基于深度神经网络的维吾尔文命名实体识别研究*

王路路[1,2]，艾山·吾买尔[1,2]，吐尔根·依布拉音[1,2]，买合木提·买买提[1,2]，卡哈尔江·阿比的热西提[1,2]

（1. 新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046；2. 新疆大学 新疆多语种信息技术实验室，新疆 乌鲁木齐 830046）

摘要：现有的维吾尔文命名实体识别主要采用基于条件随机场的统计学习方法，但依赖于人工提取的特征工程和领域知识。针对该问题，该文提出了一种基于深度神经网络的学习方法，并引入不同的特征向量表示。首先利用大规模未标注语料训练的词向量模型获取每个单词具有语义信息的词向量；其次，利用Bi-LSTM提取单词的字符级向量；然后，利用直接串联法或注意力机制处理词向量和字符级向量，进一步获取联合向量表示；然后 Bi-LSTM-CRF 深度神经网络模型进行命名实体标注。实验结果表明，以基于注意力机制的联合向量表示作为输入的Bi-LSTM-CRF方法在维吾尔文命名实体识别上F值达到90.13%。

关键词：维吾尔文命名实体识别；长短时记忆网络；条件随机场；注意力机制

中图分类号：TP391　　**文献标识码：**A

Research on Uyghur Named Entity Recognition Based on Deep Neural Network

Wang Lu-lu[1,2], Aishan Wumaier[1,2], Tuergen Yibulayin[1,2], Maihemuti Maimaiti[1,2], Kahaerjiang Abiderexiti[1,2]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China; 2. Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi, Xinjiang 830046, China)

Abstract: The current research on Uyghur named entity recognition mainly adopt statistical learning methods that are conditional random fields, but it depends on the feature engineering and domain knowledge extracted manually. To solve this problem, this paper proposes a method based on deep neural network and introduces different feature vector representation for Uyghur named entity recognition. The word embedding with semantic information about each word was obtained through the large-scale unlabeled corpus and the character-level embedding was extracted by Bi-LSTM. To further get united vector representation, we adopt direct concatenation or attention-based method to combine the character-level embedding with word embedding. Finally, the deep neural network of Bi-LSTM-CRF was used to label named entities. The experimental results show that the Bi-LSTM-CRF with the united vector representation based on attention mechanism achieves an F-value of 90.13% for Uyghur named entity recognition.

Key words: Uyghur named entity recognition; long short-term memory network; conditional random fields; attention mechanism

1 引言

随着信息化进程的加快，互联网上维吾尔文的信息资源呈逐渐增长趋势，从而维吾尔文信息化研究显得愈来愈重要，由此维吾尔语自然语言处理应运而生。命名实体识别作为自然

* **收稿日期：**　　**定稿日期：**

基金项目：国家973计划资助项目(2014CB340506)；国家自然科学基金资助项目(61462083；61262060；61662077；61331011)；新疆多语种信息技术实验室开放课题(2016D03023)

作者简介：王路路（1993—），女，博士研究生，主要研究领域为自然语言处理；艾山·吾买尔（1981—），男，硕士生导师，副教授，主要研究领域为少数民族自然语言处理与机器翻译；吐尔根·依布拉音（1958—），男，博士生导师，教授，主要研究领域为自然语言处理、社会计算。

语言处理中的一项基础性任务，旨在将非结构化文本中抽取出具有特定意义的实体，如人名、地名、机构名，并且其在信息抽取、机器翻译、问答系统等领域中有着重要作用。

随着深度学习的不断深入，基于神经网络的命名实体识别已在汉语[1][2]、英语[3][4]大规模语种上呈现了很好的性能。然而，维吾尔文命名实体识别尚处于起步阶段，面临的主要问题如下：（1）维吾尔语是形态丰富的典型性黏着语言。通过附加不同的词缀，一个词将有多种形态，使得数据稀疏，从而带来未登录词问题（OOV）；（2）维吾尔语命名实体中没有大小写特征；（3）没有公开的数据集，数据规模的有限性将会影响神经网络方法的识别性能。此外，现有维吾尔文命名实体识别研究主要采用基于统计的方法[5]或者统计与规则相结合的方法[6][7]，而这些方法严重依赖于人工提取的特征工程和领域知识。

