

文章编号: 1003-0077(2019)06-0100-08

基于 Self-Attention 和 Bi-LSTM 的中文短文本情感分析

吴小华, 陈莉, 魏甜甜, 范婷婷

(西北大学 信息科学与技术学院, 陕西 西安 710127)

摘要: 短文本情感分析用于判断文本的情感极性, 在商品评论、舆情监控等领域有重要应用。由于目前主流的基于词注意力机制的双向循环神经网络模型性能很大程度上依赖于分词的准确性, 且注意力机制需较多的参数依赖, 无法使模型更多的关注短文本的内部序列关系。针对上述问题, 该文提出了基于字向量表示方法并结合 Self-attention 和 BiLSTM 的中文短文本情感分析算法。首先, 对短文本进行字向量化表示, 采用 BiLSTM 网络提取文本上下文关系特征, 通过自注意力机制动态调整特征权重, Softmax 分类器得到情感类别。在 COAE 2014 微博数据集和酒店评论数据集的实验结果表明, 采用字向量文本表示方法较词向量更适合短文本, 自注意力机制可以减少外部参数依赖, 使模型能学到更多的文本自身关键特征, 分类性能可分别提高 1.15% 和 1.41%。

关键词: 情感分析; 字向量; 自注意力机制; 双向长短时记忆网络

中图分类号: TP391 **文献标识码:** A

Sentiment Analysis of Chinese Short Text Based on Self-Attention and Bi-LSTM

WU Xiaohua, CHEN Li, WEI Tiantian, FAN Tingting

(School of Information Science and Technology, Northwest University, Xi'an, Shaanxi 710127, China)

Abstract: Short text sentiment analysis is a better method for judging the emotions of texts. It also has important applications in the fields of commodity reviews and public opinion monitor. The performance of the bidirectional recurrent neural network model based on the word attention mechanism relies heavily on the accuracy of word segmentation. In addition, the attention mechanism has more parameter dependencies, making the model less concerned with the internal sequence relationships of short texts. Aiming at the above problems, this paper proposes a Chinese short text sentiment analysis algorithm based on the character vector representation method combined with Self-attention and BiLSTM. Firstly, the short text is vectrized, then the BiLSTM network is used to extract texts context feature. Finally, the feature weights are dynamically adjusted by the self-attention mechanism, and the Softmax classifier obtains the emotion category. Experimental results on the COAE 2014 Weibo dataset and hotel review datasets show that character vectors are more suitable for short text than word-level text vector representations. The self-attention mechanism can reduce the external parameter dependence, so that the model can learn more key features of the text itself. Classification performance can be increased by 1.15% and 1.41%, respectively.

Keywords: sentiment analysis; character vector; self-attention; BiLSTM

0 引言

随着自媒体及社交平台的发展, 社会焦点问题和突发事件讨论、电商商品评价等信息在网上广泛传播, 产生大量短文本信息。例如, 微博头条、微信留言、时事新闻的用户评论、电商买家评论等。对短

文本进行情感倾向性分析, 并从中抽取有价值的信息, 一直以来受到工业界和学术界的普遍关注^[1]。

文本情感分析是指对带有情感色彩的主观性文本信息进行分析、处理、归纳总结并判断其情感倾向^[2]。短文本情感分析主要针对微博、电商买家评价、时事新闻评论等主观性短文本进行情感分类。1990 年 Jeffrey L Elman 提出循环神经网络 (recur-

