文章编号：1003-0077(2023)04-34-11

融合字符结构特征的多任务老挝语文字识别研究

陈 琢[1]，周兰江[1]，郝永彬[1]，张建安[2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 战略支援部队 信息工程大学三院昆明大队，云南 昆明 650500)

摘 要：老挝语属于资源稀缺型语言，直接从互联网中获取老挝语文本语料较为困难，老挝语文字识别研究可在有限的图片文本资源中获取更多的老挝语文本语料。在开展老挝文字光学字符识别的研究工作中，针对老挝单字符误切分、上/下位元音以及音调识别位置存在偏差和相似老挝字符的识别问题，该文研究了老挝字符书写等级和下位辅音，提出一种有效融合老挝字符结构特征的多任务字符识别方法。首先，利用深度残差网络提取字符图片中的老挝字符结构特征，通过边框回归矫正单字符包围框；其次，将已矫正切分结果和提取的字符特征作为联合输入，通过双向长短时记忆网络预测老挝字符序列，利用连接主义时间分类对预测结果进行序列对齐；最后，根据老挝字符固定组合优化模型预测结果。实验结果表明：该方法可以精确识别已切分的老挝字符序列，字符错误率指标低至 13.06%。

关键词：老挝印刷字符识别；老挝字符结构特征；多任务识别；端到端模型

中图分类号：TP391 **文献标识码**：A

A Multi-task Approach to Lao Character Recognition with Structural Features

CHEN Zhuo, ZHOU Lanjiang, HAO Yongbin, ZHANG Jian'an

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Kunming Brarch, No.3 College, PLA Information Engineering University, Kunming, Yunnan 650500, China)

Abstract: Focused on the Optical Character Recognition of Lao script, this paper investigates the problems of Lao characters mis-segmentation, the misperception of hypernym/hyponym vowels and tone, and the confusion of similar Lao characters. According to the writing scheme and the hypo consonant of Lao characters, this paper proposes a multitasking character recognition to effectively integrate the structural features of Lao characters. The model extracts the structural features of Lao characters from character pictures via Deep Residual Network, and corrects the single character bounding box through Bounding Box Regression. Then, the corrected segmentation results and extracted character features are input jointly into Bi-directional Long-Short Term Memory network to identify the Lao character sequence, and the sequence alignment is completed by the Connectionist Temporal Classification. Finally, the result is predicted by the fixed combinatorial optimization model of Lao characters. The experimental result shows the method can reduce the Character Error Rate to 13.06%.

Keywords: Lao printed characters' recognition; Lao characters' structural features; multi-task recognition; end-to-end model

0 引言

老挝语是东南亚一种低资源语种，老挝字符识别是研究对获取的老挝单语图片文本进行准确识别的过程，以此方式扩充老挝语文本语料，为老挝语分词、词性标注、机器翻译等自然语言基础任务提供文本语料，还对研究东南亚低资源语言的文字识别问题起到推动作用。

在当前老挝语自然语言处理研究过程中，文本

收稿日期：2020-11-04 **定稿日期**：2020-11-19
基金项目：国家自然科学基金(61662040)

语料资源稀缺，通过爬虫等方式获取的文本语料多为老挝单语图片文本，且图片文本中文字难以直接使用，导致现有老挝语文本语料数量无法满足老挝语基础任务研究需求。文字识别是将图片转换为文本语料的一种方法，但与汉语[1]、日语[2]等其他语言相比，老挝语文字识别的研究寥寥无几，且词间无分割的老挝语由一个或多个字符构成，字符组合较为复杂。目前，老挝字符识别研究中主要存在的问题如下：

（1）字符切分是老挝语 OCR 研究的挑战，且切分结果对识别的准确性影响很大。老挝字符结构复杂，如图 1 所示，老挝字符图片中由于字体和文档质量等原因，存在字符粘连、断裂等情况，导致出现误切分的情况，图 1 中上面一行是字符正确情况，下面一行是字符切分错误情况；

（2）识别过程中对上、下位元音以及音调识别存在位置（相对于主辅音）偏差；

（3）老挝字符中存在很多相似字符，例如“ຣ-ຣ”“ພ-ຟ”“ມ-ນ”“ ◌ ◌ ”“ ◌ ◌ ”等，这些高度相似的字符在字符识别中存在很大障碍。

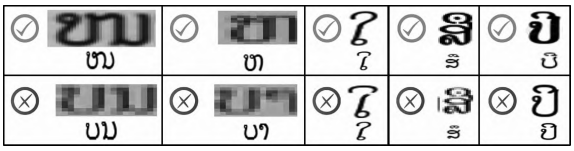


图 1　单字符切分图像对比

为此，如何构建老挝字符识别语料库以及准确识别所获取资源，使其在自然语言处理研究中成为提供基础语料的重要途径，成为亟待解决的问题。

其他语种的字符识别研究[1,3-6]，为提高训练速度和识别准确率，先对待识别图像进行二值化、去噪、旋转、增强等预处理操作，然后基于神经网络，并且结合相应语言特征或者统计分类等方法。Hossain 等人[6]通过预处理、分割、训练识别和后处理操作，由人工神经网络训练和识别孟加拉文字。He 等人[7]使用 ResNet 提取字符特征。Ling[6]以每个字符类型向量及其组合的固定词向量，通过双向 LSTM 神经网络，识别土耳其文本序列。Shi 等人[8]以端到端的可训练神经网络，识别图像中字符序列。Sarshoah 等人[9]以端到端的可训练多任务网络，从复杂文档中提取文本。

