# 基于局部对抗训练的命名实体识别方法研究

李　静，程芃森，许丽丹，刘嘉勇
(四川大学网络空间安全学院，成都 610065)

**摘　要**：命名实体识别研究中，数据集内普遍存在实体与非实体，实体内部类别间边界样本混淆的问题，极大地影响了命名实体识别方法的性能. 提出以 BiLSTM-CRF 为基线模型，结合困难样本筛选与目标攻击对抗训练的命名实体识别方法. 该方法筛选出包含大量边界样本的困难样本，利用边界样本易被扰动偏离正确类别的特性，采用按照混淆矩阵错误概率分布的目标攻击方法，生成对抗样本用于对抗训练，增强模型对混淆边界样本的识别能力. 为验证该方法的优越性，设计非目标攻击方式的全局、局部对抗训练方法与目标攻击全局对抗训练方法作为对比实验. 实验结果表明，该方法提高了对抗样本质量，保留了对抗训练的优势，在 JNLPBA、MalwareTextDB、Drugbank 三个数据集上 F1 值分别提升 1.34%、6.03%、3.65%.

**关键词**：命名实体识别；对抗训练；困难样本；目标攻击
**中图分类号**：TP391.1　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.023003

# Name entity recognition based on local adversarial training

*LI Jing*, *CHENG Peng-Sen*, *XU Li-Dan*, *LIU Jia-Yong*
(College of Cybersecurity, Sichuan University, Chengdu 610065, China)

**Abstract**: Boundary samples of different categories staggered on the boundary in the datasets of named entity recognition research, which affects the performance of named entity recognition model. A method based on local adversarial training and BiLSTM-CRF model is proposed to solve the problem above. The method selects hard examples which contain a lot of boundary samples to crafting adversarial samples. The process is based on the characteristics of boundary samples that are easily perturbed to leave from the correct category, and then get adversarial samples from the target attack step according to the confusion matrix error probability distribution. Finally, the datasets mixing with the original data and the adversarial is used to adversarial training to enhance the model's recognition ability. In order to verify the superiority of this method, global/local adversarial training based on non-target attack method and local adversarial training based on target attack are designed as comparative experiments. Experimental results show that the method proposed improves the quality of adversarial samples while retaining the advantages of adversarial training. The F1 scores on the three datasets of JNLPBA, MalwareTextDB, and Drugbank are increased by 1.34%, 6.03%, and 3.65% respectively.

**Keywords**: Named entity recognition; Adversarial training; Hard samples; Target attack

**收稿日期**：2019-06-17
**基金项目**：四川省重点研发项目(2020YFG0076)；四川大学基金(2020SCUNG205)；国家自然科学基金(U2066203,61473197)
**作者简介**：李静 (1995－)，女，硕士研究生，主要研究领域为网络数据分析与信息安全. E-mail:luyabala@qq.com
**通讯作者**：刘嘉勇. E-mail：ljy@scu.edu.cn

## 1　引　言

命名实体识别旨在文本数据中划分实体边界、检测实体类别,是自然语言处理任务中的基础研究之一. 当前命名实体识别研究已取得很多优秀成果[1-5],但多侧重于改进模型结构与特征工程,较少关注命名实体识别数据集中边界样本混淆问题. 如图 1 所示,边界样本混淆是指分类器通过类别标记划分类边界,在类边界的某一范围内邻近类样本交错分布的情况. 混淆的边界样本比远离边界的内部样本更易识别错误,故模型正确识别边界样本的程度对整体识别性能有着至关重要的意义.

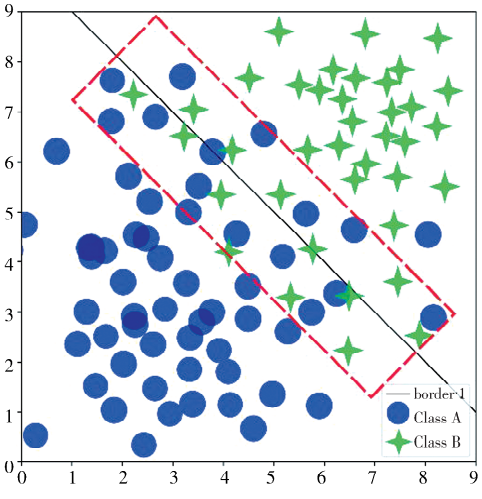


图 1　边界样本混淆示意图

Fig. 1　Overview of confusing boundary samples

传统分类问题研究中常采用基于统计学习的样本筛选方法提高模型边界样本的学习能力. 张莉等人[6]通过聚类方法分析样本离散度挑选出边界样本,剔除了对效果影响不大的冗余样本;周玉等人[7]提出了一种基于最优模糊矩阵诱导的阴影集筛选核心数据与边界数据的方法,可保证分类器的泛化能力;Chen 等人[8]提出多类实例选择(Multiple Class Instances Select, MCIS)方法选出最接近边界的实例,用来提高支持向量机的边界划分速度. 这些方法虽减少了冗余数据,提高识别速度,但也牺牲了原始文本数据的完整性,存在破坏文本结构,可能丢失重要特征的问题.

