# 面向不平衡数据的机械设备故障命名实体识别

党小超[1]，刘涧[1]，董晓辉[1*]，祝忠彦[2]，李芬芳[1]
（1.西北师范大学计算机科学与工程学院，甘肃 兰州 730000；2.金川集团股份有限公司龙首矿，甘肃 金昌 737103）

**摘　要：**命名实体识别作为构建知识图谱的基础任务，其识别效果直接影响知识图谱的质量。在实际生产中，机械故障数据通常包含大量的领域专业词汇，同时实体类型普遍存在分布不平衡的问题，这对准确识别故障实体构成了挑战。通用领域实体识别方法在这一领域效果欠佳，从而降低了知识图谱的质量。为应对上述问题，提出一种融合焦点损失（Focal Loss）函数和专业词典的实体识别方法。该方法使用 Focal Loss 函数应对实体类型不平衡问题，通过引入平衡因子和调制系数，改进传统的交叉熵损失函数，提升实体识别效果，同时将领域专业词汇嵌入到模型中，进一步提高实体识别性能，这一词典包含机械故障的领域术语，有助于模型更准确地识别机械设备故障命名实体。在自建的矿井提升机实验数据集上进行广泛实验验证，结果证明，融入 Focal Loss 后模型的 F1 值比主流模型 BERT-BiLSTM-CRF 提高了 5.57 个百分点，相比用于解决数据不平衡的典型方法 SMOTE 效果更优，在此基础上，通过嵌入领域词典，模型的 F1 值得到进一步提升，达到 89.13%。

**关键词：**命名实体识别；不平衡数据；焦点损失函数；机械设备故障；双向长短期记忆网络；条件随机场
**源代码链接：**https://github.com/buried-age/code/tree/master/ner
**中图分类号：**TP18　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0068078

# Named Entity Recognition of Mechanical Equipment Failure for Imbalanced Data

DANG Xiaochao[1], LIU Jian[1], DONG Xiaohui[1*], ZHU Zhongyan[2], LI Fenfang[1]
(1. School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730000, Gansu, China;
2. Longshou Mine, Jinchuan Group Co., Ltd., Jinchang 737103, Gansu, China)

**【Abstract】** Named Entity Recognition(NER) is a fundamental task in building knowledge graphs and directly affects graph quality. However, in practice, mechanical failure data often contain a significant amount of domain-specific vocabulary, and in general, an imbalance exists in the distribution of entity types. Thus, existing NER methods in general domains do not yield satisfactory results. To address these problems, this paper proposes an entity recognition method that integrates a Focal Loss function into domain-specific dictionaries. This method improves the cross-entropy loss function by introducing balancing and modulation coefficients for sample distributions. In addition, entity recognition is enhanced through the fusion of vocabulary features. Experimental results on a self-built dataset of mining hoist machines show that the incorporation of Focal Loss increases the F1 value by 5.57 percentage points compared with the mainstream Bidirectional Encoder Representations from Transformers (BERT)-Bidirectional Long-Short-Term Memory (BiLSTM)-Conditional Random Field (CRF) model. Furthermore, it outperforms the typical Synthetic Minority Over-sampling Technique(SMOTE) method in solving imbalanced data issues. By incorporating domain dictionaries, the F1 value is further improved, reaching 89.13%.
**【Key words】** Named Entity Recognition(NER); imbalanced data; Focal Loss function; mechanical equipment failure; Bi-directional Long Short-Term Memory(BiLSTM) network; Conditional Random Field(CRF)

## 0　引言

随着人工智能技术的发展，利用语义知识对设备故障进行分析，成为一种新的故障诊断分析方法。但是，由于机械设备故障领域的数据存在记录不规范、知识碎片化等问题，难以准确、高效地管理故障文本信息并获取其间的逻辑关系，这给专业人员的工作开展造成了困难。在智能诊断场景中，对于发生故障的设备，通过查询相关知识图谱可以将繁杂的知识结构化，帮助建立完整的知识体系，从而更快地定位故障所在。因此，对该领域的文本数据开展以命名实体识别（NER）为基础的关键信息抽取方法研究，可以帮助形成高质量的知识图谱，更好地为设备故障的智能诊断提供数据支撑。

命名实体识别是自然语言处理（NLP）中最基础

**收稿日期：**2023-07-14　　**修回日期：**2023-11-24
**基金项目：**国家自然科学基金（62162056）；甘肃省产业支撑计划（021CYZC-06）。
**通信作者 E-mail：**[*] dxh@nwnu.edu.cn

计算机工程 www.ecice06.com

的任务之一,其目的是从非结构化的语言文本中识别出需要的实体类别[1],在知识图谱构建、智能问答等下游任务中起到关键作用。

机械设备故障是一个专业性极强的领域,公开的文本数据集较少,在实际处理时样本数量有着显著性差异,存在某一类别的样本数量远大于其他一个类别或者多个类别的样本数量的现象,即数据样本的不平衡问题。通用领域的实体识别方法更偏重于多数类样本,对少数类样本的识别关注度欠缺[2]。但是,在机械故障领域,少数类样本可能蕴含更有价值的信息,实体识别错误或者无法识别可能造成一定的损失。此外,机械设备故障领域的知识包含大量专业词汇,对文字语义的特征提取要求更高。

基于以上问题,本文提出了一种面向不平衡数据的设备故障实体识别方法,具体工作如下:

1)本文以矿井提升机这一集机、电、液于一体的大型复杂设备故障数据为研究基础,通过收集西北某大型矿业公司矿井提升机故障维修记录、相关学术论文和矿井提升机相关书籍中的故障信息,采用人工标注构建实验数据集,以此来训练和评估本文提出的实体识别方法。