收稿日期: 2018-07-17 定稿日期: 2018-10-10

rent neural networks, RNN) 模型, 这种网络结构被证明在处理时间序列数据方面更加有效^[3]。1997 年 Hochreiter 等提出并实现了长短时记忆网络 (long short-term memory, LSTM), 通过加入“门”的机制控制信息的传递改善了标准 RNN 的梯度消失或爆炸问题^[4]。近年来, one-hot^[5]、LDA^[6]、CBOW 及 Skip-gram 神经语言模型^[7]、Word2Vec^[8] 等词向量化表示方法的提出, 为神经网络等深度学习模型处理自然语言奠定了基础, 此后大量学者开始将深度学习应用于自然语言处理领域。但上述方法都是基于词的映射方式。由于中文词语之间无明显分割符, 词向量映射方法的性能很大程度上依赖于分词的准确性。例如, “他说的确实在理”, 有“他/说/的/确实/在理”和“他/说/的确/实在/理”等多种不同切分方式, 从而会产生各种类型的语义歧义问题。针对中文无天然分隔符问题, 2015 年, 许云等提出一种不需分词的 n 元语法文本分类方法, 有效降低了数据稀疏带来的影响, 提供了在深度学习模型中短文本预处理更高效的方案^[9]。刘龙飞等提出基于字向量卷积神经网络的微博情感倾向性分析算法。实验证明, 使用字级别的文本向量作为原始特征会优于使用词向量作为原始特征^[10]。

上述方法虽在一般文本分类中取得了良好效果, 但传统的深度学习模型将所有特征赋予相同的权重进行训练, 无法较好地关注稀疏特征短文本中对情感类别贡献较为关键的特征。2014 年, Google Mind 团队首次提出 Attention 内容注意力机制进行图像识别。它可以聚焦于图像中局部较为重要的区域, 能充分提取图像中区别他类的关键特征, 有效提高, 图像识别精度^[11]。2016 年, Bahdanau 等将注意力机制应用到自然语言处理 (natural language processing, NLP) 领域进行机器翻译, 在翻译准确率上较传统神经网络模型有较大提高^[12]。2017 年, 谷歌提出一种自注意力机制 (self-attention mechanism) 用于机器翻译取得了比普通神经网络模型更好的性能。这种自注意力机制所需依赖参数少, 使模型能够更好地学习文本特征。但由于缺乏足够位置信息, 翻译结果中仍会出现某些词序错误现象^[13]。Pavlopoulos 等提出深层注意力机制, 将其应用到审核用户评论, 取得了比 RNN 更好的效果, 证明了注意力机制在除机器翻译外的其他方面的有效性^[14]。2018 年 Zhao 等提出基于 LSTM-Attention 模型的文本分类方法, 使用分层注意力来分别选择重要的词语及句子, 依据较为关键的词语及句

子进行分类, 模型鲁棒性得到进一步提高, 但是这种将词语和句子分开选择的做法忽略了语境信息^[15]。2018 年, 江伟等从短语特征出发, 提出了其于注意力机制的双向循环神经网络的短文本表示模型, 在斯坦福情感树库数据集的 5 分类任务上取得了目前最好的准确率 53.35%。但双向循环神经网络无法保留长距离依赖信息, 且使用标准的 Attention 模型需要人工初始化较多参数, 泛化能力较弱^[16]。Yao 等提出基于 Attention 的 BiLSTM 神经网络, 定性捕捉其中有用的语义信息。从句子的两个方向学习语义特征对短文本进行情感分析, 在斯坦福树图数据集与电影评论数据集上实验, 证明该模型较无注意力模型的 LSTM 性能提高 3%^[17]。上述文献中结合注意力机制和相关优化的循环神经网络模型在一定程度上提高了分类的准确性, 改善了循环神经网络模型对词语平均加权表示所造成的信息损失, 以及难以在小规模数据集上训练的问题。由于短文本为片段性的描述说明、观点评论或情感抒发, 区别于普通文本具有较鲜明的特性^[18]: ①短文本长度短小, 包含内容较少, 具有较强的数据稀疏性, 造成短文本词向量表示不准确; ②以微博头条、用户评论为代表的短文本语言偏向口语化, 非规范化。同时, 还可能错别字, 缩写等噪声部分。若使用普通加权的特征表示方法, 噪声会影响模型性能; ③微博头条、用户评论等短文本信息通常会夹杂着网络流行语等能显著表明情感倾向的成份。一般的分词工具面对日益繁多的网络新词, 分词性能较差。这些特性造成了一般文本情感分析模型在短文本领域性能较弱, 故提高短文本情感识别精度, 从中挖掘出更准确、更有价值的信息是非常有必要的。

