文章编号：1000-3673（2022）10-3913-10　　中图分类号：TM 721　　文献标志码：A　　学科代码：470·40

面向知识图谱构建的设备故障文本实体识别方法

田嘉鹏，宋辉，陈立帆，盛戈皞，江秀臣

（上海交通大学电子信息与电气工程学院，上海市 闵行区 200240）

Entity Recognition Approach of Equipment Failure Text for Knowledge Graph Construction

TIAN Jiapeng, SONG Hui, CHEN Lifan, SHENG Gehao, JIANG Xiuchen

(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Minhang District, Shanghai 200240, China)

ABSTRACT: Technicians have accumulated plenty of failure texts, which contain essential entity information, during the operation and maintenance of power equipment. However, such text has fuzzy entity boundaries and contains many professional terms, resulting in the traditional entity recognition methods with low training efficiency and poor performances. Therefore, an integrated algorithm of BERT based BiRNN with CRF (I-BRC) is proposed. This algorithm employs a word embedding model to convert each word in the text into the embedding vector sequences to avoid the error accumulation caused by word segmentation. The recurrent neural networks with probability graph models are introduced to extract sequence features from the text. The multiple single-type entity recognizers are integrated to learn the features of different entity types independently, and a parallel pre-training mechanism is employed to improve the training efficiency. Finally, the recognition results are integrated by the multi-type recognizer. Besides, adjusting the single-type entity recognizers can flexibly respond to different power equipment failure texts, avoiding repeated training and saving computation resources. Experiments show that the proposed algorithm reached a stable state after 3 iterations, which significantly improves the training efficiency with its F1 score, precision and recall as 88.0%、86.8% and 89.2% respectively. Compared with the traditional models, the performance is improved by 19.5% to 28.8%, which verifies the effectiveness and feasibility of the proposed model.

KEY WORDS: power equipment; failure case; Chinese entity recognition; knowledge graph; neural network

摘要：电力设备在运行维护中积累了大量包含重要实体信息的故障文本，然而文本实体边界模糊、术语较多等特点导致传统实体识别方法训练效率低下，效果难以提升。为此，该文提出一种新的实体识别方法 I-BRC(integrated algorithm of BERT based BiRNN with CRF)。该方法采用字嵌入模型将文本逐字转化为字向量序列以避免分词处理带来的误差累积；利用循环神经网络与概率图模型对文本的序列特征信息进行抽取；集成多个单一类型实体识别器分别独立学习不同类型实体的特征并采用并行预训练机制提升算法训练效率；最后利用多类型识别器对识别结果进行整合。此外，通过调整单一类型实体识别器可以灵活机动地应对不同电力设备的实体识别任务，避免重复训练，节省计算资源。实验表明，所提出的 I-BRC 仅需 3 次迭代就可收敛，训练效率大幅度提升；且该模型的 F1 值、精确率、召回率分别达到了 88.0%、86.8%与 89.2%，相比传统模型性能提升了 7.5%~29.3%，验证了所提模型的有效性与可行性。

关键词：电力设备；故障案例；中文实体识别；知识图谱；神经网络

DOI：10.13335/j.1000-3673.pst.2021.1886

0　引言

随着电网规模日趋扩大，用户对电网供电可靠性的需求日益增加，诸如变压器、发电机与气体绝缘开关柜(gas insulated switchgear，GIS)等电力设备的智能运行维护方法受到了越来越多学者的重视[1-3]。当前，电力设备的运行维护严重依赖运维人员的人工经验与应变能力，设备故障案例等大量以文本形式记录的知识需要运维人员反复记忆和查询[4-5]。知识图谱是一种基于人工智能技术的知识表示方法[4]，能够有效解决设备运行维护中人工经验无法共享、操作规范性难以确保的问题，在知识挖掘总结[6]、决策生成[7]、辅助故障定位[8]等方面具有很好的优势。作为知识图谱构建最基础的部分，实体识别能将故障文本中的信息进行蒸馏，将非结构化数据转化为结构化数据，有效提升知识存储与查询效率。

基金项目：国家重点研发计划项目(2020YFB1709701)；国家电网有限公司科技项目(5700-202119174A-0-0-00)。

Project Supported by the National Key R&D Program of China (2020YFB1709701); State Grid Science and Technology Program of China(5700-202119174A-0-0-00).

准确高效地对电力设备故障文本中各类实体信息进行挖掘并加以分析，对设备维护的智能决策起到关键的作用。然而，这类文本通常语句较长、实体边界模糊，且包含了大量数值与特殊符号[9]，导致传统实体识别模型在处理这类文本时准确度[10]与训练效率低下。其次，传统的中文实体识别方法需要对文本进行分词处理，分词模型自身的误差会累积到实体抽取模型上，导致算法的效果难以提升。另一方面，不同设备的巡检人员遵守不同的运维准则，所记录的文本侧重于不同的实体信息，采用传统实体识别方法通常需要根据各类文本特点重新训练参数，不可避免地造成了大量运算资源的浪费。

为解决上述问题，国内外学者结合信息与计算机技术，在抽取电力文本实体信息方面取得了一定的进展。文献[11-12]前瞻性地对电力领域文本实体抽取的概念与方法进行了概述，总结了实体抽取在电力领域知识图谱构建技术中的作用与地位，为该领域后续的研究构建框架；随后，文献[13]创新性地利用基于条件随机场的双向长短时记忆网络(conditional random field based bidirectional long short-term memory，BiLSTM-CRF)结构有效地对电力缺陷文本中实体信息进行抽取，但分词处理仍给模型带来了误差累积；文献[14]采用同样的结构对系统二次设备的缺陷文本进行实体抽取，与传统序列模型和注意力模型相比识别效果提升了近 20%，但设计的实体类别较少，实体边界模糊问题不明显；文献[15]在该结构的基础上增添了变换器模型(transformer)来更好地对变电站文本序列进行信息提取，由于引入注意力机制，模型效果提升了 10.6%；文献[16]则从语言表示的角度进行改进，采用预训练的嵌入模型使得算法能够更好地理解原始文本样本的语义。

上述研究对本文起到了重要的参考依据，但是现有的研究仍无法有效解决实体边界模糊以及模型灵活应用的问题。因此，本文提出了一种新的实体识别算法，其实现方法与特点如下所示：

1）采用 BERT(bidirectional encoder representation from transformers)嵌入模型对文本中文字进行嵌入式表示，舍弃分词处理避免误差累积，同时可在文本中任意 2 个字符之间直接建立注意力联系，能够更好地理解长句的语义。

2）集成多个单一类型实体识别器，各自专注于不同类型实体信息的抽取，有效解决了实体边界模糊的问题，采用并行预训练机制提升了模型训练效率，具有更高的容错率与灵活性。

3）所有实体识别器均采用双向循环神经网络(bidirectional recurrent neural network，BiRNN)对文本序列进行特征解析，并用条件随机场概率图模型(conditional random field，CRF)对序列标注进行约束。

4）利用多类型实体识别器对所有单一类型实体识别器的结果进行整合，最终抽取出设备文本中的各类实体信息。

通过实验验证，本文所提出的条件随机场下基于 BERT 双向循环网络集成算法(integrated algorithm of BERT based BiRNN with CRF，I-BRC)在训练的过程中能够迅速达到收敛，节省大量计算资源；另一方面，相比其他传统实体识别算法，该算法具有更高的精确率与召回率，证实了该算法的有效性与可行性。

