DOI：10.13733/j.jcam.issn.2095-5553.2024.08.049
杨宁，钱晔，陈健. 面向联合收割机故障领域的命名实体识别研究[J]. 中国农机化学报，2024，45(8)：338－343
Yang Ning，Qian Ye，Chen Jian. Research on named entity recognition for combine harvester fault domain [J]. Journal of Chinese Agricultural Mechanization，2024，45(8)：338－343

面向联合收割机故障领域的命名实体识别研究 *

杨宁[1]，钱晔[1,2]，陈健[1]

（1. 云南农业大学大数据学院，昆明市，650201；
2. 云南省农业大数据工程技术研究中心，昆明市，650201）

摘要：联合收割机作为一种机械化设备不可避免地会出现机械故障，为快速地找出并解决机械故障，提出一种面向联合收割机故障领域的命名实体识别模型RP－TEBC（RoBERTa－wwm－ext＋PGD＋Transformer－Encoder＋BiGRU＋CRF）。RP－TEBC使用动态编码的RoBERTa－wwm－ext预训练模型作为词嵌入层，利用自适应Transformer编码器层融合双向门控单元（BiGRU）作为上下文编码器，利用条件随机场（CRF）作为解码层，使用维特比算法找出最优的路径输出。同时，RP－TEBC模型在词嵌入层中通过添加一些扰动，生成对抗样本，经过对模型不断的训练优化，可以提高模型整体的鲁棒性和泛化性能。结果表明，在构建的联合收割机故障领域命名实体识别数据集上，相比于基线模型，该模型的准确率、召回率、$F1$值分别提高1.79%、1.01%、1.46%。

关键字：联合收割机；故障领域；命名实体识别；知识图谱；预训练模型；对抗样本
中图分类号：S225；TP391.1　　**文献标识码**：A　　**文章编号**：2095-5553（2024）08-0338-06

Research on named entity recognition for combine harvester fault domain

Yang Ning[1]，Qian Ye[1,2]，Chen Jian[1]

（1. *College of Big Data*，*Yunnan Agricultural University*，*Kunming*，650201，*China*；
2. *Agricultural Big Data Engineering Research Center of Yunnan Province*，*Kunming*，650201，*China*）

Abstract：Combine harvesters as a kind of mechanized equipment will inevitably have mechanical failure，in order to quickly find out the relevant fault entity and solve the mechanical failure，a named entity recognition model RP－TEBC（RoBERTa－wwm－ext＋PGD＋Transformer－Encoder＋BiGRU＋CRF）for combine harvester fault field is proposed. RP－TEBC uses the dynamically encoded RoBERTa－wwm－ext pre－trained model as the word embedding layer，uses the adaptive Transformer encoder layer to fuse the Bidirectional Gating Unit（BiGRU）as the context encoder，and finally uses the conditional random field（CRF）as the decoder layer，using the Viterbi algorithm to find the optimal path output. At the same time，the RP－TEBC model generates adversarial samples by adding some perturbations in the word embedding layer. Through continuous training and optimization of the model，the overall robustness and generalization performance of the model can be improved. On the constructed named entity recognition data set in the field of combine harvester faults，experiments have shown that compared with the baseline model，the accuracy，recall rate，and $F1$ value of this model have increased by 1.79%，1.01%，and 1.46% respectively.

Keywords：combine harvester；fault domain；named entity recognition；knowledge graph；pre-trained model；adversarial sample

0 引言

近年来，随着我国农业机械化的不断发展，联合收割机作为一种有效的农作物收割工具得到大量生产和广泛应用。与联合收割机维修相关的非结构化文本数据也在持续增长，如何在这些纷繁复杂的数据中高

收稿日期：2022年11月27日　　修回日期：2023年3月21日
＊基金项目：云南省科技厅科技计划项目（202002AE090010，202302AE090020）；教育部产学合作协同育人项目（202102356024）
第一作者：杨宁，男，1997年生，山东滨州人，硕士研究生；研究方向为自然语言处理。E-mail：804141529@qq.com
通讯作者：钱晔，女，1984年生，安徽巢湖人，博士，副教授；研究方向为人工智能。E-mail：33754199@qq.com

