

郭心全,李俊波,沈鹏,等. 面向普铁桥梁缺陷描述文本的命名实体识别技术研究[J]. 智能计算机与应用,2025,15(5):199-204. DOI:10.20169/j.issn.2095-2163.24112503

面向普铁桥梁缺陷描述文本的命名实体识别技术研究

郭心全¹,李俊波¹,沈鹏¹,吴霞¹,李林²,刘静²

(1 中国铁道科学研究院集团有限公司 电子计算技术研究所,北京 100081; 2 北京经纬信息技术有限公司,北京 100081)

摘要: 针对非结构化普铁桥梁缺陷描述文本数据利用度不高的问题,为有效挖掘数据价值,提出了一种面向普铁桥梁缺陷描述文本的命名实体识别方法。该方法首先以 Lattice LSTM 模型在综合考虑文本上下文信息和词语信息的基础上提取特征信息,以应对中文缺少显式分隔符和特定领域 NER 任务中专有词汇识别效率低的问题;随后,引入自注意力机制捕捉数据局部信息关键特征;接着,以全连接层网络将所有特征进行整合;最后,以 CRF 模块实现标签结果预测。选取了桥隧子系统普铁桥梁缺陷描述文本数据进行人工标注获取实验数据,对所提方法进行验证,同时对比了 HMM、CRF、BiLSTM、BiLSTM-CRF、Bert-BiLSTM-CRF 和 Lattice LSTM-CRF 六种基线模型。结果表明,该方法可有效识别普铁桥梁缺陷描述文本实体,为发挥本领域非结构化文本数据价值提供经验借鉴。

关键词: 桥梁工程;命名实体识别;Lattice LSTM;注意力机制;CRF

中图分类号: TP391.1

文献标志码: A

文章编号: 2095-2163(2025)05-0199-06

Research on named entity recognition technology for defect description text of ordinary speed railway bridges

GUO Xinquan¹, LI Junbo¹, SHEN Kun¹, WU Xia¹, LI Lin², LIU Jing²

(1 Institute of Electronic Computing Technologies, China Academy of Railway Sciences Co., Ltd., Beijing 100081, China;

2 Beijing Jingwei Information Technology Co., Ltd., Beijing 100081, China)

Abstract: Aiming at the problem of low utilization of unstructured defect description text data for ordinary speed railway bridges, a named entity recognition method for ordinary speed railway bridge defect description text is proposed to effectively explore the value of data. This method firstly uses the Lattice LSTM model to extract feature information based on a comprehensive consideration of text context information and word information, in order to address the problems of the lack of explicit separators in Chinese and low recognition efficiency of specialized vocabulary in specific domain NER tasks; Subsequently, self-attention mechanism is introduced to capture key features of local information in the data; Then, all features are integrated using a fully connected layer network; Finally, the CRF module is used to predict label results. The paper selects text data describing the defects of ordinary speed railway bridges in the bridge tunnel subsystem for manual annotation to obtain experimental data, and validates the proposed method. At the same time, the paper compares six baseline models: HMM, CRF, BiLSTM, BiLSTM-CRF, Bert-BiLSTM-CRF, and Lattice LSTM-CRF. The results indicate that this method can effectively identify the textual entities describing defects in ordinary speed railway bridges, providing experience and reference for leveraging the value of unstructured text data in this field.

Key words: bridge engineering; named entity recognition; Lattice LSTM; attention mechanism; CRF

0 引言

普通速度铁路(以下简称:普铁)是指设计速度低、只能让火车以普通速度行驶的铁路^[1]。在中国,普铁是指火车行驶速度不大于 160 km/h 速度级别的非客运专线以及大于 140 km/h 速度级别的客运专线^[2]。铁路桥梁作为铁路线路的重要组成部分

分,投资规模大、安全性要求高,对于铁路运输安全和畅通起着至关重要的作用^[3-4]。近年来,铁路客货运运量日渐增长,列车开行密度不断加大,桥梁设备数量和服役年限也在增加,且设备长期暴露于复杂的自然环境中,其缺陷问题不容忽视。因此,如何快速明确桥梁设备缺陷类型,判定设备状态至关重要。