为了避免繁琐的特征工程，本文提出了基于深度神经网络的维吾尔文命名实体识别的方法。本文的主要工作内容如下：（1）实现了对维吾尔文中的人名、地名、机构名同时识别；（2）将神经网络方法应用在维吾尔文命名实体识别上；（3）分别使用直接串联法和基于注意力机制的加权求和法将词向量和字符级向量进行联合，来动态学习形态丰富的维吾尔文字符间的特征并对比 Bi-LSTM 和 Bi-LSTM-CRF 两种模型的识别效果；（4）以联合向量表示作为输入的 Bi-LSTM-CRF 方法取得较佳的性能的同时，有效缓解了未登录词的识别。

2 相关工作

基于神经网络的方法已经成功地运用在命名实体识别序列标注任务上。Collobert 等[8]于 2011 年提出了基于 CNN-CRF 神经网络模型进行了命名实体识别研究，随后出现了一系列借鉴此方法的深度神经网络方法用于序列标注任务中。Huang 等[9]提出了一种以人工提取的特征向量和词向量的拼接向量作为输入的 Bi-LSTM-CRF 模型，在 CONLL2003 数据集上 F 值达到了 90.10%；Lample 等[3]引入了由 Bi-LSTM 获取的字符级向量，F 值达到了 90.94%；Rei 等[10]提出了利用注意力机制获取字符级向量和词向量的联合向量；Ma 等[4]构建了 BLSTM-CNNs-CRF 神经网络模型，通过 CNN 学习字符级向量，优于其他模型。 张海楠等[1]提出了一种基于深度神经网络的字词联合方法以实现中文命名实体识别，有效地解决了字词稀疏的不足；Dong 等[11]利用基于 BLSTM-CRF 神经网络模型的同时，有效结合了字向量和偏旁向量。

相比于汉语或者英语等大规模语种，维吾尔文命名实体识别研究起步较晚，近几年很多学者针对命名实体中某一类别展开研究。艾斯卡尔·肉孜等[5]利用条件随机场，引入了词性、词干、音节等特征进行人名的识别；加日拉·买买提热衣木等[12]提出了统计与规则相结合来识别维吾尔人名，主要借用边界词提取人名；塔什甫拉提·尼扎木丁等[7]从维吾尔语黏着特点出发利用条件随机场识别维吾尔文人名，然后再用基于规则的方法对汉族人名识别进行优化；买合木提·买买提等[6]采用条件随机场和规则相结合的方法研究了维吾尔文地名识别，并取得了较高的性能；麦合甫热提等[13]提出了利用语法语义知识实现了基于规则的维吾尔文机构名识别；阿依古丽·哈力克等[14]提出了基于正则表达式对维吾尔语中的时间、数字、量词进行识别。从中可以发现以上维吾尔文命名实体识别的研究主要采用基于规则的方法或者基于统计的方法，而这些方法较为传统，常常在分析语言特性的同时，需要人工编制规则或者构建复杂的特征工程，从而说明维吾尔文命名实体识别性能具有一定的改进空间。

3 特征向量表示

近年来，分布式向量表示已广泛应用于自然语言处理领域，尤其是深度学习研究。本文采用词向量作为基本的特征，引入字符级向量来验证词向量和字符级向量的联合向量表示对维吾尔文命名实体识别的影响，本文将考虑以下特征向量：

3.1 词向量

分布式向量表示能够从大规模的未标注语料中获取单词的语义信息，与 one-hot 向量表示相比，它可以有效地降低维度，获取单词间的语义相关性。word2vec[15]和 Glove[16]是目前

