·人工智能及识别技术· 文章编号:1000-3428(2018)08-0230-07 文献标志码:A 中图分类号:TP18

基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别

买买提阿依甫,吾守尔·斯拉木,帕丽旦·木合塔尔,杨文忠

(新疆大学 信息科学与工程学院,乌鲁木齐 830046)

摘 要: 为在缺乏资源和不依赖人工特征的情况下提高维吾尔文命名实体的识别性能,构建基于 BiLSTM-CNN-CRF 的神经网络模型。采用卷积神经网络训练具有维吾尔文单词的后缀、前缀等形态特征的字符向量,利用 skip-gram 模型对大规模语料进行训练,生成具有语义信息的低维度稠密实数词向量。在此基础上,将字符向量、词性向量和词向量拼接的向量作为输入,构建适合维吾尔文命名实体识别的 BiLSTM-CRF 深层神经网络。实验结果表明,该模型能够解决命名实体的自动识别问题,具有较强的鲁棒性,F1 值达到 91.89% 。

关键词: 递归神经网络;卷积神经网络;条件随机场;维吾尔文;命名实体识别

中文引用格式: 买买提阿依甫,吾守尔·斯拉木,帕丽旦·木合塔尔,等. 基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别[J]. 计算机工程,2018,44(8):230-236.

英文引用格式: Maimaitiayifu, SILAMU Wushouer, MUHETAER Palidan, et al. Uyghur named entity recognition based on BiLSTM-CNN-CRF model[J]. Computer Engineering, 2018, 44(8):230-236.

Uyghur Named Entity Recognition Based on BiLSTM-CNN-CRF Model

Maimaitiayifu, SILAMU Wushouer, MUHETAER Palidan, YANG Wenzhong

(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

【Abstract】 In order to obtain better Uyghur Named Entity Recognition (NER) performance without the need of resources and relying on artificial features is an important problem to be solved. In this paper, a neural network model based on BiLSTM-CNN-CRF is constructed. Firstly, Convolutional Neural Network (CNN) is used to train character vectors with morphological characteristics such as suffix and prefix of Uyghur words. Then, skip-gram model is used to train large-scale corpus to generate word vectors with semantic information. Finally, a BiLSTM-CRF deep neural network suitable for Uyghur NER is constructed by using concatenated vectors which includes the character vectors, part-of-speech vectors and word vectors as input. Experimental results show that the proposed model can solve the problem of automatic recognition of named entities and has good robustness. Its F1 value reaches 91.89 % .

【Key words】 recurrent neural network; Convolutional Neural Network (CNN); Condition Random Field (CRF); Uyghur; Named Entity Recognition (NER)

DOI: 10.19678/j.issn.1000-3428.0050502

0 概述

命名实体识别[1](Named Entity Recognition,NER)是自然语言处理(Natural Language Processing,NLP)工作中具有挑战性的任务之一,通过它可以准确地从文本中识别出人名、机构名、地名、时间、日期、货币、百分号等信息,为话题识别、话题跟踪、信息检索、机器翻译、舆情分析等高级 NLP 任务提供重要的特征信息。过去 NER 任务多采用基于规则的识别方法、基于统计机器学习的识别方法(包括隐马尔可夫模型、条件随机场模型、支持向量机等)和基于规则和统计相结合的混合识别方法[2]。近年来,深度神经网络在自然语言处理领域受到了广泛的关注,相比于上述方法,基于深度神经网络的方法具有泛化性更强、更少依赖人工特征的优点。因此,面向汉语和英语等大语言,研究人员已提出了许多基于深度神经网络的 NER 模型,但针对以维吾尔语为代表的低资源少数民族语言的研究较少。维吾尔命名实体识别研究大多只针对维吾尔人名的识别,关于人名、地名、机构名的通用研究较少,目前多数研究都是基于规则或统计模型的方法。

基金项目: 国家重点基础研究计划项目(2014CB340506);国家自然科学基金(61363063);新疆大学多语种重点实验室开放课题(XJDX0905-2013-01)。

作者简介: 买买提阿依甫(1981—),男,博士,主研方向为机器学习、网络舆情分析;吾守尔·斯拉木,教授、中国工程院院士、博士生导师;帕丽旦·木合塔尔(通信作者),博士;杨文忠,副教授。

收稿日期: 2018-02-15 **修回日期:** 2018-03-30 **E-mail:** 179095844@qq.com

针对维吾尔文命名实体识别问题,本文构建基于 BiLSTM-CNN-CRF 的混合深度学习模型。首先利用卷积神经网络(Convolutional Neural Network,CNN)模型捕获单词的字符级特征向量;然后将字符级特征向量、词性向量和词向量拼接的混合向量作为 BiLSTM 模型的输入进行训练,获取语句单词之间隐含的语义特征;最后通过 CRF 模型得到最优标注序列。

