DOI: 10.14188/j.1671-8836.2017.03.009

基于 CNN 和 LSTM 混合模型的中文词性标注

谢　逸，饶文碧，段鹏飞†，陈振东

（武汉理工大学 计算机科学与技术学院，湖北 武汉 430070）

摘　要：中文词性标注具有重要的作用，它的准确性和标注速度直接影响到自然语言处理的后续任务. 提出一种基于 CNN(convolutional neural network)和 LSTM(long-short term memory)混合模型进行中文词性标注. 该模型采用三层结构，用词向量和 CNN 的滑动窗口特性产生词语表示特征，LSTM 的时序性来产生词性标注的序列标签. 分别在 PFR《人民日报》语料库、CTB7.0 和 CoNLL09 语料库上对该模型进行测试，在未加入任何人工特征的条件下，对词语进行词性标注，词性标注效果好于 HMM(hidden Markov model)、MLP(multi-layer perceptron)、CNN 和 LSTM.

关　键　词：词性标注；卷积神经网络；长短期记忆；实验分析

中图分类号：TP 391　　**文献标识码**：A　　**文章编号**：1671-8836(2017)03-0246-05

A Chinese POS Tagging Approach Using CNN and LSTM-Based Hybrid Model

XIE Yi, RAO Wenbi, DUAN Pengfei†, CHEN Zhendong

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, Hubei, China)

Abstract: Chinese nature language processing (NLP) is a hot area of research currently. In Chinese NLP, part-of-speech (POS) tagging is very important to follow-up tasks. In this paper, we present a novel model based on CNN (convolutional neural network)and LSTM(long-short term memory). There are three sub-layers in this model, in which CNN layer generates word representation features with word vector and sliding windows, LSTM layer generates POS tagging using its sequence. We evaluate our hybrid model on three datasets (People's daily, CTB7.0 and CoNLL09) for POS tagging task. Without handcrafted features, the classification accuracy rate of this method is higher than the rate achieved by HMM (hidden Markov model), MLP(multi-layer perceptron), CNN and LSTM on these datasets.

Key words: POS (part-of-speech) tagging; convolutional neural network; long-short term memory; experiment analysis

0　引　言

词性标注(part-of-speech tagging，简称 POS tagging)是自然语言处理里的重要问题，它被广泛地用于机器翻译、文字识别、语音识别、信息检索等相关领域[1]. 词性标注通过上下文赋予每个词语一个正确的候选词性，从而为自然语言中的词法分析、语法分析和语义分析提供支撑.

词性标注的方法归纳起来，可以分为 4 类：

1) 基于规则的方法. 2008 年，王广正等[2]提出了基于规则优先级的词性标注方法. 基于规则的方法简单，易于实现，但手工构造规则是一项非常艰难的任务.

2) 基于统计的方法. 该方法客观性强，准确性较高，但需要处理兼类词和未登录词的问题. 常见的方法有：隐马尔科夫模型(hidden Markov model，简称 HMM)、最大熵(maximum entropy，简称 ME)、条件随机场(conditional random fields，简称 CRF)等. Kupiec 等[3]使用隐马尔科夫模型，在 50 万词的英文语料库上，词性标注准确率达到了 95%.

收稿日期：2016-08-10　† 通信联系人　E-mail: duanpf@whut.edu.cn

基金项目：国家高技术研究发展计划(863)资助项目(2015AA015403).

作者简介：谢　逸，男，博士生，现从事机器学习，自然语言处理方面的研究. E-mail: 7792221@qq.com

3) 基于规则和统计的方法,充分利用两种方法的各自优势,比单一使用一种的准确性要高[4],但该类方法依赖于建立的规则或人工特征的选取,同时也与任务领域的资源有很大的相关性,一旦领域变化,标注效果就会受较大影响.

