

列控车载设备故障知识图谱的重叠三元组联合抽取方法

张凝, 张振海

(兰州交通大学自动化与电气工程学院, 兰州 730070)

摘要:【目的】车载设备作为列控系统的核心,利用知识图谱研究故障案例,实现对其故障原因及处理方案的快速准确定位,对铁路行车安全和运输效率至关重要。【方法】针对其中关键步骤知识抽取采用传统流水线方法导致错误积累以及交互性差等问题,提出一种基于序列标注策略的重叠三元组联合抽取方法。首先,依据定义的实体与关系类型,采用“位置-实体-关系-角色”的序列标注策略,实现重叠三元组联合标注;其次,构建 FMBCR (FaultRoBERTa-MHSA-BiLSTM-CRF incorporating Regularization Methods) 模型准确抽取列控车载设备故障文本的全局语义信息,通过增加一种基于正则化模块 READ (Regularization Method with Adversarial Training and Dropout) 的 KL (Kullback-Leibler) 发散损失算法优化模型的预测性能;最后,以列控车载设备故障数据为例进行实验分析与知识图谱可视化构建。【结果】通过充分的对比实验以及消融实验,结果表明,所提方法实现小样本重叠三元组的联合抽取,且抽取性能较传统方法有显著提升,FMBCR 模型较现在广泛使用的 CASREL 联合抽取模型,准确率、召回率和 F1 值分别提升 3.0%、4.3% 和 3.7%。【结论】研究结果可为列控车载设备故障维修智能信息检索和辅助决策等高级应用提供支持。

关键词:列控系统; 车载设备故障; 知识图谱; 联合抽取; 重叠三元组; 序列标注

中图分类号:U284.95; U284.92 **文献标识码:**A **DOI:**10.13238/j.issn.1004-2954.202407180006

Joint Extraction Method for Overlapping Triples in Fault Knowledge Graph of CTCS On-Board Equipment

ZHANG Ning, ZHANG Zhenhai

(School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: [Objective] As the core of the Chinese Train Control System (CTCS), on-board equipment utilizes the knowledge graph to analyze fault cases and realize rapid and accurate identification of fault causes and treatment solutions, which is crucial for railway safety and transportation efficiency. [Methods] To address issues such as error accumulation and poor interactivity caused by traditional pipeline methods in knowledge extraction, a joint extraction method for overlapping triples based on a sequence labeling strategy was proposed. Firstly, based on the defined entity and relationship types, a “position-entity-relationship-role” sequence labeling strategy was employed to achieve the joint labeling of overlapping triples. Secondly, the FMBCR (FaultRoBERTa-MHSA-BiLSTM-CRF incorporating regularization methods) model was constructed to accurately extract the global semantic information from fault texts of CTCS on-board equipment, and its predictive performance was optimized by incorporating a KL (Kullback-Leibler) divergence loss algorithm based on the regularization module READ (Regularization method with adversarial training and dropout). Finally, experimental analysis and

knowledge graph visualization construction were conducted, using fault data of CTCS on-board equipment as an example. [Results] Through sufficient comparative experiments and ablation experiments, the results showed that the proposed method realized the joint extraction of overlapping

收稿日期:2024-07-18; 修回日期:2024-09-09; 网络首发日期:2025-03-06

基金项目:国家自然科学基金项目(61763025); 甘肃省重点研发计划项目(2YF7GA141)

作者简介:张凝(2000—),女,硕士研究生,主要从事知识图谱和故障诊断研究工作,E-mail:1026054855@qq.com。

通信作者:张振海(1983—),男,教授,2014年毕业于兰州交通大学交通信息工程及控制专业,工学博士,主要从事交通信息工程及控制和图像处理研究工作,E-mail:764411629@qq.com。

triples under small-sample conditions, and the extraction performance was significantly improved compared with traditional methods. Compared with the widely used CASREL joint extraction model, the FMBCR model improved the precision, recall, and $F1$ score by 3.0%, 4.3%, and 3.7%, respectively.

[**Conclusion**] The research findings provide support for advanced applications such as intelligent information retrieval and decision-making assistance for fault maintenance of CTCS on-board equipment.

Key words: Chinese Train Control System; on-board equipment fault; knowledge graph; joint extraction; overlapping triples; sequence labeling

引言

车载设备作为列控系统的关键组成部分,其故障频次和故障处理时间直接影响列车运行。为降低故障发生频次并缩短故障处理时间,利用智能化手段对蕴含丰富诊断和处理经验的车载设备故障维修日志进行研究,辅助维修人员精准把握故障情况和维修策略,具有重要的现实意义。知识图谱技术作为人工智能领域的研究热点,通过知识抽取自动从文本中提取实体及其关系,形成知识三元组,将复杂的故障文本转化为易于共享和集成的结构化表达形式,进而建立链接数据,帮助挖掘故障信息的内在联系。因此,构建知识图谱成为辅助故障诊断的重要手段^[1]。

知识抽取作为构建知识图谱的核心,主要依赖实体抽取和关系抽取两个关键子任务。传统的流水线方法将两个任务视为独立环节,早期通过基于规则的方法实现^[2],存在严重的特征依赖问题;随后,基于统计模型的方法在实体抽取子任务中得到应用,薛莲等^[3]利用 Bigram 模型抽取故障实体,建立列控车载设备故障知识图谱的框架。近年来,深度学习技术在图像分类和模式识别中的应用为自动提取文本信息提供新的思路^[4-5]。刘丹等^[6]采用 Graph Attention-CRF 神经网络模型,提升实体抽取的准确率;万莎^[7]引入 CNN 模型,实现限定域关系抽取;HIDALGO 等^[8]提出 Biased PromptORE 方法微调 Transformer 模型,以从专业文件中提取复杂语义关系。尽管流水线方法在灵活性和独立性方面具有优势,但存在以下问题,缺乏子任务间的交互、实体抽取错误可能传播到关系抽取阶段、部分实体可能无有效关系,造成冗余信息^[9]。

为克服流水线方法的缺陷,研究者提出实体关系联合抽取模型^[10]。MIWA 等^[11]首次提出基于端到端的树形结构模型,通过共享 LSTM 编码层序列参数,增强子任务间的交互性;ZENG 等^[12]提出 CopyMTL 联合模型通过复制机制提取实体和关系,实现多标记实体预测的同时克服以往方法对解析器性能的依赖;李代祎等^[13]将关系建模为头实体映射至尾实体的函数,解决三元组重叠问题;WEI 等^[14]提出一种级联二进制关

系抽取框架 CASREL,取得优异的关系抽取效果。上述模型通过共享编码层对实体识别和关系抽取任务进行联合训练,解决错误传播问题,但解码层仍采用独立算法,无法通过完全联合解决冗余信息问题^[15]。

