

钟茂生, 吴佳华, 罗玮, 等. 面向低资源命名实体识别的 BiLSTM-Att-BCRF 模型 [J]. 江西师范大学学报(自然科学版), 2022, 46(5): 460-467.

ZHONG Maosheng, WU Jiahua, LUO Wei, et al. The BiLSTM-Att-BCRF model for low resource named entity recognition [J]. Journal of Jiangxi Normal University( Natural Science), 2022, 46(5): 460-467.

文章编号: 1000-5862(2022)05-0460-08

# 面向低资源命名实体识别的 BiLSTM-Att-BCRF 模型

钟茂生, 吴佳华, 罗 玮, 吴水秀

(江西师范大学计算机信息工程学院, 江西 南昌 30022)

**摘要:** 在低资源场景下, 由于受训练数据量少的限制, 现有模型的参数不能拟合到预期效果, 所以导致模型识别实体的性能不佳. 该文提出一种融入伯努利分布(Bernoulli distribution)的新型损失函数, 使模型能较好地拟合数据. 此外, 该文在 BiLSTM-CRF 模型基础上融合多层字符特征信息和自注意力机制, 并结合基于伯努利分布的新型损失函数, 构建了 BiLSTM-Att-BCRF 模型. BiLSTM-Att-BCRF 模型在 20% 的 CONLL2003 和 20% 的 BC5CDR 的数据集上,  $F_1$  值在 BiLSTM-CRF 模型基础上分别提升了 7.00% 和 4.08%, 能较好地适应低资源命名实体识别任务.

**关键词:** 低资源命名实体识别; 神经网络; 伯努利分布; 自注意力机制

**中图分类号:** TP 391 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2022.05.04

## 0 引言

命名实体识别是自然语言处理的基础任务之一, 该任务旨在从非结构化的文本中自动识别出实体, 并将其标记为预定义的类别(如人名、地名和组织机构名等). 如“张无忌, 金庸武侠小说《倚天屠龙记》人物角色, 中土明教第三十四代教主.”这句话包含的实体有: 人名实体“张无忌, 金庸”, 书名实体《倚天屠龙记》, 门派实体“明教”. 由此可见, 实体识别是理解文本语义的基础. 同时命名实体识别技术在知识图谱构建、机器翻译、知识库构建等多种自然语言处理任务中有着广泛的应用.

近年来, 深度学习方法被广泛用于命名实体识别中. 如 J. Hammerton<sup>[1]</sup> 将长短期记忆网络(LSTM)应用于实体识别研究中, LSTM-CRF 结构成为实体识别的基础结构. L. Guillaume 等<sup>[2]</sup> 在 LSTM-CRF 模型的基础上, 提出双向长短期记忆网络(Bi-LSTM)和条件随机场(CRF)结合的模型<sup>[3-5]</sup>等. 这类方法

虽然在文本实体识别任务中表现较为优异, 但需要有大规模的标注数据, 需对训练语料中每个词进行人工标注. 在标注数据不足的情况下, 现有模型的参数不能较好地拟合数据, 导致模型预测最大概率标签并不一定是真实标签. 模型的识别性能下降, 很难应用到如生物、医学等这些标注语料较少的领域中. 针对上述问题, 本文提出一种融入伯努利分布的新型损失函数, 使模型参数在低资源场景下也能较好地拟合数据. 在此基础上, 为了增加模型能处理的词汇量和提升模型识别罕见词的能力, 本文在 BiLSTM-CRF 模型的基础上融合多层字符特征信息, 构建了 BiLSTM-BCRF 模型. 在 BiLSTM-BCRF 模型基础上融合了自注意力机制, 获取关键信息隐藏状态表示, 构建了 BiLSTM-Att-BCRF 模型, 进一步提升了模型的精确率和召回率.

## 1 相关工作

命名实体识别的研究方法主要有基于规则和词

收稿日期: 2022-06-17

基金项目: 国家自然科学基金(61877031)和江西省教育厅科技课题(GJJ210324)资助项目.

作者简介: 钟茂生(1974—), 男, 江西兴国人, 教授, 博士, 主要从事机器学习、自然语言处理和智能教育软件的研究.

