文章编号：1003-0077(2018)06-0124-08

基于部首和音位的情感词汇表示模型

徐琳宏[1]，林鸿飞[2]，祁瑞华[1]，关菁华[1]

(1. 大连外国语大学 软件学院，辽宁 大连 116044；2. 大连理工大学 计算机科学与技术学院，辽宁 大连 116024)

摘　要：文本情感分析是自然语言处理的热点问题之一，而词汇是情感分析的基础。汉字通过声音和形状表达意义，该文综合考虑词汇中每个字的部首和音位等信息，构建了一个情感词汇分类模型。在模型中，将词汇的字、部首和音位三种信息向量化，与原始词汇向量融合，生成新的情感词汇表示，最后采用前馈神经网络和卷积神经网络对情感词汇的极性进行分类。实验结果表明，三种细粒度特征都能有效地提高情感词汇的分类效果，并且该文在COAE评测的语料上验证了模型的有效性。

关键词：部首；音位；神经网络

中图分类号：TP391　　**文献标识码**：A

Sentiment Lexicon Embedding Based on Radical and Phoneme

XU Linhong[1], LIN Hongfei[2], QI Ruihua[1], GUAN Jinghua[1]

(1. School of Software, Dalian University of Foreign Languages, Dalian, Liaoning 116044, China;
2. School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: Text Sentiment Analysis, one of the hot topics in natural language processing, is based on the analysis of lexicon. Considering Chinese characters, the constituents of lexicon, convey their meaning through sounds and logograph, this paper aims at building a taxonomy of sentiment lexicon by the comprehensive analysis of the radicals and phonemes of each character. In our model, each Chinese character, radicals and phonemes are vectorized and then integrated with the original word vector to generate new expressions of sentiment lexicon, and finally the polarities of sentiment lexicon are categorized with feedforward neural network, convolutional neural network and other approaches. Experiment results reveal that three types of vector features have effectively improved the accuracy of sentiment lexicon classification, as well as a better sentiment sentence classification. results in COAE materials.

Key words: radical; phoneme; neural network

0　引言

基于文本的情感计算是一个多学科交叉的研究课题，研究内容涉及自然语言处理、心理学、认知与脑科学，以及语言学等多个学科。在自然语言技术不断发展的今天，人们可以通过分析语法结构、语义信息和情感词汇等方法从文本中抽取情感信息。从大量的文本中提取其中包含的情感信息在许多方面都有广阔的应用前景，如舆情监控、产品倾向性分析、语音合成、信息安全、智能机器人、模式识别、个性化文本、解析文章情感结构等。情感倾向性分析主要分为词汇、语句和篇章三个层面，其中词汇是语句和篇章计算的基础，包含丰富语义信息的词汇向量有助于对语篇情感的理解。

在词汇表示方面的主要方法有：一种是最简单的词向量，即 one-hot representation，这种方法简单，但不能有效表示词语之间的相关性。还有一种是目前广泛采用的分布式向量表示，它的每个向量由多个非零分量表示，能很好地体现词汇之间的关系。目前应用比较广泛的是 Word2vec，该模型因为训练速度快、代码容易复现等原因被广泛采用。上

收稿日期：2017-03-28　**定稿日期**：2017-08-22

基金项目：国家社会科学基金(15BYY028)；辽宁省自然科学基金(2015020017，20170540230，20170540232)；辽宁省优秀人才项目(LJQ2014127)

述词汇表示方法主要计算词汇所在的上下文的语义，忽略了词汇本身的字和部首等内在特征，下面介绍一些细粒度词汇表示的研究工作。

1　相关工作

1.1　细粒度的中英文词汇表示的相关工作

在细粒度词汇表示方面，英文主要研究构成单词的多个字母，中文方面研究构成词的字和部首等。英文的词汇表示方面，深入到字母级别的研究有：2013 年，Chrupala 等人提出一种简单的循环神经网络(SRN)学习字符级的向量表示[1]。2014 年，dos Santos 将字符级别的词汇表示用于词性标注，得到较好的效果[2]。dos Santos 在 2014 年还采用卷积网络得到词汇的字符级表示，并在 STS(stanford twitter sentiment)和 SSTb(stanford sentiment treebank)两个数据集上进行情感分类，相比 2013 年的研究结果均有 2%左右的提高[3]。

