北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2018.066

基于表示学习的情感分析研究

厉小军　施寒潇†　陈南南　柳虹　邹轶

浙江工商大学管理工程与电子商务学院，杭州 310018;
† 通信作者, E-mail: hxshory@foxmail.com

摘要　提出一个基于表示学习的文本情感分析模型 C&W-SP。首先基于 C&W 模型的词表示改进训练模型，实现在词表示训练过程中融入情感信息和词性信息的不同模型设计；然后利用 NLP&CC'2013 中的评测数据集，进行多种模型的实验对比。实验结果表明，融入情感信息和词性信息的 C&W-SP 模型性能效果最优，验证了所提出方法的有效性。

关键词　情感分析；表示学习；深度学习；词表示

Research on Sentiment Analysis Based on Representation Learning

LI Xiaojun, SHI Hanxiao†, CHEN Nannan, LIU Hong, ZOU Yi

School of Management and E-Business, Zhejiang Gongshang University, Hangzhou 310018;
† Corresponding author, E-mail: hxshory@foxmail.com

Abstract　The authors propose C&W-SP model — a text sentiment analysis model based on representation learning. Firstly, an improved training model based on C&W model is proposed which can integrate emotional information and part of speech information in the training process of word embedding. The evaluation of data sets of NLP&CC'2013 is used to compare experimental results with different models. The experimental results show that the C&W-SP model which combines emotion information and part of speech information has the best performance and confirm the effectiveness of the proposed method.

Key words　sentiment analysis; representation learning; deep learning; word embedding

随着Web2.0与互联网技术的不断发展，互联网正在从单向传播模式逐步向以用户为中心的模式转变。用户也从网络知识获取者转向更加积极的网络信息制造者，社会媒体的巨大变化使得网络媒介中出现海量的主观性文本，包含用户观点、用户情感和用户情绪等信息。该类文本大多都包含用户对待某一事件或某一事物的情感极性信息，比如喜、悲、哀、愁、赞、踩、批等情绪。通过分析含有用户主观信息的文本，可以了解大众对于某一政策、某一事件或某一产品的态度和看法，政府也可以通过这些信息更快地了解到大众的真实想法，从而辅助政府做出更好的决策；同时，商家也可以更加了解产品在市场的具体情况，进而更好地改善产品。

情感分析(sentiment analysis)主要针对人们对于产品、服务等的观点，进行有效的挖掘与分析，对信息进行归纳和推理。在传统的情感分析中，主要采用人工特征抽取和传统机器学习方法，这些方法对传统的特征抽取工作依赖度较高，同时模型的性能对标注训练集的依赖性很强，可扩展性较差。

本文采用基于表示学习的方法实现文本的情感分析。首先利用自然语言处理技术进行词性标注以及引入句子的情感类别，改进词表示的生成模型；　然后利用池化模型构建句子表示；　最后，利用公开的评测语料进行实验对比。实验结果表明，本文提出的基于表示学习方法具有非常好的性能。

国家社会科学基金(17BTQ069)资助
收稿日期: 2018–04–15; 修回日期: 2018–08–19; 网络出版时间: 2018-08-22 15:46:50

1 相关研究

1.1 情感分析研究现状

在 20 世纪 90 年代末, 学者们开始研究情感分析, 渐渐成为自然语言处理领域的热点。情感分析的目的是将具有情感倾向的主观性文本进行识别, 分为褒义和贬义两类。例如“我喜欢你”、“今天我好高兴啊”、“这间屋子真不错”、“明天又是好心情”等文本为褒义, 而“我感到很悲伤”、“这真是太无聊了”、“心情不好”等文本为贬义。目前针对情感分析的研究方法主要有传统的情感分析和基于统计学习的情感分析。

传统的情感分析方法主要采用基于规则的方法。在基于规则的方法中, 规则的分析和制定占据大部分的工作, 同时还需要相当一部分人力和物力作为支撑, 若研究的目标样本中存在语言现象过多或结构相对复杂, 则规则的制定就非常困难, 同时, 该方法和当前的研究目标之间的关系十分密切, 会使得整个方法没有很好的迁移性。因此, 学者们转向基于统计的学习方法, 该方法主要根据特征的分布, 对文本的情感类别做出正确的判断。

