◆机器学习模型◆

基于改进损失函数的实体类别平衡优化算法*

张俸玺[1],吴丞楚[1],张运泽[1],董洛兵[2]
(1. 西安电子科技大学通信工程学院,陕西西安 710071;2. 西安电子科技大学计算机科学与技术学院,陕西西安 710071)

摘要:针对自然语言处理(Natural Language Processing,NLP)任务中,命名实体识别(Named Entity Recognition,NER)存在实体类别样本不平衡的问题,提出一种基于改进损失函数的实体类别平衡优化算法。新算法是对神经网络模型中的损失函数进行优化处理,通过分析命名实体识别数据特点,在平衡正负样本的基础上引入平滑系数和权重系数,保证模型在梯度传递的过程更关注于实体类别较少和带有嵌套的难识别样本,同时减少对样本数较多的、易识别样本的关注。利用公共数据集 ACE05、MSRA 进行实验对比,结果表明改进的损失函数在数据集 ACE05 和 MSRA 上,$F1$ 值分别提高 1.53%和 0.91%。上述结果表明改进的损失函数能够较好地缓解实体中正负难易样本的不平衡。

关键词:自然语言处理;命名实体识别;损失函数;平滑系数;神经网络;难易样本

中图分类号:TP391 **文献标识码**:A **文章编号**:1005-9164(2023)01-0100-06
DOI:10.13656/j.cnki.gxkx.20230308.011

命名实体识别(Named Entity Recognition,NER)[1]是自然语言处理(Natural Language Processing,NLP)中的一项最基础的任务,其性能的好坏直接影响着下游任务,如关系抽取[2]、机器翻译[3]、知识图谱构建[4]等。目前,命名实体识别主流方法是基于神经网络的监督性学习模型,而命名实体识别数据集中常存在实体类别样本数量较少和实体嵌套现象,导致模型学习实体类别特征的能力较弱。命名实体识别数据集实体类别较少,导致模型识别实体类别时出现正负样本不平衡现象;嵌套实体是指实体内部嵌套了多个命名实体,由于识别嵌套实体属于自然语言处理中较为困难的任务,所以嵌套实体被列为难以识别的实体,嵌套实体过多则导致难易样本不平衡问题,从而影响模型的整体性能。

传统的命名实体识别方法包括基于序列标记[5]的方法、基于超图[6]的方法和基于跨度的方法。其中,基于序列标记的方法通过预测每个字对应的标签来解决嵌套实体;基于超图的方法与序列标记上的解码不同,它是基于实体嵌套结构来构建超图,从而解码超图上的实体;基于跨度的方法通过枚举[7]或识别边界[8]来提取跨度,然后对跨度进行分类。尽管上述方法在解决实体嵌套问题上具有一定的可行性,但是

收稿日期:2022-09-21 修回日期:2022-10-10
* 国家级大学生创新创业训练计划项目(202110701085)资助。

【第一作者简介】
张俸玺(2002-),男,在读本科生,主要从事计算机视觉、自然语言处理和分布式深度学习研究,E-mail:fxzhang_1@stu.xidian.edu.cn。

【引用本文】
张俸玺,吴丞楚,张运泽,等. 基于改进损失函数的实体类别平衡优化算法[J]. 广西科学,2023,30(1):100-105.
ZHANG F X,WU C C,ZHANG Y Z,et al. Entity Category Balance Optimization Algorithm Based on Improved Loss Function [J]. Guangxi Sciences,2023,30(1):100-105.

