第 46 卷 第 5 期 计 算 机 工 程 2020 年 5 月
Vol. 46 No. 5 Computer Engineering May 2020

·开发研究与工程应用· 文章编号：1000-3428(2020)05-0291-07 文献标志码：A 中图分类号：TP18

基于自注意力机制的中文标点符号预测模型

段大高，梁少虎，赵振东，韩忠明

（北京工商大学 计算机与信息工程学院，北京 100048）

摘 要： 中文标点符号预测是自然语言处理的一项重要任务，能够帮助人们消除歧义，更准确地理解文本。为解决传统自注意力机制模型不能处理序列位置信息的问题，提出一种基于自注意力机制的中文标点符号预测模型。在自注意力机制的基础上堆叠多层 Bi-LSTM 网络，并结合词性与语法信息进行联合学习，完成标点符号预测。自注意力机制可以捕获任意两个词的关系而不依赖距离，同时词性和语法信息能够提升预测标点符号的正确率。在真实新闻数据集上的实验结果表明，该模型 *F*1 值达到 85.63%，明显高于传统 CRF、LSTM 预测方法，可实现对中文标点符号的准确预测。

关键词： 标点符号预测；自注意力机制；Bi-LSTM 网络；深度神经网络；自然语言处理

开放科学（资源服务）标志码（OSID）：

中文引用格式： 段大高，梁少虎，赵振东，等. 基于自注意力机制的中文标点符号预测模型[J]. 计算机工程，2020，46(5)：291-297.

英文引用格式： DUAN Dagao，LIANG Shaohu，ZHAO Zhendong，et al. Prediction model of Chinese punctuation based on self-attention mechanism[J]. Computer Engineering，2020，46(5)：291-297.

Prediction Model of Chinese Punctuation Based on Self-Attention Mechanism

DUAN Dagao，LIANG Shaohu，ZHAO Zhendong，HAN Zhongming

（School of Computer and Information Engineering，Beijing Technology and Business University，Beijing 100048，China）

【Abstract】 Chinese Punctuation Prediction（PP） is an important task of Natural Language Pprocessing（NLP），which can help people eliminate ambiguity and understand the text more accurately. In order to solve the problem that the self-attention mechanism cannot process sequence position information，this paper proposes a Chinese punctuation prediction model based on the self-attention mechanism. This model stacks multiple layers of Bi-directional Long Short-Term Memory（Bi-LSTM） network on the basis of self-attention mechanism，and combines the part of speech and grammar information for joint learning to complete the punctuation prediction. The self-attention mechanism can capture the relationship between any two words without relying on their distance，and the accuracy of predicted punctuation can be improved by part of speech and grammatical information. Experimental results on real news datasets show that the *F*1 value of the proposed model reaches 85.63%，which is significantly higher than traditional CRF and LSTM prediction methods，and achieves accurate prediction of Chinese punctuation.

【Key words】 Punctuation Prediction（PP）；self-attention mechanism；Bi-LSTM network；Deep Neural Network（DNN）；Natural Language Processing（NLP）

DOI： 10.19678/j.issn.1000-3428.0054243

0 概述

汉语是当今世界上最主要的语言之一，全球有超过 15 亿人使用汉语。在汉语书面语中，标点符号是不可缺少的组成部分，是辅助文字记录语言的符号，用来表示停顿、语气以及词语的性质和作用，它可以帮助人们确切地表达思想感情和理解书面语言。在语音转化文本的过程中，没有添加标点符号，或者只是依据时间停顿来添加特定的标点符号，这样不符合标点符号规范，因此标点符号预测（Punctuation Prediction，PP）是一

基金项目： 国家自然科学基金（61170112，61532006）；教育部人文社会科学研究青年基金（13YJC860006）；北京市自然科学基金（4172016）。

作者简介： 段大高（1976—），男，副教授，主研方向为数据挖掘、自然语言处理；梁少虎、赵振东，硕士研究生；韩忠明，教授。

收稿日期： 2019-03-15 **修回日期：** 2019-04-19 **E-mail：** duandg@th.btbu.edu.cn

项重要的自然语言处理任务。标点符号预测指利用计算机对文本进行标点符号添加,使文本符合标点符号使用规范。

标点符号预测问题最初工作主要集中在语音识别领域,标准的语音识别系统识别出的序列既没有将输出正确地分割为句子,也没有标点符号。经语音识别系统识别后的文本没有标点符号和句子边界,带来许多理解上的问题,因此,在实际应用中,对长文本进行分割,添加标点符号是必要的。标点符号预测问题吸引了语音处理领域和自然语言处理领域学者的关注,研究人员利用统计模型中的局部特征进行预测,如词汇、韵律特征和隐藏事件的语言模型(HELM)[1-2]。HELM 对于输入用一个较小窗口的采样,这样由于上下文信息有限而不能达到满意的性能[3]。对于标点符号预测,研究全局的上下文信息是必要的,尤其是长期依赖关系,已有研究试图合并语法特征拓宽标点符号预测的视野[4]。

