文章编号：1003-0077（2017）00-0000-00

基于神经网络的片段级中文命名实体识别

王蕾，谢云，周俊生*，顾彦慧，曲维光

（南京师范大学 计算机科学与技术学院，江苏 南京 210046）

摘要：命名实体识别是自然语言处理的一个重要基础任务。传统基于统计学习模型的命名实体识别方法严重依赖特征工程，特征设计需要大量人工参与和专家知识，而且已有的方法通常大多将中文命名实体识别任务看作一个字符序列标注问题，需要依赖局部字符标记区分实体边界。为了减弱系统对人工特征设计的依赖，避免字符序列化标注方法的不足，本文对基于神经网络的片段级中文命名实体识别方法进行探索研究。通过采用深度学习片段神经网络结构，实现特征的自动学习，并通过获取片段信息对片段整体分配标记，同时完成实体边界识别和分类。基于神经网络的片段级中文命名实体识别方法在 MSRA 数据集上对人名、地名和机构名识别的总体 F1 值达到了 90.44%。

关键词：深度学习；神经网络；片段级中文命名实体识别

中图分类号：TP391 **文献标识码：**A

Segment-level Chinese Named Entity Recognition Based on Neural Network

Lei Wang, Yun Xie, Junsheng Zhou, Yanhui Gu, Weiguang Qu

(School of Computer Science and Technology, Nanjing Normal University，Nanjing, Jiangsu 210046, China)

Abstract: Chinese Named entity recognition (NER) is an important task for Chinese information processing. Traditional statistical machine learning methods often rely on hand-crafted features. However, designing feature templates is dependent heavily on human creativity and prior knowledge. Additionally, the character-level sequence labeling models are bounded by local tag dependencies. In this paper, in order to eliminate the need for most feature engineering and solve the limitation o f the character-based sequence labeling models, we attempt to exploit some segmental neural architectures for Chinese NER, by regarding the task as a joint segmentation and labelling problem. The proposed methods can learn the effective segment-level representation and contextual information and then assign tags to the segments. The experimental results on MSRA corpus show that our neural segment-level models can achieve comparable performance with the state-of-the-art systems for Chines NER.

Key words: deep learning; neural network; segment-level Chinese named entity recognition

1 引言

命名实体识别（NER）是指从文本中识别出人名、地名和机构名等专有名词，是自然语言处理的关键技术之一，也是信息抽取、问答系统、句法分析、机器翻译等应用的重要基础工作[1]。随着互联网的飞速发展和大数据时代的到来，文本数据规模更大，领域更多，内容更复杂。探索更具实用性的新的有效识别方法，成为学术界和

收稿日期：2017-06-10；**定稿日期**：2017-07-25

基金项目：国家自然科学基金(61272221，61472191)，江苏省高校自然科学研究项目(15KJA420001)，福建省信息处理与智能控制重点实验室(闽江学院)开放基金项目(MJUKF201705) ，山东省语言资源开发重点实验室开放课题(211180A41601).

*通讯作者

工业界关注的热点问题。

目前，解决命名实体识别问题的主流方法是基于统计学习模型的方法，包括基于最大熵（ME）模型、隐马尔可夫（HMM）模型、条件随机场（CRF）模型等命名实体识别方法[2-4]。传统方法通常依赖特征工程保证系统性能。然而，特征模板的制定需要人工设计和大量专家知识。特征设计需要实验进行反复修改、调整和选择，非常费时费力。传统方法中数据采用稀疏表示，容易导致参数爆炸问题。在面对大规模多领域复杂的文本数据时，传统方法则暴露出更多不足。

对于中文命名实体识别任务，现有的方法通常将该任务看作一个字符序列标注问题，通过对字符分配标记完成命名实体识别[5][6]。由于中文句子中单词间没有分隔符号，相比于字符序列标注模型，直接对中文句子中的片段进行标记分配更为合理，可以避免字符序列标注方法中依赖局部标记区分实体边界的问题。Zhou 等人提出中文命名实体边界识别与实体类别识别集成的算法模型，引入片段特征解决中文命名实体识别问题[7]。但该方法采用传统统计学习模型，仍然严重依赖具体任务的特征工程。