本文针对老挝字符结构特征，提出一种融合老挝字符结构特征的多任务字符识别方法，构建了结合字符识别损失和字符书写等级辅助损失的多任务老挝字符识别模型。通过多任务学习的方法，使字符识别主任务有效学习老挝字符书写等级分布表示[10]。首先，为解决上述老挝字符识别过程中存在的问题，在 ResNet 网络上加入 Dropout 防止过度拟合，并以多任务学习拟合老挝字符书写等级分布，利用标注语料训练模型提取老挝字符特征，获得老挝单字符切分结果，然后通过边框回归部分，矫正误切分字符；其次，以矫正字符切分结果和标注语料作为 BiLSTM（Bi-directional Long Short-Term Memory）输入，提取老挝字符图片中的特征向量，使字符识别任务得到更高级的老挝字符序列语义表示，并结合老挝字符下位辅音特征预测老挝字符序列，然后以 CTC（Connectionist Temparal Classification）进行序列对齐；最后模型根据老挝语语法规则做字符识别后处理，以修正模型误分类产生的错误结果，最终模型输出老挝文本。模型以端到端方式训练，即输入端为已预处理的老挝字符图片，输出端为模型预测的老挝字符序列，模型中间没有输出，以此减少模型复杂度且无需处理中间表示。

本文的贡献如下：

（1）针对老挝字符识别过程中存在的问题，本文提出一种融合字符结构特征的多任务老挝字符识别方法，解决单字符误切分问题以及识别老挝字符序列问题。

（2）端到端训练模型，旨在同时切分老挝单字符和识别字符序列，减少训练时长。

本文组织结构如下：第 1 节介绍印刷体文字识别研究的相关工作；第 2 节分析老挝字符结构的特征；第 3 节介绍本文提出的老挝字符识别方法；第 4 节对实验结果进行相关分析；最后，对本文所做工作进行总结，并给出下一步研究内容。

1　相关工作

目前有很多方法应用于印刷体字符识别研究，其中一些采用字符分割的方法来提取字符，并使用辅助识别器进行识别；还有通过滑动窗口的方法，以规则的间隔进行字符切分，然后由分类器确定字符边界，预测最佳字符序列。故根据模型策略可将字符识别方法分为：基于统计的方法和基于神经网络的识别方法[11]。

1.1　基于统计的方法

基于统计的方法实质上将字符识别问题看作分

类问题，首先以检测技术获取字符轮廓，然后进行版面分析定位文本区域，最后进行字符切分和识别。在以往研究中，Drup 等人[12]对印刷体藏文字符特征深入研究，对藏文字符进行局部自适应二值化处理、连通域分割和基于网格的模糊笔划特征提取，来识别藏文。王华等人[13]中采用基于置信度分析的粗、细分类两极分类策略对藏文字符进行类别判断，但模型对相似字符的分类效果欠佳。Yamina[14]提出了一种基于轮廓投影的分割方法，通过提取 56 个统计特征组合，输入 SVM(Support Vector Machine)对阿拉伯字符进行分类识别。

基于统计的字符识别方法是将字符识别问题分治为预处理、特征提取，最后通过统计特征作为特征向量，由分类器进行分类识别。其优点在于，分类器构造过程中无需语言学专家进行语言规则分析；缺点是将问题分治后，容易导致局部最优陷阱。此外，基于统计的字符识别方法对于相似字符的区别能力较弱，因为在老挝字符中，存在很多相似字符，所以难以解决老挝字符识别研究中遇到的问题。

1.2 基于神经网络的识别方法

基于神经网络的识别方法，是将待识别图像输入到神经网络中，学习直至识别的过程。在以往的工作中，Tesseract[15]是将待识别图像转换为可检索数据的一种开源 OCR(Optical Character Recognition)引擎，可以支持 100 种语言集字符识别①，但在老挝语字符识别中，因为老挝语是一种词间无分割语言，故对字符的音调、上位元音以及下位元音等级的字符位置识别存在一定偏差。Shi[5]通过对 BP(Back Propagation)神经网络添加 Momentum 梯度下降，从而提高训练速度，由于 BP 神经网络学习速率固定、隐藏层数选择无理论指导等因素，会出现训练时间过长以及权值收敛缓慢问题。DCNN[16-17](Deep Convolutional Neural Network)通常对定长序列进行输入和输出操作，但在类似老挝语等语种的不定长序列识别上，效果欠佳，故其无法直接应用于老挝语字符识别研究。Ait-Mohard 等人[11]，Ahmad[18]构建 HMM(Hidden Markov Model)字符识别模型，进行多字体印刷字符识别。Wang[19]通过无监督预训练，利用 CNN 进行端到端的文本检测和字符识别。Shi[8]将 DCNN 和 RNN 结合，构建端到端序列识别系统，解决了序列长度约束的问题。文献[20]中使用注意力机制进行序列到序列的学习来识别文本。He[7]在残差网络中使用 Identify mapping 使网络更容易学习，于 2015 年 ILSVRC 挑战赛中，在 ImageNet 和 COCO 数据集的检测、分割、识别任务上夺冠，且所使用模型性能优于 VGG-16(Oxford Visual Geometry Group)。Szegedy[21]在 Inception-V4 网络中加入残差学习，在 ImageNet 分类挑战测试集上达到 3.08%的 top-5 错误率。Ly 等人[22]使用基于注意力的端到端模型，识别日语历史图片文本。Lyu 等人[23]基于 Mask R-CNN 提出 Mask TextSpotter，通过端到端可训练的深度神经网络完成语义分段获得精确的文本检测和识别。Yu 等人[24]提出一种 PICK 框架，结合图学习与图卷积操作，有效抽取票据关键信息，产生丰富的文档语义表示，进行字符识别。

2 老挝字符结构分析

老挝文档依赖于 ISO/IEC 10646 字符集标准，其中老挝字符由 55 个字符和 10 个老挝数字组成，本文识别老挝语的基本单元为字符，老挝字符表如表 1 所示，其中元音字母可以在辅音字母的前后上下出现，构成复合元音。老挝语标点符号使用和汉语类似，有逗号、句号、叹号和问号等。

表 1 老挝语字符表

Symbols	Detail
Consonant	ກ ຈ ດ ຕ ບ ປ ຍ ອ ຄ ຊ ທ ພ ຟ ຣ ນ ມ ລ ວ ງ ຢ ຮ ຂ ສ ຖ ຜ ຝ ຫ ໜ ໝ
Vowel	ເ ະ າ ິ ີ ຶ ື ຸ ູ ໍ ໄ ໃ ໂ ຳ ແ ົ ັ ຽ
Tone	່ ້ ໊ ໋
Special Symbols	ໆ(Repetitive) ຯ(Omit) ຼ(Semivowel) ໌(Cancel)
Numeral	໐ ໑ ໒ ໓ ໔ ໕ ໖ ໗ ໘ ໙