近年来,深度学习结合对抗训练的方式在文本处理领域表现突出,成为文本研究的新趋势. Miyato 等人[9]深度学习基础上,首次在词向量层面添加扰动,用于半监督文本分类. Zhou 等人[10]在词嵌入层添加扰动提升了低资源命名实体识别模型的泛化能力. 这类方法虽然可以处理更庞大更复杂的特征,但也因对所有文本数据添加扰动合成对抗样本而极大增加了训练数据数量与计算代价.

为解决上述问题,本文在深度学习模型可处理更多特征的基础上,提出基于局部对抗训练的命名实体识别方法. 利用对抗训练既保留原始数据特征,又以对抗攻击的方式提升模型鲁棒性与泛化能力的特点,提升模型识别混淆边界样本的能力. 在困难样本挖掘思想的启发下,仅对数据中易分类错误的困难样本添加扰动,减少冗余对抗样本. 实验表明,本文方法保留对抗训练效果,增强命名实体识别任务性能的同时提高了对抗样本质量.

## 2　相关工作

深度学习可处理更庞大复杂特征的优势,在命名实体识别领域获得了蓬勃的发展. Graves[11]提出了长短时记忆模型 LSTM 解决经典文本处理模型 RNN 的长句依赖问题. Hammerton[12]结合 CRF 的优点,提出的 LSTM+CRF 模型在命名实体任务中表现优异,$F_1$ 值比基线模型提升了 5%. Huang 等人[13]提出的双向 LSTM 结构比起单向 LSTM 可以更好地捕捉前后文的双向语义特征,这种 BiLSTM+CRF 的组合在序列标注问题中表现出了极高的性能,使其逐渐成为命名实体识别中最常见的架构.

对抗训练由对抗生成网络发展而来,最初应用于提升图像处理模型的鲁棒性[14]. 随着具有连续特征的词向量的发展,对抗训练逐渐在文本处理任务中广泛应用. Alzantot 等[15]提出了一种基于种群的优化算法,通过重复随机选择相近的目标标签类的样本,从而找到最近替换词以生成扰动. Li 等[16]捕获对分类有意义的重要单词,再对这些单词添加微小扰动生成对抗样本引导深度学习分类器进行误分类. Gong 等[17]采用梯度下降的方法将词向量扰动为目标类,以此提高对抗文本的质量.

困难样本挖掘思想是将数据分为简单样本与困难样本,在训练过程中选择损失值较大的错误样本送入模型再训练,以提升网络分类性能. Shrivastava 等人[18]提出了一种在线困难样本挖掘(Online Hard Example Mining, OHEM)算法动态选择困难样本,用于解决图像中对象检测调参成本较

高的问题. Li 等人[19]考虑了训练过程不同损失分布的影响，提出根据错误分布抽样训练样本，使困难样本的再训练更有针对性.

针对命名实体识别数据集中存在边界样本混淆的问题，本文基于 BiLSTM-CRF 模型，结合对抗训练与困难样本的思想，筛选数据中损失值较大的困难样本，仅对这部分样本添加目标攻击扰动生成对抗样本；再将对抗样本与原始数据混合进行对抗训练，使模型充分学习类别边界周围困难样本的特征，提高命名实体识别效果.

## 3　局部对抗训练模型

### 3.1　基本概念

3.1.1　非目标与目标攻击　对抗训练是基于对抗攻击的训练方式，在训练过程中对模型进行对抗攻击从而提升模型的鲁棒性. 对抗攻击按照目的的不同可分为非目标攻击与目标攻击. 如图 2(a)所示，非目标攻击是使对抗样本能让模型分错，不指定具体类别. 如图 2(b)所示，目标攻击是使生成的对抗样本被模型错分到某个特定的类别上.