近几年，深度学习为解决自然语言处理问题提供了一种新的方法和途径，受到广泛关注。深度学习可以实现特征的自动学习，采用低维、稠密的实值向量表示数据，避免对人工和专家知识的严重依赖。基于深度学习的命名实体识别方法受到关注，其中，Collobert 和 Weston 构建 SENNA 系统为多项自然语言处理任务提供统一的神经网络底层结构，包括命名实体识别任务[8]；Turian 等人使用神经网络预先训练的词向量作为额外特征，与传统基于 CRF 的方法结合解决命名实体识别问题[9]；Lample 等人针对命名实体识别任务提出双向长短期记忆模型（Bi-LSTM）和 CRF 模型的组合结构[10]；Ma 等人将 Bi-LSTM、卷积神经网络（CNN）与 CRF 模型结合构建了序列标记模型[9]；Chiu 和 Nichols 利用 Bi-LSTM 和 CNN 对输入信息进行处理，完成命名实体识别任务[11]；Liu 等人以片段信息表示作为输入，采用神经网络与半马尔可夫条件随机场（semi-CRF）模型结合完成英文命名实体识别任务[12]。目前，基于神经网络的中文命名实体识别研究较少，且主要采用字符序列标注模型[13]，还没有基于神经网络的片段级中文命名实体识别研究工作。

因此，我们主要对基于神经网络的片段级中文命名实体识别方法进行探索研究，减弱对人工特征设计和专家知识的依赖，避免字符化序列标注模型的不足。在 Liu 等人的研究工作[12]基础上，我们结合中文语言特性和中文命名实体识别任务的特点，除片段内部字符和片段整体表示之外，引入离散特征与稠密向量表示结合的片段扩展特征表示，改进解码算法获取片段级上文信息，通过对片段整体分配标记完成中文命名实体识别任务。

2 基于神经网络的片段级中文命名实体识别

中文句子中词与词之间没有分隔符号，中文命名实体识别需要完成实体边界识别和实体分类任务。片段级的中文命名实体识别方法基于片段获取表示信息，对于输入的句子序列进行片段切分并对切分序列中的片段整体进行标记分配。相比于字符序列化标注方法，对片段整体进行标记分配更为合理，可以避免识别过程中依赖局部标记来区分实体边界的问题。

我们采用“PER”、“LOC”和“ORG”分别表示人名、地名和组织机构名。以句子“中华人民共和国主席习近平在北京接受中央电视台采访。”为例作为输入序列，对片段分配标记后为“中华人民共和国/LOC 主席/O 习近平/PER 在/O 北京/LOC 接受/O 中央电视台/ORG 采访/O 。/O”。例子中，“中华人民共和国”、“主席”、“习近平”等看作是句子序列中的片段。在标记集合 $T = \{\text{PER, LOC, ORG, O}\}$中选取具体的标记分配给当前片段。

Semi-CRF 模型是一种典型的对片段整体分配标记的方法[14]，但基于 semi-CRF 的命名实体识别方法具有传统统计学习模型的不足。因此，选用基于神经网络和 semi-CRF 结合的片段神经网络结构实现特征的自动学习，可以避免繁琐的人工特征设计和对大量语言先验知识的依赖。

对于输入的句子序列x，有相应的切分片段序列$s = (s_1, s_2, \dots, s_p)$。对于片段$s_j = < u_j, v_j, y_j >$，其中$u_j$表示片段起始字符在句子中的下标，$v_j$表示片段结尾字符在句子中的下标，$y_j$表示片段的标记。处理该片段时，基于片段$s_j$的信息表示作为当前神经网络模型的输入，通过神经网络计算获得当前片段的抽象表示向量代替传统方法中的稀疏特征向量。Liu 等人主要考虑片段内部单元和片段整体信息[11]。我们引入片段相关扩展特征，

从片段内部字符单元（E_{unit}）、片段整体（E_{seg}）和片段相关扩展特征表示信息（F_{extend}）三个方面，结合稠密向量表示和离散特征获取当前片段信息，模型结构如图 1 所示。