老挝字符书写结构复杂，故在字符切分及识别部分中，存在一定误差，为解决上述问题，本文提出可将老挝字符书写规则分为 4 个等级：声调等级，上位元音等级，辅音(部分元音)等级和下位元音(辅音)等级，老挝字符书写规则如图 2 所示。另外，在老挝语中，下位书写等级中一般都是元音，而在下位书写等级中为辅音的只有一个特殊情况，即ຫລ(要么)可以写成ຫຼ(要么)。因此，本文以参数形式拟合老挝字符书写等级分布规则以及下位辅音特征作为辅助任务，在字符识别模型中与主任务共享，更好地

① https://www.primaresearch.org/das2016/tutorials

学习老挝字符结构特征，以多任务学习方法提升字符识别的准确率。

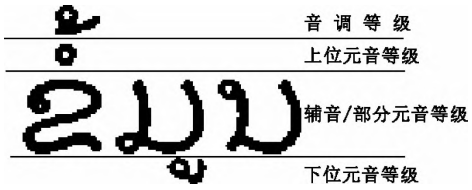


图 2　老挝语书写等级

3　多任务字符识别模型

本文以无分割老挝印刷体字符识别为背景，将其识别为可编辑文本，用作自然语言处理研究基础语料。以端到端方式训练网络，首先以 ResNet 网络作为共享特征提取层，特征共享于老挝字单字符切分和识别老挝字符序列任务。在老挝单字符切分中，以提取的共享特征进行老挝单字符切分，然后以单字符包围框的中心坐标、宽和高作为老挝单字符包围框的坐标，利用边框回归对老挝字符切分所产生的误切分单字符包围框进行矫正。其次在识别老挝字符序列任务中，针对相似老挝字符和识别结果存在误差的问题，使用 BiLSTM[25] 网络学习字符序列上下文信息，并利用 CTC 对模型预测老挝序列进行对齐。最后在字符后处理阶段，根据老挝语语法规则，对模型预测结果进行基本的拼写检查。

在模型中，融合老挝字符结构特征，进行多任务学习。具体将老挝字符书写等级特征和下位辅音特征作为辅助损失，模型损失值是将老挝字符识别模型损失值和老挝字符结构辅助损失值相加。给定一个老挝字符图片“ลาอมักแฮ่น(他喜欢跑步)”，老挝语文字识别模型图如图 3 所示。

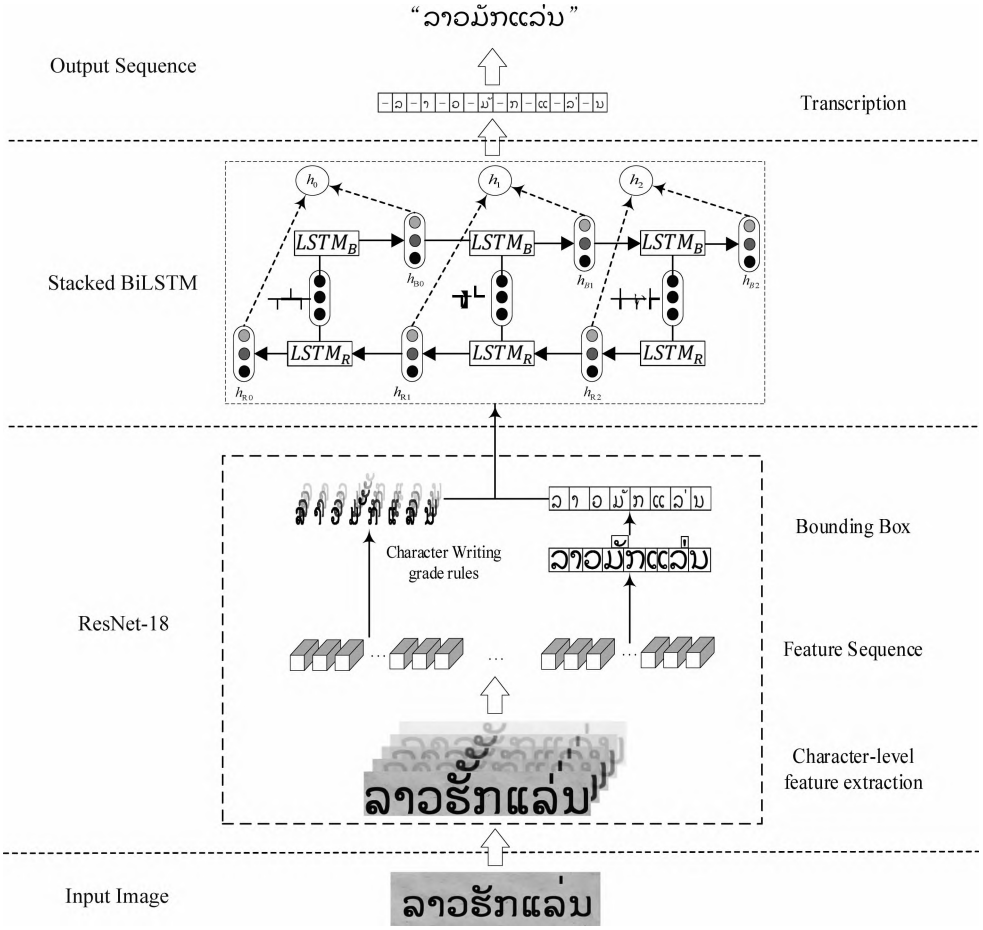


图 3　老挝语文字识别模型图

3.1　识别流程概述

在老挝字符识别所处理语料中，每句老挝文字中字符数量不固定，故视为不定长老挝字符识别过程，即以老挝文档作为输入，经过老挝字符切分、识别老挝字符和识别结果后处理，输出老挝字符序列。

