第 56 卷 第 2 期 厦门大学学报(自然科学版) Vol. 56 No. 2
2017 年 3 月 Journal of Xiamen University (Natural Science) Mar. 2017

doi:10.6043/j.issn.0438-0479.201610023

基于门循环单元神经网络的中文分词法

李雪莲,段 鸿*,许 牧
(厦门大学软件学院,福建 厦门 361005)

摘要:目前,学术界主流的中文分词法是基于字符序列标注的传统机器学习方法,该方法存在需要人工定义特征、特征稀疏等问题.随着深度学习的研究和应用的兴起,研究者提出了将长短时记忆(long short-term memory,LSTM)神经网络应用于中文分词任务的方法,该方法可以自动学习特征,并有效建模长距离依赖信息,但是该模型较为复杂,存在模型训练和预测时间长的缺陷.针对该问题,提出了基于门循环单元(gated recurrent unit,GRU)神经网络的中文分词法,该方法继承了 LSTM 模型可自动学习特征、能有效建立长距离依赖信息的优点,具有与基于 LSTM 神经网络中文分词法相当的性能,并在速度上有显著提升.

关键词:自然语言处理;中文分词;门循环单元;字嵌入;循环神经网络
中图分类号:TP 391.1 **文献标志码**:A **文章编号**:0438-0479(2017)02-0237-07

在中文中,标点符号只是对句和段进行划分,而对于词语并没有明显的分割符号,这与英文单词以空格划分存在显著差别.因此,中文自然语言处理的第一步就是将一个中文字符序列划分成词语的集合,即中文分词.中文分词是对中文进一步分析处理的基础,如词性标注、机器翻译、中文词搜索等,中文分词的效果直接影响了进一步的分析结果,因此,中文分词任务具有重要意义.然而,由于中文中存在一字多意、一词多意、不同的语境下同一个句子存在不同分词方式等问题,中文分词一直是中文自然语言处理任务中的难点.

中文分词问题从提出到现在,常用的方法可以分为以下 4 类:1) 基于字符串匹配的分词方法,又称为机械分词法[1-3];2) 基于语义分析的分词方法[4];3) 基于统计学习的分词方法[5-6];4) 基于神经网络的分词方法[7-8].基于字符串的分词存在着词典不全,对于歧义和未登录词处理效果不佳等问题.而基于语义分析的分词方法由于中文的复杂性,目前还不成熟.基于统计学习的分词则需要人工定义和提取特征,存在特征稀疏、模型复杂和容易过拟合的问题.由于神经网络可以自动学习特征,避免了传统的特征工程,近年来逐渐被应用到自然语言处理之中.2003 年,Bengio 等[9]提出了一种基于神经网络变种的概率语言模型.2011 年,Collobert 等[10]将神经网络应用到了自然语言处理中.2013 年,Zheng 等[7]开始采用神经网络解决中文分词问题.2015 年,Chen 等[8]提出了使用长短时记忆(long short-term memory,LSTM)神经网络解决中文分词问题的方法,克服了传统神经网络不能学习长距离依赖关系的问题,取得了很好的分词效果.然而,由于 LSTM 神经网络模型较为复杂,存在训练和预测时间长的问题,为解决这个问题,本文中提出了基于门循环单元(gated recurrent unit,GRU)神经网络[11]的中文分词方法.GRU 模型和 LSTM 模型均为循环神经网络(recurrent neural network,RNN)模型的扩展,但是相对于 LSTM 模型,GRU 模型将门控制单元从 3 个减少到 2 个,模型更加简单,具有更高的效率.Jozefowicz 等[12]对比了 GRU 和 LSTM 模型,发现 GRU 模型在多个问题上都能取得与 LSTM 模型相当的结果,并且更易于训练,因此,GRU 模型被越来越多地应用于自然语言处理任务中.例如,Shang 等[13]在其神经响应机的实现中采用了 GRU 模型,Kiros 等[14]也在其语言模型的实现中用 GRU 模型替换了传统的 LSTM 模型.本文中对 GRU 模型应用于中文分词任务的效果进行了验证,实验发现,基于 GRU 神