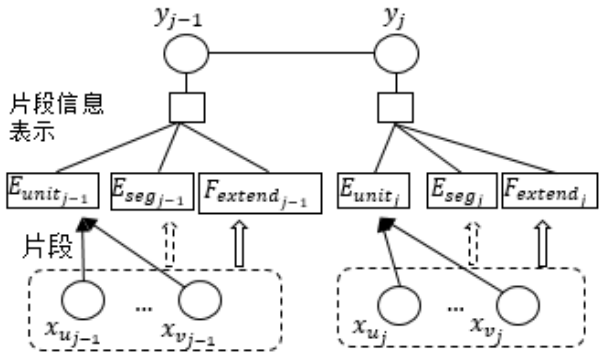


图 1 片段级中文命名实体识别模型结构

具体的，我们研究两种神经网络模型结构：（1）Bi-LSTM 和标准神经层构成的神经网络结构；（2）Bi-LSTM、Bi-RNN 和标准神经层构成的组合神经网络结构。

2.1 基于 Bi-LSTM 的片段级中文命名实体识别

对于输入的句子$x=(x_0,x_1,\dots x_n)$，通过查表（lookup）获得每个字符x_i的向量表示e_{x_i}，若字符向量表中不存在当前字符，则采用特殊符号“UNK”的字符向量表示，“UNK”取随机值初始化。字符向量表示序列$e=(e_{x_0},e_{x_1},\dots,e_{x_n})$作为 Bi-LSTM 的初始输入，经过模型编码获得字符的抽象表示。其中，对于字符x_i，通过前向 LSTM 计算获得向量表示$\overrightarrow{C_i}$，通过后向 LSTM 计算获得向量表示$\overleftarrow{C_i}$。将前向 LSTM 和后向 LSTM 的输出结果进行连接计算得到输出向量C_i，如式(1)所示：

$$C_i = relu(W_C[\overrightarrow{C_i};\overleftarrow{C_i}]+b_C) \quad (1)$$

其中，$[\overrightarrow{C_i};\overleftarrow{C_i}]$表示前向 LSTM 输出向量$\overrightarrow{C_i}$和后向 LSTM 输出向量$\overleftarrow{C_i}$连接构成一个向量，$W_C$是权值矩阵，$b_C$是偏置项。

片段内字符单元的向量表示按序连接形成片段内部特征表示E_{unit_j}，即对于片段s_j，E_{unit_j}具体表示如式(2)所示：

$$E_{unit_j} = [C_{u_j};C_{u_j+1};\dots;C_{v_j}] \quad (2)$$

其中，[; ;…;]表示各个向量依次连接构成一个向量。

由于切分片段序列中的片段长度不统一，为了使输入下一层计算的向量长度固定，模型设置最大片段长度为L。设d_C表示向量C_i的维数，若当前片段长度小于L则对E_{unit_j}向量进行末尾填充至长度为$D=L\times d_C$维的向量。

片段s_j的整体向量表示E_{seg_j}通过 lookup 操作从片段向量表中获得，如果片段向量表中不存在当前片段的向量，则选用特殊符号“UNKSEG”的向量表示，“UNKSEG”的初始向量取随机值。

片段相关的其他特征向量表示F_{extend_j}主要包含片段长度信息和片段上文已完成切分的片段相关信息，当前处理片段的前文切分片段通过查询片段向量表获得，片段长度特征向量通过查询片段长度特征向量表获得。通过神经网络模型处理输出片段的最终表示E_{s_j}，如式(3)所示：

$$E_{s_j} = relu(W_S\left[E_{unit_j};E_{seg_j};F_{extend_j};E_{y_j}\right]+b_S) \quad (3)$$

式(3)中，[; ;]表示其中各向量连接构成一个向量，W_S是权值参数，b_S是偏置项，E_{y_j}是标记y_j的向量表示。E_{s_j}是当前片段s_j通过神经网络模型输出的特征表示，也是替代传统基于 semi-CRF 模型的方法中片段特征表示的向量。图 2 是神经网络模型获得片段表示的具体结构。