具体识别流程如图 4 所示。

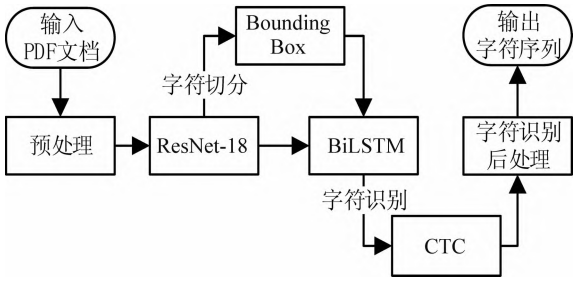
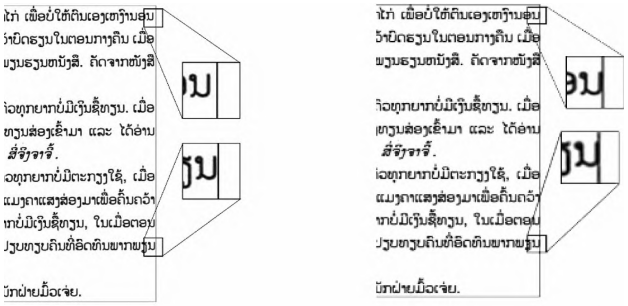


图 4 老挝语文字识别流程

3.2 预处理

本文所处理语料为从左至右书写的横排元音附标老挝文字，输入对象为印刷体老挝单语图片文本。对获取的老挝单语图片文本，首先针对老挝字符图像中存在噪点、歪斜的情况(页面细节如图 5(a)所示)，将老挝字符图片进行灰度化处理，然后根据局部自适应二值化方法[12]对灰度化的图片文本进行二值化处理，转为二值图像，接着根据式(1)及式(2)，对图片文本采取形态学开操作，先腐蚀[式(1)]后膨胀[式(2)]，消除字符图片中的小噪点，其中 z 为对 $(B)z$ 做平移操作，所有满足包含于 A



(a) 轻微偏斜的原始图像 (b) 旋转矫正后的图像

图 5 输入图像旋转矫正前后的对比

的点，A、B 为字符图片中的字符和小噪点。

$$A \odot B = \{z \mid (B)z \subseteq A\} \tag{1}$$

$$A \oplus B = \{z \mid (B \wedge)z \cap A \neq \varnothing\} \tag{2}$$

最后对图片文本利用 Hough 直线检测[26]，解决语料中老挝字符图像歪斜问题。具体以老挝字符图像中每个字符作为一系列的点，然后由式(3)将每行字符点构成的集合$\{(x_1,y_1),\cdots,(x_n,y_n)\}$转化为霍夫空间坐标$\{(r_1,\theta_1),\cdots,(r_n,\theta_n)\}$，Hough 参数空间的峰值点即为字符图像中每行文字构成的直线。其中 r，θ 是常量，角度 θ 指 r 与 x 轴之间的夹角，r 为点到直线的几何垂直距离，处理完成后图片文本细节如图 5(b)所示。

$$r = x_i\cos\theta + y_i\sin\theta \tag{3}$$

3.3 老挝单字符切分

字符切分基于 ResNet[7]构建，ResNet 是依照 VGG-style 进行的网络结构设计。使用 ResNet-18 作为共享特征提取层，如式(4)所示，$\boldsymbol{W}_i$ 为 ResNet 参数，以预处理所得图片文本 $\boldsymbol{x}_l$ 作为 ResNet 输入，在 Feature map 向下传递过程中，$F_l(\boldsymbol{x}_l,\boldsymbol{W}_i)$通过执行卷积、BN 等操作以获得固定维度输出，加入恒等映射 x_l，即为 Shortcut 连接，将图片文本中字符序列的高阶特征和低阶特征融合，$\boldsymbol{x}_{l+1}$为输出向量，$F_l(\boldsymbol{x}_l,\boldsymbol{W}_i)$为待学习残差映射，然后如式(5)所示，以 ReLU 激活函数进行线性映射，达到维度匹配的目的，以此解决图片文本中相似字符的局部差异特征，避免 Feature map 传递时信息丢失、损耗问题，其中 σ 表示 ReLU。另外，针对测试语料中多行老挝语图片文本，在 ResNet 网络中的 ConV3、ConV4、ConV5，执行放大操作，将输入的多行图片文本在垂直方向的像素上隐式展开为单行图像，得到线状 Feature map 作为老挝文字识别输入。

$$\boldsymbol{x}_{l+1} = \boldsymbol{x}_l + F_l(\boldsymbol{x}_l,\boldsymbol{W}_i) \tag{4}$$

$$F = W_2\sigma(W_1x) \tag{5}$$

本文针对老挝字符结构特征，通过拟合老挝字符书写等级规则作为辅助任务，进行多任务学习，以更好地表征老挝字符结构分布，然后以切分所得的老挝单字符包围框对应宽 w、高 h 和中心点坐标 (x_c,y_c)，通过边框回归进行老挝单字符包围框的矫正，以此解决老挝单字符存在误切分问题。

3.4 识别老挝字符序列

不定长老挝字符序列识别问题，需要单字符左右的语义信息，因此识别老挝字符序列基于栈式双向 LSTM 构建，提取图片文本中老挝字符级特征向量，得到更高级的老挝文字语义信息表示。模型训练时将老挝字符序列进行预处理，并经过老挝字符切分，得到已切分的单行老挝字符图片、ResNet 网络提取的老挝字符级特征以及分词后的老挝句子，并将其作为 BiLSTM 的输入。特别地，模型预测时没有输入上述标注信息。每个 LSTM 的输入值包括：当前的老挝字符级分布表示、前一个 LSTM 的状态值和输出值。

本文使用的栈式 BiLSTM 输出老挝字符序列特征向量为 $\{\boldsymbol{h}_0,\boldsymbol{h}_1,\boldsymbol{h}_2\}$，将老挝字符序列

