基于注意力长短时记忆网络的中文词性标注模型

司念文[1] 王衡军[1] 李 伟[2] 单义栋[1] 谢鹏程[3]
(中国人民解放军信息工程大学三院 郑州 450001)[1] (66083部队 北京 100144)[2]
(西安交通大学数学与统计学院 西安 710049)[3]

摘 要 针对传统的基于统计模型的词性标注存在人工特征依赖的问题,提出一种有效的基于注意力长短时记忆网络的中文词性标注模型。该模型以基本的分布式词向量作为单元输入,利用双向长短时记忆网络提取丰富的词语上下文特征表示。同时在网络中加入注意力隐层,利用注意力机制为不同时刻的隐状态分配概率权重,使隐层更加关注重要特征,从而优化和提升隐层向量的质量。在解码过程中引入状态转移概率矩阵,以进一步提升标注准确率。在《人民日报》和中文宾州树库CTB5语料上的实验结果表明,该模型能够有效地进行中文词性标注,其准确率高于条件随机场等传统词性标注方法,与当前较好的词性标注模型也十分接近。

关键词 词性标注,长短时记忆网络,注意力机制,上下文特征
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.04.009

Chinese Part-of-speech Tagging Model Using Attention-based LSTM

SI Nian-wen[1] WANG Heng-jun[1] LI Wei[2] SHAN Yi-dong[1] XIE Peng-cheng[3]
(The Third Institute, PLA Information Engineering University, Zhengzhou 450001, China)[1]
(66083 Army, Beijing 100144, China)[2]
(School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China)[3]

Abstract Because traditional statistical model based Chinese part-of-speech tagging relies heavily on manually designed features, this paper proposed an effective attention based long short-term memory model for Chinese part-of-speech tagging. The proposed model utilizes the basic distributed word vector as the unit input, and extracts rich contextual feature representation with bidirectional long short-term memory. At the same time, an attention based hidden layer is added in the network, and the attention probability is distributed for hidden state in different time to optimize and improve the quality of hidden vector. The state transition probability is employed in decoding process to further improve accuracy. Experimental results on PKU and CTB5 dataset show that the proposed model is able to make Chinese part-of-speech tagging effectively. It achieves higher accuracy than traditional methods and gets competitive results compared with state-of-the-art models.

Keywords Part-of-speech tagging, Long short-term memory, Attention mechanism, Contextual feature

1 引言

词性标注是中文自然语言处理领域中一项重要的基础性技术,目的是为句子中的每个词语分配特定的词性标签用于表示该词所属的类别。对于许多深层次的自然语言处理任务(如语义分析、机器翻译和信息抽取等)而言,词性特征是一种非常有效的特征,其标注准确率直接影响到后续工作的质量。因此,对词性标注进行研究具有重要意义。

对中文词性标注问题的研究早已开展,已有工作分别采用了隐马尔可夫模型(Hidden Markov Model, HMM)[1-2]、最大熵模型(Maximum Entropy, ME)[3-4]和条件随机场(Conditional Random Field, CRF)[5-6]等方法。这些基于统计的方法通过建立概率图模型,设计了大量的、针对特定任务的人工特征,在词性标注上实现了较高的准确率。然而,由于受到人工特征设计的限制,这类传统的标注方法能够利用的特征十分有限,特征数据稀疏、特征不完整问题广泛存在。同时,通过特征模板提取特征需要进行大量运算,不仅消耗时间,而且容易使模型面临过拟合的风险,降低了泛化能力。

近年来,基于深度神经网络(Deep Neural Network, DNN)的方法在自然语言处理中被广泛应用。利用DNN自动提取特征,极大地缓解了传统方法的特征依赖问题。同时,将词的分布式表示(Word Distributed Embedding)作为DNN

到稿日期:2017-05-11 返修日期:2017-07-18
司念文(1992—),男,硕士,主要研究方向为智能信息处理;**王衡军**(1973—),博士,副教授,主要研究方向为神经网络、机器学习,E-mail:wanghengjun@163.com(通信作者);**李 伟**(1990—),硕士,主要研究方向为机器学习;**单义栋**(1988—),硕士,主要研究方向为智能信息处理;**谢鹏程**(1996—),主要研究方向为机器学习、神经网络。