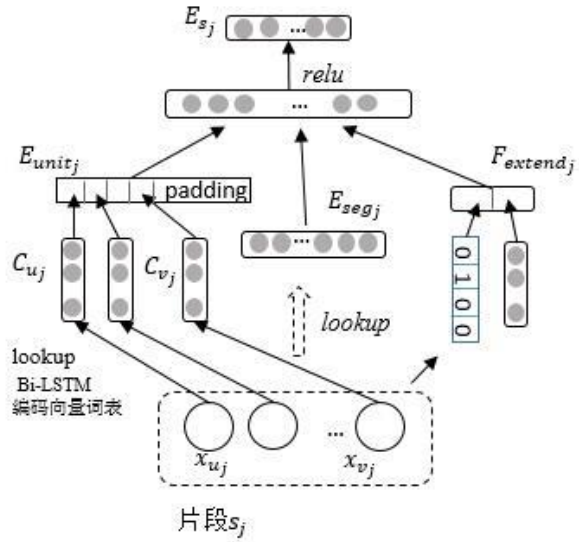


图 2 基于 Bi-SLTM 的神经网络获取片段向量的结构图

模型训练采用极大似然估计，神经网络模型优化选用 SGD 算法，初始学习率设为η_0，正则化方法采用 dropout 技术。预测过程中，处理当前切分片段时，通过神经网络模型获取片段信息的

向量表示，结合 semi-CRF 模型进行解码。

2.2 基于组合神经网络的片段级中文命名实体识别

为了避免向量填充（padding），减少人工设置参数对系统的影响和限制，我们进一步研究采用 Bi-LSTM 模型与其他神经网络模型的组合模型结构获取片段信息。随着不同的神经网络模型的组合和模型结构的加深，模型对输入的信息表示可以获得更抽象的特征信息，模型的刻画能力更强。双向循环神经网络（Bi-RNN）是序列模型，能考虑上下文信息，因此我们选用 Bi-LSTM、Bi-RNN 和普通神经层的组合神经网络结构。

句子序列$x=(x_0,x_1,\dots x_n)$相应的字符向量序列$e=(e_{x_0},e_{x_1},\dots,e_{x_n},)$，字符$x_i$通过 Bi-LSTM 模型进行编码，获得最终输出向量C_i，具体计算如式(1)所示。切分片段$s_j=<u_j,v_j,y_j>$相应的字符向量序列$C=(C_{u_j},\dots,C_{v_j})$作为 Bi-RNN 的初始输入，通过前向循环神经网络获得输出向量$\overrightarrow{S_j}$，通过后向循环神经网络获得输出向量$\overleftarrow{S_j}$，前向与后向输出连接构成片段s_j内部单元信息表示向量E_{unit_j}。获得E_{unit_j}的模型结构如图 3 所示。

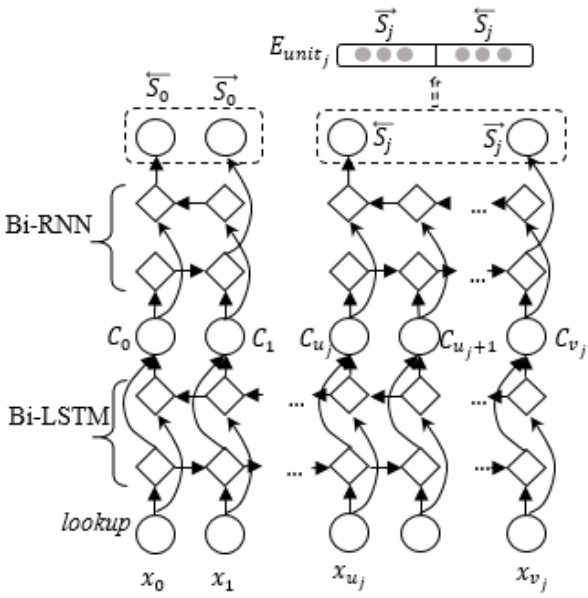


图 3 组合神经网络获得E_{unit_j}的模型结构图

对于当前片段s_j，通过 lookup 操作从片段向量表中获得该片段整体向量表示E_{seg_j}，若当前片段在片段向量表中不存在，则选取特殊符号“UNKSEG”的向量表示，“UNKSEG”的初始值选取随机值。

片段相关的其他特征向量表示F_{extend_j}主要是包含片段上文切分片段相关信息和片段本身长度信息的特征。处理当前片段时，对于前文切分产生的片段通过查询片段向量表获得前一个切分片段的向量表示，若片段向量表中不存在查询的片段，则选用特殊符号“UNKPSEG”的向量表示，“UNKPSEG”取随机值初始化。片段长度特征信息通过查询片段长度特征表获得，每个长度值对应唯一的长度表示向量，初始向量值为随机值。