基于统计学习的方法主要分为监督学习、无监督学习和半监督学习三类。

监督学习(supervised learning)模型利用大量的带标记样本的训练而得到。常见的监督学习方法有支持向量机(support vector machine, SVM)和朴素贝叶斯(Naïve Bayes, NB)等。Pang 等[1]将影评数据集分为正负两类, 并且使用 unigrams 作为数据集的分类特征。Tong[2]利用聚合的方法, 针对特定领域的词或词组进行手工编辑, 将得到的结果作为词典进行情感分类实验。Pang 等[3]使用基于图的最小割算法。Mullen 等[4]和 Xia 等[5]均采用句法关系和传统特征混合的方法。Kennedy 等[6]和 Li 等[7]使用上下文情景价态和情感移位的方法。

监督学习方法在准确率上能够取得较好的结果, 但是伴随信息的不断更新, 需要处理的数据越来越多, 所以为待处理数据分配标记成为一项非常艰巨的任务。这不仅意味着需要大量的人力和物力, 而且每个人自我判断的差异性会使标记出来的数据出现不统一等问题, 给后续处理带来麻烦。同时, 通过监督学习方法得到的模型的迁移性比较差。

无监督学习(unsupervised learning)方法能够解决使用人工标记的方法来获得有标记数据的问题。Taboada 等[8]采用基于词库的方法, 其基本思想是用一个带有相关倾向性和强度信息的情感词词典, 采用集约化的方法计算文本的最后情感值, 确定文本的最后情感倾向。Hu 等[9]提出采用 Bootstrapping 策略, 句子中所有情感词的情感倾向性分数总和决定该句子的情感倾向, 其思想也是用基于词库的方法来解决情感倾向性问题。Wiebe 等[10]提出使用部分种子词, 最后主客观分类的结果由计算梯度的方法获得, 该过程中没有任何的学习过程, 完全使用无监督学习方法解决句子级的主客观分类问题。

无监督学习方法在实际操作过程中确实可以节省大量的人力和物力, 然而模型预测的准确率比有监督学习差很多。为了使有监督学习的方法获得较高的准确率, 同时又可以借鉴无监督学习的思想, 在操作中减少大量的人工标注相关的任务, 出现了半监督学习方法。半监督学习(semi-supervised learning)方法主要是利用少量的有标记样本和大量的无标记样本共同学习实现, 大大节省了研究人员的时间和精力, 同时学到的模型通常还具有很强的泛化能力。Davidov 等[11]提出 SASI (semi-supervised sarcasm identification)算法来处理客观信息的分类和观点句的提取, 在主客观分类方面 SASI 算法是第一个比较健壮的算法, 且在试验中通常具有很好的表现。Su 等[12]针对词级别主客观信息的分类问题, 提出利用 WordNet 词典和关系结构的半监督最小割(semi-supervised minimum cut)方法进行处理。这些研究都有效推动了半监督学习方法在情感分析上的进一步应用。

随着近几年机器学习的发展, 深度学习方法在文本情感分析任务上的应用逐渐增多, 并在一定程度上开始显现效果。

1.2 深度学习及其在文本分析上的研究现状

在 1985 年, Ackley 等[13]提出将随机机制加入 Hopfield 网络, 提出玻尔兹曼机模型, 这种模型对应于真实空间, 易于理解, 但在模型训练过程中算法往往不收敛, 易发散。1986 年, Smolensky[14]首先提出限制玻尔兹曼机(restricted boltzmann machine, RBM)模型, 模型中相同的节点彼此孤立, 连接仅存在于可见节点与隐藏节点之间, 算法相对高效。2006 年, Hinton 等[15]提出两个主要观点: 1)人工神经网络的多隐层结构拥有优异的对特征学习的能力, 学习到的特征可以更加本质的刻画样本数据特征, 进而对于图像可

视化或者文本等的分类任务更加有利; 2)在深度神经网络中存在的困难可以通过“逐层初始化”(layer-wise pre-training)的方法有效克服。从此深度学习的研究浪潮在学术界和工业界同时展开。

Mikolov 等[16]将 RNN 用于语言模型建模, 2013 年, 又提出 Continue Bag-of-Word (CBOW)模型和 Continue Skip-gram (Skip-gram)模型[17]。CBOW 模型利用周边的词汇预测中间的词汇获得词表示, Skip-gram 模型正好相反, 是利用中间词汇预测周围词汇获得词表示。

在自然语言处理领域, 句子和文档的表示学习也逐渐成为研究热点。句子和文档的表示学习可以分为有监督学习方法和无监督学习方法。有监督学习方法主要面向情感分析任务、关系分类等分类任务。Huang 等[18]针对句子的表示提出深度语义相似度匹配模型(deep semantic similarity model, DSSM), 模型的目标函数根据句子之间的语义相似度来设置。在无监督学习方法中, Le 等[19]提出在训练语言模型预测目标词汇过程中, 将句子的表示(paragraph vector)加入到输入层, 即 Paragraph Vector 方法。

