基于多特征 Bi-LSTM-CRF 的影评人名识别研究*

禤镇宇[1]，蒋盛益[1,2]，张礼明[1]，包睿[1]

(1. 广东外语外贸大学 信息科学与技术学院，广东 广州 510006;

2. 广东省网络空间内容安全工程技术研究中心，广东 广州 510006）

摘要：近年来，电影行业蓬勃发展，相关的信息抽取和情报分析技术日益受到行业内的重视，其中对电影主创人物的分析尤为重要。而电影评论作为观影群体的主要反馈信息，具有重要的分析价值。如何从影评中自动抽取主创人名成为重要的基础工作。然而评论中观众对人物的称谓方式多样复杂，而且新电影的影评中往往存在大量人名未登录词，传统方法难以有效识别。针对影评的这些特点，本文提出一种基于多特征 Bi-LSTM-CRF 的影评人名识别方法。该方法通过利用外部人名语料和未标注影评提取字符级的特征；并采用 Bi-LSTM-CRF 模型进行人名字符序列标注。实验结果表明，该方法能够有效识别影评中的复杂称谓和人名未登录词，从而有效地抽取影评中的人名实体。

关键词：影评；LSTM；CRF；多特征；人名识别

中图分类号：TP391 文献标识码：A

Research on Person Name Recognition for Movie Reviews Based on Multi-Feature Bi-LSTM-CRF Model

XUAN Zhenyu[1]，JIANG Shenyu[1,2]，Zhang Liming[1]，BAO Rui[1]

(1. School of Information Science and Technology,Guangdong University of Foreign Studies, Guangzhou,Guangdong,510006,China;

2. Engineering Research Center for Cyberspace Content Security of Guangdong Province,Guangzhou,Guangdong,510006,China）

Abstract：With the rapid development of the movie industry in recent years, information extraction and intelligence analysis technologies have received more and more attention, especially for the intelligence analysis of the movie main characters. As the important feedback from the audience, movie reviews are of great value for intelligent reports, and there is a lot of information concerning movie characters.Therefore, how to automatically extract the related person name from the movie reviews has become the industry's needs. However, names of characters in reviews are always complex, like abbreviations.In addition, neologisms often occur among reviews in a new movie, which leads traditional methods (like CRF) to unsatisfied performance. Based on these characteristics of movie reviews, we propose a novel person name recognition method, called Multi-Feature Bi-LSTM-CRF Model. This model extracts relevant character-level features by using external corpora and unlabeled reviews; Then, a framework of Bi-LSTM-CRF is applied to mark the sequence of person names. The experimental results show that our model can effectively identify different forms of person names in the movie reviews as well as the unregistered word for a new person name, and thus shows a strong ability for name entity extraction from movie reviews.

Keyword: film review; LSTM; CRF; Multi-Feature; Person Name Recognition

---

*收稿日期： 定稿日期：

基金项目： 国家自然科学基金项目（61572145），广东省教育厅基础研究重大项目及应用研究重大项目（2017KZDXM031）

1. 引言

命名实体(Named Entity)[1]指的是文本中具有命名性指称的词，人名作为命名实体之一，其内部组成更复杂、识别难度较大。当前人名实体研究正从传统媒体转到社交媒体当中，如微博、Facebook 等。作为社交信息之一，影评的句法往往不规则甚至不完整；而其中的人名组成也更为多元。在影评中，“刘德华”、“周星驰”等往往会被带主观情感的称谓所替代，如“华仔”、“星爷”。这些称谓同样在“具有命名性指称的实体”的范畴当中，却常被忽略不计。此外，由于电影选角和题材上的差异，新电影中普遍存在人名新词，或称未登录词。然而关于这些问题，目前学术界仍未取得较大的突破。于此同时，对于影评的人名抽取技术日益受到工业界的关注，从影评中抽取相关的主创人物如导演、演员、角色、编剧等，能为明星营销、主创票房贡献价值分析、情感倾向分析[2]等情报技术提供支持。