基于当前片段获取的信息表示，通过神经网络模型输出片段的最终表示E_{s_j}，具体计算如式(3)所示。E_{s_j}是对于当前片段s_j通过神经网络模型输出的片段信息表示向量。图 4 是获得片段向量表示的组合神经网络模型结构。

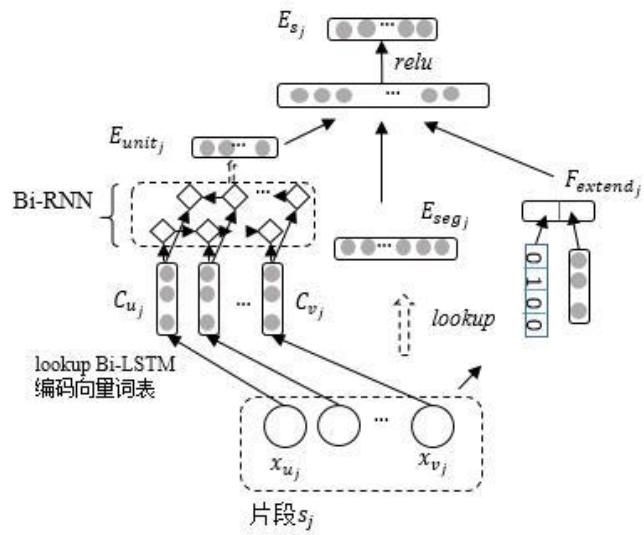


图 4 获取片段表示向量的组合神经网络结构

模型训练采用似然估计，选用 SGD 优化算法，初始学习率设为η_0，正则化方法采用 dropout 技术。预测时，与传统 semi-CRF 方法中的解码算法结合获得句子的切分片段序列和相应的片段标记序列。

2.3 片段特征表示

2.3.1 片段内部字符单元特征

中文字符是构成中文句子的最小单元，也是片段内部的基本组成单元。对于当前处理片段，针对片段内部组成单元即各字符信息提取的特征表示，本文称为片段内部单元特征，记为E_{unit}。

具体实现过程中，对于输入序列x，序列中的每个元素x_i有相应的字符向量表示e_{x_i}，字符x_i通过 Bi-LSTM 编码计算后得到向量表示C_i。对于片段$s_j =< u_j, v_j, y_j >$，其内部单元对应于输入序列中的字符子序列$(x_{u_j}, x_{u_j+1}, \dots, x_{v_j})$，经过 Bi-LSTM 编码计算输出字符表示后，片段s_j内部字符单元相应的向量序列为$C_j = (C_{x_{u_j}}, C_{x_{u_j+1}}, \dots, C_{x_{v_j}})$。片段内部单元特征表示向量通过对序列中所有向量进行处理获得。根据不同的网络结构，向量的处理方法不同。

2.3.2 片段整体特征

为了从片段整体获取片段语义信息，我们采用低维、稠密的片段向量表示片段整体，称为片段整体特征，记为E_{seg}。

对于当前处理的片段$s_j =< u_j, v_j, y_j >$，模型将片段看作一个整体获取向量表示，即片段级的向量。具体是通过查表获得当前片段的向量表示，若不存在查询的片段则选用“UNKSEG”符号的向量，取随机值初始化。片段向量表中初始的片段级词向量是利用无标注的预训练语料通过预训练获得。

2.3.3 片段相关扩展特征

中文命名实体的上下文信息具有相应的特点。如“老师”、“书记”等词常出现于人名的上下文中，“奔赴”、“境内”等词常出现在地名的上下文中。为了获取更丰富的片段信息，我们在当前片段信息基础上，引入上文片段信息。结合片段长度信息，将离散特征与稠密向量表示结合构成片段相关扩展特征，记为F_{extend}。

具体的，由于处理到当前片段时下文还未进行切分，所以我们关注当前处理片段的上文信息，选取当前处理片段的前一个切分片段。通过查询预先训练的片段向量表获取向量表示，若不存在当前片段，则采用特殊符号“UNKPSEG”的向量，该符号向量选取随机值初始化。关于片段长度特征则构建额外的特征向量表，不同长度对应唯一的离散特征向量。上文片段向量与长度特征向量连接构成F_{extend}。