的输入，使得所提取的特征包含了丰富的语义信息。这类DNN结合分布式词向量的方法，已经在许多任务中取得了成功，准确率和效率都优于传统方法。对于词性标注任务，为了尽可能避免针对特定任务的特征设计，文献[7]采用多隐层的神经网络自动提取特征，为词性标注、实体识别和组块分析等任务设计了统一的标注架构，极大地缓解了传统方法中的特征依赖问题，显著提升了各个任务上的标注结果。文献[8]针对中文分词和词性标注两个任务，设计了更加简洁高效的深层神经网络模型，在尽可能少地利用人工设计特征的情形下达到了较好的效果。然而，相比于传统标注模型，上述神经网络模型虽然减少了人工设计特征的工作量，但实际效果受到词窗口大小的限制，词性标注所能参考到的上下文信息非常有限。而目前的研究表明，词语的词性类别与其周围的上下文信息十分相关。基于此，文献[9]提出采用层次化的长短时记忆网络(Long Short-Term Memory，LSTM)来获取更大范围的上下文信息，将词性标注与分词任务相结合，互相提供辅助信息，从而提升了词性标注的准确率。文献[10]则提出在LSTM网络的输出层增加CRF层，并利用CRF层实现句子级别的标签推断，其结果优于传统的CRF模型和单独使用LSTM网络的模型。但是，该文献仅针对英文的词性标注、组块分析和实体识别，对中文词性标注及相关语料的实验未进行讨论。

注意力机制近来被引入到自然语言处理领域，并且在机器翻译[11]、句法分析[12]和自动文摘[13]等任务中取得了很好的应用效果。借助注意力机制为神经网络隐层单元分配不同的概率权重，使得隐层能够关注到对分类任务更加有利的特征信息，同时降低对一些冗余信息的关注。这样，在同样的上下文序列中，加入注意力机制的隐层能够进一步优化所提取特征的质量。注意力机制在句法分析中的应用很好地证明了这一点[12]，它使得句法分析模型能够学习到长距离的句法依赖关系信息。对于词性标注来说，词性作为词语的句法功能类别，其标注准确率明显受句子中上下文信息的影响，尤其对于一些长距离的、特定的句法结构信息，它能够很好地解决兼类词的标注问题[14]。将注意力机制加入到神经网络标注模型中，可以很好地获取这些特定的上下文信息，提升标注模型的准确率。

基于上述分析，为了更加准确地获取到丰富的上下文信息，提升中文词性的标注准确率，文中提出一种基于注意力长短时记忆网络的词性标注模型。该模型采用双向长短时记忆网络(Bidirectional Long Short-Term Memory，BLSTM)来获取当前词语的上下文信息，同时在BLSTM隐层引入注意力概率权重，通过分配不同的权重系数，优化标注过程中对目标词周围信息的利用，提升关注重点上下文特征的能力。为了验证该模型的词性标注效果，在PFR《人民日报》和CTB5两种语料库上分别进行了实验，结果表明，本文提出的词性标注模型在准确率上明显优于传统方法，在CTB5语料库上的性能与当前较好的标注器十分接近。

本文的主要工作分为以下两个方面：1)建立了基于BLSTM的词性标注模型，缓解了传统标注方法的特征依赖问题，实现了神经网络自动提取特征来进行词性标注的目标；2)在BLSTM中引入注意力机制，为各个时刻的隐层分配不同的注意力概率，提升了模型对长距离依赖信息和重要上下文特征的捕捉能力并进一步提升了标注准确率。

2　注意力 LSTM

2.1　LSTM 模型结构

长短时记忆网络是一种特殊的循环神经网络(Recurrent Neural Networks，RNNs)，通过引入记忆单元(Memory Cell)和门机制(Gated Mechanism)，解决了传统RNN存在的梯度消失问题，在表示序列数据中元素的上下文信息和提取长距离依赖关系上表现更好。图1(a)所示为单个LSTM单元，图1(b)为其内部结构。

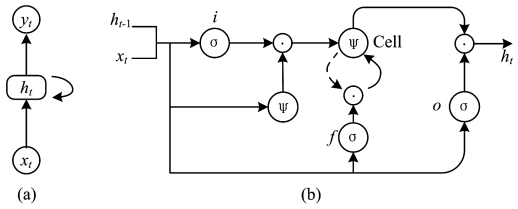

图1　LSTM单元的内部结构

Fig. 1　Internal structure of LSTM unit

LSTM单元内部共有3种门：输入门 i、遗忘门 f 和输出门 o。其中，输入门用于控制记忆单元更新的信息量；遗忘门用于控制前一时刻记忆单元信息被利用的量；输出门用于控制输出到下一隐状态的信息量。t 时刻，给定输入向量 x_t 和前一时刻的隐状态 h_{t-1}，LSTM单元通过内部循环与更新，计算出当前时刻的隐状态 h_t：

$$\begin{aligned} i_t &= \sigma(U^i x_t + W^i h_{t-1} + b^i) \\ f_t &= \sigma(U^f x_t + W^f h_{t-1} + b^f) \\ o_t &= \sigma(U^o x_t + W^o h_{t-1} + b^o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \varphi(U^c x_t + W^c h_{t-1} + b^c) \\ h_t &= o_t \odot \varphi(c_t) \end{aligned} \tag{1}$$