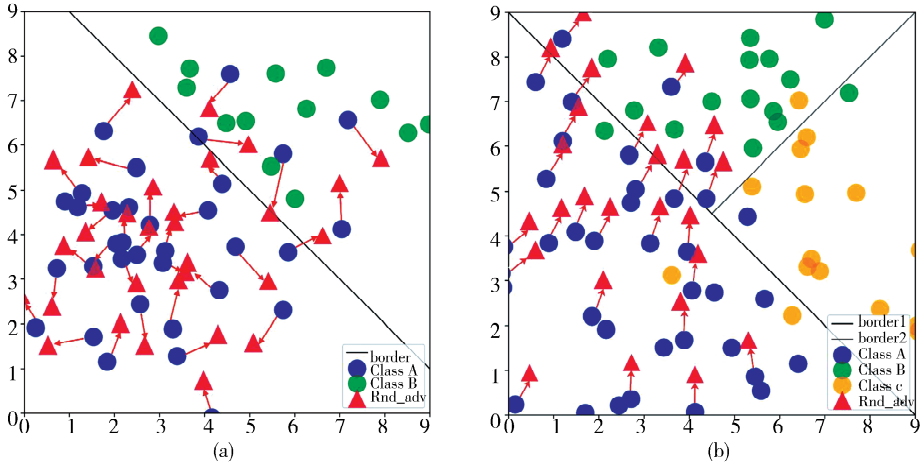


图 2　非目标攻击与目标攻击对比图

Fig. 2　Comparison of non-target and target attack

若设原始样本集为 $X$，$y_{\text{true}}$ 为样本真实类别；$y_{\text{target}}$ 为目标攻击的目标类别；$F$ 为采用的攻击方法；$\epsilon$ 控制扰动大小，非目标攻击的扰动计算为

$$r_{\text{n\_adv}} = \epsilon \cdot F(X, y_{\text{true}}) \tag{1}$$

目标攻击的扰动计算为

$$r_{\text{t\_adv}} = \epsilon \cdot F(X, y_{\text{target}}) \tag{2}$$

根据不同攻击方式对应的对抗样本生成原则[20]，非目标攻击方法生成对抗样本公式为

$$X_{\text{n\_adv}} = X + r_{\text{n\_adv}} \tag{3}$$

目标攻击方法生成对抗样本公式为

$$X_{\text{t\_adv}} = X - r_{\text{t\_adv}} \tag{4}$$

非目标攻击不需要计算扰动方向故而可以快速生成对抗扰动，但在攻击中成功率较低. 有明确指向的目标攻击命中率更高，可以生成更多的使模型分类错误的样本. 模型的反向传播机制决定了分类错误、损失较大的样本对参数权重的调整有更大的价值，故本文选择定向扰动的目标攻击方式生成对抗样本.

3.1.2　全局与局部对抗训练　从样本数据本身来说，若以样本在添加扰动后是否会被分类错误为标准，样本可分为不易被扰动的、处于类边界内部的简单样本，与容易被扰动的、位于边界周围或远离正确类边界的困难样本. 如图 3(a)所示，不进行样本筛选，直接对所有原始样本添加对抗扰动的训练为全局对抗训练；如图 3(b)所示，剔除简单样本，仅对困难样本添加扰动的训练为局部对抗训练.

设 $X_{\text{adv}}$ 为对抗样本集，$ATK$ 为生成对抗样本的攻击方法；$g$，$l$ 作为下标分别表示全局与局部的方法. 全局对抗训练中所有训练样本集合可表示为

$$X_g = X + X_{\text{g\_adv}} \tag{5}$$

其中，$X_{\text{g\_adv}} = ATK_g(X)$. 设 Hard 为困难样本筛选方法，从原始数据中筛选出困难样本，再对困难样本添加扰动生成对抗样本，局部对抗样本集可表示为

$$X_{\text{l\_adv}} = ATK_l(\text{Hard}(X)) \tag{6}$$

局部对抗训练中的所有训练样本集合为

$$X_l = X + X_{\text{l\_adv}} \tag{7}$$

对抗训练过程中，如果直接对所有样本添加扰动，大量简单样本添加扰动后仍位于类别内部，这

些处于类别内部的对抗样本因对反向传播没有贡献而变得冗余. 因此,仅对筛选出的困难样本添加扰动生成对抗样本用于梯度回传,可避免生成大量冗余对抗样本,极大减少训练的计算量.