在情感分析研究中, 深度学习也得到广泛应用。Tang 等[20]提出将情感信息融入到词表示训练中, 获得情感词表示(sentiment-specific word embedding, SSWE)。模型以 Collobert 等[21]提出的 C&W 语言模型为基础, 输入为滑动窗口内 n-gram 和相应的情感类标签, 输出一个二维标量(分别代表语义分数和情感分数), 获得情感词表示。Tang 等[22]利用 tweet 语料进行情感词表示学习, 将获得的情感词表示应用到 3 种情感分析任务(词水平的情感分析、句子水平的情感分析和建立情感词典), 进行词相似性比较。实验结果表明, 与基于上下文的词表示相比, 情感词表示拥有更好的性能。Ren 等[23]针对 twitter 情感分析, 提出将主题信息编码到词表示中的方法, 利用 LDA 主题模型来获取主题信息, 通过改进的递归自编码框架, 将主题信息结合到目标函数, 获取结合主题信息的词表示。与传统的方法相比, 该方法在 tweets 情感预测任务中有更好的性能。在目标依赖的情感分类任务中, 不同的上下文对目标词在句子中的情感极性有不同的影响。因此, 在构建学习系统时需要联合目标词和上下文词。Tang 等[24]通过扩展标准 LSTM 来构建目标依赖的 LSTM 模型(TD-LSTM), 该模型可以自动考虑目标信息。利用 twitter 基准数据集对标准的 LSTM 和 TD-LSTM 进行评估, 发现 TD-LSTM 能够显著提高精度。

2 基于情感和词性的词表示模型

词表示(word embedding)一直是自然语言处理领域的研究热点, 它作为一种新颖的特征在情感分析任务中使用, 并且表现优异。然而, 在传统模型中, 词表示的训练一般都是以语义共现作为词表示训练的基本原理, 所以获得的词表示通常只包含文本中的部分语义和语法信息, 针对情感信息任务, 还有所欠缺。

本文以 C&W 为基础模型, 由于其无监督的特性, 导致训练的词表示只利用了部分语义信息。为了使训练得到的词表示可以更好地处理情感分析任务, 本文通过改进 C&W 模型, 将情感信息和词性信息嵌入到词表示训练过程。由于信息融合的多样性, 我们提出两类基于 C&W 改进模型的词表示学习模型, 其不同点在于使用不同的策略来融合情感信息和词性信息。

2.1 C&W 模型

现有的词表示模型主要有 C&W 模型[21]和 word2vec 模型[25]等, 考虑到模型的效率和可扩展性, 本文选择 C&W 模型作为基础模型。C&W 模型是由 Collobert 等[21]提出的一种神经网络模型, 可以快速、高效地生成词表示, 并且是直接以生成词表示为训练目标。C&W 模型并没有预测当前单词 w_i 的条件概率分布 $P(w_i \mid w_{i-n+1},..., w_{i-2}, w_{i-1})$, 而是对短语 $(w_{i-n+1},..., w_{i-2}, w_{i-1}, w_i)$ 直接进行打分。即, 给定训练语料库中固定长度(一般为奇数)的任一短语 $s \in S$, $s=(w_{i-(n-1)/2},...,w_{i-1},w_i,w_{i+1},...,w_{i+(n-1)/2})$, 通过使用词典中的其他单词 w_i'将句子中心位置的单词 w_i 随机替换, 被替换的短语记为 s'。通常短语 s'遭到破坏后语法错误或语义错误, 因此 C&W 模型对 s 和 s'分别打分, 使得短语 s 的分数 $f(s)$比短语 s'的分数 $f(s')$至少高 1。所以, C&W 模型的训练目标是最小化:

$$L=\sum_{s\in D}\sum_{w\in V}\max(0,1-f(s)+f(s')), \quad (1)$$

其中, s'表示将原短语 s 的中心单词替换成 w 后的短语。该模型的输入层向量是由训练样本中所有词语拼接而成, 与神经网络语言模型相比, 最大的不同在于其只有一个输出节点, 大大地缩短了词表示的训练时间。

2.2 C&W 模型改进

虽然 C&W 模型可以快速、有效地生成词表示, 但是模型本身并没有预测当前词 w_i 的条件概

率分布 $P(w_i \mid w_{i-n+1},\cdots, w_{i-2}, w_{i-1})$，而是对短语直接进行打分。也就是说, C&W 模型训练词表示的过程中，仅仅利用了句子 $(w_{i-n+1},\cdots, w_{i-2}, w_{i-1}, w_i)$ 的语法相关性。由于模型是无监督的，所以最后获得的词表示只含有语法和语义信息。图 1 为 C&W 模型的框架，由输入层(lookup)、线性隐藏层(linear)、非线性隐藏层(hTanh)、输出层(linear)4 层构成。