1 中文电力设备故障文本实体识别

电力设备故障文本实体识别是对电力设备故障文本中各种类型的实体信息进行抽取，并对不同类型的实体例如故障位置、缺陷类型、设备种类等进行分类。本文将实体识别看作序列标注任务，通过自动标注文本中每个汉字字符来实现实体信息的抽取与分类，无需事先对文本进行分词处理，有效避免了误差累积。给定一个长度为 n 的中文电力设备故障文本 $\boldsymbol{C}_{\mathrm{S}}$，并表示为 $\boldsymbol{C}_{\mathrm{S}}=<c_{\mathrm{s}}^{1},c_{\mathrm{s}}^{2},\cdots,c_{\mathrm{s}}^{n}>$，其中 c_{s}^{i} 为 $\boldsymbol{C}_{\mathrm{S}}$ 中的第 i 个字符。则实体识别模型可以表示为一个非线性映射 $\kappa(\bullet)$，使得该文本所对应的标签序列 $\boldsymbol{T}_{\mathrm{S}}$ 表示为

$$\boldsymbol{T}_{\mathrm{S}}=\kappa(\boldsymbol{C}_{\mathrm{S}}) \tag{1}$$

式中 $\boldsymbol{T}_{\mathrm{S}}=<t_{\mathrm{s}}^{1},t_{\mathrm{s}}^{2},\cdots,t_{\mathrm{s}}^{n}>$ 与 $\boldsymbol{C}_{\mathrm{S}}$ 长度一致，且 t_{s}^{i} 为文本第 i 个字符所对应的标签。表 1 给出了某电力公司记录的设备故障案例部分文字序列，以及各个字符对应的标签，并展示了文本实体信息的标注方法。

本文采用 BIEO 序列标注法对文本中的实体进行标注，B 代表该字符为一个实体的开始，I 代表该字符在实体的内部，E 代表该字符为一个实体的结束，O 则代表该字符在任意一个实体的外面。另外，还需对实体的类别进行标注，表 1 展示了缺陷位置与缺陷类型 2 种实体，分别用标签 P、TY 来表示。通过映射 $\kappa(\bullet)$，可以得到该文本的标签序列，最终由标签序列进一步识别出位置实体“B 相电缆终端与导体触头连接处”与缺陷类型实体“悬浮放电”。

表 1　电力设备故障文本实体识别序列标注方法

Table 1　Sequence labeling method of entity recognition for power equipment failure text

字符序列	通	过	声	电	联	合	定	位	表	明	，	放	电	源	在	B	相	电	缆
标签序列	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	P-B	P-I	P-I	P-I
实体类型																	缺陷位置		
字符序列	终	端	与	导	体	触	头	连	接	处	，	类	型	为	悬	浮	放	电	。
标签序列	P-I	P-I	P-I	P-I	P-I	P-I	P-I	P-I	P-I	P-E	O	O	O	O	TY-B	TY-I	TY-I	TY-E	O
实体类型								缺陷位置									缺陷类型		

2　故障文本实体识别算法与流程

本文所提出的 I-BRC 故障文本实体识别算法共包含 3 部分：BERT 嵌入层、多个单一类别实体识别器以及多类型实体识别器，如图 1 所示。

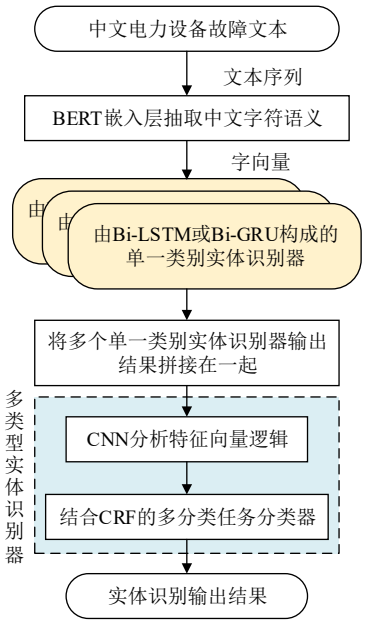


图 1　I-BRC 算法框架图

Fig. 1　Framework of I-BRC algorithm

BERT 嵌入层在文本的任意 2 个汉字之间构建注意力联系，有效提取字与句的语义特征并将文本转化为向量序列，使得算法能够充分理解文本信息；单一类型实体识别器在 CRF 的约束下利用 BiRNN 层分别解析某一特定类型的实体特征，有效应对实体类型边界模糊的问题，并在工程应用中可以模块化调整；多类型实体识别器利用卷积神经网络(convolutional neural network，CNN)分析不同识别器识别结果之间的逻辑关系，并利用分类器对向量序列进行特征解析，最终识别出各类实体信息。

本节首先对所采用的嵌入模型、循环网络与图概率模型进行了简要说明，然后再对 I-BRC 及其训练机制进行了详细的介绍。

2.1　BERT 嵌入层结构

BERT 嵌入模型[17]将文本序列转化为更易被算法理解的向量序列，具有很强的自然语言表示能力，并在实体识别[16-17]、知识问答、文本分类[18]等任务上取得了很好的效果。BERT 由多个 transformer[19]构成(如图 2 所示)，包含大量残差与归一化模块，能有效抑制梯度消失并提升训练速度。该模型利用多头注意力层对故障文本信息进行解析，解决了处理长句时信息遗失的问题。

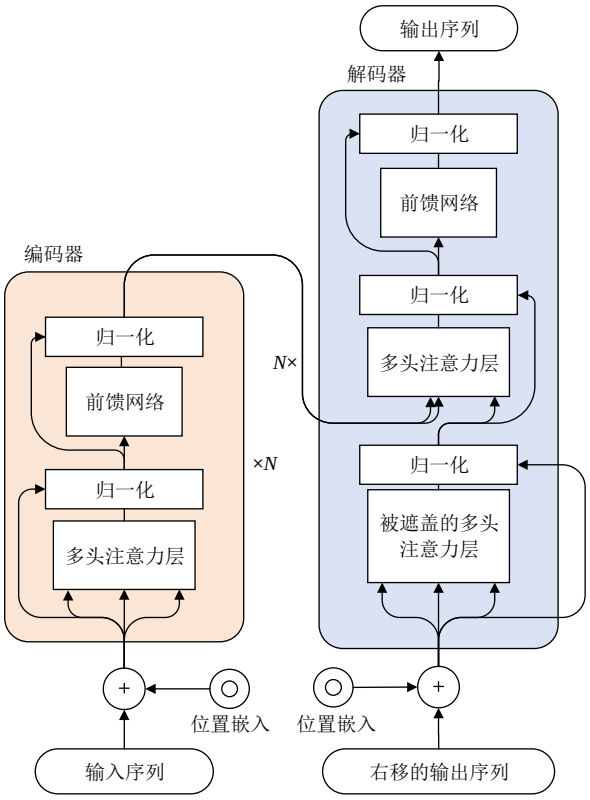


图 2　BERT 嵌入层中的 transformer 结构

Fig. 2　Structure of the transformer in BERT

BERT 嵌入模型可看作为非线性映射 $f_{\text{bert}}(\cdot)$，如式(2)所示：

$$\boldsymbol{E} = f_{\text{bert}}(\boldsymbol{C}_{\text{S}}) \tag{2}$$

式中 $\boldsymbol{E} \in \mathbb{R}^{n\times d}$ 为字向量序列，其中第 i 行为文本第 i 个字符的嵌入向量 $\boldsymbol{E}_i$，维度为 d。