其中，c_t 表示记忆单元的状态信息，参数集合 $\{U^i, W^i, U^f, W^f, U^o, W^o, U^c, W^c\}$ 对应不同门的权重矩阵，$\{b^i, b^f, b^o, b^c\}$ 表示相应的偏移项，σ 和 φ 分别为 sigmoid 和 tanh 激活函数，$\odot$ 表示向量之间逐点相乘。

一般地，LSTM网络中的信息是单向传递的，LSTM只能利用过去时刻的信息，无法利用将来时刻的信息。显然，对于某些任务如分词和词性标注而言，序列的前向和后向信息都十分重要。因此，可以在LSTM网络中增加一个反向层来构成BLSTM。BLSTM由两个方向相反的LSTM层构成，其结构如图2所示。

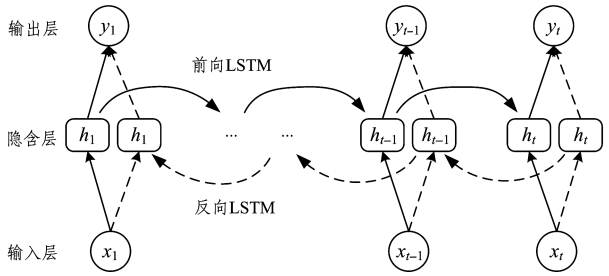

图2　展开的BLSTM网络结构

Fig. 2　Unfolded BLSTM network

图 2 中，展开的 BLSTM 网络结构分为 3 层：输入层、隐含层和输出层。x_t，h_t 和 y_t 分别表示 t 时刻的输入向量、隐状态向量和输出向量。隐含层由前向 LSTM 和反向 LSTM 构成，分别用于计算前向隐状态和反向隐状态，然后投射到共同的输出层。与单向 LSTM 相比，由于双向 LSTM 隐层信息分别沿着两个相反方向流动，能够同时获取前向和后向的历史信息，因此其在序列的特征获取与表示上效果更好，被应用到许多自然语言序列标注任务中。

2.2 注意力机制

词性作为词语的句法功能类别，其标注过程受到句子句法结构信息的影响，与有重要的句法依赖关系的词语关联更加紧密，而其他词语对当前词语的标注作用并不明显。注意力机制是一种很好的概率权重分配机制，通过计算不同时刻的注意力概率权重，使得一些与目标词的标注非常相关的节点得到更多关注，被分配到更大的概率权重，以此来提升隐含层的特征向量的质量。基本的注意力模型结构如图 3 所示。

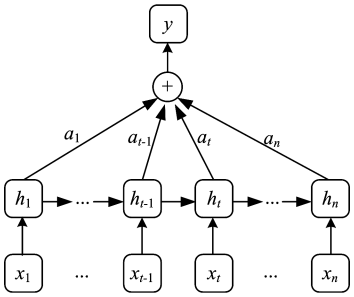

图 3 注意力模型结构

Fig.3 Structure of attention model

在加入注意力机制的神经网络模型中，新的隐状态向量 s 由各个时刻的初始隐状态向量 h_i 共同决定，计算式如下：

$$s=\sum_{i=1}^{t}\alpha_i h_i \tag{2}$$

其中，α_i 表示初始隐状态 h_i 相对于新的隐含层的权重，其计算式如下：

$$\alpha_i=\frac{\exp(e_i)}{\sum_{j=1}^{n}\exp(e_j)} \tag{3}$$

$$e_i=v\tanh(wh_i+b) \tag{4}$$

其中，e_i 表示隐状态在第 i 时刻的能量值，主要由该时刻的隐状态向量 h_i 决定。w 和 v 为权重矩阵，b 为相应的偏移值。式(2)－式(4)对应的过程实现了由初始隐含层向新的注意力层的变换，隐含层在各时刻所对应的权重系数 α_i 反映了其对当前输出的影响力大小。

3 基于注意力 LSTM 的词性标注方法

3.1 模型

本文提出的词性标注模型在 BLSTM 基础上加入了注意力机制，具体的模型结构如图 4 所示，主要包括 3 个部分：输入层、隐含层和输出层。其中，隐含层由单向 LSTM 层、双向 LSTM 层和注意力层构成，下面分别对其进行介绍。