基于循环神经网络(Recurrent Neural Networks,RNN)的标点符号预测模型技术取得了一定的进展,但由于标点符号预测任务需要对输入文本结构进行信息提取,而 RNN 不适用于处理树状结构的输入。为解决该问题,本文提出一种基于自注意力机制的多层双向 LSTM 标点符号预测模型 DABLSTM。

1 相关研究

近年来,基于文本的标点符号预测逐渐获得人们的关注,文本的标点预测不能提取声学方面的信息,需要寻找其他的特征。文献[5]基于英文文本特征,在断句中将标点插入句子。文献[6]将标点符号进行分类,提出了一种自动提取表示句末标点的线索词方法。文献[7]为了解决线索词的“远程依赖”问题,使用 F-CRF 模型,同时选取词级别特征和句子级别特征。以上方法在英语文本中取得了较好的效果,但在中文文本中效果并不理想,部分原因在于中文标点符号种类更多,较为复杂多变,中文线索词的提示性并不十分准确。只利用浅层的词法特征并不能体现中文标点符号的应用特点,因此一些学者尝试更深层次的句法特征。为了在标点预测中融合更深层的句法特征,文献[8]以网络文本为测试语料,融合多种特征,训练 CRF 模型,缺点是局限于特征的设计,当特征维度大时,容易出现过拟合现象。以上基于规则、统计以及传统机器学习模型的性能依赖于特征的选择,模型的输入是基于人为设定的特征。

神经网络具有优异的建模能力,不需要人工构建特征工程,被广泛地应用于序列标注问题上。文献[9]将神经网络模型应用到自然语言处理(NLP)。由于标点符号类别有限,中文分词任务可以看作序列标注问题,因此中文分词模型也可以应用于 PP。文献[10]提出了基于神经网络的中文分词模型。文献[11]使用 LSTM 神经网络来解决中文分词任务,一定程度上解决了传统神经网络无法长期依赖信息的问题。传统的单向 LSTM 神经网络只能处理过去的上文信息,中文句子结构复杂多变,有时需要结合下文的信息才能做出判断。文献[12]提出了一种双向 LSTM-CRF 模型,并用在序列标注问题上,取得了理想的效果。文献[13]把中文分词和标点符号预测任务用双向 LSTM 网络同时训练,得到分词和标点符号预测模型。但是由于分词标注的数据集过小,难以扩展到实际文本应用中。

本文提出一种基于自注意力机制的多层双向 LSTM 标点符号预测模型 DABLSTM,模型基于自注意力机制,直接提取输入的全局依赖关系,相比 RNN,自注意力的优点是直接提取输入文本任意两个位置的关系。因此,远程关系可以通过直接的路径解决(自注意力机制中任意两个词的距离总是 1),而远程关系对于标点符号预测是有帮助的。自注意力还可以提供更灵活的方式选择、表示和综合输入信息。双向 LSTM 网络增加了模型的表征能力,DABLSTM 模型同时利用 LSTM 和自注意力机制,捕获文本序列重要信息,联合学习文本词性与语法信息,实现准确的中文标点符号预测。

2 Bi-LSTM 序列标注模型

基于 Bi-LSTM 的通用序列标注模型结构如图 1 所示,该模型由输入层、词嵌入层、Bi-LSTM 层和 Softmax 标签预测层组成。其中,输入层为文本经过分词后的词序列 $<\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_i,\cdots,\boldsymbol{w}_L>$,该序列经过词嵌入层转换成相应的 k 维词向量序列 $<\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_i,\cdots,\boldsymbol{v}_L>$,$\boldsymbol{v}_i\in\mathbb{R}^k$,词嵌入又叫词的向量表示,一般有两种策略:用预训练的词向量查找表初始化;随机初始化,在网络训练过程中随网络同步更新。词向量序列经过 Bi-LSTM 编码为包含语义信息的表示向量 $<\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_i,\cdots,\boldsymbol{z}_L>$,$\boldsymbol{z}_i\in\mathbb{R}^d$,$d$ 为 Bi-LSTM 的输出维度。表示向量经过带有 Softmax 激活函数的全连接层得到最终的分类标签 $<\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_i,\cdots,\boldsymbol{Y}_L>$。选择的标签集为标点符号标签集时,就可以进行标点符号预测。

