计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000201

LAC-DGLU:基于 CNN 和注意力机制的命名实体识别模型

赵 丰 黄 健 张中杰
国防科技大学智能科学学院 长沙 410073
(62794258@qq.com)

摘 要 对文本进行分词和词嵌入通常是中文命名实体识别的第一步,但中文的词与词之间没有明确的分界符,专业词及生僻词等未收录词(Out of Vocabulary,OOV)严重干扰了词向量的计算,基于词向量嵌入的模型性能极易受到分词效果的影响。同时现有模型大多使用循环神经网络,计算速度较慢,很难达到工业应用的要求。针对上述问题,构建了一个基于注意力机制和卷积神经网络的命名实体识别模型,即 LAC-DGLU。针对分词依赖的问题,提出了一种基于局部注意力卷积(Local Attention Convolution,LAC)的字嵌入算法,减轻了模型对分词效果的依赖。针对计算速度较慢的问题,使用了一种带门结构的卷积神经网络,即膨胀门控线性单元(Dilated Gated Linear Unit,DGLU),提高了模型的计算速度。在多个数据集上的实验结果显示,该模型相比现有最优模型 F1 值提高了 0.2%~2%,训练速度可以达到现有最优模型的 1.4~1.9 倍。

关键词:字嵌入;局部注意力卷积;膨胀卷积;门控线性单元;残差结构
中图法分类号 TP391

LAC-DGLU:Named Entity Recognition Model Based on CNN and Attention Mechanism

ZHAO Feng,HUANG Jian and ZHANG Zhong-jie
College of Artificial Intelligence,National University of Defense Technology,Changsha 410073,China

Abstract Text segmentation and word embedding are usually the first step in Chinese named entity recognition,but there is no clear delimiter between Chinese words and words. OOV(out of vocabulary) words such as professional words and uncommon words are severely disturbing the computation of word vectors. Model performance based on word vector embedding is highly susceptible to word segmentation effects. At the same time,most of the existing models use low-speed recurrent neural network which is difficult to meet the requirements of industrial applications. Aiming at the above problems,this paper constructs a named entity recognition model based on attention mechanism and convolutional neural network:LAC-DGLU. To handel the problem of word segmentation,this paper proposes a word embedding algorithm based on Local Attention Convolution (LAC),which alleviates the dependence of the model on the effect of word segmentation. For the problem of slow calculation speed,this paper uses aconvolutional neural network with gate structure:Dilated Gated Linear Unit (DGLU) to improve the speed of model calculation. The experimental results on several datasets show that the model can increase the F1 value by 0.2% to 2%compared with the existing mainstream model,and the calculation speed can reach more than 1.4to 1.9 times of the existing mainstream model.

Keywords Character embedding,Local attention convolution,Dilated convolution,Gated linear unit,Residual structure

1 引言

命名实体识别的目标是从文本中识别出具有特定意义的实体,是自然语言处理领域的基础任务之一。由于命名实体识别是主体挖掘、关系抽取、实体链接等任务的前置步骤,其研究长期以来受到了广泛的关注。

基于深度学习的方法是当前进行命名实体识别的主流方法。现有的基于深度学习的模型多将命名实体识别转化为一个序列标注任务:基于词嵌入[1],使用循环神经网络[2](Recurrent Neural Network,RNN)和条件随机场[3](Conditional Random Field,CRF)对文本序列中的每一个词进行标注,判断该词是否是实体的一部分,最后从标注序列中提取出实体。在处理英文时,这些对每个词进行标注的模型取得了良好的效果。但中文与英文不同,英文的词与词之间有明确的分界符,而中文的词与词之间却没有明确的界限,词的边界在文本中是模糊的,往往需要使用分词工具先对原始文本进行分词,分词的结果会对模型的性能产生巨大的影响。例如"台风战斗机"这一实体,分词工具可能将该实体作为一个独立词,也有可能将实体切分为两个词,即"台风"和"战斗机",第二种切分方法引入的"台风"这一与实体相关度低的词无疑会干扰实体识别的效果。不同的分词结果甚至可能导致句子意义出现巨大差别,例如表 1 中的句子通过两种不同的方式分词后意义已经有了很大差别,最终识别出的命名实体也必然是不同的。除此之外,专业领域包含大量的专有名词和特定术语,分

到稿日期:2019-10-31 返修日期:2020-03-29 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:黄健(nudtjHuang@hotmail.com)

词工具很难将这些词切分出来，大量 OOV 在嵌入时只能用全零向量表示，这个过程损失了大量的语义信息。

表 1　分词不同导致句义不同

Table 1　Different meanings made by word-segmetation

原始句子	乒乓球拍卖完了
分词模式 1	乒乓球拍/卖 /完了
分词模式 2	乒乓球 /拍卖 /完了

相较于词嵌入，字嵌入更适应中文的特性，对文本序列中的每一个字符进行标注，就不需要考虑词的边界问题了。但单一字符难以表示有效的语义信息，如“没有”和“淹没”两个词中的“没”这一字，在不同词语中蕴含的语义信息是完全不同的，但经过字嵌入后得到的字向量是相同的。由此可见，普通字向量所包含的信息区分度不够高，词级别的语义信息仍是不可或缺的。

除了词嵌入算法对中文的不适应，计算速度慢也是当前主流模型长期难以投入工业应用的重要原因。由于 RNN 天然的序列结构，每一个时间步(Time Step)上的计算依赖于前一个时间步的输出或隐状态[4]，模型进行训练和预测时无法进行样本内的并行计算。在面对超长文本时，RNN 无法并行计算的问题尤为显著，经典的 BiLSTM-CRF[5] 模型中，长短时记忆(Long Shor-term Memory，LSTM)网络则是带门机构的 RNN，CRF 需要使用维特比(Viterbi)算法[3]进行解码，整个模型需要较高的硬件条件才能达到理想的响应速度。

