计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200074

基于上下文相关字向量的中文命名实体识别

张 栋 陈文亮
苏州大学计算机科学与技术学院 江苏 苏州 215006
(dzhang19@stu.suda.edu.cn)

摘 要 命名实体识别(NER)旨在识别出文本中的专有名词,并对其进行分类。由于用于监督学习的训练数据通常由人工标注,耗时耗力,因此很难得到大规模的标注数据。为解决中文命名实体识别任务中因缺乏大规模标注语料而造成的数据稀缺问题,以及传统字向量不能解决的一字多义问题,文中使用在大规模无监督数据上预训练的基于上下文相关的字向量,即利用语言模型生成上下文相关字向量以改进中文 NER 模型的性能。同时,为解决命名实体识别中的未登录词问题,文中提出了基于字语言模型的中文 NER 系统。把语言模型学习到的字向量作为 NER 模型的输入,使得同一中文汉字在不同语境中有不同的表示。文中在 6 个中文 NER 数据集上进行了实验。实验结果表明,基于上下文相关的字向量可以很好地提升 NER 模型的性能,其平均性能 F1 值提升了 4.95%。对实验结果进行进一步分析发现,新系统在 OOV 实体识别上也可以取得很好的效果,同时对一些特殊类型的中文实体识别也有不错的表现。

关键词:命名实体识别;语言模型;上下文相关字向量
中图法分类号 TP391.1

Chinese Named Entity Recognition Based on Contextualized Char Embeddings

ZHANG Dong and CHEN Wen-liang
School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China

Abstract Named Entity Recognition (NER) is designed to identify and classify proper nouns in text. Training data for supervised learning are usually manually annotated, and it is difficult to obtain large-scale annotated data due to time-consuming and labor-intensive. In order to solve the problem of data sparseness caused by the lack of large-scale annotation corpus and the problem of polysemy of char embedding in the Chinese NER task, this paper uses contextualized char embeddings which is pre-trained on large-scale unlabeled data to improve the performance of the Chinese NER model. Furthermore, to solve the problem of out-of-vocabulary words in named entity recognition, this paper proposes a Chinese NER system based on word language model. We use the contextualized char embeddings of generated by the language model as the input of the NER model to capture different meanings of Chinese characters in different contexts. In this paper, we conduct experiments on six Chinese NER datasets. The experimental results show that the proposed model can improve the performance and the average F1 improves by 4.95%. In addition, this paper further analyzes the experimental results and finds that the proposed model can achieve better results on OOV entities, and it has good performance for some special types of Chinese entity recognition.

Keywords Named entity recognition, Language model, Contextualized char vector

1 引言

随着互联网的迅速发展和普及,网络上的信息资源在变得越来越丰富的同时,也带来了信息过载的问题。如何从海量的文本资源中获取对用户有用的信息是目前亟待解决的一个问题。命名实体作为理解文本语义过程中的一种关键信息,在各种自然语言处理应用中,如信息检索、自动文本摘要、自动问答、机器翻译和知识库构建等,起着至关重要的作用。

对于命名实体识别任务(Named Entity Recognition, NER),目前常用的方法是把它作为有监督的序列标注问题来求解。一般来说,监督数据越多,模型训练的效果就越好。但由于代价高,难以获取大规模的标注数据。因此,在这种情况下,人工设计的特征和特定领域内的语言资源,例如地名字典,被研究者广泛地运用在 NER 任务中。在新语言和新领域下开发特定的语言资源和特征的代价很大,这使得命名实体识别的研究也更具挑战性。研究者开始研究如何从大量无标注语料学习词的表示方面来改进 NER 的性能。2011 年,Collobert 等[1]使用预训练词向量初始化神经网络的词向量表,与使用随机初始化词向量模型相比,性能有了明显的提升。预训练词向量常用的方法之一是通过 word2vec 获得[2]。word2vec

到稿日期:2019-12-09 返修日期:2020-05-28 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。
基金项目:国家自然科学基金(61876115)
This work was supported by the National Natural Science Foundation of China(61876115).
通信作者:陈文亮(wlchen@suda.edu.cn)

虽然在一定程度上可以学习到词与词之间的一些相关信息，但是它训练得到的词向量在使用时与上下文是无关的。在实际应用中，随着上下文的变化，同一个词会产生不同的表示，如“苹果”可能指的是水果，也可能指的是苹果公司。