本文针对传统词向量性能依赖分词的准确性及普通序列模型无法关注短文本稀疏特征中最关键特征的问题, 在传统双向长短时记忆网络 (bi-directional long short-term memory, BiLSTM) 基础上引入了自注意力机制, 提出一种基于字向量表示的 Self-Attention 和 BiLSTM 的中文短文本情感分析方法。注意力机制模仿人的视觉机理聚焦于句子中对类别区分度更为关键的特征, 并动态调整特征权重, 可以建模特征的依赖关系而不用考虑其在输入或输出序列中的距离^[19]。而 Self-Attention 仅需关联单个序列的不同位置以计算序列的表示。输入和输出序列中。任意位置之间的路径更短, 更容易获取文本内部依赖关系, 在短文本领域表现出更广阔

的前景,本文实验结果也表明了该方法的有效性和可行性。

1 相关工作

1.1 字向量表示方法

由于神经网络模型只能接受数值输入,所以就需要将字符文本进行数值化表示。文本数值化表示方法一般有向量空间方法、主题模型方法及 Google 提出的 Word2Vec 等,这些文本表示方法都需要对文本进行分词处理后将词语映射到多维空间中。由于中文语言的灵活性、复杂性等特点,某些句子人工断句困难,导致基于分词的向量化方法有其局限性。故本文在短文本处理过程中使用字向量表示方法,常用汉字数量较词语数量低几个数量级,也可以在一定程度上降低矩阵维度及稀疏性。同一文本其词向量和字向量表示结果区别较大。例如,“我的研究方向是短文本情感分析”。这句话如果表示为词向量形式为“我/的/研究/方向/是/短/文本/情感/分析”,而字向量表示形式为:“我/的/研/究/方/向/是/短/文/本/情/感/分/析”。

1.2 LSTM 模型

LSTM 是 1997 年 Hochreiter 等针对 RNN 的梯度消失和梯度爆炸问题而提出的改进模型^[4]。它在原始 RNN 模型之上加入“门”来控制信息的传递,可在一定程度上避免梯度消失与爆炸问题,获取文本语义的长距离依赖信息。模型内部主要包括输入门 i_t 、遗忘门 f_t 、输出门 o_t 和记忆单元 C_t 等部分,具体结构如图 1 所示。

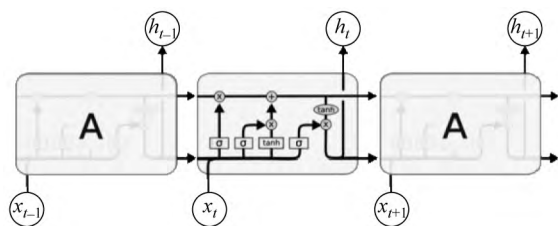


图 1 LSTM-Cell 内部结构图

首先,LSTM 必须通过“遗忘门”决定将上一个 Cell 单元中哪些信息抛弃。它由 sigmoid 完成,通过接收上一时刻输出与本时刻输入的加权和计算出一个 0 到 1 的数,0 表示完全抛弃,1 表示全部保留。其计算如式(1)所示。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

输入门控制 Cell 单元需要加入哪些信息,其计算过程如式(2)、式(3)所示。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tanh(W_f \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

输出门控制哪些信息用于此时刻的任务输出,其计算过程如式(4)、式(5)所示。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (5)$$

其中, W_i, W_f, W_o 分别为输入门、遗忘门、输出门的权重矩阵; b_i, b_f, b_o 为输入门、遗忘门、输出门的偏置矩阵; σ, \tanh 为激活函数。

1.3 Attention 模型

注意力机制来源于人的视觉处理过程,通过浏览全部信息来获得视觉注意力焦点,提取句子所表达信息的主要部分来获取句子中对于当前任务的关键信息。

Attention 函数本质上是一个由诸多 Query 和 Key-value 组成的映射函数。其计算过程主要分为以下三步:

(1) 将 query 和每个 key 进行相似度计算得到相应权重,如式(6)所示。

$$f(Q, K) = QK^T \quad (6)$$

(2) 使用 softmax 函数对权重进行归一化,如式(7)所示。

$$a_i = \text{softmax}(f(Q, K)) \quad (7)$$

(3) 将权重和相应的键值 value 进行加权求和得到最后的 Attention 值,如式(8)所示。

$$\text{Attention}(Q, K, V) = \sum a_i V \quad (8)$$

其中, Q 代表查询, $K-V$ 是文本向量键值对。

2 基于 Self-Attention 机制的 BiLSTM 模型

为了关注短文本中对于分类任务更关键的信息,本文提出基于字向量表示方法结合 Self-Attention 机制和 BiLSTM 的短文本情感分析算法(character-SATT-BiLSTM)。利用改进后的算法能够更完整保留稀疏特征的短文本的信息,以便更好地提取短文本中对分类任务更重要的特征。基于 Self-Attention 机制的 BiLSTM 算法包含字向量输入层、双向长短期记忆网络层、Self-Attention 层、Softmax 层。其模型图如图 2 所示。

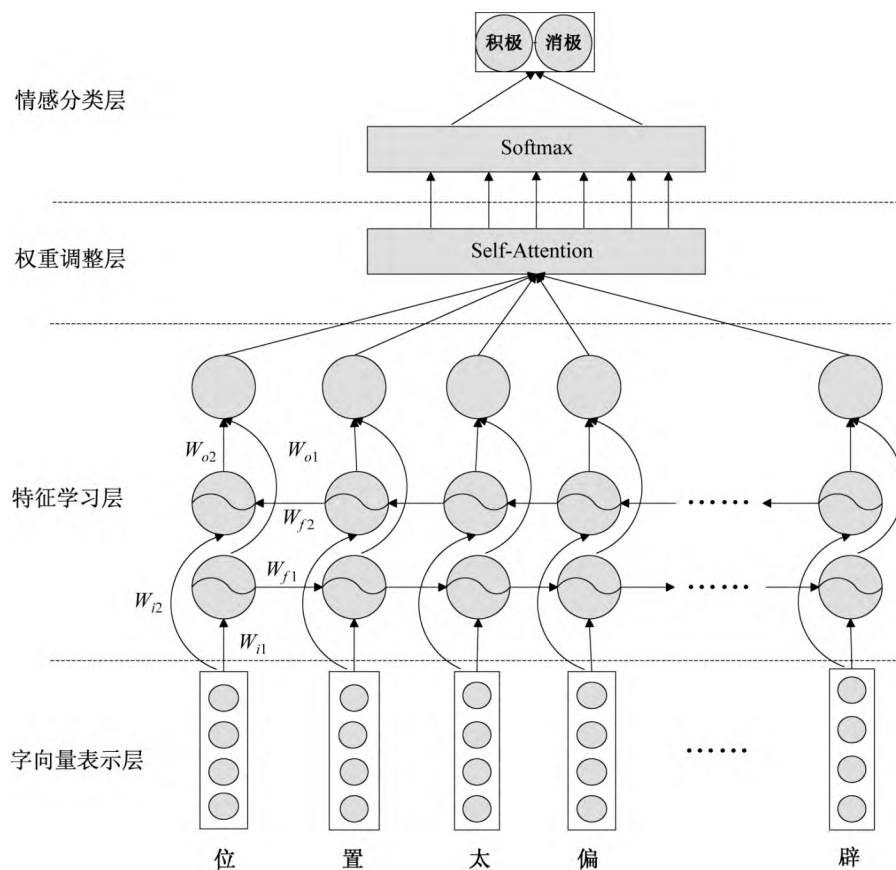


图2 基于 Self-Attention 机制的 BLSTM 模型图

2.1 任务定义

对于短文本 $t = \{w_1, w_2, \dots, w_i\}$, 其中 w_i 为文本中第 i 个字, 短文本情感分析就是对短文本 t 进行特征提取分析, 判断其所属的情感极性。考虑到一般情况下中性情感意义不大, 故本文只将情感划分为两个极性, 即: 正向情感 (Positive) 和负向情感 (Negative)。例如, “什么破酒店呀, 位置太偏僻了!”, 利用模型判断此评论情感为正向或负向。