针对上述问题,基于序列标注的联合抽取方法被提出。ZHENG 等^[16]提出 Novel Tagging 标记方案,通过同步标注每个单词的实体提及和关系类型,增强实体抽取与关系抽取的交互性,但无法处理三元组重叠问题;吴赛赛等^[17]提出“ME+R+BIESO”标注方案,通过标签匹配和映射解决单实体重叠问题;NAYAK 等^[18]基于 Encoder-Decoder 架构,通过指针网络的解码方式识别多标签对象,生成全部三元组;朱秀宝等^[19]将多标签对象的标签预测转化为多标签分类任务,但该模型需要海量训练数据的支撑,在面向小数据抽取时,效果欠佳。

上述基于深度学习的联合抽取方法应用于小规模数据集时容易出现过拟合问题,大多数研究工作采用丢弃法(Dropout)缓解^[20]。此外,ILYAS 等^[21]证明对抗性样本归因于非稳健特征的存在,这些特征具有高度预测性;束嘉伟等^[22]将对抗训练扩展到电力知识图谱构建任务,通过添加噪声扰动来增加样本的多样性。

综上,考虑到现有的列控车载设备故障知识抽取均采用传统的流水线方法,且关于小数据重叠三元组联合抽取的研究较少,提出一种结合序列标注框架和联合抽取模型的三元组抽取方法,主要工作如下。

(1)设计一种联合序列标注策略“位置-实体-关系-角色”,打破传统标注框架中实体与关系单一对应的局限,使单个实体能够同时参与并隶属于多种关系的描述。

(2)构建列控车载设备故障领域 FaultRoBERTa 预训练模型,并设计一个正则化 READ 模块,该模块首先在嵌入层引入 FGM(Fast Gradient Method)对抗训练,其次最小化相同样本的对抗性输出与原始输出分布之间的双向 KL 散度,在解决过拟合问题的同时增强模型面向小数据抽取的稳健性。

(3)将 FMBCR 联合抽取模型在列控车载设备故障数据集上进行测试,实验结果表明,该方法实现重叠

三元组抽取的同时,准确率较 CASREL 算法提高 3.0%。将抽取得到的知识三元组存储在 Neo4j 图数据库中,搭建出列控车载设备故障知识图谱。

1 列控车载设备故障数据分析及标注

1.1 数据特征分析

作为列车运行控制系统中的重要组成部分,车载设备实时从其他列车控制模块和地面设备中收集列车信

息参数,精准地生成动态速度监控曲线,以确保列车的平稳和安全运行。我国铁路线路情况复杂,所使用的车载设备类型主要有 CTCS3-300T、CTCS3-300S、CTCS3-300H、CTCS2-200C 和 CTCS2-200H 5 种。选取在我国铁路中极具代表性的 CTCS3-300T 型车载设备为研究对象,其故障信息来源于列车超速防护单元(ATPCU)、人机界面(DMI)、C2 控制单元(C2CU)、应答器信息接收单元(BTM)等模块^[20],典型故障模式见表 1。

表 1 列控车载设备典型故障模式
Tab.1 Typical fault patterns of CTCS on-board equipment

车载设备型号	故障编号	故障类型	故障现象	故障原因
CTCS3-300T	1	ATPCU 软件故障	DMI 与主机通信中断	A 核心代码模式转换无效
	2	ATPCU 硬件故障	B 系无法启动	ATPCU 模块 SPL 初始化失败
	3	BTM 软件故障	ATP 无法正常启动	BTM 端口无效

通过分析故障维修日志,发现现场记录的车载设备故障数据文本形式各异,术语专业性强、数据分布不均衡且标注数据有限。因此,为更有效地进行数据挖掘和管理,需要借助知识图谱智能技术来分析和处理这些包含丰富经验信息的故障文本,从而辅助现场人员的维修作业。知识图谱构建流程如图 1 所示,首先分析领域文本的数据特征,并采用高效的序列标注策略进行文本标注,将标注后的数据通过实体关系联合抽取模型进行知识抽取;其次,引入知识融合环节进行实体对齐和实体消歧处理,并通过知识加工环节对知识概念进行建模;然后由领域专家对抽取获得的知识元素进行质量评估,量化知识的可信度,通过舍弃置信度较低的知识,保障知识库的质量。最后,将可信度高的知识在 Neo4j 图数据库中进行存储与表示,完成列控车载设备故障知识图谱的构建。

1.2 联合标注策略

故障维修日志记录了列控车载设备的故障现象、故障原因分析和维修措施等详情,其中实体关系交织复杂。为了有效地从故障维修日志中抽取知识并构建列控车载设备故障知识图谱,需要一种能够识别和表示复杂三元组关系的抽取方法。因此,提出一种序列标注策略,该策略通过增加关系类别并丰富实体角色,克服传统序列标注中的重叠关系抽取难题。

提出的标签体系由 4 个部分组成:“位置-实体-关系-角色”。首先采用“ BIO”标签方案来表示实体中词的位置边界信息,详细含义见表 2。

表 2 位置标签含义
Tab.2 Meanings of position labels

位置标签	标签含义
B	实体类型的开始位置
I	实体类型的中间或结束位置
O	非实体类型

其次,在位置标签的基础上,引入实体类型和关系类别标签,以及针对重叠关系的特殊标签“FO”,以便更准确地描述文本中实体间的关系,详细内容见表 3。

表 3 实体类型和关系类别标签

Tab.3 Entity types and relation category labels		
标签类型	标签内容	标签含义
实体类型	TYPE	故障类型
	PHE	故障现象
	REA	故障原因
	MEA	处理措施
	FR	原因关系
关系类别	FT	分类关系
	FM	处理关系
	FO	重叠关系

最后,增添实体角色标签,形成三元组的有向性表示。标签“H”表示该实体作为三元组的头实体

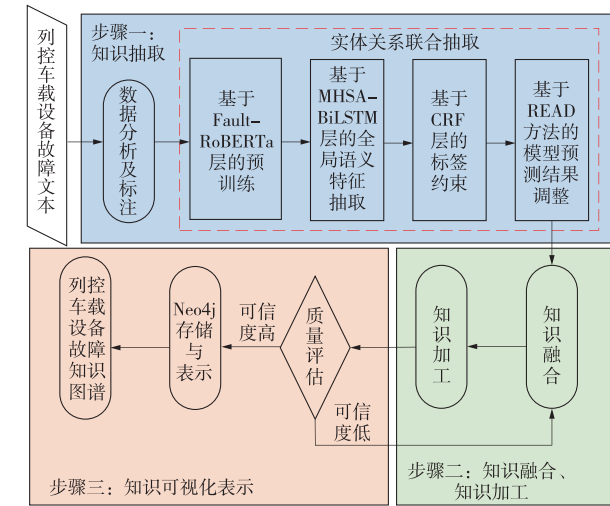


图 1 列控车载设备故障知识图谱构建流程

Fig.1 Construction process of fault knowledge graph for CTCS on-board equipment

(Head), 而“T”则表示三元组的尾实体(Tail)。针对故障文本中长句导致的重叠三元组情况, 采用组合标签表示。例如, “HT”表示该实体同时作为前 1 个三元