基金项目: 中国铁道科学研究院集团有限公司科研课题(2023YJ131)。

作者简介: 郭心全(1994—),男,硕士,助理研究员,主要研究方向:铁路信息化,智能化。Email:gxq_2019@163.com。

收稿日期: 2024-11-25

哈尔滨工业大学主办 ◆ 科技创新与应用

目前,大数据、人工智能技术与铁路行业融合越来越深,在铁路车务、工务、供电等专业方面均有应用^[5]。在自然语言处理(Natural Language Processing, NLP)领域,一些学者针对铁路行车事故^[6]、信号设备故障^[7-9]、供电故障^[10]、道岔运维^[11]等方面的文本数据开展了研究,但针对铁路桥梁缺陷文本数据的研究还较为欠缺。铁路工务安全生产管理信息系统(Railway Permanent Way Safety Production Management Information System)^[12]作为铁路工务专业唯一全国应用的安全生产系统,发展至今,系统包含了线路、桥隧、路基、调度等多个子系统。其中,桥隧子系统于2018年建设并推广应用至今,产生了大量的桥梁设备缺陷描述文本数据。但该数据作为非结构化数据,仅作为材料提供查看,用于桥梁设备缺陷履历管理模块,尚未开展大数据分析应用研究,数据价值未得到发挥。命名实体识别(Named Entity Recognition, NER)作为自然语言处理领域中的一项重要基础任务,可以从文本中识别出特定实体,如人名、地名等,对发挥文本数据价值有着重要意义^[13]。

针对以上问题,构建了一种融合网格长短期记忆网络(Lattice Long Short-Term Memory, Lattice LSTM)模型、注意力机制(Attention Mechanism, AM)

和条件随机场(Conditional Random Field, CRF)模型的普铁桥梁缺陷描述文本命名实体识别方法。该方法将中文字符和词汇的融合信息作为模型输入,以应对中文缺少显式分隔符和特定领域NER任务中专有词汇识别效率低的问题^[14]。本文选取了桥隧子系统普铁桥梁缺陷描述文本数据进行人工标注,通过对比实验分析验证了所提方法的有效性。

1 模型设计

本文提出的普铁桥梁缺陷描述文本命名实体识别模型总体框架如图1所示,主要包括Lattice LSTM层、AM层和CRF层。首先将中文字符信息作为基础,在词语最后一个字符前增加本词语信息作为输入层。嵌入层以训练好的Word2Vec模型将字符和词语分别转化为低维度字符嵌入向量和词语嵌入向量。接着,将字符和词语嵌入向量输入至Lattice LSTM层,在综合考虑文本上下文信息和词语信息的基础上提取特征信息。随后,将提取的特征信息输入AM层提取文本信息中局部重要信息特征。然后,在全连接层将提取的所有特征信息进行整合,降低特征位置信息对分类结果的影响。最后,通过CRF层输出标签预测结果。