为解决上述问题，Peters 等[3]提出基于语言模型的上下文相关的词向量表示 ELMo（Embeddings from Language Model）。ELMo 词向量是整个句子的函数，它们实际上是 embedding 层和两层双向长短时记忆网络（Bidirectional Long-Short Term Memory network，BiLSTM）输出的线性组合。由于采用 BiLSTM 的编码方式，ELMo 词向量对不同的上下文有不同的表征。ELMo 在多个英文任务上取得的成功得到了许多研究者的关注。

虽然目前英文 NER 任务已经用到了 ELMo 并取得了较好的效果，但学术界并没有对模型的结果进行深入研究。同时，由于中文处理的基本单元是字，ELMo 在字级别上的 NER 效果还不清楚。本文针对上述两个问题展开研究，提出基于 ELMo 的中文命名实体识别模型。特别的是，本文训练了一个基于汉字而不是词的语言模型，以适用中文基本单位是字的场景。本文在 6 个不同的中文命名实体识别数据集上进行了实验，并对实验结果进行了详细分析。实验结果表明，基于 ELMo 的命名实体识别方法显著优于本文的基准模型，并能有效提升 OOV 实体识别的性能。

本文的贡献如下：1）训练了一个基于中文字而不是词的 ELMo 语言模型，以适应中文命名实体识别时处理基本单元是字的场景。2）设计了基于 ELMo 的中文 NER 系统，并应用于多个中文命名实体识别数据集。实验结果表明 ELMo 可以很好地改进中文 NER 的性能。同时，在跨领域 NER 实验中也显示出了很好的效果。3）对实验结果进行了进一步的详细分析，发现 ELMo 能有效提高 OOV 实体识别的性能，同时还发现新系统对缩写、外文音译、连续实体等一些中文实体识别典型难题也具有较好的效果。

2 相关工作

命名实体识别是自然语言处理中的一项重要任务。中文命名实体识别任务最早是在 2003 年的“863”评测上被定义[4]。中文没有明显的形态学特征和词语边界，而且具有较强的多义性，实体类别丰富，因此构建中文 NER 系统相比英文等其他语言更加困难。基于特征工程的传统 NER 方法主要有支持向量机[5]、最大熵[6]、条件随机场[7]（Conditional Random Field，CRF）等，但是传统方法过于依赖人工特征，特征的设计需要很多的专家知识，特征选择的好坏更是直接影响到命名实体识别系统的性能。

随着深度学习的发展，基于深度学习的命名实体识别方法逐渐受到更多的关注。深度学习可以用低维、稠密的向量来表示数据，自动进行特征学习，从而摆脱对特征工程的依赖，也解决了高维特征空间带来的数据稀疏的问题。Hammerton[8]最早使用单向 LSTM 神经网络来解决 NER 任务。Collobert 等[1]使用了卷积神经网络来提取词的形态学特征，基于 CNN-CRF 模型取得了不错的效果。因为 LSTM（Long Short Term Memory networks）具有更好的长距离编码能力，近几年大部分的工作都是在 LSTM-CRF 框架上进行拓展。Huang 等[9]在单词特征的基础上加入拼写特征，而 Ma 等[10]使用基于字符的 CNN 来提取单词的字符特征，Lample 等[11]使用 BiLSTM 来获取单词的字符特征。然而，中文字符没有这种有效的细粒度特征，一般来说，中文的词比字具有更多的语义信息，Zhang 等[12]通过将词的信息引入到基于字的序列标注模型中，使命名实体识别的性能得到进一步提升。

Bengio 等[13]于 2001 年正式提出神经网络语言模型（Neural Network Language Model，NNLM），并提出了词向量的概念。2013 年，word2vec 将 NNLM 中的隐层和词序进一步简化，由于其简单有效，预训练的词向量已经在 NLP 系统中普遍应用。尽管可以使用预训练词向量来初始化词表示，但是模型中的参数主要还是从标注数据上学习的。因此，如何从大规模无监督数据中学习到更多丰富的语义特征逐渐成为学术界关注的一个热点。Peters 等[3]使用双向语言模型，在大规模的无标注语料库上进行预训练，在一系列英文任务中取得了较大的成功。然而，在中文命名实体识别任务中，鲜有引入中文语言模型的研究。对于中文 NER 任务，有基于词和基于字的方法。考虑到实体的边界也是词的边界的情况，在基于词的方法中，首先要对文本进行分词，再对词序列进行标注，但这种方法会因分词的错误而导致 NER 的错误。一般来说，中文的词比字具有更丰富的语义信息，但文献[14-16]的研究表明，基于字的中文 NER 方法优于基于词的中文 NER 方法。Che 等[17]使用预训练的 ELMo 在词性分析和句法任务上取得了不错的成绩，但他们预训练的是基于词的语言模型。