组的头实体和后 1 个三元组的尾实体(图 2(a))。另外, “TH”情况如图 2(b)所示, 其含义与“HT”相反。



图 2 联合标注角色标签示例

Fig. 2 Examples of role labels for joint labeling

选取车载设备故障维修日志某一语料, 其序列标注策略如图 3 所示, 实体“JRU 记录事件内容过长”存在与原因关系“FR”和分类关系“FT”多种关系相关联的复杂情况, 即实体参与重叠关系。因此使用“FO”重

叠关系标签来标识“JRU 记录事件内容过长”(B/I-REA-FO-TH)。其中, “TH”组合标签清晰地表明该实体作为前 1 个三元组的尾实体和后 1 个三元组的头实体。

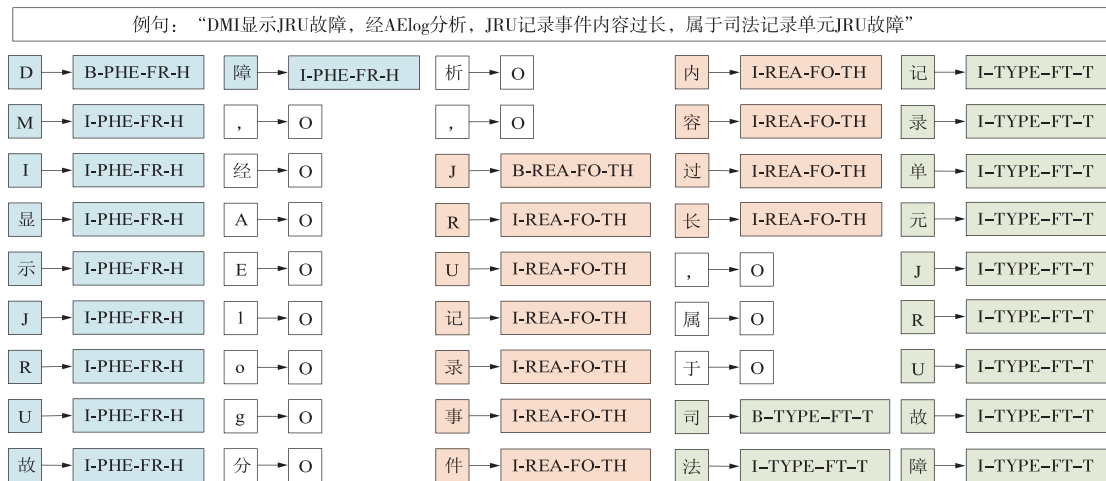


图 3 标签策略示例

Fig. 3 Example of labeling strategy

2 列控车载设备故障知识抽取模型构建

基于列控车载设备故障文本特征, 提出一种结合正则化方法的列控车载设备故障知识联合抽取模型, 命名为 FMBCR。该模型架构如图 4 所示, 首先, 利用改进交叉熵损失函数的 FaultRoBERTa 预训练模型捕获故障文本的词向量表示; 其次, 采用双向长短期记忆网络 (Bi-Directional Long Short-Term Memory, BiLSTM) 提取文本的上下文特征, 融合多头自注意力机制 (Multi-Head Self-Attention, MHSA) 获得关键信息特征, 将二者的输出进行特征融合, 从而发挥互补作用, 获得更全面的语义信息; 接着通过条件随机场 (Conditional Random Field, CRF) 解码输出故障文本序列的标签预测结果; 最后, 在训练 FMBCR 模型时, 引入一种基于 READ 的外部 KL 发散损失算法, 以进一步规范模型的预测。该算法旨在通过最小化同样样本的对抗

性输出和原始输出分布之间的双向 KL 散度来优化模型, 从而获得更可靠的列控车载设备故障知识三元组。

2.1 基于 FaultRoBERTa 的多维词嵌入

与传统的预训练模型 BERT (Bidirectional Encoder Representations from Transformers) 相比, RoBERTa (Robustly Optimized BERT Pretraining Approach) 模型采用全词掩码技术和动态掩码策略, 能够更充分地捕捉列控车载设备故障文本中词汇的信息和上下文词汇之间的关联特征, 因此在列控车载设备故障知识抽取任务中具有更好的适用性。

首先, 针对列控车载设备故障文本的复杂性和专业性, 通过故障报告记录、领域专业书籍以及领域相关文献对 RoBERTa 模块进行预训练以实现领域自适应, 构建专属于列控车载设备故障领域的 FaultRoBERTa 模型, 该模型架构如图 5 所示。通过 Embedding 层将输入的列控车载设备故障文本序列 $\mathbf{W} = \{w_1, w_2, \dots,$