在汉语词汇表示方面，近几年也有人做一些更细粒度的词汇表示，例如汉字和部首等。2010 年 Yue Zhang 等人根据词汇中某部分的特征切割整个词汇，使用到了词汇中的字特征，在准确率不降低的情况下，系统运行速度提高 10 倍多[4]。2013 年 Meishan Zhang 等人使用词汇中的字特征，构建了词性标注和短语分析等系统，取得了约 2%左右的性能提高[5]。2015 年，Xinxiong Chen 等人在 CBOW 模型的基础上，利用每个词汇中的汉字位置和形态表示词向量，将 CBOW 模型生成的原始词向量与字向量拼接，在不同数据集上计算词汇的相关性，均有 4%左右的性能提高[6]。2014 年，Yaming Sun 将部首信息引入词汇表示中，在 C&W 模型的基础上，使用 CRF 做中文分词，不同测试集上准确率均有提高[7]。2015 年，Yanran Li 在词汇表示中引入字信息和部首信息，用于词汇的相似度计算，效果较好[8]。2016 年，Yin 等人融合上下文词汇的字信息和部首信息，生成词向量，在词语相似度计算的实验中，比单纯 CBOW 方法提高了近 3%[9]。上述细粒度的词汇表示工作，主要针对普通词汇，而情感词汇作为词汇的一个特殊种类，有一些表达情感的独有特征。

1.2　情感词汇表示的相关工作

上面介绍的近几年细粒度词汇表示方面的研究成果是针对所有词汇的，通常采用词语相似度来验证词向量的有效性。而情感词汇在表示方面有自身的特点，从 20 世纪 90 年代以来，词汇倾向性的研究在国外得到了普遍的关注。Hatzivassiloglou 和 Mc Keown 在 1997 年利用词汇之间的连词(and，or，but，either，or 和 neither，nor 等)训练生成词汇间的同义或反义倾向的连接图，生成褒贬两义的词汇集[10]。2003 年，Turney 和 Littman 采用计算基准词对与词汇相似度的方法识别词汇倾向性[11]。2005 年 Vermeij 等人利用有倾向性的词汇在产品评论中出现的次数计算用户评论的倾向性，提出了一种按词频加权统计的方法[12]。基于上述方法，也构建了一些情感词汇字典[13]，用于语篇倾向性计算[14-15]。2014 年，杨亮等基于图排序做情感词汇消歧[16]，2015 年，乌达巴拉等使用 CRFs 完成短语情感分析[17]。分布式词汇表示出现后，情感词汇的表示也有新的研究成果。Duyu Tang 等人在 C&W 模型基础上，使用带标注的 Twitter 上的短文本构建词向量，将文本的标注信息带入词向量的表示中[18]。2011 年，Bespalov 等人使用 LSA(latent semantic analysis)初始化词汇表示[19]。上述情感词汇的表示主要集中在词语级别，利用词汇的上下文和语句的极性识别词汇的情感极性，而情感词汇也存在一些内部特征，帮助词汇表达情感语义，所以本文尝试在情感词汇表示中加入三种细粒度的特征，增强词汇向量的情感语义。

情感词汇与普通词汇表示不同，不仅需要考虑词汇的相似度，更要考虑词汇的情感极性。细粒度的情感词汇表示方法是否能够有效区分情感极性？情感词汇是否还有其他的有效特征能够区分词汇的情感极性和感性色彩？受到上述细粒度词汇表示和近几年分布式的情感词汇表示方法的启发，本文将中文情感词汇的表示细化到字和部首水平，并借助汉语的音位知识，增强情感词汇的表示能力，提出了一种融合多特征的情感词汇表示模型。该模型在情感词分类方面有较好的实验效果。本文的主要贡献如下：①将部首信息加入到中文情感词汇的表示中，并采用字和部首多种组合方法增强情感词汇的表示能力；②将音位信息加入到情感词汇表示中，将词中每个字的声母、韵母和声调作为特征，加入到情感词汇表示中。

