# 基于边界感知的工业设备故障命名实体识别方法

葛卫京 刘晓丽 杜亚峰
(商丘工学院机械工程学院 河南 商丘 476000)

**摘 要** 命名实体识别在识别工业设备故障方面发挥关键作用,有助于故障预测、维护管理和智能决策。针对工业设备故障数据中存在的嵌套结构和长跨度问题,提出一种边界感知的实体识别方法。该方法通过边界感知精准定位实体跨距,并结合类别预测判断实体跨距的所属类别,以提高识别性能。此外,为解决标注数据的缺乏的问题,还构建面向工业设备故障的实体识别数据集。实验结果证明了该方法在工业设备故障实体识别方面的有效性,并为后续数据分析和知识图谱的构建提供了坚实基础。

**关键词** 命名实体识别 预训练语言模型 工业设备 故障信息

**中图分类号** TP391.1 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.06.035

# NAMED ENTITY RECOGNITION METHOD FOR INDUSTRIAL EQUIPMENT FAULTS BASED ON BOUNDARY-AWARE

Ge Weijing Liu Xiaoli Du Yafeng
(*School of Mechanical Engineering, Shangqiu Institute of Technology, Shangqiu 476000, Henan, China*)

**Abstract** Named Entity Recognition plays a key role in identifying industrial equipment faults, aiding in fault prediction, maintenance management, and intelligent decision-making. Aiming at the nested structures and long spans in industrial equipment fault, this paper proposes a boundary-aware entity recognition method. This method accurately located the span of entities through boundary detection and enhanced recognition performance by combining category prediction to determine the entity span's category. To tackle the scarcity of labeled data, this paper constructed an entity recognition dataset targeted at industrial equipment faults. Experimental results demonstrate the effectiveness of this method in recognizing entities related to industrial equipment faults, which lays a solid foundation for subsequent data analysis and knowledge graph construction.

**Keywords** Named entity recognition Pre-trained language models Industrial equipment Fault information

## 0 引 言

随着智能制造的深入发展,工业设备的数字化和智能化成为现代工业研究的焦点之一。命名实体识别(Named Entity Recognition,NER),作为自然语言处理的核心技术,能快速定位工业设备故障的关键信息,为后续的预测维护、设备健康管理和智能决策提供坚实的基础。工业设备故障命名实体识别是指从涉及生产故障信息的文本中提取具有特定意义的词汇或短语,在设备状态监测、智能维护策略、故障分析等领域拥有巨大的应用潜力。例如,对特定设备故障代码的识别和归类,可以为设备维护工程师提供即时的故障解决方案、帮助生产管理人员评估设备的运行效率、为生产线工程师提供故障预测的技术支持,同时也能助力供应链人员更好地管理零部件库存,响应可能的设备更换需求。

在工业设备故障文本中,实体通常呈现出复杂的词构形式。由于中文以单个汉字作为最基本的书写单元,这经常导致一个实体包含在另一个实体之内的现象。例如,在句子"...X 型冲压机过热..."中包含了两个实体,即"X 型冲压机"和"X 型冲压机过热"。尽管这两个实体共享多个词,但它们表达不同的类别。比如,实体指称是"X 型冲压机",其实体类别应为"设备

收稿日期:2024-01-17。**葛卫京**,讲师,主研领域:深度学习,模式识别,自然语言处理。**刘晓丽**,副教授。**杜亚峰**,助教。

名称”。同理,如果实体指称是“X 型冲压机过热”,那这个实体指称的类别应为“故障原因”。因此,在识别实体时,准确感知实体指称的边界是至关重要的,这有助于正确判断实体的类别。

针对工业设备故障实体识别的难题,张阳等[1]采用了中文字符增强的实体识别模型来应对专业词汇挑战,而黄子麒等[2]则设计了一个语义感知的深度模型来从语义层面识别简单实体。然而,这些方法都无法有效识别嵌套实体。Shen 等[3]依靠手工特征,并结合语言学特性和外部知识资源来识别句子中嵌套实体,但这一过程涉及复杂的特征工程。Ju 等[4]提出了一种层级序列标注模型,该模型先识别内部实体,然后将其作为输入来识别外部实体。然而,该模型容易受到错误传播的影响,即一旦前一层提取错误实体,后续层的性能也会受损。此外,当外部实体优先被识别时,内部实体可能遗漏。Sohrab 等[5]提出了一种基于枚举跨距的分类模型,通过预测所有可能跨距的类别来识别实体,但这种方法忽略了边界信息,可能导致非实体错误提取。