“ລາວມັກແລ່ນ(他爱跑步)”的每个字符首先输入前向LSTM$_B$“ລາວ(他)”,“ຮັກ(爱)”“ແລ່ນ(跑步)”得到三个前向向量$\{\boldsymbol{h}_{B0},\boldsymbol{h}_{B1},\boldsymbol{h}_{B2}\}$。然后在后向LSTM$_R$依次输入“ແລ່ນ(跑步)”“ຮັກ(爱)”“ລາວ(我)”得到三个后向向量$\{\boldsymbol{h}_{R0},\boldsymbol{h}_{R1},\boldsymbol{h}_{R2}\}$。最后进行计算并拼接前向、后向向量得到对应输出向量。以此捕捉老挝字符序列的双向语义,并结合老挝字符结构中下位辅音特征,以多任务学习方式,辅助BiLSTM预测老挝字符序列,且在训练过程中,利用CTC损失函数实现输入的老挝字符特征序列与模型预测结果的对齐。

3.5　多任务学习

ResNet网络特征共享提取层输出值一方面分割老挝单字符,并对其结果利用边框回归矫正误切分字符,另一方面预测老挝字符书写等级,两个任务共享特征提取层的参数,因此字符书写等级辅助任务学习到的字符结构特征会以参数的形式与字符识别任务共享,加强对老挝字符结构分布的表征,提高老挝语文字识别模型效果。

在老挝单字符切分任务中,以上述多任务学习方式表征老挝字符结构特征,减少老挝单字符切分的错误率,而且在模型中,还利用边框回归对共享特征提取层输出的老挝单字符包围框P做平移和缩放处理,以获得精确老挝单字符包围框G。边框回归使用的损失函数为Smooth $L1$,如式(6)所示,其中x为真实单字符包围框与预测单字符包围框之差。

$$\text{Smooth } L1=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & x<-1 \text{ or } x>1\end{cases} \tag{6}$$

在识别老挝字符序列任务中,利用CTC loss解决输入端老挝图片文本与对应文本输出的对齐问题[27]。即在BiLSTM输入$\boldsymbol{X}=$["ແປ(翻译)"],以Beam Search方式为每个字符矩阵分数进行解码,并计算损失值以训练神经网络,找到路径中概率最大的老挝字符序列,即为其对应的最优序列$\boldsymbol{Y}=$[ແ,ປ],以此找到输入已预测的老挝序列$\boldsymbol{X}$与输出对齐后的老挝字符序列$\boldsymbol{Y}$之间精确的映射关系。

本文以多任务学习方式进行端到端老挝字符识别任务,模型的损失值是字符识别模型损失和老挝字符结构辅助损失之和,字符结构辅助任务以Softmax函数预测字符结构标签,并用交叉熵损失函数计算损失值,字符结构辅助损失如式(7)所示,其中$\boldsymbol{h}_i$为BiLSTM输出特征向量,CS代表字符结构标签;如式(8)所示,其中平衡因子$\alpha_1=1,\alpha_2=1$,L_{Bbox}为边框回归部分损失值,L_{CTC}为CTC部分损失值。

$$L_{\text{CS}}=-\sum_{i=1}^{n}\log(\text{softmax}(\text{CS}_i,h_i)) \tag{7}$$

$$L=L_{\text{CS}}+\alpha_1 L_{\text{Bbox}}+\alpha_2 L_{\text{CTC}} \tag{8}$$

3.6　识别结果后处理

老挝语以音节为基本构词单位,音节中包含一个或多个字符,其中音节内存在大量固定组合,如多个元音字母构成的元音及多个元音字母构成的复合元音(见表2)等[28],以及部分元音及音调字母位于辅音字母上方或下方,在老挝字符切分中会出现元音或音调单独切出成行的情况。

表2　多个元音字母构成元音表(○为辅音字母)

单元音	复合元音	特殊元音
○ະ	○ັຍ	○ໍາ
○○ະ	○ຍ	—
○ະ	○ັອະ	—
○າະ	○ັອ	—
○ັ	○ັອ	—
○ັ	○ັອ	—

字符识别输出预测序列后,对文本进行扫描,利用老挝语语言规则在字符序列后处理阶段执行基本拼写检查,修正消除模型误分类产生的错误结果,输出最终识别的老挝文本。因此按照音节最大切分方法对模型预测老挝文本进行扫描,当音节内出现错误搭配或过切分结果时,根据语法对其进行修正,如图1(b)所示,如出现“ແ(元音字母)”,应修正为一个“ແ”字符。此外,当“ໍ”和“າ”连续出现时,应修正为特殊元音“ໍາ”字符。

4　实验及分析

4.1　语料库构建及预处理

在老挝字符识别研究中,本文构建的语料库包括三部分:

(1) 利用实验室已有老挝语句子生成的老挝字符图片(5～15个老挝字符),以保证生成老挝语料真实性。首先利用PIL三方库,生成32×280固定大小的背景图片;其次,为已有的老挝语词典生成的

老挝句子选定字体种类、大小和颜色，然后随机选取文字贴合背景图片的坐标，将老挝句子贴到背景图片中；最后，进行图片模糊处理，并保存文本信息和对应图片名称，得到20 000张老挝图片文本。

(2) 利用爬虫获取的PDF老挝单语图片文本，通过在CRI国际在线①、老挝通讯社②、老挝人民政府③等网站爬虫获取，其中PDF文档中包含多种字体和各种背景。文档切分并去除空白页后约有1 688张老挝图片文本，平均每张图片包含14句长短不一的老挝句子，共计23 632句老挝单语文字。

(3) 网页中真实场景下和PDF文档中包含老挝图片文本，如图6所示(语料中场景图片几乎都包含人物在内，故截取仅含老挝字符部分图片展示)，该部分语料场景较为复杂，图片或文档质量偏低，且存在字体种类繁多、字体大小不一等情况，增加了模型识别的难度。共计1 022张图片。为有效抽取老挝字符结构特征，分别将本次语料中后两部分图片以9∶1的比例划分为训练和测试集，且将第一部分老挝图片文本扩充到训练集中。