2)为了解决数据集中实体类别样本数据不平衡的问题,在经典实体识别模型 BERT-BiLSTM-CRF 的基础上引入焦点损失(Focal Loss)函数,即在条件随机场(CRF)层融入焦点损失函数,引入平衡因子与调制系数来改进交叉熵损失,以减少负类样本和易分类样本的权重,缓解数据样本分布不平衡问题在实体识别中造成的影响。

3)为了进一步改善模型识别效果,采用词汇集合匹配方法构建了矿井提升机领域词典,在 BERT 底层中融入了词向量辅助确定词边界,从而取得更好的实体识别效果。

## 1 相关工作

早期的命名实体识别使用基于规则的方法[3],其主要思想是基于一组预先确定的规则,根据特定的模式和语法来识别命名实体。另一种流行的 NER 方法是使用 CRF 模型[4],CRF 使用概率图模型来预测命名实体[5]。近年来,主流的 NER 方法是基于 Transformer 的深度学习模型,常用的模型有卷积神经网络(CNN)、长短期记忆网络(LSTM)、递归神经网络(RNN)和门控递归单元(GRU)等。总体来说,基于规则的方法依赖于预定义的规则,基于 CRF 的方法使用概率模型来预测命名实体,而基于 Transformer 的方法使用自注意机制来捕捉长距离依赖关系。这些方法各自都有其优点和局限性,选择哪种方法取决于具体的 NER 任务[6]。

然而上述方法都是在样本数据较为平衡的情况下识别效果较好,但是在大多数实际问题中,数据不平衡的现象是普遍存在的,如本文涉及的机械设备故障领域,包含“故障位置”、“故障现象”、“故障原因”、“维修措施”和“故障影响”5 类实体,其中“故障位置”、“故障现象”和“故障原因”3 类实体数据占到了 90%,“维修措施”和“故障影响”数据不足 10%,样本中各类型实体占比极不平衡,然而“维修措施”决定着维修方案的生成,“故障影响”亦是故障危急评估的主要依据,两者的重要程度决定它们是无法被舍去的。针对这种不平衡问题,以上模型效果并不是很好。

数据不平衡分类问题主要有 2 类解决方法:一是数据层方法,二是算法层方法。其中数据层方法主要有欠采样、过采样和合成少数类过采样(SMOTE)[7]。欠采样是通过减少多数类样本的数量,从而使得样本的分布达到平衡,然而欠采样剔除样本可能破坏数据集的分布;过采样是增加少数类样本的数量,从而使得样本的分布达到平衡,但是有可能引入更多的噪声,带来模型的过拟合问题;SMOTE 及其相关改进算法则通过随机插值生成虚拟样本来增加少数类样本的数量,以实现数据集的均衡。如 HAN 等[8]提出了边界少数类过采样技术(Borderline-SMOTE),该方法根据 $k$ 近邻中同类样本的数量将少数类样本划分为安全样本,危险样本和噪音样本;LIU[9]提出一种 Importance-SMOTE 过采样方法,该方法根据少数类样本的邻近样本组成和分布来计算重要度,并依据得到的重要度来确定生成的样本数量;MALDONADO 等[10]通过加权的 Minkowski 距离计算少数类样本点的邻近样本来合成数据,解决了 SMOTE 算法在高维数据合成时可能产生较大误差的问题。目前,在算法层的研究中,主要有 2 种方法:集合学习与代价敏感学习。集成学习通过融合多个基分类器对均衡数据子集进行分类,以提高其泛化能力,但该方法存在计算效率不高等问题。代价敏感学习解决了数据失衡问题,它对少数类别样本赋予较高的误差成本,并以最小的分类误差总成本为优化目标,常见的方法包括对不同类别的样本赋予不同的权重以及设计特殊的损失函数等,如 CAO 等[11]提出 LDAM 损失函数,对于不同类别样本数量制定不同的间距约束,

计算机工程 www.ecice06.com

给予少数类更宽松的决策边界，从而达到提高泛化性能的目的；LI 等[12]用改进的损失函数 Dice Loss 代替标准的交叉熵处理数据不平衡的 NLP 任务，训练结果得到了显著的提升；LIN 等[13]提出 Focal Loss 损失函数，该损失函数根据不同样本的预测概率分配不同权重，使模型对难以学习的样本给予更多的关注。基于以上研究，本文在 Focal Loss 函数中引入平衡因子和调制系数改进交叉熵损失函数，以解决样本数据分布不平衡问题。

在垂直领域命名实体识别中，引入领域词典提升识别精度是一种常用的方法。ZHANG 等[14]提出的 Lattice-LSTM 在中文 NER 任务中首次引入了词汇信息，但是其计算性能低下，也会造成信息损失；LI 等[15]提出的 FLAT 引入了相对位置信息，字符可以直接与其所匹配词汇进行交互，较大程度保持了词信息；MA 等[16]提出了 Simple Lexicon 模型，该模型融合了 3 种不同的字词信息方法，可适配在不同的特征提取器中。ZHU 等[17]提出了 LexBERT 模型，通过添加实体类型的信息来提高 NER 效果；LIU 等[18]提出了 LeBERT 将词信息注入模型底部来提升实体识别准确度。借鉴上述研究成果，本文构建了矿井提升机领域词典，并在 BERT 内部以直观形式引入每个字符在该段输入文本中所匹配到的多个词信息，以提高本文模型的识别效果。

## 2 模型构建及原理

基于本文需要解决的 NER 任务，在 BERT-BiLSTM-CRF 模型的基础上，提出了一种 BERTw-BiLSTM-CRF-FL 模型，模型结构如图 1 所示，主要包括嵌入层、编码层和解码层。其中，嵌入层在 BERT 基础上，通过词典适配器(LA)模块融合故障设备领域中的词向量进行词嵌入，编码层通过双向 LSTM(BiLSTM)学习上下文特征，解码层使用 CRF 对文本向量进行解码推理，同时对传统的交叉熵函数进行改进，融合 Focal Loss 损失函数使模型更加关注文本中的少类别的实体标签。