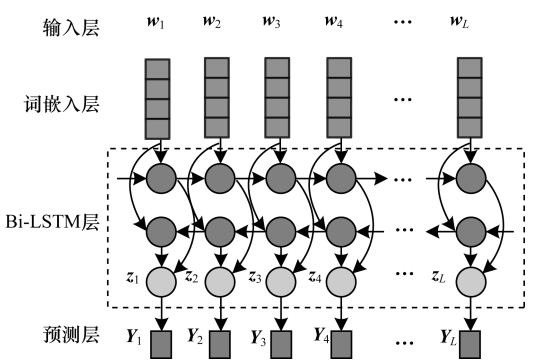

图 1 Bi-LSTM 序列标注模型结构

Fig. 1 Structure of Bi-LSTM sequence labeling model

3　DABLSTM 标点符号预测模型

本文基于 Bi-LSTM,提出 DABLSTM 模型,即深度自注意力双向 LSTM 模型,其框架如图 2 所示。

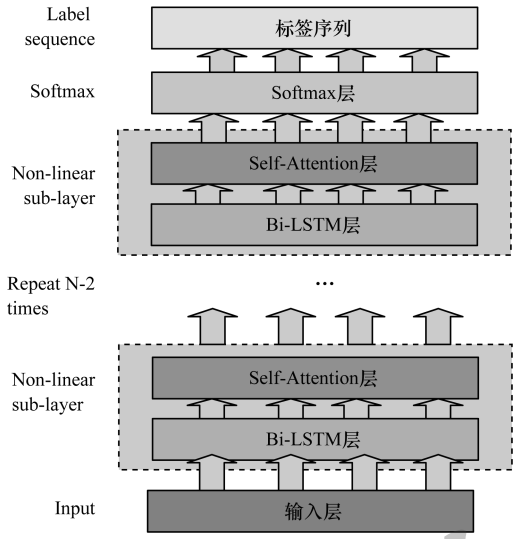

图 2　DABLSTM 神经网络模型结构

Fig. 2　Structure of DABLSTM neural network model

DABLSTM 主体堆叠 N 个相同的子层,每个子层由一个 Bi-LSTM 层和 Self-Attention 层堆叠。最上层是一个 Softmax 标签预测层构成,如图 2 所示,相比通用的 Bi-LSTM 模型,DABLSTM 模型添加了自注意力层,可以跨位置提取词之间的关系,有很强的表征句意信息。模型堆叠多次,网络更深,充分提取了句意信息来预测标点符号,实验结果表明了改进 DABLSTM 模型是有效的。

3.1　标点符号标注集

本文把标点符号标签分为 13 个类别,如果一个词的直接后继是某个类别的标点符号,则把此词标注为对应标点符号的标签,如果一个词的直接后继不是标点符号,则把此词标注为标签“0”,标点符号标签集如表 1 所示。

表 1　标点符号标签集

Table 1　Punctuation label set

标点符号	标签	标点符号	标签
,	1	“	8
。	2	”	9
?	3	《	10
!	4	》	11
;	5	$	12
、	6	无	0
:	7		

在表 1 中,标签“1”表示逗号,标签“2”表示句号,“无”表示无标点符号……“ $ ”表示段落结束标志。

3.2　自注意力层

自注意力机制是 Google 团队[14]在 2017 年提出的编码序列方案,它与一般的 RNN、CNN 同样都是一个序列编码层。自注意力机制是一种特殊的注意力机制,它只需要单独的序列来计算此序列的编码,自注意力在很多 NLP 任务中表现优异。文献[15]把自注意力机制应用在中文阅读理解任务上取得了较好的效果。文献[16]把注意力机制应用在方面级别情感分类问题上,实验结果显示了方面级别注意力机制的有效性。文献[17]提出一种融合注意力机制对评论文本深度建模的推荐方法,通过仿真实验验证了注意力机制的有效性。

