,具体约束包括以下 2 点。

(1)实体以“B-”、“O”或“S-”作为开头,而不是“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。

CRF 的基本算法利用条件概率 $P(Y/X)$ 来描述模型,给定一组输入句子序列 $X=\{x_1,x_2,\cdots,x_n\}$ 与输出预测序列 $Y=\{y_1,y_2,\cdots,y_n\}$,得分为:

$$score(X,Y)=\sum_{i=1}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{7}$$

预测序列概率计算方式如下:

$$p(Y\mid X)=\frac{e^{score(X,Y)}}{\sum\limits_{\widetilde{y}\in Y_x}e^{score(X,\widetilde{Y})}} \tag{8}$$

解码时,$\widetilde{Y}$表示标注序列,Y_x 表示可能的标注序列值,通过维特比(Viterbi)算法得到得分最大的预测标签序列,公式如下:

$$y^*=\text{argmax}(score(X,Y)) \tag{9}$$

4　实验及结果分析

4.1　数据来源

本文的数据来源于全国地质资料馆中公开的地质报告,获取了 43 篇不同类型地质报告数据(包括区域地质报告、矿产资源地质报告、水文地质报告、遥感地质报告等)。对于获取的地质报告通过数据预处理的方式采用人工标注形式形成深度学习模型所需要语料库。并将标注的语料库按照 7∶2∶1 比例形成训练集、验证集和测试集。数据集统计信息见表 2。具体的深度学习术语示例见表 3。

表 2　数据集统计信息表

Tab. 2　Dataset statistics information

分类	句子数	字符数	术语实体数
训练集	988	78 912	3 905
验证集	565	45 687	1 971
测试集	381	30 034	1 445

表 3　深度学习术语示例标注表

Tab. 3　Example labeling of deep learning

上	盘	倾	向	56°	下
O	O	O	O	S-Prop	O
盘	倾	向	是	234°	,
O	O	O	O	S-Prop	O

BERT-BiLSTM-CRF 模型基于深度学习 Keras 框架实现,开发语言为 Python3.6。实验环境采用 12 核 Intel Core i9-10700 CPU 处理器,64GB 内存,NVIDIA GeForce GTX3090GPU,Ubuntu 14.10 Linux64 位操作系统。模型中相关参数设置见表 4。

表 4　BERT-BiLSTM-CRF 模型参数设置表

Tab. 4　BERT-BILSTM-CRF model parameter settings

BERT 模型参数	参数设置	BERT-CRF 模型参数	参数设置	BiLSTM-CRF 模型参数	参数设置
Layer	12	Learn_rate	5E-5	Max_seq_length	128
Hidden	768	Warmup_proportion	0.2	Drop_rate	0.2
Heads	12	Train_batch_size	20	Clip	5
Parameters	110M	Train_epochs	35	Lstm_size	128

4.2　评价指标

本文中采用的评价指标为查准率(Precision)、召回率(Recall)和 F1 值(F1-score),其具体计算公式为:

$$P=\frac{T_p}{T_p+F_p}\times 100\%$$

$$R=\frac{T_p}{T_p+F_N}\times 100\% \tag{10}$$

$$F1=\frac{2PR}{P+R}\times 100\%$$

其中,T_P 代表模型识别正确属性数量,F_P 代表模型识别不相关属性数量,F_N 代表相关属性但未识别出的数量。对于属性抽取结果的判别只有当实

体的边界、属性类别与属性值标签都识别正确，才认为当前抽取结果是正确的。

4.3 实验结果

基于 BERT-BiLSTM-CRF 模型的实验结果见表 5。同时，笔者还对比了其他 5 类信息抽取与实体识别领域主流模型。可以看到，提出的深度学习模型在所构建的三维地质建模约束属性语料库上 P、R、$F1$ 值分别达到了 91.22%、90.88%、91.05%，均大幅领先其他模型。具体表现如下：

(1)相比于传统的条件随机场模型 CRF，P、R、$F1$ 值分别提升了 10.88%、11.00%、10.94%。可见深度神经网络 BiLSTM 的加入能够更好地理解上下文信息，获取文本特征，从而达到良好的实体识别效果。