常用于训练分布式词向量的自然语言处理开源工具，其中 word2vec 包括 CBOW 和 Skip-gram 两种模型。为了获取高质量的词向量，本文利用新疆多语种信息技术实验室自然语言处理组搜集的 385 万句的维吾尔语语料采用 word2vec 中 Skip-gram 模型获取预训练的 300 维向量，词向量表中共包含 1249649 个单词/字符及其实数值向量。本文通过词向量查找表获取输入文本中每个 token 的预训练词向量，如果某个 token 不在表中，将被映射到一个统一的向量表中。

3.2 联合向量表示

维吾尔语属于形态丰富的黏着语，通过在词根的前后附加不同的词缀来实现语法功能，因此词汇量庞大，容易造成未登录词问题。单纯的词向量对未登录词问题处理不足。但是字符级向量包含丰富的结构特征，对于形态丰富的语言来说字符级向量是非常有用的，它能够学习前缀和后缀信息等形态信息，从而能够缓解数据稀疏问题。此外，字符级向量能够有效地处理语言模型或者词性标注中的未登录词问题[17]。

首先，随机初始化包含不同字符向量的字符向量查找表；然后，将单词 *word* 中每个字符的向量通过 Bi-LSTM 模型获取单词的前向传播向量 l_{word} 和反向传播向量 r_{word}；最后将前向传播向量 l_{word} 和反向传播向量 r_{word} 进行拼接获取 $c_{word}=[l_{word};r_{word}]$。

假设单词 *word* 在词向量查找表中的向量为 w_{word}，字符级向量为 c_{word}，本文将以下两种联合向量方法。

（1）基于直接串联的联合向量表示。将 w_{word} 和 c_{word} 直接拼接构成的串联向量作为序列标注模型的输入向量 e_{word}，即 $e_{word}=[w_{word};c_{word}]$。如图 1 所示。

（2）基于注意力机制的联合向量表示。本文借鉴 Rei 等[10]的方法使用注意力机制将词向量和字符级向量加权求和进行联合，如图 2 所示。其中注意力机制的权重 a 是通过两层前馈神经网络学习的。