文中第二节介绍了我们的情感词汇分类模型；第三节中，使用前馈神经网络验证第二节中的词汇表示模型的效果；第四节总结了本文工作，并提出今

后工作的设想。

2 情感词汇的表示

分布式词汇表示的常用模型有 C&W[20]，CBOW（continuous bog-of-words）[21] 和 Skip-Gram[21] 等几种方法，这些方法都是通过上下文学习词汇的表示模型。其中 SkipGram 模型是目前应用广泛、在各种任务中表现较好的词向量表示方法，本文选择它作为 Baseline。2.2 节、2.3 节和 2.4 节分别介绍了在 SkipGram 模型基础上加入词汇的字、部首和音位三种信息的模型。

2.1 SkipGram 模型

SkipGram 模型根据目标词汇来预测源词汇，该模型将每个“上下文，目标词汇”的组合作为样本，本文使用的 Negative Sampling 模型，目标优化函数为：

$$L = \log \prod_{w \in C} \prod_{u \in \text{Context}(w)} g(u) \tag{1}$$

其中 $g(u)$ 定义为：

$$g(u) = \prod_{z \in \{u\} \cup \text{NEG}(u)} p(z \mid w) \tag{2}$$

NEG(u) 表示处理词汇 u 时产生的负样本子集，目标函数需采用梯度计算的方法进行优化。SkipGram 模型相对于 CBOW 模型在大型数据集上更为有效。

2.2 部首信息词汇表示

本文在 SkipGram 模型生成的词向量基础上，加入了汉字的部首信息。汉语的部首是表示语义的一个最小单位，也是汉字和词语构成的重要部分。汉字的部首主要起源于东汉许慎《说文解字》一书，他根据字义创建了 540 个部首，通过部首的排序表现篆书字形的意义，所以部首建立之初就具有表义的作用。同一部首的有些汉字具有相同的含义，例如：部首“心”构成的汉字多代表一定的心理活动，包含部首“心”的词汇有“愉快”“憔悴”“悲伤”“反悔”“惭愧”“恬静”“恬淡”和“恬然”等。可见很多部首具有情感色彩，能够在区分情感词汇时起到一定的甄别作用。

表 1 根据部首出现在褒、贬义情感词汇中的次数，将它们划分为褒义部首、贬义部首和中性部首。褒贬词差值是指某个部首出现在褒义词中的次数和出现在贬义词中次数之差的绝对值。当部首出现在褒义词的次数比出现在贬义词中的次数大于等于 15，则将部首统计为褒义部首，反之，为贬义部首。如果小于 15，则认为是中性部首。从表 1 可见，部分部首对情感词汇的识别有帮助作用。以褒贬词差值 15 为例，具有明显褒、贬义的部首均为 113，占据部首总数的一半左右。

表 1 部首的情感相关性

褒贬词差值	褒义部首数	贬义部首数	中性部首
5	108	67	67
10	87	49	106
15	73	40	129

表 2 给出了褒贬义差值为 15 时，具有褒、贬含义的部首。

表 2 褒义和贬义部首示例

褒义部首	贬义部首
宀，八，日，纟，讠，女，力，艹，彳，石，高，广，辶，玉，钅，几，风，戈，干，车，山，斤，口，禾，王，儿，羽，止，士，入，田，寸，水，礻，丨，二，厂，身，立，十，丿，刂，甘，里，革，彐，白，心，采，艮，飞，雨，至，牜，工，手，人，色，灬，米，香，走，鼎，丷，音，罒，文，又，青，比，冂，亠，夕	耳，目，丶，缶，巛，氵，瓦，鸟，面，豕，亅，肉，罒，母，委，弓，廾，皮，黑，舌，门，虍，小，乙，非，马，木，歹，穴，讠，火，鬼，贝，大，虫，一，亻，疒，扌，犭

本文选择两种方式生成词汇对应的部首向量，一种是取最大值的方法，另一种是拼接的方法。无论哪种方法，首先都要获取原始的词汇向量（$\boldsymbol{V}_{\text{wrd}}$）和原始的部首向量（$\boldsymbol{V}_{\text{rad}}$）。采用 2.1 节中的 Skip-Gram 模型训练得到原始的词汇向量，原始的部首向量则通过函数生成符合正态分布的随机向量。下面介绍多个字的部首向量融合成整个词汇的部首向量的方法。