2.2　双向循环神经网络

I-BRC 中各类实体识别器均采用双向循环神经网络对输入的向量序列进行特征解析，可用映射 $f_{\text{rnn}}(\cdot)$ 来表示。为避免梯度消失等问题，采用双向长短时记忆网络(bidirectional long short-term memory network，BiLSTM)与双向门控循环单元(bidirectional gate recurrent unit，BiGRU)作为循环网

络，将输入序列当前信息与此前信息相关联[20]，从而提升特征解析能力[21]。

BiLSTM 利用门控单元与细胞状态实现信息传递。令 t 时刻的输入为 $\boldsymbol{x}_t$，输出为 $\boldsymbol{h}_t$，细胞状态为 $\boldsymbol{c}_{t-1}$。则 BiLSTM 中遗忘门、输入门、输出门的输出 $\boldsymbol{f}_t$、$\boldsymbol{i}_t$ 与 $\boldsymbol{o}_t$ 可用式(3)—(5)表示：

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\mathrm{f}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{f}}) \tag{3}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{\mathrm{i}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{i}}) \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{\mathrm{o}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{o}}) \tag{5}$$

式中：$\sigma(\bullet)$ 为激活函数；$\boldsymbol{W}_{\mathrm{f}}$、$\boldsymbol{W}_{\mathrm{i}}$ 与 $\boldsymbol{W}_{\mathrm{o}}$ 为门控单元的权重；$\boldsymbol{b}_{\mathrm{f}}$、$\boldsymbol{b}_{\mathrm{i}}$ 与 $\boldsymbol{b}_{\mathrm{o}}$ 为偏置分量；$[\bullet,\bullet]$ 表示向量的拼接操作。由此，$\boldsymbol{h}_t$ 以及更新后的细胞状态 $\boldsymbol{c}_t$ 可表示为

$$\boldsymbol{c}_t=\boldsymbol{c}_{t-1}\circ\boldsymbol{f}_t+\boldsymbol{i}_t\circ\tanh(\boldsymbol{W}_{\mathrm{c}}[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_{\mathrm{c}}) \tag{6}$$

$$\boldsymbol{h}_t=\tanh(\boldsymbol{c}_t)\circ\boldsymbol{o}_t \tag{7}$$

式中：$\tanh(\bullet)$ 是 tanh 激活函数；$\circ$ 代表逐项相乘；$\boldsymbol{W}_{\mathrm{c}}$ 与 $\boldsymbol{b}_{\mathrm{c}}$ 分别为细胞状态的权重与偏置。

BiGRU 则在此基础上将遗忘门与输入门合并为更新门，并将细胞状态与输出相结合[21]，从而使得训练时间大幅缩减，但具有较差的泛化能力。为此，本文在构建单一识别器时，综合训练时间与泛化效果为其选取最佳的拓扑结构。而在多类型识别器中，则对比选取了 BiLSTM 作为分类器的解析层。

2.3 条件随机场

本文采用 CRF 概率图模型对模型序列标注进行约束，避免输出标签序列 $\boldsymbol{T}_{\mathrm{S}}$ 内部元素之间相互割裂，丢失元素之间的依赖关系。CRF 是一种典型的判别式概率图模型，通过构建条件概率模型 $P(\boldsymbol{T}_{\mathrm{S}}\mid\boldsymbol{C}_{\mathrm{S}})$[22]，并以损失函数的形式对参数迭代进行约束，从而得到概率最大的输出序列 $\boldsymbol{T}_{\mathrm{S}}$。$P(\boldsymbol{T}_{\mathrm{S}}\mid\boldsymbol{C}_{\mathrm{S}})$ 可由式(8)表示：

$$P(\boldsymbol{T}_{\mathrm{S}}\mid\boldsymbol{C}_{\mathrm{S}})=\frac{1}{Z}\exp\Big(\sum_{i,j}\lambda_j t_j(t_{\mathrm{s}}^{i+1},t_{\mathrm{s}}^{i},\boldsymbol{C}_{\mathrm{S}},i)+\sum_{i,k}\mu_k s_k(t_{\mathrm{s}}^{i},\boldsymbol{C}_{\mathrm{S}},i)\Big) \tag{8}$$

式中：t_j 为转移特征函数，用来刻画 t_{s}^{i+1} 的特征受 t_{s}^{i} 状态的影响；s_k 为状态特征函数，用来刻画 t_{s}^{i} 受 c_{s}^{i} 状态的影响；λ_j 与 μ_k 是特征函数的权重；Z 是归一化因子。

2.4 I-BRC 算法及其训练机制

I-BRC 模型的训练过程主要分为 4 步(如图 3 所示)：首先利用电力设备故障文本对 BERT 嵌入层模型进行无监督训练，从字与句 2 个层面让 BERT 充分理解故障文本；其次，对多个单一类型实体识别器进行并行预训练，识别器分别解析不同类型的实体信息以避免其他实体带来的干扰；随后，对多类型实体识别器进行训练，分析各识别器识别结果之间的逻辑关系并进行整合；最后，将训练得到的权重参数导入到 I-BRC 模型中进行微调，实现故障文本的实体识别任务。

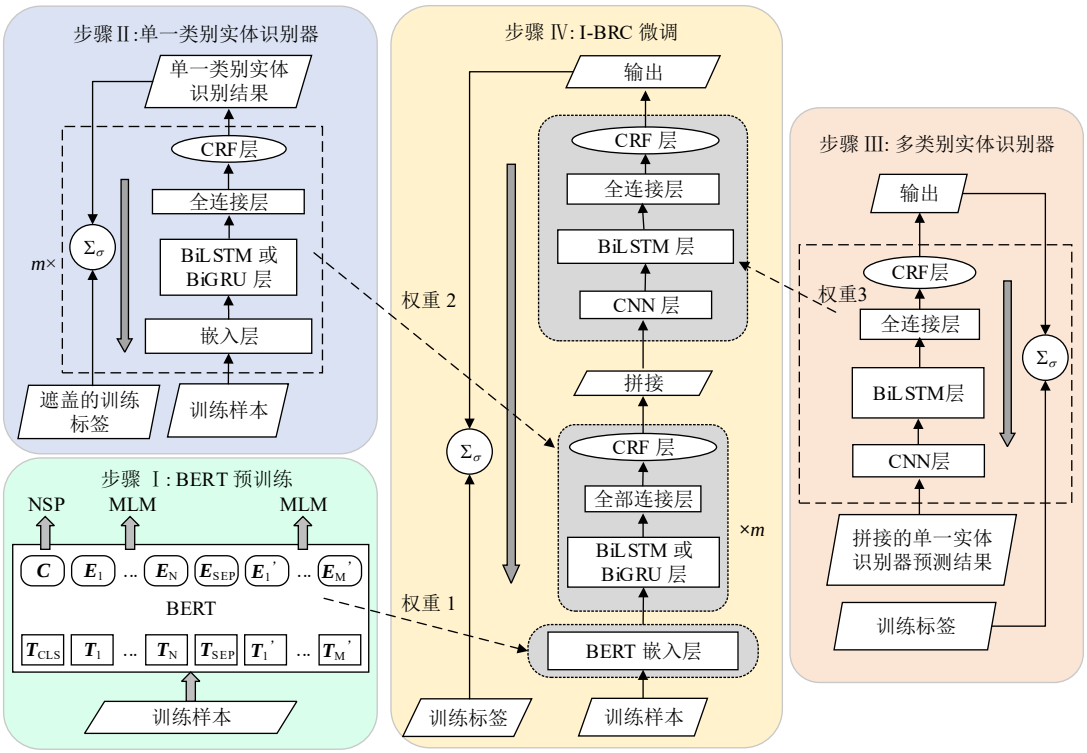


图 3 I-BRC 模型结构与训练过程

Fig. 3 Structure and training process of I-BRC

2.4.1　BERT 嵌入层预训练机制

BERT 嵌入层是 I-BRC 模型最底层的结构，直接决定了整个模型的效果。本文利用 2 个无监督学习任务来实现 BERT 的预训练过程，分别为掩码语言模型任务(masked language model，MLM)与下句预测任务(next sentence prediction，NSP)。