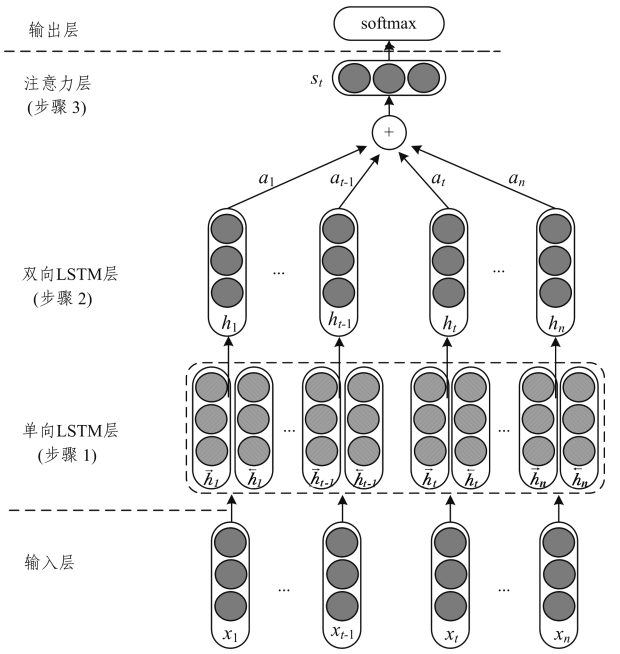

图 4 基于注意力 LSTM 的词性标注模型的结构

Fig. 4 Structure of attention-based LSTM model for POS tagging

1)输入层。传统的词向量表示方法采用独热表示(one-hot representation)，这种表示形式仅仅是将词语符号化，不包含任何语义信息；并且由于维数巨大、元素值多数为 0，因此存在严重的数据稀疏问题。分布式词向量采用低维、稠密的实数向量将词语向量化，这种表示形式包含丰富的语义信息，被广泛应用到许多自然语言处理任务中。本文采用分布式词向量表示方法，采用谷歌 word2vec 工具[1)]，通过预训练形成词向量矩阵 M，并在词向量矩阵中进行索引，将每个词语转化为其对应的词向量形式 x_t 来作为 BLSTM 网络的输入。

2)隐含层。主要分 3 个步骤进行计算。

步骤 1 根据 LSTM 模型(见 2.1 节)分别计算前向 LSTM 隐状态和反向 LSTM 隐状态。

单向 LSTM 只含一个方向的隐层，根据当前时刻的输入向量 x_t 和前一时刻的隐状态向量 h_{t-1}，计算当前时刻的隐状态 h_t。而双向 LSTM 包含前向层和反向层，需要分别计算当前时刻前向层的隐状态向量 $\overrightarrow{h}_t\in R^{m\times 1}$ 和反向隐状态向量 $\overleftarrow{h}_t\in R^{m\times 1}$：

$$\begin{aligned}\overrightarrow{h}_t&=\mathrm{LSTM}(x_t,\overrightarrow{h}_{t-1})\\ \overleftarrow{h}_t&=\mathrm{LSTM}(x_t,\overleftarrow{h}_{t-1})\end{aligned} \tag{5}$$

其中，m 为隐单元维度。LSTM()函数表示 LSTM 网络的非线性变换，其主要功能是将输入词向量编码为对应的隐状态向量，该函数的具体过程见 2.1 节。

步骤 2 根据 LSTM 前向隐状态和反向隐状态，计算 BLSTM 隐含层。

采用加权求和的方式将前向隐状态向量 $\overrightarrow{h}_t$ 和反向隐状态向量 $\overleftarrow{h}_t$ 进行线性组合，得到 BLSTM 的隐层向量 $h_t\in R^{m\times 1}$：

$$h_t=W_1\overrightarrow{h}_t+V_1\overleftarrow{h}_t+b_1 \tag{6}$$

1) https://code.google.com/p/word2vec/

其中,$W_1 \in R^{m\times m}$ 和 $V_1 \in R^{m\times m}$ 为权重矩阵,$b_1 \in R^{m\times 1}$ 为相应的偏移项。该隐含层同时聚合了输入序列中当前元素的前向和后向两个方向的序列信息,能够为词性标注提供更加丰富的上下文特征。