为了解决上述问题,本文提出一种边界感知的实体识别方法(Boundary-aware Named Entity Recognition, BNER),该方法中边界感知层采用序列标注方法定位出实体指称的边界,并结合类别预测的跨距分类方法对实体指称进行类别预测,从而进一步提高模型识别实体的准确率。此外,为了使模型具备特定领域知识,本文构建一个面向工业设备故障的实体识别数据集,用于模型的训练与评估。

本文的贡献如下:(1) 构建工业设备故障的实体识别数据集。针对工业设备故障数据的特点,构建一个复杂且多样化的实体识别数据集。该数据集可作为一个工业故障领域命名实体识别标准数据集,用于训练和评估工业设备故障实体的识别性能。(2) 提出基于边界感知的命名实体识别模型,旨在识别工业设备故障实体。该模型主要由编码层、特征共享层、边界感知层和类别预测层组成。(3) 进行充分的实验。与现有实体识别方法进行对比,验证 BNER 在工业设备故障实体识别任务中的性能,并进行详细的实验分析。

## 1 相关工作

作为自然语言处理(Natural Language Processing, NLP)中一个关键任务,命名实体识别(NER)近年来受到了广泛的关注。传统的 NER 工作依赖于领域专家定义的规则,这些规则通常建立在特定领域的名录表[6]和句法-词汇模式[7]之上。这类方法的泛化性较差难以适配新的领域。基于深度学习的方法具有更好的泛化能力,并且不依赖手工特征节省了大量成本[8-9]。随后,一系列 NER 模型相继被提出,并逐渐领域化。王春雨[10]采用 LSTM 和 CRF 进行实体识别。Qin 等[11]提出了一种新型的 CNN-BILSTM-CRF 神经网络方法,并在语料库上的实验中证明了其精确度和 F 值优于其他模型。王腾科等[12]在现有的 BiLSTM-CRF 模型基础上添加了自注意力机制,以捕获与当前词更相关的特征信息,从而提高了模型性能。Zhao 等[13]分析了实体指称的跨距关系,提出双向映射的深度模型公开数据集上的表现超越了其他深度学习模型。

最近,BERT 在 NER 领域得到了广泛应用。张阳等[1]在工业设备故障信息文本采用 RoBERTa-WWM 架构解决 NER,党小超等[14]针对机械设备故障提出融合焦点损失与词典的 NER。与以往的工作不同,本文从工业设备故障实体的结构角度出发,提出边界感知 NER 模型。该模型结合实体的边界信息并采用实体类别信息作为监督信号,进一步提高了实体识别模型的性能。此外,当前的 NER 方法尚未关注到工业设备故障领域,因此本文在工业设备故障领域属于前沿研究。

## 2 数据集构建

### 2.1 数据来源及标注

工业设备故障数据呈现专业性强、实体类别杂、稀疏性显著等特征,面向工业设备故障实体的数据集构建面临着严峻的挑战。考虑到工业设备故障数据特点,本文构建面向工业设备故障的中文命名实体识别数据集。为确保数据集的完整性和准确性,需要遵循以下步骤来完成工业设备故障实体识别数据集的构建。

**Step1** 文本语料爬取。使用网页爬虫技术,从网络上获取工业设备故障领域的设备手册、维修手册、生产厂家的技术公告等数据资源,并采用 OCR 提取文本语料。

**Step2** 清洗文本语料。首先清洗文本中的图片、表格、外部网站链接等。然后按照“描述明确、技术细节丰富、结构条理”准则筛选高质量文本。利用 Hanlp 工具对文本进行去重、空格、空行、停用词及特殊符号处理,确保待标注实体的完整性和一致性。

**Step3** 定义实体类别。本文预定义了七种面向工业设备故障的实体类别,即设备名称(Equipment

Name,EQN)、故障代码(Fault Code,FAC)、故障原因(Fault Reason,FAR)、部件厂商(Component Manufacturer,COM)、维修措施(Repair Action,REA)、生产线名称(Production Line Name,PLN)、预防操作(Prevention Action,PRA)。为了简化实体标注,本文采用缩写表示实体类别,如 EQN 表示设备名称。

**Step4**　实体标注。本文采用"BIOE"的标注策略对工业设备故障的实体进行标注。"B"标注实体的开始位置,"I"标注实体的中间部分,"E"标注实体的结束位置,"0"标注文本中非实体部分。表 1 给出一个具体文本标注样例。

**表 1　文本序列标注样例**

| 实体 | X | 型 | 冲 | 压 | 机 | 过 | 热 |
|---|---|---|---|---|---|---|---|
| 实体标注 | B-EQN | I-EQN | I-EQN | I-EQN | E-EQN | O | O |
| 实体类型 | EQN | | | | | | |
| 实体 | X | 型 | 冲 | 压 | 机 | 过 | 热 |
| 实体标注 | B-FAR | I-FAR | I-FAR | I-FAR | I-FAR | I-FAR | E-FAR |
| 实体类型 | FAR | | | | | | |