图6 截取部分场景文本图片示例

为进行老挝字符识别，需对语料库中的数据进行如下预处理操作：

(1) 针对第一部分语料图片，通过选择不同字体(见表4)、背景以及字符在图片中的位置等，增加训练集多样性。

(2) 针对第一部分语料中对应老挝图片文本信息，利用实验室开发的LaoWS分词工具[32]进行分词处理，进行模型训练。

(3) 解析并识别PDF文档中文本块和图片块对象，进行文档分页，删除空白图片页并将PDF文档中图片页单独保存到文件夹，以合并到第三部分语料。

(4) 对已分页的少量PDF文档，利用PIL模块调整色彩饱和度、色相等颜色模型，以增强图像清晰度。

(5) 在训练和测试语料中，通过随机添加模糊函数(深度平滑滤波、高斯模糊、平滑滤波)、阴影以及椒盐噪声等方式，提高老挝文字图像复杂性。

(6)二值化图像处理以及旋转矫正处理。

4.2 模型参数测试

本在NVIDIA Geforce GTX 1050Ti硬件条件下，采用Keras深度学习框架，基于老挝字符识别语料库，进行模型训练和测试。经过实验调参，模型训练参数设置如表3所示。

表3 模型参数设置

Parameter	Value
Convolution kernel	3×3
Batch size	256
Dropout	0.5
BiLSTM layer	2
Learning rate	0.000 1

其中，Convolution kernel是ResNet-18中ConV2D的卷积核大小，Batch size为共享特征层批尺寸，Dropout为共享特征提取层ConV3，ConV4和ConV5中用以防止过拟合；BiLSTM和N_hidden分别为BiLSTM层数以及隐藏神经元个数，在栈式BiLSTM中使用Dynamic_rnn跳过Padding部分的计算，以减少计算量；Learning Rate为初始学习率。

4.3 实验结果分析

本文使用方法在300 Dpi的老挝印刷图片文本上进行测试，单字符识别时间0.005 8 s，平均每张包含1 000字左右的图片文本识别时间为10.29 s。通过设计以下三组实验来验证本文使用老挝字符识别模型的有效性：①以不同训练集训练并测试模型识别效果；②端到端老挝字符识别模型性能实验；③以不同印刷字符识别方法，验证模型效果。在本实验相同语料下，实验结果表明，本文使用的老挝字符识别方法优于其他模型，CER为13.06%。具体地CER(Character Error Rate)评价指标，以字符串编辑距离实现。如式(9)所示，CER的值越低模型效果越好，C_S为替换的字数，C_D为删除的字数，C_I为插入的字数，C_H为正确的字数，N为$C_S+C_D+C_H$字数。

① China Radio International: http://laos.cri.cn/

② Khaosan Pathet Lao: http://kpl.gov.la/

③ Lao People's Government Network: http://www.na.gov.la/

$$\mathrm{CER}=\frac{C_S+C_D+C_I}{N}=\frac{C_S+C_D+C_I}{C_S+C_D+C_H} \quad (9)$$

4.3.1　不同训练集对比实验

在本次实验中，由于后两部分语料中字体种类繁多，为模型识别带来一定阻碍。因此，与 Sabir 等人[29]的方法类似，本文使用不同字体测试模型效果。在训练集中后两部分语料不变的情况下，对第一部分语料中，以下述四种老挝字体分别生成训练集：CoreSansLao Light、Lao Muang Don Regular、Noto Serif Lao Thin、Souliyo Unicode Regular，具体设置如下：单独 1 种字体、随机混合 2 种字体、随机混合 3 种字体和 4 种老挝字体。在本次实验语料库下进行模型训练，以增强老挝字符识别模型的学习能力。图 7 是使用以上四种训练集训练模型时对应的实验结果。

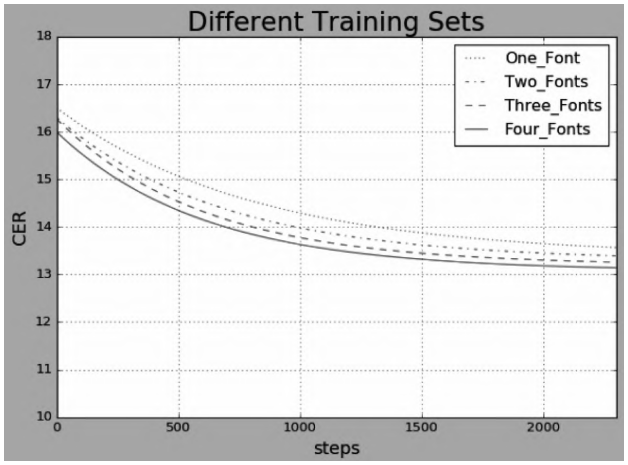


图 7　不同训练集性能实验

在不同训练集验证模型性能实验中，如表 4 所示，首先对语料中第一部分老挝字符图片进行单独 1 种字体对模型结果影响实验，其中四类字体样式以“ຊ່ວຍເຫຼືອ(帮助)”为例，结果表明不同字体的模型测试结果存在差异，如表 5 所示，CER 最高的字体 CoreSansLao Light，为 13.53%，因该字体比较潦草，对老挝字符中上/下位元音区分度较差，导致 CER 较高；Lao Muang Don Regular 字体较粗，导致部分相似字符的辨识度不高，且对老挝语上/下位元音字符在识别过程中导致存在偏差(相对于主辅音)，但相比 CoreSansLao Light 字体 CER 低了 0.28%；Noto Serif Lao Thin 因其字重较薄，且在部分 PDF 文档中由于质量原因导致字符断裂情况，致使 CER 指标相比 Souliyo Unicode Regular 字体略高 0.09%；Souliyo Unicode Regular 字体，因形态规范且边界清晰，故其 CER 最低，为 13.06%，但是由于该类字体字重较粗，在 PDF 文档中部分字符存在粘连情况导致上/下位元音字符和辅音字符部分重叠，导致有些相似字符区分度较低。然后如图 6 所示，分别使用四种不同训练集，进行不同训练集对模型性能影响实验，结果显示使用随机 1 种字体生成训练集，相比于同时四种字体生成训练集，模型结果显示，从随机一种字体的平均 CER 13.30%降至同时使用四种字体训练模型的 CER 13.06%，且模型收敛速度加快。结果表明同时使用四种字体训练模型有明显效果。

表 4　四种不同字体对模型结果的影响

Fonts	Style	CER/%
CoreSansLao Light	ຊ່ວຍເຫຼືອ	13.64
LaoMuang Don Regular	ຊ່ວຍເຫຼືອ	13.36
Noto Serif Lao Thin	ຊ່ວຍເຫຼືອ	13.15
Souliyo Unicode Regular	ຊ່ວຍເຫຼືອ	13.06