效精准地检索出所需信息，成为当前要解决的问题。

命名实体识别（NER）作为自然语言处理（NLP）信息抽取领域一项基本任务，近年来得到了广泛研究，其本质上是在句子中查找实体的开始和结束并为此实体分配类别的任务。随着深度学习技术的不断创新和发展，命名实体识别任务也从早期的基于规则和机器学习的方法朝着向深度学习的方法发展，并取得了不错的成效。Hammerton[1]首次将长短期记忆网络（LSTM）用于命名实体任务，这也是首次将神经网络模型用于命名实体识别任务。Collobert等[2]使用CNN－CRF的模型结构达到了和基于统计机器学习方法相媲美的结果。Huang等[3]首次将双向长短期记忆网络（BiLSTM）和条件随机场（CRF）相结合用于命名实体识别任务。在中文命名实体识别领域中，由于中文句子中词与词之间连接在一起，不像英语有着天然的空格分割符，所以中文的命名实体识别相较于英文更加困难，因此中文的命名识别任务的首要任务是先分词，将词级的命名实体识别模型用于分词后的句子。但是再好的分词工具也会出现一些分词错误的现象，这也会导致在进行命名实体识别模型训练的时候出现实体边界的检测和识别类别的预测错误的情况，从而影响模型的整体效果。为了解决这个问题，一些直接在字符级别执行中文命名识别的方法开始得到研究，经过相关试验证明在字符级别执行命名实体识别不仅避免了分词错误对模型的影响，而且模型的效果也得到了提升。Dong等[4]是第一个将基于字符的BiLSTM－CRF神经架构用于中文命名实体识别任务。Ma等[5]将单词词典合并到字符表示中，巧妙的结合了词典的信息，使得模型的稳健性得到进一步提升。受到在计算机视觉中对抗训练的启发，相关研究人员也开始在自然语言处理任务加入对抗训练，来提升模型的泛化能力，Zhou等[6]提出双对抗转移网络（DATNet），通过在词嵌入层增加噪声解决了在低资源下的命名实体任务。

目前大多数的命名实体模型都是针对通用领域而设计的，针对特定领域的命名实体识别模型研究较少，有关联合收割机故障领域的命名实体识别研究还尚未见报道。由于在联合收割故障诊断方面缺乏相关的数据集，因此本文需自行标注。针对现有的命名实体识别模型，实体识别准确率不高、一词多义等问题。本文在联合收割机故障领域方面提出RP－TEBC命名实体识别模型。RP－TEBC模型采用动态编码的RoBERTa－wwm－ext预训练模型作为词嵌入层，能够很好地解决一词多义的问题。通过使用自适应Transformer编码器层融合双向门控循环单元（BiGRU）可以更好地学习文本中的语义信息。同时，添加对抗训练有帮助模型提升鲁棒性和泛化能力。

1　构建数据集

通过爬虫、PDF转换文字等技术手段共收集到联合收割机故障领域领域非结构化文本文字约13万字。利用YEDDA标注工具对原始语料库进行标注。将标记的实体分为四个类别：故障名称（Fault）、故障原因（Cause）、故障部位（Position）以及故障维修（Repaies），对这四类实体均采用BIO的标注形式进行标注，“B”代表实体的开始部位，“I”代表实体的中间及结束部位，“O”代表不是实体。数据标注示例如表1所示。将构建好的数据集随机打乱后，按照8∶2的比例来划分训练集和测试集，详细数据类别分布如图1所示。

表 1　实体标注样例

Tab. 1 Entity annotation example

字	标签	字	标签	字	标签
滑	B－Position	首	O	研	I－Repaies
阀	I－Position	先	O	修	B－Repaies
芯	I－Position	应	O	复	I－Repaies
与	O	对	O	但	O
阀	B－Position	其	O	不	O
孔	I－Position	清	B－Repaies	允	O
卡	B－Fault	洗	I－Repaies	许	O
死	I－Fault	如	O	配	O
别	I－Fault	无	O	合	O
住	I－Fault	效	O	间	O
或	O	则	O	隙	O
运	B－Fault	采	O	超	O
动	I－Fault	取	O	差	O
不	I－Fault	研	B－Repaies	。	O
灵	I－Fault	磨	I－Repaies		
活	I－Fault	或	O		
时	O	配	B－Repaies		