步骤 3　根据注意力机制(见 2.2 节),为 BLSTM 隐含层分配概率权重,并计算新的注意力隐层。

由于 BLSTM 包含前向层和反向层,因此需要同时考虑前向隐状态 $\overrightarrow{h}_t$ 和反向隐状态 $\overleftarrow{h}_t$。本文采用聚合后的隐状态向量 h_t 来计算该时刻的隐层能量 e_t:

$$e_t = V_2 \tanh(W_2 h_t + b_2) \tag{7}$$

其中,$W_2 \in R^{l\times m}$ 和 $V_2 \in R^{1\times l}$ 为权重矩阵,$b_2 \in R^{l\times 1}$ 为相应的偏移项,l 为向量 V_2 的维度。然后根据各时刻隐状态向量的能量值来计算该时刻隐状态所对应的注意力概率权重:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t=1}^{n}\exp(e_t)} \tag{8}$$

其中,α_t 为隐状态 h_t 所对应的注意力概率权重。式(7)和式(8)得到的隐状态能量值和概率权重反映了各时刻的隐状态对分类所起作用的大小,利用这种概率分配为不同的上下文特征赋予不同的重要性。最后,将各时刻的隐状态及对应的概率权重相乘并累加,得到新的注意力隐层向量 $s_t \in R^{m\times 1}$:

$$s_t = \sum_{t=1}^{n}\alpha_t h_t \tag{9}$$

式(9)中得到的新的注意力隐层与初始隐层的维度相同。由于各时刻的注意力概率分布不同,使得新的注意力隐层能够关注到初始隐层与输入序列不同的部分,因此各时刻初始隐层对词性标注所起的作用也会不同。其中,对当前词语标注影响较大的隐状态的注意力概率相应地会更大。

3)输出层。采用 $softmax$ 函数计算各时刻标注集上的标签概率分布:

$$y_t = softmax(W_3 s_t + b_3) \tag{10}$$

其中,$W_3 \in R^{L\times m}$ 表示注意力隐层和输出层之间的权重矩阵,$b_3 \in R^{L\times 1}$ 为相应的偏移项。$y_t \in R^{L\times 1}$ 表示当前时刻词语在标注集上的概率分布,如 y_t 的第 k($k=1,2,\cdots,L$)维 $y_k = p(y_t = k)$ 表示当前词语分配标注集中第 k 个词性的概率,L 为词性标注集元素的总数。

3.2　训练

给定训练集 $T=\{(x^{(i)},y^{(i)})\}_{i=1}^{N}$,其中第 i 个句子 $x^{(i)}=[x_1^{(i)},x_2^{(i)},\cdots,x_n^{(i)}]$ 对应的词性序列为 $y^{(i)}=[y_1^{(i)},y_2^{(i)},\cdots,y_n^{(i)}]$。模型训练过程中采用对数似然损失函数,训练的目标是最大化对数似然,即最小化对数似然损失函数,加上 L2 正则化项。目标函数的具体定义如下:

$$L(\theta) = -\frac{1}{k}\sum_{i=1}^{m}\log P(y^{(i)}\mid x^{(i)};\theta) + \frac{\lambda}{2}\|\theta\|^2 \tag{11}$$

其中,$P(y^{(i)}\mid x^{(i)};\theta)$ 表示 $x^{(i)}$ 对应的标注序列 $y^{(i)}$ 的得分,θ 表示模型的超参数集合。训练过程中,采用 mini-batch 梯度下降法[15],k 为每个 batch 的大小。应用 Dropout 策略[16],以一定的概率随机移除部分 BLSTM 隐层单元及其权重,以防止训练数据过拟合。

3.3　解码

BLSTM 输出各时刻词性标注的概率,按照标注集给出概率分布。一般的直接解码方法直接从输出的分类概率中进行搜索,得到最高得分的标注序列,这种方法未考虑词性标签之间的转移概率特征,将各标签看作是相互独立的,也称贪婪解码法。

本文采用与文献[9]类似的方法,即引入标签之间的状态概率转移矩阵 A,将其加入到得分函数中。对于训练集 $\{(x^{(i)},y^{(i)})\}_{i=1}^{N}$,第 i 个句子 $x^{(i)}$ 对应的词性序列为 $y^{(i)}$,则该标注序列的得分为:

$$s(x_{1:n}^{(i)},y_{1:n}^{(i)}) = \sum_{t=1}^{n}(\beta A_{y_{t-1}y_t} + (y_t)_{y_t}) \tag{12}$$

其中,$A_{y_{t-1}y_t}$ 表示从标签 y_{t-1} 到 y_t 的转移概率。假设句子 $x^{(i)}$ 对应的所有可能的标注序列集合为 $Y(x^{(i)})$,则目标是从该集合中找到最大得分序列:

$$\hat{y}^{(i)} = \arg\max_{y\in Y(x^{(i)})}(s(x^{(i)},y) + \Delta(\bar{y}^{(i)},y)) \tag{13}$$