收稿日期:2016-10-30 录用日期:2017-01-15
基金项目:福建省自然科学基金(2013J01250)
* 通信作者:hduan@xmu.edu.cn
引文格式:李雪莲,段鸿,许牧.基于门循环单元神经网络的中文分词法[J].厦门大学学报(自然科学版),2017,56(2):237-243.
Citation:LI X L,DUAN H,XU M.A gated recurrent unit neural network for chinese word segmentation[J].J Xiamen Univ Nat Sci,2017,56(2):237-243.(in Chinese)

经网络模型中文分词法的分词效果与基于 LSTM 神经网络模型的方法相当,但其模型的训练和预测速度显著优于 LSTM 神经网络模型,具有更高的效率.

1 基于神经网络的中文分词法

在中文分词研究中,中文分词任务通常被看作一个字符序列标注任务,即给字符序列中的每个字符标注一个词位标签[15].目前,使用最为广泛的是四词位标签集(B,M,E,S),其中 B(Begin)标注词的开始,M(Middle)标注词的中部,E(End)标注词的结束,S(Single)则标注单字符词.通过将中文分词任务转化为一个字符序列标注任务,可以进一步将该任务看作一个分类问题,即为字符序列中的每个字符确定标签分类的问题,最后实现用神经网络解决该多分类问题.

基于神经网络中文分词法的通用框架如图 1 所示[8],并以字符序列"我们很开心"为例展示了解决分词问题的整体流程.t 时刻的输入字符 c_t="很",窗口大小 $k=5$,则输入窗口为"我"、"们"、"很"、"开"、"心" 5 个字符.第 1 步,字符序列输入到查找表中,查询其所对应的 5 个长度为 d 的字向量(d 为模型指定的字向量维度).然后将这 5 个字向量串联,形成一个长向量 $\boldsymbol{x}_t \in \mathbf{R}^{H_1}$,其中 $H_1=k\times d$,作为下一层的输入.第 2 步,将 $\boldsymbol{x}_t$ 输入到一个线性变换层得到 $\boldsymbol{z}_t$,线性变换式如式(1)所示,

$$\boldsymbol{z}_t=\boldsymbol{w}_1\times\boldsymbol{x}_t+\boldsymbol{b}_1. \tag{1}$$

其中:$\boldsymbol{w}_1\in\mathbf{R}^{H_2\times H_1}$ 是变换矩阵,H_2 为隐藏节点数;$\boldsymbol{b}_1\in\mathbf{R}^{H_2}$ 是偏置项;$\boldsymbol{z}_t\in\mathbf{R}^{H_2}$.第 3 步以 $\boldsymbol{z}_t$ 为输入,按照式(2)进行非线性变换得到 $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t=g(\boldsymbol{z}_t), \tag{2}$$

其中:g 表示非线性函数,通常采用 sigmoid 函数;$\boldsymbol{h}_t\in\mathbf{R}^{H_2}$.第 4 步对 $\boldsymbol{h}_t$ 按照式(3)进行线性变换

$$\boldsymbol{y}_t=\boldsymbol{w}_2\times\boldsymbol{h}_t+\boldsymbol{b}_2, \tag{3}$$

其中:$\boldsymbol{w}_2\in\mathbf{R}^{D\times H_2}$ 是变换矩阵,D 为词位标签个数;$\boldsymbol{b}_2\in\mathbf{R}^{D}$ 是偏置项;$\boldsymbol{y}_t\in\mathbf{R}^{D}$,$\boldsymbol{y}_t$ 中的每一个元素都代表对应词位标签的得分.通过对字符序列中的每个字符进行以上的计算,可以得到该字符序列中每个字符的标签得分矩阵.由于一个字符序列中,字符标签之间存在强依赖关系,因此,可以引入一个矩阵 $\boldsymbol{A}$ 来表示字符标签之间的转换关系,A_{ij} 表示从标签 i 转移到标签 j 的概率.第 5 步通过后向传播算法,从训练集学习得到概率矩阵 $\boldsymbol{A}$.