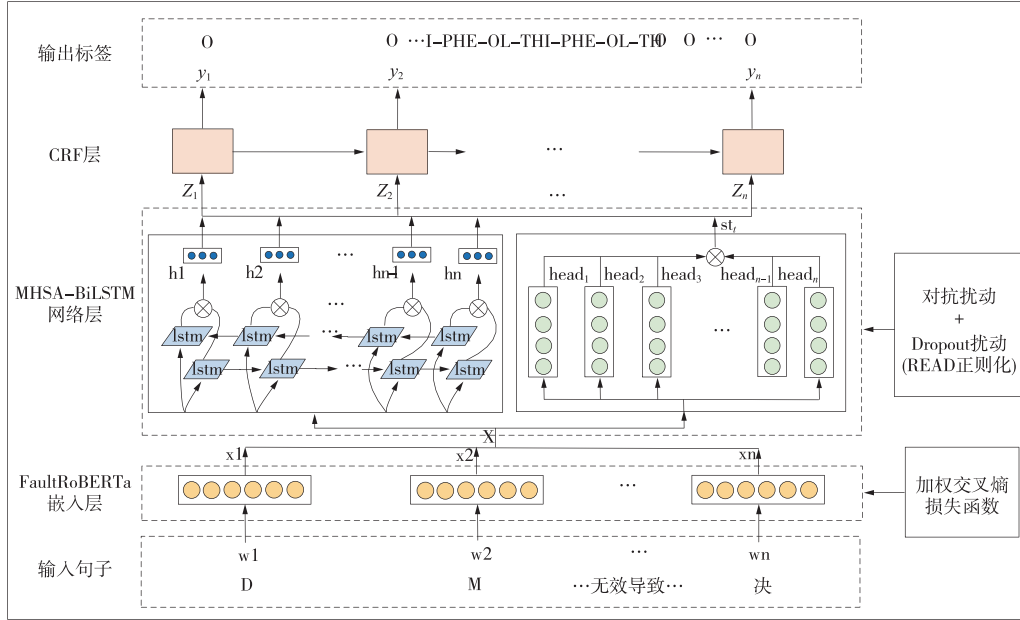


图4 FMBCR 模型架构

Fig. 4 Architecture of FMBCR model

w_n 转换为由字向量、句向量和位置向量叠加生成的向量序列。将这些向量输入模型编码层中,通过 12 层 Transformer Block 间的交互,计算句子中不同词之间的关联性,得到能够体现上下文语义的向量序列 $X = [x_1, x_2, \dots, x_n]$ 。

(2) 所示。其中 w 为权重系数,取样本在总体中占比的倒数, y_i 为样本的真实分布, p_i 为预测分布。

$$\text{loss} = - \sum_{i=1} y_i \times \ln(p_i) \quad (1)$$

$$\text{loss}^* = - \sum_{i=1} w \times y_i \ln(p_i) \quad (2)$$

2.2 基于 MHSA-BiLSTM 的全局特征提取

(1) BiLSTM 模块

长短期记忆网络 (Long Short-Term Memory, LSTM) 能够有效处理时间序列数据,缓解梯度消失、梯度爆炸问题,但 LSTM 存在特征权重随时刻推移而递减的缺陷^[22]。为了解决 LSTM 网络在不同时刻权重分配不均的问题,引入 BiLSTM 网络执行当前的三元组抽取任务。

$$i_t = \sigma(W_i \cdot [\vec{h}_{t-1}, x_t] + b_i) \quad (3)$$

$$f_t = \sigma(W_f \cdot [\vec{h}_{t-1}, x_t] + b_f) \quad (4)$$

$$\tilde{C}_t = \tanh(W_c \cdot [\vec{h}_{t-1}, x_t] + b_c) \quad (5)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (6)$$

$$O_t = \sigma(W_o \cdot [\vec{h}_{t-1}, x_t] + b_o) \quad (7)$$

$$\vec{h}_t = O_t \times \tanh(C_t) \quad (8)$$

式中, i_t 、 f_t 和 O_t 分别为输入门、遗忘门和输出门; b 为偏置项; W 为权重矩阵; C 为记忆单元; σ 为 sigmoid 激活函数。对于输入序列的每个词汇向量,由前向 LSTM 层和后向 LSTM 层根据上下文信息进行编码,得到 $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 。最后将前后向 LSTM 状态进行拼接,作为车载设备故障词汇 t 的向量编码,即

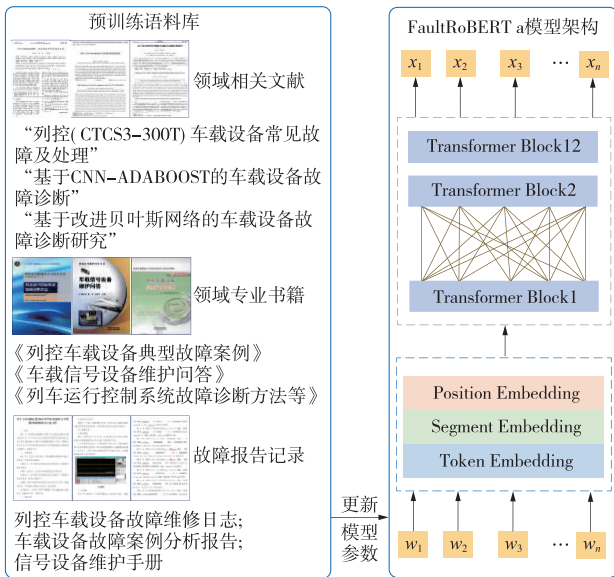


图5 FaultRoBERTa 模型架构

Fig. 5 Architecture of FaultRoBERTa model

其次,对 FaultRoBERTa 模型中的交叉熵损失函数进行更新以解决故障数据不均衡问题^[21]。原交叉熵损失函数如式(1)所示,在模型训练阶段,样本量大的类别损失成本主导总的训练损失,样本量小的类别学习效果不佳。加权交叉熵损失函数通过加入权重系数,使其对效果不好的样本给予相对较大的权重,如式

$$h_i = \vec{h_i} \oplus \overleftarrow{h_i} \quad (9)$$

(2) MHSA 模块

在列控车载设备故障维修日志的同一语料内,相同词语或字符的含义可能存在明显差异,例如,在文本“DMI 显示 DMI 与主机通信中断,导致 ATP 故障停车,经 AElog 分析,由于 A 核心代码模式转换无效,导致 ATPCU 软件故障”中,首个 DMI 仅代表列控车载设备人机界面部分,而后文“DMI 与主机通信中断”短语中 DMI 作为故障现象的组成部分,是故障知识抽取任务的关键内容。仅采用单一的 BiLSTM 模型难以捕捉文本序列中单词单元的重要性权重,可能导致关键信息的遗漏。因此,引入 MHSA 模块同时从不同的角度学习文本表示,从而优化全局特征提取性能。

MHSA 模块考虑序列内部的联系,将 FaultRoBERTa 层输出的向量矩阵经过 3 组不同的线性变换,生成 3 个维度相同的矩阵: Q 、 K 和 V ,并参与注意力函数的计算,如下

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (10)$$

式中, d_k 为神经网络的隐层单元数, $\sqrt{d_k}$ 作为缩放因子,用于调节 Q 、 K 内积的大小。

通过对这一过程进行 h 次独立的重复,每一次获得不同的结果,拼接起来形成多头自注意力机制,如下

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (11)$$

$$\text{st}_i = \text{Multihead}(Q, K, V) = \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W^0 \quad (12)$$

式中, W_i^Q 、 W_i^K 、 W_i^V 为线性映射所使用的参数矩阵; head_i 为 MHSA 模块中的第 i 个头; concat 为拼接操作; W^0 为权重矩阵,用于降低维度。