$$e_{word}=a\cdot w_{word}+(1-a)\cdot c_{word} \tag{1}$$

$$a=\sigma(W_a^3\tanh(W_a^1 w_{word}+W_a^2 c_{word})) \tag{2}$$

公式(2)中 W_a^1、W_a^2、W_a^3 分别是计算权重 a 的权重矩阵。

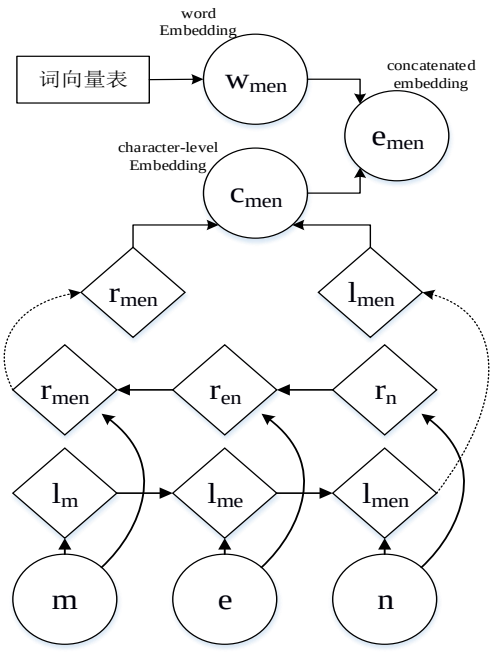

图 1 基于直接串联的联合向量表示

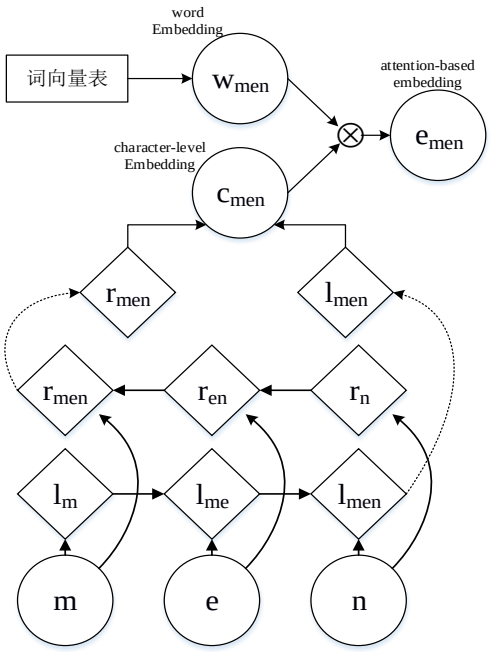

图 2 基于注意力机制的联合向量表示

4 基于 Bi-LSTM-CRF 的维吾尔文命名实体识别

将联合向量表示作为 Bi-LSTM 模型的输入，获取前向传播向量和反向传播向量；然后将两个向量的拼接向量以表示输入序列的向量，再通过 tanh 层将向量缩小至[-1,1]；最后通过条件随机场判断出最优的标记序列。为了充分理解维吾尔文命名实体识别研究，本文以拉丁维吾尔语“men junggoni söyimen” （中文意思：我爱中国）进行举例。

4.1 Bi-LSTM

循环神经网络（Recurrent neural network，RNN）是处理序列标注问题的一种神经网络语言模型，它能够利用历史信息处理长距离依赖信息，但是未能有效地解决梯度消失和梯度爆炸问题。长短时记忆网络（LSTM）[18]是 RNN 的变种，明显在该问题上表现占优，主要通过记忆单元连接各个门结构使得模型记忆有效的上下文信息。LSTM 门结构有输入门、遗忘门、输出门。LSTM 的形式化表示如下：

$$\begin{cases} f_t = \sigma(W_f x_t + U_f h_{t-1} + \mathrm{V}_f c_{t-1} + b_f) \\ \mathrm{i}_t = \sigma(W_i x_t + U_i h_{t-1} + \mathrm{V}_i c_{t-1} + b_i) \\ c_t = f_t \ \mathrm{e}\ c_{t-1} + i_t \ \mathrm{e}\ (\tanh(W_c x_t + U_c h_{t-1} + b_c)) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + \mathrm{V}_o c_{t-1} + b_o) \\ h_t = o_t \ \mathrm{e}\ \tanh(c_t) \end{cases} \tag{3}$$

其中 σ 是非线性函数 sigmoid 函数，e 代表点乘运算，tanh 表示双曲正切函数，x_t 、h_{t-1}、c_{t-1} 分别表示 t 时刻的输入、上一时刻的输出、上一时刻的单元状态。W 、U 、V 分别表示对应门或者状态的权重，b 表示偏值项。

为了充分地利用上下文信息，本文将采用 Bi-LSTM 模型。Bi-LSTM 在 LSTM 的基础上增加了反向传播层，可以将信息序列分别以两个方向出发输入到模型，然后经过隐含层保存两个方向的信息序列，即历史信息与未来信息。对于输入序列 $S=(e_1,e_2,\mathrm{L}\ ,e_n)$，Bi-LSTM 将获取前向传播向量 $l=(l_1,l_2,\mathrm{L}\ ,l_n)$ 和反向传播向量 $r=(r_1,r_2,\mathrm{L}\ ,r_n)$，则 Bi-LSTM 的最终输出为 $t_i=(l_i;r_i)$，在 Bi-LSTM 之上的 tanh 层是用于预测每个单词所有可能标记序列的置信度。