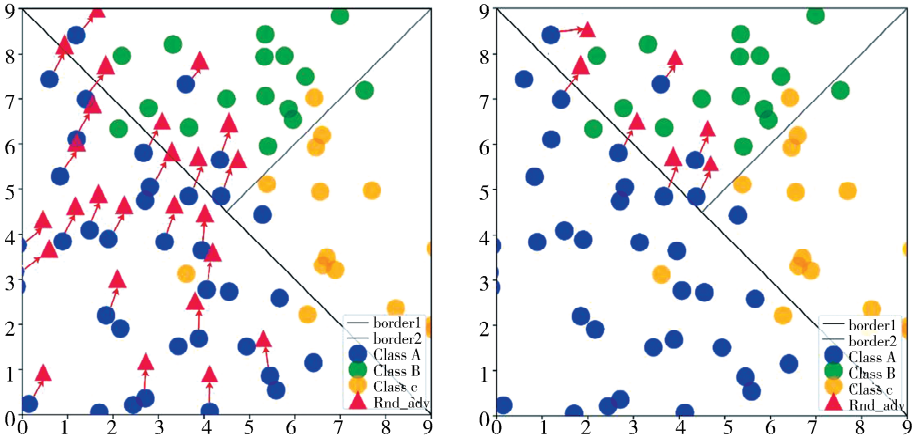


图 3　全局对抗训练与局部对抗训练对比图
Fig. 3　Comparison of global and local adversarial training

### 3.2　局部对抗训练框架

本文提出的局部对抗训练框架见图 4. 原始数据进入深度学习模型前,需将文本中的单词预处理为词向量;再对原始词向量进行损失值大小的评估,以评估结果选择与原始数据识别率相匹配的困难样本筛选比例;然后,根据混淆矩阵错误概率分布按类对困难样本计算目标攻击扰动,添加扰动后生成对抗样本;最后,将对抗样本与原始语料一起用于对抗训练,增强模型识别性能与泛化能力.

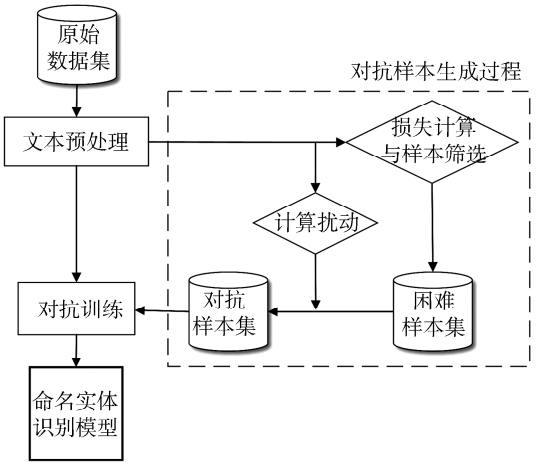


图 4　局部对抗训练模型框架图
Fig. 4　Local adversarial training model

添加扰动的神经网络结构见图 5,$x$ 代表输入文本序列;$w$ 为单词对应的词向量表示;$r$ 为词向量层的扰动;$y$ 为结果序列. Embedding 为词嵌入层,用于预处理文本数据使其向量化;BiLSTM 层同时学习过去与未来的信息,通过前向与反向传播两个隐藏状态的单元获取句子特征;CRF 层学习句子级标签的上下文信息,语句进行序列标注.

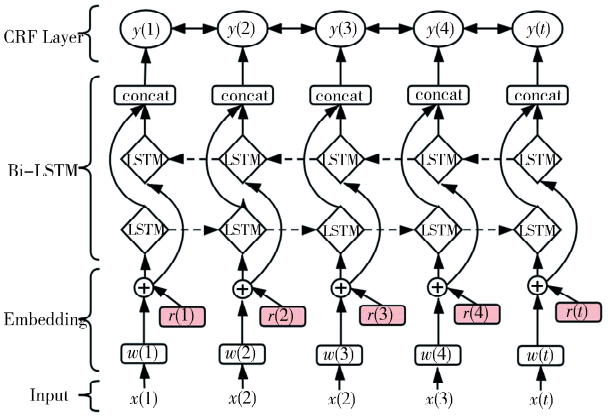


图 5　添加扰动的神经网络结构图
Fig. 5　Neural models with perturbation

3.2.1　困难样本筛选　设置 $D$ 为训练集,$\widetilde{D}$为训练集预处理后的词向量表示;$I,O$ 分别表示文本序列的输入与输出;$M$ 是文本序列长度;$\hat{\theta}$表示参数集合. 困难样本的选取采用按损失值排序筛选的方法,将样本损失值按从大到小的顺序进行排练,选择排在序列前面损失值较大的样本作为困难样本集,设置 Top 表示此筛选方法;$\rho$ 为筛选比例,样本损失值计算为

$$J(\widetilde{D},\hat{\theta})=\frac{1}{|M|}\sum_{(I,O)\in\widetilde{D}}J(I,O,\hat{\theta}) \tag{8}$$

$$J(I,O,\hat{\theta})=-\log\left(P(O|I,\hat{\theta})\right) \tag{9}$$

设最终筛选的困难样本集为$\widetilde{D}_h$，公式表示为

$$\widetilde{D}_h=\mathrm{Top}(\rho,J(\widetilde{D},\hat{\theta}),\widetilde{D}) \tag{10}$$

3.2.2　生成对抗样本　本文提出了一种基于混淆矩阵的目标攻击方式生成对抗样本，简称为 CTR 方法，该方法利用混淆矩阵可反应样本的分类错误占比的特点，对每类样本中的困难样本进行指向错误类的攻击. 设同类别困难样本集合为 $C$，样本数量为 $S$，$C=\{c^{(1)},c^{(2)},\cdots,c^{(S)}\}$. 其中，每个样本都对应共同的真实标签$l_{\mathrm{true}}$，$L$ 对应真实标签集合，标签类别的总数量为 $N$. $L$ 与 $C$ 共同组成训练集，具体表示如下.