E-mail: zhongmaosheng@sina.com

典方法、机器学习方法、深度学习等方法。基于词典和规则的方法过多依赖于语言学家制定的规则模板,容易产生错误,且移植性差。传统机器学习方法主要包括:隐马尔科夫模型(hidden Markov model, HMM)、最大熵(maximum entropy, ME)<sup>[6]</sup>、最大熵马尔科夫模型(maximum entropy Markov model, MEMM)<sup>[7]</sup>、条件随机场(conditional random fields, CRF)<sup>[8]</sup>等。这些传统的机器学习方法在特征提取方面需要人工参与,同时需要大规模的标注语料来训练模型,方法的性能主要依赖于所采用的特征是否具有辨识度。其中CRF被看作是命名实体识别的主流模型,其优点是在对一个位置进行标注的过程中可以利用内部及上下文特征信息。随着深度学习的不断发展,命名实体识别的研究重点已转向深层神经网络,R. Collobert等<sup>[9]</sup>提出基于神经网络的命名实体识别方法,在该方法中每个单词具有固定大小的窗口,但未能考虑长尾问题。为了克服这一限制,J. P. C. Chiu等<sup>[10]</sup>提出一种双向LSTM-CNNs架构,该架构可自动检测单词和字符级别的特征。J. Ham-merton<sup>[11]</sup>利用CRF关注上下文特征信息的特点,提出了LSTM-CRF模型。

近年来,大量的深度学习方法被应用于低资源命名实体识别任务中。低资源的命名实体识别技术是当前研究热点之一,其性能的提高是命名实体识别技术走向广泛实际应用的前提。相关研究工作可大致分为跨语言迁移的方法、数据增强的方法和集成自动标注语料的方法等。

跨语言迁移方法的基本思路是利用资源丰富语言的标注数据以帮助低资源语言进行命名实体识别,可大致分为数据迁移方法和模型迁移方法2大类。基于数据迁移的方法通常借助文本翻译和标签映射等把源语言中的标注数据转换成目标语言的标注数据,然后基于这些数据训练模型。Ni Jian等<sup>[11]</sup>提出了一种在语料库上进行标签映射的方法,用于创建自动标记的目标语言数据。S. Mayhew等<sup>[12]</sup>利用双语词典,使用一种类似短语机器翻译<sup>[13]</sup>的方法自动翻译源语言的标注文本。基于模型迁移的方法通常需先学习语言无关的特征,然后将在源语言的标注语料上训练的NER模型直接用于目标语言。Chen Xilun等<sup>[14]</sup>同样基于对抗学习的方法提取语言无关的特征,并动态地计算源语言和目标语言之间的相似度,从而更有效地实现了从多个源语言到目标语言的知识迁移。K. Phillip等<sup>[15]</sup>在多语言版本BERT的基础上进一步使用对抗学习<sup>[16]</sup>的方法,

以学习与语言无关的更好的特征。

数据增强方法的主要目标是在不增加人工标注成本的前提下通过增加合理的噪声来提升模型的鲁棒性,这在数据量少的场景下对模型性能的提升有较大帮助。Dai Xiang等<sup>[17]</sup>引入了一些词替换的随机操作来增加训练语料多样性;Chen Jiaao等<sup>[18]</sup>在半监督NER任务中引入了基于局部可加性的数据增强。跨语言迁移的方法和数据增强的方法虽然能够有效地缓解标注语料短缺的问题,然而具有丰富标注资源的语言是非常少的。

一些研究者提出集成自动标注语料的方法,首先通过某种方法自动标注大量语料,然后将它们集成并用于提高低资源实体识别模型的性能。Yang Yaosheng等<sup>[19]</sup>首先基于词典匹配的方法自动标注语料,然后使用Partial-CRF<sup>[20]</sup>在少量人工标注的语料和大量自动标注的语料上训练实体识别模型。此外,他们还基于强化学习<sup>[21]</sup>训练一个选择器,用于筛选具有噪声的标注数据。

除上述3类方法外,在低资源实体识别领域中还有其它方法,如Zhang Tao等<sup>[22]</sup>提出的渐进式知识提炼方法PDALN有效地将高资源域适应于低资源目标域;Chen Siqi等<sup>[23]</sup>提出了一种低资源的语言模型的微调方法,使用基于注意力机制的微调策略,从预训练的语言模型中选择相关的语义和句法信息,并将其应用于命名实体识别任务。本文主要是探索在低资源条件下基于深度学习的命名实体识别方法。

## 2 模型

### 2.1 基本架构

命名实体识别任务被看作是序列标注问题。输入句子表示为 $x = (x^1, x^2, \dots, x^i)$ ,其中 $x^i$ 表示第 $i$ 个字符(包括数字、单词、字母或标点符号等)。输出标注序列为 $y = (y^1, y^2, \dots, y^i)$ ,其中 $y^i \in \{B, M, E, S, O\}$ 是 $x^i$ 的标签, $B$ 、 $M$ 、 $E$ 、 $S$ 和 $O$ 分别代表实体首字、实体中间字、实体结尾字、实体单独字和非实体。命名实体识别就是对每个字符进行 $B$ 、 $M$ 、 $E$ 、 $S$ 、 $O$ 的分类标注。

本文在BiLSTM-CRF模型损失函数中融入伯努利分布,使模型参数在低资源场景下能较好地拟合数据,同时为增加模型能处理的词汇量和提升模型识别罕见词的能力,将多层字符信息融合到BiLSTM-CRF模型中,构建了BiLSTM-BCRF模型。在BiLSTM-

BCRF 模型的基础上,融合自注意力机制,获取了关键信息隐藏状态表示,构建了 BiLSTM-Att-BCRF 模型. BiLSTM-Att-BCRF 模型基本结构如图 1 所示,该

模型结构主要分为输入层、Bi-LSTM 层、自注意力层和 BCRF 层.