$$h_i=\tanh(\mathrm{W}_h \mathrm{t}_i) \tag{4}$$

其中 W_h 表示隐藏层的权重矩阵。

Softmax 作为 Bi-LSTM 的输出层，可以对各个位置独立进行多分类。Softmax 函数是计算每个单词的所有可能标记信息的归一化概率分布。

$$p(y_i=j|h_t)=\frac{e^{w_{o,j}h_i}}{\sum_{l\in K} e^{w_{o,l}h_i}} \tag{5}$$

$p(y_i=j|h_t)$ 表示输入序列中第 i 个单词对应的标记 y_i 是 j 的概率，K 表示标签集合。 在训练过程中通过最小化负对数似然函数优化模型。

$$\mathrm{E}=-\sum_{i=1}^{n}\log(p(y_t \mid h_t)) \tag{6}$$

4.2 Bi-LSTM-CRF

由于本文将维吾尔文命名实体识别看作序列标注任务并采用了 BIO 标注形式，而这种标记形式有很强的约束性，例如“I-ORG”之前不可能是“B-LOC”或者 O。若仅仅用 Bi-LSTM 不能充分解决此类问题，但是 CRF 能够考虑上下文标签之间的关系，从而能代替 softmax 层获取全局最优的标记序列，因此最终本文考虑将 Bi-LSTM 和 CRF 结合，即使用 Bi-LSTM-CRF 模型用于维吾尔文命名实体识别中，如图 3 所示。

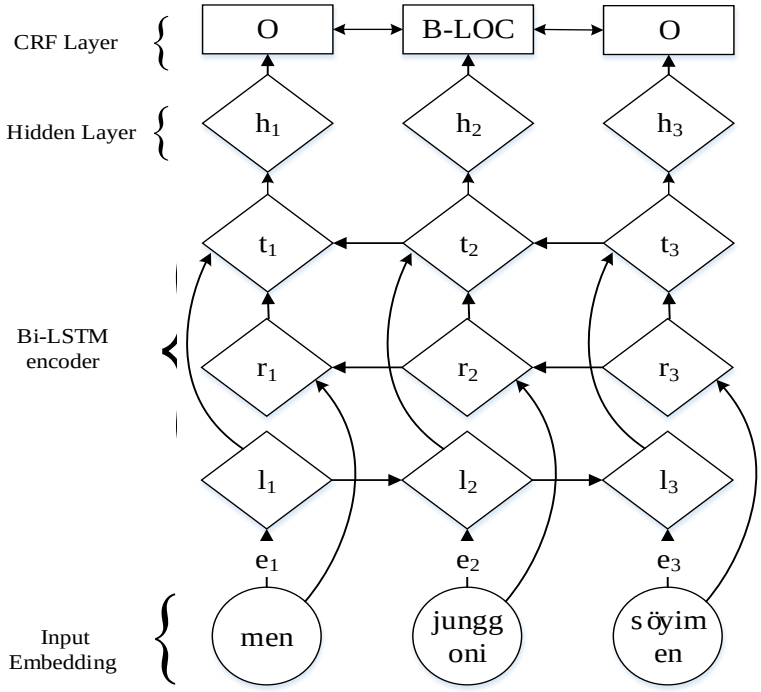

图 3 基于 Bi-LSTM-CRF 的维吾尔文命名实体识别

首先将第 3 节中的特征向量表示作为 Bi-LSTM 的输入向量，通过 Bi-LSTM 编码器获取输出结果 P （原理同 3.1），其中 P 的大小为 $n*k$ ， n 表示输入序列的长度， k 表示标签集合的大小，则其第 i 列是由公式（4）获取的向量 h_i ，则 $P_{i,j}$ 表示输入序列中第 i 个单词对应第 j 个标记的分数。通过引入转移矩阵 T 作为 CRF 模型的参数，$T_{i,j}$ 表示连续单词由标签 i 到标签 j 的转移概率。对于输入序列预测的标签序列 $y=\{y_1,y_2,\mathrm{L}\ ,y_n\}$ ，定义概率表示如下：

$$S(X,y)=\sum_{i=0}^{n}T_{\mathrm{y}_i,\mathrm{y}_{i+1}}+\sum_{i=1}^{n}P_{i,\mathrm{y}_i} \tag{7}$$

