doi:10.3969/j.issn.1673-9833.2019.05.013

基于 Attention-BiLSTM 的中文命名实体识别

冀相冰[1,2]，朱艳辉[1,2]，李 飞[1,2]，徐 啸[1,2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；
2. 智能信息感知及处理技术湖南省重点实验室，湖南 株洲 412007）

摘 要：提出一种基于 Attention-BiLSTM（attention-bidirectional long short-term memory）深度神经网络的命名实体识别方法。应用 BiLSTM 神经网络自动学习文本的隐含特征，可以解决传统识别方法存在长距离依赖等问题；引入注意力机制（attention mechanism）对文本全局特征做重要度计算，获取文本局部特征，解决了传统深度学习方法不能充分提取特征的问题；在预训练过程中加入维基百科知识，进一步提升了命名实体识别系统的性能。实验表明，所提方法在 SIGHAN 2006 Bakeoff-3 评测数据集上获得了优良的识别性能。

关键词：命名实体识别；注意力机制；BiLSTM；深度学习；局部特征

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2019)05-0073-06

引文格式：冀相冰，朱艳辉，李 飞，等 . 基于 Attention-BiLSTM 的中文命名实体识别 [J]. 湖南工业大学学报，2019，33(5)：73-78.

Entity Recognition of Chinese Names Based on Attention-BiLSTM

JI Xiangbing[1,2]，ZHU Yanhui[1,2]，LI Fei[1,2]，XU Xiao[1,2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. Hunan Key Laboratory of Intelligent Information Perception and Processing Technology，Zhuzhou Hunan 412007，China）

Abstract：This paper proposes a named entity recognition method based on Attention-BiLSTM (attention-bidirectional long short-term memory) deep neural network. Using BiLSTM neural network to automatically learn the implicit features of text can solve the problem of long-distance dependence of traditional recognition methods. Attention mechanism is used to calculate the importance of text global features, obtain local features of text, and solve the traditional deep learning method can not fully extract the feature problem; adding Wikipedia knowledge in the pre-training process further improves the performance of the named entity recognition system. Experiments show that the proposed method achieves excellent recognition performance on the SIGHAN 2006 Bakeoff-3 evaluation data set.

Keywords：named entity recognition；attention mechanism；BiLSTM；deep learning；local feature

1 研究背景

命名实体识别（named entity recognition，NER）是自然语言处理任务中关键的步骤之一，它的主要任务是识别非结构化文本数据中的具有特定含义的实体（地名、机构、事件、专用名词等），在事件检测、

收稿日期：2018-03-28

基金项目：国家自然科学基金资助项目（61402165），湖南省自然科学基金资助项目（2018JJ2098），湖南工业大学重点基金资助项目（17ZBLWT001KT006）

作者简介：冀相冰（1992-），男，山东聊城人，湖南工业大学硕士生，主要研究方向为自然语言处理与知识工程，E-mail：jxiangbing@163.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要从事自然语言处理与知识工程的教学与研究，E-mail：swayhzhu@163.com

智能客服等领域有非常广泛的应用。中文命名实体种类繁多，覆盖领域较广，从海量的文本中自动抽取出人们所需要的价值信息可以满足各行业需求，因此具有重大意义。

针对命名实体识别任务中存在的问题，学者们已经进行了许多深入的研究。M. Sundermeyer 等 [1] 通过 LSTM（long short-term memory）神经网络进行语言建模，根据困惑性和单词错误率构建评估模型并验证 2 个量的强相关性。Duan H. Z. 等 [2] 利用常用属性、窗口大小和序列标签设置不同的特征模板的功能组合，提高了中文命名实体识别的性能。A. Borthwick[3] 把最大熵统计模型应用到命名实体识别中，该研究对信息提取任务具有特殊意义。R. Collobert 等 [4] 提出了一种统一的神经网络和学习算法，该算法根据传统人工标注的特点，对大量未标注的训练数据学习内部隐含特征，取得了较好的效果。N. Greenberg 等 [5] 利用 BiLSTM+CRF 方法在多个生物医学数据集上联合抽取训练，获得了较优效果。Liu X. H. 等 [6] 通过研究推特领域的命名实体识别，提出进行两阶段标签来利用类似推文中的冗余信息，取得了较好的 F 值。Feng Y. H. 等 [7] 等利用词嵌入 + CRF 的领域专用术语识别方法，增加了词嵌入和术语嵌入的相似性，形成了特征向量，解决了传统方法忽视语义的问题。Wang G. Y. 等 [8] 利用一种混合深度神经网络（deep neural networks，DNN）进行实验，挖掘了嵌入在无标签语料库中的隐式信息，其实验效果比条件随机场模型要好。J. Mayfield 等 [9] 使用支持向量机（support vector machine，SVM）训练数据的特征，利用简单的静态函数将边际产出转换为估计得概率，对英文和德文进行了识别。D. Klein 等 [10] 讨论了字符级 HMM（hidden Markov model）和最大熵条件马尔可夫模型，他们在英语测试数据集上取得了良好的效果。D. Bahdanau 等 [11] 在神经网络中，引入注意力机制，解决了自然语言处理领域的机器翻译问题。