2.4 解码算法

片段表示引入上文片段信息时，采用传统 semi-CRF 的解码算法无法满足获取前一个已切分片段的信息[13]。解码算法需要将原解码过程中的 0 阶动态规划算法修改为 1 阶动态规划算法[15]，使得在子问题计算过程中，当前片段的前一个切分片段的信息可见。图 5 给出了算法的简要描述。

Inputs: raw sentence *sent* (length *n*)
Variables: *chart*[*b*, *e*, *t*] stores the best scored (partial)
segmentation of the characters from the beginning of the sentence to character *e*, with the last segment spanning over the characters from *b* until *e*;
character index *b* for the start of the segment;
character index e for the end of the segment;
character index p for the start of the previous segment;
tag index *t* for the tag assigned to the segment;
tag index *tp* for the tag assigned to the previous segment;
upper bound on segment length *L*.
Algorithm:
for *e* = 0... *n*-1:
  for *t* = 0...|T|:
    for *b* = max{0, *e*-*d*}...*e*, *d*=1…*L*:
      chart[*b*, *e*, *t*] ← the highest scored segmentation among those derived by combining *chart*[*p*, *b*-1, *tp*] with *sent*[*b*, *e*], for *p* = 0...*b*-1
Outputs: the highest scored segmentation among *chart*[*b*, *n*-1, *t*], for *b* = 0... *n*-1

图 5 片段级中文命名实体识别方法 1 阶动态规划解码算法

3 相关工作比较

近十几年来，对于中文命名实体识别研究主要基于传统统计学习模型，通常将任务看作一个字符序列标注问题。如廖先桃讨论了中文命名实体识别的几种方法[2]，包括规则、HMM、ME 和 CRF。史海峰以 CRF 模型为基础实现在字一级对于命名实体的识别[5]。对于中文命名实体识别任务，对片段整体分配标记更为合理，可以避免字符序列化标注方法需要依赖局部标记区分实体边界的问题。Zhou 等人提出中文命名实体边界识别与类别识别集成的算法模型[7]，引入片段级特征，同时完成实体边界识别和类别识别两个子任务。但该方法仍然基于传统统计学习模型，依赖具体任务相关的特征工程。

为了避免具体任务的特征工程，Kong 等人将神经网络与 semi-CRF 结合，提出一种片段级的循环神经网络（SRNN）模型，对于输入序列完成片段切分和片段标记分配[16]。Liu 等人在 Kong 等人的研究基础上提出 SCONCATE 模型，采用片段级神经网络结构，通过获取片段内部字符特征表示和片段整体表示对片段分配标记，解决英文命名实体识别问题。

目前还没有基于神经网络的片段级中文命名实体识别研究。由于中文句子单词间没有明显分隔符号，相比于英文命名实体识别，中文命名实体更加复杂且缺少明显的词形变化等特征，任务更困难。只考虑字符或当前片段表示不能很好的解决中文命名实体识别问题。为了更有效的获取片段信息，我们引入离散特征与稠密向量表示结合的片段扩展特征表示，改进解码算法获取片段级上文信息，通过对片段整体分配标记完成中文命名实体识别任务。

4 实验

4.1 数据

实验数据使用 MSRA 语料，基于神经网络的片段级中文命名实体识别模型利用 MSRA 训练集进行训练，在 MSRA 测试集上完成测试。针对语料在实验中的实际应用，首先对训练集进行相应的语料预处理工作。将训练集中的句子转化为“训练集句子-片段标记序列”作为模型输入的训练数据集。模型的测试集是 MSRA 测试集，是不包含任何切分信息和标记信息的中文句子。

关于模型初始输入的字符向量和片段向量，我们采用 word2vec 工具对无标注语料进行预训练[17]。初始输入向量预训练的语料集额外引入新华社 2000-2004 年和人民日报 2000 年语料。向量预训练语料规模主要分为两种：（1）MSRA 训练集；（2）MSRA 训练集、新华社和人民日报共 6 年语料数据集合。以上两种预训练语料记为 pre1 和 pre2。

