

doi: 10.3969/j.issn.1002-0268.2024.04.019

面向多模态数据的高速公路交通事故 信息识别与评价

陈娇娜^{*1}, 陶伟俊¹, 靳引利²

(1. 西安石油大学 电子工程学院, 陕西 西安 710065; 2. 长安大学 电子与控制工程学院, 陕西 西安 710061)

摘要:为实现从自然语言描述的交通事故文本中提取应急处置信息,提出了一种基于预训练模型和BiLSTM-CRF的交通事故命名实体识别方法。首先,基于陕西省高速公路2021年6月至2022年8月的多模态交通事故数据,分别比较了3种深度学习模型的识别效果和训练时长。其次,利用官方微博交通事故语料作为袋外测试集,检验实体识别模型的鲁棒性。然后,从一致性和丰富性两个维度,构建了文本信息和结构化数据的多模态交通事故信息内容评价指标。最后,以测试集为例进行交通事故信息识别,分析了应急处置实体数量与事故持续时间的相关性,计算并探讨了信息内容评价指标结果。结果表明,BERT-BiLSTM-CRF在测试集和袋外测试集的加权F1值分别为97.029 4%和69.155 5%,为模型精度、训练效率和鲁棒性3个方面综合表现最优。处置机构、处置设备、未处置、处置中、处置效果的实体数量与持续时间之间的相关系数依次为0.309, 0.151, 0.137, 0.220和0.178,呈正相关性。天气、路产损失、交通分流、事故类型和伤亡情况的信息内容一致性依次为7.06%, 45.79%, 1.59%, 67.65%和47.59%,应急处置占为36%,变异性为1.305,说明文本信息蕴含丰富的应急处置信息,然而文本信息和结构化数据对同一交通事故的信息内容一致性尚待提高。研究结果可为提高交通事故信息采集质量和有效性提供参考。

关键词:智能交通; 交通事故; 多模态数据; 预训练模型; 双向长短期记忆

中图分类号: U491.3

文献标识码: A

文章编号: 1002-0268 (2024) 04-0186-08

Recognition and Evaluation on Expressway Traffic Accident Information Using Multimodal Data

CHEN Jiao-na^{*1}, TAO Wei-jun¹, JIN Yin-li²

(1. School of Electronic Engineering, Xi'an Shiyou University, Shaanxi, Xi'an 710065, China;

2. School of Electronics and Control, Chang'an University, Shaanxi, Xi'an 710061, China)

Abstract: In order to extract emergency response information from natural language descriptions of traffic accidents, a named entity recognition method is proposed based on pre-trained models and BiLSTM-CRF. The multimodal traffic accident data on expressways from June 2021 to August 2022 in Shaanxi province are analyzed as data sources. Firstly, 3 deep learning models are compared on entity recognition effect and training time. Secondly, the traffic accident corpus from official microblog is obtained to test the robustness. Moreover, according to the dimensions of consistency and richness, the evaluation indicators are constructed to enable quantitative assessment of traffic accident content for text data and structured data. Finally, the traffic accident information recognition is carried out by using the test dataset. The result shows that the weighted F1 values of BERT-BiLSTM-CRF on both test dataset and out-of-bag dataset are 97.029 4% and 69.155 5% respectively, which have the best comprehensive performance in terms of model accuracy, training

收稿日期: 2023-03-28

基金项目: 国家自然科学基金项目 (52002315); 国家重点研发计划项目 (2019YFB1600700)

作者简介 (* 通讯作者): 陈娇娜 (1989-), 女, 云南大理人, 博士. (chenjn@xsyu.edu.cn)

efficiency, and robustness. It is verified that there is a positive correlation between the number of emergency disposal entities and the duration of accident. The correlation coefficients of disposal agency, disposal equipment, un-disposal, disposal-ing and disposal effect are 0.309, 0.151, 0.137, 0.220 and 0.178 respectively. The content consistency of weather, road loss, traffic diversion, accident type and casualty are 7.06%, 45.79%, 1.59%, 67.65% and 47.59% respectively. The proportion of emergency response is 36%, and the variability is 1.305. It is proved that text data contain rich emergency disposal information, however, the content consistency of text data and structured data for the same traffic accident should be improved. The study result can provide reference for improving the quality and effectiveness of traffic accident information.

Key words: ITS; traffic accident; multimodal data; pre-trained model; bi-directional long and short term memory

0 引言

交通运输行业是社会生产活动和公众出行的重要保障,交通安全水平的重要作用也日趋显著。随着信息技术和管理水平的不断发展,更加丰富的交通事故信息被记录。交通事故信息是典型的多模态数据,从数据类型来看,既有人工录入的结构化字段和自然语言描述文本,又有系统自动记录的视频、音频等非结构化数据。从时间维度上看,随着事故演化或者处置推进,交通事故信息可能被多次记录,包含了从事故发生、发现、响应、处置、恢复的各个阶段。

交通事故具有复杂性和随机性,难以被结构化字段完整表征。应急处置是影响交通事故持续时间、道路交通延误的关键环节,通常以自然语言形式被记录。自然语言处理技术在事故文本智能分析^[1-3]中已经取得一定进展。Li等^[4]回顾了深度学习技术在命名实体识别中的最新应用。王颖洁等^[5]总结了中文命名实体识别研究中的关键问题,一些学者探讨了军事、航空等领域的命名实体识别效果^[6-8],刘兴丽等^[9]提出了适用于小样本命名实体识别的数据增强改进算法。张鹏翔^[10]通过多维字符特征+BiLSTM+CRF模型实现了铁路设备事故报告中7类信息提取。韩广等^[11]基于432条文本和十类铁路事故,建立了结合词向量和句向量双通道的WA-S-BiLSTM模型,王向前等^[12]以239份煤矿事故案例文本为数据来源,构建了7种实体类型的ALBERT-BiLSTM-CRF模型。刘婷等^[13]通过1000份报告设计了BERT-BiLSTM混合模型以识别8种水利工程事故原因。刘昭等^[14]从微博平台爬取了近2000条数据,采用XGBoost文本分类器识别交通突发事件。樊海玮等^[15]采集了4446条交通事故文本数据,构建了BERT-BiGRU-CRF模型提取地点、时间、原因、类