其中,$\bar{y}^{(i)}$ 表示标准标注序列,$\hat{y}^{(i)}$ 表示实际预测的最佳标注序列,结构化损失函数 $\Delta(\bar{y}^{(i)},y)$ 的定义如下:

$$\Delta(\bar{y}^{(i)},y) = \sum_{j=1}^{n}\kappa l(\bar{y}_t^{(i)} \neq y_t) \tag{14}$$

其中,κ 为损失参数,$\Delta(\bar{y}^{(i)},y)$ 随着预测序列中错误标签数目的增加而增加。采用维特比算法对句子所对应的标注序列进行求解。

4　实验

4.1　实验设置

为了验证模型的词性标注效果,采用 PFR《人民日报》标注语料库和中文宾州树库 CTB5 分别进行实验。将 PFR《人民日报》1998 年 1 月的语料和 CTB5 语料按照表 1 划分为训练集、开发集和测试集,分别用于模型训练、参数调整和模型测试。

表 1　PFR 和 CTB5 数据集的统计情况

Table 1　Statistical results for PFR and CTB5 datasets

数据集		章节(比例)划分	句子数	词语数
PFR	训练集	70%	31327	771325
	开发集	10%	4476	109272
	测试集	20%	8975	224211
CTB5	训练集	1−270,400−931,1001−1151	10086	493930
	开发集	301−325	350	6821
	测试集	271−300	348	8008

实验评价指标采用词性标注准确率,其定义为:标注准确率 P = 正确标注的词数/待标注的词语总数。模型基于 theano 深度学习框架[17],采用 python 语言实现。实验主机为英特尔 Core i7 CPU,主频 3.33GHz,16G RAM。在每个 epoch 结束时调整神经网络模型的超参数,多个 epoch 结束后,选取实验中的最佳参数设置,如表 2 所列。

表 2　超参数设置

Table 2　Setting of hyper-parameters

参数	值
词向量维度	$d=50$
LSTM 隐层大小	$m=120$
梯度下降率(初始)	$\alpha=0.5$
L2 正则化参数	$\lambda=10^{-4}$
dropout 比率	$p=0.4$
mini-batch 大小	$k=50$

4.2 结果及分析

实验中,为了确定状态转移概率权重 β 值的最佳设置,需要在开发集上进行测试,并观察不同的 β 值对标注结果的影响。同时,为了验证模型的各个部分对词性标注性能的影响,分别进行了 3 组不同的实验,标记为 3 个模型:1)LSTM 模型;2)BLSTM 模型;3)BLSTM+ATT 模型。其中,LSTM 表示仅使用标准的单向 LSTM 进行词性标注;BLSTM 则在隐含层加入了前向 LSTM 和反向 LSTM,分别提取前向序列信息和反向序列信息;BLSTM+ATT 表示在 BLSTM 的基础上,在隐含层引入注意力机制。具体实验中,3 种模型在解码过程中均引入标签转移概率矩阵,采用维特比算法搜索最佳标注序列,在 CTB5 开发集上的测试准确率随 β 值变化的情况如图 5 所示。

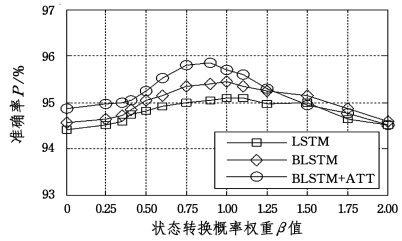

图 5 状态转移概率权重对标注结果的影响

Fig. 5 Impacts of state transition probability weight on tagging results

从图 5 可以看出,随着 β 值的增大,3 种标注模型的标注准确率整体呈现大致相同的变化趋势,但每种模型在不同的 β 值到达各自的顶点。分别选择 3 种模型的最佳测试结果进行对比,表 3 列出了各模型的最佳标注准确率,其中 β 值分别为 1.10,1.00,0.85。

表 3 基于 PFR 和 CTB5 语料的测试结果

Table 3 Experimental results on PFR and CTB5 data

模 型	状态转移概率权重 β 值	PFR /%		CTB5/%	
		开发集	测试集	开发集	测试集
LSTM	1.10	96.69	96.13	95.14	94.62
BLSTM	1.00	96.92	96.35	95.32	94.97
BLSTM+ATT	0.85	**97.33**	**96.67**	**95.85**	**95.42**