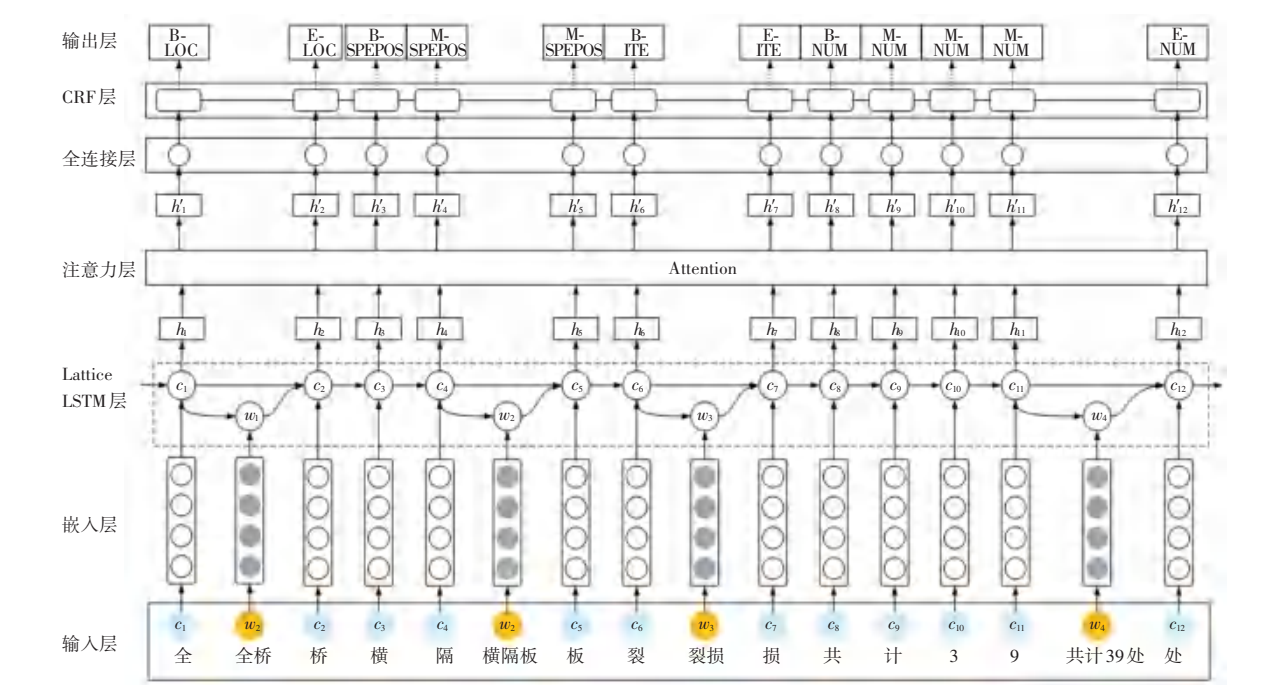


图 1 模型总体框架

Fig. 1 Overall framework of the model

1.1 Lattice LSTM 层

在中文 NER 技术研究中,实体多数由词语表

示,个别由单个字表示,因此,中文分词的准确性和中文 NER 的效果具有强相关性。当前,针对字粒度

的方法相对来说优于基于词粒度的方法,但又无法应对中文中多数情况下单字不具备完整语义的问题。尤其是特定领域中包含很多领域词汇,直接影响了分词的效果,间接影响了中文领域 NER 的效果。在普铁桥梁缺陷描述文本中甚至存在诸多行业术语缩写,线路方向如“成方”、“广方”、“昆端”等,桥台如“和台”、“沪台”、“吉台”等。为解决以上问题,文本以 Lattice LSTM 模型综合考虑字符和词语

信息。

Lattice LSTM 模型由 Zhang 等学者^[15]于 2018 年提出,解决了基于字符的中文 NER 技术无法利用词和词序信息的问题。Lattice LSTM 模型结构如图 2 所示,可以看出该模型在字符的基础上,考虑了词语和词序信息,其输入为桥梁缺陷描述文本中的字向量和对应词语向量。

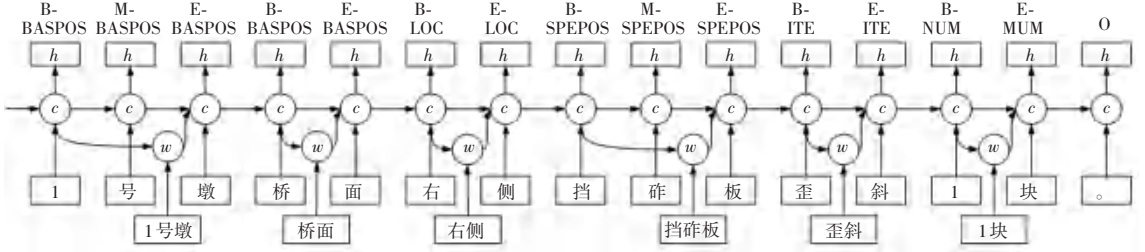


图2 Lattice LSTM 结构

Fig. 2 Lattice LSTM structure

Lattice LSTM 模型仍然针对单字进行基本计算,推得的公式具体如下:

$$\begin{pmatrix} \hat{e}_j^c \\ \hat{e}_{o_j}^c \\ \hat{e}_j^c \\ \hat{e}_{c_j}^c \\ \hat{e}_{c_j}^c \end{pmatrix} = \begin{pmatrix} \hat{e}^{\sigma} \\ \hat{e}^{\sigma} \\ \hat{e}^{\sigma} \\ \hat{e}^{\sigma} \\ \hat{e}^{\tanh} \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{u} \\ \hat{u} \\ \hat{u} \\ \hat{u} \end{pmatrix} \left(W^c \begin{pmatrix} \hat{x}_j^c \\ \hat{x}_{j-1}^c \end{pmatrix} + b^c \right) \quad (1)$$

$$c_j^c = f_j^c \odot c_{j-1}^c + i_j^c \odot \tilde{c}_j^c \quad (2)$$

$$h_j^c = o_j^c \odot \tanh(c_j^c) \quad (3)$$

其中, i_j^c 、 f_j^c 、 $\tilde{c}_j^c$ 分别表示 LSTM 模型结构中字符计算的输入门、输出门和遗忘门信息; σ 表示 Sigmoid 函数; $\tanh$ 表示双曲正切函数; W^c 、 b^c 分别表示字符的权重矩阵、偏置量,为模型参数; $\hat{x}_j^c$ 表示输入字符向量; $\hat{h}_j^c$ 表示字符隐藏状态。

相对于字符单元信息计算, LSTM 结构未考虑输出门。在开展字计算的同时,输入信息中增加词语信息,词语单元计算公式如下:

$$\begin{pmatrix} \hat{e}_{b,e}^w \\ \hat{e}_{b,e}^w \\ \hat{e}_{b,e}^w \\ \hat{e}_{b,e}^w \end{pmatrix} = \begin{pmatrix} \hat{e}^{\sigma} \\ \hat{e}^{\sigma} \\ \hat{e}^{\sigma} \\ \hat{e}^{\tanh} \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{u} \\ \hat{u} \\ \hat{u} \end{pmatrix} \left(W^w \begin{pmatrix} \hat{x}_{b,e}^w \\ \hat{x}_b^c \end{pmatrix} + b^w \right) \quad (4)$$

$$c_{b,e}^w = f_{b,e}^w \odot c_b^c + i_{b,e}^w \odot \tilde{c}_{b,e}^w \quad (5)$$

其中, $i_{b,e}^w$ 、 $f_{b,e}^w$ 、 $\tilde{c}_{b,e}^w$ 分别表示 LSTM 模型结构中词语计算的输入门、遗忘门和候选单元信息; σ 表示 Sigmoid 函数; $\tanh$ 表示双曲正切函数; W^w 、 b^w 分别表示词语的权重矩阵、偏置量,为模型参数; $\hat{x}_{b,e}^w$ 表示输入字符词向量。

进一步得到, $\hat{x}_j^c$ 和 $\hat{x}_{b,e}^w$ 的计算公式为:

$$\hat{x}_j^c = e^c(c_j) \quad (6)$$

$$\hat{x}_{b,e}^w = e^w(w_{b,e}^d) \quad (7)$$

将词语向量信息融入当前字符向量中进行特征计算。图 2 中,字符“墩” c_3^c 的信息包括字符“墩”向量 c_3^c 本身的信息和词语“1 号墩”向量 $\hat{x}_{1,2}^w$ 的信息。针对每个词语的单元结构 $c_{b,e}^w$,通过构建一个额外的门 $i_{b,e}^c$ 来控制该词语对当前字符 $c_{b,e}^c$ 的贡献度。 $i_{b,e}^c$ 计算如下所示:

$$i_{b,e}^c = \sigma \left(W^i \begin{pmatrix} \hat{x}_j^c \\ \hat{x}_{b,e}^w \end{pmatrix} + b^i \right) \quad (8)$$