基于上述相关工作的不足，本文使用基于字的语言模型 ELMo，来验证 ELMo 在多个中文命名实体识别数据集上的有效性，并对结果进行了详细分析。

3 基于字符的命名实体识别模型

近几年，基于深度学习的命名实体方法取得了突破性的进展。目前最常用的命名实体识别方法是 BiLSTM-CRF 模型。本文将命名实体识别任务转化为序列标注任务，采用的基本框架是 BiLSTM-CRF，并加入预训练的语言模型 ELMo 作为特征。整个模型的框架分为 3 个部分，即输入层、BiLSTM 网络层、CRF 层，如图 1 所示。

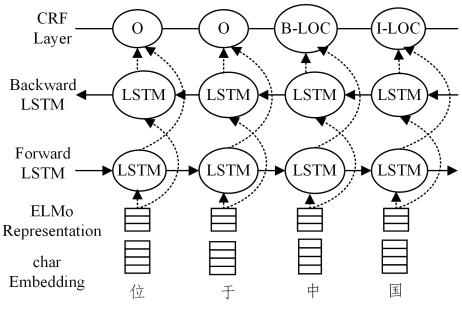

图 1 基于 BiLSTM-CRF 的中文 NER 模型

Fig. 1 Chinese NER model based on BiLSTM-CRF

3.1 输入层

给定一个包含 n 个字符的句子 $sent=(c_1 c_2 \cdots c_n)$，通过查询字向量表将每个字符映射为向量序列 $(\boldsymbol{v}_1 \boldsymbol{v}_2 \cdots \boldsymbol{v}_n)$。字向量表可以通过随机字向量或使用来自大规模无标注语料中预训

练的字向量来进行初始化。本文把 ELMo 字向量（详细内容见 3.3 节）作为额外特征拼接到输入层中。整个句子的 ELMo 字向量通过预训练的语言模型表示为向量序列（$\boldsymbol{e}_1\boldsymbol{e}_2\cdots\boldsymbol{e}_n$），如图 2 所示。最终每个字符的表示是字符向量与 ELMo 表示的拼接，即（$\boldsymbol{x}_t=[\boldsymbol{v}_t,\boldsymbol{e}_t]$，$t\in[1,n]$），整个句子表示为向量序列（$\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_n$）。

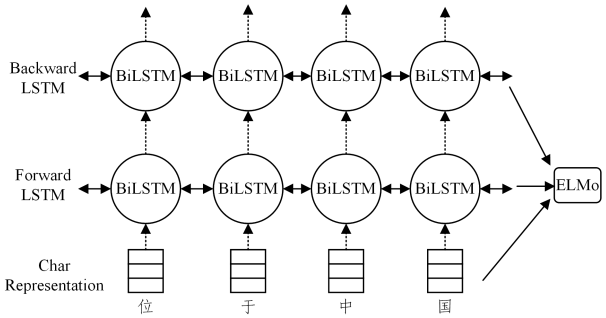

图 2　ELMo 字向量表示

Fig. 2　ELMo char vector embedding

3.2　BiLSTM 网络层

在循环神经网络（Recurrent Neural Networks，RNN）中，当前时刻的输出受到前一个时刻的输出和当前时刻的输入的影响，这一特性很适合对序列化的数据进行建模，解决了全连接网络中前后时刻不能相互依赖的问题。虽然理论上 RNN 可以保留历史信息，使模型基于过去的信息有效预测当前输出，但是由于梯度消失和梯度爆炸，实践中 RNN 并不能有效地学习到长距离依赖信息。为了解决该问题，Hochreiter 等[18]通过引入输入门、输出门等提出了 RNN 的变种——LSTM。LSTM 已被广泛应用于各种任务中，并取得了不错的效果。在许多 NLP 任务中，同时利用上下文信息是非常有用的。然而，LSTM 只能利用过去的信息，为了更好地学习到字符上下文特征，本文使用了 BiLSTM。除了使用正向 LSTM 网络学习每个字符的历史特征外，模型以逆序将序列输入到反向 LSTM 网络中学习每个字符的后续特征。LSTM 的计算方式如式（1）－式（6）所示：