传统的中文人名识别方法多是基于规则和概率计算的。李中国[3]提出了基于边界模板和局部统计的识别方法，首先从标注语料中提取边界模板以定界候选人名词汇，接着利用局部统计量和相关修正规则对候选人名进行修正。倪吉[4]通过抽取外部人名语料中的用字特征和边界特征，以计算人名内聚度、人名区分度和边界模板可信度的综合概率。而当下主流的方法多是基于机器学习模型进行训练，该类方法对标注语料进行学习，并以序列标注的形式实现人名识别。 如最大熵模型，隐马尔可夫模型，条件随机场模型等。机器学习方法的好处在于能够学习特征间的关联性和重要性。曹波[5]以词作为标注对象，先进行最大概率分词，然后利用人名角色表和词性表将句中词分为人名内部组成、上下文、无关词等，以此构造特征模板，最后利用最大熵模型进行训练和预测。该方法在 1998 年一至五月的人民日报语料中取得了 89.43%的识别精度和 94.26%的召回率。张素香[6]以原子特征、全局变量特征，复合特征等构造特征模板，并利用条件随机场模型实现人名抽取。该方法将准确率提升至 95%。上述方法均已词作为训练和标注的基本单位。然而目前大多数分词工具仅针对相对规则的人名实体，难以对影评中的人名称谓和人名未登录词进行有效的切分。另一方面，基于字符的方法在基于统计的机器学习模型中存在一定的缺陷，语言学界一般认为词是语义的最小单位，而字符往往缺乏充足的语义信息。近年来深度学习在自然语言处理领域取得丰硕成果，基于深度学习的命名实体识别方法[7,8,9]逐渐涌现，如长短期记忆网络（Long Short-Term Memory，LSTM)。LSTM 具有远距离记忆功能，能够处理标注时的长距离依赖问题。LSTM 的这一特性在一定程度上克服了字符级特征的不足。Dong C[10]首次将字符级的 Bi-LSTM-CRF 模型应用到中文命名实体识别任务中，并提出利用汉字部首作为字符的特征表示之一，然而影评中的译名和特殊称谓所包含的汉字并没有明显固定的部首，该方法并不适用于人名识别。

综合上述提到的问题，本文提出一种基于深度学习的影评人名识别方法。该方法将预训练的字向量（Character Embedding）和传统方法中常用的人工特征（边界特征和用字特征）整合为统一的字符级（Character-Level）特征；采用 Bi-LSTM-CRF 模型[11]进行字符序列标注，从而实现人名识别。

2. 模型

图 1 为所提出模型的整体架构，该模型通过构建字向量表、边界向量表和用字向量表为字符提供特征支撑。模型首先提取字符对应的三类特征，三类首尾特征拼接后作为 Bi-LSTM 层的输入，经过 Bi-LSTM 提取隐藏层特征 $h_1, h_2, \ldots h_n$ ；并以此作为 CRF 层的输入，CRF 对上下文标注以进一步约束后，输出序列标注结果 $y_1, y_2, \ldots y_n$ 。