MLM 任务是将文本中随机的 15%的字用[MASK]符号进行遮盖，并用 BERT 模型预测该未知的字符，如图 4 所示。考虑到实际的电力设备故障文本质量难以控制，因此在训练时将 10%选中的汉字随机用其他字代替，以此增加噪声提升模型的鲁棒性。MLM 任务侧重字符层面的信息解析，让 BERT 能够充分理解某字符在语境中的具体含义。

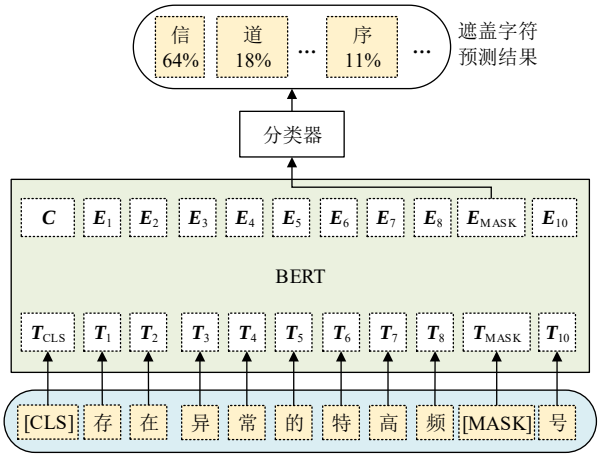


图 4　MLM 任务训练过程

Fig. 4　Training process of MLM task

NSP 任务是用来判断输入的 2 个句子是否为前后句关系。如图 5 所示，在训练时样本序列的第一个字符 $\boldsymbol{T}_{\mathrm{CLS}}$ 被设定为分类字符[CLS]，所对应的输出 $\boldsymbol{C}$ 包含了该样本整体的语义特征。输入的序列包含 2 个句子，并用分隔符[SEP]隔开。将语义特征向量 $\boldsymbol{C}$ 输入到分类器中实现语序先后的判断。NSP 任务侧重于句子级别的信息解析，使得 BERT 能够理解输入文本句子的整体含义。

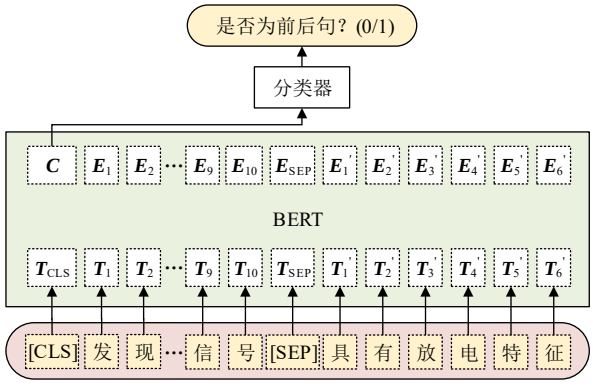


图 5　NSP 任务训练过程

Fig. 5　Training process of NSP task

值得注意的是，BERT 内部权重参数通过 MLM 与 NSP 任务能够初步学习到故障文本的特征，但针对实体识别任务还需进行微调。

2.4.2　单一类型实体识别器及其预训练机制

本文采用多个单一类型实体识别器对不同类型的实体信息进行解析以解决设备故障文本实体边界模糊的问题，并采用并行训练的方法提升算法训练效率，节省计算资源。在预训练阶段，单一实体识别器需要对样本标签中非目标实体类型的实体用符号‘O’进行遮盖。为进一步节省计算资源，采用线性嵌入层来获得单一识别器在预训练阶段中的嵌入向量序列 $\boldsymbol{E}$，如式(9)所示：

$$\boldsymbol{E}=\sigma(\boldsymbol{X}\boldsymbol{W}_{\mathrm{e}}) \tag{9}$$

式中：$\boldsymbol{X}\in\mathbb{R}^{n\times N}$ 为输入的独热向量序列；N 为故障文本字典字符数；$\boldsymbol{W}_{\mathrm{e}}\in\mathbb{R}^{N\times d}$ 为嵌入层权重。

随后，根据实际效果选定合适的 BiRNN 对向量序列信息解析，并利用全连接层对序列中的特征向量进行降维，以实现在 CRF 约束下的序列标注任务，如式(10)(11)所示：

$$\boldsymbol{F}_{\mathrm{r}}=f_{\mathrm{rnn}}(\boldsymbol{E}) \tag{10}$$

$$\boldsymbol{F}_{\mathrm{s}}=\sigma(\boldsymbol{F}_{\mathrm{r}}\boldsymbol{W}_{\mathrm{d}}+\boldsymbol{o}_{\mathrm{n}}\boldsymbol{b}_{\mathrm{d}}^{\mathrm{T}}) \tag{11}$$

式中：$\boldsymbol{F}_{\mathrm{r}}\in\mathbb{R}^{n\times d_{\mathrm{r}}}$ 为解析后的向量序列；$\boldsymbol{W}_{\mathrm{d}}\in\mathbb{R}^{d_{\mathrm{r}}\times d_{\mathrm{s}}}$ 为全连接层权重；$\boldsymbol{b}_{\mathrm{d}}\in\mathbb{R}^{d_{\mathrm{s}}}$ 为全连接层偏置；$\boldsymbol{o}_{n}\in\mathbb{R}^{n}$ 为全 1 向量，将 $\boldsymbol{b}_{\mathrm{d}}$ 扩充为矩阵方便进行矩阵加减运算；$\boldsymbol{F}_{\mathrm{s}}\in\mathbb{R}^{n\times d_{\mathrm{s}}}$ 为识别器的分类特征序列，并在式(8)的约束下可获得特定类型实体的识别结果。

2.4.3　多类型实体识别器及其训练机制

多类型实体识别器综合解析单一识别器的识别结果，梳理各识别器之间的逻辑关系与各类实体之间的边界，实现对样本各类实体信息的抽取。

令单一类型识别器个数为 m，则将单一类型识别器输出特征向量的维度从 d_{s} 扩充到 $d_{\mathrm{m}}=3m+1$ 以此来为特征向量注入实体类型信息，多出来的维度用 0 填充，并将其拼接作为多类型识别器的输入特征序列 $\tilde{\boldsymbol{F}}\in\mathbb{R}^{n\times md_{\mathrm{m}}}$。随后，利用含有 K 个卷积核的一维 CNN 对各单一识别器之间关联信息进行抽取，并对 $\tilde{\boldsymbol{F}}$ 进行卷积操作，如式(12)所示：

$$\boldsymbol{F}_{\mathrm{c}}^{i,d_{\mathrm{c}}(a-1)+j}=\sigma\left(\sum_{k=1}^{d_{\mathrm{w}}}\tilde{\boldsymbol{F}}_{i}(j+k-1)\boldsymbol{W}_{\mathrm{c}}^{a,k}\right) \tag{12}$$