表 5　四类字体识别偏差对比

Fonts	Style	Correct	Error
CoreSansLao Light	ສົ ປ ັ ໄ	ສົ ປ ັ ໄ	ສົ ປ ັ ໄ
Lao Muang Don Regular	່ ່	່ ່	່ ່
Noto Serif Lao Thin	ໄ ໆ ໄ	ໆ ໆ	ໄ ໆ ໄ
Souliyo Unicode Regular	◌ ◌ ◌	◌ ◌	Unknown Character ສົ ປ ັ ໄ

4.3.2　端到端老挝字符识别模型性能实验

为探索不同特征提取器对老挝字符结构特征的分布表示、边框回归对模型性能的影响、BiLSTM 对字符识别性能的影响以及多任务学习的有效性，本文设计如下几个模型：①不同特征提取器；②不使用边框回归；③不进行序列对齐；④不使用多任务学习。5 个模型在相同语料上进行实验对比，结果如图 8 和表 6 所示。

表 6　设计不同实验对比

Model	①	②	③	④	⑤
Bounding Box	√	√	×	√	√
CTC	√	√	√	√	×
Multitask	√	×	√	√	√
Post-Processing	√	√	√	×	√
CER/%	13.06	14.14	14.18	13.6	13.48

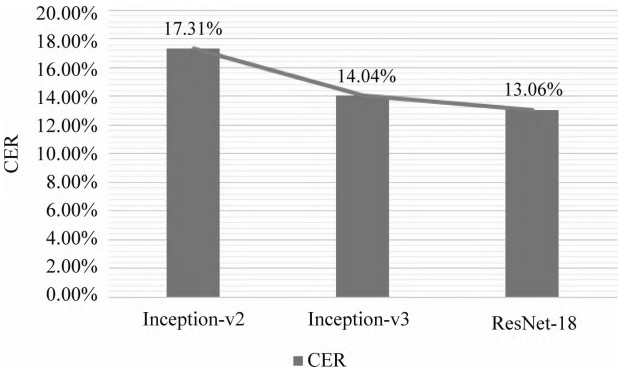


图8 不同特征提取器性能实验

本文针对老挝字符书写规则分布表示，通过比对3种基于CNN的不同特征提取器(即Inception-v2、Inception-v3和ResNet-18)对模型性能的影响。如图7所示，基于Inception-v2的CER高达16.10%，ResNet-18相比于Inception-v3的CER下降了0.51%，结果表明ResNet-18对老挝字符书写规则分布表示的性能优于其他两者。

实验③相比于实验①的CER增加了1.12%，具体表现在语料库中由于字重较粗产生的字符粘连情况以及场景文本下图片质量等原因，出现字符误切分情况严重，体现了边框回归部分对老挝单字符切分的重要性；实验⑤的CER比实验①增加了0.42%，表明在识别老挝字符序列任务中，利用CTC可对模型效果有一定程度提升；实验①比实验②的CER减少了0.95%，验证多任务学习可使模型有效学习老挝字符书写分布表示；实验①比实验④的CER减少了0.51%，证实老挝语语言规则在文字后处理阶段实验有效性，提高模型准确率。在本次实验使用语料下，经过上述模型实验对比，说明本模型在老挝字符识别研究上的有效性。

4.3.3 不同印刷体字符识别方法对比实验

在印刷体字符识别方法对比试验中，使用相同语料，分别与基于统计方法、基于神经网络方法进行比较。基于传统方法分别为：①SVM模型[14]：提取字符统计特征组合，以SVM进行字符分类；②KNN模型[30]：从图片文本提取笔划和字符轮廓，利用欧氏距离并通过KNN分类器进行印刷字符分类。基于神经网络的识别方法分别为：③Tesseract模型[17]；④HMM模型[11][18]，通过改进自适应HMM进行印刷字符识别；⑤BP模型[1]：以BP神经网络结合统计分类进行老挝印刷字符识别；⑥CRNN模型[8]，以CNN提取老挝字符特征，利用RNN进行字符识别；⑦Deep BLSTM模型[22]，以Deep BLSTM网络，并使用CTC进行训练，以识别Oriya印刷体未分词文本；⑧DenseNet模型[31]，网络以前馈方式将每一层连接到其他每一层，且所有层输出Feature map都作为后续层的输入；⑨Text-CNN模型[32]，利用多层次和丰富的监督信息识别文本区域并进行文本特征计算，且增加低层监督信息，使模型具有更好判别能力。实验结果如表7所示。

表7 不同印刷体字符识别方法对比实验

	Methods	CER/%
Statistical Methods	SVM	37.08
	KNN	28.75
Neural Network Methods	Tesseract	54.98
	HMM	40.82
	BP	28.50
	Deep BLSTM	23.64
	DenseNet	20.21
	Text-CNN	18.60
	CRNN	17.38
	Ours	**13.06**

由表7可知，在本次实验数据下，SVM模型的CER高达37.08%，因其无法准确提取老挝字符结构特征，故在老挝字符识别上，因对相似字符区分度较低，导致效果略差。此外还与基于统计的字符识别方法KNN进行对比，CER指标比SVM模型低8.33%，因语料库中图片皆为印刷体文字，故对字符的笔划和轮廓特征容易提取，故在分类过程中效果较好。Tesseract模型的CER高达54.98%，在老挝字符识别研究上效果不佳，但Tesseract在英语等语种的印刷体文字识别效果不容小觑。BP模型相比HMM模型的CER降低了12.32%，虽然在老挝字符识别上效果不错，但是在识别过程中没有自主学习字符特征分布表示的能力，且参数较多，不易收敛。Deep BLSTM模型在老挝文字识别中，误差主要出现在上下位元音修饰的主辅音存在位置偏差，及对部分老挝固定组合识别错误。DenseNet模型相比于本文所提出方法CER高7.15%，因为老挝语为词间无分隔语种，且字符结构复杂，导致部分包含元音及音调字符识别存在误差。Text-CNN模型在老挝文字识别中CER低至18.60%，但其在相似字符区分方面效果一般。本文模型在CRNN模型的基础上融合老挝字符结构特征，并利用多任务学习，