4) 基于深度学习的方法.通过对数据多层建模获得数据的特征和分布式表示,避免繁琐的人工特征抽取,具有良好的泛化能力[5].深度学习常见的模型主要有多层感知器(multi-layer perceptron,简称 MLP)、自动编码器(AutoEncoder,简称 AE)、卷积神经网络(convolutional neural network,简称 CNN)和长短期记忆模型(long-short term memory,简称 LSTM)等.这些模型中 CNN[6] 和 LSTM[7] 应用较为广泛.CNN 是目前效果最好的深度学习模型之一,利用滑动窗口,可以很好的解决词的组合特征及一定程度上的依赖问题,表达句子之间的关系较为自然,由于参数共享,因此它的计算量较小,计算速度较快.2011 年 Collobert 等[8] 在 Word Embedding 基础上,设计了一种利用卷积计算和滑动窗口的前向神经网络,并把它应用于英文词性标注和实体识别等任务上.对于词性标注这种标注任务来说,可以把输入的文本看作是线性序列,文本中字或词为序列的一个元素,每个元素的标注很大程度依赖于前面元素的信息.LSTM 隶属于循环反馈神经网络,它可以将文本中某一序列元素与某时刻模型的输入对应起来,利用隐层单元的记忆模块,保存长间隔信息,对序列元素逐一标注.2015 年,Sundermeyer 等[9] 利用 LSTM 进行语言建模,Wang 等[10] 使用 LSTM 进行英文词性标注,取得较好的效果.

本文提出了一种基于 CNN 和 LSTM 的混合模型来进行中文词性标注的方法,混合模型利用 CNN 滑动窗口和权值共享[11] 来获得局部上下文信息,从而生成词语表示特征并作为下一层的输入,LSTM 的时序性非常适合标注这种序列任务[12],将两者结合起来,充分利用两者的各自优势,中文词性标注的性能得到了显著的提升.

1　本文方法

1.1　模型结构

以"世界杯 跳水赛 中国 选手 再 夺 两 枚 金牌"为例,本文设计的基于 CNN 和 LSTM 的混合神经网络模型结构如图 1 所示.模型分为 3 个层次:第一层通过使用 Word2Vec 将文本中的词语转换成为词向量;第二层为 CNN 层,将第一层所产生的词向量输入到 CNN 层,利用滑动窗口,计算前后词对当前词的影响,生成词语表示特征;第三层为 LSTM 层,将 CNN 层生成的各词的词语表示特征依次输入 LSTM 各节点,预测最后的词性标注标签.

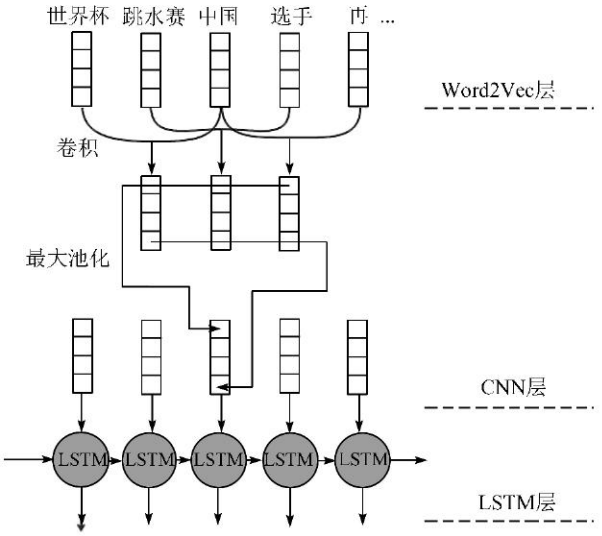

图 1　面向词性标注的 CNN 和 LSTM 混合模型

Fig. 1　Hybrid model using CNN and LSTM

1.1.1　词向量处理层

自然语言理解的问题要转化为机器能够处理的问题,第一步就必须将这些符号数字化,即将文本的表达映射到 k 维的向量空间里去.本文用 2013 年 Google 开发的 Word2Vec 将已分词的语料库中的中文词语转换为词向量.经过 Word2Vec 训练的词向量如下

$$\boldsymbol{v}_i = [a_0, a_1, \cdots, a_d] \tag{1}$$

(1)式中,d 为词向量的维度,模型初始化时设置.

1.1.2　卷积神经网络处理层(CNN 层)

使用卷积神经网络来提取词语的上下文信息,生成词语的表示特征,其结构如图 2 所示.