由于窗口大小的限制,基础的神经网络架构不能学习到窗口外的上下文信息.然而,中文语句是上下文相关的,上下文的信息对于分词结果的准确性有着重要影响.如在句子"拍卖会/上,乒乓球/拍卖/完了;商店/里,乒乓球拍/卖/完了"的分词上,如果没有词组"拍卖会上"或"商店里"的这个远距离信息,句子"乒乓球拍卖完了"就无法正确分词.而 GRU 神经网络模型则突破了窗口的限制,可以有效利用长距离的上下文信息.

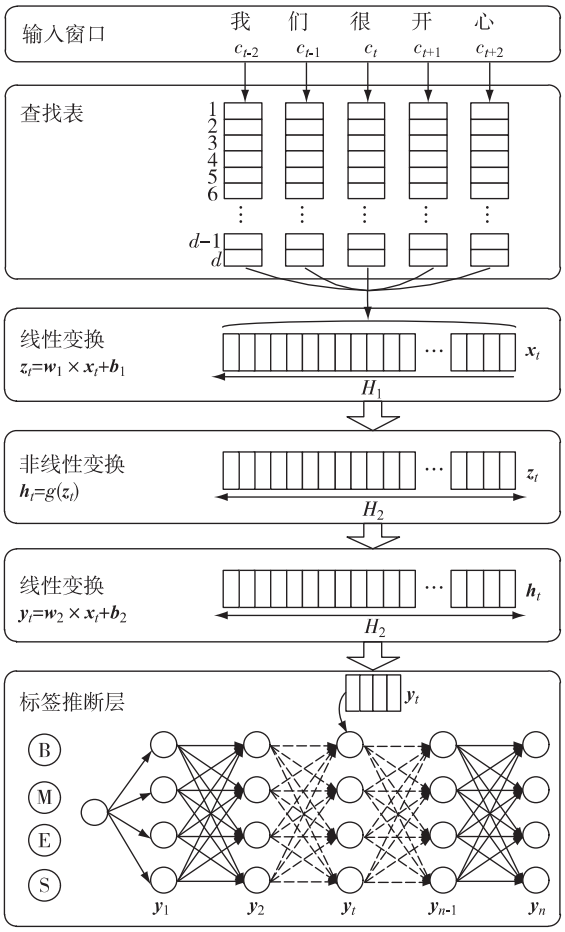

图 1 基于神经网络中文分词的通用框架

Fig. 1 General architecture of neural network for Chinese word segmentation

2 基于 GRU 神经网络的中文分词法

在这一章节中,将详细介绍基于 GRU 神经网络中文分词法的具体流程.

2.1 字符向量化

利用神经网络处理中文分词问题,第 1 步需要将字符向量化,即用一个低维的实数向量表示一个字符,该向量可以刻画字与字在语义和语法上的相关性,也作为字的特征成为神经网络的输入.这个过程通

常被称字嵌入[16-17].字符向量化构造了一个从字符到字嵌入向量的查找表,可以直接将输入的中文字符转换为字嵌入向量,作为神经网络的输入.

2.2　GRU 神经网络模型

传统的神经网络模型相邻层之间是全连接的,但是每层的各个节点是无连接的,样本的处理在各个时刻独立,使其不能对时间序列上的变化建模.然而,对于视频帧、音频以及句子中的词这样的数据,样本出现的时间顺序有着重要的意义.GRU 是 RNN 模型[11]的一种,它在隐藏层节点之间加入连接,并用一个 GRU 来控制隐藏节点的输出,可以有效建模时间序列上的变化.该单元的内部结构图如图 2 所示.

