ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 5,Sep. 2018,pp. 921－927
DOI:10.16337/j.1004-9037.2018.05.017

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

卷积神经网络下的 Twitter 文本情感分析

王煜涵[1]　张春云[1]　赵宝林[2]　裘肖明[1]　耿蕾蕾[1]　崔超然[1]

(1.山东财经大学计算机科学与技术学院,济南,250014;2.浪潮电子信息产业股份有限公司存储研发部,济南,250101)

摘　要: 随着社交网络的日益普及,基于 Twitter 文本的情感分析成为近年来的研究热点。Twitter 文本中蕴含的情感倾向对于挖掘用户需求和对重大事件的预测具有重要意义。但由于 Twitter 文本短小和用户自身行为存在随意性等特点,再加之现有的情感分类方法大都基于手工制作的文本特征,难以挖掘文本中隐含的深层语义特征,因此难以提高情感分类性能。本文提出了一种基于卷积神经网络的 Twitter 文本情感分类模型。该模型利用 word2vec 方法初始化文本词向量,并采用 CNN 模型学习文本中的深层语义信息,从而挖掘 Twitter 文本的情感倾向。实验结果表明,采用该模型能够取得 82.3％的召回率,比传统分类方法的分类性能有显著提高。

关键词: Twitter 文本;情感分析;词向量模型;卷积神经网络
中图分类号: TP391.1　　**文献标志码:** A

Sentiment Analysis of Twitter Data Based on CNN

Wang Yuhan[1], Zhang Chunyun[1], Zhao Baolin[2], Xi Xiaoming[1], Geng Leilei[1], Cui Chaoran[1]

(1. School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, 250014, China; 2. Storage R & D Department, Inspur Electronic Information Industry Co., Ltd, Jinan, 250101, China)

Abstract: With the increasing popularity of social networks, sentiment analysis based on Twitter text has become a hotspot in recent years. The sentiment tendencies contained in tweets are important for mining user needs and predicting major events. However, the existing sentiment classification methods are mostly based on hand-made text features, and it is hard to mine implicit deep semantics of texts. In addition, because of special characteristics, such as short text and arbitrariness of users' behavior, it is more difficult to improve performance of current sentiment classification. This paper presents a novel Twitter sentiment classification model based on convolutional neural network (CNN). In order to explore sentiment tendency of tweets, the proposed model utilizes a dynamic CNN architecture to learn deep semantics from tweets, which initializes input word embedding with word2vec method. Experimental results show that our proposed model can achieve a recall rate of 82.3％, which is much higher than performances of traditional classification methods.

Key words: Twitter data; sentiment analysis; word embedding model; convolutional neural network (CNN)

基金项目: 山东省高等学校优势学科人才团队培育计划资助项目;山东省自然科学杰出青年基金(JQ201316)资助项目;山东省自然科学基金(ZR2016FQ18) 资助项目;山东省高等学校科技计划(J17KA065)资助项目。
收稿日期: 2017-06-07;**修订日期:** 2017-06-26

引 言

情感分析(Sentiment analysis),又被称作观点挖掘或观点分析,其目标是通过数据挖掘得出文本的情感极性并分析判断文本的情感走向。情感分析在互联网各个应用场景中发挥着重要的作用。随着在线社交网络平台(如国内的微博、微信、国外的 Twitter 和 Facebook)的爆炸式增长,在社交网络平台上表达情感态度、发表自己的观点渐渐成为人们的生活习惯。由于这一变化,社交网络平台产生了蕴含用户观点态度、情感倾向的海量文本数据。挖掘这些信息并分析其情感倾向性,在舆情监控及优化个性化推荐等方面都有重要的意义和价值。

在基于社交网络的情感分析研究课题中,微博情感分析(Microblog sentiment analysis)被视为重要的研究课题。微博(Microblog)是社交媒体中经典的短文本来源。微博平台(如推特、新浪微博)每天都能够收集大量的微博信息,其中包含的信息资源具有较大的挖掘价值。然而,微博文本信息具有数据短小精炼、形式自由以及拼写错误较多等特点,给传统的情感分析等任务带来了巨大的挑战。目前情感分类多使用有监督学习的方法,而少使用半监督及无监督的学习方法。因此在分类过程中,特征的表示显得尤为重要。在过去的研究中,常应用词频—逆文本频率指数(Term frequency-inverse document frequency, TF-IDF),One-hot representation 等作为特征选择的算法,但由于其具有不能良好地联系文章上下文及语境和词向量稀疏等问题,因此在特征选择的过程中有一定的局限性。