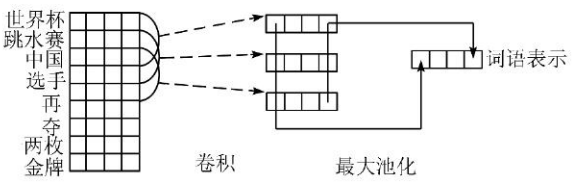

图 2　卷积神经网络处理层结构图

Fig. 2　Architecture for CNN processing layer

令 $\boldsymbol{v}_i$ 为第 i 个词的词向量,$\boldsymbol{v}_i$ 的维度为 d 维,当句子词语数为 L,卷积神经网络的滑动窗口大小为 k 时,落入第 j 个($j \leqslant L-1$)滑动窗口中的词向量依次为 $\boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \cdots, \boldsymbol{v}_{j+k-1}$,可以将它们表达为窗口向量,如下

$$\boldsymbol{X}_j = [\boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \cdots, \boldsymbol{v}_{j+k-1}] \tag{2}$$

与当前词 v_j 有关的窗口向量依次为：$\boldsymbol{X}_{j-k+1}$，$\boldsymbol{X}_{j-k+2}$，…，$\boldsymbol{X}_j$. 对于每个窗口向量 $\boldsymbol{X}_j$，用卷积核 $\boldsymbol{W}$ 进行卷积运算得到当前窗口特征，

$$\boldsymbol{Y}_j = f(\boldsymbol{X}_j \odot \boldsymbol{W} + \boldsymbol{b}) \tag{3}$$

(3)式中⊙是卷积乘，$\boldsymbol{b}$ 是偏置，f 是非线性激活函数，可以是 sigmoid 函数、tanh 函数或者 ReLU 函数，鉴于 ReLU 收敛速度快的特性，本文采用 ReLU 激活函数，ReLU 激活函数形式如下

$$g(x) = \max(0, x) \tag{4}$$

最大池化(max pooling)会有效的减少特征和参数，降低计算的复杂度，因此，在完成卷积运算后采用最大池化的方法来最大化词语特征表示. 窗口向量 $\boldsymbol{Y}_{j-k+1}, \boldsymbol{Y}_{j-k+2}, \cdots, \boldsymbol{Y}_j$ 组成窗口向量特征矩阵，对矩阵的每一行做 Max 操作，获得每一维的最大特征值，从而最大化词语的表示特征.

$$\alpha_j = \mathrm{Max}(\boldsymbol{Y}_{j-k+1}, \boldsymbol{Y}_{j-k+2}, \cdots, \boldsymbol{Y}_j) \tag{5}$$

1.1.3 长短期记忆模型处理层(LSTM 层)

LSTM 是一种特殊的 RNN(recurrent neural network). 一个 LSTM 单元是由一个 cell 和三个门(输入 input、输出 output 和遗忘 forget)组成，正是通过这种特殊的结构，LSTM 才能选择哪些信息被遗忘，哪些信息被记住. 某时刻 t，LSTM 单元各组成部分做如下更新：

$$i_t = \sigma(\boldsymbol{W}_i \boldsymbol{h}_{t-1} + \boldsymbol{U}_i \boldsymbol{\alpha}_t + \boldsymbol{b}_i) \tag{6}$$

$$f_t = \sigma(\boldsymbol{W}_f \boldsymbol{h}_{t-1} + \boldsymbol{U}_f \boldsymbol{\alpha}_t + \boldsymbol{b}_f) \tag{7}$$

$$\tilde{c} = \tanh(\boldsymbol{W}_c \boldsymbol{h}_{t-1} + \boldsymbol{U}_c \boldsymbol{\alpha}_t + \boldsymbol{b}_c) \tag{8}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{9}$$

$$o_t = \sigma(\boldsymbol{W}_o \boldsymbol{h}_{t-1} + \boldsymbol{U}_o \boldsymbol{\alpha}_t + \boldsymbol{b}_o) \tag{10}$$