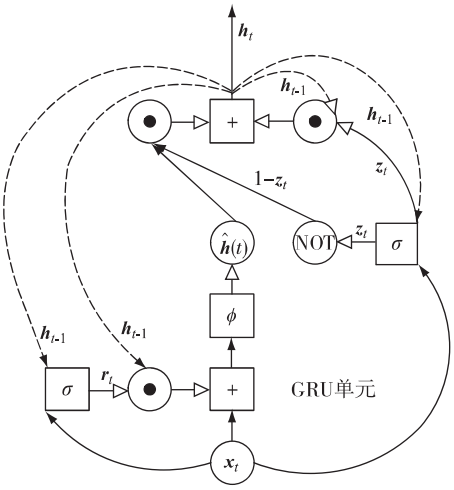

图 2　GRU 单元内部结构图

Fig. 2　GRU unit internal structure

图 2 中,虚线表示 $t-1$ 时刻的隐藏节点的激活值,实心箭头表示这条连线上有乘以一个权重.其中 $\hat{\boldsymbol{h}}_t$ 表示当前隐藏节点的候选值,$\boldsymbol{h}_t$ 表示当前隐藏节点输出的激活值.$\boldsymbol{r}_t$ 表示重置门,z_t 表示更新门.用公式表示图 2 为:

$$\boldsymbol{z}_t=\sigma(\boldsymbol{W}_{zx}\boldsymbol{x}_t+\boldsymbol{W}_{zh}\boldsymbol{h}_{t-1}),\tag{4}$$

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_{rx}\boldsymbol{x}_t+\boldsymbol{W}_{rh}\boldsymbol{h}_{t-1}),\tag{5}$$

$$\hat{\boldsymbol{h}}_t=\phi(\boldsymbol{W}_{hx}\boldsymbol{x}_t+\boldsymbol{r}_t\odot\boldsymbol{W}_{hh}\boldsymbol{h}_{t-1}),\tag{6}$$

$$\boldsymbol{h}_t=(1-\boldsymbol{z}_t)\odot\hat{\boldsymbol{h}}_t+\boldsymbol{z}_t\odot\boldsymbol{h}_{t-1}.\tag{7}$$

其中,$\boldsymbol{x}_t$ 表示当前神经网络的输入,$\boldsymbol{h}_{t-1}$ 表示上一隐藏节点输出的激活值,σ 表示 sigmoid 函数,ϕ 表示 tanh 函数,$\odot$ 表示 Hadamard 乘积,所有的 $\boldsymbol{W}$ 都是模型要训练的权重矩阵.可以看出,当重置门 $\boldsymbol{r}_t$ 接近 0 时,表示忽略之前隐藏节点的信息,只将当前时刻的输入作为输入.这个机制可以使模型丢弃一些无用信息.同理,更新门 z_t 会控制之前时刻的信息被带入到当前隐藏状态的程度,z_t 越大,之前时刻隐藏节点提供的信息越多.每个隐藏单元都会有一个独立的重置门和更新门,每个隐藏单元都会自动学习到不同时间范围的依赖关系.一般来说,学习到短距离依赖关系隐藏节点的重置门会比较活跃,而学习到长距离依赖关系隐藏节点的更新门会更活跃.

2.3　标签得分计算

前文提到,中文分词问题可以转换成字符序列中字符的标签分类问题.对于字符序列中的每个字符,基于 GRU 神经网络的中文分词模型都会给出一个它在每类标签的得分.以一个输入序列 $\boldsymbol{c}_{(1:n)}$ 为例,计算 $\boldsymbol{c}_t$ 的概率,设窗口大小为 k,字向量维度为 d,则通过查询表,可以得到一个串联的维度为 $k\times d$ 向量 $\boldsymbol{x}_t$,将其作为 GRU 神经网络的输入.可以将整个 GRU 神经网络看作是式(2)中的非线性变换 g,则通过 GRU 神经网络变换之后得到一个输出 $\boldsymbol{h}_t$,最后再经过一个线性变换,可以得到一个与标签集维度相等的向量 $\boldsymbol{y}_t$,表示 $\boldsymbol{c}_t$ 属于各个标签的得分.