型、交通方式、伤亡结果、道路类型7类信息。李昀轩等^[16]对120191条交警情提取时间、地址和车牌信息,并提出了交警情分类模型和语义分析方法。由此可知,由于实体类型、数据集规模、应用领域的异质性,事故文本实体识别尚未获得一致结论。同时,大多数研究重点分析了时间、地点、原因、类型等属性信息^[17],对于处置阶段信息的提取尚未深入开展。在前期研究中验证了处置阶段信息对高速公路交通事故持续时间具有显著的调节效应^[18]。因此,准确识别交通事故文本中的应急处置信息对应急响应效率和应急预案优化具有重要意义。

交通事故信息质量直接影响数据挖掘下游任务的效果。准确评估信息质量能够为应急处置、交通管制提供决策依据。在事故发生初期,完全掌握事故所有信息是不切实际的。随着事故演化,天气状况、道路条件等因素会发生变化,事故信息也可能在不同阶段被逐渐补充。同时,由于信息采集人员经验差异,可能产生属性划分的不准确、结构化数据与文本数据不一致等情况。针对交通事故信息所表现出的复杂、海量、多阶段等特性,有必要进行更深入的研究。目前应急信息质量评估理论尚待进一步完善,面向多模态交通事故数据的信息质量评价仍极具挑战性。朱益平等^[19]建立政务微信公众号信息质量评价模型。徐文强等^[20]从内容质量、描述质量和信息约束角度,构建大数据环境下的应急信息质量评估体系。GUO等^[21]通过BiLSTM-CRF模型提取应急预案中信息发送、消息接收和常规任务3种类型,从响应层级关系、完整性、模糊性和冗余性对突发事件应急预案文本进行评价。

综上所述,交通事故命名实体识别研究中尚未深入涉及应急处置信息识别,且对于多模态的交通事故采集信息的质量评价,尚未形成有效结论。因

此,建立一种基于预训练模型和深度学习网络的交通事故命名实体识别模型,同时,从两种模态数据的一致性和丰富性,提出交通事故信息内容评价指标,实现文本信息抽取结果与结构化数据的信息内容量化评价,是自然语言处理技术在交通应急安全领域的重要应用。研究成果可以定量评估运营管理中信息采集质量和有效性,进而为应急处置流程优化提供决策依据。

1 交通事故文本信息识别模型

基于文本信息的交通事故命名实体识别模型由文本标注模块、预训练语言模块、双向长短时记忆(Bidirectional Long Short Term Memory, BiLSTM)模块和条件随机场(Conditional Random Field, CRF)模块构成。采用 BIO 标记法对交通事故文本进行标注;在预训练语言模块中使用双向 Transformer 编码结构实现文本序列的向量化表达,生成融合字向量、句向量以及位置向量的特征向量。将特征向量输入 BiLSTM 模型提取特征,运用 CRF 对 BiLSTM 输出的序列标注进行编码和制约,以最大概率的序列标注作为最终输出结果。

1.1 文本标注

交通事故文本标注采用 BIO 标记法,其中 B 表示实体的开始, I 表示实体的中间, O 代表非实体。以表征动态演化和处置过程为业务需求,将实体分为事故位置、事故类型、事故描述和应急处置等方面,其中应急处置分为处置机构、处置设备、未处置、处置中、处置效果 5 个实体。通过 BIO 标记法分别定义各实体类别的标签,实体详细定义见表 1。

1.2 预训练语言模型

预训练语言模型显著提升了自然语言处理领域任务的当前最优结果(State of the Art, SOTA)表现,其中由谷歌在 2017 年提出的 Transformer 成为最主要的自然语言处理特征提取工具。BERT 模型使用掩码语言模型(Masked Language Model, MLM)和下句预测(Next Sentence Prediction, NSP)实现文本字词特征的预训练。它将多个 Transformer 编码器堆叠提取特征,模型结构如图 1 所示。在文本的开头和句子之间分别插入 [CLS] 和 [SEP],采用字向量、句向量和位置向量的 Embedding 编码之和作为融合的特征向量。在文本序列的特征向量输入 BERT 模型之前,随机抽取 15%的掩码词。对 80%的样本用掩码标记 [MASK] 替换掩码词,对另外 10%的样本不做任何替换,对最后 10%的样本从模