**Step5**　标注一致性检测。本文采用数据验证工具自动检测并报告可能的标注错误,并结合人工随机采样检测确保数据集的质量。

## 2.2　数据统计分析

为了深入理解工业设备故障命名实体识别数据集的特征,并为后续的模型训练提供有力的指导,本文对数据集进行了详细的统计分析。分析内容包括实体类别分布、实体频率、句子数量等关键指标,旨在揭示数据集的综合特性。具体的分析结果详见表 2。

**表 2　数据集统计分析**

| 统计项 | 数量 | 嵌套实体数量 |
|---|---|---|
| 文档 | 576 | — |
| 句子 | 4 683 | — |
| EQN | 8 753 | 5 469 |
| FAC | 9 123 | 7 881 |
| FAR | 8 892 | 3 591 |
| COM | 2 142 | 648 |
| REA | 5 673 | 2 542 |
| PLN | 6 433 | 1 385 |
| PRA | 8 034 | 5 742 |
| 平均实体长度 | 6.53 | 3.76 |
| 实体总数 | 49 051 | 27 264 |

与通用实体识别数据集相比,工业设备的实体名称在文本结构和实体数量等方面有显著特征,主要体现在以下三个方面:

(1) 存在大量嵌套结构的实体。工业设备故障实体存在实体间相互包含的现象,比如"X 型冲压机过热"中包含了"X 型冲压机过热"和"X 型冲压机"。这种嵌套结构增加了实体边界识别的复杂度。

(2) 构成实体的词数较多。工业设备故障实体常出现由多个词组成的长序列实体,如"更换了电机并重新调整了冷却系统"。这种长序列实体的存在使得模型难以准确捕获实体的开始和结束位置,从而增加了实体识别的难度。

(3) 实体分布的严重失衡。在工业设备故障实体数据集中,某些实体类别(如"EQN")的实体出现频率远高于其他类别(如"COM"),导致实体分布不均衡。这种失衡可能会使模型过度适应高频实体,而忽视低频实体,进而影响模型的整体识别性能。

# 3　方法设计

针对工业设备故障文本特点,本文提出 BNER,BNER 由四个核心部分组成,分别是编码层、特征共享层、边界感知层和类别预测层。编码层旨在将文本编码成向量,特征共享层旨在从向量中提取语言学特征,边界感知层旨在判断实体可能的跨度,类别预测层旨在依据实体跨度的向量表示预测实体类别。边界感知层和类别预测层通过一个多任务损失函数同时进行训练以优化实体边界及其分类标签的底层依赖关系识别。模型的架构如图 1 所示。

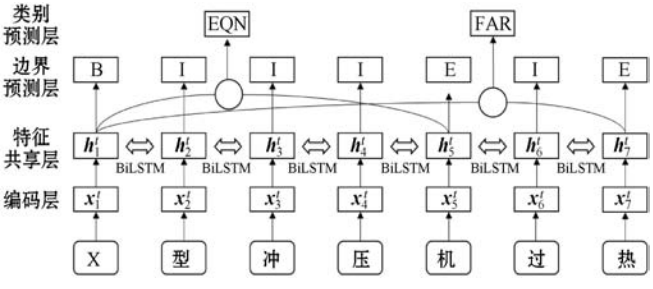


图 1　模型架构

基于边界感知的深度模型识别实体的具体流程如下。

**Step1**　编码层。利用编码器将输入的句子转化为向量形式,以获取句子的语义表达。

**Step2**　特征共享层。利用双向 LSTM 技术提取句子的深层语义特征,为后续的实体边界感知和实体类别预测提供特征的信息。

**Step3**　边界感知层。模型对提取的特征进行分

析,标注出句子中每个字符的实体跨度。跨度的开始位置标注为“B”,结束位置标注为“E”,而“B”和“E”之间的所有字符标注为“I”,其他非实体部分标注为“O”。

**Step4** 类别预测层。对检测到的带有“B”标签的字符及其对应的带有“E”标签的字符进行匹配,并进行实体类别的分类预测。

### 3.1 编码层

编码层的主要任务是将文本中的词转换成多维空间中的向量,这些向量有效地表征了单词的含义及其在语境中的关系,为后续的边界感知和类别分类提供语义支撑。具体地说,对于一个由 $n$ 个词组成的句子 $(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n)$,其中,第 $i$ 个词 $\boldsymbol{t}_i$ 的词级向量表示为 $\boldsymbol{x}_i^w$,由式(1)计算。