最后,将 MHSA 模块的输出 st_i 和 BiLSTM 模块的输出 h_i 结合起来,通过 $\tanh$ 激活函数处理,得到特征提取层的输出 z_i ,如下

$$z_i = \tanh(h_i \oplus \text{st}_i) \quad (13)$$

2.3 基于 CRF 的文本序列解码

在经过 FaultRoBERTa 模块、MHSA-BiLSTM 模块多层处理后得到大量标签序列对应的向量,为了弥补特征提取层无法获取相邻标签依赖关系和约束条件的缺陷,引入条件随机场(Condition Random Field, CRF)模块^[23],利用其中的特征函数对标签序列进行概率计算,特征函数如下

$$y^* = \arg \max_{y \in Y_X} \text{Score}(X, \tilde{y}) \quad (14)$$

式中, $\tilde{y}$ 为真实正确的标注序列; Y_X 为所有可能的标记集合,经维特比算法求得最优预测序列 y^* 。

2.4 基于 READ 正则化机制的模型优化

鉴于列控车载设备故障注释数据难收集、规模小,深度学习模型易出现过拟合的问题,采用一种新的正则化机制 READ 来正则化同一样本的 2 个分布,即原始分布和叠加对抗扰动及丢弃扰动干扰的分布,通过最小化同一样本的 2 个分布之间的双向 KL 散度来规范模型预测^[24-26]。

READ 的体系结构如图 6 所示。对于输入序列 $W = \{w_1, w_2, \dots, w_n\}$ 和输出序列 $y = \{y_1, y_2, \dots, y_n\}$,经过预训练模型 FaultRoBERTa 处理之后获得嵌入向量表示 X 。在原嵌入 X 基础上应用 FGM 对抗扰动生成对抗嵌入 X' ,如下

$$X' = X + \delta \quad (15)$$

式中, δ 为由 FGM 产生的对抗性扰动^[27]。FGM 采用 L2 归一化来调整特定的梯度,以达到更好的扰动,计算方法如下

$$\delta = \epsilon \frac{g}{\|g\|_2} \quad (16)$$

$$g = \nabla_X L_{JE}(W, \theta) \quad (17)$$

式中, ϵ 为扰动度; g 为损失的梯度; θ 为模型参数; $L_{JE}(W, \theta)$ 为输入样本的损失值,损失函数关于输入向量 X 的偏导作为梯度 g 。

如图 6 所示,在神经网络中,随机应用不同的 dropout 后,可以得到 2 个不同的训练子模型,接着,将原始嵌入 X 和对抗嵌入 X' 分别输入到 2 个子模型中,得到 2 种不同的输出标签序列分布,即 $P(y|W)$ 和 $P'(y|W)$,产生 2 种不同的损失。因此,将 2 个损失的平均值作为基本的学习对象 L_{JE} ,即

$$L_{JE} = -\frac{1}{2}(\lg P(y|W) + \lg P'(y|W)) \quad (18)$$

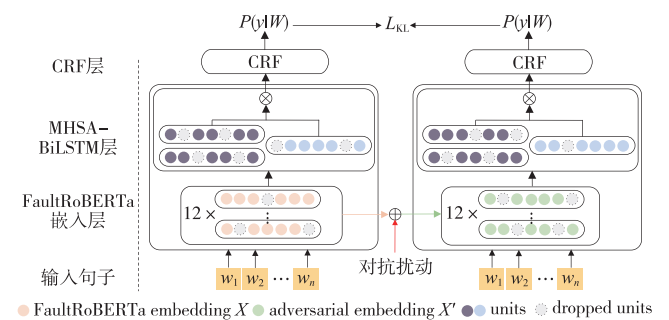


图 6 READ 模型架构

Fig. 6 Architecture of READ model

加入对抗扰动和 dropout 噪声将使表征偏离原始输入。READ 在训练过程中通过最小化同一样本的 2 个输出分布之间的双向 KL 散度来减小 2 个子模型预测之间的差异,如下

$$L_{\text{READ}} = \frac{1}{2} \left(D_{\text{KL}}(P(y|W) \| P'(y|W)) + D_{\text{KL}}(P'(y|W) \| P(y|W)) \right) \quad (19)$$

式中, L_{READ} 为正则化方法的损失函数。 $D_{\text{KL}}(P(y|W) \| P'(y|W))$ 表示 2 个分布 P 和 P' 之间的 KL 散度。

最终训练目标是最小化数据 (W, y) 的联合损失 L , 计算方法如下

$$L = L_{\text{JE}} + \lambda L_{\text{READ}} \quad (20)$$

其中, λ 为平衡 L_{JE} 和 L_{READ} 的权衡参数。

3 实验与结果分析

3.1 实验数据集

为评估联合抽取模型 FMBCR 在列控车载设备故障三元组抽取方面的性能, 选择 CTCS3-300T 型列控车载设备的典型故障案例手册, 以及近三年来自各铁路局列控车载设备故障维修日志的数据, 共计 2 450 条故障文本作为实验数据集, 由人工按照本文的标注策略进行精准标注。经过对超 35 万字符的列控车载设备故障文本语料进行深入分析与统计, 得到故障信息中涉及的基础实体和基本关系可以分别归纳为 4 种类别。表 4 展示各类实体名称、关系名称、具体数量以及分布情况。

表 4 列控车载设备故障信息实体及关系类别

Tab. 4 Fault information entities and relationship categories of CTCS on-board equipment

类别	名称	标注标签	实体数量	字符数量	字符占比/%
实体类别	故障现象	PHE	2 450	36 497	9.9
	故障原因	REA	2 635	12 624	3.0
	故障类型	TYPE	2 450	14 135	3.1
	处理措施	MEA	2 867	30 463	8.2
关系类别	原因关系	FR	2 450	36 497	9.9
	分类关系	FT	2 450	14 135	3.1
	处理关系	FM	2 867	30 463	8.2
	重叠关系	FO	2 450	10 834	2.7

实验过程中将数据集以 8 : 1 : 1 的比例划分为训练集、验证集和测试集。为全面评估模型的性能, 本研究采用准确率、精确率 P (Precision)、召回率 R (Recall) 和 $F1$ 值作为评估指标, 对 FMBCR 模型进行评价。

3.2 实验环境及参数设置

选择 Windows 10 操作系统作为实验平台, 基于 Keras 框架搭建联合抽取模型, 具体训练环境配置如下, CPU 选用 Intel i9 12900K, 配备的 GPU 为 NVIDIA GeForce RTX 2080 Ti。在编程方面, 选择 Python 3.7 作为编程语言, 并基于 Tensorflow 1.14.0 深度学习框架进行模型训练与部署。