2.2 文本向量化表示层

传统的词袋模型通常是基于统计学的文本表示

$$\arg\max_{\theta} \sum_{w, c \in D} \log P(c/w) = \sum_{w, c \in D} (\log e^{v_c \cdot v_w} - \log \sum_{c'} e^{v_c \cdot v_{c'}}) \quad (9)$$

其中, θ 为模型参数集合; w 表示字向量; c 表示 w 的语境, 即以 w 为中心, 前后 k 个字所组成的集合; $P(c/w)$ 表示字 w 出现时某一语境 c 出现的概率; D 表示所有字和它的语境 c 组成的集合; v_m 和 v_c 表示 w 和 c 的列向量; 算法训练的目的在于寻找最优参数集合 θ 。

模型, 将句子仅仅看作是词语的集合, 用词语的频率和概率进行统计, 忽略了句子原来的结构信息。且由于短文本的特点导致此类特征均不突出, 甚至会出现不可用的情况。主题模型较适用于长文本, 对包含词语较少的短文本效果较差。Word2Vec 虽已经取得了较好的效果, 但这种方法对分词准确度要求较高同时, 需要大量训练语料进行长时间训练。故本文考虑到字特征粒度更小的优点, 采用字向量表示方法将每个字作为信息的基本单位, 使用 skip-gram 算法以字为单位在大规模数据集上训练, 将其映射为 300 维的字向量作为模型的输入。skip-gram 算法的计算如式 (9) 所示。

2.3 双向长短时记忆网络层

为了能同时考虑短文本上下文语义信息, 充分提取短文本包含的所有特征, 将短文本采用字向量表示之后, 构建双向长短时记忆网络。输入单元 $X = \{x_1, x_2, \dots, x_t\}$, 其中每个 $x_i (i=1, 2, \dots, t)$ 为字向量。隐含层包含前向传播层和后向传播层。其

中,前向传播层单元用于学习当前文本上文信息,后向传播层单元用于学习下文信息。同时,连接到输出层单元。 W_{i1} 、 W_{i2} 分别为输入门到前向单元和后向单元的权值矩阵, W_{f1} 、 W_{f2} 分别为前向传播层和后向传播层隐含单元遗忘门的权值矩阵, W_{o1} 、 W_{o2} 分别为前向层隐含单元和后向传播层隐含单元到输出门的权值矩阵。双向长短时记忆网络接收 n 维字向量作为输入,分别通过隐含层前向单元和后向单元得到 t 个时刻的隐态输出向量,其矩阵表示为 \vec{H} 和 \overleftarrow{H} ,则 BLSTM 隐层输出如式(10)所示。

$$H = \{\vec{H}, \overleftarrow{H}\} \quad (10)$$

其中, $\vec{H} = \{h_1, h_2, \dots, h_t\}$, $\overleftarrow{H} = \{h_1, h_2, \dots, h_t\}$ 分别为隐层前向及后向输出。

2.4 Self-Attention 层

在权重调整层采用自注意力机制,传统的注意力模型需要依赖部分外部信息,是一种与外部信息对齐的思想,而自注意力机制只需要对自身信息训练来更新参数。如式(11)所示。

$$\text{Attention} = \text{softmax}\left(\frac{X \cdot X^T}{\sqrt{d_k}}\right)X \quad (11)$$

其中, $X \in R^n$ 为 BLSTM 输出单元的 n 维输出向量, $\sqrt{d_k}$ 是调节因子,一般为输入向量的维度。在本文指词向量或字向量的维度,其作用是避免 $X \cdot X^T$ 内积过大造成 softmax 值非 0 即 1 的情况。

2.5 模型训练

本文使用梯度下降算法进行模型训练,损失函数使用交叉熵损失。 y 为训练样本的实际值, \hat{y} 作为模型计算过程中的预测输出值。模型的目标是最小化训练样本中句子的预测输出值和实际样本值的交叉熵。交叉熵的计算如式(12)所示。

$$L = -\frac{1}{n} \sum_x [y \ln \hat{y} + (1 - y) \ln(1 - \hat{y})] \quad (12)$$