$$\boldsymbol{x}_i^{\mathrm{w}} = e^{\mathrm{w}}(\boldsymbol{t}_i) \tag{1}$$

式中:$e^{\mathrm{w}}$ 表示词向量转换器,本文采用预训练语言模型 BERT 初始化词的向量表示。

此外,为了捕获词的字形和形态特征,本文整合了字符级表示。第 $i$ 个词 $\boldsymbol{t}_i$ 的字符级向量表示为 $\boldsymbol{x}_i^c$,$\boldsymbol{t}_i$ 中每个字符的向量表示为 $e^c(\boldsymbol{c}_j)$。$e^c$ 是一个随机初始化的字符向量转换器。这些字符向量随后输入到双向 LSTM 层中,以学习相应的隐藏状态,如式(2)-式(7)所示。

$$\boldsymbol{i}_t = \sigma(\boldsymbol{w}_i[\boldsymbol{h}_{t-1}, e^c(\boldsymbol{c}_j)] + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{w}_f[\boldsymbol{h}_{t-1}, e^c(\boldsymbol{c}_j)] + \boldsymbol{b}_f) \tag{3}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{w}_o[\boldsymbol{h}_{t-1}, e^c(\boldsymbol{c}_j)] + \boldsymbol{b}_o) \tag{4}$$

$$\tilde{\boldsymbol{c}}_t = \tanh \boldsymbol{w}_c[\boldsymbol{h}_{t-1}, e^c(\boldsymbol{c}_j)] + \boldsymbol{b}_c \tag{5}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \times \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \times \tilde{\boldsymbol{c}}_t \tag{6}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \times \tanh(\boldsymbol{c}^t) \tag{7}$$

式中:$\boldsymbol{w}_i$、$\boldsymbol{w}_f$、$\boldsymbol{w}_o$ 为 LSTM 的权重矩阵,$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$ 为偏置项;$\boldsymbol{i}_t$ 为输入门、$\boldsymbol{f}_t$ 为遗忘门、$\boldsymbol{o}_t$ 为输出门;$\boldsymbol{x}_t$ 表示时间 $t$ 的输入向量,$\boldsymbol{h}_t$ 表示时间 $t$ 的隐藏状态;$\tilde{\boldsymbol{c}}_t$ 表示细胞状态的候选值,$\boldsymbol{c}_t$ 表示细胞状态。

字符的前向计算如式(8)所示。

$$\overrightarrow{\boldsymbol{h}}_i^c = \mathrm{LSTM}(e^c(\boldsymbol{c}_j), \overrightarrow{\boldsymbol{h}}_{t-1}) \tag{8}$$

字符的后向计算如式(9)所示。

$$\overleftarrow{\boldsymbol{h}}_i^c = \mathrm{LSTM}(e^c(\boldsymbol{c}_j), \overleftarrow{\boldsymbol{h}}_{t-1}) \tag{9}$$

双向 LSTM 的前向和后向输出连接起来,形成第 $i$ 个词 $\boldsymbol{t}_i$ 的字符表示,如式(10)所示。

$$\boldsymbol{x}_i^c = [\overrightarrow{\boldsymbol{h}}_i^c; \overleftarrow{\boldsymbol{h}}_i^c] \tag{10}$$

式中:$\overrightarrow{\boldsymbol{h}}_i^c$ 和 $\overleftarrow{\boldsymbol{h}}_i^c$ 表示双向 LSTM 的前向和后向输出。第 $i$ 个词 $\boldsymbol{t}_i$ 的最终向量表示为 $\boldsymbol{x}_i^t$,如式(11)所示,其中[;]表示连接操作。

$$\boldsymbol{x}_i^t = [\boldsymbol{x}_i^w; \boldsymbol{x}_i^c] \tag{11}$$

### 3.2 特征共享层

特征共享层旨在从向量表示中提取语言学特征,为后续的实体边界感知和实体类别预测提供有效信息。本文采用双向 LSTM 作为共享特征提取器从文本向量中提取特征。

具体来说,双向 LSTM 的隐藏状态可以用式(12)-式(14)表示。

$$\overrightarrow{\boldsymbol{h}}_i^t = LSTM(\boldsymbol{x}_i^t, \overrightarrow{\boldsymbol{h}}_{i-1}^t) \tag{12}$$

$$\boldsymbol{h}_i^t = LSTM(\boldsymbol{x}_i^t, \boldsymbol{h}_{i-1}^t) \tag{13}$$

$$\boldsymbol{h}_i^t = [\overrightarrow{\boldsymbol{h}}_i^t; \boldsymbol{h}_i^t] \tag{14}$$