式中：$\boldsymbol{W}_{\mathrm{c}}^{a,k}$ 为 CNN 卷积核 $\boldsymbol{W}_{\mathrm{c}}\in\mathbb{R}^{K\times d_{\mathrm{w}}}$ 中 a 行 k 列的权重；$\boldsymbol{F}_{\mathrm{c}}^{i,j}$ 为卷积特征向量序列 $\boldsymbol{F}_{\mathrm{c}}\in\mathbb{R}^{n\times Kd_{\mathrm{c}}}$ 中第 i 个向量的第 j 个特征值；$d_{\mathrm{c}}=d_{\mathrm{m}}-d_{\mathrm{w}}+1$ 为单个卷积核卷积后特征向量的维度；d_{w} 为单个卷积核大小。利用 BiLSTM 对 $\boldsymbol{F}_{\mathrm{c}}$ 的序列特征进行抽取如

式(3)—(7)所示，得到特征向量序列 $\boldsymbol{F}_{\mathrm{L}} \in \mathbb{R}^{n \times d_{\mathrm{L}}}$。最后，与单一识别器类似，在全连接层与 CRF 的约束下获得标签的概率分布向量序列 $\boldsymbol{L} \in \mathbb{R}^{n \times d_{\mathrm{m}}}$。

值得注意的是，多类别实体识别器的预训练过程需要在前 2 个训练步骤结束后进行。在实际训练过程中，可将这一步与 I-BRC 微调过程进行合并(特征抽取方法与单独训练相同)。经测试，这 2 种训练方式从性能与效率上均无明显差别。

2.4.4 I-BRC 微调

如图 3 所示，将预训练好的 BERT 嵌入层、单一类型实体识别器与多类型实体识别器的权重导入到 I-BRC，在序列标签的监督下对自身权重进行微调。经过极少的循环迭代后，I-BRC 模型的各参数达到收敛，最终在实际的电力设备故障文本实体识别任务中能够取得较好的效果。

3 实验验证

为验证本文所提实体识别算法，我们在某电网公司记录的巡检、检修案例报告上对 I-BRC 进行实验分析。先后进行各单一类别类实体识别器效果分析、算法对比分析、算法效率分析，并介绍 I-BRC 在应对不同设备故障文本的灵活模块化调整方法，最后展示了算法识别效果与设备案例知识图谱。

3.1 模型参数设定

本文采用 K 折交叉验证的方式对 I-BRC 进行验证。并通过参数扫描的方法设定 d_{r} 为 128，d_{L} 为 256，d_{w} 为 4。设定 d_{s} 为 4，d 为 768。此外，由于本文所用样本在相关运维标准下含有 7 种类型的实体信息，因此设定 $m=7$，$d_{\mathrm{m}}=22$。采用的 CNN 层卷积核个数 $K=16$，激活函数 $\sigma(\bullet)$ 为衰减率为–0.2 的衰减修整线性单元。本文算法在 tensorflow2.5 框架上实现，并在 2 块 RTX 3090 GPU 上进行训练。

3.2 数据集

本文采用的故障文本是某电力公司历经 10 年左右所记录真实设备巡检、检修记录。原始样本须经以下 3 个预处理过程：

1）原始的案例文本中包含图片、表格等结构化数据，可通过其他方式转化为知识图谱，无需进行实体识别，因此将其滤除。

2）分析样本内容，定义不同类型的实体信息。通过对该样本的分析以及相关的技术规范要求，定义 7 种实体类型：案例编号、时间、站点、设备、现象、缺陷类型以及缺陷位置。

3）按照表 1 的形式，对所有设备故障文本样本进行序列标注。

经处理后，共得 1200 个样本，每个样本以句子为单位，包含 60~200 个字符不等。

3.3 算例分析

本文采用精确度 P、召回率 R 与平衡分数 F_1 作为算法的评估指标。精确度与召回率分别表征算法正确识别各类实体与找全各类实体的能力，而 F_1 是精确度与召回率的综合表现。算法整体的精确度和召回率是不同类型实体各自的精确率与召回率的加权平均，如式(13)—(15)所示：

$$P=\frac{1}{m}\sum_{i}N_{\mathrm{TP}}^{i}/(N_{\mathrm{TP}}^{i}+N_{\mathrm{FP}}^{i}) \tag{13}$$

$$R=\frac{1}{m}\sum_{i}N_{\mathrm{TP}}^{i}/(N_{\mathrm{TP}}^{i}+N_{\mathrm{FN}}^{i}) \tag{14}$$

$$F_1=2PR/(P+R) \tag{15}$$

式中：N_{TP}^{i} 为算法识别正确的第 i 类实体的个数；N_{FP}^{i} 为其他类型实体被错误识别为第 i 类实体的个数；N_{FN}^{i} 为算法错把第 i 类实体识别为其他类别的个数；m 为实体类别个数，对单一实体识别器进行评估时 $m=1$。

3.3.1 单一类别实体识别器效果分析

单一实体识别器作为 I-BRC 最为重要的模块，直接决定了算法实体识别效果以及在实际使用中的灵活性。因此，针对本文的故障文本样本，对 7 种典型的单一实体识别器的识别效果进行分析，实验结果如表 2 所示。

表 2 各单一类别实体识别器识别结果

Table 2 Results of each single-type entity recognizer %

实体类别	P	R	F_1
案例名	99.6	99.8	99.7
时间	99.4	99.7	99.5
站点	95.0	95.0	95.0
设备	79.5	90.4	84.6
现象	68.2	71.4	69.8
缺陷类型	99.8	99.7	99.7
缺陷位置	64.3	73.0	68.4

在中文电力设备故障文本中，各类实体的复杂度、词频、字符串长度有明显区别，因此造成了不同单一识别器识别效果之间的巨大差异。在样本中，案例名、时间、站点名称与故障类型 4 类实体信息具有明显的语法特征，以及十分清晰的实体边界，识别器中的 BiRNN 能够有效地根据语境对这几类实体地信息特征进行解析，因此这 4 类识别器效果极佳。另外，对于样本中的故障现象与缺陷位置这 2 类实体信息而言，由于不同巡检、检修人员的记录习惯的差异，导致实体边界特征异常模糊。模糊的实体边界特征又使得识别器的 BiRNN 层难

以准确地解析这 2 类实体信息的语义特征，导致其抽取的特征向量序列中含有大量模糊不确定的干扰，最终导致识别器的效果较差。此时若使用一个识别器直接对所有类型实体进行抽取，则个别类型实体模糊的边界特征会通过循环网络的信息传递特性对其他实体信息进行干扰，导致算法整体的识别效果进一步劣化。然而，利用多个单一识别器对不同实体信息进行独立的解析，避免了干扰特征在序列中的传播，同时利用多类型识别器中的卷积操作还可以挖掘出不同识别器之间的内在联系进行补偿，为个别类型的实体勾勒出更加清晰的实体边界，从而提升容错率以及模型整体的识别性能。