$$i_t=\sigma(W_{ii}x_t+b_{ii}+W_{hi}h_{(t-1)}+b_{hi}) \tag{1}$$

$$f_t=\sigma(W_{if}x_t+b_{if}+W_{hf}h_{(t-1)}+b_{hf}) \tag{2}$$

$$g_t=\tanh(W_{ig}+b_{ig}+W_{hg}h_{(t-1)}+b_{hg}) \tag{3}$$

$$o_t=\sigma(W_{io}x_t+b_{io}+W_{ho}h_{(t-1)}+b_{ho}) \tag{4}$$

$$C_t=f_t\otimes C_{(t-1)}+i_t\otimes g_t \tag{5}$$

$$h_t=o_t\otimes\tanh(C_t) \tag{6}$$

其中，W 和 b 是 LSTM 单元中的参数，σ 是 sigmoid 函数，$\otimes$ 是逐元素乘法，i_t，f_t，o_t 代表 t 时刻的输入门、遗忘门和输出门，C_t，h_t，g_t 分别表示 t 时刻的细胞状态、输出状态以及新状态。给定包含 n 个字符的句子（x_1，x_2，…，x_n），然后通过拼接前向和后向 LSTM 的输出得到 t 时刻字符的特征向量表示 $h_t=[\overrightarrow{h_t},\overleftarrow{h_t}]$，这种表示有效地包含了字的上下文表示，这对序列的标注是非常有用的信息。

3.3　语言模型 ELMo

在英文 ELMo 模型中，一个英文单词表示是由英文单词的词向量和通过 CNN 对英文单词中的字符进行卷积的单词表示两部分组合而成。与英文 ELMo 不同，本文直接对字符的表示进行编码。

给定 n 个字符的序列（c_1，c_2，…，c_N），前向的语言模型通过给定（c_1，…，c_{k-1}）的条件下 c_k 的概率建模来计算整个序列的概率，如式（7）所示：

$$p(c_1,c_2,\cdots,c_N)=\prod_{k=1}^{N}p(c_k\mid c_1,c_2,\cdots,c_{k-1}) \tag{7}$$

一个后向的语言模型与前向类似，只需将输入序列进行翻转即可，相当于通过未来的信息预测过去的信息，如式（8）所示：

$$p(c_1,c_2,\cdots,c_N)=\prod_{k=1}^{N}p(c_k\mid c_{k+1},c_{k+2},\cdots,c_N) \tag{8}$$

一个双向语言模型包含一个前向语言模型和一个后向语言模型。因此，本文语言模型优化的目标是最大化前向和后向语言模型概率的和，如式（9）所示：

$$\sum_{k=1}^{N}(\log p(c_k\mid c_1,\cdots,c_{k-1};\Theta_x,\overrightarrow{\Theta_{LSTM}},\Theta_s)+\log p(c_k\mid c_1,\cdots,c_{k-1};\Theta_x,\overleftarrow{\Theta_{LSTM}},\Theta_s)) \tag{9}$$

其中，Θ_x 是字向量的参数，Θ_s 是 softmax 层的参数。

ELMo 是双向语言模型 BiLSTM 中的中间层表示的任务特定组合。对于每个字 c，一个 L 层双向语言模型可以得到 $2L+1$ 个表示，如式（10）所示：

$$\begin{aligned}R_k&=\{x_k^{LM},\overrightarrow{h}_{k,j}^{LM},\overleftarrow{h}_{k,j}^{LM}\mid j=1,\cdots,L\}\\&=\{h_{k,j}^{LM}\mid j=0,\cdots,L\}\end{aligned} \tag{10}$$

其中，$h_{k,0}^{LM}$ 是 embedding 层，$h_{k,j}^{LM}=[\overrightarrow{h}_{k,j}^{LM};\overleftarrow{h}_{k,j}^{LM}]$ 是 BiLSTM 层。

为了将 ELMo 应用在具体的 NER 任务中，本文把 R 中的所有层折叠成单个向量，$\boldsymbol{ELMo}_k=\boldsymbol{E}(\boldsymbol{R}_k;\boldsymbol{\Theta}_e)$。因此，ELMo 为每个输入的字符提供 $2L+1$ 个表示。相反，传统的字向量方法仅为查找表中的字符提供一层表示，如式（11）所示：

$$\boldsymbol{ELMo}_k^{task}=E(R_k;\Theta^{task})=\gamma^{task}\sum_{j=0}^{L}s_j^{task}h_{k,j}^{LM} \tag{11}$$