同时考虑字信息 $\tilde{c}_j^c$ 和词语信息 $\tilde{c}_{b,e}^w$ 计算模型最新单元 c_i ,可用下式来表示:

$$c_j^c = \sum_{b' \in \{b' | w_{b'}^d, j \in D\}} \alpha_{b',j}^c \odot c_{b',j}^w + \alpha_j^c \odot \tilde{c}_j^c \quad (9)$$

其中, $\alpha_{b',j}^c$ 和 α_j^c 通过对 $i_{b',j}^c$ 和 i_j^c 分别进行归一化计算获得,使其各自和为 1。运算中用到的公式为:

$$\alpha_{b',j}^c = \frac{\exp(i_{b',j}^c)}{\exp(i_j^c) + \sum_{b' \in \{b' | w_{b'}^d, j \in D\}} \exp(i_{b',j}^c)} \quad (10)$$

$$\alpha_j^c = \frac{\exp(i_j^c)}{\exp(i_j^c) + \sum_{b' \in \{b' | w_{b'}^d, j \in D\}} \exp(i_{b',j}^c)} \quad (11)$$

1.2 AM 层

深度学习中 AM 是指用于模拟人类视觉注意力的机制,使其在处理大量信息时着重关注最重要的

部分信息,而忽略相对不重要的信息^[16]。在以 Lattice LSTM 层有效捕捉文本上下文信息后,使用注意力网络重点关注文本局部关键信息,可以降低相对不重要信息对结果的影响。自注意力(Self-Attention)机制,又称内部注意力机制(Intra Attention),是注意力机制中的特例,采用点乘注意力(Scaled Dot-Product Attention)作为奖惩策略^[17]。本文采用自注意力机制对 Lattice LSTM 层的输出进行加权求和计算,可用下式来表示:

$$Attention(Q,K,V)=Softmax(\frac{QK^T}{(d_k)^{\frac{1}{2}}})V \quad (12)$$

其中, Q 、 K 和 V 分别表示查询、键和值, d_k 表示查询和键的维度。

1.3 CRF 层

CRF 是一种判别式无向图模型,表示在给定输入随机变量 X 的条件下,输出随机变量 Y 的条件概率分布,可以用捕捉上下文标签间的依赖特征信息^[18]。在全连接层整合了特征信息、并得到输出 $h'_1, h'_2, \cdots, h'_l$ 后,这里 l 表示句子字符数量,再通过 CRF 层计算标签序列的概率值 $y = l_1, l_2, \cdots, l_l$ 。计算公式如下:

$$P(y \mid s) = \frac{\exp(\sum_i (W_{CRF}^{l_i} h_i + b_{CRF}^{(l_i-1, l_i)}))}{\sum_{y'} \exp(\sum_i (w_{CRF}^{l'_i} h_i + b_{CRF}^{(l'_i-1, l'_i)}))} \quad (13)$$

其中, y' 表示任意一个标签序列, $w_{CRF}^{l_i}$ 和 $b_{CRF}^{(l_i-1, l_i)}$ 分别表示模型权重和偏置量。

在 CRF 模型训练时,以最大似然函数作为损失函数进行优化调整。其函数公式如下:

$$L = \sum_{i=1}^N \log(P(y_i \mid s_i)) + \frac{\lambda}{2} \|\Theta\|^2 \quad (14)$$

其中, λ 表示 $L2$ 正则化参数, Θ 表示模型训练所需参数集合。

2 实验分析

2.1 实验环境

本文实验计算机设备操作系统为 64 位 Windows11 家庭版,处理器为 13th Gen Intel(R) Core (TM) i9-13900HX 2.20 GHz,内存 48 G。实验开发工具及版本为 PyCharm 2021.3.3,开发语言及版本为 Python3.6,神经网络开发框架及版本为 PyTorch1.7.1。

2.2 实验数据

从桥隧子系统普铁桥梁缺陷描述文本数据中抽取部分数据进行人工标注,获得本文所用实验数据。其中,训练集 13 147 行、验证集 3 626 行、测试集 3 605 行。数据中定义实体 9 类,其具体名称、标注标签和示例见表 1。