在实体类别较少的情况下仍显不足。最近,利用外部知识增强实体类别特征学习的方法日益受到人们的关注。Li 等[9]和 Xue 等[10]将命名实体识别转换为机器阅读理解(MRC)任务,通过定义实体类别的先验知识来构造实体类型的查询。除此之外,自训练[11-15]的方法也广泛用于图像处理和自然语言处理中,并以此来解决少样本问题。Lin 等[16]在图像处理领域为损失函数增添平滑因子来缓解正负难易样本不平衡的问题。Li 等[17]为损失函数引入密度函数,使模型较少关注某些离散群点。基于前述分析,本研究提出一种基于改进损失函数的实体类别平衡优化算法,拟通过为交叉熵损失函数添加平滑系数和权重因子,在缓解正负样本不平衡的同时缓解难易样本的不平衡问题。

1 损失函数

在深度学习中,损失函数用于计算模型预测值与真实值之间的差距,以判断模型算法的优劣性。损失函数常用于回归问题与分类问题,而命名实体识别属于分类问题中的二分类问题,其对应的损失函数为 0-1 损失。0-1 损失属于不连续的分段函数,所以只有当函数为连续凸函数,并在任意取值下是 0-1 损失函数上界时,才能正确地构造出其代理函数[18]。梯度是一种迭代法,在求解模型损失函数最小值时,就是通过梯度下降来一步步迭代求解(图 1,图中星星代表梯度下降走过的路径),而且通过梯度下降,可以使损失函数最小,从而反向调节模型参数。

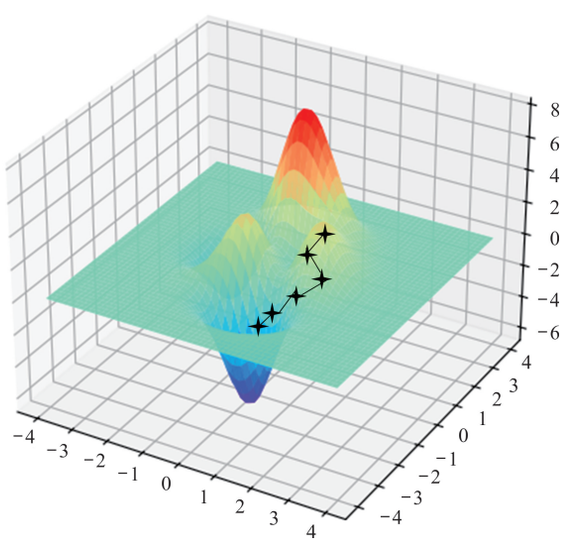

图 1 损失函数梯度下降图

Fig. 1 Loss function gradient descent diagram

1.1 梯度下降

梯度下降的计算过程就是沿着梯度的方向向下逐步求解极小值,也是求解局部最优的过程。其迭代公式为

$$a_{k+1}=a_k+\rho_k\bar{s}^{(k)}, \tag{1}$$

式中,a_k 表示上一步梯度下降的结果;$\bar{s}^{(k)}$ 表示梯度的负方向;ρ_k 表示搜索方向的步长,步长的长短直接影响模型在梯度下降过程中的收敛速度,步长过大可能导致模型无法获得局部最优,步长过小则导致收敛速度过慢。

1.2 交叉熵损失函数

命名实体识别由于涉及的是分类问题,所以常用交叉熵损失函数作为模型中用于梯度下降的损失函数。在分类的应用中,交叉熵损失函数的最小化交叉熵等价于最小化观测值和估计值的相对熵,因此它提供了一个无偏估计的代理损失,是在分类任务中应用最广的损失函数,公式如下:

$$g(p,y)=\begin{cases}-\log_2(p), & y=1\\ -\log_2(1-p), & \text{其他}\end{cases}, \tag{2}$$

式中,$y\in\{-1,1\}$,$p\in[0,1]$,y 是一个真实类,p 是模型对标签为 $y=1$ 的类的估计概率。

2 改进的交叉熵损失函数

在分析命名实体识别数据中发现,数据集常存在实体类别较少和嵌套的问题,导致模型不能很好地识别。交叉熵损失函数并没有关注数据间的不平衡,而是单纯地度量两个概率分布间的差异,因此无法缓解上述不平衡问题。本研究借鉴目标检测领域中难易样本平衡的方法来解决该类问题,目标检测中难易样本例子如图 2 所示。