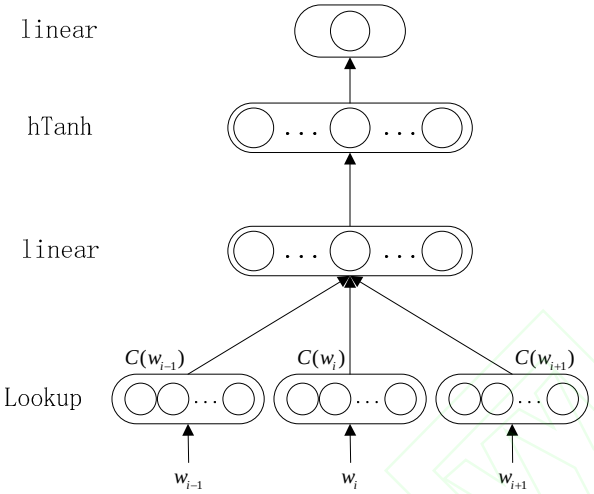

图 1　C&W 模型框架

Fig. 1　C&W model framework

模型通过对窗口内连续 n 个词打分 $f(w_{t-n+1},\cdots, w_{t-1}, w_t)$ 的形式，近似求解文档概率。f 越高说明该短语越合理，反之则不合理。模型的损失函数为

$$loss_{cw}(t,t^r)=\max(0,1-f^{cw}(t)+f^{cw}(t^r)), \quad (2)$$

其中替换的词序列用 t^r 表示, t 表示正常的词序列, $f^{cw}(\bullet)$ 代表语言模型的分数。对于式(1)中的 $f^{cw}(\bullet)$，可由式(3)~(5)计算:

$$f^{cw}(t)=w_2(a)+b_2, \quad (3)$$

$$a=hTanh\,(w_1L_t+b_1), \quad (4)$$

$$hTanh\,(x)=\begin{cases}-1 & if \quad x<-1\\ x & if \quad -1\le x\le 1\\ 1 & if \quad x>1\end{cases}, \quad (5)$$

2.2.1　SSWE模型

在情感分析任务中，往往会出现句子结构相同，但是所表达的情感不同。例如: “这是一本很棒的书！！！”和“这是一部很糟糕的手机。。。”。

“棒”和“糟糕”在句子结构中处在相同的位置，在 C&W 模型训练词表示的过程中，会将“棒”和“糟糕”两个情感相反的词映射成两个距离很近的向量，使训练出来的词表示在词表示空间模型中处于很相近的位置，所以导致情感分类模糊。

基于 C&W 模型的词表示学习框架，Tang 等[20]提出将情感信息加入到词表示训练过程中，最后获得蕴含情感信息的情感词表示学习模型 SSWE。在传统的词表示训练模型中，词表示的获得一般仅根据语法和语义的相关性获得，最后获得的词表示只含有部分语言信息，对于情感分析任务，这种模型忽略了句子中情感信息的影响。SSWE 模型结构如图 2 所示，可以看出该模型同样分为输入层(Lookup)、线性隐藏层(Linear)、分线性隐藏层(hTanh)、线性输出层(Linear)4 层。

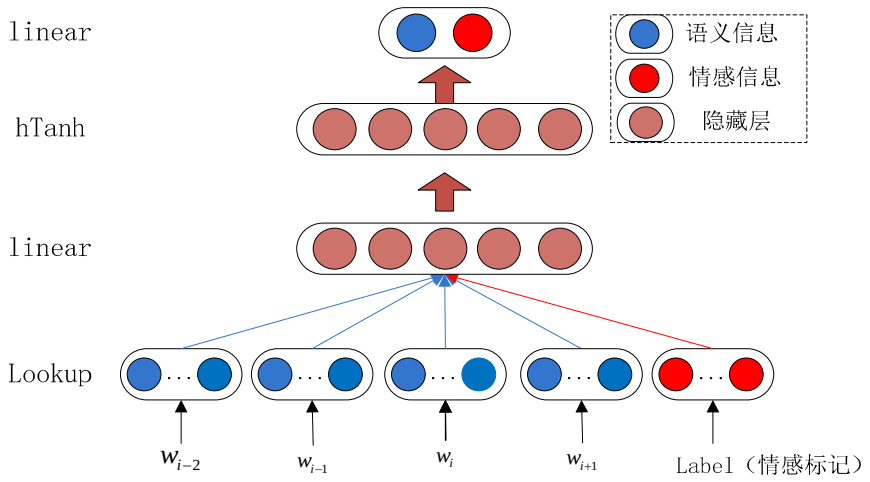

图 2 基于 C&W 模型的情感词表示学习模型(SSWE)