4.2 参数设置

实验包含多个超参数，关于神经网络模型的超参数设置具体数值如表 1 所示。

表 1 用于实验的神经网络模型超参数设置

序号	超参数名	值
1	字符向量维数	100
	片段向量维数	50
	Bi-LSTM 层输出向量维数	100
	标记向量维数	20
	输出片段向量维数	100
	初始学习率	0.1
2	Bi-RNN 输出向量维数	100
3	片段长度特征向量维数	4
	前一个片段的特征向量维数	50

表 1 中，第 1 组超参数是基于 Bi-LSTM 的片段级中文命名实体识别模型实验的参数。第 2 组是在基于组合神经网络的片段级中文命名实体识别模型中所需的参数，第 1 组和第 2 组共同组成基于组合神经网络的片段级中文命名实体识别模型的参数。第 3 组是神经网络模型初始输入包含片段扩展特征时，实验中所需的超参数。

4.3 基于神经网络的片段级中文命名实体识别方法有效性验证

为了验证基于神经网络片段级中文命名实体识别方法的有效性，我们以基于神经网络的字符级中文命名实体识别方法实现了一个基线（baseline）系统。Baseline 采用基于 Bi-LSTM 模型的字符序列标注模型结构，对于输入的句子序

列，采用“BIEOS”标注体系通过对每个字符分配标记完成中文命名实体识别。我们利用 MSRA 训练集进行模型训练，在 MSRA 测试集上进行测试。对比实验结果如表 2 所示。实验初始输入的向量预训练语料采用 pre1。从片段内部单元和片段整体两方面表示片段，基于 Bi-LSTM 的神经网络片段级模型记为 $\text{Bi-LSTM}_{\text{pre}}$，基于组合神经网络的片段级模型记为 Comb_{pre}。为了获取更丰富的片段信息提升系统性能，另一组实验选用大规模的预训练语料 pre2，同时从片段内部字符、片段整体以及片段扩展特征三个方面获取片段信息，模型记为 $\text{Bi-LSTM}_{\text{pre2+ext}}$ 和 $\text{Comb}_{\text{pre2+ext}}$，实验结果如表 3 所示。

实验结果显示，与 baseline 系统方法相比，基于神经网络的片段级中文命名实体识别方法识别效果显著提升。采用大规模预训练语料，字符向量、片段向量表示包含更丰富的语义信息[18]，可以更有效的获取片段信息提升系统性能。我们提出的两种基于不同神经网络的片段级方法获得相当的系统性能。

表 2 与 baseline 实验结果对比

模型	P	R	F	F-PER	F-LOC	F-ORG
baseline	66.12	73.91	69.80	66.42	77.83	58.84
$\text{Bi-LSTM}_{\text{pre1}}$	76.50	72.15	**74.26**	61.91	83.23	65.59

表 3 采用大规模预训练语料的实验结果

模型	P	R	F	F-PER	F-LOC	F-ORG
$\text{Bi-LSTM}_{\text{pre2}}$	91.42	88.01	89.68	90.93	91.40	84.09
$\text{Bi-LSTM}_{\text{pre2}}$+ext	92.09	88.85	**90.44**	**91.59**	92.31	84.73
$\text{Comb}_{\text{pre2}}$+ext	92.25	88.37	90.27	90.01	**92.85**	**85.12**

4.4 不同片段级中文命名实体识别方法实验比较

为了验证本文基于神经网络的片段级中文命名实体方法的有效性，我们选择与 Zhou 等人工作的实验结果进行对比。该方法集成命名实体边界识别和分类任务，针对片段级中文命名实体识别进行研究，相比于传统字符序列标注模型，在 MSRA 上获得较好的性能[7]。该方法基于传统统计学习模型，需要依赖人工特征设计和专家知识。

表 4 是在 MSRA 测试集上的测评结果对比，基于 Bi-LSTM 的片段级中文命名实体识别系统和基于组合神经网络的片段级中文命名实体识别系统分别记为 Ours1 和 Ours2。实验结果显示，与 Zhou 等人基于传统统计学习方法的片段级中文命名实体识别方法[7]相比，本文提出的基于神经网络的片段级中文命名实体识别方法中基于 Bi-LSTM 的片段级中文命名实体识别方法获得较好的系统性能，基于组合神经网络的片段级中文命名实体识别方法获得与之相当的实验结果。我们的系统在人名和地名的识别结果上分别提升了 0.9%，0.95%。