从图 2 可以观察到,目标检测领域同样存在难易样本不均衡,即难(易)样本太多,易(难)样本太少的问题,使得普通的交叉熵损失函数在对模型反向调节的过程中过度关注于不重要的信息,而忽略了重要的信息。基于上述思想,本研究在交叉熵损失函数的基础上引入一个平滑系数 p_t[16]:

$$p_t=\begin{cases}(1-p)^{\theta}, & y=1\\ p^{\theta}, & \text{其他}\end{cases}, \tag{3}$$

式中,$\theta\geqslant 0$ 是一个关注度系数,可以通过设置不同的 θ 来减少或增强对难易样本的关注度,从而缓解难易样本不平衡所导致的模型识别较差问题。

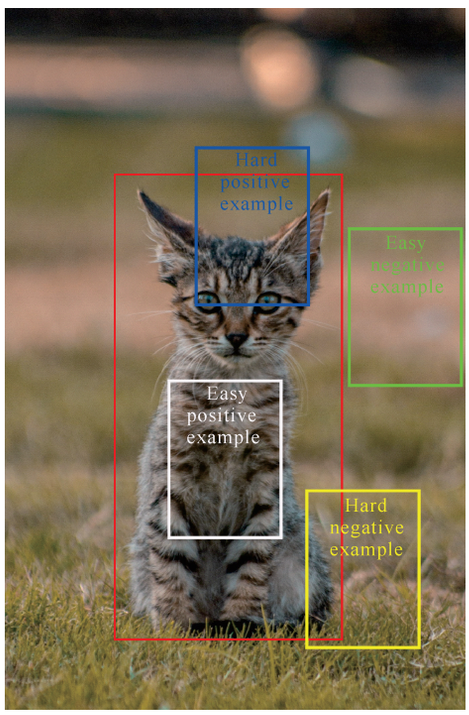

图 2　目标检测领域中难易样本识别例子

Fig. 2　Examples of hard and easy sample recognition in the field of target detection

由于命名实体识别数据集中存在较为普遍的正负样本不平衡现象,本研究在交叉熵损失函数的基础上同时引入一个权重系数 α[19],以调节正负样本的损失权重和,得到最终改进的损失函数:

$$F_{\text{loss}} = \alpha p_t g(p, y)。 \tag{4}$$

另外,可通过数学分析来直观地了解改进的损失函数的作用。例如,为了平衡难易样本,当 $\theta = 2$ 和 $p = 0.9$ 时,针对易分类样本,其关注度降低了 100 倍;而当 $p = 0.1$ 时,样本明显是一个难分类样本,模型识别存在困难,通过增添平滑系数,在降低易分类样本损失权重贡献的基础上又重点关注了难分类样本。同时,通过添加权重系数 α 来降低(提高)某一方样本过多(少)对损失权重的贡献。

3　实验仿真及分析

3.1　数据集

本研究提出基于改进损失函数的实体类别平衡优化算法,实验数据均来自公共数据集 MSRA 和 ACE05,数据样本的详细信息如表 1 所示。

3.2　评价指标

本研究利用改进的损失函数对命名实体识别任务进行验证,实验结果用精准率(P)、召回率(R)和 $F1$值来说明。其中,精准率表示为模型预测为正的样本占所有正样本的比重,召回率表示为样本中的正样本被预测正确的比重,$F1$ 值表示为精准率和召回率的调和平均,公式如下:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{7}$$

式中,TP 表示把实体正类预测为正类的数量,FP 表示把实体负类预测为正类的数量,FN 表示把实体正类预测为负类的数量。

表 1　数据集样本统计信息

Table 1　Dataset sample statistics information

数据集 Dataset	标签 Label	训练集 Train dataset	测试集 Test dataset
MSRA	LOC (Place name)	86 849	7 291
	ORG (Organization)	103 261	6 977
	PER (Person)	51 738	5 824
ACE05	PER (Person)	21 771	2 470
	ORG (Organization)	10 643	1 261
	GPE (Geography/society/politics)	13 327	2 092
	LOC (Place name)	2 350	263
	FAC (Facility)	2 568	246
	VEH (Vehicle)	1 027	155
	WEA (Weapon)	628	42

