**中文引用格式**:张念,周彩凤,万飞,等. 基于 BERT-BiLSTM-CRF 的隧道施工安全领域命名实体识别[J]. 中国安全科学学报,2024,34(12):56-63.

**英文引用格式**:ZHANG Nian, ZHOU Caifeng, WAN Fei, et al. Tunnel construction safety domain named entity recognition based on BERT-BiLSTM-CRF[J]. China Safety Science Journal, 2024, 34(12): 56-63.

# 基于 BERT-BiLSTM-CRF 的隧道施工安全领域命名实体识别*

张 念[1,2]副教授, 周彩凤[1,2], 万 飞[3]研究员, 刘 非[1], 王耀耀[1], 徐栋梁[1,2]
(1 太原理工大学 土木工程学院,山西 太原 030024;2 北京交通大学 隧道及地下工程教育部工程研究中心,北京 100044;3 交通运输部 公路科学研究所,北京 100088)

**中图分类号**:X928 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2024.12.0713

**基金项目**:中央引导地方科技发展资金资助(YDZJSX20231A021);交通运输部公路科学研究所(院)交通强国试点项目(QG2021-3-14-1);隧道及地下工程教育部工程研究中心(北京交通大学)开放研究基金资助(TUC2024-03)。

**【摘 要】** 为解决隧道施工安全领域传统命名实体识别(NER)方法存在的实体边界模糊、小样本学习困难、特征信息提取不够全面准确等问题,提出一种基于变换器的双向编码器表征(BERT)-双向长短时记忆(BiLSTM)网络-条件随机场(CRF)模型的隧道施工事故文本实体识别方法。首先,利用 BERT 模型将隧道施工事故文本编码得到蕴含语义特征的词向量;然后,将 BERT 模型训练后输出的词向量输入 BiLSTM 模型进一步获取隧道施工事故文本的上下文特征并进行标签概率预测;最后,利用 CRF 层的标注规则的约束,修正 BiLSTM 模型的输出结果,得到最大概率序列标注结果,从而实现对隧道施工事故文本标签的智能分类。将该模型与其他 4 种常用的传统 NER 模型在隧道施工安全事故语料数据集上进行对比试验,试验结果表明:BERT-BiLSTM-CRF 模型的识别准确率、召回率和 $F_1$ 值分别达到 88%、89%和 88%,实体识别效果优于其他基准模型。利用所建立的 NER 模型识别实际隧道施工事故文本中的实体,验证了其在隧道施工安全领域中的应用效果。

**【关键词】** 变换器的双向编码器表征(BERT); 双向长短时记忆(BiLSTM)网络; 条件随机场(CRF); 隧道施工; 安全领域; 命名实体识别(NER); 深度学习

# Tunnel construction safety domain named entity recognition based on BERT-BiLSTM-CRF

**ZHANG Nian**[1,2], **ZHOU Caifeng**[1,2], **WAN Fei**[3], **LIU Fei**[1], **WANG Yaoyao**[1], **XU Dongliang**[1,2]
(1 College of Civil Engineering, Taiyuan University of Technology, Taiyuan Shanxi 030024, China; 2 Research Center of Tunneling and Underground Engineering of Ministry of Education, Beijing Jiaotong University, Beijing 100044, China; 3 Research Institute of Highway Ministry of Transport, Beijing 100088, China)

**Abstract**: To solve the problems existing in the traditional NER methods in the domain of tunnel construction safety, such as fuzzy entity boundary, difficulty in small-sample learning, and insufficiently

* 文章编号:1003-3033(2024)12-0056-08; 收稿日期:2024-09-14; 修稿日期:2024-11-18

comprehensive extraction of feature information, an entity recognition method for tunnel construction accident text based on the BERT-BiLSTM-CRF model was proposed. Firstly, the BERT model was used to encode the tunnel construction accident text to obtain word vectors containing semantic features. Then, the word vectors output after the training of the BERT model were input into the BiLSTM model to further obtain the context feature of the tunnel construction accident text and conduct label probability prediction. Finally, by utilizing the constraints of the annotation rules of the CRF layer, the output result of the BiLSTM model was corrected, and the maximum probability sequence annotation result was obtained, so as to realize the intelligent classification of the labels of the tunnel construction accident texts. Comparative experiments were conducted between this model and other four commonly used traditional NER models on the tunnel construction safety accident corpus dataset. The results show that the recognition accuracy rate, recall rate and $F_1$ value of the BERT-BiLSTM-CRF model are 88%, 89% and 88% respectively, and the entity recognition effect is better than other benchmark models. By using the established NER model to recognize the entities in the actual tunnel construction accident texts, its application effect in the domain of tunnel construction safety is verified.