该模型初始输入为句子集合 $S=\{S_1,S_2,\cdots,S_m\}$，其中，$m$ 表示句子的总个数，每个句子由 $n$ 个字符组成，即 $c=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}$，每个字符都是一个单独 token，字向量在词典适配器模块与词向量融合后的输出向量作为 BiLSTM 的输入，BiLSTM 的隐藏层向量 $\boldsymbol{h}=\{h_1,h_2,\cdots,h_n\}$，$\boldsymbol{h}$ 向量与注意力机制的权重矩阵进行加权计算后进行拼接，拼接后的向量矩阵输入到融合 Focal Loss 的 CRF 层进行解码，最终得到每个 token 的对应标签。

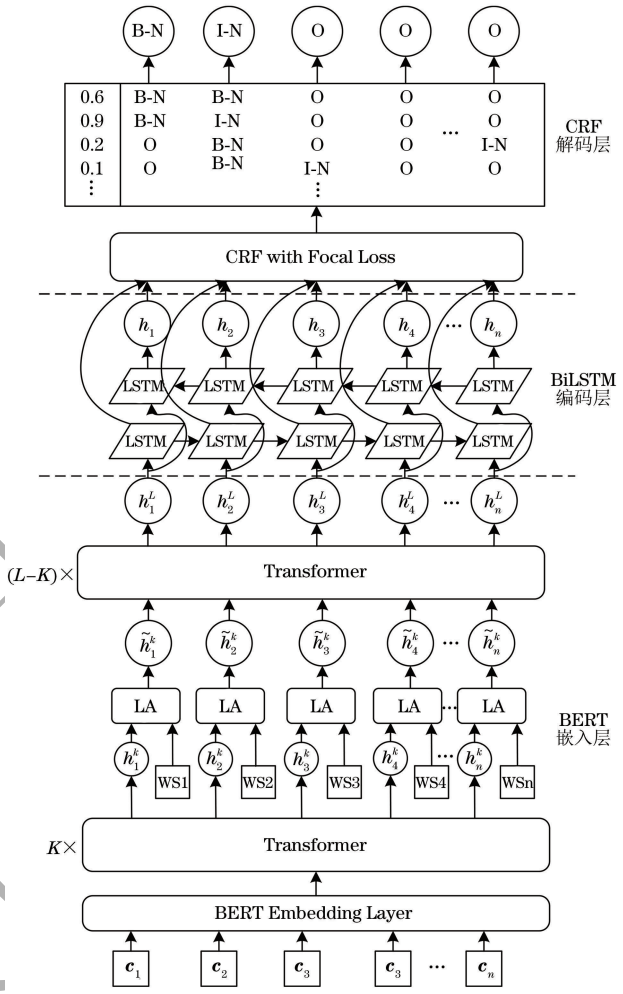



**图 1 模型结构**

**Fig. 1 Structure of the model**

### 2.1 嵌入层

嵌入层采用 BERT 预训练模型进行词嵌入编码，其基本单元由多头自注意力机制层和前馈神经网络即 Transformer 编码器组成。嵌入向量输入编码器单元后，首先进入多头自注意力机制层，通过残差连接和归一化操作处理，将该向量输入到前馈神经网络中，之后再经过一层残差连接和归一化操作，将嵌入向量输入到下一个 Transformer 编码器单元中。

#### 2.1.1 字词对序列

在本文收集的矿井提升机故障领域数据中，包含了大量的领域词汇。因此，为了增强训练效果，首先通过人工收集矿井提升机领域专业词汇形成初始领域词典，然后为该词典中专业词汇的每个字符建立词典，同时进行向量化处理。在此基础上，根据字典自动地为每一个字符匹配出一个隐含的词库，并将这些词库生成的字词库输入到该模型中进行训练。

本文采用词集合方式为每个字建立词典。按照字在匹配单词中的位置将单词集分成字符位于词的

头部(B)、字符位于词的内部(I)和单个字符(S)3 类词集合标签。矿井提升机故障领域文本 $c=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}\in W$ 中的字符 $\boldsymbol{c}_i$ 所匹配的词集合如式(1)～式(3)所示,其中 $W$ 指的是由手工构建的初始领域词典,$\omega$ 指的是领域词典中存在的词语,$i$ 指的是各字符的位置,$j$ 和 $k$ 分别表示字符开始和结束位置。

$$B(\boldsymbol{c}_i)=\{\omega_{i,k},\forall\omega_{i,k}\in W,\ i<k<n\} \tag{1}$$

$$I(\boldsymbol{c}_i)=\{\omega_{j,k},\forall\omega_{j,k}\in W,\ 1<j<i<n\} \tag{2}$$

$$s(\boldsymbol{c}_i)=\{\boldsymbol{c}_i,\exists\boldsymbol{c}_i\in w\} \tag{3}$$

如图 2 所示,“齿轮润滑油膜破裂”一句中的字符“润”通过式(1)匹配到的词集合为 $B=${润滑,润滑油,润滑油膜},设定匹配词有 3 个,不足则用<PAD>代替,如果在领域字典中找不到与某个字相匹配的字,那么这个字汇集就被定义为“None”。