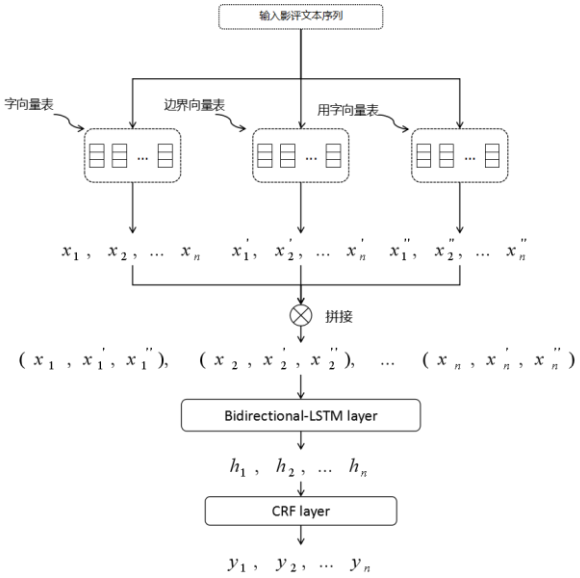

图 1 多特征 Bi-LSTM-CRF 模型框架

2.1 特征

模型采用的特征包括字向量、边界特征和用字特征。这些特征均为字符级的，下面详细描述三类特征的构造和向量化。

（一）字嵌入

字嵌入也称字向量，是对文本集合中各字符的分布式表示，字向量能够表示字符的句法和语义信息。字向量的概念源于词向量，其实质就是把语料中的每一个词映射至同一向量空间中，从而将两个词的语义距离转换为向量空间中的物理距离。当前对词嵌入的研究较广泛[12,13,14]。其中较为著名的属 Mikolov[15]提出的 word2vec 和 Jeffrey Pennington 提出的 GloVe[16]。

Skip-gram 是 word2vec 中的模型，其实质是一个三层的神经网络，它基于当前词来预测一定窗口内的上下文，模型训练目标是获取最大概率产生当前序列观测数据的隐藏层参数。而 GloVe 是一种更新的基于共现矩阵（co-occurrence matrix）的词向量模型。GloVe 通过矩阵分解的方法，不仅考虑到 word2vec 窗口的上下文信息，也考虑到全局信息，因此 GloVe 能更全面的表达词或字符的语义。利用 GloVe 对大规模影评数据进行字向量训练，使得人名之间可以进行相似性的度量，其意义在于与已有人名相似的未登录词更容易被识别，从而提高人名识别的召回率。

GloVe 首先通过滑动窗口构建词与词间的共现矩阵。定义 $X_{i,j}$，表示词 j 和词 i 共同出现在窗口内的次数。定义 $X_i = \sum_k X_{i,k}$，表示在词 i 窗口内出现的总词数，k 为窗口内的词。定义 $P_{i,k} = X_{i,k} / X_i$，表示词 k 出现在词 i 窗口内的概率。定义 $ratio_{i,j,k} = P_{i,k} / P_{j,k}$，$ratio_{i,j,k}$ 的值揭示了词 i、j、k 之间的相关性。

当 $ratio_{i,j,k}$ 很大，则词 i、k 相关，j、k 不相关；当 $ratio_{i,j,k}$ 很小，则词 j、k 相关，i、k 不相关。与条件概率相比，$ratio_{i,j,k}$ 能更好地表示相关词对与不相关词对。Glove 模型的实质是训练函数 $F(w_i, w_j, \tilde{w}_k)$，使其极大概率地拟合实际的 $ratio_{i,j,k}$。其中，w_i、w_j、$\tilde{w}_k$ 分别表示词 i、j、k 对应的词向量。当 $F(w_i, w_j, \tilde{w}_k)$ 基本拟合 $P_{i,k} / P_{j,k}$ 的分布时，意味着词向量与共现矩阵具有一致性，即词向量能表示共现矩阵中的信息。

考虑到部分词共现属于噪声，不利于模型学习参数。在构造损失函数时，引入赋权函数 $f(X_{i,j})$，完整的损失函数如式(1)所示:

$$J=\sum_{i,j=1}^{V} f(X_{i,j})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{i,j})^2 \tag{1}$$

式中，$X_{i,j}$ 表示词 i 、词 j 共同出现在窗口内的词数；w_i^T 指词 i 的词向量转置，词 i 为窗口内 d 的上下文；$\tilde{w}_j$ 指词 j 的词向量，词 j 为窗口中心词。b_i 和 $\tilde{b}_j$ 分别为 w_i^T 和 $\tilde{w}_j$ 对应的偏移量。V 指语料包含的词集。 $f(x)$ 见式(2):

$$f(x)=\begin{cases}(x/x_{\max})^a & \text{if} \quad x<x_{\max} \\ 1 & otherwise\end{cases} \tag{2}$$

通常，设 a =0.75， $x_{\max}$ =100。

（二）边界特征

中文人名一般具有边界模糊的问题。所谓边界是指与人名相邻接的词或字。传统方法[3]一般通过构建边界模板以定界候选人名。但在采用序列标注模型时，一般难以确定一个人名的长度。本文以已标人名作为种子词，在未标注语料中进行上下边界字符的提取。表 1 列举了人名上下边界中的高频字符。