**Keywords**: bidirectional encoder representations from transformers (BERT); bidirectional long short-term memory (BiLSTM); conditional random fields (CRF); tunnel construction; safety field; named entity recognition (NER); deep learning

## 0 引 言

隧道施工开挖具有高风险性和不可预见性,常常面临各种灾害事故,如塌方、突泥涌水、瓦斯爆炸等。预防这些事故的发生需依赖各类隧道施工安全知识。目前,这些知识分散存储于零散的资料里,如隧道施工技术规范、隐患排查清单以及事故调查报告等[1-2]。尽管持续积累的数据为隧道施工安全带来了极为丰富的知识来源,却难以通过有效的手段将这些信息重复利用。作为自然语言处理中的一项重要且基础的研究任务,命名实体识别(Named Entity Recognition,NER)旨在高效识别文本中具有特定意义的实体并标注其信息[3]。通过 NER 技术挖掘隧道施工事故文本数据,可自动抽取所需要的实体信息。进一步分析和利用所抽取的信息将有助于提高隧道施工安全管理水平,为隧道的数字化施工安全信息发展提供技术支持。

NER 研究经历了从基于规则和词典的方法发展到基于统计机器学习的方法,再到基于深度学习的方法的演进[4]。由神经网络衍生出的基于深度学习的方法由于不需要大量人工设计特征或规则而受到相关研究人员的关注,成为当前 NER 的研究热点。NER 任务常用基于深度学习的 NER 模型,包括循环神经网络模型(Recurrent Neural Network,RNN)[5]、卷积神经网络模型[6]、图神经网络模型[7]、长短期记忆(Long-Short Term Memory,LSTM)网络模型[8]和预训练模型等。将上述神经网络模型接入条件随机场(Conditional Random Fields,CRF)[9]已成为处理 NER 任务的主流方法。2018 年谷歌提出一种双向上下文特征编码的变换器的双向编码器表征(Bidirectional Encoder Representations from Transformers,BERT)预训练语言模型[10],在 NER 任务中为预训练模型的出现带来了显著的性能提升。目前,基于 BERT 的 NER 方法在医学、油气[11]、电力、建筑等诸多领域都有了较为成熟的应用。然而,针对隧道施工安全领域的 NER 研究多局限于铁路隧道,如王莉等[12]在构建地铁工程安全事故知识图谱时采用基于深度学习的方法获取实体;胡珉等[13]将双向长短时记忆网络(Bidirectional LSTM,BiLSTM)接入 CRF 神经网络模型自动提取盾构施工案例的关键词,人工进行校验和处理构建了盾构隧道施工领域知识图谱;张鹏翔[14]提出多维字符特征+BiLSTM+CRF 模型的实体抽取方法,并进行了铁路设备事故信息抽取;常弘[15]在数据驱动的地铁施工安全风险评估与应对研究中采用基于深度学习的 BiLSTM+CRF 中文 NER 模型进行知识抽取。上述研究采用基于深度学习的方法,在一定程度上完成了铁路隧道领域的实体识别任务,但主要依赖于手工设计的特征表示,且由于其输入的字向量是通过 Word2Vec 模型[16]生成的静态字向量,语义信息表征不全面,从而影响实体识别的准确率。

鉴于此,笔者拟采用融合注意力机制的 BERT

预训练语言模型编码隧道施工安全文本,从而得到能够表征字上下文语义信息的动态字向量,以减少误差累积。将训练后的动态字向量输入 BiLSTM-CRF 模型,以得到实体最优标签序列,提出适用于隧道施工安全领域的 BERT-BiLSTM-CRF 实体识别模型,并验证该模型的应用效果,以期解决传统 NER 方法存在的实体边界识别模糊、特征信息提取不够全面的问题。

## 1 NER 模型整体框架及流程

采用将融合注意力机制的 BERT 预训练模型与传统 BiLSTM-CRF 模型相结合的隧道施工安全领域实体识别方法,构建 BERT-BiLSTM-CRF 实体识别模型,模型结构主要由以下 3 部分组成。

1) BERT 层。BERT 模型由输入层、编码层 Transformer 编码器和输出层构成。BERT 层中首先处理隧道施工事故文本形成单个字符;同时每个输入的词向量($\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n$)都包含对字/词的嵌入、字/词所在句子的嵌入和字/词所在句中位置的嵌入;然后将 3 种嵌入和的词向量输入到双向 Transformer 编码器中;最后输出具有隧道施工安全事故文本语义特征的字向量($\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_n$)。在 BERT 中每个编码单元都引入多头注意力机制,以此提取出蕴含在事故信息里的丰富语义特征。