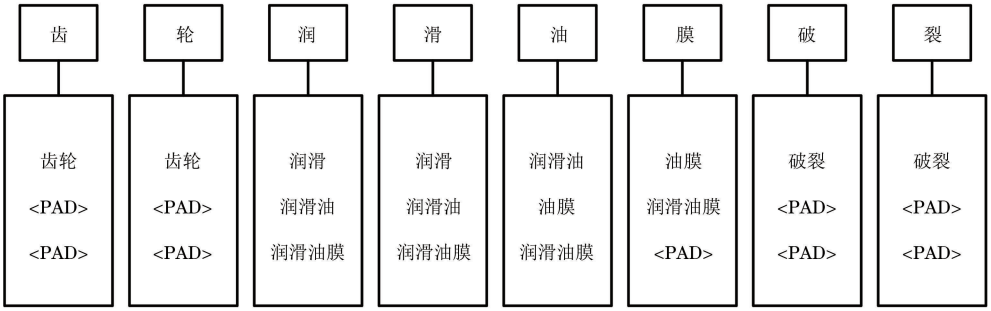



**图 2 字匹配词典**

**Fig. 2 Word matching dictionary**

对于给定的中文句子 $\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}$,利用前述构建的词典,为 $c$ 中的每个字符 $\boldsymbol{c}_i$ 匹配出对应的专业词汇,每个字符和包含该字符的词汇组成字词对集合,表示为 $S_i^{\mathrm{w}}=\{s_1^{\mathrm{w}},s_2^{\mathrm{w}},\cdots,s_i^{\mathrm{w}}\}$,其中 $i$ 表示第 $i$ 个字符匹配到的字词对,而每个字符是可以匹配到多个词汇的,每个 $S_i^{\mathrm{w}}$ 实际上是每个字符匹配到的词汇的小集合。

2.1.2 词典适配器

词典适配器模块将词汇信息注入到 BERT 中,该模块由特征词向量、双向注意力机制组成,如图 3 所示。

对于给定的 $\boldsymbol{c}_i$,将其构造出字词对向量表示为 $(\boldsymbol{c}_i,w_{i,j})$,$\boldsymbol{c}_i$ 表示为第 $i$ 个位置的字向量,$\boldsymbol{W}_{i,j}=\{w_{i,1},w_{i,2},\cdots,w_{i,j}\}$ 表示为字符 $i$ 所对应的 $j$ 个词汇的词向量。使用注意力分数 $\alpha_i$ 对 value 进行加权求和,从而获得 $\boldsymbol{c}_i$ 对应所有词汇融合后的词特征 $z_i^{\mathrm{w}}$,如式(4)所示。

$$z_i^{\mathrm{w}}=\sum_{j=1}^{m}\alpha_i\boldsymbol{W}_{i,j} \tag{4}$$

词汇信息可当作标识符直接拼接在输入之后,与原始文本起始字信息共享相同的位置嵌入,且和原始文本中的字信息共享位置 ID。匹配到的词汇信息如图 4 所示,给定中文句子“齿轮润滑油膜破裂”,经过上述方法把词向量注入到 BERT 中。

将字符特征 $h_i^{\mathrm{c}}$ 与词特征 $z_i^{\mathrm{w}}$ 相加即如式(5)所示。

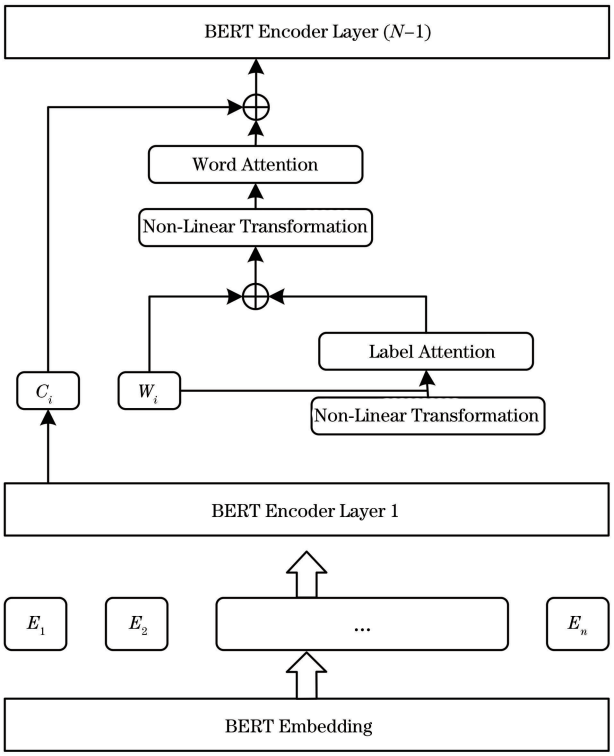



**图 3 词典适配器**

**Fig. 3 Lexicon adapter**

[CLS]齿轮断裂[SEP]立即更换[SEP][元件][现象][措施]

1 2 3 4 5 6 7 8 9 10 11 2 4 9

**图 4 词汇信息**

**Fig. 4 Vocabulary information**

$$\tilde{h}_i=h_i^{\mathrm{c}}+z_i^{\mathrm{w}} \tag{5}$$

假设第 $k$ 层 Transformer Encoder 的输出为 $h_k=\{h_{k,1},h_{k,2},\cdots,h_{k,n}\}$,使用词典适配器模块,在

第 $k$ 层与第 $k+1$ 层 Transformer Encoder 之间注入词汇信息，则在第 $k$ 层可得到具有字词特征信息的向量集合 $\tilde{h}^k$，如式(6)所示。

$$\tilde{h}^k = \mathrm{LA}(h_i^k, W_i) \tag{6}$$

### 2.2 编码层

BiLSTM 目前是命名实体识别研究中应用最为广泛的模型之一，该模型通过将输入序列分别从前向后和从后向前输入 2 个 LSTM 单元，其中前向 LSTM 层获取时间步 $t$ 之前序列的语义信息 $\overrightarrow{h}_t$，反向 LSTM 层获取时间步 $t$ 之后序列的语义信息 $\overleftarrow{h}_t$。最终将前向序列信息 $\overrightarrow{h}_t$ 和反向序列信息 $\overleftarrow{h}_t$ 相结合，所得到的结果[ $\overrightarrow{h}_t$，$\overleftarrow{h}_t$ ]作为双向 LSTM 模型在时间步 $t$ 的输出。这样，模型就能够同时利用当前时刻之前和之后的上下文信息进行预测和分类。