表 1 实体类型定义

Tab. 1 Definition of named entity

类型	序号	实体名称	实体开始 标记	实体中间 标记	文本示例
事故 位置	1	城市	B-CITY	I-CITY	西安市, 西安
	2	高速公路	B-FWAY	I-FWAY	包茂高速, 青银线, 绕城
	3	收费站	B-TS	I-TS	六村堡, 六村堡收费站
	4	服务区	B-SA	I-SA	秦岭, 秦岭服务区
	5	桥梁	B-BG	I-BG	立交桥, 大桥
	6	隧道	B-TN	I-TN	隧道
	7	匝道	B-RM	I-RM	匝道
事故 类型	8	主体类型	B-TYPE1	I-TYPE1	单方事故, 多方事故
	9	事故类型	B-TYPE2	I-TYPE2	刚蹭, 擦挂
事故 描述	10	车辆类型	B-VT	I-VT	半挂, 小车, 危化品车
	11	路产损失	B-RL	I-RL	有路损, 路产待勘验
	12	伤亡人数	B-IDR	I-IDR	1 人死亡 1 人受伤, 驾驶人重伤
	13	影响车道数	B-TD	I-TD	占用应急车道, 占用部分行车道
	14	天气	B-WT	I-WT	小雨, 大雨, 雨
环境 交通 分流	15	分流	B-SP	B-SP	分流
	16	处置机构	B-ROLE	I-ROLE	高速交警, 路政, 养护部门
应急 处置	17	处置设备	B-DVT	I-DVT	吊车, 拖车, 救护车, 清障车
	18	未处置	B-MOT1	I-MOT1	准备施救, 待救援, 正在赶往
	19	处置中	B-MOTI2	I-MOTI2	正在清理, 临时管制
	20	处置效果	B-EFF	I-EFF	已清理, 已拉走救治, 已勘验完毕
非实体	21	非实体	O	O	—

型词表中随机抽取单词来替换掩码词。利用大规模的语料数据进行自监督学习的训练,使得 BERT 模型能够正确预测这些被掩码词,以便获取输入文本数据中的语义信息。其中,[CLS] 标签对应的输出向量记为 C ,其判断 B 句是否为 A 句的下一句。是,则输出 1;否,则输出 0。 T_{Mask} 为预测的掩码词。

假设原始数据序列为 $S = \{S_1, S_2, \dots, S_n\}$, n 为样本字数。叠加并进行掩码操作后得到 BERT 模

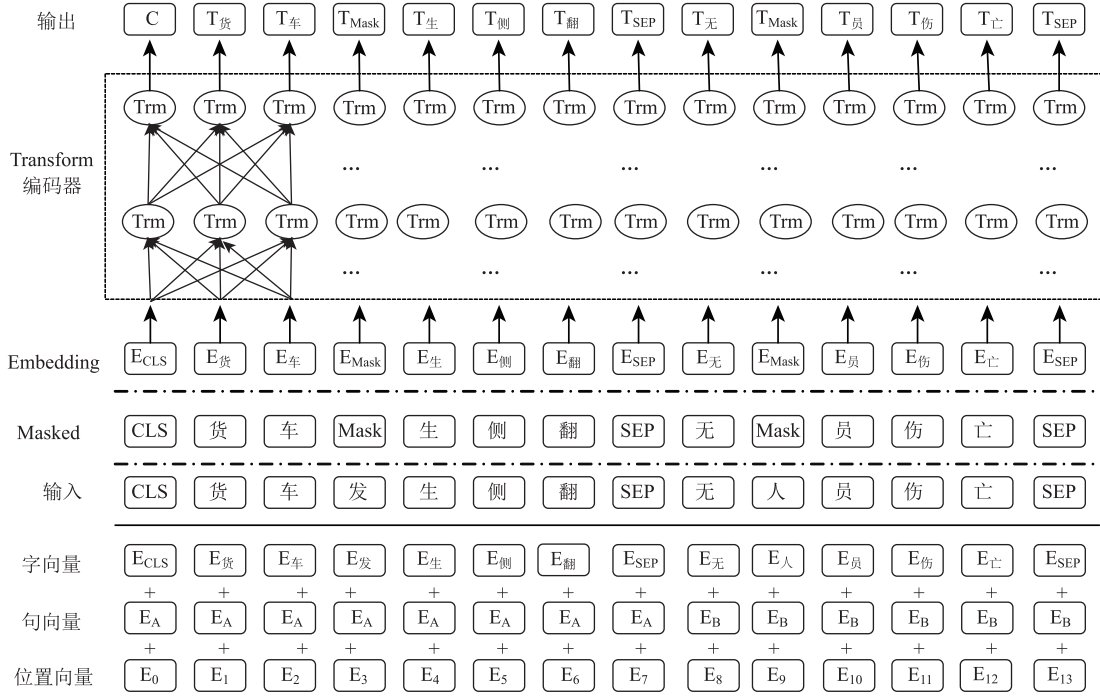


图1 BERT 模型结构

Fig. 1 Model structure of BERT

型输入为 $E = \{E_1, E_2, \dots, E_N\}$, N 为组合后词向量维度。每个 Transformer 编码器由 Self-attention 层和前馈神经网络层组成。Self-attention 是 Transformer 的核心机制。它将输入的词向量通过线性变换得到 3 组向量序列 Q, K, V 。然后使用 Softmax 函数得到特征函数 $Attention(Q, K, V)$ 。其计算过程如下:

$$Q = w_q E, \quad (1)$$

$$K = w_k E, \quad (2)$$

$$V = w_v E, \quad (3)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (4)$$

式中, E 为输入的词向量; w_q, w_k, w_v 分别为训练获得的权重矩阵; d_k 为词向量的维度。

重复多次相同的 Transformer 编码器后得到与 $E = \{E_1, E_2, \dots, E_N\}$ 向量长度相同的输出向量 $T = \{T_1, T_2, \dots, T_N\}$, 即为经过 BERT 模型预训练后的向量。