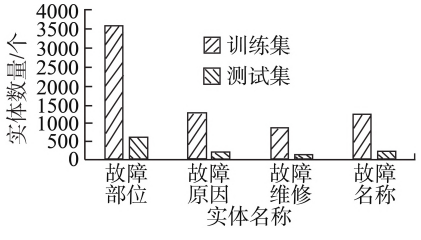

图 1　实体类别分类

Fig. 1 Entity class classification

2　命名实体识别模型

针对联合收割机故障领域命名实体识别，提出RP－TEBC模型。该模型结构主要分为四个部分，使用RoBERTa－wwm－ext预训练模型作为词嵌入层；通过在每个嵌入的词向量中加入对抗训练生成对抗样本来用

来提升模型的健壮性和泛化性;利用自适应Transformer编码器融合双向门控循环单元(BiGRU)作为上下文编码器;最后使用条件随机场(CRF)作为预测标签的输出层,得到输入文本中的实体。模型总体结构如图2所示。

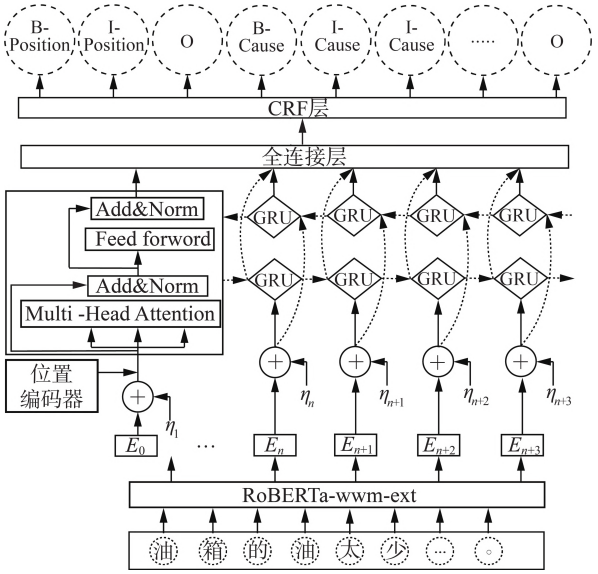

图 2 模型总体架构

Fig. 2 Model overall architecture

2.1 词嵌入层

在中文语句中相同的词语出现在句中不同的位置往往所表达语义会不相同,例如"他用小米手机扫码付钱在超市买了一袋小米。"中的"小米"在第一次出现的位置表示的意思就是手机品牌,在第二次出现的位置表示的就是粮食名称。由于Word2vec、GloVe等传统的词向量训练方法其向量表示都是恒定不变的,不能够跟随上下文变化而变化,因此很难表达一个词在不同上下文或不同语境中不同语义信息。针对这个问题Devlin等[7]基于深层的Transformer结构提出了BERT预训练模型。其在Transformer编码结构中引入多头注意力机制来获取输入文本的语义信息,运用遮蔽语言模型(MLM)和下一句预测(NSP)的子任务来训练语料的词向量表示。在这种表示方法中,词向量是由当前词所在的上下文计算获得,所以相同的词出现同一句话不同地方,词向量表示是不相同的,所表达语义也就不相同,BERT预训练模型整体结构如图3所示。

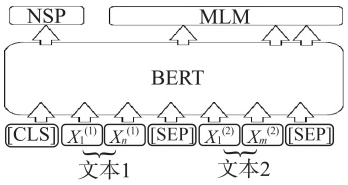

图 3 BERT模型整体结构

Fig. 3 Overall structure of the BERT model

从图3可以看出,模型的输入是由两段文本拼接而成,经过BERT的建模获取到输出文本的上下文语义表示,最终学习掩码语言模型和下一句预测。RoBERTa预训练模型是在BERT模型的基础改进而来,RoBERTa预训练模型使用动态掩码策略和更多样的训练数据,在训练过程中取消了NSP,通过改进使得RoBERTa预训练模型的性能要比BERT预训练模型的性能更加出色。RoBERTa-wwm-ext是根据Cui等[8]提出的全词遮蔽(WWM)策略使用中文数据集重新训练RoBERTa所得到的一种新的预训练模型。全词遮蔽策略是对词语作遮蔽语言训练,因为中文的词语由多个字符组成,直接遮蔽单个字符可能会导致语义信息的丢失。通过使用全词遮蔽策略,能够更好地捕捉中文词语的完整语义,从而提高模型在中文自然语言处理任务中的表现。