(2)BiLSTM-CRF 模型及 BiLSTM-Att-CRF 模型 $F1$ 值分别为 84.15%、85.98%，均落后于本文所提出的 BERT-BiLSTM-CRF 模型，说明 BERT 预训练模型的加入使得模型能够充分考虑词语本身含义及上下文中其他词语权重大小，获得强大的编码能力，有效提升实体识别精度。

表 5 不同深度学习模型地质建模约束性信息抽取性能表

Tab. 5 Different deep learning model Geological modeling constraint information extraction performance

模型	P(%)	R(%)	F1-score(%)
CRF	80.34	79.88	80.11
IDCNN-CRF	82.76	81.06	81.90
BiLSTM-CRF	84.22	84.09	84.15
BiLSTM-Att-CRF	86.98	85.01	85.98
BiGRU	88.11	87.23	87.67
BERT-BiLSTM-CRF	91.22	90.88	91.05

BERT-BiLSTM-CRF 模型地质建模约束性信息抽取结果示例见表 6。当输入句子“断层破碎带宽 30～80 m，倾向 120°，倾角 46°。下盘倾向 240°，倾角 78°。”，模型可自动抽取词语“带宽”、“倾向”、“倾角”，并识别出文本中所描述的词语属性。另外，还会自动识别出词语拓扑(Topo)、几何(Geom)等类别。

表 6 BERT-BiLSTM-CRF 模型地质建模约束性信息抽取结果展示表

Tab. 6 BERT-BILSTM-CRF model Geological modeling constraint information extraction results display

序号	输 入 句 子	正确抽取结果
1	断层破碎带宽 30～80 m，倾向 120°，倾角 46°。下盘倾向 240°，倾角 78°。	带宽：30～80 m/Prop 倾向：120°/Prop 倾角：46°/Prop 倾向：240°/Prop 倾角：78°/Prop
2	带内的矿化斑岩体的平面形态大多呈不规则的似圆形，剖面形似蘑菇状、倒水滴状、瓶状等，呈现上大下小的形态。	似圆形/Geom 蘑菇状/Geom 倒水滴状/Geom 瓶状/Geom
3	二长花岗斑岩体，作北东～南西延伸，呈长条状，长约 1 500 m，宽约 400 m，面积约 0.6 km^2。	北东—南西/Topo 长：1 500 m/Prop 宽：400 m/Prop 面积：0.6 km^2/Prop
4	贡德勇玛岩体，地表形态近圆形，直径约 2 400 m，面积约 7.5 km^2。	近圆形/Geom 直径：2 400 m/Prop 面积：7.5 km^2/Prop

注：Prop＝Property 代表属性；Topo＝Topology 代表拓扑；Geom＝Geometry，代表几何。

现有地质报告中的插图及地质数据描述信息中蕴含有大量有价值的建模数据，但是不能直接用于地质模型构建，通过机器理解和抽取方式把这些数据充分地挖掘出来，进行准确地理解，有效地融合，

并基于这些数据进行高效的三维模型构建是目前需要重点解决的难点问题。充分挖掘此类信息用于三维地质模型构建,可以为地质建模提供一种新的数据来源,进而与现有的数据进行匹配、融合并在此基础上构建三维地质模型。通过抽取地质建模的属性信息,能够结合地质剖面图构建地质实体空间特征、几何特征、构造特征的提取规则,通过地质知识库理解所提取信息,实现多模态数据约束下的区域三维地质建模。

5　结论

笔者在中文地质建模约束性信息描述特点分析的基础上,对传统的 BIOES 序列标注标签方案进行了扩展,设计了中文三维地质建模约束性信息抽取标注方案;将预训练模型 BERT、深度学习模型和统计学习模型结合起来,充分利用 3 种模型间优势;以中文地质报告为数据源,通过人工标注方式构建属性抽取数据集,设计了 CRF、IDCNN-CRF、BiLSTM-CRF、BiLSTM-Att-CRF、BiGRU 和 BERT-BiLSTM-CRF 模型,并对其抽取结果进行了对比分析。实验结果表明,本文所构建的模型 BERT-BiLSTM-CRF 在小规模数据集上抽取结果更优,且对不同类型约束下信息抽取具有良好的性能。后续的研究从领域知识的结合与语法规则融合,通过领域知识的加入,进一步提升三维地质建模约束性信息抽取精度与性能。

参考文献(References):

魏东琦,江宝得,张静雅. 非结构化地质数据内容存储方法研究[J]. 西北地质,2021,54(04):266-273.