得到概率后利用最大似然函数训练模型。

$$\begin{aligned}\log(p(y\mid X))&=\log(\frac{\exp(S(X,\mathbf{y}))}{\sum_{\mathbf{y}\in\mathbf{Y}_\mathbf{x}}\exp(S(X,\mathbf{y}))})\\&=S(X,\mathbf{y})-\log(\sum_{\mathbf{y}\in\mathbf{K}}S(X,\mathbf{y}))\end{aligned} \tag{8}$$

在预测过程中寻找条件概率最大的输出序列 y^* 。

$$y^*=\arg\max_{y}S(y\mid X) \tag{9}$$

5 实验结果与分析

本文进行了多组对比实验来验证深度神经网络对维吾尔文命名实体识别的有效性并探索不同的输入向量对识别效果的影响。

5.1 实验数据

本文采用新疆多语种信息技术实验室标注的命名实体数据集，共计 39027 条句子，包含命名实体 102360 个，人名、地名、机构名占比分别约为 27.81%、41.60%、30.58%。数据集按照交叉验证法将语料 7.5:1：1.5 的比例分为训练集、验证集、测试集。具体的分布信息如

表 1、表 2、表 3 所示，其中 NE 表示命名实体，OOV 表示未登录词（未在训练集中出现的词），ROOV 表示未登录的占比。

表 1 维吾尔文命名实体识别数据集的统计信息

Data	Sentence	Token	NE	PER	LOC	ORG
train	29270	861967	76787	21304	32011	23472
dev	3902	115689	10215	2842	4258	3115
test	5855	174989	15358	4323	6316	4719

表 2 开发集 OOA 统计信息

Dev-data	Token	NE	PER	LOC	ORG
OOV	5991	4263	1314	1222	1728
Roov	5.18%	41.73%	46.23%	28.70%	55.47%

表 3 测试集 OOV 统计信息

Test-data	Token	NE	PER	LOC	ORG
OOV	9271	6440	1987	1905	2553
Roov	5.29%	41.93%	45.96%	30.16%	54.10%

5.2 评测指标

实验采用 $F-$值（$F1$）来评测命名实体识别效果，其中 $F-$值由准确率（P）、召回率（R）来决定。计算公式如下。

$$P=\frac{\text{正确识别的实体个数}}{\text{识别的实体个数}}\times 100\% \tag{10}$$

$$R=\frac{\text{正确识别的实体个数}}{\text{测试语料中所有的实体个数}}\times 100\% \tag{11}$$

$$F-\text{值}=\frac{2\times P\times R}{(P+R)}\times 100\% \tag{12}$$

5.3 参数设置

本文参考前人的工作[10]，采用基于 batch 的梯度下降优化超参数，其中 batch 的大小为 64，使用 Adadelta 优化算法，并设置其初始学习率为 1.0；为了防止过拟合问题，设置 Dropout 参数为 0.5； LSTM 的前向传播和反向传播的字符向量维度分别为 50；LSTM 中隐藏层的大小为；在 bi-LSTM 顶部上 tanh 层的层数设置为 50。根据维吾尔文命名实体识别的词向量验证，最终确定训练词向量采用 skipgram 模型且其维度为 300 维；具体参数设置如表 4 所示。

表 4 参数设置

参数名称	值
dropout rate	0.5
初始学习率	1.0
词向量维度	300
字符级向量维度	100
隐藏层规模	200
tanh 层数	50
迭代次数	100

5.4 实验设置与结果分析

为了验证基于神经网络的维吾尔文命名实体识别方法的有效性，本文以基于 CRF 和半监督学习的维吾尔文命名实体识别方法为基线系统（新疆多语种信息技术实验室自然语言处理组提供的服务），分别以词向量、基于直接串联的联合向量表示、基于注意力机制的联合