2.2 上下文编码器层

上下文编码器层的主要构成是自适应Transformer编码器和双向门控循环单元(BiGRU)。Transformer编码器不同于传统的循环神经网络(RNN)类型的编码器,利用多头注意力机制可以更好地学习到输入文本的上下文信息,解决了传统循环神经网络在学习较长输入文本时所存在的梯度消失和梯度爆炸的问题。但是传统的Transformer编码器在命名实体识别任务中性能表现远不如在其他自然语言处理任务中的表现,因此本文利用Yan等[9]所提出自适应Transformer编码器对输入文本的字符级特征和字级特征进行编码用于命名实体识别任务。原始的Transformer编码器中的正弦位置嵌入对距离感知较为敏感,但对方向感知却不怎么敏感。针对这个问题自适应Transformer编码器采用了相对位置编码,并对相对位置编码进行改进,改进后相对位置编码不仅使用参数更少而且性能也更好。

考虑到传统的Transformer编码器的注意力分布是缩放和平滑的,但是在命名实体识别的任务中并不是所有的词都需要值得去注意,因此对于命名实体识别任务来说一个稀疏的注意力机制显然是更加合适的。对此自适应Transformer编码器使用了一个非缩放和尖锐的注意力机制,计算如式(1)~式(4)所示。

$$Q,K,V=HW_{q},H_{d_k},HW_{v} \tag{1}$$

式中:Q,K,V——查询向量、键向量和数值向量;

W_{q}、W_{v}——权重参数;

H——参数矩阵;

H_{d_k}——矩阵的运算结果;

$$R_{t-j}=\left[\cdots \sin\left(\frac{t-j}{1\,000^{2i/d_k}}\right)\cos\left(\frac{t-j}{1\,000^{2i/d_k}}\right)\cdots\right]^{T} \tag{2}$$

式中:R_{t-j}——相对位置编码,$R_{t-j}\in R^{d_k}$;

t——目标词的索引；

j——上下文词的索引；

i——索引变量，范围为$[0,\frac{d_k}{2}]$。

$$A_{t,j}^{rel}=Q_tK_j^T+Q_tR_{t-j}^T+uK_j^T+vR_{t-j}^T \tag{3}$$

式中：$A_{t,j}^{rel}$——相对注意力；

$Q_tK_j^T$——Q_t和K_j标记之间的注意力得分；

$Q_tK_{t-j}^T$——对特定相对距离的偏差；

uK_j^T——对标记的偏差；

vR_{t-j}^T——特定距离和方向的偏差项。

$$Attn(Q,K,V)=\mathrm{soft\,max}(A_{t,j}^{rel})V \tag{4}$$

$$R_t,R_{-t}=\begin{bmatrix}\sin(c_0t)\\ \cos(c_0t)\\ \vdots\\ \sin(c_{\frac{d}{2}-1}t)\\ \cos(c_{\frac{d}{2}-1}t)\end{bmatrix},\begin{bmatrix}-\sin(c_0t)\\ \cos(c_0t)\\ \vdots\\ -\sin(c_{\frac{d}{2}-1}t)\\ \cos(c_{\frac{d}{2}-1}t)\end{bmatrix} \tag{5}$$