传统深度学习在 NER 提取特征过程中，过于重视文本全局特征，因而忽视了局部特征对命名实体识别的重要影响。一段文本中命名实体的识别可能仅与局部信息有关，且每个字词对其他实体的贡献程度不同，过多的冗余信息只会对命名实体识别带来负面影响。本文提出基于 Attention−BiLSTM−CNN−CRF 的深度神经网络模型并以其进行中文命名实体识别。首先对语料集进行预训练词向量，利用卷积神经网络（convolutional neural network，CNN）提取句子中的字符表示向量，并将字嵌入向量和字符表示向量联合起来馈送到 BiLSTM 神经网络中；然后利用注意力机制在文本全局特征上获取局部特征；最后根据文本的全局特征和局部特征使用 CRF 解码整个句子的最优标注序列。

2 基于 Attention−BiLSTM 的中文命名实体识别模型

2.1 Attention−BiLSTM 模型

循环神经网络（recurrent neural network，RNN）是一个强大的连接模型族，从理论上来说，RNN 可以捕获长距离的依赖 [12]，但是在实际情况中却容易出现梯度爆炸或消失问题，长短期记忆网络（long short-term memory，LSTM）[13] 可以很好地解决这类问题。本研究提出了基于注意力机制的 BiLSTM 模型，以机构命名实体“篮球管理中心”作为实例进行表示，如图 1 所示。

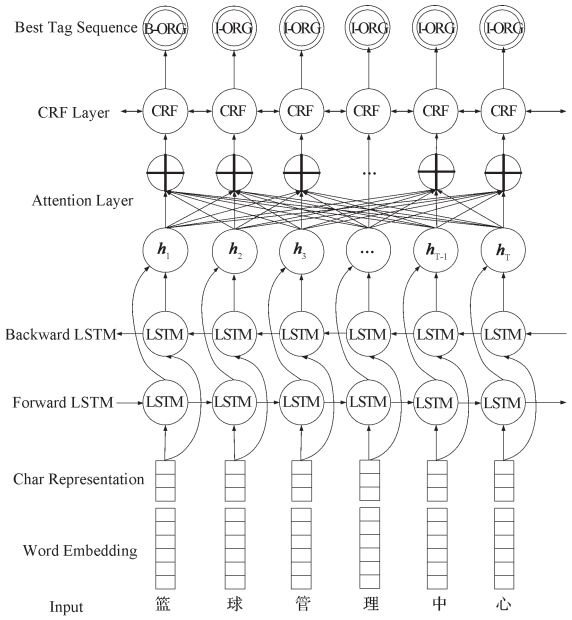

图 1 Attention−BiLSTM 模型结构

Fig.1 Attention−BiLSTM model structure

LSTM 由 3 个门组成，通过联结存储单元，利用几个门控制早期状态中需要忘记信息的比例和输入存储单元的比例，从而能够捕获长距离依赖。LSTM 强大的性能可以解决很多序列标签标注的任务，因为它可以访问文本过去和未来的情况。

它的基本思路是将所有序列展开为 2 个单独的隐藏状态，其中一个向前捕获历史的信息，另外一个向后捕获未来的信息，随后将前后两个隐藏状态联结起来，对上下文信息进行标签标记，形成全局特征输出。然后将 BiLSTM 的输出向量输入 Attention 层进一步进行局部特征提取，最后将全局特征和局部特征一起馈送到 CRF 层。

其中，所有字符都被映射到随机初始化的（预先训练的）字嵌入向量。随后字嵌入向量经过 CNN 计算，提取形成字符表示向量和字嵌入向量的联合特征表示，其过程如图 2 所示，然后它们被馈送到前向 LSTM 网络和后向 LSTM 网络。