向量作为输入向量，在 Bi-LSTM 和 Bi-Bi-LSTM-CRF 两种模型上进行实验，实验结果如表 5 所示。

表 5 不同模型的对比实验结果

模型	向量	Dev（F 值/%）				Test（F 值/%）			
		PER	LOC	ORG	NE	PER	LOC	ORG	NE
Baseline	—	—	—	—	—	91.65	85.72	85.91	87.43
Bi-LSTM	word embedding	90.70	84.42	82.69	85.60	91.03	82.82	81.51	84.67
	concatenated embedding	94.42	88.10	83.81	88.52	94.43	87.05	83.48	88.00
	attention-based embedding	94.58	88.29	83.79	88.62	93.87	86.68	83.27	87.61
Bi-LSTM-CRF	word embedding	91.05	86.37	86.80	87.82	91.15	85.22	86.68	87.35
	concatenated embedding	94.87	89.67	88.31	90.70	**95.29**	87.78	87.44	89.79
	attention-based embedding	**95.14**	**89.75**	**88.65**	**90.91**	94.85	**88.28**	**88.28**	**90.13**

从表 5 中看出，与基线系统相比，Bi-LSTM 和 Bi-LSTM-CRF 两种模型仅在词向量为输入向量的情况下在命名实体识别上表现稍弱，但是在联合向量表示为输入向量的情况下都有提高，说明引入字符级向量的联合向量表示方法进一步提高了维吾尔文命名实体识别的性能，同时能够有效地减少人工提取领域特征的工作量；从总体上看，Bi-LSTM-CRF 模型优于 Bi-LSTM，说明条件随机场能够有效学习相邻标记之间的关系，从而联合解码以得到最优序列标注；两种联合向量表示相比于词向量而言，在 Bi-LSTM 和 Bi-LSTM-CRF 模型识别效果明显提高，说明引入字符级向量，能够有效地学习形态特征，从而缓解形态丰富语言面临的问题；在 Bi-LSTM 模型上，两种联合向量表示相比时，基于注意力机制的联合向量表示在开发集上稍有提高，测试集略低；在 Bi-LSTM-CRF 上，对输入向量进行对比，发现基于注意力机制的联合向量表示在整体的命名实体识别上 F 值达到了 90.13%，且高于基于直接串联的联合向量表示，说明基于注意力机制的联合向量表示能够使 Bi-LSTM-CRF 模型动态地决定利用词向量和字符级向量中哪些信息，且适用于形态丰富的维吾尔语。

为了更好地验证深度神经网络模型的影响，本文将开发集和测试集中所有的 OOV 抽取出来，进一步对 OOV 识别进行了分析，如表 6 所示。

表 6 OOV 识别的对比实验

模型	向量	OOV—Dev（F 值/%）				OOV—Test（F 值/%）			
		PER	LOC	ORG	NE	PER	LOC	ORG	NE
Baseline	—	—	—	—	—	86.78	67.61	81.66	79.02
Bi-LSTM	word embedding	94.76	92.03	87.05	91.24	94.49	90.52	86.08	90.23
	concatenated embedding	**97.24**	93.70	86.85	92.58	96.88	93.18	86.86	92.27
	attention-based embedding	97.20	94.06	87.66	92.95	96.34	92.69	87.17	92.00
Bi-LSTM-CRF	word embedding	94.56	91.27	89.44	91.63	94.00	90.05	89.10	90.87
	concatenated embedding	97.16	**94.48**	90.90	**94.14**	**97.09**	**93.06**	90.04	93.28
	attention-based embedding	97.10	94.01	**91.11**	93.98	96.58	92.94	**90.74**	**93.29**