WEI Dongqi, JIANG Baode, ZHANG Jingya. Research on content storage method for unstructured geological data [J]. Northwestern Geology, 2021, 54(04): 266-273.

王亚辉,张茂省,师云超等. 基于综合物探的城市地下空间探测与建模[J]. 西北地质,2019,52(02):83-94.

WANG Yahui, ZHANG Maosheng, SHI Yunchao, et al. Urban underground space exploration and modeling based on integrated physical prospecting [J]. Northwestern Geology, 2019, 52(02): 83-94.

李兆亮,潘懋,韩大匡,等. 三维构造建模技术[J]. 地球科学,2016,41(12):2136-2146.

LI Zhaoliang, PAN Mao, HAN Dakuang, et al. Three-dimensional tectonic modeling techniques[J]. Earth Science, 2016, 41(12): 2136-2146.

陈麒玉,刘刚,何珍文,等. 面向地质大数据的结构-属性一体化三维地质建模技术现状与展望[J]. 地质科技通报,2020,39(04):51-58.

CHEN Qiyu, LIU Gang, HE Zhenwen, et al. Status and prospect of structure-attribute integrated 3D geological modeling technology for geological big data[J]. Geological Science and Technology Bulletin, 2020, 39(04): 51-58.

何紫兰,朱鹏飞,马恒,等. 基于多源数据融合的相山火山盆地三维地质建模[J]. 地质与勘探,2018,54(02):404-414.

HE Zilan, ZHU Pengfei, MA Heng, et al. Three-dimensional geological modeling of the Xiang Shan volcanic basin based on multi-source data fusion[J]. Geology and Exploration, 2018, 54(02): 404-414.

陈麒玉,刘刚,吴冲龙,等. 城市地质调查中知识驱动的多尺度三维地质体模型构建方法[J]. 地理与地理信息科学,2016,32(04):11-16+48+2.

CHEN Qiyu, LIU Gang, WU Chonglong, et al. A knowledge-driven approach to construct multi-scale 3D geological body model in urban geological survey[J]. Geography and Geographic Information Science, 2016, 32(04): 11-16+48+2.

侯卫生,刘修国,吴信才,等. 面向三维地质建模的领域本体逻辑结构与构建方法[J]. 地理与地理信息科学,2009,25(01):27-31.

HOU Weisheng, LIU Xiuguo, WU Xincai, et al. A logical structure and construction method of domain ontology for 3D geological modeling[J]. Geography and Geographic Information Science, 2009, 25(01): 27-31.

郭甲腾,代欣位,刘善军,等. 一种三维地质体模型的隐式剖切新方法[J]. 武汉大学学报(信息科学版),2021,46(11):1766-1773.

GUO Jiateng, DAI Xinwei, LIU Shanjun, et al. A new method for implicit sectioning of 3D geological body models [J]. Journal of Wuhan University (Information Science Edition), 2021, 46(11): 1766-1773.

代欣位,郭甲腾,刘善军,等. 基于动态四叉树索引的三维地质模型组合剖切算法[J]. 地理与地理信息科学,2020,36(04):8-13.

DAI Xinwei. GUO Jiateng, LIU Shanjun, et al. Combined profiling algorithm for 3D geological models based on dynamic quadtree indexing [J]. Geography and Geo-

graphic Information Science,2020,36(04):8-13.

王殷行. 基于凸凹理论的三维地质体空间关系模型研究[J]. 地理与地理信息科学,2019,35(01):1-5.

WANG Yinxing. Research on three-dimensional geological body spatial relationship model based on convexity-concave theory[J]. Geography and Geographic Information Science,2019,35(01):1-5.

苏丰龙，谢庆华，黄清泉，等. 基于直推式学习的半监督属性抽取 [J]. 山东大学学报：理学版，2016，51(03)：111-115.