为解决上述问题,本文采用卷积神经网络对文本进行建模,通过 word2vec 模型对互联网 web 数据进行训练获得词向量初始化数据,并基于 Twitter 文本的初始化向量采用深度卷积神经网络进行学习,挖掘出 Twitter 文本中蕴含的深层语义,从根本上解决了推文短小、上下文缺失的问题。基于该模型对 Twitter 数据进行情感倾向性分类,取得了 82.3% 的 F1 值,比传统的情感分类方法有明显提升。

1 情感分析技术的应用

常见的情感分析技术主要有两种,一是基于机器学习的有监督分类方法,二是基于情感词典的无监督分类方法。基于情感词典的方法[1]是传统情感分析中重要的模型。情感词典中包含带有较强情感意味的词语,因此可以利用此信息进行文本词汇的情感标注。以 Ku[2] 的研究为例,此方法主要是将情感词表与人工制定的规则相结合来进行分类,而采用基于情感词典的无监督分类方法训练模型时,不需要人工情感进行标注,所使用的情感词典对于分析表现具有极强的影响。但其最大的问题是情感词典中的词汇数量有限,无法判别生僻或新兴词汇的极性。

随着机器学习算法的兴起,基于机器学习的情感分析方法受到越来越多的学者青睐。Pang[3] 等首次应用机器学习方法解决情感分析问题,他们尝试使用 N-gram 模型提取特征,并且分别用支持向量机模型(Support vector machine, SVM),朴素贝叶斯(Naive Bayes, NB)模型以及最大熵(Maximum entropy, ME)模型进行测试,结果发现使用一元分词(Unigram)作为特征提取方法,并使用 SVM 进行分类,可以得到较高的准确率。随着语料库的不断扩大,N-gram 表现出了越来越好的特性[4]。机器学习方法需要解决两个关键的问题:(1)如何抽取出复杂而非简单的特征;(2)如何识别出哪一类特征最富有价值。众多学者提出了多种特征选择和提取的方法,如单一词语(Single-words)模型[5]、单一文字(Single-character)模型[6] 以及 Multi-word N-gram[7-8] 模型和词汇语法模型。但这些模型只是在挖掘句子中词语之间的词汇特征和句法特征[9],对语义特征极少研究。然而语义特征蕴含重要信息,对情感分析起着重要的作用。

随着机器学习算法的不断发展,词向量的概念也被引入到情感分析的领域中。One-hot representation 是一种十分经典的向量表示方法,它基于文本特征可以极快地生成词向量。但其明显的缺陷是应用此方法表示的词向量维度较高,会造成向量稀疏的问题,而且此方法存在"词汇鸿沟",难以表达词语之间的关系。为了解决这些问题,Hiton[10] 等提出了词向量(Word embedding)方法,通过对文本数据的训练将词语不同的语法特征及句法特征映射到向量的不同维度,用向量空间上的某个点表示相应的词

语,从而解决向量稀疏问题,并且词语在低位空间中的位置关系反映了语义层面的关系,因此本文采用 Word embedding 方法进行词语特征的学习。

近年来,随着深度学习的发展,学者们开始使用神经网络模型[11]进行建模,以解决文本中歧义与多义的问题。Collobert 等[12]将卷积神经网络(Convolutional neural network,CNN)引入了自然语言处理的领域并解决了许多实际问题,取得了较好的结果。Shen 等[13]将 CNN 引入信息检索的语义分析任务中,得到了较高的准确率。这些工作证明了 CNN 在自然语言处理中具有广泛的应用前景,因此本文基于卷积神经网络进行建模以解决 Twitter 文本的情感分析[14]。

2 基于词向量的 CNN 情感分类模型

2.1 词向量模型

基于 Bengio[15]等提出的神经网络语言模型[16](Neural network language model,NNLM),应用 Word2vec 进行词语的向量化表示。Word2vec 是一个基于 Mikolov 等[17]提出的持续词袋模型(Continuous bag-of-words,CBOW)和 Skip-gram 模型基础上的词向量处理工具。NNLM 为三层前馈神经网络结构,工作原理如图 1 所示。