3.3.2　性能对比分析

为进一步验证 I-BRC 的性能，将其与其他常见的实体识别算法进行对比实验分析。选取 BERT、连续词袋模型(continues bag of words，CBOW)与 skip-gram 作为嵌入模型；采用 BiLSTM、BiGRU、CNN 神经网络作为文本特征解析模型；采用 CRF 与隐马尔可夫模(hidden markov model，HMM)概率图模型约束算法的序列标注。对比实验结果如表 3 所示，从表中可以看出相比其他传统实体识别模型，I-BRC 在中文电力设备故障文本上进行实体信息抽取时效果要高出 7.5%~29.3%。

表 3　不同算法的性能比较

Table 3　Performance comparison between different algorithms　%

算法	P	R	F_1
BiLSTM+CRF	61.5	56.2	58.7
CBOW+BiLSTM+HMM(分词)	62.7	63.2	62.9
CBOW+BiLSTM+HMM	63.3	65.7	64.5
CBOW+CNN+CRF	63.7	66.4	65.0
CBOW+BiGRU+CRF	69.7	62.4	65.8
CBOW+BiLSTM+CRF (分词)	69.2	60.1	64.3
CBOW+BiLSTM+CRF	70.7	62.8	66.5
skip-gram+BiLSTM+HMM	65.7	67.2	66.4
skip-gram+CNN+HMM	63.4	60.5	61.9
skip-gram+CNN+CRF	65.6	63.7	64.6
skip-gram+BiGRU+CRF (分词)	72.3	69.6	70.9
skip-gram+BiGRU+CRF	74.6	71.6	73.1
skip-gram+BiLSTM+CRF(分词)	73.1	70.8	71.9
skip-gram+BiLSTM+CRF	76.6	74.4	75.5
BERT+BiLSTM+CRF(分词)	78.6	76.5	77.5
BERT+BiLSTM+CRF	81.2	79.8	80.5
I-BRC(本文)	**86.8**	**89.2**	**88.0**

嵌入模型对文本进行嵌入式表示，使算法能够正确理解文本内容，对实体识别算法至关重要。如表 3 所示，采用 BERT 的识别器效果要优于 CBOW 与 skip-gram。BERT 在句子中任意 2 个字符之间建立注意力联系，利用周围语境对某一字符的注意力程度来对该字符进行嵌入式表示，能够更好地理解相同字符在不同语境中的语义信息，并有效解决 CBOW、skip-gram 等嵌入模型在处理长句时出现的信息遗失的问题。

从特征序列解析模型来看，由于 CNN 中卷积核限制了模型的感知野，使得其对较长语句的解析能力难以提升。另一方面，BiLSTM 与 BiGRU 利用门控单元来保存序列信息并实现信息跨时间传播，使得算法对序列中某一特征向量的解析不再局限于有限的扫描窗格内，同时序列信息的遗忘机制可以在特征抽取中更加侧重距离较近的语义信息，更适合解析故障文本中的实体信息。

在概率图模型的对比分析中，判别式模型 CRF 相比生成式模型 HMM 在模型求解的过程中可以避免陷入局部最优解。同时，CRF 具有更加合理的归一化因子求取机制，能够有效避免出现偏置的问题。因此，CRF 能够更好地对序列标注进行约束和最优化，并在实际算法中具有更好的识别效果。最后，表 3 的实验结果也证实了采用字标注的方法相比传统分词处理具有更好实体识别效果，避免了误差累积，验证了本文标注方法的可行性。

3.3.3　算法效率分析

本节通过对 I-BRC 算法训练的收敛性能与时耗对其效率进行分析，如图 6、表 4 所示。从算法训练时的收敛迭代次数来看，采用线性嵌入模型的 BiLSTM-CRF 在对故障文本的实体信息进行抽取时效果较差，60 次迭代后 F_1 仍不足 60%，且由于缺少预训练机制导致算法收敛速度缓慢。采用 BERT 后，算法的 F_1 能够接近 80%，但仍需消耗大量计算资源来进行训练。当集成了多个单一类型识别器并进行预训练后，即使采用最简单的线性嵌入

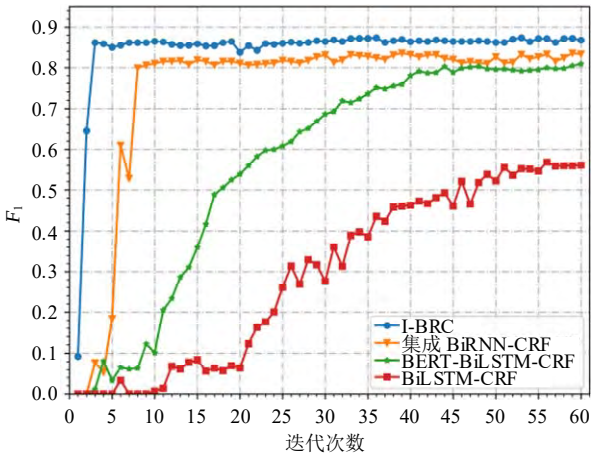


图 6　收敛性能对比分析

Fig. 6　Comparison analysis of convergence performance

表 4 各实体识别模型的训练与识别时耗

Table 4 Training and recognition timing consumption of each entity recognition model

算法	训练时耗/min	识别时耗/s
BiLSTM-CRF	43.29	0.158
BERT-BiLSTM-CRF	39.31	0.173
集成 BiRNN-CRF	14.73	0.236
I-BRC	8.54	0.249

模型，算法也能在极少的迭代次数后达到更好的识别效果，验证了单一实体识别的优越性。最终，在 BERT 与单一实体识别器的作用下，I-BRC 仅需 3 次迭代便可达到收敛，并将 F_1 突破性地提升至 88.0%。

从算法的训练时耗分析，由于 I-BRC 与集成 BiRNN-CRF 在模型结构上具有更高的复杂度，因此训练时这 2 种模型单次迭代所花费的时间要多于 BiLSTM-CRF 与 BERT-BiLSTM-CRF。但是，由于充分采用了并行预训练机制，I-BRC 迭代次数大幅减少，训练仅需 8.54min，不足 BiLSTM-CRF 的 1/5。另一方面，由于算法的识别时耗与模型结构的复杂度正相关，因此 I-BRC 在识别样本实体信息时需要耗时更多。但是，电力设备文本对实体信息抽取的时耗需求不高，因此 0.249s 的识别时间足以使 I-BRC 算法在实际工程中得到应用。

3.3.4 算法的模块化调节展示

传统的实体识别算法为适用于多种设备的运维、检修报告，会让实体识别算法学习尽可能多的实体类型。但是，针对某一种设备文本，并非所有类型的实体均可能出现，另外加上故障文本实体边界模糊的特点，采用传统方法会导致算法效果变差，甚至会出现与常识相违背的结果(例如在传输线路文本中标注出油色谱特征实体)。

为此，可采用对 I-BRC 算法进行模块化调整调整的方式来应对不同电力设备的故障文本，如图 7 所示。在使用 I-BRC 前先建立一个候选识别器库，并根据样本对不同类型的单一类别实体识别器并行地进行训练并存放该库中。候选库中的识别器通常情况下一经训练无需再次进行调参，仅在实际应用时调用即可。在对变压器(或输电线路)故障文本中的实体信息抽取时，从候选库中调用温度与油色谱识别器(或应力与天气状态识别器)，最后对 I-BRC 整体进行微调，经上节的实验验证，I-BRC 仅需 3 次迭代即可收敛，有效缩短训练时间。另一方面，该方法针对各类实体信息进行单独分析，减少了对算法参数的重复训练，从而节省了大量计算资源。