### 2.3 解码层

CRF 是一种基于概率图模型的无监督学习方法[19]。CRF 的核心思想是将输入序列作为条件随机场的观测序列，将标注序列作为隐状态序列，然后基于这些序列构建联合概率分布对序列进行标注。CRF 考虑了相邻标记之间的关系，可以更好地捕捉相邻标记之间的依赖关系，校对标签之间的相关性，减少错误标签的输出几率，得到全局最优的标签序列。如采用“BIO”策略进行标注时，B(Begin)指的是实体的起始位置，I(Inside)指的是实体的内部或末端，O(Outside)指的是这个单词是一个非实体词，会存在如下约束：

1)句子是“B-”或“O”开头，而不是“I-”。例如“O I-label”是错误的，命名实体的开头应该是“B-”而不是“I-”。

2)实体“B-label1 I-label2 I-label3”中 label1，label2，label3 应属于同一种实体类别。例如，当“B-”后面跟的为“label1”时，那么后边的“label2”和“label3”也一定是“label1”。

CRF 的损失函数包括 2 种：一种是发射分数，另一种是转移分数。发射分数是 BiLSTM 编码层的输出矩阵 $\boldsymbol{P}_{i,y_i}$，表示单词 $i$ 对应的标签为 $y_i$ 的概率，其定义为式(7)；转移分数是 CRF 层的矩阵 $\boldsymbol{A}_{y_i,y_{i+1}}$，表示标签之间的转移关系；$\tilde{y}$ 表示 $y$ 的预测值，$S(x \mid y)$ 表示该句所有可能的标签序列。损失函数定义如式(8)所示。

$$S(x \mid y) = \sum \boldsymbol{P}_{i,y_i} + \sum \boldsymbol{A}_{y_i,y_{i+1}} \tag{7}$$

$$L_{\mathrm{CRF}} = \sum_{\tilde{y} \in Y_x} \mathrm{e}^{S(x,\tilde{y})} - S(x,y) \tag{8}$$

Focal Loss 是交叉熵损失(CE)的改进版本，最初被用来处理目标检测中的不平衡问题，后来也应用于文本分类任务中。Focal Loss 通过引入 $\alpha$ 来控制正负样本对整体损失的权重，对数量较多的负类样本赋予以较低的权重，当 $\alpha$ 取值较小时，负类样本的权重程度会更大，但它难以控制难分类样本和易分类样本的权重。因此选择引入一个额外的调制因子 $\gamma$，以限制易分类的样本损失占总损失的比重，其中 $(1-P(y \mid x))^{\gamma}$ 是调制系数，调制系数可以降低易分类样本的损失函数贡献，样本被正确分类的概率越高，其系数越小，那么它对损失函数的贡献就越小[20]。当 $\gamma = 0$ 时，Focal Loss 便为最标准的交叉熵损失函数，当 $\gamma$ 增加时，调制系数增加。参数 $\gamma$ 平滑地调节了易分类样本的比例，能增强调制系数的影响。因此，与 CRF 传统的交叉熵损失函数相比，Focal Loss 可使模型更加关注少类样本和难以分类的样本。Focal Loss 定义如式(9)所示。

$$L_{\mathrm{Focal}} = -\alpha(1-P(y \mid x))^{\gamma} \ln(P(y \mid x)) \tag{9}$$

在 Focal Loss 的 2 个参数中，首先添加参数 $\gamma$ 对实体识别效果提升较多，再添加 $\alpha$ 会有微小提升，若只添加 $\alpha$ 虽然可以平衡正负样本的重要性，但是无法解决简单与困难样本的问题，所以 Focal Loss 的核心是参数 $\gamma$[21]。在训练过程中，$\alpha$ 和 $\gamma$ 可以通过交叉验证等方法进行调节，以达到最佳的性能。

## 3 实验

### 3.1 数据集的构建

本文收集了西北某大型矿业公司的矿井提升机维修记录、60 余篇相关论文和矿井提升机专业书籍中的故障信息作为实验数据集。根据样本数据特点和下游任务需要，定义了故障现象、故障原因、故障位置、维修措施和故障影响 5 类实体。

故障位置指故障发生的部位，包含系统、部件和元件层面；故障现象指设备故障的表现形式，包含机械层面的机械器件损坏和电气层面的元件状态异常等；故障原因指引起故障的设计、制造、操作和维修等有关因素；维修措施指在设备的系统、部件或零件发生磨损、性能下降以至失效之后，为了将其恢复到原来的可用状态，所采取的各种校正、调整和修补措施；故障影响指可能发生的设备失效或某种故障模式下对应的严重性相对度量。表 1 所示为故障实体类别信息。

计算机工程 www.ccice06.com

表 1 故障实体类别信息

Table 1 Faulty entity category information

| 实体类别 | 标签 | 举例 |
|---|---|---|
| 故障位置 | FLN | 齿轮箱 |
| 故障现象 | FPN | 齿面胶合 |
| 故障原因 | FCE | 滤油器堵塞 |
| 修护措施 | RMS | 清洁 |
| 故障影响 | FET | 报警 |