式中：c_0——位置标记。

d——维度。

由于传统的循环神经网络(RNN)在时间方向进行反向传播更新梯度参数时会流经tanh节点和矩阵乘积节点。$y=\tanh(x)$的导数为$\frac{dy}{dx}=1-y^2$，根据其导数可知，当导数的值小于1时，随着x的值在正数方向不断增加，导数的值是越来越接近于0的，这就意味着如果梯度经过tanh节点过多的话，导数的值就会慢慢趋近于0，从而出现梯度消失的现象。一旦出现梯度消失，权重参数将无法进行更新，这也是传统循环神经网络无法学习到长时序依赖的主要原因之一。当梯度经过矩阵乘机节点时梯度会随这时间步的增加呈现出指数级别的增长，当梯度过于庞大时就会出现非数值，导致神经网络无法进行学习，从而引发梯度爆炸。长短时记忆网络(LSTM)通过引进输入门、遗忘门和输出门在一定程度缓解了传统循环神经网络所带来的问题。LSTM在进行反向传播时是采用的是对应元素乘积的运算，对应的元素每次都会根据不同的门值进行相应的乘积运算，所以缓解了梯度消失和梯度爆炸的问题。门控循环单元(GRU)是对LSTM进行的一次升级改进，GRU由于只有重置门和更新门，所以计算成本和参数相比与LSTM更少，性能也能和LSTM相媲美。GRU计算图如图4所示，计算如式(6)～式(9)所示。

$$z=\sigma(x_{t_1}W_x^{(z)}+h_{t_1-1}W_h^{(z)}+b^{(z)}) \tag{6}$$

$$r=\sigma(x_{t_1}W_x^{(r)}+h_{t_1-1}W_h^{(r)}+b^{(r)}) \tag{7}$$

$$\tilde{h}=\tanh(x_{t_1}W_x+(r\odot h_{t-1})W_h+b) \tag{8}$$

$$h_t=(1-z)\odot h_{t-1}+z\odot\tilde{h} \tag{9}$$

式中：z——更新门；

r——重置门；

$\tilde{h}$——隐藏状态；

h_{t_1}——时间步t_1时刻的隐藏状态；

x_{t_1}——时间步t_1时刻的输入；

b——偏置项；

W——权重。

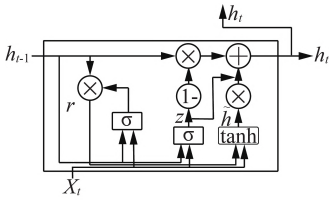

图 4　GRU计算图

Fig. 4 GRU calculation graph

对于命名实体识别任务来说目标词的词性不仅和前面的词相关，后面的词也会影响着目标词的词性。但是无论是传统的循环神经网络还是LSTM，GRU信息都是单向流动的，因此只能利用前面词的信息而利用不到后面词的信息。为解决这个问题，本文引入了双向门控循环单元(BiGRU)，一层GRU根据输入文本从头到尾对文本进行编码，另一层则从尾到头对输入文本进行编码，这样便可以同时利用到上下文的信息。BiGRU网络结构如图5所示。

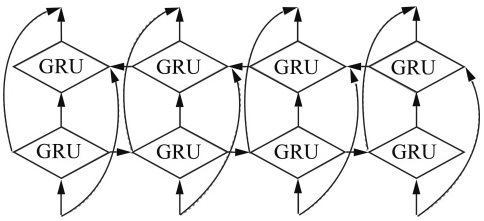

图 5　BiGRU结构图

Fig. 5 BiGRU structure diagram

自适应Transformer编码器能够更好地学习输入文本的语义信息，双向门控循环单元可以更好地区分目标词的上下文信息。通过融合两个模型的优点对输入文本进行编码使之更加适合于命名实体识别任务。

2.3　解码器层

为了利用不同标签之间的依赖性，本文使用条件随机场(CRF)作为RP－TEBC模型的解码层。条件随机场是在隐马尔可夫模型(HMM)和最大熵模型(EM)的基础提出的，打破了隐马尔可夫假设使得标签的预测更加合理，同时也修正了EM模型存在标签偏差的问题，使其可以做到全局归一化。给定一个序列$s=[s_1,s_2,\cdots,s_T]$其对应标签序列为$y=[y_1,y_2,\cdots,y_T]$，$Y_{(s)}$代表所有有效标签的序列，y的概率可由式(10)计算。在式(10)中$f(y_{t-1},y_t,s)$是计算y_{t-1}到y_t的转化分数，来最