2.4　标签推断

在中文句子中,相邻字的标签存在很强的依赖关系.如标签应该是分块的,某类标签不能在特定标签后出现.以前面提到的{B,M,E,S}标签集为例,B 标签之后只能是 M 或者 E,不能是 B 或者 S.然而,在字层面的模型训练并没有考虑到这个依赖关系.为充分利用这个依赖关系,文献[10]提出了引入标签转移权重矩阵 $\boldsymbol{A}$ 的方法,其中 A_{ij} 表示从标签 i 转移到标签 j 的权重.A_{ij} 的值越高,表示标签 i 转移到 j 的可能性越大.那么,对于训练数据集中的一个输入字符序列 $\boldsymbol{x}$,其标签序列为 $\boldsymbol{y}$,序列长度为 n,则将该字符标签序列的得分定义为:

$$s(\boldsymbol{x},\boldsymbol{y},\theta)=\sum_{t=1}^{n}(A_{y_{t-1}y_t}+\boldsymbol{y}_{[t,y_t]}),\tag{8}$$

其中,$\boldsymbol{y}_t$ 表示句子中第 t 个字符真实的标签值,θ 表示模型的参数集合,$\boldsymbol{y}$ 表示神经网络模型的预测结果矩阵,而 $\boldsymbol{y}_{[t,y_t]}$ 则表示神经网络模型预测第 t 个字符属于标签 $\boldsymbol{y}_t$ 的得分,即预测正确的得分.

2.5　模型训练

模型训练的关键是定义目标函数.设输入的句子为 $\boldsymbol{x}$,该句子正确的标签序列为 $\boldsymbol{y}$,用 $\boldsymbol{Y}(\boldsymbol{x})$ 表示所有 $\boldsymbol{x}$ 可能标签序列的集合.定义 $\hat{\boldsymbol{y}}\in\boldsymbol{Y}(\boldsymbol{x})$ 表示预测的标签序列,$\hat{\boldsymbol{y}}$ 的计算公式为:

$$\hat{\boldsymbol{y}}=\underset{\hat{\boldsymbol{y}}\in\boldsymbol{Y}(\boldsymbol{x})}{\operatorname{argmax}}s(\boldsymbol{x},\boldsymbol{y},\theta)\tag{9}$$

式(9)表示 $\hat{\boldsymbol{y}}$ 是 $\boldsymbol{Y}(\boldsymbol{x})$ 中得分最高的一个标签序列.接

着,可以定义结构损失函数:

$$\Delta(\boldsymbol{y},\hat{\boldsymbol{y}})=\sum_{t}^{n}\eta 1\{\boldsymbol{y}_t\neq\hat{\boldsymbol{y}}_t\},\tag{10}$$

其中,$1\{\boldsymbol{y}_t\neq\hat{\boldsymbol{y}}_t\}$表示当 $\boldsymbol{y}_t\neq\hat{\boldsymbol{y}}$ 时为 1,否则为 0,η 是调节比例的参数.$\Delta(\boldsymbol{y},\hat{\boldsymbol{y}})$则表示对于输入句子 $\boldsymbol{x}$,标签预测错误数的线性相关值.设训练集为 T,定义正则化的目标函数:

$$J(\theta)=\frac{1}{|T|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in T}l(\theta)+\frac{\lambda}{2}\|\theta\|_2^2,\tag{11}$$

其中,$l(\theta)=\max(0,s(\boldsymbol{x},\hat{\boldsymbol{y}},\theta)+\Delta(\boldsymbol{y},\hat{\boldsymbol{y}})-s(\boldsymbol{x},\boldsymbol{y},\theta))$.

最后,可以采用梯度下降法[18]优化目标函数,后向传播过程则可依据文献[19]实现.

3 实 验

3.1 实验数据

本文中采用的实验数据来自于三大语料数据库:CTB6,MSRA 和 PKU.其中,PKU 和 MSRA 来自 SIGHAN 举办的第二届国际中文分词大赛[20],而 CTB6 则来自 Chinese TreeBank 6.0 数据集.对于 CTB6 数据集,采用 CTB 6.0 文档中推荐的划分方法将数据集划分成训练集、开发集和测试集.对于 PKU 和 MSRA,则直接采用其提供的测试集,并在训练集中随机抽取了 10%作为开发集,剩下的 90%作为训练集.各个数据集中句子及字符数量统计如表 1 所示.