1.3 BiLSTM-CRF 模型

BiLSTM 是自然语言处理领域应用较好的一种深度学习模型。采用 BiLSTM 模型对预训练后的文本向量 T 进行特征提取, 采用 CRF 模型对网络输出结果进行条件约束, 从而得到合理的实体类型预测结果。BiLSTM-CRF 模型的结构如图 2 所示。BiLSTM 由前向 LSTM 层和后向 LSTM 层组成, 获取输入信息在两

个方向的上下文特征, 然后使用 Softmax 函数给出每个词 T_i 对应每个预定义标签 j 的概率值 p_{ij} 。同时, 在 BiLSTM 网络中加入了 Dropout 层避免过拟合现象。

LSTM 设置 3 种门结构: 遗忘门 (Forget Gate)、输入门 (Input Gate) 与输出门 (Output Gate), 三者利用 sigmoid 激活函数 (记为 σ) 的 0-1 门控特性对经过的信息进行筛选, 通过有用的信息, 阻挡无用的信息。LSTM 通过记忆细胞状态 C_t 来构建长期依赖关系, 通过隐含层状态 h_t 来构建短期依赖关系。输入门 I_t 决定是否忽略输入数据, 遗忘门 F_t 将值向 0 减少, 输出门 O_t 决定是否使用隐含层状态。LSTM 模型表示如下:

$$\text{输入门 } I_t = \sigma(x_t w_{xi} + h_{t-1} w_{hi} + b_i), \quad (5)$$

$$\text{遗忘门 } F_t = \sigma(x_t w_{xf} + h_{t-1} w_{hf} + b_f), \quad (6)$$

$$\text{输出门 } O_t = \sigma(x_t w_{xo} + h_{t-1} w_{ho} + b_o), \quad (7)$$

$$\text{临时单元状态 } \tilde{C}_t = \tanh(x_t w_{xc} + h_{t-1} w_{hc} + b_c), \quad (8)$$

$$\text{当前单元 } C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t, \quad (9)$$

$$\text{隐含层状态 } h_t = O_t \odot \tanh C_t, \quad (10)$$

式中, x_t 为当前输入; h_{t-1} 为前一时刻隐含层输出; $w_{xi}, w_{hi}, w_{xf}, w_{hf}, w_{xo}$ 和 w_{ho} 为权重; b_i, b_f, b_o 和 b_c 为偏差; $\odot$ 为对应元素点乘; $\tanh$ 函数的作用是

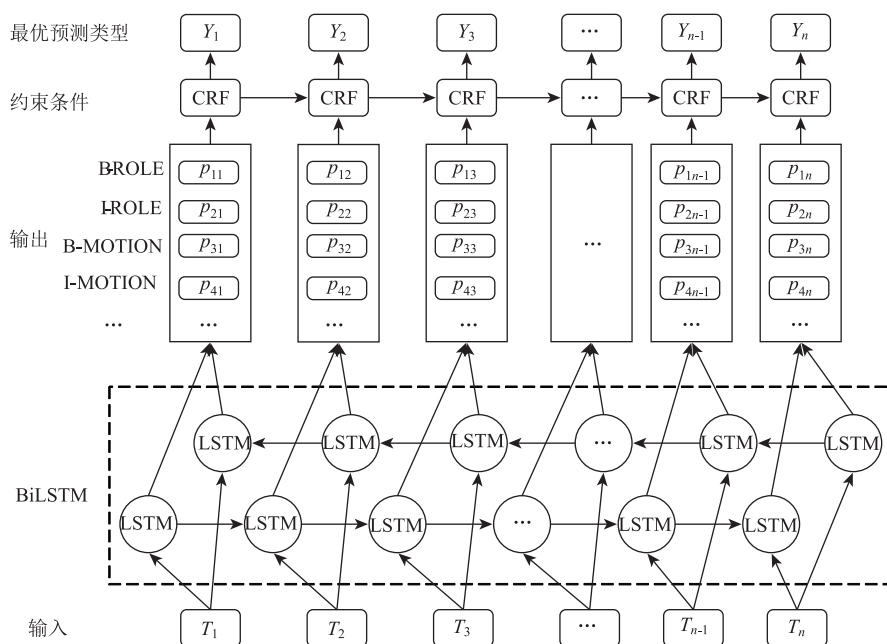


图2 BiLSTM-CRF模型结构

Fig. 2 Model structure of BiLSTM-CRF

保证值 $\in (-1, 1)$ 。

最后将 Bi LSTM 层输出的隐藏层状态 h_i 连接起来, 形成特征向量 H 。将 H 作为 Softmax 函数的输入来预测标签。由于 Softmax 函数的输出值之间是相互独立的, 可能存在不合理序列的情况。因此需要采用 CRF 模型, 对 BiLSTM 的输出进行修正。CRF 模型考虑了各个字标签之间的约束关系, 通过各标签得分与标签之间的转移得分来计算不同标签的出现概率。选出现概率最大的标签序列作为最优的预测结果。

令 $X = \{x_1, x_2, \dots, x_n\}$ 为观测序列, $y = \{y_1, y_2, \dots, y_n\}$ 为与之相应的标记序列, 则构建条件概率模型:

$$P(y|X) = \frac{1}{Z} \exp \left[\sum_j \sum_{i=1}^{n-1} \lambda_j t_j(y_{i+1}, y_i, x, i) + \sum_k \sum_{i=1}^n u_k s_k(y_i, x, i) \right], \quad (11)$$

式中, $t_j(y_{i+1}, y_i, x, i)$ 为定义在观测序列的两个响铃标记位置上的转移特征函数, 用于刻画相邻标记变量之间的相关关系以及观测序列对其影响; $s_k(y_i, x, i)$ 为定义在观测序列的标记位置 i 上的状态特征函数, 用于刻画观测序列对标记变量的影响; λ_j 和 u_k 为参数; Z 为规范化因子, 用以确保式 (11) 是正确定义的概率。