大化$P(y|s)$,使用维特比算法找到最优的标签路径输出。

$$P(y|s)=\frac{\sum_{t=1}^{T}e^{f(y_{t-1},y_t,s)}}{\sum_{y'}^{Y_{(s)}}\sum_{t=1}^{T}e^{f(y'_{t-1},y'_t,s)}} \tag{10}$$

2.4 对抗训练

在训练深度神经网络模型的时候很容易受到对抗性示例的影响,这些对抗数据很难让模型和正常数据相区分,从而造成错误分类的结果。围绕对模型的对抗性和鲁棒性进行的优化的思路,Madry等[10]提出的PGD(Projected Gradient Descent)对抗训练算法为解决这个问题提供了很好地帮助。假设考虑一个标准的分类任务的数据分布为D,数据$x\in R^d$,标签$y\in[k]$,损失函数为$L(\theta,x,y)$,θ为固定的参数,利用经验风险最小化(ERM)找到模型最优的参数即:$\min_{\theta}E_{(x,y)\sim D}[L(x,y,\theta)]$。但是ERM不会产生对对抗数据具有鲁棒性的模型,因此需要来扩充ERM范式。通过对每个数据点x引入一组扰动,扰动的大小为S,利用原始样本生成对抗样本,再利用对抗样本求得期望,这就是著名的Min－Max公式,如式(11)所示。Min－Max主要有内部损失最大化和外部经验最小化组成,内部最大化问题的目的是找到给定数据点x的对抗样本,做到最高损失;外部经验风险最小化是找到模型的最优参数,使对抗性损失最小化。PGD通过“小步幅的走,一点一点靠近”的策略来保证扰动不要太大,如果走出扰动半径就重新映射会“球面”上,计算如式(12)所示。

$$\min_{\theta}\rho(\theta),where\rho(\theta)=E_{(x,y)\sim D}[\max_{\partial\in S}L(\theta,x+\delta,y)] \tag{11}$$

$$x^{t+1}=\prod_{x+S}(x^t+\infty\,\mathrm{sgn}(\nabla_x L(\theta,x,y))) \tag{12}$$

式中:x——原始输入文本;

y——真实的标签;

∞——小步步长;

θ——固定参数。

RP－TEBC通过利用PGD对抗训练算法,在词嵌入层输入到上下文编码器前,通过添加扰动来增加模型的鲁棒性和泛化性。

3 试验与分析

3.1 评估指标

本文试验中采用精确率P(Precision),召回率R(Recall)和$F1$($F1$－measure)值作为评价指标。精确率是指在预测的结果中预测正确的数量占全部结果的比重,召回率是指在预测正确样本被找出来的比重。由于召回率和精确率难以平衡,因此引入调和平均$F1$值,只有精确率和召回率比较高的情况下才能有较高的$F1$值。P、R、$F1$计算如式(13)～式(15)所示。

$$P=\frac{TP}{TP+FP}\times 100\% \tag{13}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{14}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{15}$$

式中:P——阳性;

N——阴性;

TP——预测是P,答案果然是P;

FP——预测是P,答案是N,因此是假的P;

FN——预测是N,答案是P,因此是假的N。

3.2 试验配置

试验所使用的编程语言为Python3.9,使用一块NVIDIA 3090显卡,在CUP型号为Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz,操作系统为Linux的服务器,利用Pytorch1.12.1深度学习框架进行命名实体识别试验。试验中所使用的超参数配置如表2所示。

表 2 超参数配置表

Tab. 2 Hyperparameter configuration table

参数	数值
Learning Rata	0.000 1
Epochs	100
Heads	8
Max Length	128
Dropout	0.5
Batch Size	8
Hidden Dim	128

3.3 与其他模型的对比试验

为了验证RP－TEBC模型对联合收割机故障领域命名实体识别的有效性,本文在相同的试验环境下进行了对比试验,试验结果如表3所示。

表 3 模型对比试验结果

Tab. 3 Model comparison experiment results %

模型	P	R	$F1$
BERT＋BiLSTM ＋CRF	74.53	86.09	79.90
RoBERTa－wwm－ext＋BiLSTM＋CRF	74.62	86.56	80.15
RoBERTa－wwm－ext＋BiGRU＋CRF	74.95	86.28	80.22
RoBERTa－wwm－ext＋FGM＋BiGRU＋CRF	74.60	87.02	80.33
RoBERTa－wwm－ext＋Transformer－Encoder＋BiGRU＋CRF	75.18	86.56	80.47
RP－TEBC	76.33	87.11	81.36

