文章编号：1003-0077（2011）00-0000-00

基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别*

李丽双，郭元凯

（大连理工大学 计算机科学与技术学院，辽宁 大连 116023）

摘要：命名实体识别是自然语言处理任务的重要步骤。近年来，不依赖人工特征的神经网络在新闻等通用领域命名实体识别方面表现了很好的性能。然而在生物医学领域，许多实验表明基于领域知识的人工特征对于神经网络模型的结果影响很大。因此，如何在不依赖人工特征的情况下获得较好的生物医学命名实体识别性能是有待解决的问题。本文提出一种基于 CNN-BLSTM-CRF 的神经网络模型。首先利用卷积神经网络（CNN）训练出单词的具有形态特征的字符级向量，并从大规模背景语料训练得到具有语义特征信息的词向量，然后将二者进行组合作为输入，再构建适合生物医学命名实体识别的 BLSTM-CRF 深层神经网络模型。实验结果表明，不依赖任何人工特征，本文方法在 Biocreative II GM 和 JNLPBA2004 生物医学语料上都达到了目前最好的结果，F-值分别为 89.09%和 74.40%。

关键词：生物医学命名实体识别；词向量；LSTM; CNN；CRF；

中图分类号：TP391　　**文献标识码：**A

Biomedical Named Entity Recognition with CNN-BLSTM-CRF

LI Lishuang, GUO Yuankai

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China)

Abstract: Named entity recognition (NER) is one of important stages in natural language processing (NLP). In recent years, End-to-End neural network models for named entity recognition have shown effective performances on general domain datasets (e.g. news), without additional hand-crafted features. However, in biomedical domain, recent studies indicate that hand-designed features have great impact on the model's performance. In this paper, we propose a novel end to end neural network model: CNN-BLSTM-CRF, which does no relay on the hand-designed features and domain knowledge. Firstly, CNN (convolutional neural network) extracts the character vectors with shape features from each word, which are concatenated with the word embeddings and fed to the BLSTM-CRF network.We evaluate our approach by comparing against existing neural network models for NER using Biocreative II GM dataset and JNLPBA2004 dataset. The experimental results show that out system reaches 89.09% and 74.40% F-scores and outperforms other state-of-the-art of methods.

Keywords: Biomedical NER;Word Embeddings; LSTM; CNN; CRF

1.引言

命名实体识别是自然语言处理任务的重要步骤。近年来，神经网络在通用领域的命名实体识别表现了很好的性能。相比于统计机器学习方法或基于规则的方法，基于神经网络的深

*收稿日期：　　定稿日期：

基金项目： 国家自然科学基金项目(no.61672126)。

作者简介：李丽双（1967—），女，教授，博士生导师，本文通信作者，主要研究领域为自然语言处理、信息抽取与文本挖掘。郭元凯（1994—），男，硕士，主要研究领域为自然语言处理。

度学习方法具有泛化性更强，更少依赖人工特征的优点。因此，许多基于神经网络的通用领域命名实体识别模型被提出。例如 Collobert[1]等首次使用 CNN 与 CRF 结合的方式在通用命名实体识别领域的 CONLL2003 语料上取得了较好的效果。Huang[2]等构造了一个采用人工设计的拼写特征的 BLSTM-CRF 模型，在 CONLL2003 语料上达到了 88.83%的 F-值。Chiu 和 Nichols[3]等建立了 CNN-LSTM 模型在 CONLL2003 语料上达到了 91.62%的 F-值。虽然神经网络在通用命名实体识别领域中展现了较好的性能，但在生物医学命名实体识别领域中的应用仍存在问题。相比于一般领域的命名实体，生物医学命名实体识别有以下几个难点：1、包含的实体数量和种类多；2、待识别的实体可能会由许多单词修饰，导致实体的边界难以划分；3、生物医学语言没有一套统一的命名方式，所以待识别的实体可能会有多种的表述方式；4、待识别的实体经常存在缩写，嵌套，大小写混合，含有特殊字符的情况。也正是因为如此，生物医学命名实体识别的许多方法依旧依赖人工特征和领域知识。