本文采用文献[14]提出的多头注意力机制,图 3 所示为多头注意力机制的计算过程。

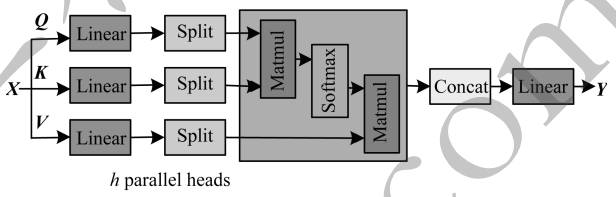

图 3　多头注意力机制计算过程

Fig. 3　Process of multiple attention mechanism

计算图的关键是缩放的点积注意力,它是点积注意力[18]的一种变体,和标准的加性注意力[19]相比,多头注意力机制添加了一层前馈神经网络,点乘机制利用矩阵乘法可以更高效地完成运算。给定 n 个查询向量组成的查询矩阵 $\boldsymbol{Q}\in\mathbb{R}^{n\times d}$、键矩阵 $\boldsymbol{K}\in\mathbb{R}^{n\times d}$ 和值矩阵 $\boldsymbol{V}\in\mathbb{R}^{n\times d}$,缩放的点乘注意力通过下式计算注意力分数:

$$\text{Self-Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \quad (1)$$

其中,d 是网络隐藏层神经元数,Softmax( ) 为非线性激活函数。

多头注意力机制首先把输入矩阵 $\boldsymbol{X}\in\mathbb{R}^{t\times d}$ 通过不同的线性变换分别映射为查询矩阵、键矩阵和值矩阵。然后将 h 个并行头聚焦于不同通道的值向量。具体地,对第 i 个头应用可学习的权重矩阵 $\boldsymbol{W}_i^{Q}\in\mathbb{R}^{d\times d/h}$、$\boldsymbol{W}_i^{k}\in\mathbb{R}^{d\times d/h}$、$\boldsymbol{W}_i^{v}\in\mathbb{R}^{d\times d/h}$ 分别线性映射为查询矩阵、键矩阵和值矩阵。之后用缩放的点乘注意力来计算查询矩阵和键矩阵的相关性得到混合表示的结果。计算公式如下:

$$\text{head}_i=\text{Self-Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}) \quad (2)$$

最后,将所有并行计算出的头拼接成一个矩阵。同样一个线性映射用来从不同的头混合不同的通道:

$$\boldsymbol{M}=\text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_n) \quad (3)$$

$$\boldsymbol{Y}=\boldsymbol{M}\boldsymbol{W} \quad (4)$$

其中,$\boldsymbol{M}\in\mathbb{R}^{n\times d}$,$\boldsymbol{W}\in\mathbb{R}^{d\times d}$。

自注意力机制和 RNN 或 CNN 相比具有很多的优势。首先,它从根本上解决了远程依赖问题,任何位置的输入和输出的距离为 1,而在 RNN 中可能是 n(输入序列长度)。与 CNN 相比,自注意力不局限于固定窗口的大小,它可以学习序列任意两个位置的关系信息。其次,自注意力机制使用加权和计算产生输出向量,它的梯度传播比 RNN 或者 CNN 更容易,不会出现 RNN 梯度弥散或膨胀问题。最后,点乘注意力可以高度并行,矩阵乘法可以在 GPU 上优化,而 RNN 由于递归机制难以并行处理。

3.3 位置编码层

自注意力机制不能处理序列位置信息,因此对输入序列向量位置信息编码是极为重要的。位置编码有很多方式可以训练得出位置向量,也可构造位置编码得出位置向量,本文使用文献[14]提出的位置编码方式:

$$\begin{cases}\mathrm{PE}_{2i}(p)=\sin(p/10\ 000^{2i/k})\\ \mathrm{PE}_{2i+1}(p)=\cos(p/10\ 000^{2i/k})\end{cases} \tag{5}$$

其中,表达式将编号为 p 位置映射为一个 k 维位置向量,向量的第 i 个元素数值为 $\mathrm{PE}_i(p)$。

3.4 Bi-LSTM 层

自注意力通过加权和得到输出向量,有着很强的表征序列能力,但不能表示序列顺序信息,循环神经网络(RNN)[20]在近些年来被广泛应用于语音识别和自然语言处理领域。文献[21]使用循环神经网络对天气等时间序列进行预测,相比传统方法预测准确度有很大提升。文献[22]提出一种基于双向 LSTM 的结构识别方法,对流式文档进行结构识别,实验结果表明双向 LSTM 具有更好的识别效果。RNN 这种基于时间序列的循环计算方式使得模型能够捕捉输入序列中长距离的特征信息,具有十分强大的信息记忆能力。由于 RNN 适用的时间序列与文本序列有着相同的特性,因此 RNN 网络相对于其他深度学习模型更适用于文本处理。在 RNN 中,每个处理单元在 t 时刻的隐藏状态均由其外部输入 x_t 和上一时刻的隐藏状态 h_{t-1} 共同决定,表示如下:

$$h_t=\begin{cases}0, t=0\\ f(h_{t-1},x_t), t>0\end{cases} \tag{6}$$

然而这种强大的记忆能力在模型的训练过程中会导致梯度消失和梯度爆炸问题,一个简单的 RNN 并不能有效处理序列的长距离依赖关系。为了解决该问题,一种 RNN 的变体长短时记忆网络(LSTM)[23]应运而生。LSTM 神经网络模型通过引入门机制和细胞状态来解决长期依赖问题。