（1）取最大值的方法获取词汇的部首向量：首先，解析出词汇中每个字的部首信息，将部首对应的向量按分量取最大值，得到词汇对应的部首向量。矩阵 $\boldsymbol{R}_{n \times m}$ 表示词汇 w 对应的部首矩阵，其中 n 表示词汇中包含的字数，m 表示原始部首向量的长度：

$$\boldsymbol{R}_{n \times m} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_i, \cdots, \boldsymbol{r}_n)^{\mathrm{T}} \tag{3}$$

向量 $\boldsymbol{r}_i$ 表示词汇中第 i 个字对应的部首向量 $\boldsymbol{r}_i \in \boldsymbol{V}^{\text{rad}}$，词汇的部首向量（$z_{\text{rad}}$）的分量 z_i 为：

$$z_i = \max_{j=1}^{m}(a_{ji}), \quad a_{ji} \in \boldsymbol{r}_i \tag{4}$$

通过上述方法得到词汇的部首向量后与词汇的原始向量拼接，采用前馈神经网络分类，最后通过

softmax 层得到最终的分类结果。图 1 详细描述了整个模型的情感词汇分类的过程。

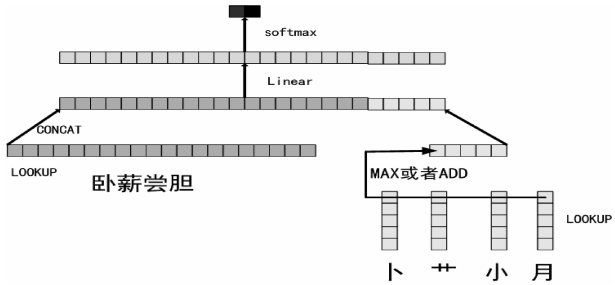

图 1　MAX 方法的词汇表示模型

(2) 多部首拼接的方法获取词汇的部首向量：首先，解析出词汇中每个字的部首信息，将词中多个字的部首拼接得到词汇对应的部首向量。矩阵 $\boldsymbol{R}_{n\times m}$ 表示词汇 w 对应的部首矩阵，其中 n 表示词汇中包含的字数，m 表示原始部首向量的长度：

$$\boldsymbol{R}_{n\times m}=(\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_i,\cdots,\boldsymbol{r}_n)^{\mathrm{T}} \tag{5}$$

向量 $\boldsymbol{r}_i$ 表示词汇中第 i 个字对应的部首向量 $\boldsymbol{r}_i\in\boldsymbol{V}^{\mathrm{rad}}$，词汇的部首向量为：

$$z_{\mathrm{rad}}=\mathrm{CONCAT}_{i=1}^{n}(\boldsymbol{r}_i) \tag{6}$$

一个词汇中包含的字数可能不同，为了使每个词汇的部首向量长度相等，模型选取词汇中最后两个字的部首信息。所以实际系统中，上述公式中 i 的取值为：$n-1$ 和 n。多部首拼接的方法分类过程如图 2 所示。

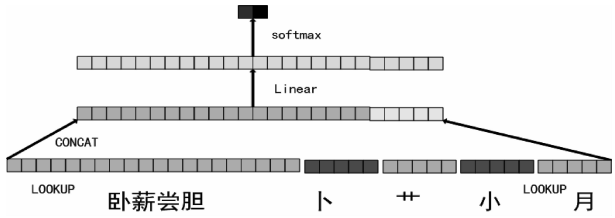

图 2　CONCAT 方法的词汇表示模型

通过与原始词汇向量拼接，可以得到如下新的词汇向量，同样采用单层前馈神经网络分类，其中损失函数使用交叉熵损失函数：

$$L_{\text{cross-entropy}}(Y',Y)=-\sum_i y_i\log(y_i') \tag{7}$$

2.3　按字切分的情感词汇表示

中文的词汇由汉字组成，汉字的意义在一定程度上表示了词汇的情感含义，例如“快乐”中的“乐”，“悲伤”中的“悲”，都表示了词汇的基本情感倾向。所以我们尝试将语料按字分割，首先生成词汇的字向量，多个原始字向量拼接得到词汇的字向量，然后再与词汇的原始向量连接，得到最终的词汇向量，如式(8)所示。

$$\boldsymbol{v}_w^{\mathrm{new}}=\mathrm{CONCAT}(\boldsymbol{v}_w,\boldsymbol{c}_w) \tag{8}$$