其中, L 为损失值, x 为样本, n 为样本数, y 为样本实际值, \hat{y} 为模型预测输出值。

3 实验

本文在字向量表示方法的基础上结合 Self-Attention 机制和 BLSTM 网络进行短文本情感分类。将训练文本采用字向量化表示,输入模型迭代训练,最后输出为预测值。实验环境如表 1 所示。

表 1 实验环境

硬件环境	CPU: Intel i5-4130 3.4GHz
	内存: 4G DDR3L
软件环境	windows7 64bit Python 3.5.2
	tensorflow 1.3.0 numpy1.14.0

3.1 实验数据

数据集采用中文信息学会推出的中文倾向性分析评测 (Chinese Opinion Analysis Evaluation, COAE) 2014 年任务四微博文本数据集和酒店评论数据集,数据标注均由数据提供平台完成。COAE 2014 数据集总共有 40 000 条微博数据,已人工标注 10 000 条数据。其中,带有积极情感的数据 5 132 条,带有消极情感的数据 4 868 条;酒店评论数据集选自中科院计算所谭松波提供的酒店评论语料,包含已标注的 3 039 条正向情感数据,2 813 条负向情感数据。本文在已标注数据集上采用五折交叉验证进行多次实验,尽可能地降低随机性对实验结果的影响。部分数据样例如表 2 所示。

表 2 实验数据样例

Positive	Negative
手机性能不错,反应快	屏幕较差,拍照很粗糙
离单位很近,从价格来说,性价比很高	酒店脏、旧、破!房间厕所都是下水道的臭味
离火车站近,交通方便	房间空气难闻,有点小

3.2 评价标准

本文采用国际通用评价标准准确率 (Precision)、召回率 (Recall)、 F 值 (F-Measure)^[19] 对实验结果进行评价。准确率指分类器正确分类样本在数据集中的比例,它反映了分类器对分类样本的正确识别能力;召回率反映了分类器可以检测出来的所有正确样本占数据集中此类样本的比例; F 值是准确率和召回率的调和均值,当 $\alpha = 1$ 时,即最常见的 F_1 值。其相关计算如式(13)~式(15)所示。

$$P = \frac{TP}{TP + FP} \quad (13)$$

$$R = \frac{TP}{TP + FN} \quad (14)$$

$$F_1 = \frac{2 * P * R}{P + R} \quad (15)$$

其中,TP、TN、FP、FN 含义为表 3 所示。

表 3 混淆矩阵表

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

3.3 实验参数设置

深度学习模型的参数设置非常关键,本文实验主要参数与参数值如表 4 所示。

表 4 实验参数表

参数	参数描述	参数值
r	学习率	0.001
batch_size	数据批处理量	128
hidden_units	隐含层单元数	100
dropout	神经元丢弃率	0.2

3.4 对比实验设置

为了证明模型的有效性,将本文提出的字向量表示方法结合 Self-Attention 和 BiLSTM 的模型与下列目前较好的情感分析模型在相同环境下进行对比实验。模型中词向量或字向量分别使用 Word2Vec 和 skip-gram 算法在本文数据集上进行预训练得到,为了保证对比实验单一变量原则,均采用同一类预训练的词向量或字向量。

(1) LSTM。虽然 LSTM 是在 RNN 的基础上加入门限控制,但是几乎所有使用 LSTM 的论文其 LSTM 模型结构与参数设置均有某些不同,本对比实验使用的是 Zaremba 等描述的网络结构作为标准模型^[21]。

(2) BiLSTM。采用 Zaremba 等描述的 LSTM 网络来构建 BiLSTM 网络模型。基于词的文本向量表示,结合双向长短时记忆网络捕捉句子的上下文情感语义信息。

(3) ATT-BiLSTM。Wang 等提出的情感分析模型^[22],在 BiLSTM 网络中引入 Attention 机制,采用 BiLSTM 提取文本序列化信息,Attention 内容注意力机制依据特征对分类任务的重要程度区别表示。

(4) Word2Vec-ATT-BiLSTM。江伟等提出的基于 Word2Vec 对文本向量化表示方法,结合 Attention 和 BiLSTM 模型的短文本情感分析算法^[16]。