1　神经网络体系结构

本节将详细描述 BiLSTM-CNN-CRF[1] 神经网络体系结构的各组成部分,从下至上逐一介绍神经网络中的各神经层。

1.1　字词向量特征

1.1.1　字符特征

维吾尔语是典型的黏着语[3],具有复杂的形态变化。从文字信息处理的角度出发,维吾尔文字属于复杂文本信息处理的范畴。现行维吾尔文使用的文字为基于阿拉伯字符的文字,该类文字的形状特征为不等宽的字符,每个字符根据在词中出现的位置又有不同的形状,书写特征是自右向左书写(数字和其他非阿拉伯字符保持自己的书写顺序),与英语和汉语顺序相反。

现行维吾尔文有 32 个字母。每个字母按出现在词首、词中、词末的位置有不同的形式。字母表中的单式除代表该字母的独立形式外,一般出现在词末不可连字母之后,前式出现在可连字母之前,中式出现在词中 2 个可连字母中间,末式出现在词末可连字母之后,有些字母只有单式和末式,这样维吾尔文 32 个字母实际共有 126 种写法。为了降低字符向量维度,本文通过设计现行维吾尔文到拉丁维吾尔文的转换算法,从而将维吾尔文字符转换为一个拉丁字符,这样只用 32 个拉丁字母就可以表示维吾尔文。

1.1.2　词向量

词向量[4]的主要设计思想是通过神经网络学习词语的联合概率分布,将语料中的单词映射到指定的 d 维稠密实数向量。word2vec 用到了 2 个重要的模型:CBOW 模型和 Skip-Gram 模型[5]。

维吾尔句子中单词之间用空格或标点符号来分割。维吾尔文单词在结构上可以分为词根和词干:词根是不可分割的最小语义单元;词干是由几个词根或词根和词缀连接构成,单词一般由词干和词缀(附加成分)连接构成,每个词的变化形式最多可达到数百种。例如:词根为 ish(事宜,事情),通过对其连接构词词缀 qi,可以得到词干 ish + qi = ishqi(工人),可以将单词结构表示为:单词 = 词缀 + 词干 + 后缀 1 + 后缀 2 + 后缀 3 + …。例如:ish + qi + lar + ning = ishqilarning(工人们的)。为了获取维吾尔单词隐含的丰富信息,本文未对维吾尔词语进行词干提取,保留了词干与词缀,对语料库中的每个单词进行训练预先生成了对应的词向量。

本文利用 word2vec 工具的 Skip-Gram 模型对从网络上下载的无标注维吾尔语料库进行训练生成了词向量,假设语料库由 $w_1, w_2, \cdots, w_m$ 个单词组成,Skip-Gram 模型的目标是使以下函数最大化:

$$F = \frac{1}{M}\sum_{m=1}^{M}\sum_{\substack{-n\leqslant j\leqslant n\\ j\neq 0}} \lg p(w_{t+j}\mid w)$$

其中,n 是训练窗口大小,训练时根据窗口大小获取当前词语的上下文相关词语。

经过 word2vec 生成的词向量为:$\boldsymbol{v}_i = [a_0, a_1, \cdots, a_d]$,其中 $d = 300$ 表示词向量的维度。通过 word2vec 训练生成的词向量每一维都包含丰富的上下文信息。例 1 给出了维吾尔语料库中一个单词用 word2vec 生成的词向量。

例 1　چىرايلىق(漂亮) = [0.721 4　-1.105 4　-0.159 6 …　-1.241 2　-0.163 8　0.3247]。

通过 word2vec 的相似度计算算法获取以下单词的相似单词,如图 1 所示。可以看出,训练出的维吾尔文单词“چىرايلىق”(漂亮)的词向量与“سالاپەتلىك”(有风度)和“قاملاشقان”(英俊)等词的词向量相似度很近。“ياخشى”(好)与“كۆڭۈلدىكىدەك”(理想的)和“قالتىس”(了不起)等单词的词向量相似度较近,具有很近的语义关系。

```
In [4]: word_vectors.most_similar('چىرايلىق')

Out[4]: [('0.7424442768096924', 'سالاپەتلىك'),
         ('0.7226890921592712', 'قاملاشقان'),
         ('0.721005916595459', 'كۆزۈمىمىز'),
         ('0.7178435325622559', 'نازاكەتلىك'),
         ('0.7155851125717163', 'جەزىبدار'),
         ('0.7105744481086731', 'كېلىشكەن'),
         ('0.7031242847442627', 'جەلىپكار'),
         ('0.6791760921478271', 'ناخشا'),
         ('0.6757470965385437', 'ئەتنىز'),
         ('0.6710501909255981', 'لاتاپەتلىك')]

In [9]: word_vectors.most_similar('ياخشى')

Out[9]: [('0.6661605834960938', 'كۆڭۈلدىكىدەك'),
         ('0.6452983021736145', 'ياخشىمۇ'),
         ('0.6363388299942017', 'ناچار'),
         ('0.5860540866851807', 'قالتىس'),
         ('0.5591778755187988', 'يامان'),
         ('0.5448861718177795', 'ياخشىراق'),
         ('0.5268281698226929', 'مۇكەممەل'),
         ('0.5254047513008118', 'چىنىنىڭ'),
         ('0.5196686387062073', 'ئاقىلانە'),
         ('0.4989717602797974', 'ئوبدان')]
```