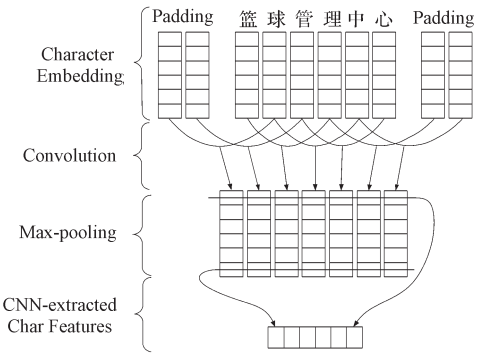

图 2　CNN 结构图

Fig. 2　CNN structure

BiLSTM 层输入的是一个综合字嵌入和字符表示联合表达的向量序列，表示为 $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$。BiLSTM 层输出为输入词向量的一系列最终隐藏状态 $\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n$，表示前向$\overleftarrow{\boldsymbol{h}_i}$隐藏状态和后向$\overrightarrow{\boldsymbol{h}_i}$隐藏状态的联合状态，其公式如下：

$$\boldsymbol{h}_i = \left[\overleftarrow{\boldsymbol{h}_i}; \overrightarrow{\boldsymbol{h}_i}\right]。\tag{1}$$

BiLSTM 网络在 i 时刻的输入状态实现：

$$\boldsymbol{x}_i = \boldsymbol{v}_{\text{word}}。\tag{2}$$

式中 $\boldsymbol{v}_{\text{word}}$ 表示字词向量。

BiLSTM 网络在 i 时刻的输出状态实现：

$$t_{t_i} = f(\boldsymbol{w}_t \boldsymbol{h}_i + b_t);\tag{3}$$

$$t_{d_i} = f(\boldsymbol{w}_d \boldsymbol{h}_i + b_d);\tag{4}$$

$$t_{e_i} = f(\boldsymbol{w}_e \boldsymbol{h}_i + b_e);\tag{5}$$

$$t_{m_i} = f(\boldsymbol{w}_m \boldsymbol{h}_i + b_m);\tag{6}$$

$$t_{p_i} = f(\boldsymbol{w}_p \boldsymbol{h}_i + b_p)。\tag{7}$$

式（3）~（7）中：f 为 softmax 函数；

t_{t_i}、t_{d_i}、t_{e_i}、t_{m_i} 和 t_{p_i} 为模型中各神经元在 i 时刻的预测值；

$\boldsymbol{w}_t$、$\boldsymbol{w}_d$、$\boldsymbol{w}_e$、$\boldsymbol{w}_m$ 和 $\boldsymbol{w}_p$ 为隐含状态的映射矩阵；

b_t、b_d、b_e、b_m 和 b_p 为隐含状态的偏置。

2.2　注意力机制

注意力机制是从众多信息中选择对当前任务目标更关键的信息，然后对需要重点关注的目标区域投入更多的注意力资源。在 BiLSTM 神经网络中加入注意力机制，选择性地对文本中不同文字赋予不同的权重，再利用基于上下文的语义关联信息可以有效弥补深度神经网络获取局部特征方面的不足。文本局部特征表示文本中部分内容之间的关联特征。例如在句子“在有英格兰队比赛的近两小时电视转播时段”中，“英格兰队”是一个命名实体，各个字之间的关联更加密切，而实体前面的“有”和后面的“比”与它的关联较弱。因为各个字符对命名实体的影响程度不一样，所以可以为它们分配不同的权重。

基于注意力机制的神经网络模型如图 3 所示。

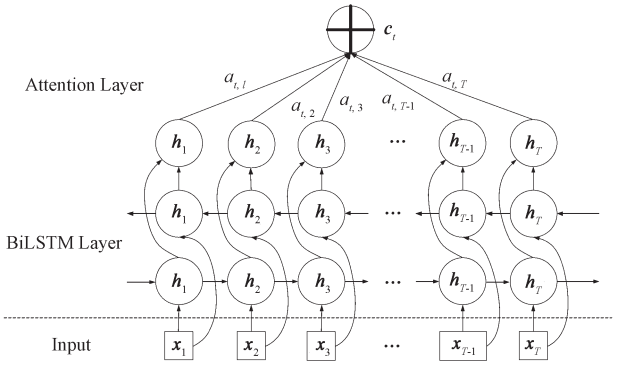

图 3　基于注意力机制的神经网络模型

Fig. 3　Neural network model based on attention mechanism

利用该机制进行特征提取的过程如下：首先，把字嵌入和字符表示联合向量序列输入 BiLSTM 网络提取全局特征；然后，通过 Attention 机制给全局特征中不同的特征向量赋予不同的权重，以提取局部特征 [14]；最后，生成包括全局特征和局部特征的联合特征向量序列。