本模型根据多次试验结果, 最终确定 FMBCR 模型中的相关超参数配置如下, FaultRoBERTa 作为该模型的上下文嵌入层, 词嵌入维度为 768, 输入文本最大长度为 256; BiLSTM 隐藏层维数设置为 256, 用 Adam 优化器进行参数更新, 批处理大小为 15, 学习率取 1.0×10^{-5} , CRF 学习率设为 0.001, 并在实验中引入 Dropout 机制缓解过拟合问题。在模型训练过程中, 通过调整 Dropout 取值 0.1~0.8 范围内研究该参数对准确率的影响, 结果表明, 当设定 Dropout 取值为 0.5 时, 准确率最优; 另外, 在保持其他超参数不变的情况下, 对比正则化损失权重参数 λ 在 6 个不同取值 (1、2、3、4、5 和 10) 下的模型表现。实验结果显示, λ 取值为 2 时, 模型获得最佳准确率。Dropout 和 λ 取值寻优过程如图 7 所示。

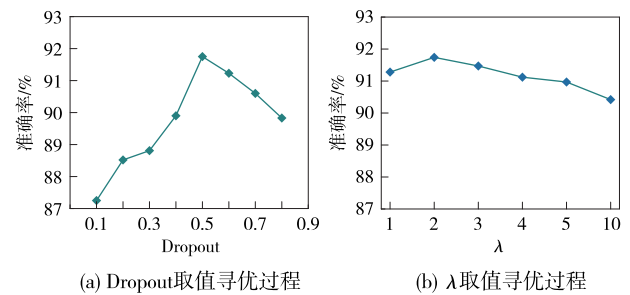


图 7 FMBCR 模型关键参数寻优过程

Fig. 7 Optimization process of key parameters in FMBCR model

3.3 对比实验

(1) 模型抽取性能对比实验

为了验证所提模型在列控车载设备故障数据集上的知识抽取有效性, 选取抽取性能较好的 4 种经典模型, Bigram+模式匹配、BiLSTM-CRF+CNN、Novel Tagging 和 CASREL 与 FMBCR 联合抽取模型进行比较。不同模型抽取性能比较如图 8 所示。

实验数据充分验证 FMBCR 模型在抽取性能上的卓越表现, 其精确率、召回率及 $F1$ 值均优于对比算法。下面对实验结果进行分析: Bigram+模式匹配、BiLSTM-CRF+CNN 两种模型在处理三元组抽取任务时, 均遵循先识别实体再抽取关系的传统流程, 无法保证实体及其关系特征在整个序列中的连续性和一致性, 对比结果表明实体关系联合抽取的优越性; 基于原始标注策略的 Novel Tagging 模型无法正确识别重叠的三元组, 作为列控车载设备故障知识抽取中的关键部分, 重叠三元组识别中的一个错误可能会影响多个三元组的正确抽取, 因此 Novel Tagging 模型的精确率、召回率和 $F1$ 值均不及另外 5 个模型。基于参数共享的 CASREL 联合抽取模型虽然有效解决了重叠三元

组的抽取难题,但本质上是一种两阶段方法,会受到曝光偏差和误差传播的影响。对于所提 FMBCR 模型而言,对序列标注和抽取模型均进行改进,使得更加适合于重叠实体关系联合抽取任务,相较于 CASREL 模型,精确率、召回率和 $F1$ 值分别提升 3.0%、4.3% 和 3.7%。

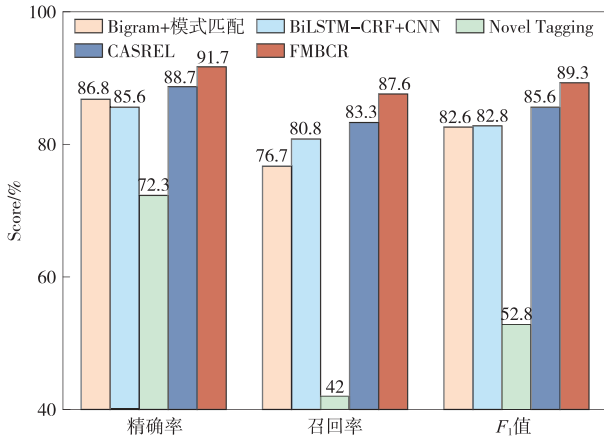


图 8 不同模型抽取性能比较

Fig. 8 Comparison of extraction performance across different models

(2) 重叠关系抽取对比实验

为了深入评估 FMBCR 模型在提取重叠关系方面的性能,根据三元组中关系重叠的不同模式,将列控车载设备故障文本数据集划分为两个子集,即正常类别 (Normal) 和单一实体重叠类别 (Single Entity Overlap, SEO),选取联合抽取的经典模型 CopyRE^[28-29]、CopyMTL^[14]、GraphRel^[30] 和 CASREL 对不同类型的数据进行扩展实验。

图 9 显示 5 种不同模型在列控车载设备故障数据集不同重叠模式下进行知识抽取的结果。实验结果表明, CopyRE、CopyMTL、GraphRel 三种模型在 Normal 类和 SEO 类不同重叠模式下的性能存在显著差异,反映 SEO 类是模型相对难提取的模式,暴露了模型在提取重叠三元组方面的缺陷。而 CASREL 模型倾向于从给定句子中尽可能提取多个关系三元组,因此在 SEO 类数据集上表现出色。然而, CASREL 采取就近匹配策略,导致处理过程中较远的尾实体容易被忽略,因此限制了其在解决复杂关系重叠场景下的性能。对于所提 FMBCR 模型而言,精准的三元组标注,以及 MHSA 机制的加入,使得模型能够捕获句中任意位置字符之间的关联关系,解决长序列的远距离依赖问题。因此,更加适用于列控车载设备故障数据集的重叠三元组抽取任务。FMBCR 模型在 Normal 类和 SEO 类车载设备故障数据集上的 $F1$ 值均超越其他模型。相较于 CAS-

REL 模型,分别提升 5.2% 和 1.6%,展现出 FMBCR 模型在该领域内的显著优势。

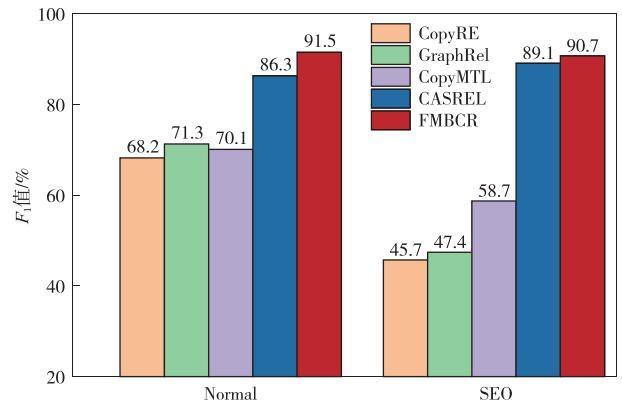


图 9 基于列控车载设备故障数据集模型对比

Fig. 9 Model comparison based on fault dataset of CTCS on-board equipment

3.4 消融实验

为了深入探究 FMBCR 模型中各个组件的作用及其对整体知识抽取性能的贡献,实施一系列消融实验。依次移除 MHSA 模块, READ 模块中的对抗性扰动 AP, READ 模块中的 Dropout 扰动 DP 以及 READ 模块,观察这些变化对知识抽取性能的影响,表 5 展示基于列控车载设备故障数据集的模型消融实验结果。