图 1　通过维吾尔文词向量获取的相似单词

1.1.3　词性特征

本文采用词向量很方便地添加了其他新的特征。例如,对于词语可以添加词性特征、字符特征等信息,通过这些信息可以对词语进行抽象化,能够进一步发现语句中词语的结构联系。因此,本文加入了字符特征和词性特征,进一步提高了命名实体识

别的性能。维吾尔文词性有 2 种标注方法:一级词性和二级词性,如表 1 所示。

表 1 维吾尔文词性标注设置

类别	词性标识
一级词性(1-POS)	名词(n),动词(v),形容词(a),副词(d),连词(c),代词(r),数词(m),量词(q),时间词(t),叹词(e),象声词(o),标点符号(q),后缀词(h),语气词(y)
二级词性(2-POS)	动词(av,nv,dv),名词(na,nr,ns,nt,nz,nk),副词(dh,dw,do,dd),数词(md,mm,mo,mr),代词(ra,rs,rk,re,ry,rb,ro),时间词(tt)

由于本文语料库没有二级词性标注数据,因此实验只使用一级词性特征。使用维度为 4 的实数向量表示词性向量,最后与词向量和字符向量拼接构成混合向量作为 BiLSTM 模型输入,提高了模型的命名实体识别性能。

1.2 卷积神经模型

卷积神经网络[6]中卷积层能够提取文本数据的局部特征信息,通过使用卷积和最大池化层可以提取局部特征信息中最具有代表性的部分作为特征向量。现有研究表明,CNN 是一种从词的字符中提取形态信息(如词的前缀或后缀)并将其编码为神经表示的最有效方法,文献[1,7]采用 CNN 提取字符级特征在命名实体识别领域达到了很好的效果,因此,本文利用 CNN 提取维吾尔文单词的字符特征,通过使用字符级特征、单词词性和词向量相结合的方法提高模型的命名实体识别性能,但维吾尔文中不存在大小写的问题,在本文中并没有用到字符类型等特征,采用 CNN 提取的维吾尔文形态特征信息作为词向量的补充,从而模型的识别率得到了很好的提高。

CNN 模型结构如图 2 所示,其由字符向量表、卷积层和池化层组成。

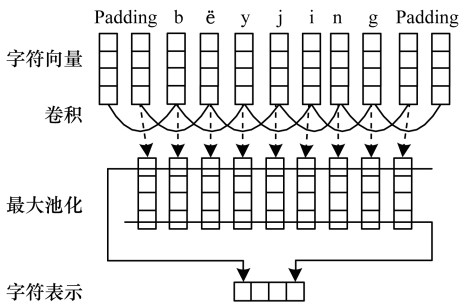

图 2 字符级 CNN 模型结构

对 CNN 网络中包括 32 个维吾尔文字母和37 个标点符号,再加上一个表示不在字符集中的不确定字符的共 70 个字符分别生成对应的字符向量,由这些向量构成字符向量表。字符向量表的作用是将单词中的每个字符转换成为对应的字符向量,然后生成单词对应的字符向量矩阵。由于单词长度不一样,因此生成的字符向量矩阵的大小也不一样。为解决该问题,本文以最长的单词长度为标准,利用 Padding 占位符补全单词两端[8],使字符向量矩阵的长度一致。此方法同样可以用于句子长度不一致的问题,最后字符向量表在卷积神经网络训练过程中通过反向传播机制自动更新字符向量矩阵。通过实验发现 CNN 网络可以有效获取维吾尔文单词中的前缀或后缀等形态特征信息。