目前生物医学命名实体识别的方法主要分为浅层机器学习和深层神经网络的方法。浅层机器学习方法主要包括，条件随机场模型（CRF）隐马尔可夫模型（HMM），最大熵模型（ME），支持向量机（SVM）等。例如，Li[4]等通过使用丰富的人工特征基于 CRF 进行实体识别，在 Biocreative II GM 语料上达到了 87.28%的 F-值。Manabu [5]等将 CRF，HMM，ABNER，LingPipe 等模型融合，在 Biocreative II GM 语料上达到了最高 F-值 88.87%。此外 Wang[6]等验证了基于 CRF 的 Gimli 方法，在 JNLPBA2004 语料上 F-值达到了 72.23%。Zhou 和 Su[7]通过丰富的领域知识和人工特征采用 CRF 在 JNLPBA2004 语料上 F-值提高到了 72.55%。Liao[8]等构建了 skip-chain CRF 模型用于生物医学命名实体识别，该模型能够充分考虑到较远距离具有依赖关系的生物医学信息，在 JNLPBA2004 语料上达到了 73.20%的 F-值。但是传统的浅层机器学习方法在很大程度上依赖于人工特征的设计，人工特征和领域知识在提高模型性能的同时也导致整个模型的鲁棒性和泛化能力下降。

为了减少复杂的人工特征，有相关研究利用词向量结合浅层机器学习方法进行生物实体识别。如 Tang[9]等采用 CRF 模型进行生物实体识别，在基本人工特征的基础上加入不同的词向量特征，在 BioCreative II GM 和 JNLPBA 语料上的 F-值分别为 80.96%和 71.39%。Chang[10]等利用少量人工特征和词向量结合的方式构建 CRF 模型并添加后处理，在 JNLPBA 语料上达到了 71.39%的 F-值。虽然词向量在一定程度能够提高浅层机器学习方法的性能，但是与其他最好的系统相比仍然存在一定的差距，这主要是因为这些词向量本身包含的特征信息有限，并不能完全取代复杂的人工特征，而且难以模式化距离较远信息的依赖关系。

在使用深度神经网络进行生物医学命名实体识别的研究中，Yao[11]等首先在无标注的生物文本上利用神经网络生成词向量，然后建立多层神经网络，在 JNLPBA 语料上 F-值为 71.01%。Li [12]等采用双向长短期记忆网络（BLSTM）方法在 Biocreative II GM 的语料上达到了 88.6%的 F-值，同时在 JNLPBA 语料上达到了 72.76%的 F-值。上述研究虽然没有使用领域知识和人工特征，但是词向量对于字符级特征不能很好表示，因此识别性能有待提高。本文提出一种基于 CNN-BLSTM-CRF 神经网络模型。该模型首先利用 CNN 训练出单词的字符级特征，然后与从大规模背景语料训练得到的词向量进行组合，再将组合的词向量送入 BLSTM-CRF 深层神经网络进行训练，从而得到一个利用字符级特征和词向量的生物实体识别模型。在 Biocreative II GM 和 JNLPBA2004 语料上的实验结果表明，在未使用任何人工特征的情况下，该模型在两个语料上都达到了目前的最好效果，F-值分别是 89.09%和 74.40%。

2.基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别

2.1 模型整体框架

图 1 为本文的 CNN-BLSTM-CRF 模型框架。CNN-BLSTM-CRF 共有三部分组成，CNN

模块，BLSTM 模块和 CRF 模块。首先通过查询词向量表将输入的语句转换为相应的词向量序列。然后对于语句中的每一个单词，通过查询字符向量表获得每个字符的字符向量。由字符向量组成单词的字符向量矩阵。CNN 对字符向量矩阵进行卷积和池化，获得每个单词的字符级特征。每个单词的字符向量和词向量进行拼接，拼接后的词向量输入 BLSTM 进行实体识别。最后 CRF 模块将 BLSTM 的输出解码出一个最优的标记序列。