1.4 模型评价指标

对于第 i 个预定义的实体类别, 采用准确率 P_i 、

召回率 R_i 、 $F1_i$ 值作为模型的评价指标。准确率 P_i 描述正例预测准确程度, 召回率 R_i 衡量模型识别真实正例的程度, $F1$ 值是准确率和召回率的调和平均值, $F1 \in (0, 1)$, 越接近 1 模型效果越佳。计算式为:

$$P_i = \frac{TP}{TP + FP}, \quad (12)$$

$$R_i = \frac{TP}{TP + FN}, \quad (13)$$

$$F1_i = \frac{2(P_i \cdot R_i)}{P_i + R_i}, \quad (14)$$

式中, TP 为真正例个数; FP 为假正例个数; FN 为假反例个数。

考虑各实体类别数量可能不均衡, 采用加权的 $F1$ 值作为模型的综合评价指标, 计算式为:

$$F1_w = \sum_{i=1}^n \frac{d_i}{D} F1_i \times 100\%, \quad (15)$$

式中, d_i 为样本中第 i 类标签的数量; n 为训练数据集中标签类别数; D 为训练数据集中标签总数。

2 多模态交通事故信息内容评价

2.1 数据来源及预处理

以陕西省高速公路 2021 年 6 月至 2022 年 8 月的交通事故数据集为数据来源, 其中 2021 年 6 月至 2022 年 5 月为训练集, 2022 年 6—8 月为测试集。陕西省高速公路交通事故数据集中记录了交通事故的

结构化属性,包括事故开始时刻、事故清除时刻、事故类型、受伤人数等,并保存了每次报送的自然语言文本。结合信息内容评价需求,根据数据集中属性的划分提取结构化字段如表2所示。其中,交通事故持续时间定义为系统记录的事故开始时刻与事故清除时刻的时间差^[17]。BIO标注后不同数据集规模如表3所示。

表2 交通事故结构化数据
Tab.2 Structured traffic accident data

类别	字段名称	数据类型	说明
持续时间	duration	定量	持续时间
事故基本信息	vcType	定类	事故类型,分为单方事故和多方事故两类
	Injured	定量	受伤人数
	Death	定量	死亡人数
	weather	定类	天气,分为雾、雪、阴、雨、晴5类
交通分流	loss	定类	有路产损失,1;无路产损失,0
	diversion	定类	采取分流,1;不采取分流,0

表3 样本数据集规模

Tab.3 Sample size

	训练集	测试集
样本数量	27 000	6 951
命名实体数量	329 950	94 169

2.2 信息内容评价指标

信息内容是数据价值的体现,是数据质量评价的前提和关键性问题。以信息内容评价为目标,考虑应急信息质量的特点和需求,从多模态数据的一致性和信息描述的丰富性两个维度,对交通事故信息内容进行评价,指标定义见表4。

表4 交通事故信息内容评价指标

Tab.4 Evaluation indicators of traffic accident information content

一级指标	二级指标	计算来源	指标说明
一致性 A	一致性 A1	Data; Text	数据库内逻辑的一致性以及不同模态数据间的一致性。
丰富性 B	应急处置占比 B1	Text	应急处置实体类型在文本数据中占总体识别实体的比例。
	变异性 B2	Text	被识别出的各类命名实体数量的波动性。

令

$$Data = \begin{Bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{Bmatrix}$$

表示结构化属性, $Text = \{I_1, \dots, I_m\}$ 表示文本信息, 式中, a_{ij} 为第 i 条样本中第 j 个结构化属性; I_i 为第 i 条样本的文本信息; n 为属性个数; m 为样本数。预定义的实体类别为 K , 交通事故实体识别结果为

$$Entity = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1K} \\ k_{21} & k_{22} & \cdots & k_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ k_{m1} & k_{m2} & \cdots & k_{mK} \end{bmatrix},$$

式中, k_{ij} 为第 i 条样本中第 j 个实体类型出现的频数。

(1) 一致性 A1 表征结构化数据与文本信息对相同内容要素记录的一致程度。计算式:

$$A1 = \frac{\sum_{i=1}^m \sum_{j=1}^{K_c} ifcon(a_{ij}, k_{ij})}{m \times K_c} \times 100\%, \quad (16)$$

式中, K_c 为结构化数据记录的实体类型个数, $K_c \in 0, 1, 2, \dots, K$; $ifcon(a_{ij}, k_{ij})$ 为判断结构化数据 a_{ij} 与文本信息实体内容 k_{ij} 是否一致, 若一致则为 1, 否则为 0。若数据为空而导致无法比较时, 记为 0。

(2) 应急处置占比 B1 表征应急处置实体数量在样本实体数量的占比, 计算式:

$$B1 = \frac{\sum_{i=1}^m \sum_{j=1}^{K_a} k_{ij}}{\sum_{i=1}^m \sum_{j=1}^K k_{ij}}, \quad (17)$$

式中 K_a 为实体类型中属于应急处置实体的个数。

(3) 变异性 B2 表示被识别的不同实体类型的分布情况, 量化各要素在数据集中的波动程度。计算式为:

$$B2_j = \frac{\text{std}(\cdot)}{\text{mean}(\cdot)}, \quad (18)$$

式中, $\text{std}(\cdot)$ 和 $\text{mean}(\cdot)$ 分别为各类型实体数量的均值和标准差。

3 实例分析

3.1 交通事故信息识别模型性能分析

ALBERT 和 RoBERTa 是两种 BERT 的改进模型。ALBERT 模型是一种轻量级 BERT, 降低了内存和算力消耗, 提高了训练速度。RoBERTa 与 BERT 相比, 在模型参数量、batch size 和训练数据上规模更大。采用相同的训练集和试验环境, 比较 BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF 和 RoBERTa-BiLSTM-CRF 这 3 种深度学习模型的识别结果。以式 (15)