3.3　实验环境及参数设置

本研究是在 tensorflow 1.14 和 python 3.6 的环境下进行实验。设置模型在训练、测试时的 batch_size 为 128;为降低在模型训练过程中出现的过拟合现象,将 dropout 设置为 0.5;学习率设置为 0.001,Adam 作为模型的优化器;改进损失函数中的 θ 分别设置为 0,1,2,4,α 设置为 0.25 以验证其有效性。BERT 预训练模型选择 base 版本。

3.4　实验结果

为验证改进损失函数的实体类别平衡优化算法的有效性,本研究利用 BERT 和 BERT + F_{loss} 进行实验对比,如表 2 所示。从表 2 可以看出,改进损失函数算法的实体类别平衡优化有效地提高了实体识别的性能。具体表现为在 MSRA 数据集中 $F1$ 值提高了 0.91%;在 ACE05 数据集中,数量最少的实体类别 VEH 和 WEA,$F1$ 值分别提高 1.36% 和

5.05%,总体 $F1$ 值提高了 1.53%。原因在于:MSRA 数据集嵌套实体较少,实体类别样本较多,正负难易样本不平衡的情况较少,所以在识别性能上提升较小;ACE05 数据集中存在大量的嵌套实体和个别实体类别样本较少的情况,通过改进损失函数反向调节模型,缓解了模型在正负难易样本识别上的不平衡。

表 2　改进损失函数的实体类别平衡优化实验对比

Table 2　Experimental comparison of entity category balance optimization based on improved loss function

数据集 Dataset	模型 Model	标签 Label	精准率(%) P(%)	召回率(%) R(%)	$F1$ (%)
MSRA	BERT	LOC	91.88	94.36	93.10
		ORG	84.70	95.03	89.57
		PER	97.49	97.44	97.47
		Total	91.21	95.45	93.28
	BERT + F_{loss}	LOC	93.93	94.24	94.08
		ORG	89.17	93.60	91.33
		PER	97.64	96.88	97.26
		Total	**93.56**	**94.82**	**94.19**
ACE05	BERT	PER	84.69	90.28	87.40
		ORG	74.55	83.33	78.70
		GPE	80.78	83.17	81.96
		LOC	61.82	64.15	62.96
		FAC	69.23	61.36	65.06
		VEH	69.34	71.43	70.37
		WEA	60.53	69.70	64.79
		Total	79.35	83.96	81.59
	BERT + F_{loss}	PER	87.72	89.61	88.66
		ORG	77.40	85.17	81.10
		GPE	82.96	82.63	82.80
		LOC	63.94	62.74	63.33
		FAC	69.78	71.36	70.56
		VEH	81.73	63.91	71.73
		WEA	73.33	66.67	69.84
		Total	**82.18**	**84.09**	**83.12**

Note:the bolded data in the table is the optimal value

结合改进的损失函数,在两个数据集上对模型进行实验观察(图 3)。从图 3 可以观察出,结合了改进损失函数的命名实体识别模型相比于结合交叉熵损失函数的命名实体识别模型,损失函数收敛速度明显加快,其原因在于为交叉熵损失函数添加平滑系数和权重系数后,模型更加关注于难识别样本,同时平衡正负样本,可以有效提高模型识别性能。