表 1 上下边界字符频数

上边界	我	待	爱	为	给	看	着	@	了	有	#	欢	是	的	和
	22	28	32	35	41	42	51	73	74	80	91	101	111	121	126
下边界	帅	很	这	去	还	在	太	#	真	好	也	是	演	和	的
	22	25	27	27	30	31	33	33	35	36	37	38	114	118	340

统计发现，上边界共有 2601 种不同字符，前 15 种字符占了总频率的 26.2%，下边界则有 2633 种不同字符，前 15 种字符占了总频率的 22.1%。其中，“帅”、“爱”、“@”、“#”、“演”、“太”、“很”等高频边界表明影评的强领域性。而人名边界的集中分布情况也说明边界信息具有一定的人名区分能力。考虑到高频字符中存在常见的停用词，如表 1、2 中的“和”、“的”等，本文采用以可信度作为边界特征的衡量标准，可信度定义见式(3)：

$$P(c_i)=\frac{\log(f_{c_i}+1)}{\log(w_{c_i}+1)} \tag{3}$$

式中，c_i 表示训练语料中的第 i 个字符， f_{c_i} 表示字符 c_i 作为上文（下文）边界的频率。 w_{c_i} 表示 c_i 在未标注语料中的频数。为了将特征融入神经网络模型，对 $P(c_i)$ 进行标准化和离散化处理，并随机生成向量，作为网络的输入。见式(4)：

$$F(c_i)=\text{round}(\frac{P(c_i)-P_{\min}}{P_{\max}-P_{\min}}\cdot k) \tag{4}$$

式中，round 函数为四舍五入计算，k 为切割值，控制离散化后的特征数。离散化后的边界特征可参照字向量的形式映射至向量空间当中，作为神经网络的输入。

（三）用字特征

在中文人名识别中，用字特征一般以布尔值或可信度进行衡量[5,20]。本文在此基础上进行了

改良。本文将用字特征分为七类，包括姓用字、单名字、双名首字、双名尾字、译名首字、译名中字、译名尾字等。这七类用字不仅对中文人名和国外译名的识别有帮助，大多数人名称谓也存在这七类用字，例如“华仔”和“吴先生”。其中“华”是双名尾字或者单名尾字，“吴”则是姓用字。在衡量特征值时，对大规模的中文人名和国外人名语料进行字符频数统计。与边界特征一样，离散化后随机映射至向量空间中，并作为神经网络的输入。用字特征的计算和离散化过程见式(5)：

$$v_c = \text{round}((\frac{\log(f_c - f_{min}+1)}{\log(f_{max} - f_{min}+1)}) \cdot k) \tag{5}$$

其中 c 为字符，v 为字符 c 对应特征值，fc 为 c 的字符频率，f_{max}为频率最大值，f_{min}为频率最少值，k 为切割值，控制离散化后的特征数量。

2.2 Bi-LSTM

LSTM 网络以上一时刻的隐藏层输出向量和当前字符向量作为当前标注的衡量信息，计算上一时刻的标注对当前标注的影响。LSTM 网络能较好的控制信息的输出、输入和保存，在 LSTM 神经元中，状态的保存与更新由输入门、忘记门和输出门决定。输入门控制从输入信息中那些可以保存到状态中，忘记门决定历史状态的保留信息，输出门控制更新后的状态中哪些信息被输出。LSTM 具体工作流程如式(6)(7)(8)(9)(10)。

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}[x_t, x_t', x_t''] + b_i) \tag{6}$$

$$f_t = \sigma(W_f \cdot h_{t-1} + W_{fx}[x_t, x_t', x_t''] + b_f) \tag{7}$$

$$o_t = \sigma(W_o \cdot h_{t-1} + W_{ox}[x_t, x_t', x_t''] + b_o) \tag{8}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_C \cdot h_{t-1} + W_{Cx}[x_t, x_t', x_t''] + b_C) \tag{9}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{10}$$