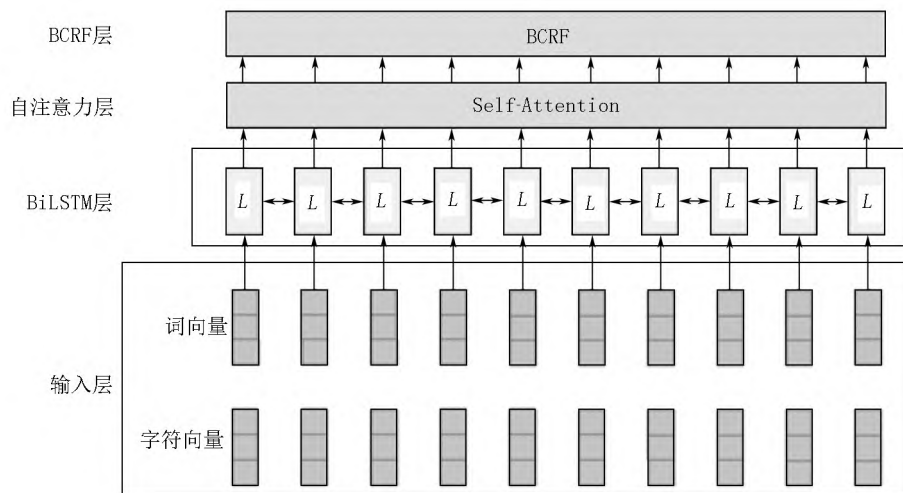


图1 BiLSTM-Att-BCRF 模型基本结构

## 2.2 输入层

图 2 为模型输入层的结构图,其中  $x$  表示词向量,是使用 J. Pennington 等<sup>[24]</sup>提出 Glove 英文词向量文件生成的  $c_1, c_2$  表示字符;  $m, m_1$  由 BiLSTM 训练生成的字符向量  $\hat{x} = (x \parallel m \parallel m_1)$ . 最后将词向量与字符向量拼接输入到 BiLSTM 层.

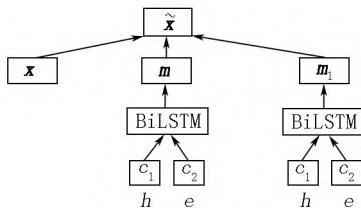


图2 模型输入层基本结构图

## 2.3 BiLSTM 层

LSTM 神经网络在命名实体识别任务中表现出良好的建模能力,能较好地学习单词与字符的特征信息, BiLSTM 层结构主要由 2 个 LSTM 组合而成. LSTM 的网络结构分为遗忘阶段、选择记忆阶段和输出阶段. LSTM 单元结构如图 3 所示,其中  $i_t, f_t, o_t$  分别表示 LSTM 单元的输入门、遗忘门、输出门在  $t$  时刻的状态.  $h_{t-1}$  表示在  $t-1$  时刻的隐藏状态,  $\hat{c}_t$  表示在  $t$  时刻的细胞记忆状态.  $\sigma$  表示 Sigmoid,  $\tanh$  表示双曲正切激活函数:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \\ f_t &= \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f), \\ \hat{c}_t &= \sigma(W_{ch}h_{t-1} + W_{cx}x_t + b_c), \\ c_t &= f_t c_{t-1} + i_t \hat{c}_t, \\ o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \end{aligned}$$

$$h_t = o_t \odot \tanh(c_t),$$

在 BiLSTM 神经网络中的输出隐藏状态  $h_t = [\vec{h}_t; \overleftarrow{h}_t]$ , 其中  $\vec{h}_t$  和  $\overleftarrow{h}_t$  分别为前向输出和后向输出.