为了实验效果,本文参照文献[8]中的方法,对数据进行了预处理.将数据中的成语替换成^,英语单词替换成*,数字替换成 $.为了评估模型,采用精确率 P、召回率 R 和综合指标 F 值 3 个指标.最后,分别在数据训练集和数据预测集中的采用单个句子平均训练时间和预测时间来比较 GRU 和 LSTM 的速度.

3.2 模型超参

为了比较 GRU 和 LSTM 神经网络中文分词的效果和性能,对 2 个模型采用同样的参数.将最小批处理尺寸设置为 20,隐藏层节点数设置为 150,字嵌入向量的维度为 100.对于输入窗口,将窗口分为左右两边,左窗口设置为 0,右窗口设置为 2,即将 t 到 $t+2$ 的 3 个字符同时输入.为防止神经网络过拟合,在输入层将采用 dropout,并设置 dropout 的丢弃率为 0.2.

3.3 实验结果与分析

本文中实现了基于 GRU 神经网络的中文分词算法,并将其与文献[8]中基于 LSTM 神经网络模型的中文分词算法以及文献[5]中传统的基于条件随机场(CRF)的中文分词算法的结果对比,3 个算法的分词效果如表 2 所示.从表 2 可以看出,基于 GRU 神经网络的中文分词算法在 3 个数据集上的分词效果,在精确率、召回率和综合指标 F 值 3 个指标上都与基于 LSTM 神经网络的中文分词算法以及传统的基于 CRF 的中文分词算法相当.但是,传统的基于 CRF 的中文分词算法需要人工定义特征,受限于特征的选择和提取.GRU 和 LSTM 神经网络模型都可以自动学

表 1 数据集信息统计

Tab.1 The information of datasets

数据集	CTB6		MSRA		PKU	
	行数	字符数	行数	字符数	行数	字符数
训练集	23 420	1 055 583	78 231	3 650 013	17 150	1 645 048
开发集	2 079	100 316	8 693	400 456	1 906	181 400
测试集	2 796	134 149	3 985	184 355	1 944	172 733

表 2 在三大数据集上的分词效果对比

Tab.2 Word segmentation results comparison on three datasets %

模型	CTB6			MSRA			PKU		
	P	R	F	P	R	F	P	R	F
LSTM[8]	95.0	94.8	94.9	96.7	96.2	96.4	95.8	95.5	95.7
CRF[5]				96.6	96.2	96.4	95.4	94.6	95.0
GRU	94.9	94.7	94.8	96.5	96.1	96.3	95.6	95.5	95.5

习特征,而从图 3 中可以看出,GRU 神经网络在各个数据集中单个句子平均训练时间和预测时间都明显小于 LSTM 神经网络.为进一步考察 GRU 神经网络在速度上的优势,对比了不同字向量维度下基于 GRU 神经网络和基于 LSTM 神经网络的中文分词法的句子平均训练和预测时间,如图 4 所示.可以看出,在不同字向量维度下,GRU 神经网络分词法在速度上都具有显著优势.

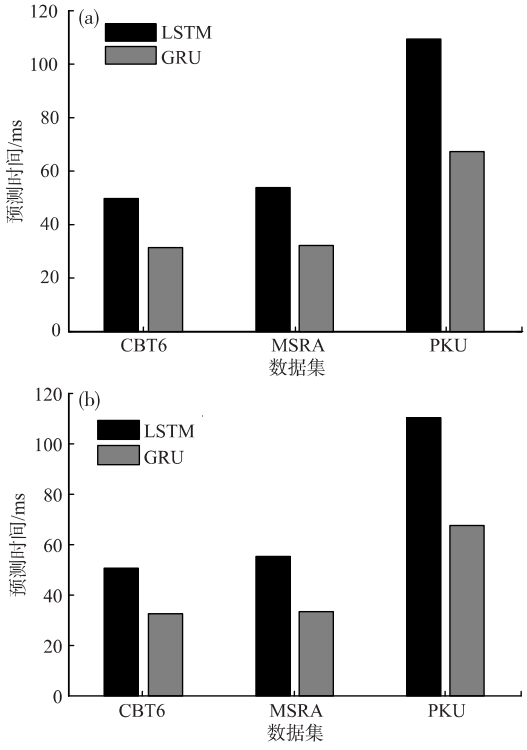

图 3　单个句子平均训练时间(a)平均预测时间(b)对比

Fig. 3　Comparison of a sentence′s average training time (a) and average predicting time(b)