的加权 $F1$ 值评价每个模型的性能,保存多次训练中性能较优的模型作为最终识别模型。

SET0 表示原始的测试集,SET1~SET5 表示相同样本量、有放回抽样的测试集。同时,以陕西交通

微博为语料来源,采用网络爬虫随机获取 2 000 条文本信息作为袋外测试集,分析模型对“新”数据的识别效果,检验模型的鲁棒性。各深度学习模型的性能如表 5 所示。

表 5 测试集和袋外测试集的模型性能比较

Tab. 5 Model performance of test dataset v. s. out-of-pocket test dataset

测试集 编号	测试集			袋外测试集		
	BERT-BiLSTM-	RoBERTa-BiLSTM-	ALBERT-BiLSTM-	BERT-BiLSTM-	RoBERTa-BiLSTM-	ALBERT-BiLSTM-
	CRF	CRF	CRF	CRF	CRF	CRF
SET 0	96.989 6	97.115 4	76.919 4	71.989 8	71.799 1	56.365 2
SET 1	96.797 8	96.917 8	77.008 7	65.991 1	65.980 6	54.982 5
SET 2	97.186 7	97.330 9	77.089 2	72.057 9	72.010 9	55.580 0
SET 3	96.885 6	97.180 4	77.512 9	66.601 1	66.504 5	52.867 9
SET 4	97.026 3	97.155 9	77.038 6	66.255 2	66.188 3	56.862 5
SET 5	97.290 6	97.339 0	76.874 6	72.038 0	71.873 0	56.870 2
均值	97.029 4	97.173 2	77.073 9	69.155 5	69.059 4	55.588 0

由表 5 可知,测试集上 RoBERTa-BiLSTM-CRF 的识别效果最佳。同时,袋外测试集上 BERT-BiLSTM-CRF 效果最佳,说明其对不同数据的适应性最好。BERT,ALBERT 和 RoBERTa 每次测试集训练时间约 10,5 和 20 h。无论测试集还是袋外测试集上,采用 BERT 和 RoBERTa 预训练模型的效果极为接近,采用 ALBERT 的模型性能均最差。综合考虑模型精度和训练时长,BERT-BiLSTM-CRF 是一个在识别准确度、学习效率和鲁棒性较优的深度学习模型,以此作为本研究的交通事故命名实体识别模型。

3.2 交通事故应急处置实体识别结果分析

利用 BERT-Bi LSTM-CRF 模型对测试集进行交通事故信息识别,应急处置实体的识别结果示例见表 6。计算持续时间和处置机构、处置设备、未处置、处置中、处置效果的 Spearman 相关系数,依次为 0.309,0.151,0.137,0.220 和 0.178。结果表明,应急处置实体数量与持续时间之间存在正相关性。

表 6 应急处置实体识别结果示例

Tab. 6 Example of emergency disposal identified entity

处置机构	处理设备	未处置	处置中	效果
分公司	消防	待救援	临时交通管制,	正常通行,
			交通管制,救援	已到现场
路政、高速	拖车	—	预计 1 h 处理完	
交警			毕,预计 1 h,	已到现场
			等待	
分公司、交	吊车,	—	进行救援,救援	已经到位
警、路产巡	拖车			
查员、巡查				
交警	救援车	—	预计 1 h,救援	暂时封闭

3.3 交通事故信息内容的评价结果与分析

以测试集 6 916 条样本为例,选取天气、路产损失、交通分流、事故类型和伤亡情况作为多模态数据一致性评价内容,评价指标计算结果如表 7 所示。

表 7 交通事故信息内容评价指标结果

Tab. 7 Indicator result of traffic accident information

content evaluation

	一致性 A1/%				应急处置占比 B1		变异性 B2	
	路产 损失	交通 分流	事故 类型	伤亡 情况	平均 值	标准 差	平均 值	标准 差
天气	7.06	45.79	1.59	67.65	47.59	0.360	0.233	1.305
	0.292							

由表 7 可知,结构化数据和文本数据的一致性较低,且不同信息内容之间波动性较大,说明两种模态数据对同一交通事故的信息内容描述不一致问题显著,需要在运营管理中加强工作人员对外部环境和控制措施的记录质量,尤其是天气和交通分流的信息采集质量。同时,应急处置实体占比达到 36%,说明该数据集中以自然语言记录的文本信息具有较高的内涵,是结构化数据的有效补充,对应急处置信息命名实体识别具有实际意义。综上可知,交通事故数据集中不同模态的信息一致性尚有较大提升空间,文本信息蕴含丰富的应急处置信息,规范信息记录质量标准具有重要作用。

4 结论

建立了一种高速公路交通事故命名实体识别模型,是自然语言处理技术在交通事故文本信息抽取的新方法,弥补了现阶段应急处置信息识别的不足。

比较了 BERT-BiLSTM-CRF, ALBERT-Bi LSTM-CRF 和 RoBERTa-BiLSTM-CRF 这 3 种深度学习模型, BERT 和 RoBERTa 预训练模型的效果最佳, 且 BERT 训练时间较 RoBERTa 短。结果表明, BERT-BiLSTM-CRF 模型在识别效果、训练速度和鲁棒性方面具有明显优势。