表 5 FMBCR 模型消融实验结果

Tab. 5 Ablation experimental results of FMBCR model

MHSA	DP	AP	P /%	R /%	$F1$ /%
	✓	✓	91.1	86.8	88.7
✓	✓		91.7	86.2	88.9
✓		✓	91.0	87.0	88.4
✓			90.3	86.9	88.2
✓	✓	✓	91.7	87.6	89.3

相较于完整的 FMBCR 模型,移除 MHSA 模块后,精确率、召回率和 $F1$ 值分别下降 0.6%、0.8% 和 0.6%,揭示了对于知识抽取任务而言,有效提取标签的语义特征至关重要,结合 MHSA 模块更加关注关键特征信息,学习全局信息能力更强的特性,将其引入到神经网络中能够显著增强知识抽取模型的优越性。

移除 READ 模块中的 AP 或 DP 机制也导致知识抽取性能的降低,与 FMBCR 模型相比, $F1$ 值分别下降 0.4% 和 0.9%,分析性能下降的原因是,训练神经网络时采用小的故障文本训练集,删除 AP 或 DP 机制,由于降低样本多变性,模型缺乏海量训练样本,通常对测试数据的泛化能力较差,从而出现过拟合现象,影响知识抽取性能。其中 DP 表现优于 AP 的原因是, AP 仅作用于 FaultRoBERTa 嵌入层,而 DP 应用于整个模型,对样本干扰范围更广。

最后,移除 FMBCR 模型中的 READ 模块后,精确率、召回率和 $F1$ 值分别下降 1.4%、0.7% 和 1.1%。可以明确,移除 READ 模块后,精确率和 $F1$ 值的下降幅度最显著。主要原因是,加入的每个扰动都较简单,仅采用 AP 或 DP 一种机制只能引入很小的扰动,加入不同的扰动可增加同一样本表示的差异。从而证明 READ 正则化方法可以有效解决小样本数据造成的过

拟合问题,显著提高知识抽取的准确性与全面性。

3.5 三元组抽取实例分析

进一步讨论传统流水线模型 Bigram+模式匹配、基于参数共享的联合抽取模型 CopyMTL 以及 FMBCR 联合抽取模型在故障知识三元组抽取方面的效果。表 6 展示示例结果,其中“[]”代表正确识别的实体,下标指明该实体所属的关系类别。

表 6 故障三元组抽取结果示例
Tab.6 Examples of fault triple extraction results

模型	抽取结果
示例 1	如果重启仍旧[报 BTM 端口无效] _{FM} ,建议[更换 BTM 且软件升级] _{FM} 解决
Bigram+模式匹配	如果重启仍旧[报 BTM 端口无效] _{FR} ,建议[更换 BTM 且软件升级] _{FR} 解决
CopyMTL	如果重启仍旧[报 BTM 端口无效] _{FM} ,建议[更换 BTM 且软件升级] _{FM} 解决
FMBCR	如果重启仍旧[报 BTM 端口无效] _{FM} ,建议[更换 BTM 且软件升级] _{FM} 解决
示例 2	发车前[ATP 报 C2 级故障后换系重启正常] _{FR} 。经检查,由于 ATP 车载设备启动时,C 级模块在后台工作读取[MVB 总线应答器端口信息无效] _{FO} ,所以报 C2 故障,属于[C2CU 软件故障] _{FT} ...
Bigram+模式匹配	发车前[ATP 报 C2 级故障] _{FR} 后换系重启正常。经检查,由于 ATP 车载设备启动时,C 级模块在后台工作读取[MVB 总线应答器端口信息无效] _{FR,FT} ,所以报 C2 故障,属于[C2CU 软件故障] _{FT} ...
CopyMTL	发车前[ATP 报 C2 级故障后换系重启正常] _{FR} 。经检查,由于 ATP 车载设备启动时,C 级模块在后台工作读取[MVB 总线应答器端口信息无效] _{FR} ,所以报 C2 故障,属于 C2CU 软件故障...
FMBCR	发车前[ATP 报 C2 级故障后换系重启正常] _{FR} 。经检查,由于 ATP 车载设备启动时,C 级模块在后台工作读取[MVB 总线应答器端口信息无效] _{FO} ,所以报 C2 故障,属于[C2CU 软件故障] _{FT} ...

针对示例 1 中的列控车载设备故障文本序列,相较于标准的抽取结果,三类模型均能够正确识别出相应实体。但是,采用流水线方法的模型在基于此实体识别结果进行关系预测时,由于未能充分考虑实体识别与关系抽取任务间的交互性,错误地将原本的语义关系“FM”分类为“FR”。相比之下,两种联合抽取模型能够更准确地捕捉并表达实体间的语义关系。

针对示例 2,标准抽取结果包含 3 个实体,其中实体“MVB 总线应答器端口信息无效”隶属于多个关系。对于该标注结果,模型在准确识别所有实体的同时,能够准确表达所有实体的关系类别。而关系域映射实体域的 CopyMTL 模型未能识别出“C2CU 软件故障”实体及其参与的“FT”关系,因此,该模型对于单句中存在多个三元组的抽取性能较差。流水线模型对于实体“ATP 报 C2 级故障后换系重启正常”的识别存在边界错误,且该错误实体被传递给下一级关系抽取任务,导致错误实体与其他实体和关系进行匹配,造成误差累积。以上结果表明,模型在复杂文本的三元组抽取任务中表现最佳。

3.6 知识三元组图形化展示

基于提出的列控车载设备故障知识抽取方法,共抽得到 1 014 对故障知识三元组,包含 304 组原因关系,304 组分类关系,406 组处理关系,以及 764 个车载设备故障实体,其中故障现象实体 198 个,故障原因实体 237 个,故障类型实体 15 个,处理措施实体

314 个。接着,采用 Neo4j 图数据库构建 CTCS3-300T 型列控车载设备故障知识图谱,局部示意如图 10 所示。图中不同颜色的圆圈代表不同类型的实体,通过知识图谱将复杂的故障文本转化为简洁的三元组形式,便于分析故障信息间的链接关系,为后续应用提供数据支撑。