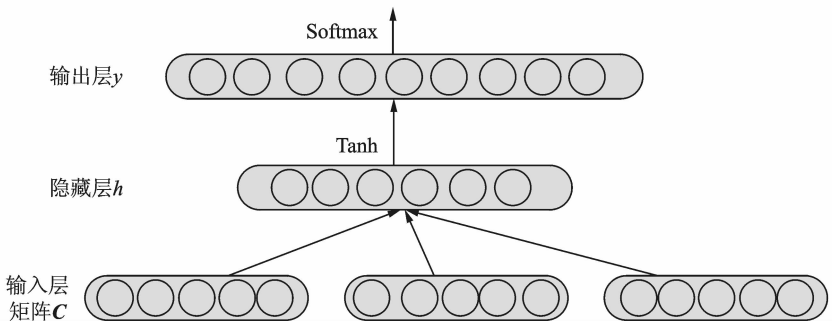

图 1 词向量生成图

Fig. 1 Schematic diagram of word embedding produce

假设需要通过此语言模型预测的词语为ω_i,可根据其前 $n-1$ 个词语$\omega_{i-n+1},\cdots,\omega_{i-2},\omega_{i-1}$个词语进行预测。$\omega$ 所对应的词向量用 $C(\omega)$表示,矩阵 $\boldsymbol{C}$ 中存储整个模型中使用的词向量,$|V|\times M$ 为 $\boldsymbol{C}$ 的维度,其中词表的大小用$|V|$表示,词向量的维度用 M 表示。网络的输入层是一个$(n-1)\times m$ 维的向量,该向量由 $C(\omega_{i-n+1}),\cdots,C(\omega_{i-2}),C(\omega_{i-1})$这 $n-1$ 个向量首尾相连拼接起来,记为 $\boldsymbol{x}$,隐藏层和输出层可分别表示为

$$h=\tanh(b+\boldsymbol{H}\boldsymbol{x}) \tag{1}$$

$$y=d\boldsymbol{H}+\boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}h \tag{2}$$

式中:b,d 为偏置项,初始化值随机;tanh 为激活函数;$\boldsymbol{U}$ 为隐藏层到输出层的权重矩阵;$\boldsymbol{H}$ 为输入层到隐藏层的权重矩阵,$\boldsymbol{W}$ 包含了从输入层到输出层的直连边。模型最终通过 Softmax 函数归一化输出层 y,将其变为目标词的概率分布,即

$$p(\omega_i\mid\omega_{i-n+1},\cdots,\omega_{i-2},\omega_{i-1})=\frac{\exp(y(\omega_i))}{\sum_{k=1}^{|V|}\exp(y(\omega_k))} \tag{3}$$

训练时,优化的目标最大化式为

$$\sum_{\omega_{i-n+1}}\log P(\omega_i\mid\omega_{i-n+1},\cdots,\omega_{i-2},\omega_{i-1}) \tag{4}$$

最后使用梯度下降法进行模型的优化,优化结束后某个词对应的词向量可从输出的 y 值中获取。

2.2 卷积神经网络模型

应用CNN方法可以充分利用上下文,且可以避免人工设置特征。本文所使用的卷积神经网络由词向量映射层、卷积层、Max-pooling层、全连接层和Softmax层5部分组成,其结构图如图2所示。

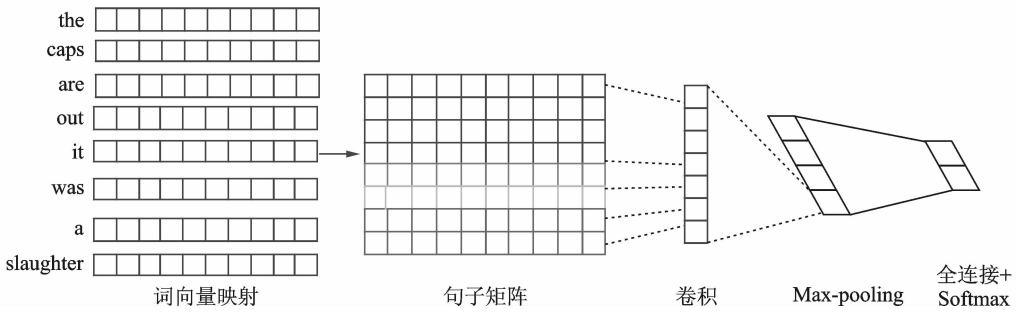

图2 卷积神经网络结构图

Fig. 2 Structure diagram of CNN

2.2.1 词向量映射层

卷积神经网络最早应用于图像的处理和识别。图像由像素点矩阵组成,因此需要将文本信息转换为二维矩阵的形式进行输入。词向量是基于大量的语料库集合利用神经网络算法经过无监督学习得到的词语的低维表示形式,因此可以充分地表示词语之间的相似性和上下文相关的特征。本文应用Twitter数据集,并且以句子为一个样本单位进行处理。如果文本中最长的句子包含 n 个词语,且每个向量有 k 维特征,则卷积神经网络的输入是由 n 个 k 维矩阵组成的 $n\times k$ 的二维矩阵,如图1所示。