LSTM 网络能够按需要决定模型对于历史信息所要记忆的长度,但该模型中默认序列最后元素所记忆的信息最多,因此序列最后元素占有较重要地位。而序列标注问题要求序列中每个单元都应是平等的,因此本文采用双向 LSTM(Bi-LSTM)作为基础模型。Bi-LSTM 包含了输入序列中来自双向的信息,正向 LSTM 捕获了上文的特征信息,而反向 LSTM 捕获了下文的特征信息,所以相对单向 LSTM 来说能够获取到更多的序列信息,因此在通常情况下,Bi-LSTM 的表现比单向 LSTM 或者单向 RNN 要好。

给定输入序列向量 $\{\boldsymbol{x}_i\}$,Bi-LSTM 分别以正向、反向方向处理输入序列,然后通过两个方向表示的向量之和来得到输出向量:

$$\overrightarrow{\boldsymbol{h}}_t=\mathrm{LSTM}(\boldsymbol{x}_t,\overrightarrow{\boldsymbol{h}}_{t-1}) \tag{7}$$

$$\overleftarrow{\boldsymbol{h}}_t=\mathrm{LSTM}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}}_{t+1}) \tag{8}$$

$$\boldsymbol{y}_t=\overrightarrow{\boldsymbol{h}}_t+\overleftarrow{\boldsymbol{h}}_t \tag{9}$$

3.5 输入层

本文采用三类特征作为模型输入:词向量,词性向量,句法向量。通过对词向量、词性向量、句法向量联合学习,可以得到多模态的语义句法信息,对标点符号预测有提高作用。

输入层由三部分构成:

1)词向量和位置编码,词向量是用 word2vector 工具在搜狗新闻语料库上预训练得到[24],词向量包含当前词的语义信息,本文中词向量维度设置为 300。按照上文给出的编码方案把位置信息编码为位置向量维度为 300,然后将词向量和位置向量加和作为网络部分输入。

2)词性向量,表示当前词在语义环境中的词性,可以是名词、动词、形容词与副词等,由于标点符号大多出现在名词后,因此将词性作为模型的输入特征之一。

3)句法向量,表示当前词在语境中的句法作用,包括主语、谓语、宾语等,句法信息对提升标点符号预测有巨大提升作用。将词向量和位置向量加和后的词性向量、句法向量拼接在一起作为网络的最终输入。

4 实验结果与分析

4.1 数据集

本文实验采用以下两个数据集:

1)搜狐新闻数据(SogouCS)版本,此数据集包含来自搜狐新闻 2012 年 6 月—7 月期间国内、国际、体育、社会、娱乐等 18 个频道的新闻数据,数据集大小为 648 MB。

2)当代文学作品 50 部,数据集大小为 58 MB。首先对两个数据集依据标点符号标注规则进行标

注,过滤掉在标注集之外的标点,然后以去除标点符号得到纯文本序列 X 和相对应的标签序列 Y 作为训练数据。按照 8:2 的比例把数据集划分为训练集和测试集,实验结果在测试集上得到。

在对中文标点符号预测任务中,使用分类问题的评价指标精确度(Precision)、召回率(Recall)和 $F1$ 值来评价模型整体性能,以 $F1$ 值作为主要评价指标。

4.2　实验设计

在实验中,使用分词工具 HanLP 对文本进行分词处理、DABLSTM 神经网络进行实现,使用 python3.6 和深度学习框架 TensorFlow 1.50 来构建深度网络。所用工作站参数:CPU 为 Inter Core i7 6800K,GPU 为 Nvidia GTX1080Ti 图形处理卡,操作系统为 Ubuntu16.04。

对于初始化网络参数,以均值为 0、方差为 $\frac{1}{\sqrt{b}}$ (实验中 b 为 100)的高斯分布采样来初始化网络的参数矩阵。对于词向量,采取嵌入预训练的词向量。网络超参数设置如表 2 所示,并对一些主要参数进行说明。