其中，s^{task} 是 softmax 正则化权重，γ^{task} 是缩放因子，允许任务模型缩放整个 ELMo 向量。

3.4　CRF 层

一个简单且有效的解码方法是直接用 h_t 作为特征，对输出 o_t 进行独立的标签预测。此方法在一些简单问题（如词性标注）中取得了成功[19]，但是无法处理具有较强依赖关系的序列标注问题。NER 就是这样的任务，需要对标签序列的“语法”强加几个硬约束（如 I-PER 不可能跟着 B-LOC），这些约束不可能使用独立假设进行建模。因此，本文不是对标签进行独立的决策，而是对整个序列进行全局最优的解码。基于此，本文与 Lafferty 等[20]一样使用条件随机场对它们进行联合建模。

本文将提取到的特征 h_t，$t\in[1,n]$ 输入到 CRF 层进行解码，CRF 涉及到两部分。首先，根据当前字符特征 h_t 计算每个标签得分，如式（12）所示：

$$o_t=W_oh_t+b_o \tag{12}$$

其中，o_t 的维度是标签的个数。然后需要定义一个转移矩阵 $\boldsymbol{A}$，$\boldsymbol{A}_{i,j}$ 表示从标签 i 转移到标签 j 的得分。整个序列（$c_1c_2\cdots c_n$）的得分如式（13）所示：

$$s(X,y)=\sum_{t=0}^{n}A_{y_t,y_{t+1}}+\sum_{t=1}^{n}o_{t,y_t} \tag{13}$$

其中，y_0 和 y_{n+1} 是一个句子的开始和结束标签，因此这里 $\boldsymbol{A}$ 是维度为（$k+2$）的一个方阵。

整个序列的概率如式（14）所示：

$$p(y|X)=\frac{e^{s(x,y)}}{\sum_{\tilde{y}\in Y_x} e^{s(X,\tilde{y})}} \tag{14}$$

训练过程中，最大化正确标签序列的对数概率，如式(15)所示：

$$\log((y|X))=s(X,y)-\log(\sum_{\tilde{y}\in Y_x} e^{s(X,\tilde{y})}) \tag{15}$$

其中，Y_x表示所有可能的标签序列。在解码时，通过式(16)预测输出最可能的标签序列。

$$y^*=\arg\max_{\tilde{y}\in Y_x} s(X,\tilde{y}) \tag{16}$$

4 实验

4.1 实验介绍

(1)数据集。本文实验使用了 6 个中文 NER 数据集，包括 Boson，LDC2009，PKU，MSRA[21]，CityU[22] 和 Literature[23]，其中前 5 个来自新闻领域，最后 1 个来自文学作品领域。由于部分数据集在前人工作中并未有明确切分，本文对数据进行如下的预处理：将 Boson，LDC2009 和 PKU 按照 8∶1∶1 随机划分为训练集、验证集、测试集；Literature 根据文献[23]进行划分；由于 CityU 和 MSRA 常见切分方案中没有验证集，本文从训练集中随机划分 20%作为验证集，测试集保持一致。各个数据集的实体统计如表 1 所列。

表 1 数据集实体数量统计

Table 1 Statistics of entity number in dataset

Dataset	Train	Dev	Test
Boson	18522	2307	2348
MSRA	60323	14736	6190
CityU	89748	22613	16407
LDC2009	34067	4294	4253
Literature	133105	10571	16186
PKU	214785	53480	54299

(2)预训练。为了获得预训练的 word2vec 字向量和 ELMo 字向量，本文使用了大约 1181 万句来自新闻领域的中文 Giga 数据作为无标注语料。本文最终获得的词表含有 6127 个字符(包括符号)。此外，本文在 NER 模型训练期间对字向量进行了微调。

(3)评价方法。本实验的评价指标采用 F1 值，如式(17)—式(19)所示：

$$P=\frac{\text{预测出的正确实体}}{\text{预测出的实体}} \tag{17}$$

$$R=\frac{\text{识别出的正确实体}}{\text{样本中的实体数}} \tag{18}$$

$$F1=\frac{2*(P*R)}{P*R} \tag{19}$$

(4)实验设置。实验构建了 4 个系统，分别如下：

1)LSTM-CRF-Rand，LSTM-CRF 使用随机初始化的字向量，即基线系统(baseline)。

2) LSTM-CRF-PreEmb，LSTM-CRF 使用 word2vec 字向量。

3)LSTM-CRF-Rand＋ELMo，在 LSTM-CRF-Rand 的基础上加入 ELMo 字向量。

4)LSTM-CRF-PreEmb＋ELMo，在 LSTM-CRF-PreEmb 的基础上加入 ELMo 字向量。

(5)超参设置。本实验中使用的主要超参数如表 2 所列。此外，对于 LSTM-CRF 模型，为了防止过拟合，本文设置了早停止条件，即在验证集上的 F1 值连续 5 轮不再提高，则模型停止训练。