2.2.2 卷积层

卷积层的核心为滤波器,由不同大小的卷积核提取多组局部的特征图,进而挖掘出Twitter文本中不同的特征。每个卷积层都要定义一个大小固定的滑动窗口,且每次都处理窗口内的信息。本文选择N-gram作为卷积的基本单位,窗口长度定义为 l,l 即为N-gram中的 n。卷积操作中连续 l 个词的词向量组成N-gram向量 $\boldsymbol{c}_i$,即

$$\boldsymbol{c}_i = v_i \oplus v_{i+1} \oplus \cdots \oplus v_{i+l-1} \tag{5}$$

式中:$\oplus$ 表示连接操作(Concatenate),即将窗口内的词向量首尾相接组成一个更长的N-gram向量。设包含填充向量的句子长度为 N,则输入句子的N-gram矩阵为

$$\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_j, \cdots, \boldsymbol{c}_{N-l+1}] \tag{6}$$

式中:输入句子 $\boldsymbol{C}\in\mathbf{R}^{dl\times(N-l+1)}$,定义权重矩阵为 $\boldsymbol{W}^{\mathrm{m}}\in\mathbf{R}^{rl\times d}$,$rl$ 为该滤波器的区域大小,b_{c} 为偏置因子,卷积操作可以表示为

$$\boldsymbol{X} = f_{\mathrm{c}}(\boldsymbol{W}^{\mathrm{m}} \cdot \boldsymbol{C} + b_{\mathrm{c}}) \tag{7}$$

式中:$\boldsymbol{X}\in\mathbf{R}^{r\times(N-l+1)}$,$f_{\mathrm{c}}$ 表示非线性激活函数。使用非线性激活函数会增强网络的拟合能力,常用的激活函数有Tanh函数、Sigmod函数和Relu函数。为加快收敛速度,此模型中应用Relu非线性激活函数。

2.2.3 Max-pooling

Max-pooling层是在卷积层的基础上,对文本特征进行全局筛选,降低信息冗余,选择最能体现Twitter文本的特征。本文选择最大池化法对特征进行筛选得到 $\hat{\boldsymbol{X}}$,即

$$\hat{\boldsymbol{X}} = \max\{\boldsymbol{X}\} \tag{8}$$

池化层不仅可以起到减少参数数量和滤过噪声的作用,并且可最终产生文本向量 $\boldsymbol{S}$,作为下一层的输入。

2.2.4 Softmax层和全连接层

Softmax层和全连接层起到分类的作用,句子向量 $\boldsymbol{S}$ 经过矩阵 $\boldsymbol{W}^{\mathrm{o}}\in\mathbf{R}^{L\times fK}$ 的线性映射,成为类别向

量$\boldsymbol{S}^{'}\in\mathbf{R}^{L}$,其中 L 表示任务所需的类别总数,即有

$$\boldsymbol{S}^{'}=\boldsymbol{W}^{\circ}\cdot\boldsymbol{S}+b_{\mathrm{a}} \tag{9}$$

为了得到观点类别的估计值,需要使用分类函数对其进行处理。通过 Softmax 函数决策出概率最大的类,即有

$$p(y^{i}=j\mid S^{'(i)},\boldsymbol{\theta})=\frac{\exp(\boldsymbol{\theta}_{j}^{\mathrm{T}}S^{'(i)})}{\sum_{k=1}^{L}\exp(\boldsymbol{\theta}_{j}^{\mathrm{T}}S^{'(i)})} \tag{10}$$

式中:$\boldsymbol{\theta}$ 代表深度网络需要学习的参数集合,$S^{'(i)}$ 代表第 i 条推文,$p(y^{i}=j\mid S^{'(i)},\boldsymbol{\theta})$ 表示推文 i 的观点类别为 j 的概率,最终将推文归为概率最大的那个类别中,θ 可定义为

$$\boldsymbol{\theta}=[\boldsymbol{X},\boldsymbol{W}^{\mathrm{e}},\boldsymbol{W}^{\mathrm{m}},\boldsymbol{W}^{\circ},b_{\mathrm{e}},b_{\mathrm{c}},b_{\mathrm{a}}] \tag{11}$$