其中 $\boldsymbol{v}_{\mathrm{w}}^{new}$ 为拼接后的词汇向量，$\boldsymbol{v}_{\mathrm{w}}$ 表示原始词向量，$\boldsymbol{c}_w$ 表示由词汇中的字信息生成的词汇字向量。本文采用两种方法生成原始字向量，一种是用原始语料按字分词训练得到，另一种是符合正态分布的随机向量。最后，新生成的词汇向量 $\boldsymbol{v}_w^{\mathrm{new}}$，使用单层前馈网络和交叉熵损失函数分类。

2.4　音位信息的情感词汇表示

音位是从一定言语连续体的众多音素中归纳出来的能区别语素的最小语音单位。音位标音是用符号把归纳出来的音位标示出来，使它成为可读的东西。汉字是包含音节的语素文字，有时候是“因形以得其音，因音以得其义[22-23]。汉语词汇是一个音形义的统一整体，“汉语的声调语言，调形曲线遵循一定之规，具有区别词义的作用”[24]。一般传统的方法把语音归纳为声母、韵母和声调三部分[25]。目前，有人认为：“用什么方法来归纳普通话音位系统，有两种针锋相对的意见。一是按元音辅音系统归纳，得出的是元音音位、辅音音位和声调音位；一是主张按声韵调体系归纳，得出的是声位(声母音位)、韵位(韵母音位)和调位(声调音位)”[26]，但无论哪种归纳方法，声母、韵母和声调都是其中的重要部分。不同极性的情感词汇在不同音位上的音调也有所差别，以韵母“ai”为例，它与消极情感相关性较大，如“哀怨”“灾难”“歪曲”“痴呆”。

获取词汇音位向量的方法与 2.2 节和 2.3 节类似，首先得到情感词汇中每个字的音位信息，本文采用词汇中最后两个字的声母、韵母(同一韵母声调不同用不同的向量表示)四部分信息组成音位向量。首先为不同的声母和不同声调的韵母分配向量，向量还是使用符合正态分布的随机向量。先将每个字不同声母向量和韵母向量拼接，生成音位向量，然后将情感词汇中不同字的音位连接，最后添加原始词汇向量如式(9)所示。

$$\boldsymbol{v}_{\mathrm{w}}^{\mathrm{new}}=\mathrm{CONCAT}(\boldsymbol{v}_{\mathrm{w}},\boldsymbol{o}_1,\boldsymbol{r}_1,\boldsymbol{o}_2,\boldsymbol{r}_2) \tag{9}$$

其中 $\boldsymbol{o}_1,\boldsymbol{r}_1,\boldsymbol{o}_2,\boldsymbol{r}_2$ 分别表示两个字中的声母向量和韵母向量。

3　实验及分析

3.1　实验语料及设计

本文采用了两个数据集训练原始的词向量，一

个是维基百科中文语料(WikiData),共1GB大小。另一个是从数据堂上获取的200万条微博情感句(WeiboData)①,共97.7M。使用Word2vec②,训练基本的词向量,词向量的长度为200。情感词汇选取大连理工大学信息检索实验室的情感词汇本体[15],选取2 521个积极情感词汇和1 776个消极情感词汇,共4 297个词汇做情感分类。80%作为训练集,20%作为测试集。部首和声调采用GB2312字库中的信息,包括汉字的字、音、义、部首和笔画等信息③。为了验证加入特征后词向量的有效性,本文还选取COAE2014任务4提供的5 000条微博语句作为语料集,完成语句级的分类。

实验主要分为三部分:首先在不同语料上实验原始词向量的分类结果,找到一个准确率较高的Baseline。然后在此基础上分别添加字、部首和音位信息,检验细粒度的特征是否能提高情感词汇的分类结果。最后在COAE评测的语句级分类任务中验证细粒度情感词汇表示模型的有效性。