从表 6 中可知，神经网络模型在 OOV 识别上优于基线系统；无论是哪种神经网络模型，引入字符级向量，OOV 识别性能几乎都提高 2%左右，说明联合向量表示可以有效缓解未登录词的识别；基于直接串联的联合向量表示与基于注意力机制的联合向量表示相比，在 OOV 识别上相差不大，由此可以说明基于直接串联的联合向量表示在非 OOV 上识别效果较好。

6 结束语

现有的维吾尔文命名实体识别研究依赖于人工的特征工程和领域知识，针对该问题，本文提出了基于深度神经网络的方法，主要采用基于不同输入向量的 Bi-LSTM-CRF 的神经网络模型。首先通过大规模的无监督学习语料训练词向量以建立词向量查找表，从而获取每个

单词具有语义的词向量；然后由 Bi-LSTM 获取的字符级向量进行联合，分别获取基于直接串联的联合向量表示和基于注意力机制的联合向量表示；最后通过 Bi-LSTM-CRF 神经网络模型对进行实体标注。实验表明，基于注意力机制向量表示的 Bi-LSTM-CRF 方法的识别效果最佳，由此说明基于注意力机制的联合向量表示能够使模型动态地利用字符级向量或者词向量中的有效信息。

在未来的研究工作中，我们将继续研究基于深度神经网络的维吾尔文命名实体识别，探索其他神经网络模型组合或者在模型中引入注意力机制，验证出最适合于维吾尔文命名实体识别的模型；此外，将利用迁移学习实现其他黏着语种的命名实体识别。

参考文献

[1] 张海楠,伍大勇,刘 悦,程学旗. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28-35.

[2] 王蕾,谢云,周俊生,顾彦慧,曲维光. 基于神经网络的片段级中文命名实体识别[J]. 中文信息学报, 2018, 32(3): 84-90,100.

[3] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. 2016:260-270.

[4] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.

[5] 艾斯卡尔·肉孜, 宗成庆, 姑丽加玛丽·麦麦提艾力,等. 基于条件随机场的维吾尔人名识别方法[J]. 清华大学学报(自然科学版), 2013(6):873-877.

[6] 买合木提·买买提,卡哈尔江·阿比的热西提,艾山·吾买尔,吐尔根·依布拉音,王路路. CRF 与规则相结合的维吾尔文地名识别研究[J]. 中文信息学报, 2017, 31(6): 110-118.

[7] 塔什甫拉提·尼扎木丁, 汪昆, 艾斯卡尔·艾木都拉, 帕力旦·吐尔逊. 统计与规则相结合的维吾尔语人名识别方法. 自动化学报, 2017, 43(4): 653-664.

[8] Collobert R, Weston J, Karlen M, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[9] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[10] Rei M, Crichton G K O, Pyysalo S. Attending to Characters in Neural Sequence Labeling Models.[C].// the 26th International Conference on Computational Linguistics. 2016: 309–318.

[11] Dong C, Zhang J, Zong C, et al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[C]// International Conference on Computer Processing of Oriental Languages. Springer International Publishing, 2016:239-250.

[12] 加日拉·买买提热衣木, 吐尔根·依布拉音, 艾山·吾买尔. 基于统计和规则混合策略的维吾尔人名识别研究[J]. 新疆大学学报(自然科学版), 2014(3):319-324.

[13] 麦合甫热提, 米日姑·肉孜, 等. 基于语法语义知识的维吾尔文机构名识别[J]. 计算机工程与设计, 2014(8):2944-2948.

[14] 阿依古丽·哈力克, 艾山·吾买尔, 吐尔根·伊布拉音,等. 汉维时间数字和量词的识别与翻译研究[J]. 中文信息学报, 2016(6):190-200.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[16] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[17] Ling W, Luís T, Marujo L, et al. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation[J]. Computer Science, 2015:1899-1907.

[18] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

作者联系方式：王路路　新疆维吾尔自治区乌鲁木齐市天山区胜利路 666 号新疆大学 830046　13899948293 wanglulu@stu.xju.edu.cn