式中，i_t、f_t、o_t、C_t 分别表示 t 时刻神经元的输入门、忘记门、输出门和细胞状态。x_t、x_t'、x_t''分别表示 t 时刻的字向量、边界向量和用字向量。h_{t-1} 表示 t-1 时刻的隐藏层特征，W 和 b 为各门的权重矩阵以及偏置向量。σ 表示 sigmoid 激活函数，tanh 表示 tanhyperbolic 激活函数；激活函数将神经网络转换为非线性模型，以逼近样本的空间分布。

双向长短期记忆网络(Bi-LSTM)在 LSTM 网络的基础上进行了扩展。Bi-LSTM 包含两层异向的 LSTM 网络。通过前向 LSTM 层网络获得前向特征$\overrightarrow{h_t}$，通过后向的 LSTM 获得反向特征$\overleftarrow{h_t}$。Bi-LSTM 通过拼接前后两层 LSTM 的特征以表示词或字符，$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$。Bi-LSTM 相较于单向的 LSTM，能更充分地考量上下文依赖信息。

2.3 Bi-LSTM-CRF

Bi-LSTM-CRF 在 Bi-LSTM 的基础上扩充了 CRF 层，其结构如图 2。CRF[17]模型在序列标注任务中的优越性能已被多次验证。在 Bi-LSTM-CRF 模型中，CRF 的主要作用是进一步增强前后标注的约束，避免不合法的标注情况出现，如标签“B-nr”后面接标签“E-nr”的情况。对于 Bi-LSTM 的输出序列$h = (h_1, h_2, ..., h_n)$，通过概率模型 CRF 获得候选标签序列$y = \{y_1, y_2, ..., y_n\}$，CRF 原理如式(11)：

$$P(y \mid h; \mathrm{W}, \mathrm{b}) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, h)}{\sum_{y' \in Y(h)} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_i, h)} \tag{11}$$

式中，$p(y \mid h; W, b)$为给定隐藏层序列 h,标记为序列 y 的概率，$\psi_i(y' y, h) = \exp(\mathrm{W}_{y',y}^{T} h_i + \mathrm{b}_{y',y})$为势函数，$W_{y',y}^{T}$和$b_{y',y}$分别为权重向量和偏移量。在 CRF 训练时，使用极大似然估计的方法。对于训练集$\{(h_0, y_0), (h_1, y_1)$，，．$(h_n, y_n)\}$,其优化目标如式(12)：

$$L(W, b) = \sum_{i=1}^{n} \log P(y_i \mid z_i; \mathrm{W}, \mathrm{b}) \tag{12}$$

最大似然估计的目标是调整相关参数 W 和 b,使得$L(W, b)$最大化。在使用 CRF 进行标注时，选取概率最大的候选标注序列作为最终标注结果。

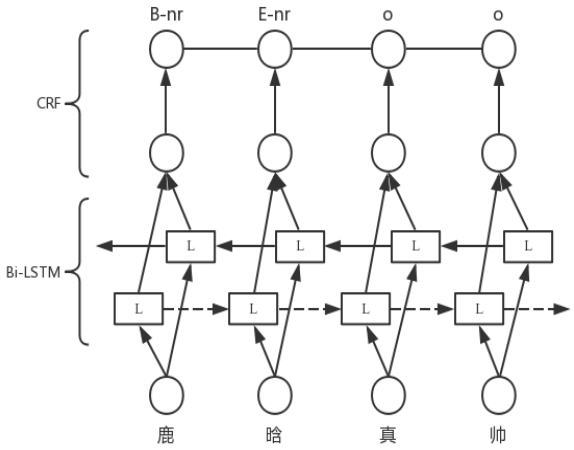

图 2 Bi-LSTM-CRF

2.4 模型训练

训练时采用 Adam[18]作为优化器，学习速率为 0.001；dropout[19]为 0.3；预训练的字嵌入设为 200 维，上下边界特征和各类用字特征均设为 32 维。每一层 LSTM 网络设 256 个神经元。模型训练时，若连续迭代 5 次后，验证集对应的损失值均未下降，则训练完成。