表 2 超参设置

Table 2 Hyper-parameter settings

参数	LSTM-CRF	语言模型
字向量维数	200	512
LSTM 输出维数	100	4096
Dropout	0.5	0.1
优化器	Adam	Adagrad

4.2 实验结果

表 3 列出了各个系统在 6 个数据集上的实验结果。从表中可以看出，使用预训练的字向量模型优于使用随机初始化的字向量模型，说明从大规模预训练得到的字向量中包含了有用的信息。从整体来说，LSTM-CRF-PreEmb＋ELMo 取得了最好的性能。多个数据集的平均 F1 值比 LSTM-CRF-PreEmb 提高了 4.95%，说明 ELMo 得到的上下文相关字向量优于传统预训练的字向量。ELMo 可以从大规模无监督语料中学习到大量的语义和语法特征，说明上下文相关字向量比传统的预训练字向量能更有效地帮助中文 NER 系统。从 Literature 数据集的结果来看，从新闻领域得到的 ELMo 字向量同样对文学作品领域有很大的帮助，这在一定程度上体现了基于上下文相关字向量的领域迁移能力。

表 3 命名实体识别的 F1 值

Table 3 F1 value of named entity recognition

(单位：%)

数据集	LSTM-CRF-Rand		LSTM-CRF-PreEmb		LSTM-CRF-Rand＋ELMo		LSTM-CRF-PreEmb＋ELMo	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Boson	55.88	58.25	59.80	61.79	70.03	71.53	71.68	72.86
Literature	72.33	69.97	73.49	71.65	76.11	73.49	76.27	74.15
LDC2009	76.09	74.59	77.82	77.03	82.10	80.73	82.56	81.40
PKU	94.69	94.90	95.25	95.40	97.08	97.08	97.24	97.10
MSRA	90.36	86.16	91.69	88.48	95.63	93.89	95.85	93.73
CityU	91.13	86.61	91.65	87.76	94.37	91.84	94.72	92.56
平均值	80.08	78.41	81.62	80.35	85.89	84.76	86.39	85.30

4.3 与其他相关研究的对比

CityU 和 MSRA 是前人工作中使用得较多的数据集，因此本文在这两个数据集上与前人相关工作进行了对比，对比结果如表 4 所列。Chen 等[21]使用两个 CRF 识别人名和地名，使用最大熵(ME)模型识别组织机构名。Chen 等[22]除了提取字符本身的特征外，还提取了列表特征和边界特征等，并且加入了一些对实体识别结果的后处理。Zhang 等[24]基于最大熵将多个外部知识引入到 NER 系统中。Yang 等[25]则

提取了汉字更细粒度的特征（五笔）到模型中。Zhou 等[26]通过联合学习的方法进行命名实体识别。Zhang 等[12]则是在字的命名实体识别方法的基础上，通过网格网络加入词典的信息。从表 4 的结果可以看出，本文的结果均优于已有研究的结果，并且都有较大的提升。

表 4　与之前工作的比较

Table 4　Comparison with previous works

（单位：%）

CityU		MSRA	
模型	F1	模型	F1
Chen 等[21]	88.53	Chen 等[21]	86.20
Chen 等[22]	89.03	Zhang 等[24]	89.03
Zhang 等[24]	89.78	Zhou 等[26]	90.28
Yang 等[25]	**90.68**	Zhang 等[12]	**93.18**
LSTM-CRF-Rand	86.16	LSTM-CRF-Rand	86.61
LSTM-CRF-PreEmb	87.76	LSTM-CRF-PreEmb	88.48
LSTM-CRF-Rand＋ELMo	91.84	LSTM-CRF-Rand＋ELMo	**93.89**
LSTM-CRF-PreEmb＋ELMo	**92.56**	LSTM-CRF-PreEmb＋ELMo	93.73