SU Fenglong, XIE Qinghua, HUANG Qingquan, et al. Semi-supervised attribute extraction based on direct push learning [J]. Journal of Shandong University: Science Edition, 2016, 51(03): 111-115.

余丽，陆锋，刘希亮. 开放式地理实体关系抽取的 Bootstrapping 方法 [J]. 测绘学报，2016a，45(5)：616-622.

YU Li, LU Feng, LIU Xiliang. Bootstrapping method for open geographic entity relationship extraction [J]. Journal of Surveying and Mapping, 2016a, 45(5): 616-622.

余丽,陆锋,刘希亮,等.稀疏地理实体关系的关键词提取方法[J].地球信息科学学报,2016b,18(11):1465-1475.

YU Li, LU Feng, LIU Xiliang, et al. A bootstrapping algorithm for geo-entity relation extraction from online encyclopedia[J]. International Conference on Geoinformatics. IEEE, 2016b,18(11):1465-1475.

吕鹏飞，王春宁，朱月琴. 基于文献的地质实体关系抽取方法研究 [J]. 中国矿业，2017，26(10)：167-172.

LV Pengfei, WANG Chunning, ZHU Yueqin. Research on literature-based relationship extraction method for geological entities [J]. China Mining, 2017, 26(10): 167-172.

张雪英，张春菊，朱少楠. 中文文本的地理空间关系标注 [J]. 测绘学报，2012，41(3)：468-474.

ZHANG Xueying, ZHANG Chunju, ZHU Shaonan. Geospatial relationship annotation of Chinese texts [J]. Journal of Surveying and Mapping, 2012, 41(3): 468-474.

宋卿，戚成琳，杨越. 基于 Bootstrapping 的新闻事件型实体关系抽取方法 [J]. 中国传媒大学学报：自然科学版，2017，24(04)：46-50.

SONG Qing, QI Chenglin, YANG Yue. A Bootstrapping-based method for extracting news event-based entity relationships [J]. Journal of Communication University of China: Natural Science Edition, 2017, 24(04): 46-50.

Zhong D, Wang L, Lin B I, et al. Implicit modeling of complex orebody with constraints of geological rules[J]. Transactions of Nonferrous Metals Society of China, 2019, 29(11): 2392-2399.

Lyu M, Ren B, Wu B, et al. A parametric 3D geological modeling method considering stratigraphic interface topology optimization and coding expert knowledge[J]. Engineering Geology, 2021: 106300.

Wu X, Liu G, Weng Z, et al. Constructing 3D geological models based on large-scale geological maps[J]. Open Geosciences, 2021, 13(1): 851-866.

Etzioni O, Banko M, Soderland S, et al. Open information extraction from the web [J]. Communications of the Acm, 2008, 51(12):68-74.

Etzioni O,Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the Web: An experimental study [J]. Artificial Intelligence, 2005, 165(1): 91-134.

Li Y U, Feng L U, Liu X, et al. A Method of Context Enhanced Keyword Extraction for Sparse Geo-entity Relation [J]. Journal of Geo-Information Science, 2016, 18(11): 1465-1475.

Hwang J, Nam K W, Ryu K H. Designing and implementing a geologic information system using a spatiotemporal ontology model for a geologic map of Korea [J]. Computers & Geosciences, 2012, 48: 173-186.

Caers J. Modeling uncertainty in the earth sciences [M]. John Wiley & Sons, 2011.

Elia A, Guglielmo D, Maisto A, et al. A Linguistic-Based Method for Automatically Extracting Spatial Relations from Large Non-Structured Data[M]. Algorithms and Architectures for Parallel Processing. Springer International Publishing, Springer. 2013.

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.

Qiu Q, Xie Z, Wu L, et al. BiLSTM-CRF for geological named entity recognition from the geoscience literature [J]. Earth Science Informatics, 2019a, 12(4): 565-579.

Qiu Q, Xie Z, Wu L, et al. Geoscience keyphrase extraction algorithm using enhanced word embedding[J]. Expert Systems with Applications, 2019b, 125: 157-169.