3.2 实验结果

3.2.1 不同语料训练词向量的分类效果

本实验分别采用维基百科和微博情感句两个语料训练词向量,对情感词汇分类,结果如表3所示。

表3 不同维度及领域语料的分类效果

语料	向量长度	分类方法	准确率/%
维基百科(1GB)	400	单层前馈神经网络	80.60
维基百科(1GB)	200	单层前馈神经网络	81.43
微博情感(97MB)	200	单层前馈神经网络	78.93
维基百科十等分	400	单层前馈神经网络	73.33
维基百科十等分	200	单层前馈神经网络	79.54

和维基百科的语料比,微博情感语料里包含的情感词汇更多,与情感领域更相关,但是分类的准确率比维基百科语料低了2.5%左右。两个语料的大小差异较大,为了验证语料的大小对词向量的影响,我们将维基百科语料分为10份,每份大约100MB,分别训练词向量,用于情感词汇分类。结果显示当语料大小相当时,如果向量长度取200,维基百科和微博情感语料效果基本相近,相差0.5%左右,都低于1GB维基百科的训练结果。可见语料大小对词向量的质量有影响。另外,在10等分情况下,400长的词向量效果远小于200长的词向量,可见训练语料规模较小时,长度过大,会影响词向量效果。

另外,表3还对比了不同维度词向量在情感词汇分类中的效果,向量长度由200增加到400,词向量的分类准确率相差不大,400维的准确率略低于长度为200的词向量。

3.2.2 部首和音位信息等对分类结果的影响

通过随机正态分布为242个部首和96个声母、韵母和声调信息分配长度为50的向量,根据字和部首的对应表,以及字和音位对应表,将部首向量和音位向量连接在每个词向量后,词向量和字向量选择维基百科语料的训练结果。因为一个词中包含多个字,每个字都有部首和声调,所以本文尝试了多种向量的组合方式,获取整个词汇的字、部首和音位向量的方法主要有以下几种:词中多个部首向量按分量取最大值(WOMAXRAD);词中多个部首向量拼接(WOCONRAD);前两个字的部首向量连接(WOSERAD);最后两个字的韵母向量连接(WOVOWEL);最后两个字的声母和韵母向量连接(WOPHONEME);语料训练生成的字向量(WOCHA);随机生成字向量(WOCHARANDOM)。以上获取的向量再与原始词汇向量连接。另外还尝试了去除原始词汇向量的几种方法:单独字向量连接(CHACONCAT);单独部首向量(RADICAL)和单独词向量(WORD)。最后尝试了词、字、部首和音位向量拼接(WOCHARADPHOCAT);词、字、部首和音位向量按列取最大值(WOCHARADPHOMAX)。

本文采用单层前馈神经网络实现词汇的情感分类。将字信息、部首信息和音位信息依次加入到原始词向量中。表4列出每个特征对分类结果的影响。

从实验结果看,添加语料训练的字信息,分类结果提高了4%;添加部首信息,能在原始词向量的基础上提高2%;添加音位信息,分类结果能提高1.5%。将字、部首和音位信息同时融合到原始词向量中,效果最好,比单纯的词向量提高了5.3%。这里的字向量是通过语料训练得到的,不是随机生成的。如果采用随机生成的字向量与词汇向量拼接

① http://www.datatang.com/datares/go.aspx?dataid=619757

② http://code.google.com/p/word2vec

③ http://more.datatang.com/data/44078

(WOCHARANDOM)，分类的效果与单纯使用词向量几乎没有差别。可能部首信息和音位信息是随机生成，所以效果没有字向量好，如果能找到训练部首和音位向量的方法，结果可能更好。但无论哪种方法，都对原始词向量有补充作用，提高了情感词汇分类的准确率。

表 4　不同特征的分类效果

表示模型	词向量的生成方式	准确率/%
字、部首和音位向量融合模型	WOCHARADPHOCAT	**86.79**
	WOCHARADPHOMAX	86.19
字向量模型	WOCHA	**85.60**
	CHACONCAT	78.21
	WOCHARANDOM	81.31
部首向量模型	WOMAXRAD	**83.69**
	WOSERAD+ WOPHONEME	83.69
	WOCONRAD	83.33
	WOSERAD	83.21
	RADICAL	57.14
音位向量模型	WOPHONEME	**82.98**
	WOVOWEL	82.26
原始词向量	WORD	**81.43**