定义 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T$ 为 BiLSTM 层输入的字嵌入和字符表示联合向量序列；a_{tj} 为 Attention 机制给所有特征向量赋予的权重，计算公式如下：

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})},\tag{8}$$

式中：e_{ij} 为输入和输出之间的关联能量；

$$e_{ij} = \boldsymbol{v}_a^{\mathrm{T}} \tanh(w_a \boldsymbol{c}_{t-1} + u_a \boldsymbol{h}_j),\tag{9}$$

其中，$\boldsymbol{v}_a$ 为全局权值向量；

$\boldsymbol{c}_{t-1}$ 为注意力模型的上一时刻状态向量；

$\boldsymbol{h}_i$ 为 BiLSTM 层输出的特征向量序列；

u_a 为上一时刻特征向量的权值；

w_a 为注意力机制上一时刻的权值。

注意力机制最后的输出状态为 $\boldsymbol{c}_t$，其公式为

$$\boldsymbol{c}_t = \sum_{j=1}^{T} a_{tj} \boldsymbol{h}_j。\tag{10}$$

2.3　标记预测

中文命名实体识别是一个多标签分类问题，可以利用条件概率　进行描述 [15]，如下式所示：

$$p(l \mid x,\theta)=\frac{e^{f(x,l,\theta)}}{\sum_{j} e^{f(x,j,\theta)}}。\quad (11)$$

它表示字 x 被标记为 l 标签的概率，其中，j 为相邻标签，θ 为利用随机梯度下降算法进行训练之后所得到的参数。

命名实体识别与其他序列标注任务一样，需要依据标记路径的得分对标注结果进行判断。语料中的相邻标签之间存在特殊关联，例如人名的开始标签不会连接组织的中间标签，这种依存关系叫做标签之间的转移概率。因此，句子的序列标注问题就可以转化为寻找最优路径问题。

语料的标记路径得分由 $f(x_{[1:T]}, l, t, \theta)$ 和 A_{ijk} 组成，$f(x_{[1:T]}, l, t, \theta)$ 为 t 时刻句子 $x_{[1:T]}$ 被标注为 l 的得分，A_{ijk} 为标记 k 与相邻标记 (i, j) 之间的关系表达。语料集中句子标注的整体得分情况为

$$S\left(x_{[1:T]}, l_{[1:T]}, \tilde{\theta}\right)=\sum_{t=1}^{T}\left(A_{(lt)-(2lt)+(1lt)}+f(x_{[1:T]}, l_{[t]}, t, \theta)\right)。\quad (12)$$

式中 $\tilde{\theta}$ 表示 A_{ijk} 和 θ 构成的参数集，在训练时，使用 log-add 方法计算所有的训练样本，把 $\sum_{x,y\in T}\log p(y_{[1:T]} \mid x_{[1:T]}, \tilde{\theta})$ 最大化，表达式如下：

$$\log p\left(y_{[1:T]} \mid x_{[1:T]},\ \tilde{\theta}\right)=S\left(x_{[1:T]}, y_{[1:T]},\ \tilde{\theta}\right)-\log\operatorname*{add}_{\forall l_{[1:T]}} S\left(x_{[1:T]}, l_{[1:T]},\ \tilde{\theta}\right)。\quad (13)$$

最后，采用 Viterbi 算法计算句子中最优标注序列 y^{*}，最终结果如下：

$$y^{*}=\arg\max_{l_{[1:T]}} S\left(x_{[1:T]},\ l_{[1:T]},\ \tilde{\theta}\right)。\quad (14)$$

3 实验与分析

3.1 数据集及标注模式

为了验证本研究所提的方法在 SIGHAN 2006 Bakeoff-3 语料集上的识别效果，课题组将实验数据集按照 6:2:2 的比例进行划分，如表 1 所列。同时在维基百科上面抽取 1.44 GB 的知识语料库融入 SIGHAN 数据集中进行训练。

表 1 实验数据集划分

Table 1 Experimental data set division

类别	训练集	验证集	测试集
数据量 /MB	6.18	1.53	1.55

SIGHAN 2006 Bakeoff-3 语料中命名实体的类别如表 2 所列。

表 2 实体类别及编码

Table 2 Entity category

实体类别	人名	地名	组织机构
标注编码	PER	LOC	ORG

实验采用的标注模式为 BIO 模式，B 表示实体的起始点，I 表示实体的中间部分，O 表示非实体。例如：组织机构相关实体标注为 B-ORG/I-ORG。

3.2 实验环境配置

实验软硬件环境配置如表 3 所示。

表 3 软硬件环境

Table 3 Hardware and software description

项目	环境
系统	Ubuntu16.04 LTS
GPU	NVIDIA Quadro K1200
硬盘	1 TB
内存	16 GB
Python 版本	Python3
TensorFlow 版本	TensorFlow1.2.1
分词系统	NLPIR 中文分词系统