3 实验结果与分析

3.1 实验数据集

实验采用的数据是从 Twitter 平台通过其 API 接口爬取的数据,共包含 10 万条。因为数据正负样本数量不成比例并且存在超短文本,会导致分类结果出现偏差,所以为保证数据的质量,对 10 万条 Twitter 数据进行预处理操作,包括去除无意义的以及重复的数据,最终选取 48 418 条负面情感数据和 43 078 条正面情感数据作为情感分类的实验数据集,其中,训练数据集中包括 36 313 条负面情感数据和 32 308 条正面情感,其余的为测试数据集。

3.2 实验评价指标及参数

为避免模型过于复杂所致的过拟合情况的发生,本文仅设计了 1 层卷积层和 1 层 Max-pooling 层,用以提高模型的泛化能力。词向量初始化时,此模型设置词向量的维度为 300,卷积层滤波器个数为 100,并且设滤波器对应的窗口大小分别为 3,4,5,最后采用 Relu 激活函数进行计算。为防止局部最优,使用L_2正则化和 Dropout 策略控制模型训练的过程,且设 Dropout 值为 0.5,每批处理的最小样本数为 50。

参数基于 Twitter 文本数据(从 Twitter 平台 API 爬取后进行处理)进行选取,通过网络搜索方法进行优化,最终得到最优解。为使实验结果更具说服力,另外采用梯度下降法以及 10 折交叉验证进行模型的训练。

3.3 实验设计及结果

(1)基于传统机器学习方法的情感分类模型

TF-IDF 和 One-hot presentation 都是自然语言处理中最为经典的特征表示方法。将上述预处理后的推文作为输入,分别用词袋模型和 One-hot representation 进行特征表示,然后将训练数据输入到 logistics 分类器中进行参数的学习,测试数据输入到分类器中进行预测,分别得到了 74%和 39%的 $F1$ 值。

(2)word2vec+logistics

将 Google 提供的训练数据集输入到 word2vec 的工具中进行词向量的学习,表 1 展示了部分词语的相似词语以及相对应的余弦相似度。在上述所训练词向量的基础上,计算每条 Twitter 文本的句子向量,即将每条 Twitter 文本中的词向量相加取平均值得到句子

表 1 相似词向量及相似度

Tab. 1 Similar word vectors and similarity

情感词	相似情感词	余弦相似度
Bad	Horrible	0.596 59
	Terrible	0.562 88
	Shabby	0.538 51
	Nasty	0.534 81
Happy	Pleased	0.518 17
	Lovely	0.513 85
	Celebrating	0.508 35
	Pleasing	0.508 14
Boring	Lonely	0.692 77
	Quiet	0.682 90
	Miserable	0.660 75
	Depressing	0.647 26

向量。将训练数据的句子向量输入到 logistics 分类器中进行参数学习，测试数据输入到分类器中进行测试，最终得到的分类准确率为 69%。虽然用到了神经网络算法计算词向量，但因为 logistics 分类器不是上下文相关的分类算法，所以结果并不理想。

(3) word2vec+CNN

将 Google 提供的训练数据集输入到 word2vec 的工具中，进行词向量学习。再将上述进行过预处理的每条 Twitter 文本用矩阵表示，可以表示为动态矩阵或静态矩阵。动态矩阵是指在 word2vec 词向量训练的基础上，通过 CNN 对词向量的参数进行优化，得到的句子矩阵。而静态矩阵则是直接使用 Word2vec 训练的向量生成的矩阵，然后再经过卷积层、池化层以及全连接层对 Twitter 文本矩阵进行分类得出相应的分类结果。CNN(静态)得到的准确率为0.820，CNN(动态)得到的准确率为 0.823。可以看出动态矩阵比静态矩阵的分类效果更佳，且比传统机器学习方法的分类效果有显著的提高。

3.4 实验结果分析

基于词向量的 CNN 模型在 Twitter 文本分类中可以得到较好的结果，原因在于：(1)基于神经网络的词向量模型可以根据词语之间的关系及上下文的语境充分表达语义信息，它克服了 TF-IDF 中文章相关性差的缺陷，同时也解决了 One-hot 向量表示词特征时矩阵稀疏的问题；(2)基于 CNN 的神经网络模型充分学习了文本的语义信息，不再仅通过独立的词语对文本进行分类。从表 2 的结果可以看出，仅用 word2vec 进行词向量学习后再通过传统方法进行 Twitter 文本的分类，准确率并没有发生显著的变化，但应用 CNN 进行文本参数的学习及分类后，准确率却有了显著的提升。因此基于词向量的 CNN 模型是解决 Twitter 文本分类有效的方法。