2) BiLSTM 层。BiLSTM 模型为双向长短时期记忆网络,由正向 LSTM 和反向 LSTM 组合而成。它凭借巧妙的门设计改进 RNN,成功解决了 RNN 存在的梯度爆炸和长期依赖问题。BiLSTM 模型对每个序列按照时间步分别采用顺逆序计算得到隐藏状态,再在每个时刻结合正向 LSTM 层和反向 LSTM 层的相应输出结果,利用向量拼接得到最终的输出。语义提取计算如下式:

$$\boldsymbol{h}_t = f(w_1\boldsymbol{x}_t + w_2\boldsymbol{h}_{t-1}) \tag{1}$$

$$\boldsymbol{h}_t' = f(w_3\boldsymbol{x}_t + w_5\boldsymbol{h}_{t+1}') \tag{2}$$

$$\boldsymbol{o}_t = g(w_4\boldsymbol{h}_t + w_6\boldsymbol{h}_t') \tag{3}$$

式中:$f$ 和 $g$ 分别为 sigmoid 和 tanh 激活函数;$w_n$ 为不同位置的权重参数;$\boldsymbol{x}_t$ 为 $t$ 时刻输入 BiLSTM 模型的事故文本向量;$\boldsymbol{h}_t$ 和 $\boldsymbol{h}_t'$ 分别为正向 LSTM 获取的上文序列信息和逆向 LSTM 获取的下文序列信息;$\boldsymbol{o}_t$ 为 $t$ 时刻输出 BiLSTM 模型的字标签得分向量。采用 BiLSTM 层正反双向编码前一层输出的字向量,通过学习隧道施工事故案例文本中的依赖关系,深度理解隧道施工事故文本,综合提取出文本的特征向量,并将得到的字标签得分向量传递给下一层。

3) CRF 层。CRF 模型是一种广泛用于 NER 领域的 CRF 模型,CRF 接收来自 BiLSTM 的字标签得分向量 $\boldsymbol{H}=(\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_n)$ 作为输入,通过模型概率计算得到输出状态序列 $\boldsymbol{R}=(\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_n)$。

CRF 层能够利用相邻标签间的约束关系,过滤掉不符合规则的序列标注,获取最优的隧道施工事故文本实体的标注序列。如一个句子只能是以 B-label 或 O-label 作为开始,若以 I-label 作为开始则是错误的;若前一个标签为 B-label1,则只能是 I-label1 紧随其后,而不是 B-label1、I-label2 等不符合规则的形式,以此减少预测错误标签的出现。

采用 BERT-BiLSTM-CRF 实体识别模型的总体框架如图 1 所示。

图 1 中,以四川都江堰董家山隧道瓦斯爆炸事故为例作为模型的输入序列,将文本分割得到的“董”“家”“山”“隧”“道”“瓦”“斯”“爆”“炸”“事”“故”11 个不同字符,依据向量表转换成对应的向量形式,输入到 BERT 层得到每个字的动态字向量表示;BiLSTM 拼接 BERT 层输出的字向量,进一步提取序列特征和编码,并给出每个字对应标签的概率,但是输出的字标签相互之间是独立的,未考虑连续标签之间存在的依赖关系,所以存在不符合标注规则的情况;通过 CRF 层剔除不合理序列,修正后得到实体最优标签序列。

## 2 隧道施工安全领域语料集及数据

### 2.1 隧道施工安全领域语料集构建

在进行隧道施工安全领域的 NER 之前,先需要建立一个事故案例全面、分类科学的数据库。从中国知网数据库、百度百科、新闻网页及各省应急管理厅进行事故信息提取,是一种较为科学和高效的获取事故案例的途径。为尽可能地避免遗漏,以隧道施工事故、隧道塌方、隧道突涌水、隧道瓦斯爆炸等关键字为主题进行检索,筛选过滤掉与隧道施工事故案例无关的文献,总计得到 114 篇期刊文献、10 篇百度百科文本及 18 篇相关事故报告。

为提高文本的准确性以保证试验的可靠,从多个平台获取的文献资料经过简单复制后不能直接作为标注的源文件,要清洗数据文本。首先,删除文本中的无效信息,仅保留其中的事故信息相关文段,将内容较少的文本合并,共形成 25 篇 word 文本;然后,删除空格、处理复制过程中的错误字符并进行换行处理,另存为 txt 文本格式,编码格式采用 UTF-8。通过数据清洗过程,在使数据完整的基础