本文使用 Label Studio 标注工具对样本数据进行标注，采用“BIO”策略处理标注的实体，实体标签由实体边界和实体类别组成。如图 5 所示，在句子“连接螺栓出现松动”一句中，“螺栓”一词被标记为“故障位置”这一标签，“螺”字作为首字母，由“B-”作开头，以其标签的英文简写“FLN”作为后缀。

| 连 | 接 | 螺 | 栓 | 出 | 现 | 松 | 动 |
|---|---|---|---|---|---|---|---|
| 0 | 0 | B-FLN | I-FLN | 0 | 0 | B-FPN | I-FPN |

图 5 BIO 标注示例

Fig. 5 Example of BIO labeling

经过如上标注得到实体数量共 4 597 个，其实体类型分布占比如图 6 所示，可以发现不同实体类型间的数量占比差异很大，数据不平衡现象较为严重。

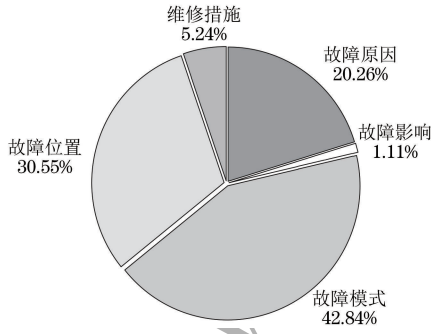



图 6 实体比例

Fig. 6 Proportion of entities

### 3.2 领域词汇

本文通过收集矿井提升机书籍中的专业词汇，得到初始领域词典如表 2 所示。

表 2 初始领域词典

Table 2 Initial field dictionary

| 实体类别 | 实例 |
|---|---|
| 故障位置 | 液压站、制动器、闸瓦等 |
| 故障模式 | 过热、漏液、异常震动等 |
| 维修措施 | 更换受损部件、清洗设备等 |
| 故障原因 | 使用不当、设备老化等 |
| 故障影响 | 设备寿命降低、生产停滞 |

### 3.3 实验设置与评估指标

实验环境使用的是 Ubuntu 操作系统，GPU 选用 NVIDIA GeForce GTX 2080Ti，基于 PyTorch 框架，使用 Python3.6 版本进行搭建，其中 BERT 使用的是谷歌提供的中文 BERT 预训练模型 BERT-Base-Chinese。在 Focal Loss 损失函数中，经过文献参考和实际实验结果，选取平衡因子 $\alpha=0.96$，$\gamma=2$。超参数设置如表 3 所示。

表 3 超参数设置

Table 3 Hyper parameter setting

| 参数 | 设置 |
|---|---|
| Max sentence lenth | 256 |
| Epoch | 100 |
| Dropout | 0.2 |
| Batch size | 64 |
| Word embedding size | 50 |
| Character embedding size | 50 |
| Learning rate | 0.005 |
| Learning rate decay | 0.05 |

采用精确率($P$)、召回率($R$)和 F1 值($F_1$)作为评估模型性能的指标，以 F1 值对模型性能进行综合评价。计算方法如式(10)～式(12)所示。

$$P=\frac{n_{TP}}{n_{TP}+n_{FP}}\times 100\% \tag{10}$$

$$R=\frac{n_{TP}}{n_{TP}+n_{FN}}\times 100\% \tag{11}$$

$$F_1=\frac{2PR}{P+R}\times 100\% \tag{12}$$

式中：$n_{TP}$ 是指被成功识别到的正确实体数目；$n_{FN}$ 是指未被成功识别到的正确实体数目；$n_{FP}$ 是指非实体的数目。

### 3.4 实验结果

为了验证 BERTw-BiLSTM-CRF-FL 模型的识别效果，将 BERT-BiLSTM-CRF 作为基线模型[22]，针对数据不平衡问题，采用了 3 种方法验证本文方法的有效性：首先通过 SMOTE 方法的模型和融入 Focal Loss 的模型[23]比较，验证数据层方法与算法层方法的优劣；其次通过融入 Dice Loss 损失函数优化的模型与融入 Focal Loss 的模型比较，探究不同损失函数对整个模型的适配效果[24]；最后在融入 Focal Loss 的模型的基础上融入领域词典[25-27]，以验证本文提出的 BERTw-BiLSTM-CRF(FL)模型的最终识别效果。实验结果如表 4 所示，其中，DL 表示融入 Dice Loss 损失函数，FL 表示融入 Focal Loss 损失函数。

计算机工程 www.cie06.com

表 4 各模型性能对比

Table 4 Comparison of the performance of each model %

| 模型 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| BERT+BiLSTM+CRF | 81.52 | 82.62 | 82.86 |
| BERT+BiLSTM+CRF(SMOTE) | 89.80 | 85.62 | 87.66 |
| BERT+BiLSTM+CRF(DL) | 72.20 | 82.89 | 77.22 |
| BERT+BiLSTM+CRF(FL) | 86.52 | 90.43 | 88.43 |
| BERT-FLAT+BiLSTM+CRF(FL) | 87.36 | 86.95 | 87.15 |
| BERTw+BiLSTM+CRF(FL) | 90.89 | 87.43 | 89.13 |

如表 4 所示，在不平衡的数据集上，相较于基线模型，采用 SMOTE 的数据层处理方法可以明显提高实体识别的效果，但是，在算法层面使用引入 Focal Loss 损失函数优化的模型，其识别效果更优，比主流模型 BERT-BiLSTM-CRF 提高了 5.57 个百分点；在算法层面，同样是对模型损失函数的改进，使用融入 Focal Loss 损失函数优化的模型比融入 Dice Loss 损失函数优化的模型效果更好；在融入领域词典后，本文提出的 BERTw-BiLSTM-CRF-FL 模型进一步提高了实体识别精度，相对于 FLAT 方法也有着更好的结果表现。