式中:$\overrightarrow{\boldsymbol{h}}_t$ 和 $\overleftarrow{\boldsymbol{h}}_t$ 分别表示 LSTM 前向隐态和 LSTM 后向隐态;[*;*]表示拼接操作。

### 3.3 边界感知层

边界感知层旨在从句子中判断出潜在实体的边界,为后续实体类别分类提供有效的跨距表示。具体来说,给定一个句子 $(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n)$ 和句中的一个实体,本文将实体表示为 $R(i,j)$,即由连续的词序列 $(\boldsymbol{t}_i,\boldsymbol{t}_{i+1},\cdots,\boldsymbol{t}_j)$ 组成的实体。如图 2 所示,边界感知层将边界标记 $\boldsymbol{t}_i$ 标为“B”和 $\boldsymbol{t}_j$ 标为“E”。实体内的词被分配标签“I”,非实体标记被分配标签“O”。

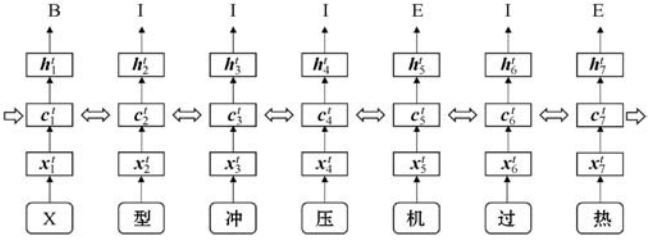


图 2 边界感知示意图

边界感知的具体流程如下:

**Step1** 对于句子中的每个词 $\boldsymbol{t}_i$,首先获取其对应的共享特征表示 $\boldsymbol{h}_i^t$。

**Step2** 将共享特征表示 $\boldsymbol{h}_i^t$ 输入 ReLU 激活函数,并采用 Softmax 分类器来预测一个边界标签,如式(15)-式(16)所示。

$$\boldsymbol{o}_i^t = \boldsymbol{U}\boldsymbol{h}_i^t + \boldsymbol{b} \tag{15}$$

$$\boldsymbol{d}_i^t = \mathrm{softmax}(\boldsymbol{o}_i^t) \tag{16}$$

式中:$\boldsymbol{U}$ 和 $\boldsymbol{b}$ 是可训练参数,本文采用 KL 散度多标签损失函数来量化真实分布 $\hat{\boldsymbol{d}}_i^t$ 和预测分布 $\boldsymbol{d}_i^t$ 之间的差异,如式(17)所示。

$$\boldsymbol{L}_{\mathrm{bcls}} = -\sum(\hat{\boldsymbol{d}}_i^t)\log(\boldsymbol{d}_i^t) \tag{17}$$

### 3.4 类别预测层

类别预测层的主要目的是基于边界感知层识别出的边界表示来判断实体的类别。具体而言,给定一个输入词序列 $X=(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n)$,以及对应的边界标签序

列 $L=(\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_n)$,本文将每个带有标签 B 的字符与带有标签 E 的字符匹配,以构造候选实体跨距,并基于实体跨距的向量表示判断实体的类别。

类别预测层的具体流程如下:

**Step1**　本文首先将带有标签 B 的标记与其自身匹配以召回只包含单个标签的实体。然后对边界区域内的每个词的表示取平均以召回由多个标签构成的实体。实体 $R(i,j)$ 的表示如式(18)所示。

$$\boldsymbol{R}_{i,j}=\frac{1}{j-i+1}\sum_{k=i}^{j}\boldsymbol{h}_k^t \tag{18}$$

**Step2**　将实体的最终表示送入 ReLU 激活函数,并采用 Softmax 层来预测实体的类别。类别标签预测损失的计算如式(19)－式(20)所示。

$$\boldsymbol{d}_{i,j}^{e}=\mathrm{softmax}(\boldsymbol{U}_{i,j}^{e}\boldsymbol{R}_{i,j}+\boldsymbol{b}_{i,j}^{e}) \tag{19}$$

$$\boldsymbol{L}_{\mathrm{ecls}}=\sum(\hat{\boldsymbol{d}}_{i,j}^{e})\log(\boldsymbol{d}_{i,j}^{e}) \tag{20}$$

式中:$\boldsymbol{U}_{i,j}^{e}$ 和 $\boldsymbol{b}_{i,j}^{e}$ 是可训练参数;$\hat{\boldsymbol{d}}_{i,j}^{e}$ 和 $\boldsymbol{d}_{i,j}^{e}$ 分别表示实体分类标签的真实分布和预测分布。