$$h_t = o_t * \tanh(c_t) \tag{11}$$

其中，σ 表示 sigmoid 激活函数，$*$ 是元素乘，α_t 为 t 时刻的输入向量，$\boldsymbol{h}_t$ 代表隐藏状态，$\boldsymbol{U}_i, \boldsymbol{U}_f, \boldsymbol{U}_c, \boldsymbol{U}_o$ 分别为 x_t 不同门的权值矩阵，而 $\boldsymbol{W}_i, \boldsymbol{W}_f, \boldsymbol{W}_c, \boldsymbol{W}_o$ 则为 $\boldsymbol{h}_t$ 不同门的权值矩阵，$\boldsymbol{b}_i, \boldsymbol{b}_f, \boldsymbol{b}_c, \boldsymbol{b}_o$ 为各门的偏置. i_t，f_t，c_t 和 o_t 分别代表了输入门、遗忘门、记忆单元状态和输出门.

LSTM 层的输入为 CNN 层的输出词向量 $\boldsymbol{\alpha}_j$ ($j \leqslant L$)，CNN 层的一个 $\boldsymbol{\alpha}_j$ 输出对应于一个时刻 t 的 LSTM 输入. LSTM 层的输出送入 Softmax 分类器进行分类，计算每个词对应的标注标签最大的概率，从而获得词语的词性标注标签，Softmax 公式如下：

$$P(y = i \mid x, \theta) = \frac{\mathrm{e}^{o_i}}{\sum_{k=1}^{N} \mathrm{e}^{o_k}} \tag{12}$$

其中，$P(y=i|x,\theta)$ 为样本 x 属于 i 类的概率.

1.2 标注算法

利用以上的中文词性标注混合模型，我们设计了模型训练的算法(算法 1)，描述如下：

算法 1 模型训练算法

输入：训练集 D

初始化参数：词数 L，词向量维度 d，滑动窗口大小 k，隐层节点数 n，迭代次数 epochNum

训练过程：

Word2Vec 训练获得词向量 $\boldsymbol{v}_i$；

While 误差＞阈值 or 迭代次数＜epochNum

for $t=1\cdots L$

- 使用滑动窗口，获得窗口向量 $\boldsymbol{X}_j$(公式(2))，卷积并最大池化，产生词语表示特征 $\boldsymbol{\alpha}_j$

$$\boldsymbol{\alpha}_j = \mathrm{Max}(\boldsymbol{Y}_{j-k+1}, \boldsymbol{Y}_{j-k+2}, \cdots, \boldsymbol{Y}_j)$$

- 一个 $\boldsymbol{\alpha}_j$ 对应于一个时刻 t，输入 LSTM，产生概率向量

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

- LSTM 输出送入 Softmax 分类器，产生每个词语的分类标签

$$P(y=i|x,\theta) = \frac{\mathrm{e}^{o_i}}{\sum_{k=1}^{N} \mathrm{e}^{o_k}}$$

- 计算对数损失函数 L：

$$L(y, P(y|x)) = -\log P(y|x)$$

输出：权值确定的多层模型

2 实验与结果分析

2.1 数据集

本文采用 PFR《人民日报》1998 年 1 月份的语料库(*http://www.icl.pku.edu.cn/icl_res/*)、CoNLL09(*http://ufal.mff.cuni.cz/conll2009-st/index.html*)和 CTB7.0(*https://catalog.ldc.upenn.edu/LDC2010T07*)作为实验数据. PFR《人民日报》语料库是由北京大学计算语言学研究所与富士通研究开发中心有限公司共同制作. 各语料库规模如表 1 所示.

表 1 实验用语料库规模

Table 1 Experiment corpus

语料库	句子数	总词数
PFR	31 668	776 817
CTB7.0	2 796	59 955
CoNLL09	2 577	73 154

实验中，语料库将分割为 3 部分：训练集、开发集和测试集. PFR 语料库的训练集、开发集和测试集的比例为 7∶1∶2. 在进行词性标注时，根据现代汉语语料库加工规范，PFR 语料库有 39 种不同的标注[13]. CoNLL09 和 CTB7.0 按照各自官方文档提供的划分方法进行训练集、开发集和测试集的

划分.