表 1 普铁桥梁缺陷文本描述实体定义

Table 1 Definition of entity for defect description text data of ordinary speed railway bridges

实体名称	标注标签	示例
线名	Line;-LIN	南昆线/滇藏线/坪田联络线
里程	Distance;-DIS	K114+057
线路方向	Direction;-DIR	红方/成方/广方/昆明端/上行/下行/大里程
桥名	Bridge;BRI	中寨 1#大桥
缺陷位置	Location;-LOC	全桥/左侧/上行侧/内外侧/第 6 孔/距离梁端 15 cm 处/底部/~间/吉侧
缺陷基本部位	Basic Position;-BASPOS	桥面/支座/右片梁/1#墩/桥墩/墩台/桥台/和台(和端桥台)/若台(若端桥台)/耳墙/检查梯/护锥/限高架/围栏
缺陷具体部位	Specific Position;-SPEPOS	连接杆/挡铁/圆钢/盖板/钢部件/螺栓/挡砟板/桥面板/排水管/支撑垫石/下翼缘
缺陷项目	Item;-ITE	锈蚀/掉块/脱落/外露/缺失
缺陷数量	Number;-NUM	共计 13 处/约 6 平/一处

2.3 数据预处理

在中文 NLP 领域 NER 任务中,数据序列标注方法包括 BIO 标注法、BIOES 标注法、BMES 标注法等^[19-21]。针对本领域实验数据特性,选择 BMES 标

注法进行数据标注。BMES 标注法是基于 BIO 标注法进行扩展的,可更准确地描述实体内部结构,有效应对存在较长实体的情况。其中 B(Begin)表示一个词的词首位置,M(Middle)表示一个词的中间位

置, E(End) 表示一个词的末尾位置, S(Single) 表示单字成词的实体, O(Outside) 表示非实体部分。部分标注数据见表 2。

表 2 部分标注数据

Table 2 Partially annotated data

字符及标注	字符及标注	字符及标注	字符及标注
第 B-LOC	蚀 M-ITE	孔 E-SPEPOS	刷 E-ITE
一 M-LOC	严 M-ITE	流 O	2 B-NUM
孔 E-LOC	重 E-ITE	下 O	0 M-NUM
南 B-LOC	约 B-NUM	部 O	平 M-NUM
头 E-LOC	6 M-NUM	位 O	方 M-NUM
吊 B-BASPOS	m E-NUM	, O	米 E-NUM
栏 E-BASPOS	, O		。 O
托 M-BASPOS	上 O	下 B-LOC	
架 E-BASPOS	方 O	游 E-LOC	南 B-LOC
局 B-ITE	为 O	护 B-SPEPOS	侧 E-LOC
部 M-ITE	泄 B-SPEPOS	底 E-SPEPOS	水 B-ITE
锈 M-ITE	水 M-SPEPOS	冲 B-ITE	平 M-ITE

经标注后, 训练集、验证集和测试集中各实体统计信息见表 3。

表 3 各实体统计信息

Table 3 Statistical information of each entity

实体类别	实体数量统计		
	训练集	验证集	测试集
线名	25	9	17
里程	23	12	25
线路方向	180	44	54
桥名	39	15	26
缺陷位置	661	166	153
缺陷基本部位	518	142	128
缺陷具体部位	468	112	142
缺陷项目	838	223	221
缺陷数量	378	104	87

2.4 评价指标

为验证模型性能, 本文采用 NER 任务中常用评价指标准确率 $P(Precision)$ 、召回率($Recall$) 和 $F1$ 值($F1_score$) 对模型进行评价和对比。具体地, 精确率是预测为正样本的数据中实际为正样本的数据所占比例; 召回率是实际为正样本的数据中预测为正样本的数据所占比例; $F1$ 值是精确率和召回率的调和平均值。各指标计算公式见如下:

$$Precision = \frac{TP}{TP + FP}$$

(15)

$$Recall = \frac{TP}{TP + FN}$$

(16)

$$F1_score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

(17)