3.3 评价指标

评价指标体系采用准确率（P）、召回率（R）和 F 值的方法，具体公式如下：

$$P=\frac{correct}{correct+missing},\quad (15)$$

$$R=\frac{correct}{correct+spurious},\quad (16)$$

$$F=\frac{2PR}{P+R}。\quad (17)$$

式（15）~（17）中：P 为系统正确标注的实体占系统识别到的实体总量的比值；

correct 为正确标注的实体数；

missing 为识别错误的实体数；

spurious 为未被识别到的正确实体数；

R 为正确标注的实体占测试集实体总量的比值；

F 为 P 和 R 的加权几何平均值。

3.4 BiLSTM 参数获取实验

课题组在开发集上进行了多组实验来选择最优参数，其中学习率和隐藏节点的实验结果如图 4 和 5 所示。从图 4 和 5 的实验结果可以观察到，当学习率和隐藏节点数分别达 0.002 和 200 时，F 值取得最好的结果；当学习率与隐藏节点数再增加或减少时，F 值有不变或者降低的趋势。因此本模型将最优学习率设置为 0.002，最优隐藏节点数设置为 200。利用以上实验方法可以获得其他参数的最优值，如表 4 所示。为了避免过拟合问题，采用了 L2 正则化算法和

Dropout 技术。

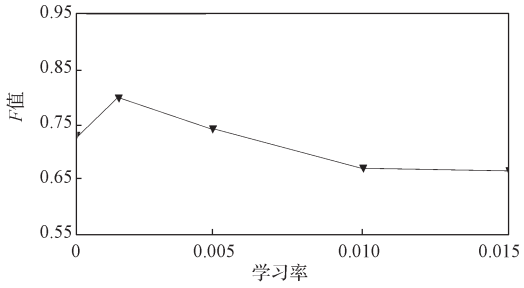

图 4　学习率对 *F* 值的影响

Fig.4　Effect of learning rate on *F* values

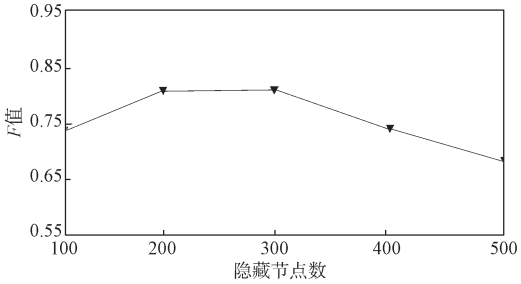

图 5　隐藏节点数对 *F* 值的影响

Fig. 5　Influence of hidden node number on *F* values

表 4　NER 参数设置

Table 4　NER parameter setting

参　数	取　值
学习率	0.002
词向量维	300
Batch size	10
隐藏节点个数	200
词向量窗口大小	10
迭代次数	100

3.5　实验结果及分析

为了验证本文方法的识别效果，分别利用 CRF、LSTM−CRF、BiLSTM−CRF 和 BiLSTM−CNN−CRF 的方法在 SIGHAN 2006 Bakeoff−3 语料上进行命名实体识别实验。实验对比结果如表 5 所列。

表 5　不同命名实体识别方法对比

Table 5　Comparison of different named entity recognition methods　　%

方　法	准确率	召回率	*F* 值
CRF	79.5	73.3	76.3
LSTM−CRF	86.9	80.5	83.6
BiLSTM−CRF	86.3	81.3	83.7
BiLSTM−CNN−CRF	88.1	83.2	85.6
Attention−BiLSTM−CNN−CRF	89.3	88.8	89.5