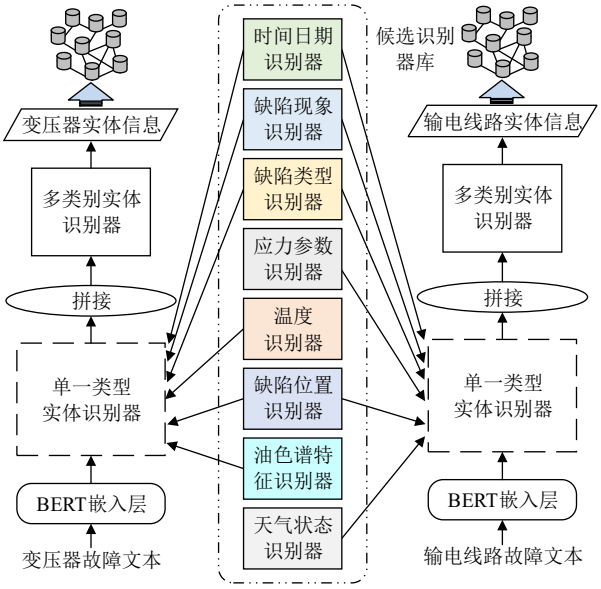


图 7 I-BRC 多种设备灵活应用示例

Fig. 7 Flexible application illustration of I-BRC on multiple equipment

3.3.5 实体识别效果展示及案例知识图谱

本文所提出的 I-BRC 能够有效识别各类设备故障文本中的实体信息并构建案例知识图谱。图 8 展示了 I-BRC 对某电缆故障报告中的实体(用粗体字标出)进行抽取的效果。I-BRC 中 6 个单一实体识别器对文本中不同类型的实体信息进行抽取，由于缺陷位置与现象实体的边界比较模糊，单一实体识别器无法准确的识别出实体信息，其中识别有误的

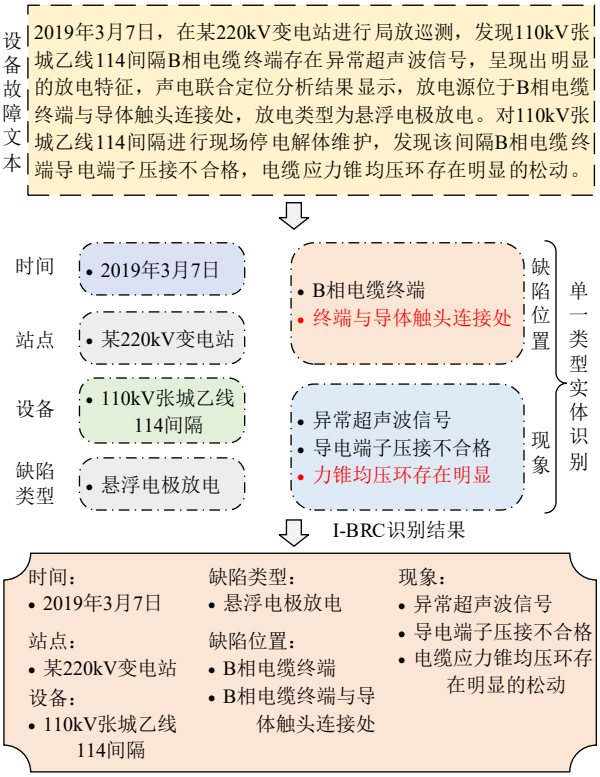


图 8 I-BRC 实体抽取效果展示

Fig. 8 Results demonstration of I-BRC entity recognition

实体用红色字体标出。利用多类型识别器的整合作用，I-BRC 综合多个单一识别器的识别结果，通过横向比较实体特征勾勒出更加清晰的实体边界信息，高效、准确地识别出文本中各类实体信息。

图 9 为 I-BRC 从该文本抽取实体后构建的案例图谱，其中黄色节点为该案例的各类实体信息。大量的案例图谱与相关维修预案、规范图谱汇集为规模庞大的设备运维知识图谱。当设备发生故障时，工作人员利用检索技术迅速找到与当前事件相似的案例进行初步的原因分析。随后，算法结合相关规范知识与知识推理技术智能地给出维修决策，辅助运维人员确保设备的安全稳定地运行。最后，I-BRC 会抽取出当前案例的实体信息并将其加入到运维图谱中实现图数据库的更新与完善。

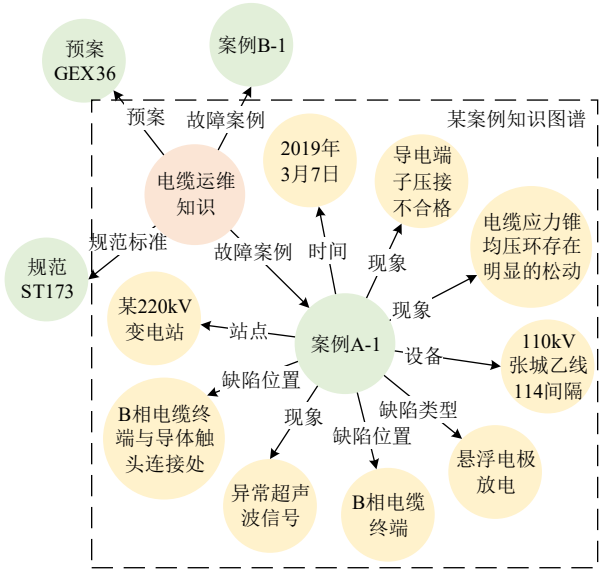


图 9　I-BRC 抽取并构建的某案例知识图谱

Fig. 9　A case knowledge graph extracted and constructed by I-BRC

4　结论与展望

本文提出了一种全新的 I-BRC 实体识别算法，能够有效地对各类电力设备故障案例文本的实体信息进行抽取，将非结构化数据转化为结构化数据。通过实验验证得到以下结论：

1）论文所提出的 I-BRC 算法在实验样本上 F_1 达到了 88.0%，能有效识别出样本中各类实体信息，与传统算法性能提升了 7.5%~29.3%，证实了该算法的可行性。

2）论文所提出的算法通过字嵌入方式避免了误差累积，利用集成的方式有效解决了各类实体边界模糊的问题，提升了算法性能；本文采用的并行训练方式，使得算法仅需 3 次迭代就可收敛，有效提升了算法训练效率。

3）论文所提出算法在应对多种设备文本时可利用模块化调节的方式进行调节，并可在极短的时间内微调参数，在确保准确率的同时又能广泛灵活地应用于各类设备故障文本。

此外，本文所提算法能够有效抽取故障文本中的实体信息并构成案例图谱，但是对于文本中包含多个案例事件的情况效果较差。因此，在后续工作中将进一步研究能够有效解决事件边界模糊问题的实体事件联合抽取方法，从而完善设备故障文本信息的提取。