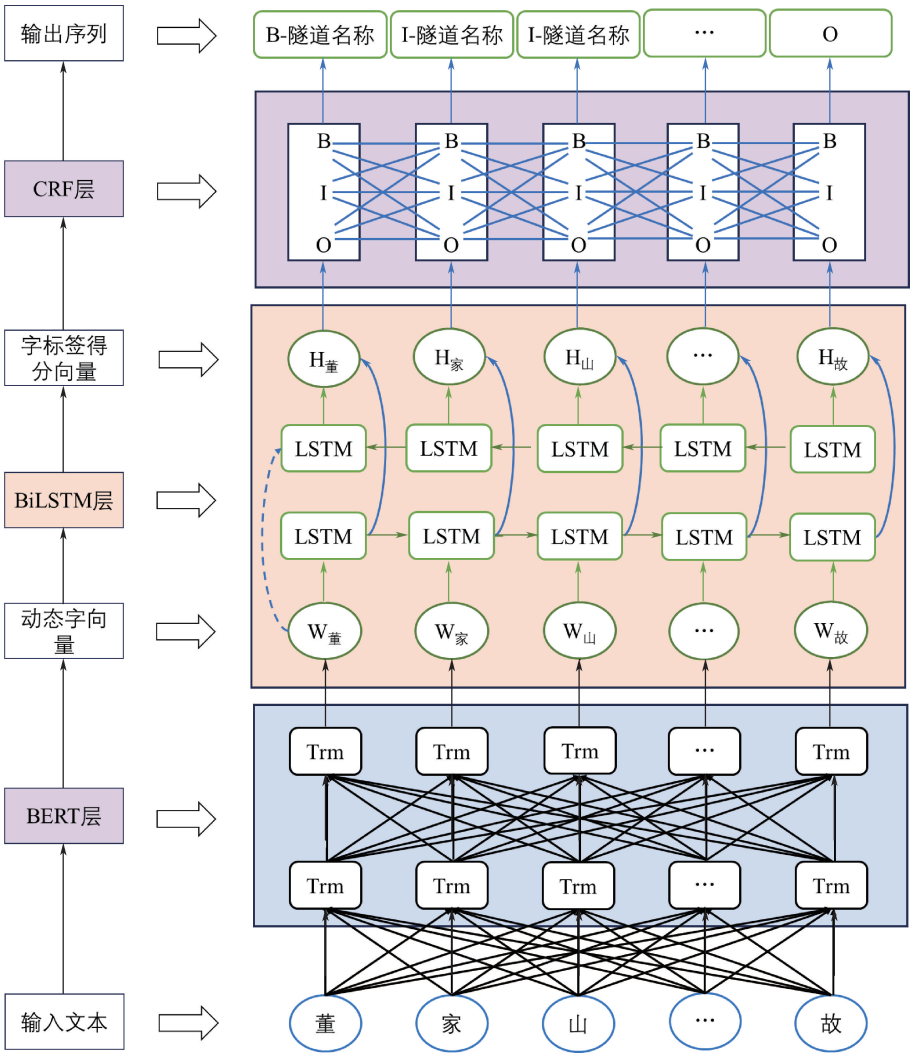


**图 1 模型整体框架**

**Fig. 1 Overall frame of model**

上去除原始文本中图片、空格、表格、杂乱字符等,并最大程度地保留原始文本内容,最终形成 1 293 条隧道施工事故语料,这些语料中包含丰富的事故信息,如事故时间、事故类型、事故地点、隧道名称、伤亡损失等。最后,将语料集中的文本随机打乱顺序,按照 8∶2 的比例划分为训练集和测试集,得到的训练集和测试集分别包含 1 078、215 条语句。

## 2.2 隧道施工安全领域数据标注

brat 标注工具具有简单易用、协作性、灵活性、支持多种标注类型等特点,因此,数据标注任务采用基于网页的 brat 文本标注工具,用于结构化处理非结构化的原始隧道施工事故文本。进行标注前,在 Ubuntu 系统中完成 brat 标注软件的环境配置及安装启动,并将定义好的实体类型添加到相关配置文件中。借助 brat 文本标注工具人工标注事故类型、事故时间、事故地点、隧道名称、伤亡损失 5 类实体,原始 txt 文本经标注后会产生相应 ann 文件,用来记录标注语料的位置信息,标注过程如图 2 所示。