### 3.5　多任务训练

考虑到边界感知层和实体类别预测层共享相同的实体边界表示,本文设计一个多任务损失函数,用以同时优化这两个任务。多任务训练能显著降低过拟合的风险,并增强实体边界感知与类别预测之间的相关性。

在训练阶段,为了确保类别预测的准确性,真实边界标签被直接输入到实体类别预测模块,这样可以保证分类器在没有错误边界感知影响的条件下进行训练。对于测试阶段,模型依据边界感知模块的输出来确定待预测分类标签的实体区域。多任务损失函数的定义如式(21)所示。

$$\boldsymbol{L}_{\mathrm{multi}}=\alpha\sum\boldsymbol{L}_{\mathrm{bcls}}+(1-\alpha)\sum\boldsymbol{L}_{\mathrm{ecls}} \tag{21}$$

式中:$\boldsymbol{L}_{\mathrm{bcls}}$ 表示边界感知模块的分类交叉熵损失;$\boldsymbol{L}_{\mathrm{ecls}}$ 表示实体分类标签预测模块的损失;$\alpha$ 是一个超参数,用来控制每个任务的重要性程度。

## 4　实　验

### 4.1　实验数据集

经过细致且系统地标注后,本文构建一个面向工业设备故障的实体识别数据集。该数据集包含了 4 683 条句子和 49 051 个实体,覆盖了 7 种不同的实体类别。为了全面评估所提模型的有效性,本文按照 8∶1∶1的比例将数据集划分为训练集、验证集和测试集。各个数据子集中实体的分布情况详细列于表 3。

**表 3　数据子集的实体分布情况**

| 类别 | 训练集 | 验证集 | 测试集 |
|---|---|---|---|
| EQN | 5 387 | 2 289 | 1 077 |
| FAC | 6 492 | 1 134 | 1 497 |
| FAR | 6 112 | 1 234 | 1 546 |
| COM | 1 273 | 334 | 535 |
| REA | 3 123 | 1 045 | 1 505 |
| PLN | 3 632 | 1 345 | 1 456 |
| PRA | 5 123 | 1 545 | 1 366 |

### 4.2　实验环境与超参数

实验环境为一块 24 GB NVIDIA GeForce RTX 3090 显卡,Ubuntu 22.04,Python 3.9 和 PyTorch 框架。实验中设置预训练词向量和字符向量分别为 200 和 50 维,Dropout 为 0.5。超参数在训练集上训练并在验证集上的微调后确定,训练迭代设置为 150、学习率为 3E-5 和批次为 24。采用 Adam 作为优化器。

### 4.3　对比方法

为了验证 BNER 的有效性,我们选取了四种主流的实体识别方法,并在构建的数据集上进行对比分析。

(1) 基于 CRF 的方法:这一类方法利用深度学习模型提取句子的语义特征,并采用 CRF 进行实体的解码。具体包括:CNN-CRF、LSTM-CRF、CNN-LSTM-CRF、BERT-CRF、BERT-LSTM-CRF、BERT-LSTM-CNN-CRF。

(2) 基于 Softmax 的方法:这类方法同样基于深度学习模型提取句子语义特征,但使用 Softmax 进行实体的解码。具体方法包括:CNN-Softmax、LSTM-Softmax、CNN-LSTM-Softmax、BERT-Softmax、BERT-LSTM-Softmax、BERT-LSTM-CNN-Softmax。

(3) 基于 GRAPH 的方法:此方法将句子中的每个字符视为图的节点,通过特征抽取实现实体的识别。典型方法:BERT-LSTM-Graph。

(4) 基于 SPAN 的方法:该方法通过匹配句子中的字符对,并对匹配间的字符向量进行类别预测,从而识别实体。典型方法:BERT-Exhausted-Span。

### 4.4　评价指标

为了确保对比的公平性,本文采用严格的评估标准来验证 NER 的有效性,即一个实体只有在其边界和类别标签均正确识别时才被视为正确识别。性能评估主要依赖于三个标准指标,即精确度(Precision,P)、召回率(Recall,R)和 F1 值(F1-score,F1)。