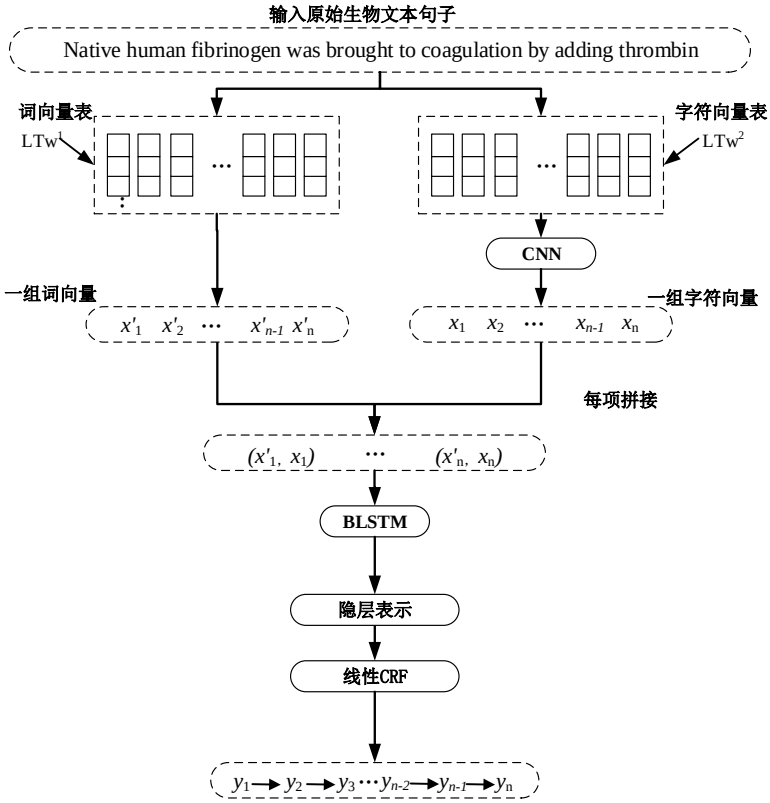

图 1　生物医学命名实体识别的 CNN-BLSTM-CRF 模型

2.2 CNN 模块

卷积神经网络中的卷积层能够很好地描述数据的局部特征，通过池化层可以进一步提取出局部特征中最具有代表性的部分。Santos[13]等利用 CNN 对字符进行处理得到 CharWNN 用于词性标注工作（POS），并取得了较好的效果。Chiu 和 Nichols[3]等采用 CNN 抽取字符级特征在通用实体识别领域达到了很好的效果。因此，本文提出利用 CNN 抽取生物医学文本中单词的字符级特征，通过字符级特征与词向量相结合的方法来提高模型的性能。这里使用的 CNN 模块与 Chiu[3]不同之处有：1、本文并没有采用 Chiu[3]额外设计一些人工的字符特征与字符向量拼接的方法。2、本文对于不同类型的字符设置并随机初始化了不同的字符向量，以区分字符的大小写，字符类型（字母，数字，标点，特殊字符）。例如大写字母 A 与小写字母 a 分别对应了两组不同的字符向量。

CNN 的结构如图 2 所示。主要由字符向量表，卷积层，池化层组成。字符向量表将一个单词中的每个字符转化成为对应的字符向量。首先，由单词的每个字符的字符向量组成单词的字符向量矩阵。其次，为了解决由于单词长度不同导致字符向量矩阵大小不同的问题，以最长的单词为准，在单词的左右两端补充占位符（padding）使得所有字符向量矩阵大小一致。最后，字符向量表在模型的训练过程中通过反向传播算法不断更新。

卷积层使用一个大小是 T 的卷积核在单词的字符向量矩阵上进行卷积来提取出局部特征，卷积核大小 T 决定了可以提取单词周围 T 个词的特征。最后通过池化获得单词的字符级特征向量。