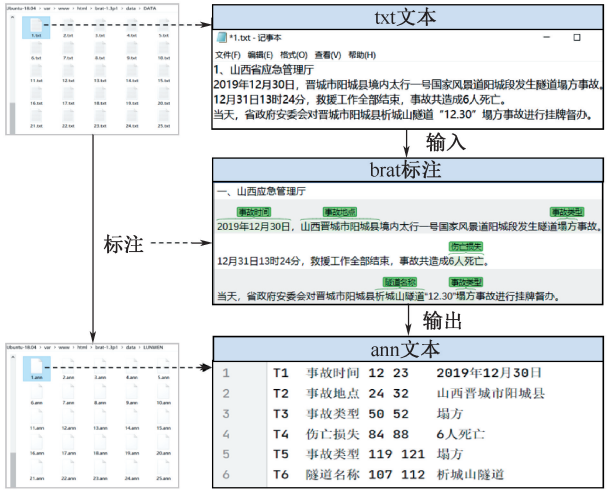


**图 2 标注过程**

**Fig. 2 Annotating process**

图 2 中,ann 文本中从左至右,每一列分别代表实体的编号、实体类别、实体起始位置索引、实体终止位置索引及实体内容。如第一行中,T1 表示实体 1,其实体类别为事故时间,实体位于第 12~23 个字符,实体具体内容为 2019 年 12 月 30 日。本次数据标注任务包括标注事故时间、事故类型、事故地点、隧道名称、伤亡损失等 5 类实体,共计产生 2 484 个被标注实体。

语料标注完毕后,使用 BIO 实体标注法进行数据预处理,其中,B(Begin)代表实体的起始字符,I(Intermediate)代表实体的中间字符,O(Other)代表不属于任何实体。事故类型、事故时间、事故地点、隧道名称、伤亡损失这 5 类实体依次产生 11 类标签:B-事故类型、I-事故类型、B-事故时间、I-事故时间、B-事故地点、I-事故地点、B-隧道名称、I-隧道名称、B-伤亡损失、I-伤亡损失和 O;再给每个字符打上相应标签,将 ann 转换成逐个字的 BIO 标注。以宜万铁路马鹿菁隧道施工工地发生透水事故为例,标注结果形式如图 3 所示。

{"text": ["宜", "万", "铁", "路", "马", "鹿", "箐", "隧", "道", "施", "工", "工", "地", "发", "生", "透", "水", "事", "故"],
"label": ["O", "O", "O", "O", "B-隧道名称", "I-隧道名称", "I-隧道名称", "I-隧道名称", "I-隧道名称", "O", "O", "O", "O", "O", "O", "B-事故类型", "I-事故类型", "O", "O"]}

**图 3 语料标注结果示例**

**Fig. 3 Example of corpus annotation result**

## 3 NER 模型试验结果及分析

### 3.1 模型试验环境及其参数设置

基于 Python 语言和 Pytorch 深度学习框架进行试验,隧道施工安全领域的中文 NER 模型的具体训练运行环境设置见表 1。

**表 1 试验环境**

**Table 1 Experimental environment**

| 试验环境名称 | 配置 |
|---|---|
| 操作系统 | 64 位 Windows10 系统 |
| 编程语言 | Python3.7 |
| 深度学习框架 | Pytorch1.10.2+cu113 |
| GPU | NVIDIA GeForce GTX 1650 |
| GPU 加速器 | cuDA 11.3,CUDNN 8.0 |
| 内存/GB | 16 |
| 代码编写平台 | PyCharm Professional Edition 2021.2.4 |

注:图形处理器(Graphics Processing Unit,GPU);计算统一设备架构(Compute Unified Device Architecture,CUDA);CUDA 深度神经网络库(CUDA Deep Neural Network,cuDNN)。

试验前,根据自建数据集中的句长统计,将隧道施工事故文本数据语料最大句长设为 256。试验选用对于大多数任务通用的 Adam 优化器,因其对于不同参数有不同的学习率,具有较好的收敛性[17]。具体试验训练参数设置见表 2。

**表 2 参数配置**

**Table 2 Parameter configuration**

| 参数名称 | 参数值 | 参数注释 |
|---|---|---|
| max_seq_len | 256 | 最大序列长度 |

**续表 2**

| 参数名称 | 参数值 | 参数注释 |
|---|---|---|
| lstm_hidden | 128 | LSTM 隐藏层单元 |
| dropout | 0.5 | 随机失活率 |
| learning rate | $3\times10^{-5}$ | 学习率 |
| optimizer | Adam | 优化器 |
| batch_size | 6 | 每批次训练样本数 |

### 3.2 模型评价指标

隧道施工安全领域的实体识别试验采用深度学习中常用的准确率 $P$、召回率 $R$ 和 $F_1$ 值作为模型识别效果的评价标准[18],分别评价所定义 5 类实体,其计算公式具体如下:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$F_1=\frac{2PR}{P+R} \tag{6}$$

式中:TP 为模型正确识别的实体数;FP 为模型错误识别的实体数;FN 为模型未识别出的实体数。准确率 $P$ 用于评估模型识别实体的准确性,召回率 $R$ 用于评估模型识别实体的全面性,$F_1$ 值用来评估模型时则兼顾了准确率和召回率,在提高两者结果的前提下尽可能缩小两者之间的差异。