4 结论

(1)针对列控车载设备故障知识抽取中存在的错误传递、信息冗余、重叠三元组难抽取等问题,提出一种基于序列标注的联合抽取方法,该方法利用序列标注策略实现实体关系的同步标注,采用 FMBCR 联合抽取模型预测序列标签,其中 FaultRoBERTa 发挥出极大的领域适应能力,可提升实体边界识别的准确率,MHSA 模块可补充 BiLSTM 忽略的字词单元重要性权重,增强网络对文本语义的学习能力,而 READ 模块通过增加扰动的多样性放大大一样本的可变性,能够提高模型的鲁棒性,最终在列控车载设备故障维修日志语料中取得优异的抽取效果。

(2)以近 3 年各铁路局列控车载设备故障维修日志为基础,通过充分实验,证明所提方法实现优异的抽取性能,在 Normal 类和 SEO 类不同的关系重叠模式下,其 $F1$ 值分别较 CASREL 模型提高 5.2%和 1.6%。

(3)通过所提方法,实现 CTCS3-300T 型列控车载设备故障知识图谱的构建,证明该方法的实用性。

- 究[J]. 小型微型计算机系统, 2022, 43(12): 2479-2486.
- LI Daiyi, LI Zhongliang, YAN Li. Research on Chinese-oriented Entity Relation Joint Extraction Method[J]. Journal of Chinese Computer Systems, 2022, 43(12): 2479-2486.
- [14] WEI Z P, SU J L, WANG Y, et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[EB/OL]. 2019: arXiv: 1909.03227. <https://arxiv.org/abs/1909.03227>
- [15] 王传栋, 徐娇, 张永. 实体关系抽取综述[J]. 计算机工程与应用, 2020, 56(12): 25-36.
- WANG Chuandong, XU Jiao, ZHANG Yong. Survey of Entity Relation Extraction[J]. Computer Engineering and Applications, 2020, 56(12): 25-36.
- [16] ZHENG S C, WANG F, BAO H Y, et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 1227-1236.
- [17] 吴赛赛, 周爱莲, 谢能付, 等. 基于深度学习的作物病虫害可视化知识图谱构建[J]. 农业工程学报, 2020, 36(24): 177-185.
- WU Saisai, ZHOU Ailian, XIE Nengfu, et al. Construction of Visualization Domain-specific Knowledge Graph of Crop Diseases and Pests Based on Deep Learning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(24): 177-185.
- [18] NAYAK T, NG H T. Effective Modeling of Encoder-decoder Architecture for Joint Entity and Relation Extraction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8528-8535.
- [19] 朱秀宝, 周刚, 陈静, 等. 基于增强序列标注策略的单阶段联合实体关系抽取方法[J]. 计算机科学, 2023, 50(8): 184-192.
- ZHU Xiubao, ZHOU Gang, CHEN Jing, et al. Single-stage Joint Entity and Relation Extraction Method Based on Enhanced Sequence Annotation Strategy[J]. Computer Science, 2023, 50(8): 184-192.
- [20] LIU F Y, VULIĆ I, KORHONEN A, et al. Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders[EB/OL]. 2021: arXiv: 2104.08027. <https://arxiv.org/abs/2104.08027>
- [21] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial Examples Are Not Bugs, they Are Features[C]//Neural Information Processing Systems. , 2019.
- [22] 束嘉伟, 杨挺, 耿毅男, 等. 面向电力知识图谱构建的重叠实体关系联合抽取方法[J]. 高电压技术, 2024, 50(11): 4912-4922.
- SHU Jiawei, YANG Ting, GENG Yinan, et al. Joint Extraction Method for Overlapping Entity Relationships in the Construction of Electric Power Knowledge Graph[J]. High Voltage Engineering, 2024, 50(11): 4912-4922.
- [23] 魏伟, 赵小强, 丁艳华, 等. 基于 BiLSTM 和改进注意力机制的高铁车载设备故障诊断[J]. 铁道学报, 2022, 44(11): 53-62.
- WEI Wei, ZHAO Xiaoqiang, DING Yanhua, et al. Fault Diagnosis of Vehicle On-board Equipment Based on BiLSTM and Attention Mechanism[J]. Journal of the China Railway Society, 2022, 44(11): 53-62.
- [24] 周长玉, 李想, 焦润海, 等. 基于领域知识图谱的变压器故障联合推理[J]. 中国电机工程学报, 2025, 45(6): 2160-2170.
- ZHOU Changyu, LI Xiang, JIAO Runhai, et al. Joint Reasoning for Transformer Faults Based on Domain Knowledge Graphs[J]. Proceedings of the Chinese Society for Electrical Engineering, 2025, 45(6): 2160-2170.
- [25] 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.
- LIN Haixiang, BAI Wansheng, ZHAO Zhengxiang, et al. Knowledge Extraction Method for Operation and Maintenance Texts of High-speed Railway Turnout[J]. Journal of Railway Science and Engineering, 2024, 21(7): 2569-2580.
- [26] SHARMA R, MORWAL S, AGARWAL B. Named Entity Recognition Using Neural Language Model and CRF for Hindi Language[J]. Computer Speech and Language, 2022, 74: 101356.
- [27] WU Lijun, et al. R-drop: Regularized dropout for neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 10890-10905.
- [28] 许鑫, 马文政, 张浩, 等. 基于融合对抗训练的农作物品种信息抽取方法[J]. 农业机械学报, 2023, 54(12): 272-279, 337.
- XU Xin, MA Wenzheng, ZHANG Hao, et al. Crop Variety Information Extraction Method Based on Integrated Adversarial Training[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(12): 272-279, 337.
- [29] ZHOU R G, CHANG S D, LI Y C. A Neural Network Architecture for Information Extraction in Chinese Drug Package Insert[J]. IEEE Access, 2020, 8: 51256-51264.
- [30] FU T J, LI P H, MA W Y. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 1409-1418.

(编辑:夏梦然)

(上接第210页)

- [25] 武小年, 张楚芸, 张润莲, 等. WSN 中基于改进粒子群优化算法的分簇路由协议[J]. 通信学报, 2019, 40(12): 114-123.
- WU Xiaonian, ZHANG Chuyun, ZHANG Runlian, et al. Clustering Routing Protocol Based on Improved PSO Algorithm in WSN[J]. Journal on Communications, 2019, 40(12): 114-123.

- [26] 刘帅, 李正炜, 吴元昊, 等. 基于能耗梯度的无线传感器网络路由算法[J]. 传感技术学报, 2016, 29(8): 1247-1252.
- LIU Shuai, LI Zhengwei, WU Yuanhao, et al. Routing Algorithm for Wireless Sensor Network Based on Energy Consumption Gradient[J]. Chinese Journal of Sensors and Actuators, 2016, 29(8): 1247-1252.

(编辑:夏梦然)