2.2　实验参数设置

要获得一个较优的深度模型,模型里参数的设置是否恰当是关键因素.CNN 层主要涉及的参数有:词向量维度、滑动窗口大小、卷积核数和激活函数;LSTM 层主要涉及的参数有:节点数、Dropout 值、优化方法及学习率.本文选择了词向量维度(取值:50,100,200,300,400)、滑动窗口大小(取值:3,5,7,9)、节点数(取值:100,200,300)、学习率(取值:0.001,0.005,0.01,0.015)进行实验来取优,其余参数为默认值,实验结果见图 3.

通过实验对比,从图 3 中可知,模型中 CNN 层和 LSTM 层的参数分别设置如表 2 和表 3 所示,模型性能达到最佳.

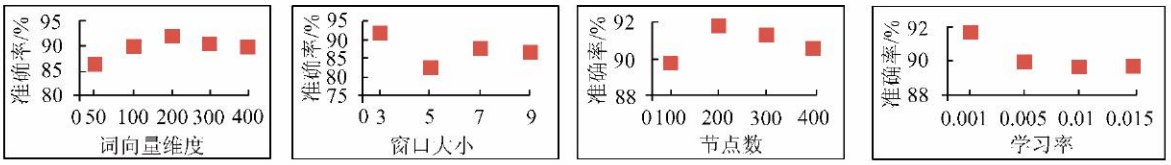

图 3　各参数对实验效果的影响

Fig. 3　Effects of different parameters

表 2　CNN 层参数设置

Table 2　Parameters of CNN layer

词向量维度	200
滑动窗口大小	3

表 3　LSTM 层参数设置

Table 3　Parameters of LSTM layer

节点数	200
学习率	0.001

2.3　中文词性标注实验及结果分析

为了验证混合模型的有效性,本文在未加入任何人工特征的前提条件下,在相同的语料库上分别采用 HMM、MLP、CNN、LSTM 以及本文提出的混合模型进行词性标注,HMM 采用一阶马尔科夫模型,对出现兼类词(即一词多性)的情况不做特殊处理.实验从准确率和模型标注速度两方面对模型进行评价,准确率和模型标注速度的实验结果如表 4 和表 5 所示.实验使用的机器配置为:Intel(R) Xeon E5-2620@2.0GHz,24GB RAM,GPU Quadro k2000.

表 4　各模型词性标注准确率

Table 4　Accuracy of different model

模型	PFR	CTB7.0	CoNLL09
HMM	0.910 3	0.912 2	0.909 6
MLP	0.908 8	0.908 3	0.902 5
CNN	0.904 7	0.903 2	0.899 4
LSTM	0.905 8	0.906 7	0.902 2
CNN+LSTM	**0.918 0**	**0.915 2**	**0.916 8**

从表 4 可以看出,在未加入任何人工特征的条件下,对 PFR 语料库进行 39 种词性标注的准确率按高到低的顺序依次为:CNN + LSTM、HMM、MLP、LSTM、CNN.CNN+LSTM 混合模型的准确率比隐马尔科夫(HMM)高 0.77 个百分点,比 LSTM 高 1.22 个百分点,比 CNN 高 1.33 个百分点,这得益于混合模型有效的利用了 CNN 的滑动窗口的特性来获取上下文相关信息产生词语表示特征,LSTM 的时序性来产生标注的序列标签.相同的情况也出现在 CTB7.0 和 CoNLL09 语料库上.

表 5 中的数据是给定 560KB 的文本,让三种模型对其进行标注而计算得出的.实验表明,混合模型标注的速度较 CNN 或 LSTM 的标注速度略低,分析其原因,是因为混合模型在层次上比 CNN 或 LSTM 有所增加,计算复杂度有所增加,标注速度就略有降低.

表 5　各模型的标注速度对比

Table 5　POS tagging speed of each model

模型	标注速度/KB·s^{-1}
CNN	1.323 8
LSTM	1.355 9
CNN+LSTM	1.233 4

由于汉语语料库加工规范规定的词性标签较多,词性标注的问题最终仍然是分类的问题,所以标签的个数直接影响到最后的标注准确率.标签越多,准确率就会下降,标签越少,准确率就会越高.如果在充分研究中文词性标注的语言特点之上,人工加入某些特征,那么 CNN+LSTM 模型对中文词性标注的准确率将会有很大的提升空间.