(1) 准确率指的是被正确预测的实体占所有预测为实体的样本比例。

$$p = \frac{T_P}{T_P + F_P} \times 100\%$$

（2）召回率是指所有实际正确的样本中被预测为正确样本的数量。

$$R = \frac{T_P}{T_P + F_N} \times 100\%$$

（3）F1 值是准确率和召回率的调和平均数的一种测量方法，用于衡量 NER 的综合性能。

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

## 5 实验结果

为了充分验证 BNER 的性能，本文在工业设备故障实体识别集上进行一系列详尽的实验。这些实验包括：实体识别的对比实验、边界感知的对比实验和消融实验。

### 5.1 实体识别的对比实验

为了评估 BNER 与当前主流实体识别模型在工业设备故障的性能，本文选择了 14 种基线方法进行比较分析。实验结果如表 4 所示。

**表 4 现有方法的对比实验（%）**

| 模型 | P | R | F1 |
|---|---|---|---|
| CNN-CRF | 60.32 | 67.54 | 63.73 |
| LSTM-CRF | 61.01 | 66.89 | 63.81 |
| CNN-LSTM-CRF | 61.23 | 68.01 | 64.44 |
| BERT-CRF | 64.32 | 69.32 | 66.73 |
| BERT-LSTM-CRF | 64.57 | 69.41 | 66.90 |
| BERT-LSTM-CNN-CRF | 64.66 | 69.55 | 67.02 |
| CNN-Softmax | 60.55 | 67.85 | 63.99 |
| LSTM-Softmax | 61.23 | 67.04 | 64.00 |
| CNN-LSTM-Softmax | 61.27 | 68.33 | 64.61 |
| BERT-Softmax | 64.67 | 68.91 | 66.72 |
| BERT-LSTM-Softmax | 64.66 | 69.56 | 67.02 |
| BERT-LSTM-CNN-Softmax | 64.88 | 69.98 | 67.33 |
| BERT-LSTM-Graph | 74.32 | 71.89 | 73.08 |
| BERT-Exhausted-Span | 70.52 | 91.45 | 79.25 |
| BNER | 78.53 | 80.53 | 80.03 |

可以看出，在工业设备故障实体识别任务中，Softmax 解码器的表现优于 CRF 解码器。这一结果可能源于该数据集中实体标注方式的特点，减少了标签间的依赖性，使得 CRF 解码器难以有效捕捉标签之间的连续性。此外，比较了 CNN 和 LSTM 的特征抽取能力，发现 LSTM 优于 CNN。这可能是因为 CNN 更擅长捕捉局部语义，而 LSTM 能更好地处理句子的全局语义，这对于实体的准确识别至关重要。基于 BERT 的 NER 模型表现出较 Softmax 和 CNN 更优的性能，这可能归功于 BERT 在通用语料上的预训练，它为句子中的每个词提供了恰当的向量表示。最终，BNER 模型在所有主流 NER 模型中表现最佳，这主要得益于 BNER 不仅考虑了句子的语义信息，还综合考虑了实体边界信息。通过采用多任务学习框架，BNER 增强了实体跨距检测与实体类别之间的相关性，显著提高了准确率。

### 5.2 边界感知的对比实验

为了评估 BNER 在识别实体边界的性能，本文选择了四种有代表性的方法进行比较。比较结果如表 5 所示。

**表 5 边界感知的对比分析（%）**

| 模型 | P | R | F1 |
|---|---|---|---|
| BERT-LSTM-CNN-CRF | 70.56 | 71.34 | 70.95 |
| BERT-LSTM-CNN-Softmax | 72.34 | 71.45 | 71.89 |
| BERT-LSTM-Graph | 75.62 | 76.12 | 75.87 |
| BERT-Exhausted-Span | 71.32 | 91.87 | 80.30 |
| BNER | 81.33 | 83.25 | 82.28 |

可以看到，BNER 在实体边界感知方面优于现有方法。特别值得一提的是，BERT-Exhausted-Span 方法在实体边界感知性能上与 BNER 相近。BERT-Exhausted-Span 的主要策略是使用枚举方法来识别所有可能的实体跨度，随后利用实体类别作为监督信号来筛选出不正确的候选跨度。虽然这种枚举策略在短句中效果良好，但它的计算复杂度会随着句子长度的增加而显著增长，从而影响其效率和实用性。

### 5.3 消融实验

为了探究 BNER 中每个模块对工业设备故障实体识别的有效性，本文进行了一系列消融实验，实验结果表 6 所示。

**表 6 模型的消融实验（%）**

| 模型 | P | R | F1 |
|---|---|---|---|
| BNER | 78.53 | 80.53 | 80.03 |
| -Pre-trained | 76.23 | 77.87 | 77.04 |
| -BiLSTM | 75.47 | 76.32 | 75.89 |
| -Dropout | 75.21 | 75.98 | 75.59 |
| -char | 74.89 | 74.32 | 74.60 |