Fig. 2 Emotional word representation learning model (SSWE) based on C&W model

SSWE 模型将情感信息和语义信息融合在同一维度空间进行训练，经过线性隐藏层和非线性隐藏层，输出一个二维标量(分别代表语义分数和情感分数)。与 C&W 模型类似，SSWE 使用连续的 n-gram 作为模型输入，以当前 n-gram 所在句子的极性作为类标，通过随机梯度下降和反向传播算法更新模型中的参数以及输入的词表示。

SSWE 模型的损失函数为

$$loss_{sswe}(t,t^r)=a\cdot loss_{cw}(t,t^r)+(1-a)\cdot loss_{us}(t,t^r), \quad (6)$$

其中，超参数 a 为权重，$loss_{cw}(t,t^r)$ 为 C&W 模型损失函数(式(2))，$loss_{us}(t,t^r)$ 为情感损失函数，即

$$loss_{us}(t,t^r)=\max(0,1-\delta_s(t)f_1^u(t)+\delta_s(t)f_1^u(t^r)), \quad (7)$$

其中，$f_1^u(t)$ 为 t 的情感分数，$f_1^u(t^r)$ 为 t^r 的情感分数，$\delta_s(t)$ 为反映句子情感倾向的指示函数:

$$\delta_s(t)=\begin{cases}1 & if \quad f^g(t)=[1,0]\\ -1 & if \quad f^g(t)=[0,1]\end{cases}, \quad (8)$$

其中，$f^g(t)$ 为标准情感倾向, [1, 0]表示积极, [0, 1]表示消极。

基于 SSWE 模型的分析与研究，本文提出两种词表示学习模型，进一步探究词表示的有效性。第一种基于词性信息的词表示模型(C&W Based Part-of-speech Word Embedding, C&W-P)，是在不同的维度空间内将词性信息与原有语义信息分开训练；第二种是混合词表示模型(C&W Based Sentiment and Part-of-speech Word Embedding, C&W-SP)，将 SSWE 和 C&W-P 相结合，将词表示空间分为情感–语义空间和词性空间两个维度。

2.2.2 C&W-P模型

在自然语言处理任务中，一词多义、多词性现象在中文语料中经常出现。例如: “老师鼓励我们好好学习”和“这是对我们的一种鼓励”。

虽然“鼓励”在两个句子中都是“激励，激发”的意思，但是在句子“老师鼓励我们好好学习”中是以动词的形式出现，而在句子“这是对我们的一种鼓励”中的出现形式是名词。Zhang 等[26]总结了一般不同的词性需要不同的对待，大部分的情感词为形容词或副词，有时名词或动词也会表述情感。因此，针对句子中可能出现的这种歧义，词性标注显得尤为重要。

基于以上考虑， C&W 模型在训练时，将每个词的词性标记融入进去，使得到的词表示可以同时兼顾语义信息和词性信息。C&W-P 模型将语义信息和词性信息在两个不同的维度空间内进行训练，语义空间部分继续使用 C&W 模型中的训练方法，词性空间在对目标词上下文预测的同时，将出现在窗口内的所有词的词性信息通过加和的方式映射到一起，作为 Softmax 层的输入，通过随机梯度下降和反向传播算法来更新模型中的参数以及输入的词表示。模型如图 3 所示。

C&W-P 模型的损失函数为

$$loss_{cw-p}(t)=-\sum_{k\in(0,1)}f_k^g(t)\cdot\log(f_k^{cw-p}(t)), \quad (9)$$

其中 $f_k^g(t)$ 为标准的词性类标分布，$f_k^{cw-p}(t)$ 为经过模型预测的词性类标分布。

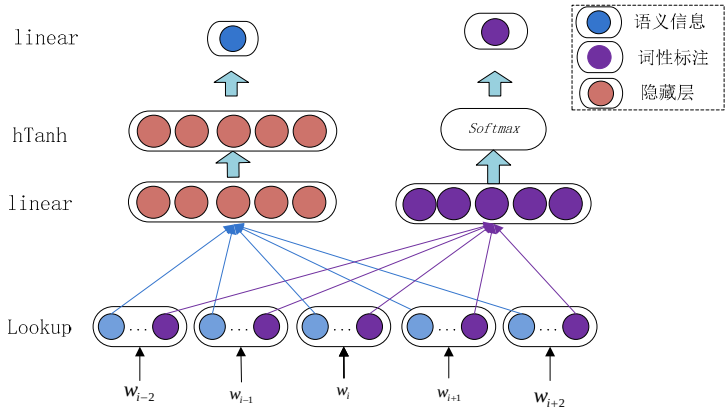

图 3 基于 C&W 模型的词性词表示学习模型 C&W-P

Fig. 3 Part-of-speech representation learning model (C&W-P) based on the C&W model