### 3.5 Focal Loss 参数组合优化实验

为了更好地了解调制系数与平衡因子对模型性能的影响，本文通过比较实验，得到了调制系数与平衡因子不同组合下的实验结果，并将其性能变化显示在图 7 折线中。可以看出，当 $\gamma=1.5$ 与 $\gamma=2$ 两条曲线趋势较为一致，且 $\alpha=0.96$ 时 F1 值都达到最高，而 $\alpha=0.96$、$\gamma=2$ 时，模型能够获得最好的性能。

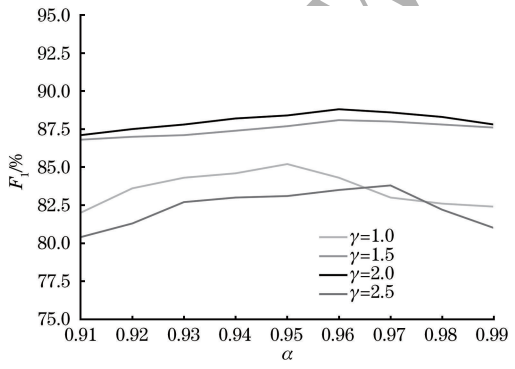



图 7 Focal Loss 不同参数组合对比

Fig. 7 Comparison of different parameter combinations for Focal Loss

### 3.6 细粒度实验结果分析

为了更深入探究融入 Focal Loss 和词典对不同实体类别的具体影响，将 5 类实体的实验结果进行对比，如表 5 所示，可以发现将 CRF 层的损失函数改为 Focal Loss 后，实体数量较大的“故障位置”、“故障现象”和“故障原因”识别准确度基本持平或者会略微下降，而实体数量较小的“修护措施”和“故障影响”识别精度则明显有着较大的提升，因而总体识别准确度有一定提升。说明 Focal Loss 在不平衡实体的识别中，大样本实体类别识别精度基本保持不变，其影响最大的是小样本实体类别识别的准确度，小样本标签实体识别准确度的大幅度提升带来了整体实体标签识别准确度的提升。而在词典融入后，所有实体类别识别准确度均有一定提升，说明在专业领域中融入领域词典能够显著提升识别效果。

表 5 各模型细粒度识别实验结果

Table 5 Experimental results of fine-grained recognition for each model %

| 实体类别 | 指标 | BERT+BiLTM+CRF | BERT+BiLSTM+CRF(FL) | BERTw+BiLSTM+CRF(FL) |
|---|---|---|---|---|
| 故障位置 | $P$ | 91.91 | 92.17 | 92.71 |
| | $R$ | 85.84 | 85.15 | 90.03 |
| | $F_1$ | 88.77 | 88.52 | 91.35 |
| 故障现象 | $P$ | 87.27 | 92.39 | 87.72 |
| | $R$ | 85.70 | 81.75 | 88.63 |
| | $F_1$ | 86.47 | 86.74 | 88.17 |
| 故障原因 | $P$ | 92.24 | 89.80 | 92.17 |
| | $R$ | 83.84 | 85.62 | 85.65 |
| | $F_1$ | 87.84 | 87.66 | 88.79 |
| 修护措施 | $P$ | 55.61 | 68.29 | 68.76 |
| | $R$ | 77.79 | 72.39 | 73.88 |
| | $F_1$ | 64.86 | 70.27 | 71.23 |
| 故障影响 | $P$ | 78.56 | 76.72 | 77.79 |
| | $R$ | 48.39 | 55.29 | 55.61 |
| | $F_1$ | 59.88 | 64.21 | 64.86 |

## 4 结束语

针对机械设备故障领域命名实体识别中样本数据不平衡的问题，本文在 CRF 层引入 Focal Loss 函数，有效缓解了数据不平衡问题，在此基础上，嵌入了领域专业词典，进一步提高了实体识别效果，对细粒度实体的实验结果分析可以清晰看出上述 2 项工作对于不同标签实体识别的具体影响。实验表明，本文提出的 BERTw+BiLSTM+CRF(FL)方法对于机械设备故障的实体识别效果显著，在自建的矿井提升机数据集上 F1 值达到 89.13%，以此可形成较高质量的机械设备故障知识图谱，为故障定位分析、维修方案生成、故障知识问答等下游任务提供支

计算机工程 www.ecice06.com

撑。未来可以进一步探索机械设备故障诊断的其他关键问题，例如关系抽取和知识融合等问题，并尝试应用于其他专业领域，以提高模型的泛化性。

## 参考文献


[1] 倪骥，王宇嘉，赵博. 融合部首特征和 BERT 的汽车生产设备故障领域命名实体识别[J]. 小型微型计算机系统，2024，45(6)：1370-1375.
NI J，WANG Y J，ZHAO B. Named entity recognition for automotive production equipment fault domain fusing radical feature and BERT [J]. Journal of Chinese Computer Systems，2024，45(6)：1370-1375.(in Chinese)

[2] 江婉榕. 基于文本数据增强和特征选择的不平衡分类算法研究[D]. 合肥：中国科学技术大学，2022.
JIANG W R. Research on imbalanced classification algorithm based on text data augmentation and feature selection[D]. Hefei：University of Science and Technology of China，2022.(in Chinese)

[3] ZHOU G D，SU J. Named entity recognition using an HMM-based chunk tagger[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg，USA：Association for Computational Linguistics，2001：473-480.

[4] AKBIK A，BERGMANN T，BLYTHE D，et al. FLAIR：an easy-to-use framework for state-of-the-art NLP [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg，USA：Association for Computational Linguistics，2019：54-59.

[5] LIU T Y，JIANG Y E，MONATH N，et al. Autoregressive structured prediction with language models[C]//Proceedings of EMNLP 2022. Stroudsburg，USA：Association for Computational Linguistics，2022：993-1005.

[6] SHAH A，VITHANI R，GULLAPALLI A，et al. Finer：financial named entity recognition dataset and weak-supervision model[EB/OL]. [2023-06-10]. https://arxiv.org/pdf/2302.11157.