### 3.3 BERT-BiLSTM-CRF 模型试验结果分析

为验证 BERT-BiLSTM-CRF 模型的性能,在相

同试验环境下,采用以下 4 种常用模型 BERT、BiLSTM-CRF、BiLSTM 和 BERT-CRF 在上述相同的数据集上进行对比试验。在模型实际训练期间,从 Python 的代码数据库 seqeval. metrics 导入序列标注模型结果评估模块 classification_report 来实现以上评估数据的输出。试验结果对比见表 3。

**表 3 不同模型试验结果对比**

**Table 3 Comparison of experimental results of different models**

| 模型 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| BERT | 0. 87 | 0. 87 | 0. 87 |
| BiLSTM-CRF | 0. 82 | 0. 78 | 0. 80 |
| BiLSTM | 0. 70 | 0. 70 | 0. 70 |
| BERT-CRF | 0. 88 | 0. 88 | 0. 88 |
| BERT-BiLSTM-CRF | 0. 88 | 0. 89 | 0. 88 |

表 3 中,所有试验数据是在不同的迭代次数下所取得的最优值,通过比较发现,BERT-BiLSTM-CRF 模型在各个测量指标上都能达到最优,智能识别准确率 $P$、召回率 $R$ 和 $F_1$ 值分别达到 88%、89% 和 88%,充分证明了该模型对隧道施工事故文本实体识别的有效性。采用 BERT-BiLSTM-CRF 模型的评价指标随迭代次数的变化趋势如图 4 所示。

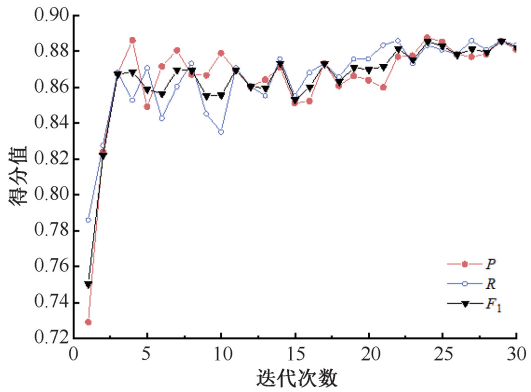


**图 4 BERT-BiLSTM-CRF 模型评价指标变化趋势**

**Fig. 4 Change trend of evaluation indexes of BERT-BiLSTM-CRF model**

从图 4 中可以看出,在迭代次数小于 3 时,BERT-BiLSTM-CRF 模型的各评价指标值均随着迭代次数的增加急速上升;在迭代 3~23 次时,该模型的各评价指标上升速度有了明显放缓,且变化曲线上下小幅度地波动;在迭代 23 次以后,该模型的各指标值几乎不再波动,曲线逐渐趋于平稳,表明该模型训练开始收敛,不会出现大幅度波动。模型在第 24 次迭代时 $F_1$ 值达到最优,为 88%。

采用 BERT-BiLSTM-CRF 模型在训练过程中损失值随迭代次数的变化曲线如图 5 所示。从图 5 中可以看出,该模型自开始训练起,损失值开始大幅度下降,当到达 10 个迭代次数后,开始趋于稳定状态,也验证了图 4 的评价指标变化趋势是在一定阶段后模型的评价指标开始在一定范围内上下波动,没有出现大幅度上升或下降。在模型训练过程中,模型的状态变化为从最开始的不拟合状态,进入优化拟合状态。在模型试验的训练过程中,迭代次数为 24 时,$F_1$ 值达到最高,表明此时模型训练效果较好,同时为避免模型发生过拟合,该模型的迭代次数应设置 24 次左右。得到的模型试验结果见表 4。

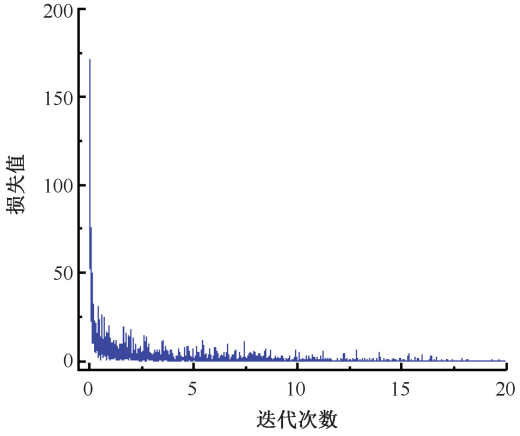


**图 5 BERT-BiLSTM-CRF 模型训练损失函数变化曲线**

**Fig. 5 Change curve of training loss function of BERT-BiLSTM-CRF model**

**表 4 模型试验结果**

**Table 4 Results of model experiment**

| 实体类别 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 事故地点 | 0. 81 | 0. 86 | 0. 86 |
| 事故时间 | 0. 91 | 0. 94 | 0. 92 |
| 事故类型 | 0. 84 | 0. 95 | 0. 84 |
| 伤亡损失 | 0. 88 | 0. 90 | 0. 89 |
| 隧道名称 | 0. 93 | 0. 93 | 0. 93 |