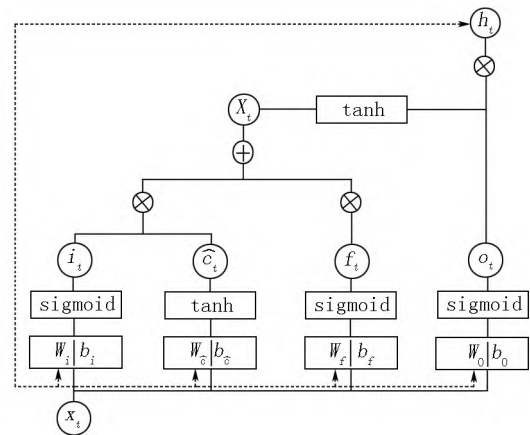


图3 LSTM 单元结构图

## 2.4 自注意力层

自注意力层的功能是给予上下文的局部信息,使模型加强对重要信息的捕捉,减少非必要信息的噪声影响. 它将重点放在序列的特定部分,同时不丢弃编码器状态的中间值,充分利用它生成上下文向量,以便解码器给出输出结果. 其机制公式为

$$H = \text{softmax}(W_2 \tanh(W_1 h_t)),$$

其中  $W_1, W_2$  为权重参数  $h_t$  为 BiLSTM 层输出的隐藏状态.

## 2.5 BCRF 层

原有命名实体识别模型解码层主要是条件随机场(conditional random field, CRF). CRF 是由状态

特征函数和状态特征转移函数组成, 状态特征函数也被称为发射概率, 状态特征转移函数在模型中可以用一个状态转移矩阵表示, 最后得到的条件概率为

$$p(y|x) = \exp(w\varphi(x, y)) / \left( \sum_{y \in Y^n} \exp(w\varphi(x, y)) \right),$$

其中  $\varphi(x, y)$  为  $x$  和  $y$  一组特征向量的映射.  $p(y|x)$  为模型在给定文本序列  $x$  条件下得到标签序列  $y$  的概率. 损失函数计算公式为

$$L(w, x) = - \sum_i \lg p(y^{(i)} | x^{(i)}; w).$$

CRF 方法的优点是可以进一步考虑序列标签的依赖关系. 同时, 在训练过程中, 采用 Viterbi 算法进行最大似然估计, 使模型对输入文本预测出标签的最大概率为  $\hat{y} = \operatorname{argmax}_y p(y | x^{(i)}; w)$ , 其中  $\hat{y}$  表示模型预测标签的最大概率. 但是在低资源的场景下, 由于受标注数据量少的限制, 模型的参数没有拟合到预期效果, 所以输出预测概率最大的标签序列并不一定是真实的标签序列, 导致模型最后的识别性能下降. 因此, 本文借鉴 Jie Zhanming 等<sup>[25]</sup> 在不完全标注实体识别任务上采用交叉验证方式训练数据的方法, 将伯努利分布融入条件随机场损失函数中构建新的损失函数, 称对应的解码模型为 BCRF.

伯努利分布 (Bernoulli distribution) 又被称为 2 点分布或 0-1 分布, 它是指对于随机变量  $X$ , 概率为  $P$  ( $0 < P < 1$ ), 当  $P = 1$  或  $1 - P = 0$  时, 该分布是一个离散型概率分布. 是在  $N = 1$  时二项分布的特例. 将伯努利分布函数融入 CRF 中, 构建新的损失函数为

$$L(w, x) = - \sum_i \lg \sum q(y | x^{(i)}; w) p(y | x^{(i)}; w). \quad (1)$$

在原有的损失函数计算公式中加入一个分布函数  $q$ , 得计算公式为

$$q(y | x^{(i)}; w) = \begin{cases} 1 & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$$

分布函数  $q$  取值为 0 或 1, 呈伯努利分布. 其中  $\hat{y}$  为模型预测最大概率的标签, 表示真实标签. 在式 (1) 中, 当句子中词预测标签与真实标签一致时, 得到的损失值也就较小. 相反, 若在一个句子中预测标签错误越多, 则其损失值也越大. 采用新的损失函数公式进行计算, 当在只有少量训练样本时, 模型参数能较好地拟合, 提升了模型的精确率和召回率, 提高了模型的识别效果.

## 3 实验

### 3.1 数据集及评价指标

本文选择 CONLL2003<sup>[26]</sup> 数据集和 BC5CDR<sup>[27]</sup> 数据集来验证所提出模型的有效性. CONLL2003 数据集包含 4 种实体类型以及英语和德语 2 种语言, BC5CDR 数据集包含 2 种实体和 1 500 篇医药文章. 由于实体识别的任务主要是对实体的边界和类别的识别, 所以只有当边界及实体的类别都被正确识别时, 才判断正确. 本文通过使用精确率 (precision) 和召回率 (recall) 来求得  $F_1$  值, 用于衡量该模型的性能, 其计算方法分别为

$$\text{精确率 } r_p = M/N, \text{ 召回率 } r_e = M/K, F_1 = 2r_p r_e / (r_p + r_e) \times 100\%,$$

其中  $N$  为模型所预测出的实体总数,  $M$  为在模型预测的实体中正确预测实体的总数,  $K$  为在数据集中所标注的实体总数. 在模型中超参数的设置如表 1 所示.