参考文献

[1] 胡美玉，胡志坚，汪祥，等．基于双向层级结构的计及开关故障的配电系统可靠性评估[J]．电网技术，2016，40(5)：1476-1481．
HU Meiyu，HU Zhijian，WANG Xiang，et al．Reliability evaluation of distribution network taking switch faults into account based on bidirectional hierarchical structure[J]．Power System Technology，2016，40(5)：1476-1481(in Chinese)．

[2] 廖瑞金，王有元，刘航，等．输变电设备状态评估方法的研究现状[J]．高电压技术，2018，44(11)：3454-3464．
LIAO Ruijin，WANG Youyuan，LIU Hang，et al．Research status of condition assessment method for power equipment[J]．High Voltage Engineering，2018，44(11)：3454-3464(in Chinese)．

[3] 田嘉鹏，宋辉，罗林根，等．基于均值漂移的局部放电边缘计算方法研究[J]．电网技术，2021，45(6)：2449-2456．
TIAN Jiapeng，SONG Hui，LUO Lin'gen，et al．Edge computing method of partial discharge based on mean-shift[J]．Power System Technology，2021，45(6)：2449-2456(in Chinese)．

[4] 郭榕，杨群，刘绍翰，等．电网故障处置知识图谱构建研究与应用[J]．电网技术，2021，45(6)：2092-2100．
GUO Rong，YANG Qun，LIU Shaohan，et al．Construction and application of power grid fault handing knowledge graph[J]．Power System Technology，2021，45(6)：2092-2100(in Chinese)．

[5] 王骏东，杨军，裴洋舟，等．基于知识图谱的配电网故障辅助决策研究[J]．电网技术，2021，45(6)：2101-2112．
WANG Jundong，YANG Jun，PEI Yangzhou，et al．Distribution network fault assistant decision-making based on knowledge graph[J]．Power System Technology，2021，45(6)：2101-2112(in Chinese)．

[6] SONG Qi，WU Yinghui，LIN Peng，et al．Mining summaries for knowledge graph search[J]．IEEE Transactions on Knowledge and Data Engineering，2018，30(10)：1887-1900．

[7] 阮聪，齐林海，王红．融合知识图谱与神经张量网络的需求响应智能推荐[J]．电网技术，2021，45(6)：2131-2140．
RUAN Cong，QI Linhai，WANG Hong．Intelligent recommendation of demand response combined with knowledge graph and neural tensor network[J]．Power System Technology，2021，45(6)：2131-2140(in Chinese)．

[8] LV Ke，GAO Caixia，SI Jikai，et al．Fault coil location of inter-turn short-circuit for direct-drive permanent magnet synchronous motor using knowledge graph[J]．IET Electric Power Applications，2020，14(9)：1712-1721．

[9] 杜修明，秦佳峰，郭诗瑶，等．电力设备典型故障案例的文本挖掘[J]．高电压技术，2018，44(4)：1078-1084．
DU Xiuming，QIN Jiafeng，GUO Shiyao，et al．Text mining of typical defects in power equipment[J]．High Voltage Engineering，2018，

44(4)：1078-1084(in Chinese).

[10] 佟佳弘，武志刚，管霖，等．电力调度文本的自然语言理解与解析技术及应用[J]．电网技术，2020，44(11)：4148-4156.
TONG Jiahong，WU Zhigang，GUAN Lin，et al．Power dispatching text analysis and application based on natural language understanding[J]．Power System Technology，2020，44(11)：4148-4156(in Chinese).

[11] 陶洪铸，翟明玉，许洪强，等．适应调控领域应用场景的人工智能平台体系架构及关键技术[J]．电网技术，2020，44(2)：412-419.
TAO Hongzhu，ZHAI Mingyu，XU Hongqiang，et al．Architecture and key technologies of artificial intelligence platform oriented for power grid dispatching and control application scenarios[J]．Power System Technology，2020，44(2)：412-419(in Chinese).

[12] 乔骥，王新迎，闵睿，等．面向电网调度故障处理的知识图谱框架与关键技术初探[J]．中国电机工程学报，2020，40(18)：5837-5848.
QIAO Ji，WANG Xinying，MIN Rui，et al．Framework and key technologies of knowledge-graph-based fault handling system in power grid[J]．Proceedings of the CSEE，2020，40(18)：5837-5848(in Chinese).

[13] LI Jiabin，FANG Suwan，REN Yuqi，et al．SWVBiL-CRF：selectable word vectors-based BiLSTM-CRF power defect text named entity recognition[C]//Proceedings of 2020 IEEE International Conference on Big Data．Atlanta，USA：IEEE，2020：2502-2507.

[14] 戴宇欣，张俊，季知祥，等．基于功能缺陷文本的电力系统二次设备智能诊断与辅助决策[J]．电力自动化设备，2021，41(6)：184-191.
DAI Yuxin，ZHANG Jun，JI Zhixiang，et al．Intelligent diagnosis and auxiliary decision of power system secondary equipment based on functional defect text[J]．Electric Power Automation Equipment，2021，41(6)：184-191(in Chinese).

[15] YANG Q Y，JIANG J，FENG X Y，et al．Named entity recognition of power substation knowledge based on transformer-BiLSTM-CRF network[C]//Proceedings of 2020 International Conference on Smart Grids and Energy Systems (SGES)．Perth，Australia：IEEE，2020：952-956.

[16] 蒋晨，王渊，胡俊华，等．基于深度学习的电力实体信息识别方法[J]．电网技术，2021，45(6)：2141-2149.
JIANG Chen，WANG Yuan，HU Junhua，et al．Power entity information recognition based on deep learning[J]．Power System Technology，2021，45(6)：2141-2149(in Chinese).

[17] CAI Qing．Research on Chinese naming recognition model based on BERT embedding[C]//Proceedings of the IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)．Beijing，China：IEEE，2019：1-4.

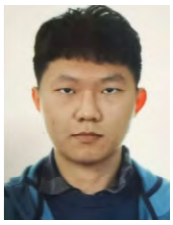
[18] SAHA T，JAYASHREE S R，SAHA S，et al．BERT-Caps：a transformer-based capsule network for tweet act classification[J]．IEEE Transactions on Computational Social Systems，2020，7(5)：1168-1179.

[19] VASWANI A，SHAZEER N，PARMAR N，et al．Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems．Long Beach，CA，USA：ACM，2017：6000-6010.

[20] DAI Jiejie，SONG Hui，SHENG Gehao，et al．LSTM networks for the trend prediction of gases dissolved in power transformer insulation oil[C]//Proceedings of the 12th International Conference on the Properties and Applications of Dielectric Materials (ICPADM)．Xi'an，China：IEEE，2018：666-669.

[21] DENG Yaping，WANG Lu，JIA Hao，et al．A sequence-to-sequence deep learning architecture based on bidirectional GRU for type recognition and time location of combined power quality disturbance[J]．IEEE Transactions on Industrial Informatics，2019，15(8)：4481-4493.

[22] LAFFERTY J D，MCCALLUM A，Pereira F C N．Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning (ICML)．Williamstown，MA，USA：ACM，2001：282-289.



田嘉鹏

在线出版日期：2021-12-08。
收稿日期：2021-09-15。

作者简介：
田嘉鹏(1997)，男，博士研究生，研究方向为输变电设备状态监测与智能化研究，E-mail：jiapengtian@sjtu.edu.cn；
宋辉(1987)，男，通信作者，博士，助理研究员，主要从事输变电设备状态监测与智能化研究，E-mail：songeos@163.com。

（责任编辑　宋钰龙）