BERT-BiLSTM-CRF 模型在 5 个实体类别上的识别效果都优于其他基准模型,并且在每一个类别实体上的 $P$、$R$ 和 $F_1$ 值都达到 80%以上。表 4 中,事故时间和隧道名称这 2 类实体的 $P$、$R$、$F_1$ 值均超过 92%。究其原因,一方面,这 2 类实体的字数长短较为合适,不会因为长距离依赖现象导致模型性能下降;另一方面,它们的形式固定、边界特征较为显著,例如:×年×月×日、××隧道等实体的结尾容易辨识。而事故地点、事故类型、伤亡损失这 3 类实体因类型较多且表述方式各异,$F_1$ 值略微下降,但均在

84%以上。

### 3.4 模型应用效果

为验证隧道施工事故文本实体识别模型的应用效果,以搜集到的实际隧道工程施工事故案例文本为例,进行实体信息预测,预测结果如图 6 所示。图中“文本”行表示被预测的语句,方框中为事故案例的真实的实体信息,“实体”行表示利用 BERT-BiLSTM-CRF 模型识别出的文本中包含的具体实体信息,包含实体类型、实体名称及实体在文本中的位置。


**图 6 BERT-BiLSTM-CRF 模型识别结果**

**Fig. 6 Recognition results of BERT-BiLSTM-CRF model**

从图 6 中可以看出,该文本中含有 9 个实体,且涵盖全部(5 种)定义的实体类型,所有的实体边界及实体类型均被正确识别,模型的智能识别正确率达到 100%。试验预测结果表明:基于 BERT-BiLSTM-CRF 模型的 NER 方法在一定程度上解决了传统 NER 方法存在的实体边界识别模糊、特征信息提取不够全面的问题,在隧道施工安全领域具有较好的适用性。在此过程中得到的隧道施工事故实体将促进隧道施工安全领域知识图谱的构建,并更好地指导施工安全管理的安全培训。

## 4 结 论

1)构建全面的隧道施工安全事故语料库,这些语料能够为后续的模型训练提供数据,形成的语料库具有较高的价值密度,可在隧道施工安全领域的不同任务中重复利用。

2)针对隧道施工安全领域事故文本实体抽取不准确、不全面等问题,构建 BERT-BiLSTM-CRF 算法模型,在自建数据集上对模型进行多次迭代训练,优化模型的超参数。试验结果证明了该模型在隧道施工安全领域信息抽取的有效性,准确率、召回率及 $F_1$ 值均达到 80%以上。

3)利用 BERT-BiLSTM-CRF 模型识别输入的隧道施工事故文本中的实体,文本中的实体均被正确识别,并给出所属的实体类别及位置信息,实际应用效果良好。

4)尽管 BERT-BiLSTM-CRF 模型在各类实体的识别上都比其他模型表现更优,但是针对每一个类别实体的识别效果依旧有进一步探究的价值。后续的研究可从收集更多隧道领域数据、优化模型超参数以及改进模型算法等方面着手,以更好地提高模型识别准确率,为构建高品质的隧道施工安全领域知识图谱筑牢根基。

## 参 考 文 献

[1] 薛亚东, 董宏鑫, 李彦杰. 山岭公路隧道施工安全风险评估理论体系[J]. 天津大学学报:自然科学与工程技术版, 2019, 52(增 1):84-91.
XUE Yadong, DONG Hongxin, LI Yanjie. Theoretical System for the safety risk assessment of mountain tunnel construction[J]. Journal of Tianjin University: Science and Technology, 2019, 52(S1): 84-91.

[2] 周泽林, 张凯, 张恒, 等. 属性识别理论下的岩溶隧道地表塌陷风险评价[J]. 中国安全科学学报, 2022, 32(11): 105-112.
ZHOU Zelin, ZHANG Kai, ZHANG Heng, et al. Risk assessment of surface subsidence in karst tunnels under attribute recognition theory[J]. China Safety Science Journal, 2022, 32(11): 105-112.