其中, TP 表示实际为正样本数据且预测为正样本数据的数量, TN 表示实际为负样本数据且预测为负样本数据的数量, FP 表示实际为负样本数据但预测为正样本数据的数量, FN 表示实际为正样本数据但预测为负样本数据的数量。

2.5 结果分析

针对本文所用的实验数据集, 选取了隐马尔可夫(Hidden Markov Model, HMM)、CRF、双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)、BiLSTM-CRF、Bert-BiLSTM-CRF 和 Lattice LSTM-CRF 六种基线模型进行了实验对比, 验证了本文所提算法模型的优越性, 仿真实验结果见表 4。由表 4 可知, 本文所提模型评价指标精确率、召回率和 $F1$ 值分别达到了 89.45%、89.07% 和 89.26%, 均优于相比较的基线模型。同时由表 4 可以观察到各模型在本数据集的 NER 任务上的表现。融合字符和词语信息的模型优于以字符信息为基础的模型。在以字符信息为基础的模型中, 以 BiLSTM 模型捕捉上下文重要信息做 NER 任务效果最差; 常用于 NER 任务的生成式概率模型 HMM 和鉴别式概率模型 CRF 相对较好一些; 结合模型 BiLSTM+CRF 效果相对于 BiLSTM 模型有一定提升, 但仍不如单一的概率模型达到的效果; 而在结合了预训练模型 Bert 之后, 效果较优一些。在字符信息结合词语信息后, 模型效果又有了一个提升。

表 4 各模型对比结果

Table 4 Comparison results of various models

%

模型	精确率	召回率	$F1$ 值
HMM	80.30	79.49	79.51
CRF	86.42	85.80	85.85
BiLSTM	73.56	73.33	72.70
BiLSTM-CRF	77.99	77.87	77.28
Bert-BiLSTM-CRF	87.09	87.09	87.09
Lattice LSTM-CRF	88.27	88.75	88.51
本文模型	89.45	89.07	89.26

本文所提方法经训练后测试集数据损失值、精确率、召回率和 $F1$ 值随着迭代次数变化如图 3 所示, 其中为实现同图对比, 损失值是在缩放 500 倍后进行展示。由图 3 可知, 在模型迭代训练 20 次后基本收敛, 各值提升幅度大幅下降; 在训练 50 次后基

本稳定,模型精确率也达到 89%以上。

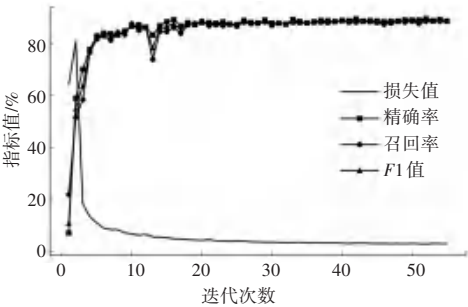


图 3 测试集各指标随迭代次数变化

Fig. 3 Changes in various indicators of the test set with the number of iterations

3 结束语

为实现普铁桥梁缺陷“一句话”识别的目标,有效发挥非结构化普铁桥梁缺陷描述文本数据的价值,本文针对普铁桥梁设备缺陷描述文本开展了NER技术研究,提出了一种融合 Lattice LSTM 模型、AM 和 CRF 模型的普铁桥梁缺陷描述文本 NER 方法。该方法以 Lattice LSTM 模型为网络起始模块,从中文词法入手,以字词信息融合作为模型输入,以有效应对中文缺少显式分隔符和领域 NER 任务中专有词汇问题。接着引入 AM 以进一步捕捉文本局部信息特征。随后以全连接层将所有特征信息进行融合,最终以 CRF 模块实现标签结果预测。

从桥隧子系统普铁桥梁缺陷描述文本数据中抽取了部分数据进行人工标注获取实验数据,对所提方法进行实验验证,然后选择了 HMM、CRF、BiLSTM、BiLSTM-CRF、Bert-BiLSTM-CRF 和 Lattice LSTM-CRF 六大基线模型进行实验对比。结果表明,本文所提方法在普铁桥梁缺陷描述文本数据集上进行 NER 任务效果更好,可为本领域开展文本挖掘研究提供支撑。未来,将继续优化数据集实体定义,持续提升模型准确率,同时,将考虑引入知识图谱技术,以提高非结构化数据表达效果,为实际应用做有效铺垫。

参考文献

[1] 苏学培. 普速铁路运输能力挖掘增效探讨与实践[J]. 百科论坛电子杂志,2018(17):164.