1.3 BiLSTM 模块

1.3.1 LSTM 模块

递归神经网络[9]具有一定的记忆功能,可以被用来解决很多 NLP 问题,但是它并不能很好地处理长时依赖问题,存在梯度消失和梯度爆炸的问题。

为了解决传统递归神经网络的梯度消失等问题,研究者提出了 RNN 的特殊形式:长短期记忆网络(Long Short-Term Memory, LSTM)[9-10],传统 RNN 每一步的隐藏单元只是执行一个简单的 tanh 或 ReLU 操作[11]。LSTM 是递归神经网络的一种特殊形式,同样考虑时序关系,只是 LSTM 每个隐层节点还加一些特殊的结构,如图 3 所示。可以看出,LSTM 增加了记忆单元,主要由 3 个控制门,即遗忘门、输入门、输出门与一个记忆单元(cell)组成。LSTM 利用记忆单元对历史信息进行记录,并且这个记录是由3 个控制门来控制 LSTM 单元应写入、读取、输出的内容。因此,通过这些控制门,LSTM 能够缓解原始 RNN 所面临的“梯度消失”或“梯度爆炸”问题。

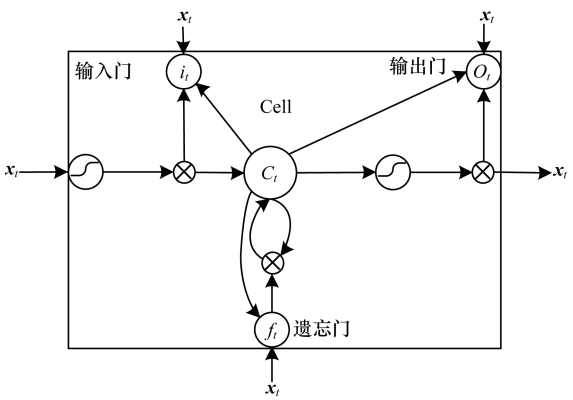

图 3 LSTM 单元结构

LSTM 单元在 t 时刻更新的公式如下:

$$i_t = \sigma(\boldsymbol{W}_i \boldsymbol{h}_{t-1} + \boldsymbol{U}_i \boldsymbol{x}_t + \boldsymbol{b}_i) \tag{1}$$

$$f_t = \sigma(\boldsymbol{W}_f \boldsymbol{h}_{t-1} + \boldsymbol{U}_f \boldsymbol{x}_t + \boldsymbol{b}_f) \tag{2}$$

$$\tilde{c}_t = \tanh(\boldsymbol{W}_c \boldsymbol{h}_{t-1} + \boldsymbol{U}_c \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4}$$

$$o_t = \sigma(\boldsymbol{W}_o \boldsymbol{h}_{t-1} + \boldsymbol{U}_o \boldsymbol{x}_t + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = o_t \odot \tanh(c_t) \tag{6}$$

其中,i_t为输入门,f_t为遗忘门,$\tilde{c}_t$ 为新记忆单元,c_t为最终记忆单元,o_t为输出门,h_t为隐藏层,σ 表示神

经网络中的 sigmoid 激活函数,输出范围为(0,1),tanh 表示双曲正切激活函数,⊙是对应元素点积,$\boldsymbol{x}_t$是在 t 时刻的输入词向量,$\boldsymbol{h}_t$是在 t 时刻的输出,$\boldsymbol{U}_i$、$\boldsymbol{U}_f$、$\boldsymbol{U}_c$、$\boldsymbol{U}_o$是在不同控制门对应输入向量的权重矩阵,$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$是隐藏层的权重矩阵,而 $\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_c$、$\boldsymbol{b}_o$是偏差向量。新记忆单元使用当前单词 $\boldsymbol{x}_t$和上一时刻隐藏层状态 $\boldsymbol{h}_{t-1}$产生当前新信息的 c_t。

在维吾尔文中,人名、机构名和地名中由 3 个以上的单词构成的情况较多,通过 LSTM 网络能够记忆单词间的长距离依赖关系的特点,有效识别出维吾尔文中的较长的人名、地名和机构名。例如:"samat bilan kvrash xinjiang aptonum rayonluk helik dohturhanisida ishlaydu."(赛买提和库莱西在新疆自治区人民医院工作。),针对这句话中的机构名"xinjiang aptonum rayonluk helik dohturhanisida"(新疆自治区人民医院),用传统的统计模型 CRF 进行识别时出现了无法完全识别的问题,而 LSTM 模型巧妙地识别出了类似长机构名。在这句话中 kvrash 是个兼类词(人名和动词),通过 LSTM 模型根据上下文历史信息正确识别 CRF 统计模型无法识别的兼类词。

1.3.2　BiLSTM 模块

在句子中命名实体的正确识别取决于词的上下文[12]。前后 2 个词对预测标签都很重要,如果能够获取过去和将来的上下文信息,对命名实体识别任务很有帮助。然而,LSTM 的隐藏状态 $\boldsymbol{h}_t$仅从过去获取信息,对未来一无所知。双向 LSTM[7,13](简称为 BiLSTM)是一种较好的解决方案,其有效性已在前人的工作中得以证明,基本思想是将每个顺序序列和逆序序列呈现到 2 个单独的隐藏状态,以分别捕获过去和将来的信息,然后将连接 2 个隐藏状态作为最终输出。BiLSTM 已经被证明在许多机器翻译、问题回答、序列标注等 NLP 任务中很有用。