3. 实验结果

3.1 实验语料

影评数据获取自微博电影[1]。共获取 1224 部电影，总计 600 多万评论。接着对 80 部电影（除动画）进行了标注，最终获得有效评论 2247 条。标注时，将影评中的人名实体分为中文人名、国外译名和人名称谓。表 2 列出了人名实体的定义。外部人名语料获取自网络语料，共获取中文人名 120 万和国外译名 48 万。

为验证本文方法，设置两组数据集：数据集 A，忽略影评所属电影，将已标评论集进行随机的切分；数据集 B，从 80 部电影中随机抽取 65 部作为封闭集，15 部作为开放集，以模拟在

[1] http://movie.weibo.com

已有电影的情况下对新电影评论进行人名识别。表 3 给出两组数据集中各类人名的情况。两组数据中各类人名占比基本一致，主要差异在于未登录词的数量。未登录词指开放集中存在而封闭集和外部人名语料中均不存在的人名实体。数据集 B 中的人名未登录词约占总数的 31%，远高于数据集 A 的 14%。

表 2 影评中的人名定义

类别	定义	例子
中文人名	姓氏+单名	陈坤
	姓氏+双名	吴亦凡
国外译名	姓氏	亚伯拉罕
	名称	珍妮
	名称+“·”+姓氏	范·迪塞尔
人名称谓	双名	亦凡
	名用字+名用字	凡凡、鹿鹿
	姓氏+称谓	吴先生、冯导演
	昵称字+名用字	老吴、啊包、我凡
	名用字+昵称字	凡哥
	特殊称谓	水果姐、小绵羊

表 3 数据集的人名分布情况

数据集	训练集				测试集				
	中文名	译名	称谓	总计	中文名	译名	称谓	总计	未登录词
A	1363	355	810	2528	321	76	183	576	**81**
B	1312	343	811	2469	372	88	182	640	**199**

3.2 评价指标

实验结果采用识别准确率（P）、召回率（R）和二者的调和平均 F1 值（F）作为评判指标。P 指正确识别的人名占总计识别的人名的百分比，R 指正确识别的人名占测试集中所有人名的百分比，F 是 P 和 R 的调和平均值，综合考量模型的性能。

3.3 实验结果与分析

首先以 Bi-LSTM-CRF 为基础分别对字嵌入(E)、边界特征（B）、用字特征（U）等特征进行测试，测试在数据集 A 中进行。实验基线为基于字符的 Bi-LSTM-CRF 模型，模型随机生成向量作为字符特征。字嵌入的测试对比了 skip-gram 以及 GloVe；边界特征和用字特征的测试则分别设置不同的 k 值，以对比特征带来的增益。实验结果（表 4）表明，相较于 skip-gram，GloVe 的字嵌入表示效果更优，而在边界特征和用字特征方面，当 k_B=2 和 k_U=5 时，特征对模型带来的增益达到最高。当 k 值继续增大时，特征泛化能力减弱，F1 值逐渐下降。当组合各特征（EBU）时,模型的综合 F1 值达到 89.8%，高于所有单特征模型。后续实验均在特征参数最优的情况下进行。

本文进一步对比了 CRF、CRRM[20]、Bi-GRU-CRF、Bi-LSTM-CRF、Bi-GRU-CRF（EBU）、Bi-LSTM-CRF（EBU）在数据集 A、B 上的表现（表 5）。CRRM 在传统模型 CRF 的基础上加入了可信度衡量和规则的方法。GRU[21]也是 RNN 中的一种主流结构，相比 LSTM，其结构更简单、参数更少。GRU 只有两个门，分别为更新门（update gate）和重置门(reset gate)。该门结构

能起到信息保存的作用，使得依赖信息不会由于长距离的传播而完全丢失。实验结果显示，Bi-LSTM-CRF（EBU）在综合指标上表现最佳，在数据集 A 和 B 上的 F1 值分别为 89.8%和 81.9%，远高于传统方法 CRF 和 CRRM，神经网络模型对比方面，在一般情况下（数据集 A）Bi-LSTM-CRF（EBU）的 F1 值与 Bi-GRU-CRF（EBU）相当，仅高出 0.3%。而在面对未登录词更多的情况（数据集 B）表现更佳，比 Bi-GRU-CRF（EBU）高出 0.8%。