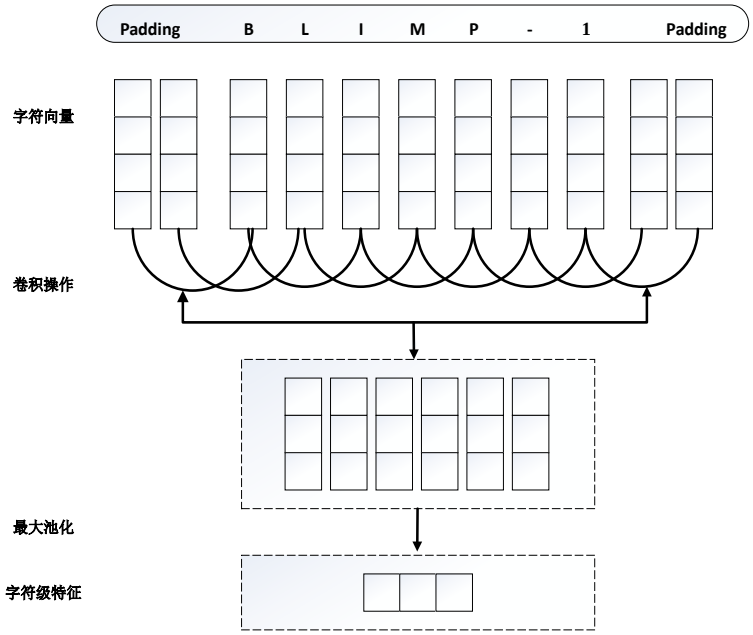

图 2　字符级卷积神经网络模型

2.3 BLSTM 模块

长短时记忆网络（LSTM）[14] 是一种特殊的循环网络（RNN）模型，克服了传统 RNN 模型由于序列过长而产生梯度弥散问题。LSTM 模型通过特殊设计的门结构使得模型可以有选择的保存上下文信息，因此 LSTM 具有适合生物医学命名实体识别的特点。LSTM 网络的主要结构可以形式化地表示为：

$$
\begin{aligned}
i_t &= \sigma(x_t \cdot w_{xh}^i + h_{t-1} \cdot w_{hh'}^i + b_h^i) \\
f_t &= \sigma(x_t \cdot w_{xh}^f + h_{t-1} \cdot w_{hh'}^f + b_h^f) \\
o_t &= \sigma(x_t \cdot w_{xh}^o + h_{t-1} \cdot w_{hh'}^o + b_h^o) \\
\tilde{c}_t &= \tanh(x_t \cdot w_{xh}^c + h_{t-1} \cdot w_{hh'}^c + b_h^c) \\
c_t &= i_t \otimes \tilde{c}_t + f_t \otimes c_{t-1} \\
h_t &= o_t \otimes \tanh(c_t)
\end{aligned}
\qquad (1)
$$

其中 σ 是激活函数 *sigmod*；$\otimes$ 是点乘运算，*tanh* 表示双曲正切激活函数，i_t，f_t，o_t，分别表示在 *t* 时刻的输入门，忘记门，输出门。C_t *表示 t 时刻的*状态。h_t 表示 *t* 时刻的输出。

为了能够有效利用上下文信息，我们采用双向 LSTM（BLSTM）结构。双向 LSTM 对每个句子分别采用顺序（从第一个词开始，从左往右递归）和逆序（从最后一个词开始，从右向左递归）计算得到两套不同的隐层表示，然后通过向量拼接得到最终的隐层表示。

2.4 线性 CRF 模块

CRF 能够考虑到相邻标签的关系获得一个全局最优的标记序列。本文将 CRF 融合到

BLSM 模块中，对 BLSTM 的输出进行处理，获得全局最优的标记序列。对于一个句子 $S=\{W_1,W_2,\cdots\cdots,W_n\}$送入网络中训练，定义矩阵 P 是 BLSTM 层的输出结果，其中 P 的大小 $N\times K$，N 是单词个数，K 是标签的种类。定义 p_{ij} 代表句子中第 i 个单词的第 j 个标签的概率。对于一个预测序列 $y=\{y_1,y_2,\cdots\cdots,y_n\}$，它的概率可以表示为：

$$S(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{2}$$

矩阵 A 是转移矩阵，例如 A_{ij} 表示由标签 i 转移到 j 的概率。y_0，y_n 则是句子起始和结束的标记，因此 A 是一个大小为 $K+2$ 的方阵。所以在原语句 S 的条件下产生标记序列 y 的概率为：

$$p(y\mid S)=\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Y_X}S(X,\tilde{y})} \tag{3}$$

在训练过程中标记序列的似然函数：

$$\log(p(y\mid S))=s(X,y)-\log(\sum_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}) \tag{4}$$

其中 Y_X 表示所有可能的标记集合，包括不符合 BIOES 标记规则的标记序列。通过公式（4）得到有效合理的输出序列。预测时，由公式（5）输出整体概率最大的一组序列。

$$y^{*}=\arg\max_{\tilde{y}\in Y_X}s(X,\tilde{y}) \tag{5}$$

2.5 训练参数

训练过程中，优化器采用 RMSprop，相比于随机梯度下降 SGD 模型的训练速度更加快速；学习率选取 0.001。同时，通过实验发现在双向 LSTM 的输入和输出部分增加 Dropout 可以减轻模型过拟合的问题，dropout[15]值选取了 0.5。整体的模型训练通过 GTX1080 进行加速。