表 1 超参数设置

参数	取值
隐层向量维度	200
词向量维度	100
字符向量维度	50
Dropout	0.5
学习率	0.1
批尺寸	10
训练轮数	100
$L_2$ 正则化	$1 \times 10^{-8}$

### 3.2 实验结果与分析

实验采用的数据集主要是 CONLL-2003 英语数据集和 BC5CDR 专业医学领域数据集. BiLSTM-Att-BCRF 模型、BiLSTM-BCRF 模型和 BiLSTM-CRF 模型实验结果如图 4 和图 5 所示.

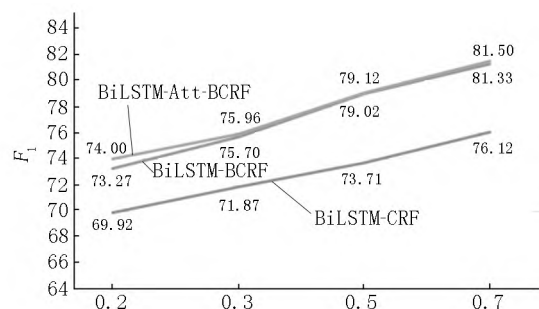


图 4 在 BC5CDR 数据集上 BiLSTM-Att-BCRF 模型、BiLSTM-BCRF 模型与 BiLSTM-CRF 模型实验结果对比

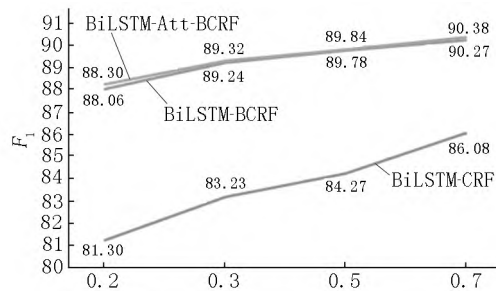


图5 在 CONLL2003 数据集上 BiLSTM-Att-BCRF 模型、BiLSTM-BCRF 模型与 BiLSTM-CRF 模型实验结果对比

从图4和图5可以看出:该模型在少量标注数据集上的性能高于 BiLSTM-CRF 模型的性能,更适合于低资源领域的实体识别任务.同时该模型在20%的 CONLL2003 数据集上  $F_1$  值达到了 88.30%,这说明该模型在不需要大量的标注语料的情况下也能取得比较好的识别效果.

TMN 是 L. Y. Bill 等<sup>[28]</sup>提出的一种基于实体和触发词标注的命名实体识别模型, DualNER 是 ZHONG Maosheng 等<sup>[29]</sup>提出的一种基于对偶学习和触发词标注的命名实体识别模型. BiLSTM-Att-BCRF 模型与 DualNER 模型和 TMN 模型实验对比结果如表2和表3所示.

表2 BiLSTM-Att-BCRF 模型与其他模型实验结果对比

模型	20% CONLL2003		
	精确率/%	召回率/%	$F_1$
BiLSTM-CRF	82.17	80.35	81.30
BiLSTM-Att-CRF	85.50	80.81	83.09
TMN	85.65	85.38	85.51
DualNER	86.55	86.69	86.62
BiLSTM-BCRF	87.49	88.63	88.06
BiLSTM-Att-BCRF	88.28	88.31	88.30

从表2可以看出:在20%的 CONLL-2003 数据集上, BiLSTM-Att-BCRF 模型的  $F_1$  值高于 DualNER 模型和 TMN 模型的  $F_1$  值.

表3 BiLSTM-Att-BCRF 模型与其他模型实验结果对比

模型	20% BC5CDR		
	精确率/%	召回率/%	$F_1$
BiLSTM-CRF	79.09	62.66	69.92
BiLSTM-Att-CRF	73.63	69.15	71.32
TMN	74.30	72.44	73.36
DualNER	76.06	73.66	74.84
BiLSTM-BCRF	74.35	72.22	73.27
BiLSTM-Att-BCRF	75.45	72.60	74.00

如表3所示,在20%的 BC5CDR 数据集上, BiLSTM-Att-BCRF 模型  $F_1$  值比 TMN 模型的高 0.64%,比 DualNER 模型的低 0.84%,这主要是该模型识别一些专有名词的性能略低于 DualNER 模型所致.但是, BiLSTM-Att-BCRF 模型并不需要标注触发词,它所需要的人工成本只是 TMN 模型和 DualNER 模型的 3/4,同时 BiLSTM-Att-BCRF 模型在 CONLL2003 数据集上的  $F_1$  值比 DualNER 模型的高 1.68%.然而,如何提高 BiLSTM-Att-BCRF 模型识别专有名词的性能,是本文后续工作的重点.