表 4 各特征对识别效果的影响

特征	配置	中文人名			国外译名			特殊称谓			总计		
		P	R	F	P	R	F	P	R	F	P	R	F
Char	random	78.9	91.0	84.5	71.8	73.7	72.7	77.8	73.9	75.8	82.9	83.4	83.2
E	skip-gram,dim	89.9	91.0	90.4	80.1	77.9	79.0	76.8	72.9	74.8	84.7	83.7	84.1
	GloVe	92.1	94.1	93.1	74.4	84.2	79.0	78.7	76.1	77.4	85.6	87.2	86.4
B	k_B=2	88.4	93.8	91.0	**83.8**	75.0	79.2	80.3	77.2	78.8	85.7	85.9	85.8
	k_B=5	90.0	92.3	91.1	77.5	72.4	74.8	84.7	71.7	77.7	86.9	84.4	85.6
	k_B=8	89.4	91.3	90.4	79.7	77.6	78.7	78.4	76.7	77.5	84.9	85.0	84.9
U	k_U=2	91.7	96.0	93.8	79.2	80.3	79.7	81.4	75.6	78.4	86.9	87.4	87.2
	k_U=5	93.3	94.8	94.0	76.3	80.3	78.2	79.7	82.7	81.2	86.7	89.1	87.9
	k_U=8	92.8	96.0	94.4	74.7	81.6	78.0	78.5	78.9	78.7	86.0	88.8	87.4
	k_U=20	92.8	96.0	94.4	74.7	81.6	78.0	78.5	78.9	78.7	83.0	88.5	85.6
EBU	Glove,dim=200,k_B=2,k_U=5	**94.6**	**96.9**	**95.7**	78.8	**82.9**	**80.8**	**85.7**	**80.0**	**82.8**	**89.7**	**89.9**	**89.8**

表 5 各模型的识别效果

数据集	A（%）			B（%）		
指标	P	R	F	P	R	F
CRF	86.4	76.7	81.3	71.7	63.6	68.4
CRRM[20]	**88.9**	87.7	87.7	76.3	74.5	75.4
Bi-GRU-CRF	84.1	80.8	82.4	75.3	67.4	71.2
Bi-LSTM-CRF	82.9	83.4	83.2	75.5	70.4	72.8
Bi-GRU-CRF（EBU）	88.7	**90.5**	89.5	83.0	79.2	81.1
Bi-LSTM-CRF（EBU）	89.7	89.9	**89.8**	**83.1**	**80.8**	**81.9**

为了进一步验证并对比字嵌入、用字特征、边界特征在未登词识别时的增益，本文对数据集 B 进行了对比实验。实验结果表明（表 6），字嵌入和边界特征虽然对模型的准确率提升不大，但能够使模型识别更多的未登录词，召回率分别提高了 7.4%、2.7%。相较于二者，用字特征为模型带来了更大的增益，召回率提升高达 16.8%，可见通过外部语料提取相关用字知识能够有效提升模型的人名新词识别能力。

我们抽取部分识别结果来对比本文模型和 CRF 在某些人名称谓上的识别差异(表 7)。多特征 Bi-LSTM-CRF 能完整识别出“我林哥哥”，而 CRF 仅识别出“我林哥”；可见本文模型能够较好的克服称谓中的边界模糊问题。此外，本文模型还从语料中学习到“吴宝”、“兴宝”这类以“宝”为结尾的称谓模式，可见模型能较好的适应人名称谓的组成多元性。

表 6 各特征对人名未登录词识别的增益

特征	P(%)	R(%)
None	55.7	48.1
E	53.5（-2.2）	55.5（+7.4）
B	55.7（+0.0）	50.8（+2.7）
U	**59.0**（3.3）	64.9（+16.8）
EBU	58.6（2.9）	**67.4**（+19.3）