假设输入句子为 $S=\{w_1,w_2,\cdots,w_n\}$,$H=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$是与之对应的隐式状态序列,BiLSTM 编码器按向前和向后 2 个方向分别生成隐式状态序列$\{h_1^f,h_2^f,\cdots,h_n^f\}$和$\{h_1^b,h_2^b,\cdots,h_n^b\}$。其中句子 S 中单词 w_i的隐式状态序列由向前和向后对应位置的隐式状态序列拼接生成,即 $\boldsymbol{h}_k=(h_k^f:h_k^b)$。

1.4　CRF 模块

条件随机场(Condition Random Field,CRF)[7,12]模型是一种用于标注和切分有序数据的条件概率模型。该模型结合了隐马尔可夫模型和最大熵模型的优点[13],避免了这些模型本身存在的一些缺点,能够有效地解决序列标注问题。可以把命名实体识别任务转化成一个序列标注任务,本文采用 SBIEO 标记策略(如表 2 所示)对语料进行标注,表 3 是使用 SBIEO 标记策略对给定现行维吾尔文句子进行转换为拉丁维吾尔句子后的标注示例。

表 2　CRF 模块 SBIEO 标签集

实体标记	开始标记	中间标记	结束标记
人名	B-PER	I-PER	E-PER
机构名	B-ORG	I-ORG	E-ORG
地名	B-LOC	I-LOC	E-LOC
非实体标记	O	O	O

表 3　维吾尔文命名实体标注方法示例

原文	alim	beijing	shehiridiki	chingxua	uniwersitida	oquydu
标注序列	S-PER	B-LOC	E-LOC	B-ORG	E-ORG	O

对于命名实体识别任务,本文使用 CRF 模型联合建模标注决策,而不是独立建模决策。将 CRF 层作为神经网络架构的最后一层,对 BiLSTM 模块的输出结果进行处理,获得最优的全局标注序列。

对于一个给定维吾尔文句子,本文用 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$表示对应的输入单词序列,假设 $\boldsymbol{P}$ 是大小为 $n\times k$ 的 BiLSTM 网络输出的分数矩阵,其中 k 是不同标签的数量,$P_{i,j}$对应第 i 个单词的第 j 个标签的分数。对于一个标签预测 $y=\{y_1,y_2,\cdots,y_n\}$,定义其分数为:

$$s(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{7}$$

其中,$\boldsymbol{A}$ 是转移分数矩阵,$A_{i,j}$表示从标签 i 转移到标签 j 的分数,y_0和 y_n是在句子开始和结束为位置添加的标签,因此,$\boldsymbol{A}$ 是一个大小为 $k+2$ 的方阵。

对于序列 y,本文采用 softmax 来生成所有:

$$p(y|X)=\frac{e^{s(X,y)}}{\sum\limits_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}} \tag{8}$$

在训练过程中最大化正确标签序列的对数概率:

$$\lg(p(y|X))=s(X,y)-\lg\Big(\sum_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}\Big) \tag{9}$$

其中,Y_X 是对于输入句子 X 的所有可能标签序列。从上式可以明显看出,本文的神经网络产生有效的输出标签序列。最终解码时,通过以下公式预测得分最大的输出序列:

$$y^*=\underset{\tilde{y}\in Y_X}{\operatorname{argmax}}\,s(X,\tilde{y}) \tag{10}$$

1.5　BiLSTM-CNN-CRF 模型

通过将 BiLSTM 的输出向量输入到 CRF 层来构造神经网络模型[11]。本文神经网络架构由 BiLSTM 模块、CNN 模块和 CRF 模块组成。第 1 层是输入层,主要负责将输入的句子进行字词向量的映射,为了便于后期处理首先通过转换算法将现行维吾尔文句子转换成拉丁维吾尔文,然后通过查询词向量表将文本转换为词向量序列,再对于文本中的每个单词,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 模块对字符向量矩阵进行卷积和最大池化,获得每个单词的字符级特征,每个单词的字符向量和词性向量与词向量拼接组合后的混合向量作为第 2 层神经网络模块 BiLSTM 的输入,最后用第 3 层CRF 模块将第 2 层的输出解码出一个最优的标记序列。本文神经网络的体系结构如图 4 所示。