从表 3 可以看出,双向 LSTM 在标注准确率上比 LSTM 更优,因为其利用了更多的上下文信息。在加入注意力机制的 BLSTM+ATT 模型中,注意力概率分配进一步提升了 BLSTM 模型的性能,达到了 3 种模型的最佳标注准确率。同时,为了比较本文方法相对于其他模型的优势,进一步在 CTB5 语料上将其与已有工作进行了比较,如表 4 所列。从表 4 可以看出,与传统基于统计的单独词性标注模型(如文献[2]的半监督 HMM、文献[6]的 CRF 模型)相比,本文模型实现了较高的标注准确率,在单独标注模型下其最高达到 95.42%。同时,文献[18-20]在级联情形下词性标注模型的准确率也比本文方法低。其中,文献[18]实现了基于无向图的分词和词性标注一体化模型,文献[19]实现了基于转换的词性标注和句法分析联合模型,文献[20]实现了神经网络下的基于转换的词性标注句法分析联合模型。将词性标注与其他任务联合处理时,从联合模型的词性标注结果来看,分词和句法信息能够为词性标注提供辅助信息。另外,文献[19]在联合模型中加入了一些额外的特征,如非监督词聚类特征,这也在一定程度上提升了词性标注的性能。本文的词性标注模型在使用最少的特征(词向量)情形下,达到了较高的标注准确率。

表 4 本文模型与先前工作的比较

Table 4 Comparisons of our model and previous models

模型	方法	单独/联合模型	准确率/%
文献[2]模型	半监督 HMM	单独词性标注	95.48
文献[6]模型	CRF	单独词性标注	92.07
文献[18]模型	无向图	单独词性标注	94.96
	无向图	分词词性联合	95.37
文献[19]模型	基于转换	单独词性标注	94.0
	基于转换	词性句法联合	**96.0**
文献[20]模型	基于转换,LSTM	单独词性标注	95.06
	基于转换,LSTM	词性句法联合	95.58
本文模型	注意力 LSTM	单独词性标注	95.42

结束语 本文通过建立基于长短时记忆网络下的词性标注模型,解决了传统词性标注模型对人工特征依赖的问题。在网络的隐层加入注意力机制后,进一步提升了词性标注准确率。由于词语间的句法关系对词性标注有一定的辅助作用,未来我们将尝试通过联合中文词性标注和句法分析任务来改善当前串行模型的效率问题,同时提升两个任务的准确率。

参 考 文 献

[1] LIU Q,ZHANG H P,YU H K,et al. Chinese lexical analysis using cascaded hidden markov model[J]. Journal of Computer Research and Development,2004,41(8):1421-1429. (in Chinese)
刘群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展,2004,41(8):1421-1429.

[2] HAN X,HUANG D G. Research on Chinese Part-of-speech tagging based on semi hidden Markov model [J]. Journal of Chinese Computer Systems,2015,36(12):2813-2816. (in Chinese)
韩霞,黄德根. 基于半监督隐马尔科夫模型的汉语词性标注研究[J]. 小型微型计算机系统,2015,36(12):2813-2816.

[3] ZHAO Y,WANG X L,LIU B Q,et al. Fusion of clustering trigger-pair features for POS tagging based on maximum entropy model [J]. Journal of Computer Research and Development,2006,43(2):268-274. (in Chinese)
赵岩,王晓龙,刘秉权,等. 融合聚类触发对特征的最大熵词性标注模型[J]. 计算机研究与发展,2006,43(2):268-274.

[4] HE J Z,WANG H F. Chinese word sense disambiguation based on maximum entropy model with feature selection [J]. Journal of Software,2010,21(6):1287-1295. (in Chinese)
何径舟,王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报,2010,21(6):1287-1295.

[5] HONG M C,ZHANG K,TANG J,et al. A Chinese part of speech tagging approach using conditional random fields [J]. Computer Science,2006,33(10):148-151. (in Chinese)
洪铭材,张阔,唐杰,等. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学,2006,33(10):148-151.

[6] YU D J,GE Y Q,YU Z T. Chinese Part-of-speech tagging based on conditional random field [J]. Microelectronics & Computer,2011,28(10):63-66. (in Chinese)

(下转第 82 页)

结束语 本文研究了混合车辆调度问题，该问题是具有NP难的限载车辆路径问题和限载弧路径问题的综合模型。基于该问题的高复杂性，提出了一种混合进化算法。该算法采用了一种基于5种邻域结构的禁忌搜索和随机扰动来提高解的质量，设计了一个基于路径的交叉算符以尽可能保留高质量解的特性。在经典的车辆调度算例上对所提出的算法进行了测试和对比，实验结果表明该混合进化算法可以在较短的时间内得到满意的结果，尤其在4个算例上可以达到最优解。

参 考 文 献

[1] GOLDEN B, RAGHAVAN S, WASIL E. The Vehicle Routing Problem: Latest Advances and New Challenges [M]. Springer US, 2008.