针对模型性能依赖分词效果的问题，我们提出了一种基于局部注意力卷积(Local Attention Convolution，LAC)的字嵌入算法，在减轻模型对分词效果依赖的同时能够利用词级别的语义信息。该算法以字为基础计算单元，在字向量中融入了词级别的语义信息，这些信息主要通过两种方式获得：加入了一个带局部注意力机制的卷积层，该层可以提取字符序列的字符间(Bi-Char)依赖关系这一局部结构特征，将字符间的依赖关系编码到字向量中，通过整合字向量和字符间依赖关系来获取词级别语义信息；将预训练词向量作为一个特征拼接到字向量后面，将其作为词级别语义信息的补充。

针对计算速度慢的问题，我们使用了一种新型结构，即 DGLU，该结构是在 CNN 的基础上改造得到的。与 RNN 相比，DGLU 可以进行样本内的并行计算，且计算速度更快；与普通 CNN 相比，门控、膨胀卷积等机制的加入使 DGLU 能够在层数较小的情况下获得较大的感受野，能够提取长距离依赖关系，且训练的稳定性更强。

通过上述两个方面的改进，我们构建了一种基于卷积神经网络和注意力机制[6](Attention Mechanism)的命名实体识别模型，即 LAC-DGLU。实验结果显示，该模型在多个数据集上相比主流模型取得了 2%～8%的提高，且训练速度达到了主流模型的 1.4～1.9 倍。

2　相关工作

伴随着深度学习的快速发展，基于深度学习的方法成为了命名实体识别的主流方法。Huang 等[5]利用 LSTM 和 CRF 构建了基于词向量的 BiLSTM-CRF 这一主流模型，该模型将命名实体识别任务转变为序列标注任务的思想也为后续研究广泛接受。Lample 等[7]将 BiLSTM-CRF 模型改造为端到端模型，融合了字级别和词级别的信息，在 CoNLL-2003[8] 数据集上 F1 值高达 90.94。Strubell 等[9]提出了迭代膨胀卷积神经网络(Iterated Dilated Con-volution Neural Network，IDCNN)，首次将 CNN 用于命名实体工作，在 CoNLL-2003 数据集[8]上 F1 值为 90.54，计算速度也取得了一定程度的提高。Vaswani 等[6]提出的注意力机制在不使用 RNN 和 CNN 的情况下在自然语言处理领域的多个任务中都取得了良好的效果，引发了广泛的关注。Shen 等[10]提出了将自注意力(Self-Attention)用于提取文本中的长距离依赖，为在命名实体识别领域使用注意力机制提供了技术途径。也有许多研究聚焦迁移学习和预训练模型。Peters 等[11]提出的 ELMo 是基于 LSTM 使用大量未标注语料训练得到的预训练模型，在多个任务榜单刷新了记录。Devlin 等[12]提出的预训练模型 BERT 则将 LSTM 替换为双向注意力机制，有效解决了多义词问题。

在中文命名实体识别领域，BiLSTM-CRF 模型也得到了广泛的应用。Feng 等[13]将 BiLSTM-CRF 用于中文新闻等通用领域数据集，F1 值为 90.5。Shan 等[14]将注意力机制融入了 BiLSTM-CRF 模型，在军事文本这一专业领域数据集上 F1 值达到了 90.79。Peng 等[15]构建了一种联合训练模型，同时训练词向量和实体识别模型。Zhang 等[16]提出了一种带网格结构的 LSTM，将来自预训练模型的词级别信息动态地添加至字向量，可以处理已收录词相关的分词错误问题；Cao[17]等提出了一种新型的对抗迁移学习框架，将分词任务和命名实体识别任务结合在一起共享信息。Dong 等[18]对字符进行拆解，获取了偏旁部首级别的特征，在 MSRA 数据集上将其与字向量配合使用，F1 值为 90.95%。Yang 等[19]提取了更细粒度的特征，将每个汉字的笔画也作为特征输入到模型中，进一步提高了识别效果。

现有中文命名实体识别模型主要通过两种方法提高模型的性能：提取更细粒度的特征和缓解分词错误问题。

Dong 等[18]、Yang 等[19]分别提取了偏旁部首和汉字笔画级别的特征。通过这些研究可以发现，虽然细粒度特征对模型识别的效果有一定提升，但是提取更细粒度的特征需要建立专用的偏旁部首及笔画数据库，数据标注工作量也有所提高，计算量大且泛用性较差。因此，我们主要针对分词错误问题开展研究。

Peng 等[15]、Cao 等[17]分别将词嵌入模型和分词模型与命名实体识别模型结合进行联合训练，联合训练能够缓解误差传播，降低分词错误的概率，从而缓解分词错误问题，但该方法无法保证分词完全正确，只要模型仍基于词嵌入，分词错误的问题就依然存在。Zhang 等[16]提出的网格 LSTM 基于字嵌入，部分解决了分词错误问题，但该模型的词级别信息只来源于已收录词的预训练词向量，无法处理专业词、生僻词及流行语等未收录词。上述中文命名实体识别模型全部基于 LSTM 实现，LSTM 由于无法进行样本内并行计算的天然特性，计算速度较慢，现有模型在 LSTM 的基础上进行扩展，很难达到工业应用的要求，因此设计可以替代 LSTM 的高速计算结构具有一定的研究意义。使用字嵌入可以避开分词错误

这一问题，但单个字符所包含的语义信息量少且不完备，需要将词级别信息编码到字向量中。

针对分词错误问题，我们使用字嵌入来略去分词这一步骤，同时使用局部注意力卷积将词级别信息编码到字向量中，从而保证信息的高区分度；针对计算速度慢的问题，我们借鉴LSTM的设计思路，将门控机制加入到CNN中，引入膨胀卷积、残差结构等机制，赋予CNN提取长距离依赖的能力，设计实现可以替代LSTM的高速计算结构。