$$L=\{l_{\mathrm{true}},\underbrace{\cdots}_{S-2},l_{\mathrm{true}}\} \tag{11}$$

$$\widetilde{D}_h=\mathrm{Top}(\rho,J(\widetilde{D},\hat{\theta}),\widetilde{D})=\{C^{(n)},L^{(n)}\}_{n=1}^{N} \tag{12}$$

设 $C$ 对应的对抗攻击标签序列为$L_{\mathrm{tar}}$，使用 conf($L$)表示按混淆矩阵的错误概率分布排列的标签集合，其关系如下.

$$L_{\mathrm{tar}}=\{l_{\mathrm{tar}}\mid l_{\mathrm{tar}}\in L,l_{\mathrm{tar}}\neq l_{\mathrm{true}}\}=\mathrm{conf}(L) \tag{13}$$

设 $r_{\mathrm{tar}}$为 $C$ 的目标攻击扰动集合，由若干个单位扰动$r_{\mathrm{tar}}^{(s)}$组成. 其中，$\varepsilon$ 用于控制扰动大小；$g$ 表示梯度；$\hat{\delta}$表示所有参数集合，设 $J$ 表示损失计算函数. 关系表示如下.

$$r_{\mathrm{tar}}=\{r_{\mathrm{tar}}^{(s)}\}_{s=1}^{S} \tag{14}$$

$$r_{\mathrm{tar}}^{(s)}=\varepsilon\cdot\frac{g^{(s)}}{\|g\|_2} \tag{15}$$

其中，$g^{(s)}=\nabla_{c^{(s)}}J(C,L_{\mathrm{tar}},\hat{\delta})$. 设 $C_{\mathrm{tar}}$为生成的目标攻击对抗样本集合，与 $L$ 一起组成对抗样本集$\widetilde{D}_{\mathrm{tar}}$，有$(C_{\mathrm{tar}},L)\in\widetilde{D}_{\mathrm{tar}}$，对抗样本计算公式如下.

$$C_{\mathrm{tar}}=C-r_{\mathrm{tar}}=\{c^{(s)}-r_{\mathrm{tar}}^{(s)}\}_{s=1}^{S} \tag{16}$$

3.2.3　对抗训练　训练的最终目的找到最大化真实标签的预测概率，使数据总损失值最小的参数集合.

对抗样本的损失函数计算公式为

$$J(\widetilde{D}_{\mathrm{tar}},\hat{\delta})=\frac{1}{|N|}\sum_{\langle C,L\rangle\in\widetilde{D}_h}J(C_{\mathrm{tar}},L,\hat{\delta}) \tag{17}$$

$\alpha$ 用于控制原始语料与对抗样本损失值比例，对抗训练总损失为

$$\mathrm{Loss}_{\mathrm{L\_ADV}}=\alpha\cdot J(\widetilde{D},\hat{\delta})+(1-\alpha)\cdot J(\widetilde{D}_{\mathrm{tar}},\hat{\delta}) \tag{18}$$

对抗训练最优参数计算为

$$\hat{\delta}=\mathrm{argmin}\{\mathrm{Loss}_{\mathrm{L\_ADV}}\} \tag{19}$$

# 4　实验结果与分析

## 4.1　数据集

为了验证本文方法的性能，选择 3 个专业领域的公开数据集进行了实验. 其中 JNLPBA[21] 为生物领域的数据集，标注了分子生物领域的专业实体，该数据集样本数量相对较多，可用于对比本文方法在不同规模数据集的表现. MalwareTextDB[22] 为恶意软件领域的数据集，其中数据来源于恶意软件报告，数据集中标记了 APT 攻击和恶意软件等实体. Drugbank 为医药领域的数据集[23]. 该数据集收集了大量医药信息，标注了各种药物数据，是医药领域最详细的数据集之一. 实验中对 3 个数据集划分训练集/验证集/测试集. 各数据集的统计信息如表 1.

**表 1　数据集统计信息**

Tab. 1　Static information of datasets

| 数据集 | 领域 | 句子数(实体数) | | |
|---|---|---|---|---|
| | | 训练集 | 验证集 | 测试集 |
| JNLPBA | 生物 | 14 837 (35 431) | 3 709 (10 302) | 3 856 (11 139) |
| MalwareTextDB | 恶意软件 | 4 180 (6 859) | 1 393 (2 204) | 1 393 (2 224) |
| Drugbank | 医药 | 3 483 (8 642) | 1 120 (2 756) | 1 144 (2 821) |

## 4.2　评价指标

本文采用准确率(Precision)、召回率(Recall)和 $F_1$ 值评估各个数据集的学习情况，计算公式如下.

$$precision=\frac{\text{识别正确的实体数}}{\text{识别出的实体数}}\times100\% \tag{10}$$

$$Recall=\frac{\text{识别正确的实体数}}{\text{样本集中的实体数}}\times100\% \tag{11}$$

$$F_1=\frac{2\cdot Precition\cdot Recall}{Precision+Recall}\times100\% \tag{12}$$

## 4.3　实验设置

4.3.1　实验环境　本文实验基于 Tensorflow 深度学习框架设计，采用 Python 语言实现，实验运行平台为 Ubuntu16. 04(64 位)，显存为 8 GB，GPU 为 GTX 1070.