3.实验结果

为了说明本文 CNN-BLSTM-CRF 模型的有效性和泛化性，分别选用了 Biocreative II GM 和 JNLPBA2004 语料进行了实验。所有的实验都是基于相同预训练 200 维词向量和相同参数。

3.1 语料介绍

Biocreative II GM 和 JNLPBA2004 语料详细信息见表 3.1。此外 JNLPBA 的语料不同于 Biocreative II GM 语料，JNLPBA 待识别的实体有 5 种，分别是 DNA，RNA，Cell_line，Cell_type，Protein。所以相对于 Biocreative II GM 语料只需识别出基因实体，JNLPBA 则还需对于识别出的实体给出准确的类别。

表 1 语料介绍

语料	训练集	开发集	测试集
Biocreative II	15000	-	5000
JNLPBA 2004	2000	-	404

在语料处理方面，为了能够清楚的表示语料中待识别的命名实体，我们采用了 BIOES 标记的方式代替 BIO2 去标记实体。因为根据 Ratinov 和 Roth[16]，Dai[17]，Lample[18]等的研究采用 BIOES 的标记效果好于 BIO2 方式，能更加清楚的划分实体的边界。BIOES 的标记规则如下：B 代表一个实体的开头，I 则代表实体中间的单词，E 则是一个实体中最后一个单词，S 则代表只有一个单词的实体，O 则是其他非实体的单词。

3.2 实验结果及分析

表 2 Biocreative II GM 结果

模型	精确率(%)	召回率(%)	F-值(%)
LSTM	83.28	74.20	78.48
BLSTM	84.17	81.38	82.75
BLSTM-CRF	84.65	85.59	85.12
CNN-BLSTM	88.17	83.38	85.71
CNN-BLSTM-CRF	**89.31**	**88.88**	**89.09**

CNN-BLSTM-CRF 模型在 Biocreative II GM 语料上的进行生物命名实体识别结果如表 2 所示，下面通过对比实验的结果来分析各个模块在模型中起到的作用。

（1）BLSTM 模块

为了验证 BLSTM 结构的有效性，进行了 BLSTM 模型与 LSTM 模型的对比实验。根据实验结果，BLSTM 模型在 Biocreative II GM 语料上的 F-值为 82.75%，精确率 84.17%，召回率 81.38%，比 LSTM 模型的 F-值高出了 4.27%。无论是召回率还是精确率，双向 LSTM 递归神经网络明显优于单向的网络，主要由于 BLSTM 模型相对于 LSTM 模型，更加充分的利用了上下文信息。

（2）CNN 模块

为了验证 CNN 模块抽取的字符级特征的有效性，进行了 CNN-BLSTM 模型与 BLSTM 模型的对比实验。实验结果表明，CNN-BLSTM 相对于 BLSTM 精确率提升了 4.00%，召回率提升了 2.00%，F-值提升 2.96%。表明了 CNN 抽取的字符级特征的有效性。由于我们通过 CNN 模块抽取的字符级向量能够一定程度上表示形态特征，所以对于具有大小写混合，包含特殊字符，边界模糊特点的这类生物实体能够充分获取相关特征，从而提高识别的 F-值。例如‘LTRran1+ kinase’，‘fructose-2,6-bisphosphatase’，‘G alpha i-2’等实体被 CNN-BLSTM 正确识别，而 BLSTM 则不能正确识别。由此可见 CNN 模块的加入使得模型对存在含有特殊字符的实体效果提升。

（3）线性 CRF 模块

为了验证 CRF 模块的有效性，进行了 BLSTM-CRF 模型与 BLSTM 模型的对比实验。实验结果表明，BLSTM-CRF 模型相比于 BLSTM 模型，召回率、精确率、F-值分别提高了 4.21%，0.48%，2.73%。由于线性 CRF 能够充分利用相邻标签的关系，在全局优化输出的标签序列，对长度较大以及带有修饰词汇的生物命名实体，识别性能较高。例如‘mammalian glycoprotein hormone receptors’，‘P-ITIM-compelled multi-phosphoprotein complex’，‘human urokinase-type plasminogen activator gene’这类生物医学实体被 BLSTM-CRF 正确识别，而 BLSTM 模型则不能正确识别。CRF 模块的加入解决了一部分含有修饰词，长度较大实体识别的问题。