3 LAC-DGLU 模型结构

命名实体识别的过程可以理解为如下形式：对于一个输入的文本序列 $T=(t_1,t_2,\cdots,t_n)$，t_i 为该序列中的第 i 个字符，我们需要生成对应的输出序列 $Output=\{o_1,o_2,\cdots,o_n\}$，$o_i\in L$，$L$ 为所有标签的集合，最后通过解码算法从 $Output$ 中解码出实体。模型由嵌入层、编码层、注意力层、输出层4部分组成，整体结构如图1所示，后面将对每一部分依次进行介绍。

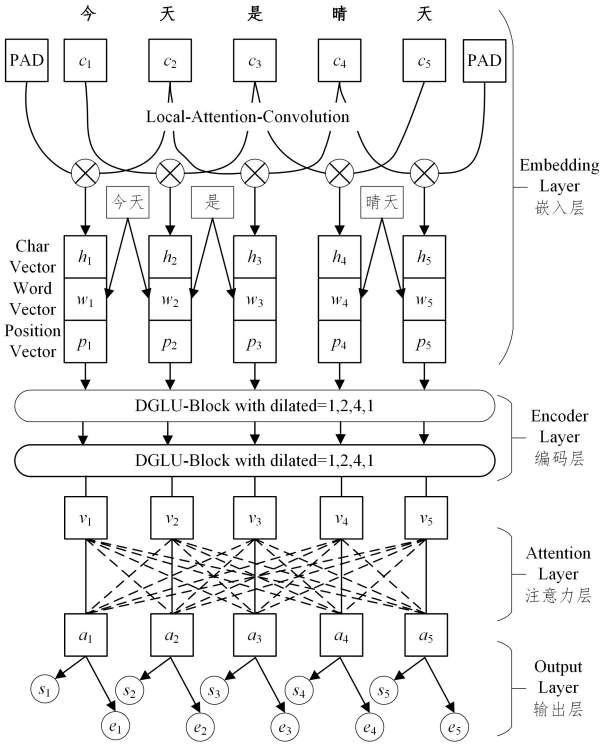

图1 模型总体结构图

Fig. 1 Overall structure of our model

3.1 基于 LAC 的嵌入层

字嵌入算法以字符为基础计算单元，不需要提前对文本分词，是解决分词依赖问题最直观的方法。但字符级别的语义信息区分度差，难以处理单一字符多义的问题，以前文提到的“没有”和“淹没”两词为例，两词中的“没”这一字符语义完全不同但字向量相同，普通字向量无法体现同一字符不同语境下的差异性。因此，嵌入层需要解决的问题就变成了：多义字符的向量化表示如何在不同语境下表现出差异性。

通过对句子结构和字向量进行分析发现，字符周边的局部结构特征可以很好地体现这种差异性。如“没有”这一词中的“没”，其周边结构特征为“没-有”这一连接关系，而“淹没”中的“没”，其周边结构特征为“淹-没”这一连接关系，如果能够将这种结构信息编码到字向量中，就可以体现出多义字在不同语境下的差异性。除了与相邻词的连接关系，也存在更大范围的结构特征，如成语或流行语中一个字的含义往往与周围多个字都有相关性。

为了获取这种结构信息并将其整合到字向量中，我们使用了一种局部注意力卷积（Local Attention Convolution，LAC）结构（见图2）来提取序列的字符间依赖关系，将依赖关系作为结构特征融入字向量。

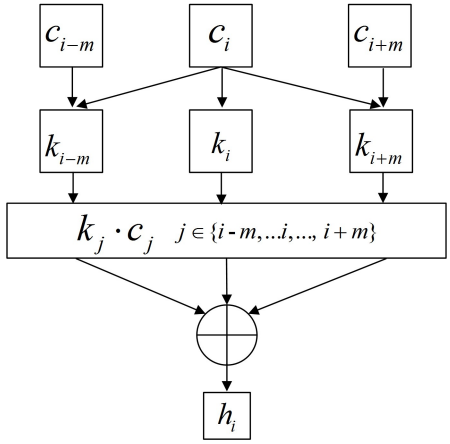

图2 局部注意力卷积

Fig. 2 Local attention convolution

对于通道数为 d_h 的卷积网络，图2以第 i 个字符为中心，大小为 $2m+1$ 的卷积窗为例。将窗内 $2m+1$ 个字向量 $c_{i-m},\cdots,c_i,\cdots,c_{i+m}$ 作为输入，首先计算各个字向量与中心字向量 c_i 的注意力权重 $k_{i-m},\cdots,k_i,\cdots,k_{i+m}$：

$$k_j=\frac{\exp(\langle c_iW_1,c_jW_2\rangle)}{\sum\limits_{n\leqslant i+m}^{n\geqslant i-m}\exp(\langle c_iW_1,c_nW_2\rangle)} \tag{1}$$

其中，$W_1,W_2\in R^{d_c*d_h}$，作为网络权重是可训练参数，$\langle c_iW_1,c_jW_2\rangle$ 表示 c_iW_1 和 c_jW_2 的内积。将注意力权重与对应的字向量相乘并进行加和池化（Sum Pooling），即可得到通道数为 d_h 的局部注意力卷积层在这一卷积窗的输出 h_i，该输出也是 c_i 通过局部注意力卷积层后的对应输出。

$$h_i=\sum_{j\geqslant i-m}^{j\leqslant i+m}(W_3*(k_jc_j)+b) \tag{2}$$

$W_3\in R^{(2m+1)*d_h*d_c}$，$b\in R^{k*d_h}$，都是可训练的参数，$*$ 表示二者之间的对应元素积（Element-Wise Product）。对于整个字向量序列 C，可以得到一个局部注意力卷积层输出序列 $H=(h_1,h_2,\cdots,h_n)$，h_i 即为融合了字符间依赖关系的字向量。