4.3.2　参数设置　本文采用 GLOVE 方法[24] 训练所得的 100 维预训练词向量 glove. 6B. 100d 对文本数据进行预处理. 为更好学习数据特征，批量大小的设置根据数据集的数据量变化，Drugbank、

MalwareTextDB 批量大小设置为 64，JNLPBA 的批量大小设置为 128. 设置 LSTM 隐藏层数为 100，参数优化由 Adam 优化器[25]执行. 根据 Srivastava 等人[26]的经验，设置初始学习率为 0.01，梯度裁剪率为 5.0；为防止过拟合，在嵌入层与 LSTM 输出层设置 dropout 为 0.5. 在对抗样本生成过程中，$\rho$ 表示困难样本筛选比例，根据的 3 个数据集在基线方法的效果，$\rho$ 在 JNLPBA，MalwareTextDB，Drugbank 数据集中分别设置为 30%，50%，20%；根据 Zhou 等人[10]的经验，$\alpha$ 依次从 0.1 至 0.9 中取值，用于平衡原始语料与对抗样本的损失值影响；$\varepsilon$ 从 0.01 至 1 中取值用于控制扰动大小. 最终可从测试集不同参数的训练效果中选出最合适的 $\alpha$，$\varepsilon$ 参数组合.

4.3.3　对比实验　本文以 BiLSTM-CRF 模型为基线方法的同时，设置 3 个对比实验用于证明局部目标对抗训练方法于提升命名实体识别效果的优越性. 快速梯度符号下降(Fast Gradient Sign Method，FGSM)方法[17]为最常用的非目标攻击算法，故 3 组对比实验分别为基于 FGSM 的全局对抗训练、基于 FGSM 的局部对抗训练，基于 CTR 方法的全局对抗训练. 基线方法可用于对比本文方法与其余对抗训练方法在命名实体识别任务的提升效果；全局与局部方法的对比用于证明局部对抗训练是否保持了全局对抗训练的效果，并展示识别率的损失情况. 不同攻击方式的对比用于展现本文中 CTR 攻击方法与局部对抗训练结合的优越性.

### 4.4　结果与分析

4.4.1　实验结果　表 2 展示了基线方法与各种对抗训练方法的实验结果，Baseline 表示基线方法，FGSM_GOL 表示采用 FGSM 方法的全局对抗训练，FGSM_LOC 表示采用 FGSM 方法的局部对抗训练，CTR_GOL 表示采用 CTR 方法的全局对抗训练，CTR_LOC 表示采用 CTR 方法的局部对抗训练.

(1) 表 2 显示，较于基线方法，基于 FGSM 方法与基于 CTR 方法的对抗训练都显著地提升了实体识别 $F_1$ 值. 同时，表 3～表 5 中表现出 3 个数据集的准确率与召回率都有明显的提升，证明了对抗训练对增强命名实体识别效果的有效性. 在 3 个数据集的表现中，最优识别率均出现在运用了 CTR 方法的对抗训练方法. JNLPBA 数据集中，CTR_GOL 方法 $F_1$ 值比基线方法高 1.63%；MalwareTextDB 数据集中，CTR_LOC 方法的 $F_1$ 值比基线方法高 6.03%；Drugbank 数据集中，CTR_LOC 方法的 $F_1$ 值比基线方法高 3.65%. 其中 CTR_LOC 方法在 3 个数据集的召回率分别提升 0.88%、8.23%、3.74%. 召回率与 $F_1$ 值的明显提高，说明了该方法有效缓解了边界样本因混淆而难以识别的问题，增强了模型的泛化能力.

(2) 不同攻击方式的对抗训练方法之间具有差异. 在采用 FGSM 方法的两个实验中，JNLPBA、MalwareTextDB 和 Drugbank 等 3 个数据集的局部对抗训练 $F_1$ 值均低于全局模式的效果，分别降低 0.49%，0.15%和 0.55%. 在采用 CTR 方法的两个实验中，MalwareTextDB 和 Drugbank 数据集的局部对抗训练效果较于全局对抗训练分别增加 2.40%和 0.47%，JNLPBA 数据集在局部对抗训练的效果较全局对抗训练降低 0.29%. 局部对抗训练相比于全局对抗训练的识别效果虽然具有细小的波动，但基本维持了全局对抗训练的效果，并且在 3 个数据集中分别减少了 70%，50%，80%(困难样本筛选中的简单样本淘汰比例为 1-$\rho$)的生成对抗样本的计算量，极大地减少了冗余对抗样本的生成，提升了对抗训练的质量.