### 3.3 词与多层字符信息的融合分析

图2提出了模型输入层结构,为验证在模型输入层中哪些因素会影响模型进行命名实体识别的性能,本文在 BiLSTM-BCRF 模型上进行了如下实验.

在表4中,实验所使用的模型为 BiLSTM-BCRF 模型,采用的数据集为20%的 BC5CDR 数据集.其中 word 为词向量信息,维度设置为 100, char 为字符向量,维度设置为 50, “char\* 2”为2个字符向量矩阵拼接,“+”表示拼接.

表4 词向量与字符向量拼接顺序对模型性能的影响

方法	20% BC5CDR 数据集		
	精确率/%	召回率/%	$F_1$
word + char	71.82	72.50	72.16
char + word	71.23	73.07	72.14
BiLSTM-BCRF (word + char* 2)	74.35	72.22	73.27
char + word + char	72.25	71.73	71.99

为探究词向量与字符向量拼接顺序是否会影响模型的性能,在模型输入层采用词向量和一个字符向量矩阵拼接时,将它们的拼接顺序调换,模型  $F_1$  值与之前相比略有下降.当词向量与2个字符向量矩阵拼接并将词向量放至2个字符向量矩阵中间时,模型的精确率、召回率、 $F_1$  值与之前相比都有所下降.由此可知,词向量与字符向量的拼接顺序会影响模型的性能.

在表5中,实验所使用的模型为 BiLSTM-BCRF 模型,采用的数据集为20%的 BC5CDR 数据集.其中 word 为词向量信息,维度设置为 100, “word (200)”表示维度为200的词向量, char 为字符向量,维度设置为 50, “char\* 3”为3个字符向量矩阵拼接,“+”表示拼接.

在低资源场景下,可以通过在词向量后拼接字

符向量来提高模型处理罕见词的能力,以提高模型的识别性能.为探究在拼接字符向量矩阵数量为多少时模型的识别性能提升最大,本文做了以下实验:首先将词向量维度设置为 100,当拼接 1 个字符向量矩阵、2 个字符向量矩阵、3 个字符向量矩阵时,通过实验结果对比,发现在词向量后拼接 2 个字符向量矩阵情况下,模型识别效果最好.然后将词向量维度设置为 50、200、300,再拼接不同数量的字符向量矩阵.通过实验结果对比,发现将词向量维度设置为 100,拼接 2 个字符向量矩阵,模型的识别性能提升的最多(见表 5).

3.4 消融实验

为探究伯努利分布和多层字符信息以及自注意力机制对模型性能的影响,将 BiLSTM-CRF 模型设置为基准模型.首先将伯努利分布融入基准模型的损失函数中,构建 BiLSTM-BCRF(1)模型;然后在

BiLSTM-CRF(1)模型基础上融合多层字符信息,构建 BiLSTM-BCRF 模型;最后在 BiLSTM-BCRF 模型基础上融合自注意力机制,构建了 BiLSTM-Att-BCRF 模型.以上模型在 2 个数据集的实验结果如表 6 所示.

表 5 词向量拼接字符向量矩阵的个数对模型性能的影响

方法	20% BC5CDR 数据集		
	精确率/%	召回率/%	$F_1$
BiLSTM-BCRF ( word + char* 2)	74.35	72.22	73.27
word( 200) + char* 4	73.12	73.20	73.16
word( 300) + char* 6	70.74	73.87	72.27
word( 50) + char	70.42	70.70	70.56
word + char	71.82	72.50	72.16
word + char* 3	71.07	74.21	72.60
( word + char* 2) * 2	73.48	71.37	72.41

表 6 消融实验结果

方法	20% CONLL2003			20% BC5CDR		
	精确率/%	召回率/%	$F_1$	精确率/%	召回率/%	$F_1$
BiLSTM-CRF	82.17	80.35	81.3	79.09	62.66	69.92
BiLSTM-BCRF( 1)	86.43	87.75	87.08	73.77	71.98	72.87
BiLSTM-BCRF	87.49	88.83	88.06	74.35	72.22	73.27
BiLSTM-Att-BCRF	88.28	88.31	88.30	75.45	72.60	74.00

注: BiLSTM-CRF 模型为基准模型, BiLSTM-BCRF( 1) 模型表示将伯努利分布融入 BiLSTM-CRF 模型损失函数中.

表 6 消融实验结果表明: BiLSTM-BCRF( 1) 模型在 2 个数据集上取得的  $F_1$  值均高于 BiLSTM-CRF 模型的  $F_1$  值,这表明新型损失函数对于模型的性能具有提升作用.从表 6 中的模型 BiLSTM-Att-BCRF 在 2 个数据集取得的精确率、召回率、 $F_1$  值可以看出本文提出的 BiLSTM-Att-BCRF 模型识别效果是比較好的.