4.4　结果分析

本节主要从两个角度分析 ELMo 对 NER 系统性能的影响：1）ELMO 对 OOV 实体的影响；2）实例分析。

4.4.1　ELMO 对 OOV 实体的影响

OOV 实体是指未出现在训练集中的实体，IV 实体是指在训练集中出现过的实体。各个数据集的 OOV 实体比例统计结果如表 5 所列。

表 5　测试集 OOV 实体比例统计

Table 5　OOV entity proportion statistics in test set

（单位：%）

数据集	OOV	数据集	OOV
Boson	44.7	PKU	17.7
Literature	28.1	MSRA	29.0
LDC2009	23.8	CityU	25.7

结合表 1 和表 3 可以看出，OOV 实体的比例越高，训练集实体的数目越少，命名实体识别的性能就越差。

为了分析 ELMo 对 OOV 实体的影响，本文对各个数据集的测试集识别结果进行了分类分析，结果如表 6 所列。从结果可以看出，ELMo 对 IV 实体和 OOV 实体的性能都有提升作用，其中 IV 实体性能提升较小，而 OOV 实体的性能提升较大。IV 实体性能代表了一个模型的拟合能力，而 OOV 实体性能表示一个模型的泛化能力。从结果可以看出，基于 ELMo 模型的泛化能力较强。

表 6　IV 实体和 OOV 实体结果

Table 6　IV and OOV entity results

数据集	LSTM-CRF-PreEmb		LSTM-CRF- PreEmb＋ELMo	
	test_iv	test_oov	test_iv	test_oov
Boson	75.10	47.90	82.56(＋7.46)	62.14(＋14.24)
Literature	78.16	54.51	79.98(＋1.82)	58.34(＋3.83)
LDC2009	83.61	57.28	86.33(＋2.72)	65.95(＋8.67)
PKU	97.70	85.57	98.40(＋0.70)	91.47(＋5.90)
MSRA	94.81	73.61	97.18(＋2.37)	85.31(＋11.70)
CityU	93.69	71.77	96.13(＋2.44)	82.95(＋11.18)

4.4.2　实例分析

本文进一步对系统标注结果进行了实例分析，发现基于 ELMo 的模型可以较好地纠正表 7 中几种类型的错误。例如，在缩写类型中，“民进”表示的是组织名“中国民主促进会”，基于 word2vec 字向量的 NER 系统是无法识别的，而基于 ELMo 的 NER 系统可以识别“民进”代表的是组织名。然而，有些情形是基于 ELMo 的系统也无法识别的。例如，句子：也就是在这个地方，素有“[芒市@LOC]谷子[遮放@LOC]米”之说。由于上下文信息不足，系统难以将“芒市”和“遮放”识别为地名。还有一些实体识别需要背景知识，例如，句子：当时正值困难时期，官兵们发扬艰苦奋斗的精神，战胜种种难以想象的困难，创造了闻名于世的“[北大荒@LOC]精神”，把沉睡的雪原变成了金色的粮仓，令人钦佩。这里的“北大荒”是指中国黑龙江省北部的三江平原、黑龙江沿河平原及嫩江流域的广大荒芜地区。

表 7　实例分析

Table 7　Case study

类型	句子	基于 word2vec	基于 ELMo
缩写	谨向[民进@ORG]的全体同志致以崇高的敬意	未识别	[民进@ORG]
	[旅协@ORG]亦推出了“环球盛事汇香江———迈向 21 世纪”的大型推广计划	未识别	[旅协@ORG]
外文音译	这里还展出了许多[毕加索@PER]制作的陶器及雕刻作品	未识别	[毕加索@PER]
	[略伦特@PER]拥有超强的空中优势，同时他的控球能力也非常出色	未识别	[略伦特@PER]
连续实体	记希腊船王[康斯坦塔科普洛斯@PER][符小渤@PER]	[康斯坦塔科普洛斯@PER]	[康斯坦塔科普洛斯@PER][符小渤@PER]
	身边还有[张龙@PER][赵虎@PER][王朝@PER][马汉@PER]四个卫士	[张龙赵@PER]王朝马汉	[张龙@PER][赵虎@PER]王朝[马汉@PER]
线索错误	事实上，[马拉维@LOC]已经在吸引着越来越多的人到这个一度极为封闭的国家一游	[马拉维@PER]	[马拉维@LOC]
	而戴安娜，不论她起初是否自觉，至少在当上王妃以后是欣然地接受了“明星”的角色的	[王妃@PER]	无实体
边界	民间资金托起乡韩[黄河@LOC]大桥今年十月底可望通车	[韩黄河@LOC]	[黄河@LOC]
	因为[中国@LOC]人民会永远铭记他的努力曾经为他们带来进步与发展	[中国人民会@ORG]	[中国@LOC]