为了使用来自外部的词级别信息，我们将预训练词向量作为特征拼接到新的字向量后面。首先利用自动分词工具对文本序列T进行分词，并利用预训练模型将词转换为固定的词向量，将OOV转换为全零向量，得到一个词向量序列 $(w_1,w_2,\cdots,w_p)$，将词向量拼接到该词包含的字的字向量的后面。

$$l_i=Concat(h_i,w_j),char_i\in word_j \tag{3}$$

其中，$char_i\in word_j$ 表示第 i 个字符，是第 j 个词的一部分。由于模型在后面的编码层没有使用RNN结构，因此模型完全没有位置感，整个模型类似于一个池袋（Bag of Words）模型。因此，本文在嵌入层也使用了位置向量，将序列中的每一个位置也编码为一个向量，并将其融入到字向量中。本文所

用位置向量为 Vaswani 等[6]给出的位置向量构造公式，不需要进行训练，将位置 p 映射为一个 d_{pos} 维的向量 PE 的公式为：

$$\begin{cases} PE_{2i}(p)=\sin(p/10000^{\frac{2i}{d_{pos}}}) \\ PE_{2i+1}(p)=\cos(p/10000^{\frac{2i}{d_{pos}}}) \end{cases} \tag{4}$$

其中，$i\geqslant 0$ 且 $2i+1<d_{pos}$。利用该公式，序列中每一个位置都对应一个位置向量 PE_i。将位置向量也拼接至词向量后面。

$$o_i=Concat(l_i,PE_i) \tag{5}$$

o_i 即为一个字符经过嵌入层后的输出向量，一个文本序列经过嵌入层可以得到一个输出向量序列 $O=(o_1,o_2,\cdots,o_n)$。该输出向量由 3 部分组成：1）经过 LAC 处理的字向量，LAC 整合了字向量和字符周边结构特征信息来获取词级别信息；2）来自预训练模型的词向量，这种词向量包含来自外部的词级别信息；3）位置向量，处理文本这类序列型数据时，位置信息仍然十分重要。

由于嵌入层的输出仍是字向量序列，模型后面的计算都是以字符为基本单位的。相比基于词向量的模型，这种嵌入方法减少了模型对分词结果的依赖，同时将通过两种途径获得的词级别信息有效融入了字向量。

3.2　基于 DGLU 的编码层

文本经过嵌入层处理后，将生成的字向量序列送入基于膨胀门控线性单元（Dilated Gated Linear U-nit，DGLU）的编码层，以提取文本序列的高维特征。

DGLU 是在 Dauphin 等[20]提出的门控线性单元（Gated Linear Unit，GLU）的基础上进行改造得到的。GLU 模仿 LSTM 中的门结构，为卷积网络加入了门控机制，门控机制的形式为：

$$Y=Conv1(X)*sigmoid[Conv2(X)] \tag{6}$$

Conv1 和 Conv2 为两个超参数相同但权重不同的卷积核，$sigmoid[Conv2(X)]$ 的值为 0～1。相比普通卷积，GLU 的输出效果是普通卷积的输出 $Conv1(X)$ 乘以一个 0 到 1 之间的值，从而实现门控，体现了信息的选择性流动。DGLU 将残差[21]（Residual）结构也加入到门控机制中，其结构如图 3 所示，计算公式为：

$$\begin{cases} Y=Conv2(X)*\alpha+X*(1-\alpha) \\ \alpha=sigmoid(Conv1(X)) \end{cases} \tag{7}$$

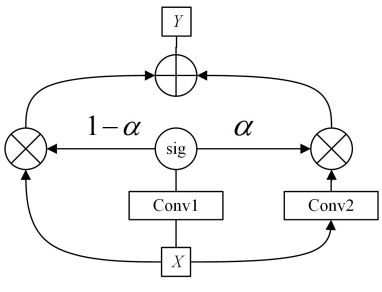

图 3　DGLU 的计算结构

Fig. 3　Calculation structure of DGLU

加入残差结构不仅可以解决梯度消失问题，提高训练稳定性，更重要的是提供了一种信息跨层传播的途径。由于 DGLU 中输入直接连接输出，信息可以实现跨层传播，即信息经过多层网络处理后仍保有一定量的原始输入信息。这种信息跨层传播的机制所起的作用与 LSTM 的记忆门类似，带残差结构的门控机制赋予了 CNN 一定程度的记忆能力。

为了处理长序列，CNN 需要较大的感受野（Rceptive Field）。一般通过两种方法增大感受野：增大卷积核大小（Kernel Size）和堆叠网络。增大卷积核大小会极大地提高计算量，因此主要使用堆叠网络的方法来增大感受野。如果使用普通卷积，网络需要叠得很深才能获得满足要求的感受野，如 QANet[22]使用了 21 层 CNN 作为编码层。网络层数过多不仅意味着模型规模庞大，模型的训练难度也会明显提高，容易出现训练稳定性下降及在小数据集上不收敛的问题。

为了在获取理想感受野的前提下尽量减少模型的层数，我们引入了 Yu 等[23]提出的膨胀卷积，普通卷积与膨胀卷积的对比如图 4 所示。

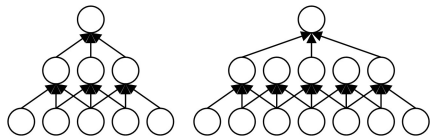

图 4　普通卷积与膨胀卷积的对比

Fig. 4　Comparison between two convolutions

图中左侧为两层普通卷积，右侧为膨胀率（Dilated Rate）依次为 1，2 的两层膨胀卷积。左侧网络的感受野为 5，右侧网络的感受野为 7，这是由于右侧网络第二层跳过了与中心直接相邻的输入，直接连接与中心次邻的输入。膨胀卷积的工作机制为跳过与中心直接相邻的 $p-1$ 个输入，与中心及与中心相邻的第 p 个输入连接，p 为膨胀卷积的膨胀率。图 1 中的 DGLU-Block 使用了卷积核大小为 3、膨胀率依次为 1，2，4，1 的 4 层膨胀卷积，编码层使用了两个 DGLU-Block。模型在使用了 8 层网络的情况下取得了 33 的感受野，使用相同卷积核大小的普通卷积需要堆叠 15 层才能获得相同的感受野。通过使用膨胀卷积，模型可以在网络层数较少的情况下取得较大的感受野。