**表 2　实验结果比较($F_1$ 值)**

Tab. 2　Compared with other experiments ($F_1$ score)

| 方法 | 数据集 | | |
|---|---|---|---|
| | JNLPBA | MalwareTextDB | Drugbank |
| Baseline | 72.57 | 50.57 | 83.16 |
| FGSM_GOL | 74.02 | 56.40 | 86.66 |
| FGSM_LOC | 73.53 | 56.25 | 86.21 |
| CTR_GOL | 74.20 | 54.20 | 86.34 |
| CTR_LOC | 73.91 | 56.60 | 86.81 |

**表 3　JNLPBA 数据集上不同实验结果**

Tab. 3　Experimental result in JNLPBA

| 方法 | 准确率/% | 召回率/% | $F_1$ 值/% |
|---|---|---|---|
| Baseline | 69.42 | 76.01 | 72.57 |
| FGSM_GOL | 71.35 | 76.89 | 74.02 |
| FGSM_LOC | 70.37 | 76.98 | 73.53 |
| CTR_GOL | 71.84 | 76.73 | 74.20 |
| CTR_LOC | 71.15 | 76.89 | 73.91 |

**表 4　MalwareTextDB 数据集上不同实验结果**

Tab. 4　Experimental result in MalwareTextDB

| 方法 | 准确率/% | 召回率/% | $F_1$值/% |
|---|---|---|---|
| Baseline | 47.69 | 53.81 | 50.57 |
| FGSM_GOL | 54.50 | 58.43 | 56.40 |
| FGSM_LOC | 51.49 | 61.97 | 56.25 |
| CTR_GOL | 53.90 | 54.58 | 54.20 |
| CTR_LOC | 52.03 | 62.04 | 56.60 |

**表 5　Drugbank 数据集上不同实验结果**

Tab. 5　Experimental result in Drugbank

| 方法 | 准确率/% | 召回率/% | $F_1$值/% |
|---|---|---|---|
| Baseline | 83.53 | 82.84 | 83.16 |
| FGSM_GOL | 86.57 | 86.75 | 86.66 |
| FGSM_LOC | 85.75 | 86.69 | 86.21 |
| CTR_GOL | 86.10 | 86.58 | 86.34 |
| CTR_LOC | 87.06 | 86.58 | 86.81 |

4.4.2　结果分析　(1) JNLPBA 数据集在对抗训练中的提升低于其数据集，分析原因应为 JNLPBA 数据集中样本数量更大，为模型学习提供了更加充足的特征，故对抗训练在此类大样本数据中不能发挥最优作用；而样本数量相对较少的 MalwareTextDB 与 Drugbank 数据集在合成对抗样本的环节中变相扩充了语料数据，仅添加微小扰动的对抗样本分布在原始样本周围，对模型充分学习样本特征具有积极的意义. 除此之外，在 MalwareTextDB 数据集的全局对抗训练中，FGSM 方法高出 CTR 方法 2.20%. 这种明显的差异可能源于该数据集中原始样本的识别率较低，使指向错误分类的目标攻击对抗样本超过最合适的对抗训练比例，导致效果明显低于其他对抗训练方法.

(2) FGSM 为非目标攻击方法，训练效果的提升主要依赖于大量随机方向的对抗样本对模型充分学习样本特征，故采用局部对抗训练时，对抗样本的减少与非目标攻击成功率低的双重作用下，造成识别效果的损失. CTR 是基于目标攻击思想的方法，效果提升主要依赖于错误分类样本在模型参数优化机制上的重要性. 与困难样本筛选结合后不影响分类错误的对抗样本的生成，反而降低类别内部的对抗样本对训练效果的影响，从而能出现对抗训练效果不降反升的情况. 从实验中可得出，对抗训练是提升命名实体识别模型性能的有效手段，困难样本筛选是提高对抗训练质量的辅助办法.

## 5　结　论

本文从边界样本的角度出发，提出了一种基于混淆矩阵错误概率分布的目标攻击方法，并结合困难样本的思想提出了局部对抗训练方案，用于命名实体识别研究. 该方法以 BiLSTM-CRF 模型为基线模型，采用困难样本筛选的思想，筛选出对模型性能有关键影响的，包含大量边界样本的困难样本；利用边界样本易被扰动的特性，结合基于混淆矩阵错误概率分布的目标攻击方法生成对抗样本用于对抗训练. 实验结果证明了 CTR 方法在对抗训练的有效性，也证明了本文提出的 CTR 结合困难样本的局部对抗训练方法的优异性. 该方案不仅有效缓解了边界样本混淆限制命名实体识别性能的问题，极大提升命名实体识别效果，而且减少了常规对抗训练中增加计算成本的冗余对抗样本，保留了对抗训练效果的同时提高了对抗样本质量. 下一步工作将考虑进一步优化对抗攻击方法，使对抗样本在对抗训练中发挥更积极的作用.