[2] 巩江峰,王伟,王芳,等. 截至 2023 年底中国铁路隧道情况统计及 2023 年新开通重点项目隧道情况介绍[J]. 隧道建设(中英文),2024,44(2):377-392.

[3] 李超,朱宏伟,姜子清,等. 铁路桥梁病害信息化管理系统设计[J]. 铁路计算机应用,2024,33(4):39-44.

[4] 朱肖. 铁路桥梁三维设计系统研发及模型深化应用[J/OL]. 铁道标准设计. (2024-04-15). <https://doi.org/10.13238/j.issn.1004-2954.202311130004>.

[5] 郭心全,吴霞,李俊波,等. 基于 GBDT 模型的接触网异物分类研究[J]. 智能计算机与应用,2024,14(6):41-49.

[6] 韩广,卜桐,王明明,等. 基于双通道双向长短时记忆网络的铁路行车事故文本分类[J]. 铁道学报,2021,43(9):71-79.

[7] 林海香,陆人杰,卢冉,等. 基于文本挖掘的铁路信号设备故障自动分类方法[J]. 云南大学学报(自然科学版),2022,44(2):281-289.

[8] 杨连报,李平,薛蕊,等. 基于不平衡文本数据挖掘的铁路信号设备故障智能分类[J]. 铁道学报,2018,40(2):59-66.

[9] 李新琴,史天运,李平,等. 基于文本的高速铁路信号设备故障知识抽取方法研究[J]. 铁道学报,2021,43(3):92-100.

[10] 杨涛存,郭剑峰,杜文然,等. 基于有限状态机模型的供电故障断电区间快速提取方法[J]. 中国铁路,2020(8):7-12.

[11] 林海香,白万胜,赵正祥,等. 面向高速铁路道岔运维文本的知识抽取方法[J]. 铁道科学与工程学报,2024,21(7):2569-2580.

[12] 孙美,郭年根,沈鹏. 铁路工务安全生产管理信息系统的设计与实现[J]. 铁路计算机应用,2013,22(8):23-27.

[13] 丁建平,李卫军,刘雪洋,等. 命名实体识别研究综述[J]. 计算机工程与科学,2024,46(7):1296-1310.

[14] 赵振宇,朱静静,张宇馨,等. 基于汉字上下文信息增强词典知识融入的中文命名实体识别[J]. 四川大学学报(自然科学版),2024,61(4):104-112.

[15] ZHANG Yue, YANG Jie. Chinese NER using lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. ACL,2018:1554-1564.

[16] 卢禹冰. 深度学习注意力机制设计与应用研究[D]. 兰州:兰州理工大学,2023.

[17] 李鸿飞,刘盼雨,魏勇. 基于自注意力和 Lattice-LSTM 的军事命名实体识别[J]. 计算机工程与科学,2021,43(10):1848-1855.

[18] 李军怀,陈苗苗,王怀军,等. 基于 ALBERT-BGRU-CRF 的中文命名实体识别方法[J]. 计算机工程,2022,48(6):89-94.

[19] 郑杜福,黄蔚,任祥辉. 一种基于 ERNIE 的军事文本实体关系抽取模型[J]. 信息技术,2021(2):38-43.

[20] 任刚. 基于知识图谱的政府采购咨询问答系统的研究与实现[D]. 西安:西安石油大学,2023.

[21] 刘盼,郭延明,雷军,等. 基于边界定位与纠偏的中文命名实体提取规则研究[J]. 计算机科学,2023,50(3):276-281.