3.3　注意力层

注意力层用来获取句子级别的信息，可以提取长距离依赖，并对编码层信息进行整合。我们在注意力层使用了全局自注意力（Global Self Attention）机制，形式为：

$$a_i=\sum_{j=1}^{n}k_{i,j}v_j \tag{8}$$

其中，n 为文本长度，v_j 表示第 j 个字符通过编码层后的输出，a_i 表示第 i 个字符通过注意力层后的输出，注意力权重的计算公式为：

$$k_{i,j}=\frac{\exp(\langle v_iW_4,v_jW_5\rangle)}{\sum_{\varepsilon=1}^{n}\exp(\langle v_iW_4,v_\varepsilon W_5\rangle)} \tag{9}$$

$W_4,W_5\in R^{(d_c+d_w+d_{PE})*d_h}$，是可训练参数。全局注意力形式与局部注意力形式基本一致，只是作用域不同，局部注意力仅作用于卷积窗内，全局注意力则作用于整个文本。

3.4　输出层

我们没有在输出层使用主流的 BEMS 标注和条件随机场解码，而是选用了更简单的切片式（span）解码方法。相比 CRF，切片式方法的训练速度和预测速度都更快。

对于注意力层的输出 $A=(a_1,a_2,\cdots,a_n)$，分别通过两个不同的以 sfotmax 函数为激活函数的全连接层(Full Connected Layer，FCL)，得到一个实体首部预测序列 $SP=(sp_1,sp_2,\cdots,sp_n)$和一个实体尾部预测序列 $EP=(ep_1,ep_2,\cdots,ep_n)$：

$$\begin{cases} SP=softmax[FCL_1(A)] \\ EP=softmax[FCL_2(A)] \end{cases} \tag{10}$$

其中，sp_i，ep_i表示第 i 个位置的字符是某类实体首部、尾部的概率，取概率值最大的那一类作为该位置的标签：

$$\begin{cases} s_i=argmax(sp_i) \\ e_i=argmax(ep_i) \end{cases} \tag{11}$$

其中，$s_i,e_i\in\{0,1,\cdots,q\}$，$q$ 为实体类别数，0 表示该位置的字符不是任何一类实体的首部或尾部。我们使用一种切片式解码算法，对 S 和 E 进行处理，将实体从原始文本 T 中提取出来。切片式解码算法利用预测得到的实体首部位置和尾部位置在文本上进行切片以获得实体。这种方法可以取得接近 CRF 的解码效果，且计算量相比 CRF 小了很多。

算法 1 切片式解码算法

输入：T，S，E

输出：Set={$Entity_1$，$Entity_2$，…，$Entity_n$}

```
1. Set←{ }
2. len←len(S)
3. for i←0 to(len−)by 1 do
4.    if S[i]≠0 then
5.       for j←i to(len−1)by 1 do
6.          if S[i]=E[j] then
7.             Entity_k←T[i:j]
8.             Set.append(Entity_k) /*将Entity_k加入集合 Set 中*/
9.       end
10. end
11. final
12. return Set
```

模型训练使用基于交叉熵(Cross Entropy)的损失函数，其形式为：

$$Loss=CE(SP,TrueS)+CE(EP,TrueE) \tag{12}$$

其中，$CE(S,TrueS)$，$CE(E,TrueE)$分别为首部位置预测、尾部位置预测与真实值的交叉熵，将两个交叉熵相加作为模型最终的损失函数。

4 实验及结果分析

为了分析模型的性能和计算速度，我们在多个中文命名实体识别数据集上设置了多组对照实验。

4.1 实验设置

本文在 3 个数据集上进行实验：1) MSRA 数据集是 SIGHAN 提供的中文通用领域数据集，由 20 864 条训练语料、4 636 条测试语料以及 2 318 条验证语料组成，包含“地点”“人物”“机构”3 种实体，标注质量较高；2) Weibo 命名实体识别数据集由 1 890 条标注语料组成，语料来源于微博，口语化程度较高，数据集包含 4 个类型的实体，即 ORG(组织)、PER(人物)、LOC(地点)和 GPE(政治实体)，所有实体还分为明确的命名实体(Named Entity)和泛指实体(Nominal Mention)，如“张三”这一人名，“学生”这一宽泛概念两个大类；3)军事文本数据集是本课题组自行标注的军事专业领域数据集，包含 8 000 条军事新闻语料，语料主要来自环球网、参考消息等主流媒体的军事板块及相关军事爱好者社区，该数据集未对实体分类，仅包含一种“军事相关实体”，但该数据集中专业词汇及生僻词较多，且标注质量一般，我们使用 6 000 条语料作为训练集，剩下的 2 000 条语料作为测试集。

实验代码基于 Keras 框架实现，由于 LSTM 有多种实现方式，且不同实现方式的运算速度差异较大，实验中所使用的 LSTM 实现均为速度最快的 CuDNNLSTM[24]。优化器使用 Radam[25]，初始学习率设置为 1×10^{-3}，其他超参数的设置如表 2 所列。

表 2 超参数设置

Table 2 Hyperparameter setting

参数	数值
字向量维度	256
词向量维度	128
位置向量维度	128
嵌入层卷积核大小	5
编码层卷积核大小	3
Batch-Size	32
Epoch-Num	15

为了避免训练随机性的影响，使用相同超参数对模型进行 30 次训练及测试，取 30 次实验结果的均值作为最终的实验结果。实验结果中，MSRA 数据集和 Weibo 数据集上引用模型的性能数据来源于所引论文，军事文本数据集上的性能数据及所有数据集上的计算耗时数据均来源于实验记录。