从一致性和丰富性两个维度提出了交通事故信息内容的评价指标, 提供了一种多模态的信息内容量化评价方法, 为解决文本信息和结构化数据一致性提供了技术路径, 并实证了文本信息的分析价值。研究成果可以量化评估多模态数据采集质量, 进而提取真实完整的应急处置流程, 为降低交通事故持续时间提供改进依据。

信息质量评估是提高信息资源管理效率的关键, 有利于提高信息价值, 促进应急安全管理的发展。本研究主要从多模态数据的一致性进行评价, 后续需要从信息效用、传播过程等角度深入探讨交通应急处置中信息报送的质量和有效性。

参考文献:

References:

- [1] JI K K, LI Z Z, CHEN J, et al. Freeway Accident Duration Prediction Based on Social Network Information [J]. Neural Network World, 2022, 32 (2): 93-112.
- [2] RAKHMAWATIN N A, AWWAB Y, NAJIB A C, et al. Ontology-based Traffic Accident Information Extraction on Twitter in Indonesia [J]. Inteligencia Artificial, 2022, 25 (70): 1-12.
- [3] ZHANG L Y, ZHANG M, TANG J Z, et al. Analysis of Traffic Accident Based on Knowledge Graph [J]. Journal of Advanced Transportation, 2022: 3915467.
- [4] LI J, SUN A, HAN J, et al. A Survey on Deep Learning for Named Entity Recognition [J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34 (1): 50-70.
- [5] 王颖洁, 张程烨, 白凤波, 等. 中文命名实体识别研究综述 [J]. 计算机科学与探索, 2023, 17 (2): 324-341.
- [6] WANG Ying-jie, ZHANG Cheng-ye, BAI Feng-bo, et al. Review of Chinese Named Entity Recognition Research [J]. Journal of Frontiers of Computer Science and Technology, 2023, 17 (2): 324-341.
- [7] 童昭, 王露笛, 朱小杰, 等. 基于预训练模型的军事领域命名实体识别研究 [J]. 数据与计算发展前沿, 2022, 4 (5): 120-128.
- [8] TONG Zhao, WANG Lu-di, ZHU Xiao-jie, et al. Research on Military Domain Named Entity Recognition Based on Pre-training Model [J]. Frontiers of Data and Computing, 2022, 4 (5): 120-128.
- [9] 孙安亮, 时宏伟, 王金策. 基于字符与单词嵌入的航空安全命名实体识别 [J]. 计算机技术与发展, 2022, 32 (9): 148-153.
- [10] SUN An-liang, SHI Hong-wei, WANG Jin-ce. Named Entity Recognition Based on Character and Word Embedding in Aviation Safety [J]. Computer Technology and Development, 2022, 32 (9): 148-153.
- [11] CHANG J, HAN X H. Multi-level Context Features Extraction for Named Entity Recognition [J]. Computer Speech & Language, 2023, 77: 101412.
- [12] 刘兴丽, 范俊杰, 马海群. 面向小样本命名实体识别的数据增强算法改进策略研究 [J]. 数据分析与知识发现, 2022, 6 (10): 128-141.
- [13] LIU Xing-li, FAN Jun-jie, MA Hai-qun. Improvement of Data Augment Algorithm for Named Entity Recognition with Small Samples [J]. Data Analysis and Knowledge Discovery, 2022, 6 (10): 128-141.
- [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.
- [15] ZHANG Peng-xiang. Information Extraction Method for Railway Equipment Accidents Based on Multi-dimensional Character Feature Representation [J]. China Safety Science Journal, 2022, 32 (6): 109-114.
- [16] 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.
- [17] HAN Guang, BU Tong, WANG Ming-ming, et al. Text Classification of Railway Traffic Accidents Based on Dual-channel Bidirectional Long Short Term Memory Network [J]. Journal of the China Railway Society, 2021, 43 (9): 71-79.
- [18] 王向前, 李敏敏, 孟祥瑞. 基于 ALBERT-BiLSTM-CRF 的煤矿事故案例文本命名实体识别方法 [J]. 阜阳师范大学学报 (自然科学版), 2022, 39 (3): 56-64.
- [19] WANG Xiang-qian, LI Min-min, MENG Xiang-rui. Named Entity Recognition Method of Coal Mine Accident Case Text Based on ALBERT-BiLSTM-CRF [J]. Journal of Fuyang Normal University (Natural Science), 2022, 39 (3): 56-64.
- [20] 刘婷, 张社荣, 王超, 等. 水利施工事故文本智能分析的 BERT-BiLSTM 混合模型 [J]. 水力发电学报, 2022, 41 (7): 1-12.
- [21] LIU Ting, ZHANG She-rong, WANG Chao, et al. Text Intelligent Analysis for Hydraulic Construction Accidents Based on BERT-BiLSTM Hybrid Model [J]. Journal of Hydroelectric Engineering, 2022, 41 (7): 1-12.
- [22] 刘昭, 何赏璐, 刘英舜. 基于社交网络数据的交通突发事件识别方法 [J]. 交通信息与安全, 2021, 39 (2): 53-60.
- [23] LIU Zhao, HE Shang-lu, LIU Ying-shun. A Method to Identify Traffic Incidents Based on Social Network Data [J]. Journal of Transport Information and Safety, 2021, 39 (2): 53-60.