[3] 康怡琳, 孙璐冰,朱容波,等. 深度学习中文命名实体识别研究综述[J]. 华中科技大学学报:自然科学版, 2022, 50(11):44-53.
KANG Yilin, SUN Lubing, ZHU Rongbo, et al. Survey on Chinese named entity recognition with deep learning[J]. Journal of Huazhong University of Science and Technology: Nature Science Edition, 2022, 50(11): 44-53.

[4] 高翔, 王石, 朱俊武,等. 命名实体识别任务综述[J]. 计算机科学, 2023, 50(增 1):26-33.

GAO Xiang, WANG Shi, ZHU Junwu, et al. Overview of Named Entity Recognition Tasks[J]. Computer Science, 2023, 50(S1): 26-33.

[5] HAMMER B. Learning with recurrent neural networks[J]. Assembly Automation, 2001, 21(2): 178-183.

[6] CUN Y L, BOSER B, DENKER J S, et al. Handwritten digit recognition with a back-propagation network[J]. Advances in Neural Information Processing Systems, 1990, 2(2): 396-404.

[7] BARCZ A, SZYMAŃSKI Z, JANKOWSKI S. Implementation aspects of graph neural networks[J]. Proceedings of SPIE-the International Society for Optical Engineering, 2013, 8903(12): 405-408.

[8] SUNDERMEYER M, NEY H, SCHLUTER R. From feedforward to recurrent LSTM neural networks for language modeling[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(3): 517-529.

[9] DIESNER J, CARLEY K M. Conditional random fields for entity extraction and ontological text coding[J]. Computational and Mathematical Organization Theory, 2008, 14(3): 248-262.

[10] LI Fei, JIN Yonghao, LIU Weisong, et al. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study[J]. JMIR Medical Informatics, 2019, 7(3):DOI:10.2196/14830.

[11] 高国忠, 李宇, 华远鹏,等. 基于 BERT-BiLSTM-CRF 模型的油气领域命名实体识别[J]. 长江大学学报:自然科学版, 2024, 21(1):57-65.
GAO Guozhong, LI Yu, HUA Yuanpeng, et al. Named entity recognition in oil and gas field based on the BERT-BiLSTM-CRF model[J]. Journal of Yangtze University: Natural Science Edition, 2024, 21(1):57-65.

[12] 王莉, 王建平, 许娜,等. 基于知识图谱的地铁工程事故知识建模与分析[J]. 土木工程与管理学报, 2019, 36(5):109-114,122.
WANG Li, WANG Jianping, XU Na, et al. Knowledge graph-based metro engineering accidents knowledge modeling and analysis[J]. Journal of Civil Engineering and Management, 2019, 36(5): 109-114,122.

[13] 胡珉, 胡星辰, 宋云,等. 基于知识图谱的盾构施工辅助决策系统设计和开发[J]. 现代隧道技术, 2023, 60(1): 66-75.
HU Min, HU Xingchen, SONG Yun, et al. Design and development of the shield construction assistant decision-making system based on the knowledge graph[J]. Modern Tunnelling Technology, 2023, 60(1): 66-75.

[14] 张鹏翔. 多维字符特征表示的铁路设备事故信息抽取方法[J]. 中国安全科学学报, 2022, 32(6): 109-114.
ZHANG Pengxiang. Information extraction method for railway equipment accidents based on multi-dimensional character feature representation[J]. China Safety Science Journal, 2022, 32(6): 109-114.

[15] 常弘. 数据驱动的地铁施工安全风险评估与应对研究[D]. 徐州: 中国矿业大学, 2023.
CHANG Hong. Research on data-driven safety risk assessment and response in metro construction[D]. Xuzhou: China University of Mining and Technology, 2023.

[16] JATNIKA D, BIJAKSANA A M, SURYANI A A. Word2Vec model analysis for semantic similarities in English words[J]. Procedia Computer Science, 2019, 157: 160-167.

[17] YUAN Wei, HU Fei, LU Liangfu. A new non-adaptive optimization method: stochastic gradient descent with momentum and difference[J]. Applied Intelligence, 2021, 52(4): 1-15.

[18] 夏成魁, 李少波. 基于 BERT-BiLSTM-MHA-CRF 的中文命名实体识别方法[J]. 计算机与数字工程, 2023, 51(9):2 087-2 091,2 102.
XIA Chengkui, LI Shaobo. Chinese named entity recognition method based on BERT-BiLSTM-MHA-CRF model[J]. Computer & Digital Engineering, 2023, 51(9): 2 087-2 091,2 102.

**作者简介:** 张 念 (1984—),男,湖北襄阳人,博士,副教授,主要从事隧道及地下工程安全与防灾方面的研究。E-mail: zhangnian@tyut.edu.cn。