实验中主要使用精确率(Precision，P)、召回率(Recall，R)及 Micro-F1 值来评价模型的性能，指标定义如下：

$$P=TP/(TP+FP) \tag{13}$$

$$R=TP/(TP+FN) \tag{14}$$

$$F1=2*P*R/(P+R) \tag{15}$$

其中，P 表示识别正确的正类占识别出的所有正类的比例，R 表示识别正确的正类占样本中真实正类的总数的比例，$F1$ 对 P 和 R 都进行了加权，是结合了二者的一个综合指标。

4.2 模型性能分析

为了分析模型性能，我们在前面提到的 3 个数据集上进行了实验，并将实验结果与当前最优模型进行对比。但 LAC 和 DGLU 主要在嵌入层和编码层进行了改造，当前先进模型的工作也集中在嵌入层和编码层，因此无法将 LAC 或 DGLU 直接应用于现有最优模型以验证本文提出的两种新方法的有效性。为了更全面地分析模型性能，我们分两部分进行性能分析：首先将 LAC-DGLU 与当前最优模型进行对比，以验证模型的整体优越性；其次，利用控制变量的思想，基于经典的 BiLSTM 模型，通过仅将嵌入层的词嵌入方法替换为 LAC 或仅将编码层的 LSTM 替换为 DGLU 的控制变量方法，分别证明了 LAC 和 DGLU 的有效性。

4.2.1 整体性能验证

(1)MSRA 数据集

MSRA 数据集上的实验结果如表 3 所列。Dong 等[18]将汉字拆解为偏旁部首，提取了更细粒度的特征，F1 值为

90.95%；Yang 等[19]进一步将汉字拆解为笔画，F1 值为 91.67%；Cao 等[17]使用一种对抗迁移学习架构来实现分词任务和命名实体识别任务联合训练，F1 值为 90.64%。Zhang 等[16]提出的网格 LSTM 是当前的最优模型，通过字序信息匹配句子中可能出现的词汇，将潜在词信息融合到基于字符的 LSTM-CRF 模型中，在 MSRA 数据集上 F1 值为 93.18%。实验结果显示，本文模型在 MSRA 数据集上的表现超越了现有最优模型，将 F1 值进一步提高了 0.22%。提升幅度很小是由于 MSRA 数据集中的文本比较规范，句子中未收录的词极少，网格 LSTM 无法处理未收录词的缺点在 MSRA 数据集上没有暴露出来。

表 3　MSRA 数据集的实验结果

Table 3　Experimental results on MSRA dataset

（单位：%）

模型	P	R	F1
文献[18]中的模型	91.28	90.62	90.95
文献[17]中的模型	91.30	89.58	90.64
文献[19]中的模型	92.04	91.31	91.67
文献[16]中的模型	**93.57**	92.79	93.18
Our model	93.23	**93.58**	**93.40**

(2)Weibo 数据集

Weibo 数据集上的实验结果如表 4 所列，由于 Weibo 数据集将实体分为命名实体和泛指实体两大类，我们使用明确的命名实体 F1 值（NE）、泛指实体 F1 值（NM）以及总体 F1 值（Overall）作为评价模型性能的指标。He 等[26]提出了一种 F1 值驱动训练方法，将 F1 值的计算公式融入到损失函数中，从而直接以 F1 值驱动模型训练，F1 值为 58.23%。Zhang 等[16]提出的最优模型的 F1 值为 58.79%。实验结果显示，本文模型的总体 F1 值比当前最优模型（Zhang[16]等提出的模型）取得了 0.44%的提高，Weibo 数据集训练数据量小，作为补充信息引入的预训练词向量在此时对模型性能有较大作用。在训练数据量较小的 Weibo 数据集上，本文模型相比其他模型提高并不明显，说明其不适用于小样本训练下的命名实体识别，这一应用限制是由以下几方面造成的：过小的样本量会加剧模型的过拟合现象，样本量越小，模型过拟合得越快；数据集体量小，其包含的信息量本身不足以支撑模型学习到一个完整的行为模式；样本量小时训练集内的样本类别不平衡现象会被进一步放大，容易出现模型对不同类别实体识别差异较大的问题。

表 4　Weibo 数据集的实验结果

Table 4　Experimental results on Weibo dataset

（单位：%）

模型	NE	NM	Overall
文献[26]中的模型	54.40	62.17	58.23
文献[17]中的模型	54.34	57.35	58.70
文献[16]中的模型	53.04	62.25	58.79
Our model	**55.18**	**62.65**	**59.23**

(3)军事文本数据集

为了获取当前最优模型在军事文本数据集上的表现，我们复现了 Zhang 等[16]、Cao 等[17]的模型并在军事命名实体识别数据集上进行了实验。军事文本数据集上的实验结果如表 5 所列，与 Cao 等[17]提出的模型相比，本文模型取得了 6.32%的提高，这是由于 Cao 等[17]的模型性能严重依赖于分词任务的效果，军事文本数据集中包含的大量专业词汇会导致分词出现较多错误；与 Zhang 等[16]提出的模型相比，本文模型取得了 2.06%的提高，Zhang 等[16]提出的模型对分词错误问题有一定的应对能力，但只能处理已收录词，无法处理军事文本中未被外部词典收录的部分专业术语。

表 5　军事文本数据集上的实验结果

Table 5　Experimental results on Military-text dataset

（单位：%）

模型	P	R	F1
文献[17]中的模型	58.07	61.32	59.65
文献[16]中的模型	63.50	64.33	63.91
Our model	**65.27**	**66.69**	**65.97**