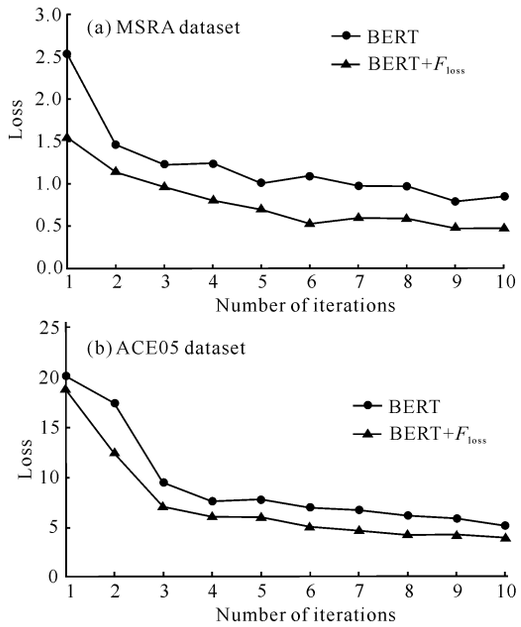

图 3　损失函数迭代变化

Fig. 3　Iterative change of loss function

平滑系数中不同的 θ 值,会对模型产生不同的影响,结果如表 3 所示。从表 3 可以看出,当 $\theta=0$ 时,改进后的损失函数与原损失函数相同;当 $\theta=1$ 时,由于平滑系数变化不大,导致其对难识别样本关注度不高,性能提升较小;当 $\theta=2$ 时,两个数据集均达到了最优性能,原因在于此时的损失函数能够很好地抑制易识别样本,而且也能很好地关注难识别样本;当 $\theta=4$ 时,由于平滑系数调节过大,导致模型过度关注于难识别样本,反而忽略了易识别样本,导致模型识别性能降低。因此,从实验结果可以看出,当数据集难易样本不平衡时,改进损失函数所呈现出的性能提升

表 3　平滑系数 θ 对实验结果的影响

Table 3　Effect of θ on experimental results

数据集 Dataset	平滑系数 θ	精准率(%) P(%)	召回率(%) R(%)	$F1$(%)
MSRA	0	91.21	95.45	93.28
	1	92.04	95.32	93.65
	2	**93.56**	**94.82**	**94.19**
	4	90.65	95.02	92.78
ACE05	0	79.35	83.96	81.59
	1	80.79	83.02	81.89
	2	**82.18**	**84.09**	**83.12**
	4	78.32	83.13	80.65

Note:the bolded data in the table is the optimal value

较为明显;当数据集中难易样本较为平衡时,模型提升性能较小。由此说明,本研究改进的损失函数适合于数据集存在难易样本不平衡的情况,并表明了其有效性。

为验证权重系数和改进损失函数的通用性,在其他交叉熵模型 MRC[9] 上对数据集 ACE05 进行验证,实验结果如表 4 所示。实验结果表明,为交叉熵损失函数添加权重系数后,$F1$ 值有一定提升,并且改进损失函数在其他交叉熵模型上同样取得了不错的性能提升,由此验证了改进损失函数的通用性。上述结果说明公式(4)中权重系数对于平衡样本存在一定贡献,并且改进损失函数在其他交叉熵模型上也同样有效。

表 4 权重系数和改进损失函数通用性实验结果

Table 4 Experimental results of generalization of weighting coefficients and improved loss functions

模型 Model	权重系数 α	精准率(%) P(%)	召回率(%) R(%)	$F1$(%)
MRC	0.1	79.09	80.12	79.60
	0.2	81.62	80.94	81.28
	0.4	82.99	82.66	82.82
	0.8	**83.41**	**83.54**	**83.47**
MRC + F_{loss}	0.1	83.78	84.01	83.89
	0.2	**84.47**	**84.81**	**84.63**
	0.4	83.89	83.65	83.77
	0.8	83.55	83.74	83.64

Note:the bolded data in the table is the optimal value

4 结论

本文提出了一种基于改进损失函数的实体类别平衡优化算法,缓解了命名实体识别中的正负难易样本不平衡所导致的模型识别性能较低的问题。通过在两个公共数据集上的实验,证明了本文所提出的改进策略在命名实体识别分类模型中的有效性。在未来工作中,将根据不同数据集的特点,探索进一步的改进方式。

参考文献

[1] 许力,李建华. 基于 BERT 和 BiLSTM-CRF 的生物医学命名实体识别[J]. 计算机工程与科学,2021,43(10):1873-1879.

[2] XIONG C Y,LIU Z,CALLAN J,et al. Towards better text understanding and retrieval through kernel entity salience modeling [C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor,MI:SIGIR,2018:575-584.