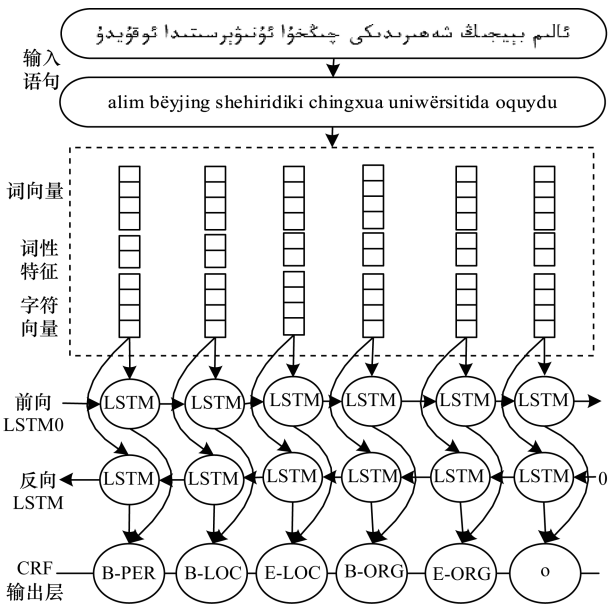

图 4 深度神经网络结构

2 神经网络训练

2.1 参数初始化

2.1.1 词向量

文献[10,14]已指出,词向量在提高序列标记任务性能中起到了至关重要的作用。目前缺乏公开的已训练好的维吾尔文词向量数据集。为了构建维吾尔文词向量,本文首先从知名度较高的几个新闻网站上下载了新闻数据(如表 4 所示),然后对收集到的 9.5 万条新闻数据(包含多余 3 500 万条词条,40 多万单词)用 gensim 的 skip-gram 模型进行训练[15],生成了维度为 300 的词向量,本实验设置窗口的大小为 5,单词频率最小值设为 4。

表 4 下载数据统计

网站名称	网站域名	新闻数量	词条
天山网	uy. ts. cn	42 556	15 405 272
Hawar 新闻网	hawar. cn	37 580	16 986 160
Nur 新闻	nur. cn	15 250	3 522 750

2.1.2 字符向量

实验对 70 个维吾尔文字符和标点符号使用随机均匀分布来初始化字符向量查询表,字符向量的维度设为 10,并且其取值范围为[-0.5,0.5]。

2.2 优化算法

目前神经网络中流行的优化算法有随机梯度下降(Stochastic Gradient Descent,SGD)、Momentum、Adagrad、Adadelta、RMSprop、Adam、Adamax 等[13,16-17],每个优化算法都有自己的特点,本文实验中使用了 SGD 算法进行优化,实验结果表明 SGD 优化算法提高了模型性能,学习率 η_0 初始值设为 0.001,momentum 设为0.9,每个训练周期学习率 η_0 通过公式:$\eta_t = \eta_0/1 + \rho t$ 来自动更新,其中延迟率为 $\rho = 0.5$,t 是已经完成的训练循环数。

2.3 Dropout 参数

在正则化方法中,Dropout[16] 是非常有用和成功的一种技术。一般来说,它会随机删除一些神经元,以在不同批量上训练不同的神经网络架构。在实验中 Dropout 的值和在模型中的位置很关键,直接影响到模型的性能。在多数神经网络研究中,Dropout 值设为 0.5 时的性能较好,能够有效防止过拟合问题,但在本文实验中,用不同的 Dropout 的值对模型进行了交叉验证。实验结果表明,Dropout 值为 0.63 时达到了最好的识别效果,在 BiLSTM 模型输入输出端两端都用了 Dropout 机制[17]。本文神经网络参数设定如表 5 所示。

表 5 神经网络参数设置

参数	值	参数	值
字符向量维度	10	词向量维度	300
字符特征窗口大小	5	最小单词频率	4
学习率	0.01	词特征窗口大小	5
Dropout	0.63	1 级词性个数	14

3 实验与结果分析

3.1 实验数据集

由于目前维吾尔命名实体识别缺乏公开的标注数据集,因此本文人工建立了一个维吾尔文命名实体识别数据集。所使用的语料是从政府新闻网站天山网下载的维吾尔语新闻数据,从中挑选 22 150 个维吾尔语句子,然后对其进行人工标注词性和命名实体标记,作为本文实验的维吾尔文命名实体识别语料库,如表 6 所示。

表 6 维吾尔文命名实体识别标注语料库

数据集	语料数量	人名数量	地名数量	机构名数量
训练集	17 750	5 126	3 576	2 683
开发集	2 250	574	468	561
测试集	2 250	467	395	378

3.2　实验结果

本文进行了 5 组实验对维吾尔文人名、地名、机构名进行命名实体识别,5 组实验都在实验室的 UNERDATA 数据集上进行命名实体识别。实验的评测方法是 F1 值、准确率、召回率[18]。