表 2　实验参数设置

Table 2　Experimental parameter settings

超参数	参数值	超参数	参数值
N	7	Batch_size	64.0
K	300	Drop_prob	0.2
Posi_d	300	Max_epoch	20.0
Len_LSTM	100	Head	4.0
LSTM_d	200		

将网络深度 N 设置为 7,即有 7 层重复的非线性子层,输入词向量维度 K 设置为 300,位置编码维度 Posi_d 设置为 300,LSTM 相关模型输入长度 Len_LSTM 设定置为 100,即在网络输入层有 100 个 LSTM 处理单元,LSTM 输出维度 LSTM_d 设置为 200,多头注意力 Head 设置为 4。通过在词嵌入层和自注意力层添加 Dropout[25-26] 层来使隐藏层节点不工作,增强网络的泛化能力,丢弃率 Drop_prob 设置为 0.2。采用随机梯度下降法来优化网络参数,训练的 Batch_size 设置为 64。具体地,本文采用 Adadelta[27] 作为参数优化器。

4.3　实验结果

本文为探究网络结构、自注意力机制对实验结果的影响,分别进行两个实验。

实验 1　为了验证 DABLSTM 网络模型的优越性,首先将本文模型与传统 CRF、LSTM、Bi-LSTM 模型分别在搜狗新闻数据和当代文学数据两个数据集上进行对比实验,如表 3 所示。

表 3　4 种模型的识别性能对比

Table 3　Comparison of recognition performance of four models　%

数据集	模型	精确度	召回率	$F1$
新闻	CRF	63.41	51.87	57.02
	LSTM	70.47	64.14	67.16
	Bi-LSTM	76.72	70.93	73.71
	DABLSTM	87.57	83.78	85.63
文学	CRF	62.41	50.23	55.66
	LSTM	69.18	67.18	68.17
	Bi-LSTM	73.47	70.25	71.82
	DABLSTM	85.97	82.64	84.27

本文对比实验模型介绍如下:

CRF:模型使用的特征包括线索词、句型特征、句法特征及主题词特征。

LSTM:传统的 LSTM 网络,将 PP 问题看作序列标注问题,模型的输入包括词向量和位置向量。

Bi-LSTM:双向 LSTM 网络,将 PP 问题看作序列标注问题,模型的输入包括词向量和位置向量。

DABLSTM:深度自注意力双向 LSTM 网络,有多个子层堆叠而成,每个子层包括一个双向 LSTM 层一个自注意力层,模型的输入包括词向量和位置向量。

实验 2　为了验证自注意力机制在解决序列长期依赖的有效性,在 DABLSTM 网络中用前馈网络层代替自注意力层,前馈网络层定义为:

$$\mathrm{FFN}(\boldsymbol{X}) = f(\boldsymbol{X}\boldsymbol{W}_1)\boldsymbol{W}_2 \tag{10}$$

其中,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 为可学习的参数矩阵,$f(\,)$ 为非线性激活函数,本文选用 RELU。

为研究位置向量编码对自注意力机制影响,本文做了对比实验,在网络结构超参数都相同的情况下,一组输入为词向量与位置向量的加和,另一组输入仅为词向量。为确定网络深度是否能够提高评价指标,设置了不同的网络深度 N,为探究隐藏层神经元个数即网络宽度对评价指标的影响,进行了对照实验。以上实验都是在新闻数据集上,用搜狗新闻语料库训练的 300 维词向量进行的对比实验。不同参数对比结果如表 4 所示。

表 4　不同参数对比结果

Table 4　Comparison results of different parameters

层数	非线性层	位置编码	网络深度	网络宽度	$F1$/%
1	自注意力	√	7	200	84.48
2	自注意力	√	3	200	74.23
3	自注意力	√	6	200	81.40
4	自注意力	√	9	200	82.25
5	自注意力	√	12	200	83.45
6	自注意力	√	7	400	84.53
7	自注意力	√	7	600	84.53
8	前馈网络	√	7	200	73.31
9	自注意力	×	7	200	80.12
10	无	√	7	200	70.28

在表 4 中,√表示自注意力层可以捕获序列的位置信息,×表示自注意力层不能捕获序列的位置信息。

4.4 结果分析

通过表 3 可以看出,在两个不同的数据集上,本文的 DABLSTM 网络模型在精确度、召回率和 *F*1 值上均是最优的,尤其是召回率和 *F*1 值,这是由于深度神经网络能够编码更多的潜在语义信息。纵观深度模型,Bi-LSTM 在两个数据集上的 *F*1 值分别高于 LSTM 模型 4.73% 和 4.87%,这是因为在 LSTM 网络中添加双向连接能够使模型学习到句中双向的语义信息,从而给模型带来更佳的预测结果。