通过分析表 5 中实验数据发现，和 CRF、LSTM−CRF、BiLSTM−CRF 等传统命名实体识别方法相比，本文方法的实验效果优于前者方法。在 CRF 中加入 LSTM 神经网络模型之后的方法优于 CRF 方法，因为 LSTM 解决了长距离依赖问题，并且可以深度挖掘隐含在文本中的特征信息，说明深度学习方法在命名实体识别的效果上优于传统统计学习方法的。而 BiLSTM−CRF 方法与 LSTM−CRF 方法差距不是特别明显，可能是因为参数设置没有达到最优或者是受到其他微小因素影响。加入 CNN 后的实验效果相比之前的结果有了较大提升，这是因为 CNN 的高计算能力可以计算出字符表示向量，有利于句子上下文信息的表示。从以上实验数据还可以观察到，加入维基百科语料集之后，准确率、召回率和 *F* 值相比其他方法都有了小幅度提升。另外，本文方法加入 Attention 机制，进一步加强了模型的标记预测能力，一个句子中不同的字词对上下文的贡献程度不一样，在特征提取过程中加入文本的局部特征，弥补了传统方法仅注重全局特征提取的缺陷，可以有效抽取更多的上下文特征。以上实验结果表明，基于注意力机制的 BiLSTM 方法在命名实体识别系统中具有优良的性能。

4　结语

综上所述，本研究采用基于注意力机制的 BiLSTM 神经网络进行中文命名实体识别，能够学习语料上下文中的复杂关系，不需要大量人工的特征和数据预处理，并且改善了长距离依赖问题。引入注意力机制可以获取文本局部特征，抑制无价值的信息，有效解决了传统方法在特征提取过程中因注重提取全局特征导致抽取特征不全面问题；在系统中加入了维基百科知识库语料之后，进一步增强了系统的识别能力。本文方法针对通用语料集进行了命名实体识别，未来考虑加入 BERT 模型，进一步提升 NER 系统的应用识别能力，以便模型可以更好地应用到其他研究领域中。

参考文献：

[1] SUNDERMEYER M，NEY H，SCHLUTER R. From Feedforward to Recurrent LSTM Neural Networks for Language Modeling[J]. IEEE/ACM Transactions on Audio Speech & Language Processing，2015，23(3)：517−529.

[2] DUAN H Z，ZHENG Y. A Study on Features of the CRFs-Based Chinese Named Entity Recognition[J]. International Journal of Advanced Intelligence Paradigms，2011，3(2)：287−294.

[3] BORTHWICK A. A Maximum Entropy Approach to Named Entity Recognition[D]. New York：New York

University，1999.

[4] 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.

[5] GREENBERG N，BANSAL T，VERGA P，et al. Marginal Likelihood Training of Bilstm-Crf for Biomedical Named Entity Recognition from Disjoint Label Sets[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels：ACL，2018：2824−2829.

[6] LIU X H，ZHOU M. Two-Stage NER for Tweets with Clustering[J]. Information Processing and Management，2013，49(1)：264−273.

[7] FENG Y H，YU H，SUN G，et al. Domain-Specific Terminology Recognition Method Based on Word Embedding and CRF[J]. Journal of Computer Applications，2016，36(11)：3146−3151.

[8] WANG G Y，CAI Y Q，GE F J. Using Hybrid Neural Network to Address Chinese Named Entity Recognition [C]//2014 IEEE International Conference on Cloud Computing and Intelligence Systems. Shenzhen：IEEE，2014：433−438.

[9] MAYFIELD J，MCNAMEE P，PIATKO C. Named Entity Recognition Using Hundreds of Thousands of Features [C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg：Association for Computational Linguistics，2003：184−187.

[10] KLEIN D，SMARR J，NGUYEN H，et al. Named Entity Recognition with Character-Level Models[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL. Stroudsburg：Association for Computational Linguistics，2003：180−183.

[11] BAHDANAU D，CHO K，BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]// ICLR 2015. San Diego：[s.n.]，2015：1−15.

[12] MA X Z，HOVY E. End-to-End Sequence Labeling Via Bi-Directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.[S.l.]：Association for Computational Linguistics，2016：1064−1074.

[13] GRAVES A，SCHMIDHUBER J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks，2005，18(5/6)：602−610.

[14] QAMAS G K S，尹继泽，潘丽敏，等 . 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(10)：29−35.
QAMAS G K S，YIN Jize，PAN Limin，et al. Research on Named Entity Recognition Method Based on Deep Neural Network[J]. Netinfo Security，2017(10)：29−35.

[15] 姚 霖，刘 轶，李鑫鑫，等 . 词边界字向量的中文命名实体识别 [J]. 智能系统学报，2016，11(1)：37−42.
YAO Lin，LIU Yi，LI Xinxin，et al. Chinese Named Entity Recognition via Word Boundary Based Character Embedding[J]. CAAI Transactions on Intelligent Systems，2016，11(1)：37−42.

（责任编辑：申 剑）