**参考文献：**


[1] Santos C N, Guimaraes V. Boostingnamed entity recognition with neural character embeddings [C]// Proceedings of NEWS 2015 the 5th Named Entities Workshop. Stroudsburg, PA: Association for Computational Linguistics, 2015.

[2] Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2017.

[3] Maryam H, Leon W, Mariana N, *et al*. Deep learning with word embeddings improves biomedical named entity recognition [J]. Bioinformatics, 2017, 14: 14.

[4] GUL K S Q, 尹继泽, 潘丽敏. 基于深度神经网络的命名实体识别方法研究[J]. 信息网络安全, 2017, 17: 29.

[5] 许丽丹, 刘嘉勇, 何祥. 一种解决命名实体识别数据集类别标记失衡的方法[J]. 四川大学学报：自然科学版, 2020, 57: 82.

[6] 张莉, 郭军. 基于边界样本的训练样本选择方法[J]. 北京邮电大学学报, 2006, 29: 77.

[7] 周玉, 朱安福, 周林. 一种神经网络分类器样本数据选择方法[J]. 华中科技大学学报：自然科学版,

2012, 40: 39.

[8] Chen J, Zhang C, Xue X, *et al*. Fast instance selection for speeding up support vector machines [J]. Knowl-Based Syst, 2013, 45: 1.

[9] Miyato T, Dai A M, Goodfellow I. Adversarial-training methods for semi-supervised text classification [J]. arXiv preprint, 2016, 1605: 07725.

[10] 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. Stroudsburg, PA: Association for Computational Linguistics, 2019.

[11] Graves A. Supervised sequence labelling with recurrent neural networks [M]. Berlin: Springer, 2012.

[12] Hammerton J. Namedentity recognition with long short-term memory [C]//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL. Stroudsburg, PA: Association for Computational Linguistics, 2003.

[13] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF-models for sequence tagging [J]. arXiv preprint, 2015, 1508: 01991.

[14] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [C]// Proceedings of the International Conference on Machine Learning. Lille, France: International Machine Learning Society, 2015.

[15] Alzantot M, Sharma Y, Elgohary A, *et al*. Generating natural language adversarial examples [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Procssing. Stroudsburg, PA: Association for Computational Linguistics, 2018.

[16] Li J, Ji S, Du T, *et al*. TextBugger: Generating adversarial text against real-world applications [C]// Proceedings of the Network and Distributed System Security Symposium. San Diego, CA: Internet Society, 2019.

[17] Gong Z, Wang W, Li B, *et al*. Adversarial texts with gradient methods [J]. arXiv preprint, 2018, 1801: 07175.

[18] Shrivastava A, Gupta A, Girshick R. Trainingregion-based object detectors with online hard example Mining [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE Computer Society, 2016.

[19] Li M, Zhang Z, Yu H, *et al*. S-OHEM: Stratified online hard example mining for object detection [C]// Proceedings of the 2nd Chinese Conference on Computer Vision. Singapore: Springer, 2017.

[20] Kurakin A, Goodfellow I J, Bengio S. Adversarial machine learning at scale [C]//Proceedings of the International Conference on Learning Representations 2017. [S. l.: s. n.], 2017.

[21] Kim J D, Ohta T, Tsuruoka Y, *et al*. Introduction to thebio-entity recognition task at JNLPBA [C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). Stroudsburg, PA: Association for Computational Linguistics, 2004.

[22] Lim S K, Muis A O, Lu W, *et al*. Malwaretextdb: A database for annotated malware articles [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017.

[23] Vivian L, Craig K, Yannick D, *et al*. DrugBank 4.0: Shedding new light on drug metabolism [J]. Nucleic Acids Res, 2014, 42: D1091.

[24] Pennington J, Socher R, Manning C. Glove: global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics, 2014.

[25] Kingma D P, Ba J. Adam: Amethod for stochastic optimization [J]. arXiv preprint, 2014, 1412: 6980.

[26] Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting [J]. J Mach Learn Res, 2014, 15: 1929.

**引用本文格式:**

中　文: 李静, 程芃森, 许丽丹, 等. 基于局部对抗训练的命名实体识别方法研究[J]. 四川大学学报: 自然科学版, 2021, 58: 023003.

英　文: Li J, Cheng P S, Xu L D, *et al*. Name entity recognition based on local adversarial training [J]. J Sichuan Univ: Nat Sci Ed, 2021, 58: 023003.