(5) Word2Vec-SATT-BiLSTM。使用 Word2Vec 对文本向量化表示,结合 Self-Attention 和 BiLSTM 的短文本情感分析算法。

(6) character-SATT-BiLSTM。本文提出的使用字向量表示方法对文本向量化表示,结合 Self-Attention 和 BiLSTM 网络的短文本情感分析算法。

3.5 实验结果与分析

本文算法与其他对比算法分别在相同数据集上的实验结果如表 5、表 6 所示。

表 5 COAE 数据集实验结果

数据集 评价标准	COAE(2014 年)中文评测数据集		
	准确率/%	召回率/%	F ₁ 值/%
LSTM	82.02	83.15	82.58
BiLSTM	82.67	83.59	83.13
ATT-BLSTM	86.35	85.27	85.81
Word2Vec-ATT-BiLSTM	87.22	86.22	86.71
Word2Vec-SATT-BiLSTM	87.67	89.52	88.59
character-SATT-BiLSTM	90.04	89.45	89.74

表 6 酒店评论数据集实验结果

数据集 评价标准	酒店评论数据集		
	准确率/%	召回率/%	F ₁ 值/%
LSTM	84.09	83.11	83.59
BiLSTM	85.76	85.24	85.50
ATT-BiLSTM	88.22	87.56	87.88
Word2Vec-ATT-BiLSTM	89.13	87.43	88.27
Word2Vec-SATT-BiLSTM	90.35	89.04	89.69
character-SATT-BiLSTM	91.67	90.55	91.10

(1) 通过对比 LSTM 和 BiLSTM 算法的实验结果可知,BiLSTM 由于增加了后向传播单元可以同时考虑上下文信息,通常较单向 LSTM 更优。

(2) 通过对比 BiLSTM 和 ATT-BiLSTM 算法的实验结果可知,通过加入注意力机制,能很大程度上获取句子的局部信息,模型准确率、召回率、F₁ 值

分别提高了 2.46%、2.32%、2.38%。

(3) 通过对比 ATT-BiLSTM 和 Word2Vec-ATT-BiLSTM 算法的实验结果,表明使用 Word2Vec 词向量表示能保留词语之间的语义信息,降低语义信息丢失,有一定的性能提升。

(4) 通过对比算法 Word2Vec-ATT-BiLSTM 和 Word2Vec-SATT-BiLSTM 算法在两个数据集的实验结果可知,Self-Attention 机制相比标准的 Attention 有不错的性能提升。因为 Self-Attention 可以减少普通注意力机制的外部参数依赖,在文本中通过关联单个序列的不同位置以计算序列的表示来获得更加准确的句子内部关键特征,同时输入和输出序列中任意组合位置之间路径更短,学习远距离依赖性就更容易,结合 BiLSTM 提供的位置信息来建模序列前后关系,是自注意力机制与普通神经网络在短文本情感分类领域结合的探索,算法最终的 F_1 值提高了 1.42%。

(5) 通过对比算法 Word2Vec-SATT-BiLSTM 和 character-SATT-BiLSTM 的实验结果可知,因本文使用的数据集均为评论性文本数据,有内容短、用词随意性强等特点。故本文利用字级别的文本向量表示方法可以更加细粒度的提取到短文本稀疏特征,更完整的保留某些细节信息,利用 Self-Attention 对局部关键特征进行捕捉。同时,也可以降低分词错误产生的噪声影响,在两类数据集上的 F_1 值分别较 Word2Vec-SATT-BiLSTM 算法提高 1.15%和 1.41%。本文方法在 COAE 2014 数据集上的准确率、召回率、 F_1 值高于文献[23]的最好性能;在酒店评论数据集上的 F_1 值优于文献[24],充分证明了本文算法的有效性。

4 结论

本文提出了基于 Self-Attention 机制的 BiLSTM 短文本情感分析算法。通过字向量对短文本向量化表示,利用 BiLSTM 提取短文本的上下文特征,引入 Self-Attention 机制对特征重要程度动态调整,在两类数据集上实验证明了字向量表示方法的基于 Self-Attention 和 BiLSTM 算法在短文本情感分类方面的有效性。