[2] CORBERÁN A, PRINS C. Recent results on arc routing problems: An annotated bibliography[J]. Networks, 2010, 56(1): 50-69.

[3] PRINS C, BOUCHENOUA S. A memetic algorithm solving the vrp, the carp and general routing problems with nodes, edges and arcs[M] // Recent Advances in Memetic Algorithms. Springer Berlin Heidelberg, 2005, 166: 65-85.

[4] PANDI R, MURALIDHARAN B. A capacitated general routing problem on mixed networks [J]. Computers & Operations Research, 1995, 22: 465-478.

[5] DELL'AMICO M, HASLE G, CARLOS J, et al. An Adaptive Iterated Local Search for the Mixed Capacitated General Routing Problem [J]. Transportation Science, 2014, 50(4): 1223-1238.

[6] KOKUBUGATA H, MORIYAMA A, KAWASHIMA H. A practical solution using simulated annealing for general routing problems with nodes, edges, and arcs[M] // Engineering Stochastic Local Search Algorithms: Designing, Implementing and Analyzing Effective Heuristics. Springer Berlin Heidelberg, 2007.

[7] BRANDIO J, EGLESE R. A deterministic tabu search algorithm for the capacitated arc routing problem [J]. Computers & Operations Research, 2008, 35(4): 1112-1126.

[8] PRINS C, BOUCHENOUA S. A memetic algorithm solving the vrp, the carp and general routing problems with nodes, edges and arcs[M] // Recent Advances in Memetic Algorithms. Springer Berlin Heidelberg, 2005.

[9] BOSCO A, LAGANA D, MUSMANNO R, et al. Modeling and solving the mixed capacitated general routing problem [J]. Optimization Letters, 2013, 7(7): 1451-1469.

[10] POTVIN J Y, BENGIO S. The vehicle routing problem with time windows part II: genetic search [J]. INFORMS Journal on Computing, 1996, 8(2): 165-172.

[11] CHEN Y, HAO J K, GLOVER F. A hybrid metaheuristic approach for the capacitated arc routing problem[J]. European Journal of Operational Research, 2016, 253(1): 25-39.

[12] BACH L, HASLE G, WØHLK S. A lower bound for the node, edge, and arc routing problem [J]. Computers & Operations Research, 2013, 40(4): 943-952.

(上接第70页)

于江德，葛彦强，余正涛. 基于条件随机场的汉语词性标[J]. 微电子学与计算机，2011，28(10)：63-66.

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

[8] ZHENG X, CHEN H, XU T. Deep learning for Chinese word segmentation and POS tagging[C] // Conference on Empirical Methods in Natural Language Processing. 2013.

[9] ZHOU Q, WEN L, WANG X, et al. A Hierarchical LSTM Model for Joint Tasks[M] // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2016.

[10] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging [J]. arXiv Preprint. arXiv: 1508.01991.

[11] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C] // Proceeding of International Conference on Learning Representations. 2015.

[12] CHENG H, FANG H, HE X, et al. Bi-directional Attention with Agreement for Dependency Parsing[C] // Conference on Empirical Methods in Natural Language Processing. 2016.

[13] RUSH A M, CHOPRA S, WESTON J. A Neural Attention Model for Abstractive Sentence Summarization [C] // Conference on Empirical Methods in Natural Language Processing. 2015.

[14] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008.

[15] COTTER A, SHAMIR O, SREBRO N, et al. Better Mini-Batch Algorithms via Accelerated Gradient Methods[C] // Advances in Neural Information Processing Systems. 2011: 1647-1655.

[16] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.

[17] BASTIEN F, LAMBLIN P, PASCANU R, et al. Theano: new features and speed improvements[C] // Deep Learning and Unsupervised Feature Learning, IPS 2012 Workshop. 2012.

[18] ZHU C H, ZHAO T J, ZHENG D Q. Joint Chinese word segmentation and pos tagging system with undirected graphical models [J]. Journal of Electronics & Information Technology, 2010, 32(3): 700-704. (in Chinese)
朱聪慧，赵铁军，郑德权. 基于无向图序列标注模型的中文分词词性标注一体化系统[J]. 电子与信息学报，2010，32(3)：700-704.

[19] WANG Z, XUE N. Joint POS Tagging and Transition-based Constituent Parsing in Chinese with Non-local Features[C] // Meeting of the Association for Computational Linguistics. 2014: 733-742.

[20] YANG L, ZHANG M, LIU Y, et al. Joint POS Tagging and Dependency Parsing with Transition-based Neural Networks[J]. arXiv Preprint. arXiv: 1704.07616.