1)网络深度的影响:通过表 4 的第 1 行~第 5 行可以看出,随着网络深度的增加,模型的效果增强显著,这也验证了文献[28-29]的工作,即深度是网络表征能力的关键。从第 3 层增加到第 6 层,效果提升 9.6% 比较显著,从第 6 层增加到第 9 层效果有较小提升,而在增加到一定层数后,再增加层数反而效果下降,原因是当模型太深时反向传播算法梯度难以传导,使用随机梯度优化参数的方法效果就要打折扣,另一部分原因是网络参数量大,出现了过拟合现象。对于深度网络难以训练问题,可以采用添加残差模块[29],跨层传导梯度来解决,这也是模型改进的方向,对于网络过拟合问题,可以在损失函数中引入参数正则化,这也是模型另一个改进的方向。对比第 1、6、7 行可以看出,网络宽度从 200 增加到 400、600,网络提升并不显著。因此,网络的深度比宽度对模型的提升更有效。

2)自注意力的影响:通过表 4 的第 1、8 行可以看出,当自注意力层替换为前馈网络层时,模型 *F*1 相对下降 13.2%,而当去掉自注意力层时,模型 *F*1 相对下降 16%,这一结果充分证明了自注意力层的必要性和有效性,它是对 Bi-LSTM 的必要补充,从根本上解决了序列的长期依赖问题(任意两个位置距离为 1)。

3)位置编码的影响:通过表 4 的第 1 行和 9 行看出,当输入没有编码位置信息时,模型 *F*1 值相对下降5.1%,可以看到位置信息对模型的重要性,不难理解,自注意力层不能捕获序列的位置信息,所以在词向量中融合位置信息会有较大提升,位置编码总是伴随自注意力。

5 结束语

本文建立一种基于自注意力机制的中文标点符号预测模型 DABLSTM。通过自注意力机制构建 DABLSTM 网络,提取文本添加标点的信息,根据对词的词性信息和句法信息进行联合学习,利用词的词法信息和句法信息预测标点符号。实验结果表明,对比传统的 CRF 模型和 Bi-LSTM 模型,DABLSTM 模型在中文标点符号预测上取得了较好的效果,大幅提升了预测正确率。DABLSTM 网络是 Bi-LSTM 网络的增强,对序列标注问题有较好的建模能力,下一步将对该模型进行泛化能力验证,以避免出现过拟合现象。

参考文献

[1] LIU Y,STOLCKE A,SHRIBERG E,et al. Using conditional random fields for sentence boundary detection in speech[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.[S.l.]:Association for Computational Linguistics,2005:451-458.

[2] STOLCKE A,SHRIBERG E. Automatic linguistic segmentation of conversational speech [C]//Proceedings of the 4th International Conference on Spoken Language. Washington D.C., USA:IEEE Press,1996:1005-1008.

[3] FAVRE B,HAKKANI-TUR D,PETROV S,et al. Efficient sentence segmentation using syntactic features [C]// Proceedings of Spoken Language Technology Workshop. Washington D.C.,USA:IEEE Press,2008:77-80.

[4] ROARK B,LIU Y,HARPER M,et al. Reranking for sentence boundary detection in conversational speech [C]// Proceedings of 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Washington D.C.,USA:IEEE Press,2006:1-10.

[5] TOMANEK K,WERMTER J,HAHN U. Sentence and token splitting based on conditional random fields [EB/OL]. [2019-03-20]. https://www.researchgate.net/publication.

[6] GUO Y,WANG H,GENABITH J V. A linguistically inspired statistical model for Chinese punctuation generation[J]. ACM Transactions on Asian Language Information Processing,2010,9(2):1-27.

[7] LU W,NG H T. Better punctuation prediction with dynamic conditional random fields [C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.[S.l.]:Association for Computational Linguistics, 2010:177-186.

[8] HUANG J, ZWEIG G. Maximum entropy model for punctuation annotation from speech[C]//Proceedings of the 7th International Conference on Spoken Language Processing. Washington D.C.,USA:IEEE Press,2002:1-9.

[9] COLOLBERT R, WESTON J, KARLEN M, et al. Natural language processing from scratch[J]. Journal of Machine Learning Research,2011,12(1):2493-2537.

[10] ZHANG D,WU S,YANG N,et al. Punctuation prediction with transition-based parsing[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Washington D. C., USA: IEEE Press, 2013: 752-760.