表 2 文本分类效果

Tab. 2 Effect of text classification

文本分类模型	准确率/%	召回率/%	*F*1 值
TF-IDF+ logistics	74	74	0.74
One-hot+ logistics	63	52	0.39
Word2vec+logistics	69	69	0.69
Word2vec+CNN(静态)	82	81.4	0.821
Word2vec+CNN(动态)	82.3	82.3	0.823

4 结束语

为解决传统机器学习方法在 Twitter 文本分类中未能充分利用上下文及语义和数据稀疏的问题，本文提出了利用上下文对 Twitter 文本进行分类的基于词向量的 CNN 模型。为适应 Twitter 文本简短、不规范和表情符号过多的特点，首先应用 word2vec 对词语进行向量化表示，进而对 Twitter 文本进行向量化表示，再通过 CNN 对 Twitter 文本进行学习，最终得到较为优化的分类器。实验显示运用此方法可以得到较好的预测结果。但是，因为 Twitter 文本较为个性化，所以在今后的工作中应该加以考虑用户以及主题因素以进一步提高分类质量。

参考文献：

[1] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2): 267-307.

[2] Ku L W, Wu T H, Lee L Y, et al. Construction of an evaluation corpus for opinion extraction[C]//Proceedings of NTCIR-5 Workshop Meeting. Tokyo, Japan: [s. n.], 2005.
[3] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2002: 79-86.
[4] Ekmekcioglu F C, Lynch M F, Willett P. Stemming and n-gram matching for term conflation in Turkish texts[J]. Information Research News, 1996, 7(1): 2-6.
[5] Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis: The good the bad and the omg[C]//International Conference on Weblogs and Social Media. Barcelona, Spain: [s. n.], 2011, 11: 164.
[6] Bouamor D, Semmar N, Zweigenbaum P. Identifying bilingual multi-word expressions for statistical machine translation[C]//International Conference on Language Resources and Evaluation. Istanbul, Turkey: [s. n.], 2012: 674-679.
[7] Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters[J]. Computational Intelligence, 2006, 22(2): 110-125.
[8] Xavier G, Antoine B, Yoshua B. Domain adaptation for large-scale sentiment classification: A deep learning approach[C]//Proceedings of the 28th International Conference on Machine Learning. Washington: ACM, 2011: 97-110.
[9] Paccanaro A, Hinton G E. Learning distributed representations of concepts using linear relational embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(2): 232-244.
[10] Tomas M, Stefan K, Lukas B, et al. Extensions of recurrent neural network language model[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic: IEEE, 2011: 5528-5531.
[11] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.
[12] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168-177.
[13] Shen Y, He X, Gao J, et al. Learning semantic representations using convolutional neural networks for web search[C]//Proceedings of the 23rd International Conference on World Wide Web. Korea: ACM, 2014: 373-374.
[14] 王伟,周咏梅,阳爱民,等. 基于种子词的微博表情符情感倾向判定方法[J]. 数据采集与处理, 2017, 32(1): 198-204.
Wang Wei, Zhou Yongmei, Yang Aimin, et al. Determination method for sentiment orientation of microblog smileys based on seed words[J]. Journal of Data Acquisition and Processing, 2017, 32(1): 198-204.
[15] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
[16] Ronan C, Jason W. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008: 160-167.
[17] Mikolov T, Sutskeer I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

作者简介:

王煜涵(1996-),女,本科生,研究方向:自然语言处理,E-mail: hanhanwinny@126.com。

张春云(1986-),通信作者,女,讲师,研究方向:信息抽取、自然语言处理和机器学习等,E-mail: zhangchunyun1009@126.com。

赵宝林(1986-),男,研发工程师,研究方向:嵌入式系统开发与研究、存储软件开发、存储介质研究与开发。

裘肖明(1987-),男,讲师,研究方向:机器学习与数据挖掘、模式识别和医学图像处理。

耿蕾蕾(1984-),女,讲师,研究方向:基于字典学习的图像处理应用研究、遥感数据评价等。

崔超然(1987-),男,教授,研究方向:信息检索、推荐系统、多媒体分析与处理等。

(编辑:王 静)