单纯使用部首或者字信息分类，比单纯词向量(Baseline)降低了 24%。因为缺少基础词向量的信息，部首向量也是随机生成的，对词汇的表示能力不强，所以效果较差。在多种添加部首信息的方法中，词中每个部首向量按分量取最大值的方法效果最好。在多种添加音位信息的方法中，包含的信息较全面的方法是多个字的声母和韵母向量拼接，准确率较高。

汉字能表达含义，但可能单独的字形体上所体现的意义并非是单一的，孕育着表示一个以上汉语词义的能力[27]。单纯考虑字向量会丢失词语组合的信息，导致歧义较大。所以当去除原始词向量，单纯使用字向量分类时，比单纯的词向量准确率低了 3%。

除了上述的单层神经网络，本文还尝试了其他神经网络模型，选择表 4 中的词向量与字向量连接的模型(WOCHA)，分别使用单层神经网络、多层神经网络和卷积神经网络三种方法，做了几组对比实验，结果如表 5 所示。

表 5　多种神经网络分类效果比较

词向量的生成方式	分类方法	准确率/%
词向量与字向量融合(WOCHA)	单层神经网络	**85.60**
	卷积神经网络	81.19
	多层神经网络	57.12

从实验结果看，多层神经网络比单层神经网络低了近 30%。卷积神经网络也比前馈神经网络准确率低了 5%左右。另外，在实验中还发现，卷积神经网络的过滤器越多，池化范围越大，效果越低。

3.2.3　词汇表示模型在句子分类中的效果

为了验证情感词汇表示模型的有效性，本文选取 COAE2014 任务 4 提供的 5 000 条微博语句作为语料集。这些句子分别标注为褒、贬两义，实验选择 1 666 个句子作为训练集，3 334 个句子作为测试集。本文采用两种方法生成句子向量：一个是句子中多个词汇按分量取最大值的方法(MAX)，另一个是多个词汇取平均的方法(AVERAGE)，实验结果如表 6 所示。

表 6　情感句的识别效果

生成语句向量的方法	词向量	准确率/%
多个词汇向量按分量取最大值的方法(MAX)	WOCHA	**89.25**
	WOPHONEME	87.62
	WOMAXRAD	87.53
	WORD	87.17
多个词汇向量取平均的方法(AVERAGE)	WOCHA	**90.81**
	WORD	89.55
	WOMAXRAD	89.34
	WOPHONEME	89.25

在 MAX 方法生成语句向量中，添加字信息的词向量比单纯词向量的准确率增加 2%左右。采用 AVERAGE 方法，语句分类效果更好，比原始词向量也提高 1.5%左右。可见，添加了字信息的情感词汇向量在句子分类中的效果较好。

3.2.4　情感词汇和非情感词汇的分类结果

除了情感词汇的褒、贬极性分类，我们还将上述的情感词汇表示方法用在情感词汇和非情感词汇的识别中，为了使正例数与负例数平衡，选择情感词汇 4 295 个，非情感词汇 5 000 个，实验结果如表 7 所示。

表 7 非情感词汇的识别效果

表示模型	词向量	准确率/%
字向量模型	WOCHA	**89.84**
部首向量模型	WOMAXRAD	89.18
音位向量模型	WOPHONEME	88.97
原始词向量	WORD	87.77

从实验结果看，添加字向量、部首向量和音位向量，对情感词汇的识别效果都有提高，其中加入训练后的字向量效果最好，比单纯词汇向量的分类结果提高 2%，可见添加三种细粒度特征能够提高情感词汇和非情感词汇的分类准确率，从而能在句子和篇章中，更高效地识别出情感词汇。

3.3 实验分析

通过上述的实验结果可以得出以下几点结论：①在情感词汇分类中，训练情感词汇的语料规模更重要；②添加字、部首和音位信息都能有效地提高情感词汇的极性的分类效果，其中字、部首和音位信息与词汇信息融合的模型效果最好；③添加字、部首和音位信息有助于情感词汇和非情感词汇的分类结果；④加入三种细粒度特征的情感词汇，在 COAE 评测语料的语句级情感分类中也有较好效果。

4 结论与展望

本文将字、部首和音位信息加入词向量的表示中，不同的词向量表示方式，分别使用前馈神经网络和卷积神经网络两种分类方法，完成情感词汇的极性分类。实验结果表明，字、部首和音位信息包含一定的情感含义，能有效区分情感词汇的极性。汉字是图形表意的，未来可以在词向量中尝试添加“形声字”和“会意字”的信息，也可以考虑寻找一些合理的方法来预训练原始部首和音位向量，分类方法上可以尝试多个分类方法的融合。

参考文献

[1] Chrupala Grzegorz. Text segmentation with character-level text embeddings[C]//Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing,2013.