对老挝字符图像进行端到端训练。在老挝字符识别研究上，相比于上述印刷文字识别方法，其在字符错误率和识别速度方面均有一定提升，在老挝语光学字符识别方面，具有一定的创新性意义。

5　结束语

针对老挝语文本语料稀缺，目前老挝语文字识别方法效果较差，本文提出了一种融合老挝字符结构特征的多任务老挝字符识别方法，在本实验所构建语料库上，本文的方法达到了 86.94％的准确率，在老挝字符识别任务中效果优于当前其他主流方法。下一步将对模型进一步优化，使用 Attention 机制识别老挝字符序列，同时在字符后处理研究中利用老挝语语法、词法等特征，研究词法拼写检查机制。

参考文献

[1] HU X, PENG J, WANG, M, et al. A printed Chinese character recognition method[C]//Proceedings of the International Conference on Computer Science and Service System, 2011: 2904-2907.

[2] RUSU A, GOVINDARAJU V. CAPTCHA: Using the difference in the abilities of humans and machines in reading handwritten words[C]//Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition -Handwritten,2004: 226-231.

[3] SANKARAN N, JAWAHAR C V. Recognition of printed devanagari text using BLSTM neural network [C]//Proceedings of the 21st International Conference on Pattern Recognition. IEEE, 2012: 322-325.

[4] LING W, LUÍS T, MARUJO L, et al. Finding function in form: Compositional character models for open vocabulary word representation[C]//Proceedings of the EMNLP, 2015:1899-1907.

[5] SHI Y, FAN W, SHI G, The research of printed character recognition based on neural network[C]// Proceedings of the 4th International Symposium on Parallel Architectures, Algorithms and Programming, Tianjin, 2011.

[6] HOSSAIN S K A, TABASSUM T. Neural net based complete character recognition scheme for Bangla printed text books[C]//Proceedings of the 16th Int'l Conf. Computer and Information Technology. IEEE, 2014: 71-75.

[7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 770-778.

[8] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11):2298-2304.

[9] SARSHOGH M R, HINES K. A multi-task network for localization and recognition of text in images[C]// Proceedings of the International Conference on Document Analysis and Recognition. IEEE, 2019: 494-501.

[10] PLANK B,SØGAARD A, GOLDBERG Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016: 412-418.

[11] AIT-MOHAND K, PAQUET T, RAGOT N. Combining structure and parameter adaptation of HMMs for printed text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36 (9): 1716-1732.

[12] DRUP N, ZHAO D, REN P, et al. Study on Printed Tibetan Character Recognition[C]//Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence, 2010: 280-285.

[13] 王华，丁晓青. 多字体印刷藏文字符识别[J]. 中文信息学报，2003，17(6)：48-53.

[14] YAMINA O J, El Mamoun M, KADDOUR S. Printed Arabic optical character recognition using support vector machine[C]//Proceedings of the International Conference on Mathematics and Information Technology. IEEE, 2017: 134-140.

[15] SMITH R,ANTONOVA D, LEE D S. Adapting the tesseract open source OCR engine for multilingual OCR[C]//Proceedings of the International Workshop on Multilingual OCR. 2009: 1-8.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communication of the ACN, 2017, 60 (6): 84-90.

[17] LECUN Y, BOTTOU L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[18] IRFAN B, AHMAD A, S A M A, G A F B. Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models[J]. Pattern Recognition, 2016(51): 97-111.

[19] WANG T, WU D J, COATES A, et al. End-to-end

text recognition with convolutional neural networks [C]//Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, 2012: 3304-3308.

[20] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4168-4176.

[21] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-V4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelgence, 2017: 4278-4284.

[22] LY N T, NGUYEN C T, NAKAGAWA M. An attention-based end-to-end model for multiple text lines recognition in japanese historical documents[C]// Proceedings of the International Conference on Document Analysis and Recognition. IEEE, 2019: 629-634.

[23] LIAO M, LYU P, HE M, et al. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43 (2): 532-548.

[24] YU W, LU N, QI X, et al. PICK: Processing key Information extraction from documents using improved graph learning-convolutional networks[C]// Proceedings of the 25th International Conference on Pattern Recognition, 2020: 4363-4369.

[25] RAY A, RAJESWAR S, CHAUDHURY S. Text recognition using deep BLSTM networks[C]//Proceedings of the 8th International Conference on Advances in Pattern Recognition. IEEE, 2015: 1-6.

[26] AGGARWAL N, KARL W. C. Line detection in images through regularizedhough transform[J]. IEEE Transactions on image processing, 2005, 15(3): 582-591. March 2006, doi: 10.1109/TIP.2005.863021.

[27] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning, 2006: 369-376.

[28] ZHANG L M. Practical grammar of Lao language [M]. Foreign Language Teaching and Research Press, 2001.

[29] SABIR E, RAWLS S, NATARAJAN P. Implicit language model in LSTM for OCR[C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. IEEE, 2017, 7: 27-31.

[30] BV D, RG B, HANGARGE M. Printed and Handwritten kannada numerals recognitionusing directional stroke and directional density with KNN[J]. International Journal of Machine Intelligence, 2011, 3(3): 121-125.

[31] HUANG G, LIU Z, VAN DERMAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[32] HE T, HUANG W, QIAO Y, et al. Text-attentional convolutional neural network for scene text detection [J]. IEEE Transactions on Image Processing, 2016, 25(6): 2529-2541.

[33] 何力,周兰江,周枫等. 基于双向长短期记忆神经网络的老挝语分词方法[J]. 计算机工程与科学,2019,41(07): 1312-1317.

陈琢(1995—),硕士,主要研究领域为文字识别。
E-mail: 1178988731@qq.com

郝永彬(1994—),硕士,主要研究领域为自然语言处理。
E-mail: 845415849@qq.com

周兰江(1964—),通信作者,硕士,副教授,主要研究领域为信息检索、机器学习和自然语言处理。
E-mail: 915090822@qq.com