3.5 定性分析

为更好地对比 BiLSTM-Att-BCRF 模型与 BiLSTM-CRF 模型在命名实体识别任务上的差异,本文从数据集中选取 2 个实例句子: “OnlyFranceand-Britain backed Fischler’sproposal.” 和 “Rare Hendrix song draft sellsforalmost \$ 1 700”,人工标注、BiLSTM-CRF 模型和 BiLSTM-Att-BCRF 模型的标注结果如表 7 和表 8 所示.

表 7 模型识别实例 1

	Only	France	and	Britain	backed	Fischler	’s	proposal	.
人工标注	O	B-LOC	O	B-LOC	O	B-PER	O	O	O
BiLSTM-CRF	O	B-LOC	O	B-LOC	O	O	O	O	O
BiLSTM-Att-BCRF	O	B-LOC	O	B-LOC	O	B-PER	O	O	O

表 8 模型识别实例 2

	Rare	Hendrix	song	draft	sells	for	almost	\$	1 700
人工标注	O	B-PER	O	O	O	O	O	O	O
BiLSTM-CRF	B-PER	I-PER	O	O	O	O	O	O	O
BiLSTM-Att-BCRF	O	B-PER	O	O	O	O	O	O	O

从表 7 可以看出该句子共有 3 个实体,原来的 BiLSTM-CRF 模型只识别出了 2 个实体,本文提出

的 BiLSTM-Att-BCRF 模型将句中包含的 3 个实体全部识别出来了. 在表 8 中, 该句子只有 1 个人名实体, 但 BiLSTM-CRF 模型把句子中前 2 个单词错误地识别为 1 个人名实体, 而本文提出的模型能将句中人名实体准确地识别出来.

## 4 总结

针对低资源领域标注语料较少的问题, 本文提出了一种低资源场景下命名实体识别模型 (BiLSTM-Att-BCRF). 该模型通过在损失函数中融入伯努利分布, 使模型参数在低资源场景下也能得到较好地拟合. 同时将多层字符特征信息融合到模型中, 提升了模型处理罕见词的能力, 使模型在标注数据量少时也有较好的识别性能, 能更好地适用于低资源命名实体识别任务. 但该模型识别专有名词的性能有待提升, 同时模型的知识迁移和跨领域的性能提升也是本文以后的研究重点.

## 5 参考文献

- [1] HAMMERTON J. Named entity recognition with long short-term memory [EB/OL]. [2022-04-03]. [http://dl.acm.org/ft\\_gateway.cfm?id=1119202&type=pdf](http://dl.acm.org/ft_gateway.cfm?id=1119202&type=pdf).
- [2] GUILLAUME L, Miguel B, Sandeep S, et al. Neural architectures for named entity recognition [EB/OL]. [2022-04-05]. <https://arxiv.org/pdf/1603.01360.pdf>.
- [3] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2022-04-02]. <https://arxiv.org/abs/1508.01991>.
- [4] 殷章志, 李欣子, 黄德根, 等. 融合字词模型的中文命名实体识别研究 [J]. 中文信息学报, 2019, 33(11): 95-100, 106.
- [5] 林广和, 张绍武, 林鸿飞. 基于细粒度词表示的命名实体识别研究 [J]. 中文信息学报, 2018, 32(11): 62-71, 78.
- [6] RATHAPARKHI A. A maximum entropy part of speech tagger [EB/OL]. [2022-04-06]. <https://www.mendeley.com/catalogue/51884aa1-bab9-3acc-8e85-51606f7087ce/>.
- [7] MCCALLUM A, FREITAG D, Pereira F C N. Maximum entropy Markov models for information extraction and segmentation [EB/OL]. [2022-04-02]. <https://dl.acm.org/doi/10.5555/645529.658277>.
- [8] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [EB/OL]. [2022-04-07]. <https://dl.acm.org/doi/10.5555/645530.655813>.
- [9] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.
- [10] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs [EB/OL]. [2022-04-09]. <https://arxiv.org/pdf/1511.08308.pdf>.
- [11] NI Jian, GEORGIANA D, RADU F. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection [EB/OL]. [2022-04-02]. <https://arxiv.org/abs/1707.02483>.
- [12] MAYHEW S, TSAI C T, ROTH D. Cheap translation for cross-lingual named entity recognition [EB/OL]. [2022-04-05]. <https://aclanthology.org/D17-1268.pdf>.
- [13] TOOLKIT O S, HOC B, NG M, et al. Moses: open source toolkit for statistical machine translation [EB/OL]. [2022-04-02]. <https://dl.acm.org/doi/10.5555/1557769.1557821>.
- [14] CHEN Xilun, AHMED H A, HANYH, et al. Multi-source cross-lingual model transfer: learning what to share [EB/OL]. [2022-04-04]. <https://aclanthology.org/P19-1299/>.
- [15] PHILLIP K, LU Yichao, VIKAS B. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER [EB/OL]. [2022-04-02]. <https://arxiv.org/abs/1909.00153v3>.
- [16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [EB/OL]. [2022-04-02]. <http://de.arxiv.org/pdf/1406.2661>.
- [17] DAI Xiang, ADELH. An analysis of simple data augmentation for named entity recognition [EB/OL]. [2022-04-04]. <https://arxiv.org/abs/2010.11683>.
- [18] CHEN Jiaao, WANG Zhenghui, TIAN Ran, et al. Local additivity-based data augmentation for semi-supervised NER [EB/OL]. [2022-04-06]. <https://arxiv.org/abs/2010.01677>.
- [19] YANG Yaosheng, CHEN Wenliang, LI Zhenghua, et al. Distantly supervised NER with partial annotation learning and reinforcement learning [EB/OL]. [2022-04-07]. <https://aclanthology.org/C18-1183.pdf>.
- [20] TSUBOI Y, KASHIMA H, MORI S, et al. Training conditional random fields using incomplete annotations [EB/OL]. [2022-04-08]. [http://dl.acm.org/ft\\_gateway.cfm?id=1599194&type=pdf](http://dl.acm.org/ft_gateway.cfm?id=1599194&type=pdf).
- [21] FENG Jun, HUANG Minlie, ZHAO Li, et al. Reinforcement learning for relation classification from noisy data