2.2.3 C&W-SP模型

SSWE 模型和 C&W-P 模型通过不同的方式将情感信息和词性信息与语义信息结合，为了更进一步的探究语义信息与情感信息和词性信息的关系，本文提出 C&W-SP 模型。

类似 C&W-P 模型， C&W-SP 模型将生成的词表示分解为情感–语义空间和词性空间两个维度。情感–语义空间部分等同于 SSWE 模型的整个部分，即将情感和语义信息融合在一起。模型中的词性空间的构造方法与 C&W-P 模型相同，在该部分仅考虑词性信息。C&W-SP 模型结构如图 4 所示。

C&W-SP 模型的损失函数为

$$loss_{cw-sp}(t,t^r)=a\cdot loss_{sswe}(t)+(1-a)loss_{cs-p}(t,t^r), \quad (10)$$

其中，$loss_{sswe}(t)$、$loss_{cw-p}(t,t^r)$ 分别代表情感–语义空间损失函数和词性空间损失函数，a 为权值。

3 实验与分析

为了验证本文提出的两种词表示训练模型学习得到的词表示在情感分析任务中能的表现，同时也为了甄选出更好的词表示训练模型，得以在接下来的情感分析模型中使用，首先使用本文提出的两种词表示训练模型和 SSWE 词表示训练模型在中文和英文公开评论数据集上分别训练，获得相应的词表示，然后对句子进行句子表示的构建，并进行情感分类任务的对比实验，从而验证各种词表示的性能。

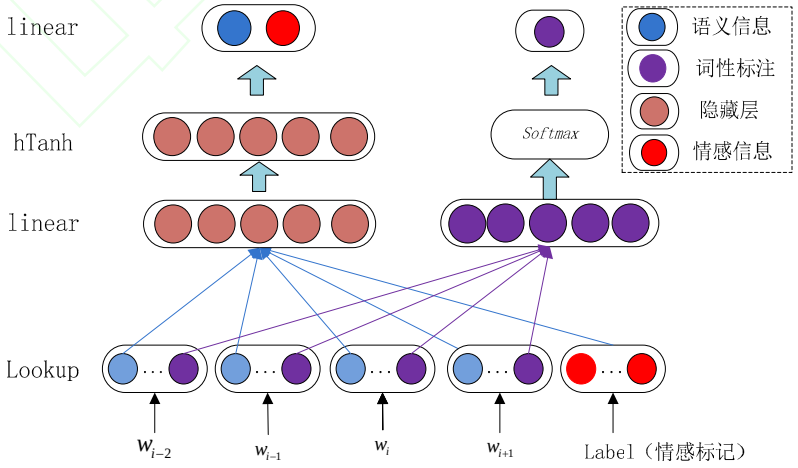

图 4 基于 C&W 模型的情感和词性向量学习模型 C&W-SP

Fig. 4 C&W-based sentiment and part-of-speech vector learning model C&W-SP

为了使用本文提出的词表示模型训练出词表示，首先对词表示矩阵进行初始化 $M \in R^{N\times d}$ (N 为训练数据中词的数目， d 为词表示的维度), M 中的每一行代表一个词表示。将模型的滑动窗口大小设置为 5，词频阈值设置为 3，训练迭代次数设置为5，词表示维度设为100(d=100)，学习速率为0.1。使用反向传播算法和随机梯度下降，将训练误差传递到各个词表示中。在训练过程中，几个模型均通过最小化各自的损失函数来得到词表示。

3.1 数据集

实验数据集为第三届自然语言处理与中文计算 (nature language processing and chinese computing, NLP&CC)会议中基于深度学习的情感分类(sentiment classification with deep learning)任务中公开的商品评论数据集，数据分别从中文和英文电商网站上获取，数据来自多领域(包括书籍、DVDs和电子产品等)。数据集中文和英文评论各12500条，其中10000条(5000条为积极，5000条为消极)作为训练数据，2500条作为验证数据。

3.2 实验设置

针对中英文词表示训练数据和模型训练与验证数据，第一步是要对数据进行预处理操作:

1) 对于英文评论数据，预处理的第一步是将所有字母转化为小写，并使用斯坦福大学的Stanford parser工具进行词性标注。

2) 对于中文评论数据，预处理的第一步是使用opencc工具，将所有数据中的繁体字转换为简体字； 然后使用Python结巴分词中的精确模式，对所有文本进行分词处理； 最后使用斯坦福大学的Stanford parser工具进行词性标注。

预处理完成后，取其中训练数据并将情感类标和词性标注类标分别输入本文提出的两种词表示训练模型和SSWE词表示训练模型，分别得到维度为100的3种中文词表示和英文词表示。

参考 Collobert 等 [27] 和 Socher 等 [28] 的 min, average和max池化方法进行句子建模。最终句子向量由 $V_{\min}(t)$， $V_{average}(t)$ 和 $V_{\max}(t)$ 3个向量拼接而成: $V(t)=[V_{\min}(t),V_{average}(t),V_{\max}(t)]$。由词表示构建句子向量的过程如图5所示。

为了评估词表示训练模型在情感分类任务中的效果，使用F1参数作为评估标准。机器学习分类算法使用Scikit-learn工具包中的支持向量机训练，其中核函数为线性核，惩罚系数c为1。另外，为确保结果的稳定性，试验重复10次，以平均值作为最后的预测结果。

3.3 基准系统

为了对比本文提出的词表示训练模型与传统的词表示训练模型在情感分析任务中的表现，本实验设立了3个基准系统：n元模型(n-gram)、CBOW模型和C&W模型。其中，CBOW模型和C&W模型的词表示训练数据和标准实验数据与所有对比实验数据集相同，不同的是输入数据不含有任何情感和词性标签。

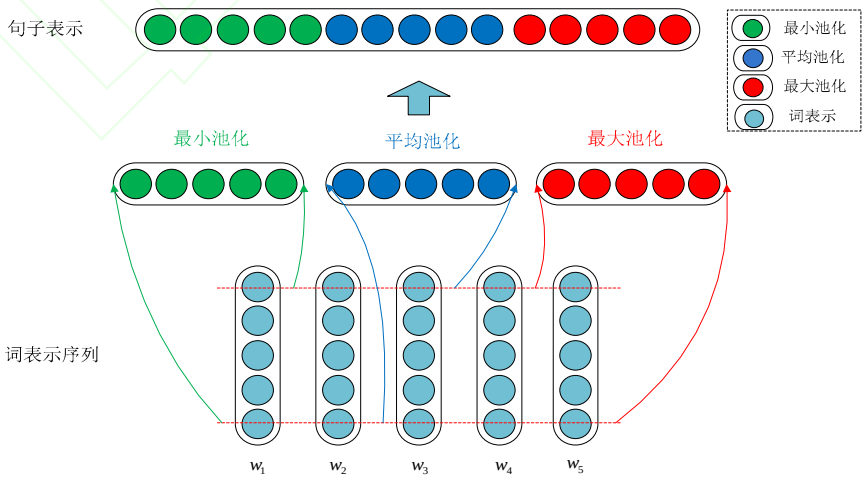

图5 由词表示构建句子向量过程

Fig. 5 Process of constructing sentence vectors by word embedding

表1 不同词表示在中英文评论数据上的情感分类结果(%)
Table 1 Sentiment classification results on Chinese and English review data based on different words embedding (%)

模型	中文评论			英文评论		
	Precision	Recall	F1	Precision	Recall	F1
n-gram(一元)	62.27	65.31	63.75	67.69	69.05	68.36
n-gram(一元/二元/三元)	64.76	65.51	65.13	68.95	70.73	69.83
CBOW	66.40	69.54	67.93	69.19	72.51	70.81
C&W	68.76	69.53	69.14	70.83	73.88	72.32
SSWE	70.41	72.20	71.29	73.16	75.96	74.53
C&W-P	69.93	70.69	70.31	72.25	75.17	73.68
C&W-SP	**72.52**	**74.41**	**73.45**	**74.64**	**76.98**	**75.79**