实验 1　实验目的:1)将 CRF 模型作为基准模型,测试统计模型 CRF 在实验室提供的 UNERDATA 数据集上的性能;2)汇总使用 CRF 模型进行命名实体识别时发现的一系列问题。实验中使用了目前较流行的统计模型 CRF ++[18],由于标记数据集是基于句子的,因此对于 CRF ++ 工具,只考虑了词级特征。使用 CRF 工具进行 UNER 任务后发现了以下问题:

1)CRF 统计模型对语料中没有出现的人名、地名无法正确识别。由于目前维吾尔文中尚缺少大型人名、地名和机构名称的标注语料库,导致统计模型无法正确识别命名实体。

2)维吾尔文中机构名称存在大量的缩写情况,CRF 模型对这种由单独字符组成的机构名缩写无法进行识别。例如:"ürümchi shehᵑ irlik j x idarsi"(乌鲁木齐市公安局)里面的"j x"是公安局的缩写,CRF 模型对这种缩写无法准确识别。

3)维吾尔文中的人名存在缺乏统一的写作风格,有些人名有几种写法。例如:人名"memetqasim"(买买提喀斯木)的另一个写法是"matqasim"(买提喀斯木)。

4)维吾尔文中存在大量兼类词,有些人名兼有其他含义。例如:人名"yalqun"(亚力坤)的另一个意思是火焰,CRF 模型无法利用上下文对这种兼类词进行正确识别,有时将类似人名识别标记为 O(其他)。

5)维吾尔人姓名基本上由 2 个词组成,但也存在一个人名由 3,4 个人名组成的情况,例如:"nurmemetoბulqasim"(努尔麦麦提吾布力卡斯木),CRF 模型无法对这些长人名正确识别。

6)维吾尔地名中大量存在长地名和长机构名,例如:"shinjang uyghur aptonom rayoni"(新疆维吾尔自治区),CRF 模型无法对类似长地名准确识别。

7)维吾尔文论坛、微博、新闻等网络文本中存在大量的拼写错误问题,CRF 模型无法对拼写错误的单词准确识别,其准确率为 78.35%,召回率为 75.78%,F1 值为 77.04%,

实验 2　实验目的:1)研究深度神经网络模型相对于统计模型在维吾尔命名实体识别上是否有优势;2)研究深度神经网络能否解决 CRF 统计模型中发现的问题。实验中分别用简单 RNN 模型、LSTM 模型和 BiLSTM 模型进行 UNER 任务。从表 7 中可以看出,简单 RNN 模型的性能和 CRF 模型基本一样,LSTM 模型和 BiLSTM 模型的性能都比 CRF 模型好,其中 BiLSTM 模型的 F1 值比 CRF 模型提高了 5.03%。

表 7　神经网络模型的实验结果　%

模型	准确率	召回率	F1 值
RNN	76.87	77.34	77.10
LSTM	80.91	79.41	80.15
BiLSTM	82.79	81.54	82.13

实验 3　实验目的:验证 CNN 模型的有效性。实验在 LSTM 模型和 BiLSTM 模型的基础上加入了 CNN 模型,使用 CNN 模型获取字符特征,然后将字符向量和词向量拼接后作为 LSTM 或 BiLSTM 模型的输入进行训练。从表 8 的实验结果可以看出,LSTM、BiLSTM 模型加入 CNN 网络后系统的识别能力都得到了提高,LSTM-CNN 模型的 F1 值比 LSTM 模型提高了 1.3%,BiLSTM-CNN 模型比 BiLSTM 模型 F1 值提高了 2.69%。

表 8　加入 CNN 模型后的实验结果　%

模型	准确率	召回率	F1 值
LSTM-CNN	82.91	80.04	81.45
BiLSTM-CNN	84.42	85.23	84.82

实验 4　实验目的:验证 CRF 模型加入到 BiLSTM-CNN-CRF 框架后系统的性能,并进一步提升系统 UNER 任务中的识别性能。在实验 3 的基础上,对 BiLSTM 模型的输出进行 CRF 层,输出概率最大的最优标记序列。从表 9 的实验结果可以看出,加入 CRF 层后 LSTM-CNN-CRF 模型和 BiLSTM-CNN-CRF 模型准确率都得到了提高,其中 BiLSTM-CNN-CRF 模型的 F1 值比 BiLSTM-CNN 模型提高了 4.3%。

表 9　加入 CRF 模型后各模型的实验结果　%

模型	准确率	召回率	F1 值
LSTM-CNN-CRF	87.43	85.69	86.55
BiLSTM-CNN-CRF	90.98	87.34	89.12

实验 5　实验目的:进一步提高系统的命名实体识别性能。在实验 4 的基础上,对系统的输入向量增加了词性向量,本文实验中由于 UNERDATA 数据集中未提供维吾尔二级词性的标记,只使用了一级词性作为特征进行了模型训练,将 CNN 模型提取出来的字符特征向量和词性向量与词向量拼接生成最终特征向量作为 RNN-CNN-CRF 模型的输入进行训练。从表 10 的实验结果中可以看出,词性向量加入到词向量后 BiLSTM-CNN-CRF 模型准确率有了提升,其中 BiLSTM-CNN-CRF 模型的准确率达到了 91.46%,F1 值达到了 91.89%,相对于基线 CRF 方法,其准确率提高了 13.11%,F1 值提高了 14.85%。