[7] 陈旭，刘鹏鹤，孙毓忠，等. 面向不均衡医学数据集的疾病预测模型研究[J]. 计算机学报，2019，42(3)：596-609.
CHEN X，LIU P H，SUN Y Z，et al. Research on disease prediction model for unbalanced medical datasets[J]. Journal of Computing，2019，42(3)：596-609.(in Chinese)

[8] HAN H，WANG W，MAO B. Borderline-SMOTE：a new over-sampling method in imbalance data sets learning[C]//Proceedings of ICIC 2005. Berlin，Germany：Springer，2005：878-887.

[9] LIU J. Importance-SMOTE：a synthetic minority oversampling method for noisy imbalanced data[J]. Soft Computing，2022，26(3)：1141-1163.

[10] MALDONADO S，VAIRETTI C，FERNANDEZ A，et al. FW-SMOTE：a feature-weighted oversampling approach for imbalanced classification [J]. Pattern Recognition，2022，124：108511.

[11] CAO K，WEI C，GAIDON A，et al. Learning imbalanced datasets with label-distribution-aware margin loss[EB/OL]. [2023-06-10]. https://arxiv.org/abs/1906.07413.

[12] LI X Y，SUN X F，MENG Y X，et al. Dice Loss for data-imbalanced NLP tasks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg，USA：Association for Computational Linguistics，2020：465-476.

[13] LIN T Y，GOYAL P，GIRSHICK R，et al. Focal Loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2017：2980-2988.

[14] ZHANG Y，YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). Stroudsburg，USA：Association for Computational Linguistics，2018：1554-1564.

[15] LI X N，YAN H，QIU X P，et al. FLAT：Chinese NER using flat-lattice Transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg，USA：Association for Computational Linguistics，2020：6836-6842.

[16] MA R，PENG M，ZHANG Q，et al. Simplify the usage of lexicon in Chinese NER[EB/OL]. [2023-06-10]. https://arxiv.org/abs/1908.05969.

[17] ZHU W，CHEUNG D. Lex-BERT：enhancing BERT based NER with lexicons[EB/OL]. [2023-06-10]. https://arxiv.org/abs/2101.00396.

[18] LIU W，FU X Y，ZHANG Y，et al. Lexicon enhanced Chinese sequence labeling using BERT adapter [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1：Long Papers). Stroudsburg，USA：Association for Computational Linguistics，2021：5847-5858.

[19] 司逸晨，管有庆. 基于 Transformer 编码器的中文命名实体识别模型[J]. 计算机工程，2022，48(7)：66-72.
SI Y C，GUAN Y Q. Chinese named entity recognition model based on Transformer encoder[J]. Computer Engineering，2022，48(7)：66-72.(in Chinese)

[20] 贾鑫桐. 不平衡样本的多标签毒性文本分类——基于改进损失函数的神经网络模型[D]. 上海：上海财经大学，2021.
JIA X T. Multi-label toxic text classification of unbalanced samples—neural network model based on improved loss function[D]. Shanghai：Shanghai University of Finance and Economics，2021.(in Chinese)

[21] 张汝佳，代璐，王邦，等. 基于深度学习的中文命名实体识别最新研究进展综述[J]. 中文信息学报，2022，36(6)：20-35.
ZHANG R J，DAI L，WANG B，et al. Recent advances of Chinese named entity recognition based on deep learning[J]. Journal of Chinese Information Processing，2022，36(6)：20-35.(in Chinese)

[22] JIA C，SHI Y F，YANG Q R，et al. Entity enhanced BERT pre-training for Chinese NER[C]//Proceedings of EMNLP 2020. Stroudsburg，USA：Association for Computational Linguistics，2020：6384-6396.

[23] 郭渊博，李勇飞，陈庆礼，等. 融合 Focal Loss 的网络威胁情报实体抽取[J]. 通信学报，2022，43(7)：85-92
GUO Y B，LI Y F，CHEN Q L，et al. Fusion of Focal Loss's cyber threat intelligence entity extraction[J]. Journal on Communications，2022，43(7)：85-92.(in Chinese)

[24] 吴广硕，樊重俊，陶国庆，等. 基于 LEBERT-BCF 的电子病历实体识别[J]. 计算机时代，2023(2)：92-97.
WU G S，FAN C J，TAO G Q，et al. Entity recognition of electronic medical records based on LEBERT-BCF [J]. Computer Era，2023(2)：92-97.(in Chinese)

[25] LEE L H，LU C H，LIN T M. NCUEE-NLP at SemEval-2022 task 11：Chinese named entity recognition using the BERT-BiLSTM-CRF model [C]//Proceedings of the 16th


计算机工程 www.ecice06.com

International Workshop on Semantic Evaluation. Stroudsburg, USA: Association for Computational Linguistics, 2022: 1597-1602.

[26] 张应成,杨洋,蒋瑞,等. 基于 BiLSTM-CRF 的商情实体识别模型[J]. 计算机工程,2019,45(5):308-314.
ZHANG Y C, YANG Y, JIANG R, et al. Commercial intelligence entity recognition model based on BiLSTM-CRF [J]. Computer Engineering, 2019, 45(5): 308-314. (in Chinese)

[27] 买买提阿依甫,吾守尔·斯拉木,帕丽旦·木合塔尔,等. 基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别[J]. 计算机工程,2018,44(8):230-236.
Maimaitiayifu, SILAMU Wushouer, MUHETAER Palidan, et al. Uyghur named entity recognition based on BiLSTM-CNN-CRF model[J]. Computer Engineering, 2018, 44(8): 230-236. (in Chinese)

编辑 金胡考

计算机工程 www.ecice06.com