3.3 与现有其他工作的对比

（1）Biocreative II GM 语料

表 3 给出了 Biocreative II GM 语料上本文模型与先进系统的对比结果。下面从是否采用人工特征进行分析。

表 3 在 Biocreative II GM 语料上的模型对比

模型	精确率(%)	召回率(%)	F-值(%)
Ando et al.	88.48	85.97	87.21
Li et al.	90.38	84.39	87.28
Li and Jin et al.	89.54	87.69	88.61
Manabu Torii et al.	88.21	89.52	88.87
Tang et al.	-	-	80.96
Ours	**89.31**	**88.88**	**89.09**

在使用人工特征的方法中，Ando[19]等从大规模未标注数据中学习新的特征表示，结合字典和大量设计的人工特征在 Biocreative II GM 评测上取得了第一名，达到了 87.21%的 F-值。Li 等通过使用丰富的人工特征例如：词性，形态特征，词干等用 CRF 进行实体识别，达到了 87.28%的 F-值。Manabu[5]等通过将 CRF，HMM，ABNER，LingPipe 等模型融合，在 Biocreative II GM 语料上达到了 88.87%的 F-值。而本文方法自动学习词向量和字符向量，未采用任何基于人工总结的特征，取得了比采用大量领域知识和人工特征方法更好的结果。

在不依赖人工特征的方法中，Li 和 Jin[12]通过构建带有双词向量和句子向量的 BLSTM 模型达到了 88.87%的 F-值。该方法充分利用了词向量所表示的语义但未能表达字符级形态特征。本文的 CNN-BLSTM-CRF 模型则通过 CNN 卷积获得了字符级的词的形态特征，并与词向量组合，达到了 89.09%的 F-值，比 Li 和 Jin[9]的结果高了 0.48%。

通过以上对比分析可以看出，我们的模型在未使用任何人工特征的情况下，在 Biocreative II GM 取得了目前的最好结果。

（2）JNLPBA 语料

为了说明我们模型的泛化能力，表 4 给出了 JNLPBA 语料上与其他先进模型的对比实验。同样从是否采用人工特征方面进行分析。

在使用人工特征方面，良好设计的人工特征起到了很好的作用。例如，Chang[10]等通过一些人工设计的特征和词向量送入 CRF 模型进行训练达到了 71.85%的 F-值。此外 Wang[20]等验证了基于 CRF 的 Gimli 方法达到了当时的最高 F-值 72.23%，Zhou 和 Su[7]通过丰富的领域知识和人工特征使 F-值提高到了 72.55%。Liao[8]等构建了能充分考虑到长距离依赖关系的 skip-chain CRF 模型，在 JNLPBA2004 语料上达到了 73.20%的 F-值。 而本文提出的 CNN-BLSTM-CRF 模型，未采用任何人工特征，在 JNLPBA 语料上取得了更好的结果。由表 4 可知，CNN-BLSTM-CRF 模型比目前最好的系统 Liao[8]F-值提高 1.20%。

同样也有探索不依赖人工特征的深层神经网络方法，Yao[11]等通过使用多层神经网络学习特征表示，达到了 71.01%的 F-值。Li 和 Jin[9]等人通过构建带有双词向量和句子向量的 BLSTM 模型，达到了 72.76%F-值。本文的方法 F-值为 74.40%，比 Yao[11]，Li 和 Jin[12]等分别高出了 3.39%，1.64%。

表 4 在 JNLPBA 语料上的模型对比

模型	精确率(%)	召回率(%)	F-值(%)
Yao et al.	76.13	66.54	71.01
Chang et al.	-	-	71.85
Liao et al.	72.80	73.60	73.20
Zhou and Su et al.	75.99	69.42	72.55
Li and Jin	74.77	70.85	72.76
Tang et al	70.78	72.00	71.39
ours	**79.58**	**69.86**	**74.40**

4.结论

本文针对生物医学命名实体识别任务，提出了通过 CNN 网络获得字符级特征来补充词向量，进而构建 CNN-BLSTM-CRF 神经网络模型的方法，在 Biocreative II GM 和 JNLPBA 语料上取得了目前最好的性能。主要结论如下：