4.2.2　方法有效性验证

为了验证 LAC 和 DGLU 两种新方法的有效性，我们以命名实体识别领域最流行的双向 LSTM（Bidirectional Long Short Term Memory Network，BiLSTM）模型为基线模型，设置了对照实验。通过在基础模型上分别独立使用 LAC 和 DGLU，递进式地证明两种新结构的有效性，对照组实验设置如下：1）Baseline 模型基于词嵌入方法和双向 LSTM，使用了全局注意力机制，是命名实体识别领域的主流模型；2）Baseline-DGLU 模型将 Baseline 模型中编码层的双向 LSTM 替换为膨胀门控线性单元，用于验证 DGLU 的有效性；3）Baseline-LAC 模型将 Baseline 模型中的词嵌入方法替换为本文提出的基于局部注意力卷积的字嵌入方法，用于验证 LAC 的有效性。

表 6　对照实验结果

Table 6　Control experimental results

（单位：%）

模型	MSRA 数据集			Weibo 数据集			军事文本数据集		
	P	*R*	*F*1	NE	NM	Overall	*P*	*R*	*F*1
Baseline	91.37	88.66	89.99	49.32	59.25	54.22	57.13	60.05	58.55
Baseline-DGLU	90.55	89.97	90.26	49.86	59.55	54.35	58.22	60.27	59.23
Baseline-LAC	93.21	92.95	93.08	54.98	62.37	58.44	64.33	66.31	65.30
Our Model	**93.23**	**93.58**	**93.40**	**55.18**	**62.65**	**59.23**	**65.27**	**66.69**	**65.97**

对比实验结果如表 6 所列，实验结果显示，本文模型相比 Baseline 在 3 个数据集上的 F1 值取得了 3%～7%的提高，效果明显。通过观察对照组实验可以发现，Baseline-LAC 相比 Baseline 取得了接近 7%的提高，这是由于 LAC 使用字嵌入回避了分词错误问题，同时在字向量中融入了多种来源的词级别信息，字向量中的信息量甚至高于预训练词向量。Baseline-DGLU 的实验结果则与 Baseline 比较接近，DGLU 更多的是模仿 LSTM 功能对 CNN 进行改造，其功能与 LSTM 基

本一致，因此对模型性能的提升意义不大。本文模型性能的提高主要来源于 LAC，DGLU 对于模型性能的提升意义较小。对比不同数据集上的结果，使用 LAC 带来的提升效果也有一定区别，LAC 在军事文本数据集上的提升效果最大，达到了 7.42%，这是由于军事文本数据集中存在大量的专业词，导致分词任务出现错误的概率增大，LAC 通过回避分词错误问题有效提高了模型性能。MSRA 数据集上 LAC 带来的提升效果较小，这是由于 MSRA 数据集中的文本形式比较正式，分词错误的情况较少，但使用 LAC 得到的字向量比使用普通词嵌入得到的词向量包含了更丰富的语义信息，因此也取得了 3.41%的提高。

由于专业文本和社交媒体文本中含有大量的未收录词，本文模型在这些数据集上可以取得比在通用领域文本数据集上更大的提高。

表 7 MSRA 数据集实体分类别实验结果

Table 7 Sub-category experimental results on MSRA dataset

（单位：%）

模型	地点			人物			机构		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
Baseline	92.52	90.03	91.26	91.22	88.69	89.94	89.02	85.55	87.25
Baseline-DGLU	92.61	91.01	91.80	89.86	89.12	89.49	86.10	88.94	87.50
Baseline-LAC	93.68	93.39	93.53	**93.17**	92.61	92.89	92.22	92.48	92.35
Our Model	**93.86**	**94.25**	**94.05**	92.74	93.09	92.91	92.57	92.83	92.70

由于 Weibo 数据集的测试集小且实体分类后单类别的样本量少，而军事文本数据集没有对实体进行分类，因此我们只对 MSRA 数据集上的实验结果进行实体分类统计，分析本文方法对不同类别实体的作用。

实验结果如表 7 所列，结果显示本文模型对 3 种类型的实体都有一定的提升效果，对实体类别也有较强的适应性。在 3 个类别中，本文模型对机构类实体的提升效果最明显，提高了 5.45%。这是由于机构形式多变，用词灵活，易出现分词错误，且有一定量的未收录机构实体；人物名称虽然同样灵活多变，但 MSRA 数据集中的人物实体几乎全部为知名公众人物，外部词典多有收录，分词工具能够将这些人名正确分出。本文模型相比 Baseline 也更为平衡，Baseline 上的实验结果中，3 类实体上的准确率和召回率的差值较大，本文模型在 3 类实体上的准确率和召回率都较为接近。

4.3 计算速度分析

MSRA 数据集中文本长度大多小于 100，可以作为短文本代表性数据集；军事命名实体数据集中文本长度多为 100～300，可以作为长文本的代表性数据集。我们在上述两个数据集上进行实验并记录了训练过程中一个 Step 即模型处理一个批次样本并更新一次权重的耗时，对模型的计算速度进行对比分析。

单个 Step 耗时对比如表 8 所列，我们复现了两个当前先进的模型与本文模型进行对比，表 8 中的速度比表示该模型与最慢模型的计算速度比值。结果显示，本文模型的训练速度达到了当前最优模型的 1.44～1.87 倍，加速效果明显。在军事文本数据集上的加速效果优于在 MSRA 数据集上的加速效果，这是由于军事文本数据集中文本序列长度较大，LSTM 无法在样本内并行计算的缺点在处理长序列时被进一步放大。由于 Zhang 等[16]、Cao 等[17]的模型都是基于 LSTM 实现的，这两个模型在长文本数据集上的耗时与短文本数据集上的耗时相比涨幅较大，达到了 2.06 倍。由于局部注意力卷积和膨胀门控线性单元都可以并行计算，本文模型的涨幅只有 1.58 倍，对不同长度的文本数据有更好的适应性。

表 8 单个 Step 耗时的对比

Table 8 Time consumption comparison of single step

模型	MSRA 数据集		军事文本数据集	
	耗时/ms	速度比	耗时/ms	速度比
文献[16]中的模型	89	1×	183	1×
文献[17]中的模型	82	1.09×	164	1.12×
Our model	62	**1.44×**	98	**1.87×**