具体的,当字向量维度为 100 时,在 3 个数据集上,GRU 神经网络的训练速度平均是 LSTM 神经网络训练速度的 2.02 倍(图 3(a)),测试集上的预测速度平均是 LSTM 神经网络的 1.63 倍(图 3(b)).当取不同的字向量维度时,GRU 神经网络的训练速度平均是 LSTM 神经网络的 2.25 倍(图 4(a)),而预测速度平均是 LSTM 神经网络的 1.62 倍(图 4(b)).相对于 LSTM 神经网络模型,GRU 神经网络模型用更新门替换了 LSTM 中的输入门和输出门,用一个重置门控制之前隐藏节点信息的输入,舍弃了 LSTM 中的内部记忆单元和输出门.LSTM 在计算当前隐藏节点时需要前一个隐藏节点的内部记忆单元以及输出值,而 GRU 神经网络只需要前一个隐藏节点的输出值,GRU 神经网络模型得到了简化,在每一次隐藏节点

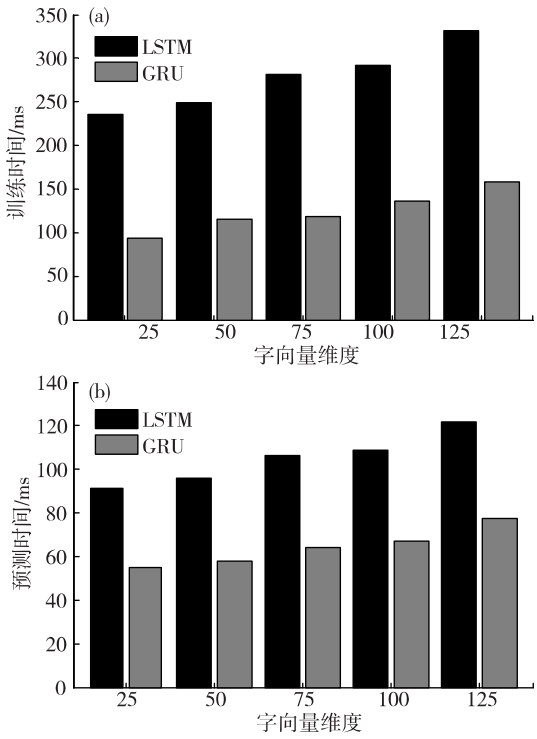

图 4　不同字向量维度的句子平均训练时间(a)平均预测时间(b)对比

Fig. 4　Comparison of a sentence′s average training time (a) and average predicting time (b) by different embedding size

的计算中,比 LSTM 少两个非线性变换的计算,因此,模型训练和预测速度更快.实验结果很好地验证了这一结论.

为了考察 GRU 神经网络获取长距离依赖信息在中文分词法上的优势,将实验结果与文献[22]中只采用了局部窗口信息的基于最大间隔张量神经网络(MMTNN)的中文分词结果对比,如表 3 所示.

表 3　GRU 模型与 MMTNN 模型的中文分词效果比较

Tab.3　Performance comparison of Chinese word segmentation between GRU and MMTNN　%

Models	MSRA			PKU		
	P	*R*	*F*	*P*	*R*	*F*
MMTNN	95.2	94.6	94.9	94.4	93.6	94.0
GRU	96.5	96.1	96.3	95.6	95.5	95.5

MMTNN 模型的主体神经网络框架与 GRU 模型类似,但是 MMTNN 模型只能获取局部窗口信息,而 GRU 模型可以获取长距离依赖信息,虽然 MMTNN 模型引入了标签向量信息和张量信息对神经网络进行优化,但从表 3 可以看出,在中文分词的

效果上该模型还是不如可以获取长距离依赖信息的 GRU 模型,可以说明,获取长距离依赖信息这一特性使得 GRU 模型在中文分词法上具有显著优势.