可以看出，移除了预训练语言模型后，BNER 的整体性能下降了 2.99 百分点。删除了 BiLSTM，BNER 的整体性能下降了 1.15 百分点。去除了 Dropout，BNER

（下转第 249 页）

2011:471 – 478.
[15] Xu Y, Zhong Z F, Yang J, et al. A new discriminative sparse representation method for robust face recognition via $l_2$ regularization[J]. IEEE Transaction on Neural Networks and Learning Systems,2017,28(10):2233 – 2242.
[16] Wang J, Yang J C, Yu K, et al. Locality-constrained linear coding for image classification[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010:3360 – 3367.
[17] 刘建伟,崔立鹏,刘泽宇,等. 正则化稀疏模型[J]. 计算机学报,2015,38(7):1307 – 1325.
[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//International Conference on Learning Representations,2015.
[19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM,2017,60(6):84 – 90.
[20] Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems [EB]. arXiv:1603.04467,2016.
[21] Toch L. Phone recognition with deep sparse rectifier neural networks[C]//IEEE International Conference on Acoustics, Speech, and Signal,2013:6985 – 6989.
[22] 屈永康,冀俊忠,梁佩鹏,等. 基于正则化 Softmax 回归的全脑功能性磁共振成像数据特征选择框架[J]. 模式识别与人工智能,2016,29(7):641 – 649.
[23] Zgenel F, Sorgu A G. Performance comparison of pretrained convolutional neural networks on crack detection in buildings [C]//35th International Symposium on Automation and Robotics in Construction,2018:693 – 700.

(上接第 242 页)

的整体性能下降了 0.3 百分点。删除了字符表示,BNER 的整体性能下降了 0.99 百分点。由此,表明了预训练语言模型与特征抽取在工业设备故障实体识别中的重要性。此外,实验结果显示了学习单词的字符表示对实体识别的益处,这也表明字符级别的表示能够改善对领域特定词汇的理解和表达。

## 6 结　语

随着工业领域数字化和智能化的发展,快速准确地识别设备故障信息变得至关重要。本文提出一种边界感知实体识别模型,该模型通过先准确定位实体跨距,后对跨距进行类别分类的方式识别文本中实体。此外,本文构建面向工业设备故障的实体识别数据集,填补了该领域高质量标注数据的空缺。经过实验验证,该模型在工业设备故障实体识别任务中表现出色,对未来的数据分析和知识图谱构建具有重要学术和实践价值。

## 参 考 文 献

[1] 张阳,刘瑾. 基于字符增强的工业设备故障命名实体识别[J/OL]. 电子科技. [2024 – 01 – 16]. https://doi.org/10.16180/j.cnki.issn1007-7820.2024.10.007.
[2] 黄子麒,胡建鹏. 基于语义感知的工业制造领域知识抽取方法[J/OL]. 计算机工程与应用. [2024 – 01 – 16]. http://kns.cnki.net/kcms/detail/11.2127.TP.20230601.1624.004.html.
[3] Shen D, Zhang J, Zhou G, et al. Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain[C]//ACL 2003 Workshop on Natural Language Processing in Biomedicine,2003.
[4] Ju M, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition[C]//2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2018.
[5] Sohrab M G, Miwa M. Deep exhaustive model for nested named entity recognition[C]//2018 Conference on Empirical Methods in Natural Language Processing,2018.
[6] 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.
[7] 王红,李浩飞,邸帅. 民航突发事件实体识别方法研究[J]. 计算机应用与软件,2020,37(3):166 – 172.
[8] Zhao J, Liu C, Liang J, et al. A novel cascade instruction tuning method for biomedical NER[C]//IEEE International Conference on Acoustics,Speech and Signal Processing,2024.
[9] 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.
[10] 王春雨. 基于 CRF 的农业命名实体识别研究[D]. 保定:河北农业大学,2014.
[11] Qin Y, Shen G W, Zhao W B, et al. A network security entity recognition method based on feature template and CNN-BiLSTM-CRF[J]. Frontiers of Information Technology & Electronic Engineering,2019,20:872 – 884.
[12] 王腾科,朱广丽,李瀚臣,等. 基于字词融合和多头注意力的专利实体识别[J]. 计算机工程与设计,2023,44(12):3778 – 3783.
[13] Zhao J, Li Z, Xiao Y, et al. HTMapper: Bidirectional Head-Tail mapping for nested named entity recognition[C]//32nd ACM International Conference on Information and Knowledge Management,2023.
[14] 党小超,刘涧,董晓辉,等. 面向不平衡数据的机械设备故障命名实体识别[J/OL]. 计算机工程. [2024 – 01 – 16]. https://doi.org/10.19678/j.issn.1000-3428.0068078.