为了准确分析本文提出的方法对模型计算速度的影响，我们设置了与性能验证实验中相同的对照组，以便分别独立分析 LAC 和 DGLU 所带来的效果。对照组单个 Step 耗时的对比结果如表 9 所列，可以发现 Baseline-DGLU 的速度达到了 Baseline 的 1.42～1.80 倍，使用 DGLU 替换 LSTM 可以显著提高模型的训练速度。Baseline-LAC 模型的速度为 Baseline 的 1.21～1.33 倍，LAC 机制增加了计算量，造成了一定量的训练耗时增加，但局部注意力仅仅计算卷积窗序列上的注意力权重，且各个卷积窗上的计算可以并行计算，因此训练时间的增加较少，处于可接受的范围。本文模型在 MSRA 数据集上与 Baseline 的耗时接近，但在军事文本数据集上耗时仅为 Baseline 的 0.72 倍，这是由于 LAC 的局部注意力卷积和 DGLU 的膨胀门控线性单元都可以在序列上的多个位置同时进行计算，耗时不会伴随序列长度的增加而大幅提高，而 Baseline 基于 LSTM，在处理序列时只能在序列的一个位置上进行计算，耗时会伴随着序列长度的增加快速增长。

表 9 本文模型与对照组单个 Step 耗时的对比

Table 9 Time consumption comparison of single step between the proposed model and control group

模型	MSRA 数据集		军事文本数据集	
	耗时/ms	速度比	耗时/ms	速度比
Baseline	64	1×	137	1×
Baseline-DGLU	45	**1.42×**	76	**1.80×**
Baseline-LAC	85	0.75×	166	0.83×
Our Model	62	1.03×	98	1.40×

结束语 本文提出了一种基于局部注意力卷积的卷积算法，且在卷积神经网络的基础上引入了门控机制和膨胀卷积，构造了一种基于 CNN 和注意力机制的命名实体识别模型。

该模型适用于训练数据充足情况下的命名实体识别任务，相比经典的 BiLSTM 模型及当前最优模型，其 F1 值和计算速度都有一定提高，且在专业领域数据集上的提高尤为明显。后续工作将对本文方法与预训练模型的结合展开研究，利用更高质量的外部信息，搭建适用于小样本命名实体识别任务的模型。

参 考 文 献

[1] LEVY O，GOLDBERG Y. Neural word embedding asimplicit matrix factorization[C]// Advances in Neural Information Processing Systems. 2014：2177-2185.

[2] MIKOLOV T，KARAFIÁT M，BURGETL，et al. Recurrent neural network based language model[C]// Eleventh Annual Conference of the International Speech Communication Association. 2010.

[3] YAO K，PENG B，ZWEIG G，et al. Recurrent conditional random field for language understanding[C]// 2014 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2014：4077-4081.

[4] SUNDERMEYER M，SCHLÜTER R，NEY H. LSTM neural networks for language modeling[C]// Thirteenth Annual Conference of the International Speech Communication Association. 2012.

[5] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv：1508. 01991，2015.

[6] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017：5998-6008.

[7] LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural Architectures for Named Entity Recognition[C]// Proceedings of NAACL-HLT. 2016：260-270.

[8] SANG E F，DE MEULDER F. Introduction to the CoNLL-2003 shared task：Language-independent named entity recognition [J]. arXiv：cs/0306050，2003.

[9] STRUBELL E，VERGA P，BELANGER D，et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017：2670-2680.

[10] SHEN T，ZHOU T，LONGG，et al. Disan：Directional self-attention network for rnn/cnn-free language understanding [C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[11] PETERS M E，NEUMANN M，IYYER M，et al. Deep contextualized word representations[C]// Proceedings of NAACL-HLT. 2018：2227-2237.

[12] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810. 04805，2018.

[13] FENG Y H，YU H，SUN G，et al. Named Entity Recognition Method Based on BLSTM[J]. Computer Science，2018，45(2)：261-268.

[14] SHAN Y D，WANG H J，HUANG H，et al. Study on Named Entity Recognition Model Based on Attention Mechanism—Taking Military Text as Example[J]. Computer Science，2019，46(S1)：111-114，119.

[15] PENG N，DREDZEM. Named entity recognition for chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015：548-554.

[16] ZHANG Y，YANG J. Chinese NER Using Lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). 2018：1554-1564.

[17] CAO P，CHEN Y，LIUK，et al. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018：182-192.

[18] 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，Cham，2016：239-250.

[19] YANG F，ZHANG J，LIU G，et al. Five-Stroke Based CNN-BiRNN-CRF Network for Chinese Named Entity Recognition [C]// CCF International Conference on Natural Language Processing and Chinese Computing. Springer，Cham，2018：184-195.

[20] DAUPHIN Y N，FAN A，AULIM，et al. Language modeling with gated convolutional networks[C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org，2017：933-941.

[21] 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.

[22] YU A W，DOHAN D，LUONG M T，et al. Qanet：Combining local convolution with global self-attention for reading comprehension[J]. arXiv：1804. 09541，2018.

[23] YU F，KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv：1511. 07122，2015.

[24] BRAUN S. LSTM Benchmarks for Deep Learning Frameworks [J]. arXiv：1806. 01818，2018.

[25] LIU L，JIANG H，HE P，et al. On the variance of the adaptive learning rate and beyond[J]. arXiv：1908. 03265，2019.

[26] HE H，SUN X. A unified model for cross-domain and semi-supervised namd entity recognition in chinese social media[C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017.

ZHAO Feng，born in 1997，postgraduate. His main research interests include natural language processing and so on.

HUANG Jian，born in 1971，Ph.D，professor，Ph.D supervisor. Her main research interests include complex system modeling and so on.