表 10　加入词性向量后各模型的实验结果　%

模型	准确率	召回率	F1 值
LSTM-CNN-CRF + POS	87.86	86.23	87.04
BiLSTM-CNN-CRF + POS	91.46	92.32	91.89

以上 5 组实验结果表明,本文建立的 BiLSTM-CNN-CRF 模型通过使用字符向量、词性向量和词向量组合的混合向量,在维吾尔文命名实体识别任务中达到了最好的性能。

4 结束语

针对维吾尔文命名实体识别任务,本文以传统的 CRF 统计模型作为基准进行实验,总结维吾尔文命名实体识别中出现的问题,进而构建基于 BiLSTM-CNN-CRF 框架的神经网络模型。该模型在 CNN 层捕获字符级特征向量,在 BiLSTM 层获取当前词语的过去和将来的上下文信息,在 CRF 层对 BiLSTM 层的输出进行解码,最终输出最优的标记序列。基于 UNERDATA 语料的实验结果进一步验证了 BiLSTM-CNN-CRF 框架对维吾尔文命名实体识别的有效性。

本文构建的 BiLSTM-CNN-CRF 深度学习模型能够在维吾尔文命名实体识别语料库上得到较好的实验结果,并已应用于维吾尔文网络舆情分析系统,有效识别出了文本中的人名、地名和机构名,提高了舆情系统分析能力。后续将进一步完善语料库,加入二级词性标注特征信息,并在新语料库的基础上测试本文模型的性能。

参考文献

[1] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs [EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1511.08308v1.pdf.

[2] 张海楠,伍大勇,刘 悦,等. 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.

[3] 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),2013,53(6):873-877.

[4] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1310.4546.pdf.

[5] REI M,CRICHTON G K O,PYYSALO S. Attending to characters in neural sequence labeling models[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1611.04361v1.pdf.

[6] DERNONCOURT F,LEE J Y,SZOLOVITS P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1705.05487.pdf.

[7] MA X Z,HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1603.01354.pdf.

[8] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1603.01360.pdf.

[9] LI L S,MAO T,HUANG D,et al. Hybrid models for Chinese named entity recognition[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. Beijing,China:[s.n.],2006:72-78.

[10] COLLOBERT R,BOTTOU J W L,KARLEN M,et al. Natural language processing(almost) from scratch[J]. The Journal of Machine Learning Research,2011,12:2493-2537.

[11] HUANG Z H,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1508.01991v1.pdf.

[12] 陈 斌,周 勇,刘 兵. 基于卷积长短期记忆网络的事件触发词抽取方法[J/OL]. 计算机工程:1-7[2018-02-17]. http://kns.cnki.net/kcms/detail/31.1289.TP.20180206.1410.002.html.

[13] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1310.4546.pdf.

[14] 喻靖民,向凌云,曾道建. 一种基于 Word2vec 的自然语言隐写分析方法[J/OL]. 计算机工程:1-7[2018-02-17]. http://doi.org/10.19678/j.issn.1000-3428.0050407.

[15] 王洪亮. 基于词向量聚类的中文微博产品命名实体识别[J]. 兰州理工大学学报,2017,43(1):104-110.

[16] YANG Z, SALAKHUTDINOV R, COHEN W. Multi-task cross-lingual sequence tagging from scratch[EB/OL]. [2017-05-11]. https://arxiv.org/pdf/1603.06270v1.pdf.

[17] MIKOLOV T, KARAFIAT M, BURGET L, et al. Recurrent neural network based language model[C]//Proceedings of INTERSPEECH'10. Makuhari, Japan:[s.n.],1045-1048.

[18] HAMMERTON J. Named entity recognition with long short-term memory[C]//Proceedings of HLT-NAACL'03. [S.l.]:ACL,2003:172-175.

编辑 金胡考

(上接第 229 页)

[12] 岳 昆,王晓玲,周傲英,等. Web 服务核心支撑技术:研究综述[J]. 软件学报,2004,15(3):428-442.

[13] GRUBERT R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition,1993,5(2):199-220.

[14] 吴 健,吴朝晖,李 莹,等. 基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机学报,2005,28(4):595-602.

[15] 孙 萍,蒋昌俊. 利用服务聚类优化面向过程模型的语义 Web 服务发现[J]. 计算机学报,2008,31(8):1340-1353.

编辑 顾逸斐