- [EB/OL]. [2022-04-04]. <https://arxiv.org/pdf/1808.08013.pdf>.
- [22] ZHANG Tao ,XIA Congying ,PHILIP S Y ,et al. PDALN: progressive domain adaptation over a pre-trained model for low-resource cross-domain named entity recognition [EB/OL]. [2022-04-04]. <https://aclanthology.org/2021.emnlp-main.442/>.
- [23] CHEN Siqi ,PEI Yijie ,KE Zunwang ,et al. Low-resource named entity recognition via the pre-training model [J]. Symmetry 2021 ,13( 5) : 786.
- [24] PENNINGTON J ,SOCHER R ,MANNING C D. Glove: global vectors for word representation [EB/OL]. [2022-04-05]. <https://aclanthology.org/D14-1162/>.
- [25] JIE Zhanming ,XIE Pengjun ,LU Wei ,et al. Better modeling of incomplete annotations for named entity recognition [EB/OL]. [2022-04-02]. <https://www.xueshufan.com/publication/2945214158>.
- [26] ERIK F T ,Fien D M. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition [EB/OL]. [2022-04-04]. <https://aclanthology.org/W03-0419/>.
- [27] LI Jiao ,SUN Yueping ,JOHNSON R J ,et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction [EB/OL]. [2022-04-02]. <https://pubmed.ncbi.nlm.nih.gov/27161011/>.
- [28] BILL Y L ,LEE D H ,SHEN Ming ,et al. TriggerNER: learning with Entity triggers as explanations for Named Entity recognition [EB/OL]. [2022-04-03]. <https://github.com/INK-USC/TriggerNER>.
- [29] ZHONG Maosheng ,LIU Ganlin ,XIONG Jian ,et al. Dual-NER: a trigger based dual learning framework for low-resource Named Entity recognition [J]. IEEE Intelligent Systems 2022 ,37( 4) : 79-87.

## The BiLSTM-Att-BCRF Model for Low Resource Named Entity Recognition

ZHONG Maosheng ,WU Jiahua ,LUO Wei ,WU Shuixiu

( School of Computer and Information Engineering ,Jiangxi Normal University ,Nanchang Jiangxi 330022 ,China)

**Abstract:** In low-resource scenarios ,the existing models are limited by the small amount of training data ,and the parameters are not fitted to the expected effect ,resulting in poor performance of the model in recognizing entities. In this paper ,a new loss function incorporating Bernoulli distribution is proposed to allow the model to fit the data better. In addition ,a BiLSTM-Att-BCRF model based on the BiLSTM-CRF model is constructed by fusing multi-layer character feature information and self-attention mechanism ,combined with a novel loss function based on Bernoulli distribution. The BiLSTM-Att-BCRF model proposed in this paper improves the  $F_1$  values by 7.00% and 4.08% based on the BiLSTM-CRF model on the datasets of 20% CoNLL2003 and 20% BC5CDR ,respectively. The model is better adapted to low resource named entity recognition tasks.

**Key words:** low resource named entity recognition; neural network; Bernoulli distribution; self-attention mechanism

( 责任编辑: 冉小晓)