表 7 CRF 和本文模型所识别出的人名称谓

人物	Bi-LSTM-CRF（EBU）	CRF
吴亦凡	二凡、凡宝、吴先生、老吴、吴皇、凡凡、吴凡	二凡、老吴、吴皇、凡凡、吴凡
林俊杰	老林、我林哥哥、俊杰、林二	老林、我林哥、林二
张艺兴	小绵羊、绵羊兴、兴宝、艺兴	小绵羊、艺兴
未知	阿福、铁叔、李××	铁叔

4. 总结与展望

针对影评中称谓复杂多样，人名未登录词普遍存在的特点，本文提出一种混合多特征的 Bi-LSTM-CRF 模型，该方法利用外部人名语料和未标注影评提取字向量、用字特征、边界特征等，并将这些特征规整为统一的字符向量以作为模型的输入。实验结果表明，Bi-LSTM-CRF 模型相较于传统方法和 Bi-GRU-CRF，具有更优的人名识别效果。而且字向量、用字特征、边界特征的引入有效的提高了模型对人名未登录词和人名称谓的识别能力。然而，本文模型对用字特征和边界特征进行更精确的学习和表示。如何对这些外部特征进行有效的学习，将是后续研究的重点。同时，我们还会对提出的方法作进一步提炼，提高其泛化能力，并应用至相似的社交媒体文本中，以提高对这类不规则文本的人名实体识别能

参考文献

[1]Sundheim B M. Name Entity task definition, verson2.1[J]. In:Proc.of the Sixth Message Understanding Conf, 1995, 319-332.

[2]刘鸿宇,赵妍妍,秦兵,刘挺.评价对象抽取及其倾向性分析[J].中文信息学报,2010,24(01):84-88.

[3]李中国,刘颖.边界模板和局部统计相结合的中国人名识别[J].中文信息学报,2006(05):44-50.

[4]倪吉,孔芳,朱巧明,李培峰.基于可信度模型的中文人名识别研究[J].中文信息学报,2011,25(03):45-50.

[5]曹波,苏一丹,邓琦.基于最大熵模型的中国人名自动识别[J].计算机工程与应用,2009,45(04):227-228.

[6]张素香,高国洋,戚银城.基于条件随机场的中国人名识别方法[J].郑州大学学报(理学版),2009,41(02):40-43.

[7]Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, vol. 4:357-370.

[8]Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics, 2017, 33(14):i37-48.

[9]Pham T H, Le-Hong P. End-to-end Recurrent Neural Network Models for Vietnamese Named Entity Recognition: Word-level vs. Character-level[J]. Computational Linguistics, 2017:219-232.

[10]Dong C, Zhang J, Zong C, et al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[M]// Natural Language Understanding and Intelligent Applications: Springer International Publishing, 2016:239-250.

[11]Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

[12]Levy O, Goldberg Y. Dependency-Based Word Embeddings[C]// Meeting of the Association for Computational Linguistics. 2010:302-308.

[13]Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings[C]// International Conference on Artificial Intelligence. AAAI Press. 2015:1236-1242.

[14]Tang D, Wei F, Yang N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]// Meeting of the Association for Computational Linguistics. 2014:1555-1565.

[15]Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[16]Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers. Inc. 2001:282-289.

[17]Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[18]Kingma, D. P., & Ba, J. L. Adam: a Method for Stochastic Optimization[C]// International Conference on Learning Representations. 2015: 1-15.

[19]Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[20]王祖兴,吕钊,顾君忠.基于混合方法的中文人名识别研究[J].计算机工程与应用,2015,51(08):211-217.

[21]Dey R, Salemt F M. Gate-variants of Gated Recurrent Unit (GRU) neural networks[C]// IEEE, International Midwest Symposium on Circuits and Systems. IEEE, 2017:1597-1600.

褶镇宇（1995-），男，本科，主要研究领域为数据挖掘、自然语言处理。Email: xuanzhenyu@foxmail.com

蒋盛益（1963-），男， 教授，硕士生导师，主要研究领域是数据挖掘、自然语言处理。本文通信作者。Email：jiangshenyi@163.com

张礼明（1994-)，男，硕士，主要研究领域为数据挖掘、自然语言处理。Email：zhangliming134@foxmail.com

包睿（1997-)，男，本科，主要研究领域为数据挖掘、自然语言处理。Email：13160835526@163.com