[11] CHEN Xinchi, QIU Xipeng, ZHU Chenxi, et al. Long short-term memory neural networks for Chinese word segmentation[C]//Proceedings of International Conference on Empirical Methods in Natural Language Processing. Washington D.C.,USA:IEEE Press,2015:1197-1206.

[12] HUANG Zhiheng, XU Wei, YU Kai. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2019-03-20]. https//www. arXiv preprint arXiv: 1508.01991.

[13] LI Yakun,PAN Qing. Joint Chinese word segmentation and punctuation prediction based on improved multilayer BLSTM network[J]. Journal of Computer Applications, 2018,38(5):1278-1282,1314.(in Chinese)

李雅昆，潘晴. Everett X. WANG. 基于改进的多层 BLSTM 的中文分词和标点预测[J]. 计算机应用，2018，38(5)：1278-1282，1314.

[14] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [EB/OL]. [2019-03-20]. https://www.douban.com/note/732780452/.

[15] ZHANG Haoyu，ZHANG Pengfei，LI Zhenzhen，et al. Self-attention based machine reading comprehension [J]. Journal of Chinese Information Processing，2018，32(12)：125-131. (in Chinese)
张浩宇，张鹏飞，李真真，等. 基于自注意力机制的阅读理解模型[J]. 中文信息学报，2018，32(12)：125-131.

[16] ZEN Yifu，LAN Tian，WU Zufeng，et al. Bi-memory based attention model for aspect level sentiment classification [J]. Chinese Journal of Computers，2019，42(8)：1845-1857. (in Chinese)
曾义夫，蓝天，吴祖峰，等. 基于双记忆注意力的方面级别情感分类模型[J]. 计算机学报，2019，42(8)：1845-1857.

[17] HUANG Wenming，WEI Wancheng，ZHANG Jian，et al. Recommendation method based on attention mechanism and review text deep model [J]. Computer Engineering，2019，45(9)：176-182. (in Chinese)
黄文明，卫万成，张健，等. 融合注意力机制对评论文本深度建模的推荐方法[J]. 计算机工程，2019，45(9)：176-182.

[18] LUONG M T，PHAM H，MANNING C D. Effective approaches to attention-based neural machine translation [EB/OL]. [2019-03-20]. https://arxiv.org/abs/1508.04025.

[19] CHO K，VAN MERRIENBOER B，BAHDANAU D，et al. On the properties of neural machine translation：Encoder-decoder approaches [EB/OL]. [2019-03-20]. https://arxiv.org/abs/1409.1259.

[20] GRAVES A，MOHAMED A，HINTON G. Speech recognition with deep recurrent neural networks [C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2013：6645-6649.

[21] WANG Huijian，LIU Zheng，LI Yun，et al. Trend Prediction Method of Time Series Trends Based on Neural Network Language Model [J]. Computer Engineering，2019，45(7)：13-19，25. (in Chinese)
王慧健，刘峥，李云，等. 基于神经网络语言模型的时间序列趋势预测[J]. 计算机工程，2019，45(7)：13-19，25.

[22] ZHANG Zhen，LI Ning，TIAN Yingai. Stream document structure recognition based on bidirectional LSTM network [J]. Computer Engineering，2020，46(1)：60-66，73. (in Chinese)
张真，李宁，田英爱. 基于双向 LSTM 网络的流式文档结构识别[J]. 计算机工程，2020，46(1)：60-66，73.

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

[24] LI Shen，ZHAO Zhe，HU Renfen，et al. Analogical reasoning on Chinese morphological and semantic relations [EB/OL]. [2019-03-20]. https://arxiv.org/abs/1805.06504? context = cs.

[25] SRIVASTAVA N，HINTON G，KRIZHEVSKY A，et al. Dropout：a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research，2014，15(1)：1929-1958.

[26] HINTON G E，SRIVASTAVA N，KRIZHEVSKY A，et al. Improving neural networks by preventing co-adaptation of feature detectors [EB/OL]. [2019-03-20]. https://arxiv.org/abs/1207.0580v1.

[27] ZEILER M D. ADADELTA：an adaptive learning rate method [EB/OL]. [2019-03-20]. https://arxiv.org/abs/1212.5701.

[28] ZHOU Jie，XU Wei. End-to-end learning of semantic role labeling using recurrent neural networks [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Washington D. C.，USA：IEEE Press，2015：1127-1137.

[29] HE Kaiming，ZHANG Xiangyu，REN Shaoqiang，et al. Deep residual learning for image recognition [C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：[s. n.]，2016：770-778.

编辑　索书志