首先，在生物医学命名实体识别任务中，人工特征和领域知识对于结果的影响很大。但是构建合适的人工特征需要大量的特征选择实验，导致了系统的成本提升，泛化能力下降。因而本文构建了 CNN-BLSTM-CRF 深层神经网络模型，在不使用任何人工特征的情况下，获得了比使用大量丰富特征和领域知识的浅层机器学习方法更好的结果。

其次，本文提出了利用 CNN 网络卷积来获得表示单词形态特征的字符向量，用以补充词向量的不足。通过字符向量的加入，使得模型对于含有特殊字符，大小写混合这类实体能够更有效识别，从而提高了模型的性能。

最后，为了获得更加准确的识别结果，我们通过 CRF 对 CNN-BLSTM 网络的输出进行解码，获得最优的标记序列。CRF 的融入提升了对于含有多修饰词，边界模糊的生物医学实体的识别性能。

综上，在生物医学命名实体识别任务上，本文提出的通过 CNN 网络获得字符级特征来补充词向量，以及 BLSTM 与 CRF 模型的融合都是有效提高识别性能的有效途径。

参考文献：

[1]Pinheiro P H O, Collobert R. Recurrent Convolutional Neural Networks for Scene Parsing[J]. 2014, 1: 82-90.

[2] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. 2015, arXiv preprint arXiv:1508.01991.

[3] Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. 2015, arXiv preprint arXiv:1511.08308.

[4] Li Y, Lin H, Yang Z. Incorporating Rich Background Knowledge for Gene Named Entity Classification and Recognition[J]. BMC Bioinformatics, 2009, 10(1): 223.

[5] Torii M, Hu Z, Wu C H, et al. BioTagger-GM: a Gene/Protein Name Recognition System[J]. Journal of The American Medical Informatics Association, 2009, 16(2): 247-255.

[6] Wang X, Yang C, Guan R. A Comparative Study for Biomedical Named Entity Recognition[J]. International Journal of Machine Learning & Cybernetics, 2015: 1-10.

[7] Zhou G, Zhang J, Su J, et al. Recognizing Names in Biomedical Texts: a Machine Learning Approach[J]. Bioinformatics, 2004, 20(7): 1178-1190.

[8]Liao Z, Wu H. Biomedical Named Entity Recognition Based on Skip-Chain CRFS[C]//Proceedings of the Industrial Control and Electronics Engineering (ICICEE). 2012: 1495-1498.
[9] Tang B, Cao H, Wang X, et al. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks[J]. BioMed research international, 2014: 1-6.
[10] Chang, F, Guo, J, Xu, W, et al. Application of Word Embeddings in Biomedical Named Entity Recognition Tasks [J]. Digital Inf. Manage.2015, 13(5): 321-327.
[11] Yao L, Liu H, Liu Y, et al. Biomedical Named Entity Recognition Based on Deep Neutral Network[J]. International Journal of Hybrid Information Technology, 2015, 8(8): 279-288.
[12]Li L, Jin L, Jiang Y, et al. Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM[C]//China National Conference on Chinese Computational Linguistics. Springer International Publishing. 2016: 165-176.
[13] Santos C D, Zadrozny B. Learning Character-Level Representations for Part-of-speech Tagging[C] //Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014: 1818-1826.
[14] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 2014, 9(8):1735-1780.
[15] 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.
[16] Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009: 147-155.
[17] Dai H J, Lai P T, Chang Y C, et al. Enhancing of Chemical Compound and Drug Name Recognition using Representative Tag Scheme and Fine-grained Tokenization[J]. Journal of Cheminformatics, 2015, 7(1):S14.
[18] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J], 2016: 260-270.
[19] Ando R K. BioCreative II Gene Mention Tagging System at IBM Watson[C]//Proceedings of the Second BioCreative Challenge Evaluation Workshop. 2007,(23): 101-103.
[20] Wang X, Yang C, Guan R. A Comparative Study for Biomedical Named Entity Recognition[J]. International Journal of Machine Learning and Cybernetics, 2015: 1-10.

李丽双（1967—），女，教授，博士生导师，主要研究领域为自然语言处理。本文通讯作者。Email: lilishuang314@163.com。

郭元凯（1994—），男，硕士，主要研究领域自然语言处理。Email: guoyuankai@outlook.com。