然而,本文算法主要考虑了短文本特征稀疏性问题,后期工作将会考虑对其他特征建模,提高算法性能。由于 BiLSTM 等序列模型属于递归模型,并行能力较弱,后期将在此方面做进一步改进。

参考文献

- [1] 王仲远,程建鹏,王浩勋,等. 短文本理解研究[J]. 计算机研究与发展,2016,53(2): 262-269.
- [2] Picard R W. Affective computing: challenges[J]. International Journal of Human-Computer Studies,2003,59(s1-2): 55-64.
- [3] Jeffrey L Elman. Finding structure in time[J]. Cognitive Science,1990,14(2): 179-211.
- [4] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation,1997,9(8): 1735-1780.
- [5] Esin Y E, Alan O, Alpaslan F N. Improvement on corpus-based word similarity using vector space models[C]//Proceedings of the International Symposium on Computer and Information Sciences. IEEE, 2009: 280-285.
- [6] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research Archive,2003,3: 993-1022.
- [7] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[C]//Proceedings of the Innovations in Machine Learning. Springer Berlin Heidelberg,2006: 137-186.
- [8] Mikolov T, Corrado G, Chen K, et al. Efficient estimation of word representations in vector space[C]//Proceedings of the International Conference on Learning Representations, 2013: 1-12.
- [9] 许云,樊孝忠,张锋. 一种不需分词的中文文本分类方法[J]. 北京理工大学学报,2015,25(9): 778-781.
- [10] 刘龙飞,杨亮,张绍武,等. 基于卷积神经网络的微博情感倾向性分析[J]. 中文信息学报,2015,29(6): 159-165.
- [11] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Proceedings of Advances in neural information processing systems, 2014: 2204-2212.
- [12] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409.0473,2016: 1-15.
- [13] Vaswani, Ashish, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017,22(7): 139-147.
- [14] Pavlopoulos J, Malakasiotis P, Androutsopoulos I. Deeper attention to abusive user content moderation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1125-1135.
- [15] Zhao Q L, Cai X D, et al. Text feature extraction

- method based on LSTM-Attention neural network [J]. Modern Electronics Technique, 2018; 41 (8): 167-170.
- [16] 江伟,金忠. 基于短语注意机制的文本分类[J]. 中文信息学报,2018,32(2): 102-109.
- [17] Yao X. Attention-based BiLSTM neural networks for sentiment classification of short texts[J]. Proceedings of Science, 2017: 1-8.
- [18] Cheong M, Lee V. Dissecting Twitter: A review on current microblog research and lessons from related fields[M]. From Sociology to Computing in Social Networks. Springer, Vienna, 2010: 343-362.
- [19] Kim Y, Denton C, Hoang L, Alexander M. Rush. Structured attention networks [C]//Proceedings of International Conference on Learning Representations, 2017: 1-21.
- [20] Chen Y, Zhao X, Jia X. Spectral-spatial classification of hyperspectral data based on deep belief network [J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2015, 8 (6): 2381-2392.
- [21] Wojciech Zaremba, Ilya Sutskever, Oriol Vinyals. Recurrent neural network regularization [J]. Eprint Arxiv, 2014: 1049-1329.
- [22] Wang Y Q, Huang M, Zhao L, et al. Attention-based LSTM for aspect-level sentiment classification[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing, 2016: 606-615.
- [23] 陈钊,徐睿峰,桂林,等. 结合卷积神经网络和词语情感序列特征的中文情感分析[J]. 中文信息学报, 2015,29(6): 172-178.
- [24] 冯兴杰,张志伟,史金钊. 基于卷积神经网络和注意力模型的文本情感分析[J]. 计算机应用研究, 2018(5): 1-4.



吴小华(1993—), 硕士, 主要研究领域为机器学习与自然语言处理。
E-mail: xiaohua9303@sina.cn



陈莉(1963—), 通信作者, 教授, 博士生导师, 主要研究领域为大数据与数据挖掘。
E-mail: chenli@nwu.edu.cn



魏甜甜(1993—), 硕士, 主要研究领域为机器学习、推荐系统。
E-mail: weitian@stumail.nwu.edu.cn