[3] SONG J,KIM S,YOON S. AligNART:non-autoregressive neural machine translation by jointly learning to estimate alignment and translate [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL,2021:1-14.

[4] JI S,PAN S,CAMBRIA E,et al. A survey on knowledge graphs:representation,acquisition,and applications [J]. IEEE Transactions on Neural Networks and Learning Systems,2021,33(2):494-514.

[5] HUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging [Z/OL]. (2015-08-09) [2022-09-12]. https://arxiv.org/pdf/1508.01991.pdf.

[6] LU W,ROTH D. Joint mention extractionand classification with mention hypergraphs [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 857-867.

[7] SOHRAB M G,MIWA M. Deep exhaustive model for nested named entity recognition [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2843-2849.

[8] ZHENG C,CAI Y,XU J,et al. A Boundary-aware neural model for nested named entity recognition [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg,PA:ACL,2019:357-366.

[9] LI X,FENG J,MENG Y,et al. A unified MRC framework for named entity recognition [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL,2020:5849-5859.

[10] XUE M,YU B,ZHANG Z,et al. Coarse-to-Fine pre-training for named entity recognition [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg,PA:ACL,2020:6345-6354.

[11] NIU Y,JIAO F,ZHOU M,et al. A Self-Training method for machine reading comprehension with soft evidence extraction [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:ACL,2020:3916-3927.

[12] MENG Y,ZHANG Y,HUANG J,et al. Distantly-su-

pervised named entity recognition with noise-robust learning and language model augmented self-training [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL,2021:10367-10378.

[13] ZOPH B,GHIASI G,LIN T Y,et al. Rethinking pre-training and self-training [C]//LAROCHELLE H,RANZATO M,HADSELL R,et al. Advances in Neural Information Processing Systems 33. [S. l.]:NeurIPS,2020:3833-3845.

[14] MI F,ZHOU W,KONG L,et al. Self-training improves pre-training for few-shot learning in task-oriented dialog systems [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL,2021:1887-1898.

[15] DU J,GRAVE É,GUNEL B,et al. Self-training improves pre-training for natural language understanding [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg,PA:ACL,2021:5408-5418.

[16] 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. Piscataway,NJ:IEEE,2017:2980-2988.

[17] LI B,LIU Y,WANG X. Gradient harmonized single-stage detector [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park,CA:AAAI,2019:8577-8584.

[18] 周志华. 机器学习[M]. 北京:清华大学出版社,2016:121-139,298-300.

[19] XIE S,TU Z. Holistically-nested edge detection [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2015:1395-1403.

Entity Category Balance Optimization Algorithm Based on Improved Loss Function

ZHANG Fengxi[1],WU Chengchu[1],ZHANG Yunze[1],DONG Luobing[2]

(1. School of Telecommunications Engineering, Xidian University, Xi′an, Shaanxi, 710071, China; 2. College of Computer Science & Technology,Xidian University,Xi′an,Shaanxi,710071,China)

Abstract:Aiming at the problem of unbalanced entity category samples in Named Entity Recognition (NER) in Natural Language Processing (NLP) tasks,an entity category balance optimization algorithm based on improved loss function is proposed. The new algorithm is to optimize the loss function in the neural network model. By analyzing the characteristics of named entity recognition data,the smoothing coefficient and the weight coefficient are introduced on the basis of balancing the positive and negative samples to ensure that the model pays more attention to the difficult recognition samples with fewer entity categories and nesting in the process of gradient transfer,while reducing the focus on easy-to-identify samples with more samples. Using the public datasets ACE05 and MSRA for experimental comparison,the results show that the improved loss function is on the data sets ACE05 and MSRA,and $F1$ value increases by 1.53% and 0.91%,respectively. The above results show that the improved loss function can better alleviate the imbalance of positive and negative difficult and easy samples in the entity.

Key words:natural language processing; named entity recognition; loss function; smoothing coefficient; neural networks; difficult and easy examples

责任编辑:米慧芝