表 4 不同方法的实验结果对比

模型	P	R	F	F-PER	F-LOC	F-ORG
Zhou	91.86	88.75	90.28	90.69	91.90	**86.19**
Our1	92.09	**88.85**	**90.44**	**91.59**	92.31	84.73
Our2	**92.25**	88.37	90.27	90.01	**92.85**	85.12

5 结束语

中文命名实体识别是中文自然语言处理领域中的重要基础任务之一。本文针对传统统计学习方法和字符序列化标注模型的不足，主要研究基于神经网络的片段级中文命名实体识别方法，采用两种神经网络模型结构与半马尔可夫条件随机场模型结合，通过对片段整体分配标记完成中文命名实体识别。据我们所知，这是首次针对基于

神经网络的片段级中文命名实体识别进行研究。实验结果显示，该算法的识别效果明显优于baseline，并且获得与当前其他最优的中文命名实体识别系统相当的识别性能。

在下一步的研究工作中，我们将继续研究获取表示片段信息的方法，使得输入的片段信息表示可以更加完整有效，提升系统性能；另外，我们将探索不同的神经网络模型或不同神经网络模型的组合模型调整现有的模型结构，设计更适用于中文命名实体识别任务的模型结构，从而获得更好的识别性能。

参考文献

[1] 宗成庆. 统计自然语言处理[M]. 清华大学出版社, 2008: 150-178.

[2] 廖先桃. 中文命名实体识别方法研究[D], 哈尔滨工业大学, 2006.

[3] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of HLT-NAACL, 2003: 188-191.

[4] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87-94.

[5] 史海峰, 姚建民. 基于 CRF 的中文命名实体识别研究[D]. 苏州: 苏州大学, 2010.

[6] 王志强. 基于条件随机域的中文命名实体识别研究[D]. 南京：南京理工大学，2006.

[7] Zhou J, Qu W, and Zhang F. Chinese Named Entity Recognition via Joint Identification and Categorization[J]. Chinese Journal of Electronics, 2013: 225-230.

[8] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[9] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of ACL, 2010: 384-394.

[10] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]//Proceedings of NAACL-HLT, 2016: 260-270.

[11] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//Proceedings of ACL, 2016: 1064-1074.

[12] Liu Y, Che W, Guo J, Bing Qin, Ting Liu. Exploring Segment Representations for Neural Segmentation Models[C]//Proceedings of IJCAI, 2016: 2880-2886.

[13] 王国昱. 基于深度学习的中文命名实体识别研究[D]. 北京工业大学, 2015.

[14] Sarawagi S, Cohen W W. Semi-Markov Conditional Random Fields for Information Extraction[C]//Proceedings of NIPS, 2004, 17:1185-1192.

[15] Zhang Y, Clark S. Syntactic processing using the generalized perceptron and beam search[J]. Computational linguistics, 2011, 37(1): 105-151.

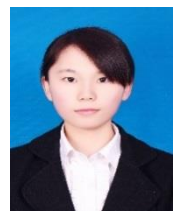
[16] Kong L, Dyer C, and Noah A. Segmental recurrent neural networks[C]//Proceedings of ICLR, 2016.

[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of Workshop at ICLR, 2013.

[18] Lai S, Liu K, He S, et al. How to generate a good word embedding[J]. IEEE Intelligent Systems, 2016, 31(6): 5-14.

王蕾（1992—），硕士研究生，主要研究领域为自然语言处理。

E-mail：wangleinlp@163.com



谢云（1993—），硕士研究生，主要研究领域为自然语言处理。

E-mail：1316480114@qq.com

周俊生（1972—），博士，教授，主要研究领域为自然语言处理、人工智能。

E-mail：zhoujs@njnu.edu.cn

顾彦慧（1978—），博士，副教授，主要研究领域为自然语言处理、信息检索。

E-mail：gu@njnu.edu.cn

曲维光（1964—），博士，教授，主要研究领域为自然语言处理、计算语言学、人工智能。

E-mail：wgqu@njnu.edu.cn