4 结 论

本研究的主要工作是将 GRU 神经网络应用到中文分词算法中,用 PKU、MSRA 和 CTB6 这 3 个数据库做实验,并与基于 LSTM 神经网络的中文分词算法对比.通过实验结果发现,基于 GRU 神经网络的中文分词算法在 P、R 和 F 值 3 项指标上都能取得和基于 LSTM 神经网络的中文分词算法相当的结果,但 GRU 神经网络在速度上表现出了很好的优势,训练速度比 LSTM 神经网络模型提高了 1.02 倍,预测速度提高了 0.63 倍.同时,GRU 神经网络模型对于字符序列标注问题具有一定的通用性,下一步工作会将 GRU 神经网络模型应用到藏文等类似语言的分词任务中.

参考文献:

[1] 骆正清,陈增武,胡尚序.一种改进的 MM 分词方法的算法设计[J].中文信息学报,1996,10(3):30-36.

[2] 张华平,刘群.基于 N-最短路径方法的中文词语粗分模型[J].中文信息学报,2002,16(5):1-7.

[3] 吴春颖,王士同.基于二元语法的 N-最大概率中文粗分模型[J].计算机应用,2007,27(12):2902-2905.

[4] WU A D, JIANG Z X. Word segmentation in sentence analysis[C]// Proceedings of the 1998 International Conference on Chinese Information Processing. Beijing: Tsinghua University, 1998: 169-180.

[5] TSENG H, CHANG P, ANDREW G, et al. A conditional random field word segmenter for sighan bakeoff 2005[C]// Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. Jeju: Association for Computational Linguistics, 2005: 168-171.

[6] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown: IEEE, 2001: 282-289.

[7] ZHENG X Q, CHEN H Y, XU T Y. Deep learning for chinese word segmentation and pos tagging[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013: 647-657.

[8] CHEN X C, QIU X P, ZHU C X, et al. Long short-term memory neural networks for chinese word segmentation [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 1385-1394.

[9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[10] 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.

[11] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1724-1734.

[12] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [J]. Journal of Machine Learning Research, 2015, 37(1): 2342-2350.

[13] SHANG L F, LU Z D, LI H. Neural responding machine for short-text conversion[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 1577-1586.

[14] KIROS R, ZHU Y K, SALAKHUTDINOV R, et al. Skip-thought vectors[C]// Proceedings of Advances in Neural Information Processing Systems. Quebec: MIT, 2015: 2394-3302.

[15] XUE N W. Chinese word segmentation as character tagging [J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-48.

[16] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: Association for Computational Linguistics, 2013: 746-751.

[17] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proceedings of Advances in Neural Information Processing Systems. Nevada: MIT, 2013, 3111-3119.

[18] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization

[J].Journal of Machine Learning Research,2011,12(1):2121-2159.

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

[20] EMERSON T. The second international chinese word segmentation bakeoff[C]// Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. Jeju: Association for Computational Linguistics, 2005: 123-133.

[21] PEI W Z,GE T,CHANG B B.Max-margin tensor neural network for chinese word segmentation [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Maryland: Association for Computational Linguistics, 2014: 293-303.

A Gated Recurrent Unit Neural Network for Chinese Word Segmentation

LI Xuelian,DUAN Hong*,XU Mu

(Software School of Xiamen University,Xiamen 361005,China)

Abstract: Currently, the common method for Chinese word segmentation is traditional machine learning on character-based sequence labeling. However, this method faces disadvantages such as manual feature engineering and sparse features. With the increasing research and application of deep learning, researchers have proposed a method by applying long short-term memory (LSTM) to Chinese word segmentation task. This method is capable of learning features automatically and capturing long-distance dependence as well. However, this method is complicated, and has defects in speed. Therefore, we propose a gated recurrent unit (GRU) neural network for Chinese word segmentation, which are also associated with advantages of learning features automatically and the ability of capturing long-distance dependence. Finally, our method performs comparably well as the LSTM neural network for Chinese word segmentation, and exhibits a great improvement in training and predicting speeds.

Key words: natural language processing; Chinese word segmentation; gated recurrent unit (GRU); character embedding; recurrent neural networks