结束语　本文主要研究语言模型 ELMo 在中文命名实体识别中的有效性。在没有引入外部资源和额外特征等情况下，本文提出基于字的语言模型 ELMo 的 NER 系统来提高中文命名实体识别的性能。实验结果表明，ELMo 对 6 个不同来源的数据集的性能都有较大幅度的提升。本文进一步对实验结果进行了分析，发现新系统对中文 OOV 实体有很大的帮助，并对缩写、外来音译、连续实体等一些中文实体识别的典型难题产生了较好的效果。未来可以在深度学习中合理

引入外部知识，使中文命名实体识别的性能得到进一步提升。此外，最近研究人员提出的新的语言模型如 GPT，BERT，均在多个自然语言任务中体现出了非常好的效果，但是它们需要消耗更多的计算资源，因此计划未来增加计算资源，在中文命名实体识别中尝试这些新的语言模型进行比较分析。

参 考 文 献

[1] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(8): 2493-2537.

[2] MIKOLOU T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proc of NIPS. Cambridge: MIT Press, 2013: 3111-3119.

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

[4] SUN Z, WANG H L. Summary of research progress in named entity recognition [J]. New Technology of Library and Information Service, 2010, 26 (6): 42-47.

[5] LI L, MAO T, HUANG D, et al. Hybrid models for Chinese named entity recognition [C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 72-78.

[6] ZHANG S, QIN Y, WEN J, et al. Word segmentation and named entity recognition for SIGHAN Bakeoff3 [C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 158-161.

[7] ZHOU J, HE L, DAI X, et al. Chinese named entity recognition with a multi-phase model [C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 213-216.

[8] HAMMERTON J. Named entity recognition with long short-term memory [C] // Proceedings of the seventh conference on Natural language learning at HLT-NAACL. Association for Computational Linguistics. Sapporo: ACL, 2003: 172-175.

[9] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging [J]. arXiv: 1508. 01991.

[10] MA X, 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. Berlin: ACL, 2016: 7-12.

[11] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. Berlin: ACL, 2016: 260-270.

[12] ZHANG Y, YANG J. Chinese NER Using Lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018: 1554-1564.

[13] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(2): 1137-1155.

[14] HE J, WANG H. Chinese named entity recognition and word segmentation based on character [C]// Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing. Columbus: ACL, 2008: 128-132.

[15] LIU Z, ZHU C, ZHAO T. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? [C]// International Conference on Intelligent Computing. Berlin: Springer, 2010: 634-640.

[16] LI H, HAGIWARA M, LI Q, et al. Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese [C] // Proceedings of the Ninth International Conference on Language Resources and Evaluation. Reykjavik: LREC, 2014: 2532-2536.

[17] CHE W, LIU Y, WANG Y, et al. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation [C] // Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Melbourne: ACL, 2018: 55-64.

[18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[19] LING W, LUÍS T, MARUJO L, et al. Finding function in form: Compositional character models for open vocabulary word representation [C]// Proceeding of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015: 1520-1530.

[20] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]// Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[21] CHEN A, PENG F, SHAN R, et al. Chinese named entity recognition with conditional probabilistic models [C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 173-176.

[22] CHEN W, ZHANG Y, ISAHARA H. Chinese named entity recognition with conditional random fields [C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 118-121.

[23] XU J, WEN J, SUN X, et al. A discourse-level named entity recognition and relation extraction dataset for Chinese literature text [J]. CoRR, 2017, 11(7): 100-104.

[24] ZHANG S, QIN Y, WEN J, et al. Word segmentation and named entity recognition for SIGHAN Bakeoff3 [C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: ACL, 2006: 158-161.

[25] 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. Hohhot: Springer, 2018: 184-195.

[26] ZHOU J, QU W, ZHANG F. Chinese named entity recognition via joint identification and categorization [J]. Chinese journal of electronics, 2013, 22(2): 225-230.

ZHANG Dong, born in 1992, postgraduate, is a member of China Computer Federation. His main research interests include natural language processing and named entity recognition.

CHEN Wen-liang, born in 1977, professor, doctoral supervisor, is a member of China Computer Federation. His main research interests include natural language understanding, information extraction and knowledge graph.