(下转第 213 页)

- Road Traffic Event Detection Algorithm Based on CNN [J]. *Journal of Highway and Transportation Research and Development*, 2022, 39 (3): 176-182.
- [13] 张雄涛, 郑景玉, 申倩, 等. 基于混合图卷积的多通道时空交通流预测模型 [J]. *电信科学*, 2023, 39 (9): 97-110.
- ZHANG Xiong-tao, ZHENG Jing-yu, SHEN Qing, et al. Multi-channel Spatial-temporal Traffic Flow Prediction Based on Hybrid Static-dynamic Graph Convolution [J]. *Telecommunications Science*, 2023, 39 (9): 97-110.
- [14] 贺文武, 裴博彧, 李雅婷, 等. 基于双向自适应门控图卷积网络的交通流预测 [J]. *交通运输系统工程与信息*, 2023, 23 (1): 187-197.
- HE Wen-wu, PEI Bo-yu, LI Ya-ting, et al. Traffic Flow Forecasting Based on Bi-direction Adaptive Gating Graph Convolutional Networks [J]. *Journal of Transportation Systems Engineering and Information Technology*, 2023, 23 (1): 187-197.
- [15] ZHAO L, SONG Y J, ZHANG C, et al. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2019, 21 (9): 3848-3858.
- [16] ZHU J, WANG Q, TAO C, et al. AST-GCN: Attribute-augmented Spatiotemporal Graph Convolutional Network for Traffic Forecasting [J]. *IEEE Access*, 2021, 9: 35973-35983.
- [17] LIAO B B, ZHANG J Q, WU C, et al. Deep Sequence Learning with Auxiliary Information for Traffic Prediction [C] // *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. London: Association for Computing Machinery, 2018, 537-546.
- [18] ZHANG D, KABUKA M R. Combining Weather Condition Data to Predict Traffic Flow: A GRU-based Deep Learning Approach [J]. *IET Intelligent Transport Systems*, 2018, 12 (7): 578-585.
- [19] YU B, YIN H T, ZHU Z X. Spatio-temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting [C] // *Proceedings of the Twenty Seventh International Joint Conference on Artificial Intelligence*. Stockholm: IJCAI, 2018: 3634-3640.
- [20] VLAHOIANNI E I, KARLAFTIS M G, GOLIAS J C. Short-term Traffic Forecasting: Where We Are and Where We're Going [J]. *Transportation Research Part C: Emerging Technologies*, 2014, 43: 3-19.
- [21] LI Y G, YU R, SHAHABI C, et al. Diffusion Convolutional Recurrent Neural Network: Data-driven Traffic Forecasting [C] // *ICLR 2018*. Vancouver: arXiv, 2018.
- [22] YUAN H, LI G. A Survey of Traffic Prediction: From Spatio-temporal Data to Intelligent Transportation [J]. *Data Science and Engineering*, 2021, 6 (1): 63-85.
- (上接第193页)
- [15] 樊海玮, 秦佳杰, 孙欢, 等. 基于BERT与BiGRU-CRF的交通事故文本信息提取模型 [J]. *计算机与现代化*, 2022 (5): 10-15.
- FAN Hai-wei, QIN Jia-jie, SUN Huan, et al. Traffic Accident Text Information Extraction Model Based on BERT and BiGRU-CRF Fusion [J]. *Computer and Modernization*, 2022 (5): 10-15.
- [16] 李昀轩, 李萌, 陆建, 等. 基于多任务迁移学习的交通警情信息自动处理方法 [J]. *中国公路学报*, 2022, 35 (9): 1-12.
- LI Yun-xuan, LI Meng, LU Jian, et al. An Auto-processing Method of Traffic Safety Information Based on a Multi-task Transfer Learning Algorithm [J]. *China Journal of Highway and Transport*, 2022, 35 (9): 1-12.
- [17] 李硕, 马玉坤, 韩晖, 等. 山区高速公路货车事故严重程度致因及随机参数分析 [J]. *公路交通科技*, 2023, 40 (4): 228-236.
- LI Shou, MA Yu-kun, HAN Hui, et al. Analysis on Causes and Random Parameters of Truck Accident Severity in Mountainous Expressway [J]. *Journal of Highway and Transportation Research and Development*, 2023, 40 (4): 228-236.
- [18] 陈娇娜, 靳引利, 陶伟俊, 等. 处置阶段信息对高速公路交通事故持续时间的调节效应分析 [J]. *安全与环境学报*, 2023, 23 (4): 1169-1177.
- CHEN Jiao-na, JIN Yin-li, TAO Wei-jun, et al. Moderating Effect Analysis of Information in the Processing Phase on Expressway Traffic Accident Duration [J]. *Journal of Safety and Environment*, 2023, 23 (4): 1169-1177.
- [19] 朱益平, 杜海娇, 张佳, 等. 基于RS-BP神经网络的政务微信公众号信息质量评价模型研究 [J]. *情报科学*, 2021, 39 (2): 54-61.
- ZHU Yi-ping, DU Hai-jiao, ZHANG Jia, et al. Information Quality Evaluation Model Research of Government Wechat Public Account Based on RS-BP Neural Networks [J]. *Information Science*, 2021, 39 (2): 54-61.
- [20] 徐文强, 刘春年, 周涛. 大数据环境下应急信息质量评估体系研究 [J]. *图书情报工作*, 2020, 64 (2): 50-58.
- XU Wen-qiang, LIU Chun-nian, ZHOU Tao. Research on Emergency Information Quality Evaluation System in Big Data Environment [J]. *Library and Information Service*, 2020, 64 (2): 50-58.
- [21] GUO W Y, ZENG Q T, DUAN H, et al. Text Quality Analysis of Emergency Response Plans [J]. *IEEE Access*, 2020, 8: 9441-9456.