3　结　论

本文设计了一种用于中文词性标注的 CNN 和 LSTM 混合模型,在未使用任何人工特征的条件下,对中文词语进行词性标注.模型有效的利用

CNN滑动窗口以及LSTM时序性的特性.实验结果表明,使用基于CNN和LSTM的混合模型的中文词性标注方法在不加入任何其他人工特征的基础上能够取得很好的标注效果.在同样的实验条件下,混合模型的标注效果要优于HMM、MLP、CNN以及LSTM.今后将针对未登录词和兼类词,研究其对词性标注的影响,使用更多的特征,提高标注的准确率;在进行中文词性标注时,将加入某些特定的中文语言特征,使得准确率进一步提升.

参考文献:

[1] 王丽杰,车万翔,刘挺. 基于SVMTool的中文词性标注[J].中文信息学报,2009,**23**(4):16-21.
WANG L J, CHE W X, LIU T. An SVMtool-based Chinese POS tagger[J]. *Journal of Chinese Information Processing*, 2009, **23**(4):16-21(Ch).

[2] 王广正,王喜凤.一种基于规则优先级的词性标注方法[J].安徽工业大学学报(自然科学版),2008,**25**(4):426-429.
WANG G Z, WANG X F. A method of POS tagging based on priority of rules[J]. *Journal of Anhui University of Technology (Natural Science)*, 2008, **25**(4):426-429(Ch).

[3] KUPIEC J. Robust part-of-speech tagging using a hidden Markov model[J]. *Computer Speech & Language*, 1992, **6**(3):225-242.

[4] 王素格,张永奎. 汉语词性自动标注系统的设计与实现[J]. 计算机工程,2001,**27**(3):7-8.
WANG S G, ZHANG Y K. The design and implementation of the Chinese part-of-speech automatic tagging system[J]. *Computer Engineering*, 2001, **27**(3):7-8(Ch).

[5] ZHENG X, CHEN H, XU T. Deep learning for Chinese word segmentation and POS tagging[C]// *Proceedings of the* 2013 *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: The Association for Computational Linguistics, 2013: 647-657.

[6] BOUVRIE J. *Notes on convolutional neural networks* [R]. Cambridge: MIT, 2006.

[7] SEPP H, JURGEN S. Long short-term memory[J]. *Neural Computation*, 1997, **9**(8): 1735-1780.

[8] COLLOBERT R, WESTON J, BOTTOU, *et al*. Natural language processing (Almost) from scratch[J]. *Journal of Machine Learning Research*, 2011, **12**(1):2493-2537.

[9] SUNDERMEYER M, NEY H, SCHLUTER R. From feedforward to recurrent LSTM neural networks for language modeling[J]. *IEEE/ACM Transactions on Audio Speech & Language Processing*, 2015, **23**(3):517-529.

[10] WANG L, TIAGO L, LUIS M, *et al*. Finding function in form: Compositional character models for open vocabulary word representation[C]//*Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: The Association for Computational Linguistics, 2015:1520-1530.

[11] MA X, HOVY E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF[C]//*Proceedings of the* 54*th Annual Meeting of Association for Computational Linguistics*. Stroudsburg: The Association for Computational Linguistics, 2016:1064-1074.

[12] PLANK B, SOGGARD A, GOLDBEERG Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[C]// *Proceedings of the* 54*th Annual Meeting of Association for Computational Linguistics*. Stroudsburg: The Association for Computational Linguistics, 2016: 412-418.

[13] 俞士汶,段慧明,朱学锋,等. 北京大学现代汉语语料库基本加工规范(续)[J]. 中文信息学报,2002,**16**(6):49-64.
YU S W, DUAN H M, ZHU X F, *et al*. The basic processing of contemporary Chinese corpus at Peking University specification[J]. *Journal of Chinese Information Processing*, 2002, **16**(6):49-64(Ch).

□