3.4 实验结果及分析

我们把几种模型引入到在中文评论和英文评论两种语料集上，进行对比实验，结果如表1所示。根据表1，可以得到以下结论。

1) 本文提出的C&W-SP词表示模型的结果明显高于其他模型。这说明在词表示训练过程中，将情感因素和词性信息加入到模型训练，能够提高情感分析任务的精确度。

2) C&W-SP模型和SSWE模型的实验结果优于C&W-P模型，原因可能是在情感分析的任务中，与词性信息相比，句子中的情感信息对情感分类结果的影响更大。

3) 比较C&W，SSWE，C&W_P和C&W_SP的结果可知，在处理情感分析任务时，句子情感类标签和句子中各词的词性标签都会对分析结果产生影响。

4) n-gram模型的结果最差，原因可能是无论一元还是多元模式都会导致学习到的特征是高维稀疏的，不能很好获取词的上下文信息。

综上所述，在中文和英文商品评论领域，本文提出的C&W-SP词向量训练模型均有最好的表现，同时，句子中的情感信息和词性信息均会影响情感分类的效果。

4 结论

在情感分析任务中，为了减少传统词表示忽略情感和词性信息对结果的影响，本文提出了基于C&W模型的词表示改进模型C&W_P和C&W_SP。在C&W模型的基础上分别以不同的方式将情感信息和词性信息加入到模型训练中，使获得的词表示蕴含词性和情感信息。为了验证词表示学习框架的有效性，在中文和英文商品评论公开数据上进行了词表示对比实验，实验结果表明，融入词性信息和情感信息词表示训练模型在文本情感分析任务中性能最好。

本文在句子表示生成过程中只是简单地利用最大、平均和最小池化方法来进行，而没有将句子中的词序信息考虑进去，未来的研究我们将引入卷积神经网络、长短时间记忆网络等方法，进行句子表示向量的构建，进一步研究基于深度学习的情感分析方法。

参考文献

[1] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques // Proceedings of ACL. Stroudsburg, 2002: 79–86

[2] Tong R M. An operational system for detecting and tracking opinions in on-line discussion // Proceedings of the ACM SIGIR Workshop on Operational Text Classification. New Orleans, 2001: 1–6

[3] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts // Proceedings of ACL. Barcelona, 2004: 271–278

[4] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources // Proceedings of EMNLP. Barcelona, 2004: 412–418

[5] Xia R, Zong C. Exploring the use of word relation features for sentiment classification // Proceedings of ACL. Uppsala, 2010: 1336–1344

[6] Kennedy A, Inkpen D. Sentiment classification of

movie reviews using contextual valence shifters. Computational intelligence, 2006, 22(2): 110–125

[7] Li S, Lee S Y M, Chen Y, Huang C R, et al. Sentiment classification and polarity shifting // Proceedings of COLING. Beijing, 2010: 635–643

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

[9] Hu M, Liu B. Mining and summarizing customer reviews // Proceedings of KDD. Seattle, 2004: 168–177

[10] Wiebe J. Learning subjective adjectives from corpora // Proceedings of AAAI. Austin, 2000: 735–740

[11] Davidov D, Tsur O, Rappoport A. Semi-supervised recognition of sarcastic sentences in twitter and amazon // Proceedings of CoNLL. Uppsala, 2010: 107–116

[12] Su F, Markert K. Subjectivity recognition on word senses via semi-supervised mincuts // Proceedings of HLT-NAACL. Boulder, 2009: 1–9

[13] Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for Boltzmann machines. Cognitive Science, 1985, 9(1): 147–169

[14] Smolensky P. Information processing in dynamical systems: foundations of harmony theory. Parallel Distributed Processing: Explorations in The Microstructure of Cognition, 1986, 1: 194–281

[15] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507

[16] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model // Proceedings of INTERSPEECH. Makuhari, 2010: 1045–1048

[17] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations // Proceedings of NAACL-HLT. Atlanta, 2013: 746–751

[18] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data // Proceedings of CIKM. San Francisco, 2013: 2333–2338

[19] Le Q V, Mikolov T. Distributed representations of sentences and documents. Computer Science, 2014, 4: 1188–1196

[20] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for twitter sentiment classification // Proceedings of ACL. Baltimore, 2014: 1555–1565

[21] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning // Proceedings of ICML. Helsinki, 2008: 160–167

[22] Tang D, Wei F, Qin B, et al. Sentiment embeddings with applications to sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(2): 496–509

[23] Ren Y, Wang R, Ji D. A topic-enhanced word embedding for Twitter sentiment classification. Information Sciences, 2017, 369: 188–198

[24] Tang D, Qin B, Feng X, et al. Effective LSTMs for target-dependent sentiment classification [EB/OL]. (2016–09–29)[2018–05–04]. https://arxiv.org/abs/1512.01100

[25] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013–09–07)[2018–04–01]. https://arxiv.org/abs/1301.3781

[26] Zhang L, Liu B. Sentiment analysis and opinion mining // Sammut C, Webb G. Encyclopedia of Machine Learning and Data Mining. Boston: Springer, 2016.

[27] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12(1): 2493–2537

[28] Socher R, Huang E H, Pennington J, et al. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Advances in Neural Information Processing Systems, 2011, 24: 801–809