试验以目前主流的命名实体识别模型BERT＋BiLSTM＋CRF为基线模型，从表3中可以看出，RP－TEBC模型相比于基线模型准确率、召回率和$F1$值分别提高了1.79%、1.01%、1.46%，证明RP－TEBC模型对联合收割机故障领域命名实体识别的效率均优于传统的模型。

3.4　消融试验

为了验证加入自适应Transformer编码器和引入对抗训练对于联合收割机故障领域命名实体识别的有效性。故进行了消融试验，结果如表4所示。＋TR表示加入了自适应Transformer编码器层，＋PGD表示加入对抗训练。

表 4　消融试验

Tab. 4 Ablation experiment　　%

序号	设置		P	R	$F1$
	＋TR	＋PGD			
1	×	×	74.95	86.28	80.22
2	√	×	75.12	86.74	80.51
3	×	√	75.78	86.93	80.97
4	√	√	76.33	87.11	81.36

由表4可知，2号模型在加入自适应Transformer编码器后使得模型可以更好的学习到输入文本的语义信息。相比于1号模型（RoBERTa－wwm－ext＋BiGRU＋CRF），2号模型在准确率上提升了0.17%，召回率提升了0.46%，$F1$值提升了0.29%。3号模型引入对抗训练使得模型的泛化性和鲁棒性得到了提升，相比于1号模型3号模型的准确率提升了0.83%，召回率提升了0.65%，$F1$值提升了0.75%。4号模型（RP－TEBC）同时加入了自适应Transformer编码器和对抗训练使得模型，不仅可以很好地学习到输入文本的语义信息而且还增加了模型的泛化性和鲁棒性，使得4号模型的准确率、召回率和$F1$值都比1号、2号、3号模型要高。由此可以得出，加入自适应Transformer编码器的同时引入对抗训练是可以提高联合收割机故障领域命名实体识别的效果。

4　结论

1）本文为实现联合收割机故障诊断命名实体识别任务构建一套专门的数据集。

2）提出RP－TEBC命名实体识别模型。利用自适应Transformer编码器使得模型对输入文本的编码更加适合于命名实体识别任务，通过引入对抗训练使得模型在泛化性和鲁棒性上得到提升。相比于传统的BERT－BiLSTM－CRF模型，实体识别的准确率提升1.79%，召回率提升1.01%，$F1$值提升1.46%。RP－TEBC模型的提出为农机故障领域的命名实体识别模型研究提供参考，同时也为构建相关农机故障领域的知识图谱提供一种新的模型工具。

3）考虑到原始数据不足对模型性能的影响，未来还应标注更多农机故障领域数据，为农机故障领域命名实体识别研究提供可靠的数据支撑。

参　考　文　献

[1] Hammerton J. Named entity recognition with long short-term memory [C]. Proceedings of the Seventh Conference on Natural Language Learning at HLT－NAACL 2003, 2003: 172－175.

[2] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12: 2493－2537.

[3] Huang Z, Xu W, Yu K. Bidirectional LSTM－CRF models for sequence tagging [J]. arxiv preprint arxiv: 1508.01991, 2015.

[4] Dong C, Zhang J, Zong C, et al. Character-based LSTM－CRF with radical-level features for Chinese named entity recognition [C]. Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2－6, 2016, Proceedings 24. Springer International Publishing, 2016: 239－250.

[5] Ma R, Peng M, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER [J]. arxiv preprint arxiv: 1908.05969, 2019.

[6] Zhou J T, Zhang H, Jin D, et al. Dual adversarial neural transfer for low-resource named entity recognition [C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3461－3471.

[7] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arxiv preprint arxiv: 1810.04805, 2018.

[8] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504－3514.

[9] Yan H, Deng B, Li X, et al. TENER: Adapting transformer encoder for named entity recognition [J]. arxiv preprint arxiv: 1911.04474, 2019.

[10] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [J]. arxiv preprint arxiv: 1706.06083, 2017.