[2] Cicero Nogueira dos Santos, Bianca Zadrozny. Learning character-level representations for part of speech tagging[C]//Proceedings of the 31st International Conference on Machine Learning, 2014.

[3] Cicero Nogueira dos Santos, Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts[C]//Proceedings of the 25th International Conference on Computational Linguistics,2014: 69-78.

[4] Yue Zhang, Stephen Clark. A fast decoder for joint word segmentation and POS tagging using a single discriminative model[C]//Proceedings of the EMNLP, 2010: 843-852.

[5] Meishan Zhang, Yue Zhang, Wanxiang Che, et al. Chinese parsing exploiting characters[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,2013(1): 125-134.

[6] Xinxiong Chen, Lei Xu, Zhiyuan Liu, et al. Joint learning of character and word embeddings[C]//Proceedings of the 24th International Conference on Artificial Intelligence,2015: 1236-1242.

[7] Yaming Sun, Lei Lin, Nan Yang, et al. Radical-enhanced chinese character embedding[C]//Proceedings of the 21st International Conference on Neural Information Processing,2014: 279-286.

[8] Yanran Li, Wenjie Li, Fei Sun, et al. Component-enhanced Chinese character embeddings[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 829-834.

[9] R Yin, W Quan, L Rui, et al. Multi-granularity chinese word embedding[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 981-986

[10] Hatzivassiloglou V, Mc Keown K R. Predicting the semantic orientation of adjectives[C]//Proceedings of ACL297,35th Annual Meeting of the Association for Computational Linguistics,1997: 174-181.

[11] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003,21(4): 315-346.

[12] M J, Vermeij M. The orientation of user options through advers, verbs and nouns[C]//Proceedings of the 3rd Twente Student Conference on IT,2005.

[13] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008,27(2): 180-185.

[14] 徐琳宏，林鸿飞，杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报，2007,21(1): 96-100.

[15] 徐琳宏，林鸿飞. 基于语义特征和本体的语篇情感计算[J]. 计算机研究与发展，2007(Z2): 356-360.

[16] 杨亮，张绍武，林鸿飞，等. 基于图排序的词汇情感消歧研究[J]. 中文信息学报，2014,28(6): 129-136.

[17] 乌达巴拉，汪增福. 一种扩展式 CRFs 的短语情感倾向性分析方法研究[J]. 中文信息学报，2015，29(1)：155-162.

[18] Duyu Tang, Furu Wei. Learning sentiment-specific word embedding for twitter sentiment clssification [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014：1555-1565.

[19] Bespalov Dmitriy, Bai Bing, Qi Yanjun, et al. Sentiment classification based on supervised latent n-gram analysis[C]//Proceedings of the Conference on Information and Knowledge Management, 2011：375-382.

[20] R Collobert，J Weston. A unified architecture for natural language processing：Deep neural networks with multitask learning [C]//Proceedings of the ICML, 2008.

[21] Tomas Mikolov，Kai Chen，Greg Corrado and Jeffrey Dean. Efficient estimation of word representations in vector space[C]//Proceedings of the ICLR Workshop Track，2013.

[22] 段玉裁. 说文解字注[M]. 北京：中华书局，2013.

[23] 王世华. 文字假借不是词义引申[J]. 中国语文，2003(5)：477-478.

[24] 曹剑芬. 汉语声调与语调的关系[J]. 中国语文，2002(3)：195-286.

[25] 陈其光. 音位标音的几种选择[J]. 中国语文，1994(4)：266-273.

[26] 王理嘉. 音位归纳的多重可能性[J]. 汉语学习，1988(3)：1-7.

[27] 董琨. 汉语的词义蕴含与汉字的兼义造字[J]. 中国语文，1994(3)：226-230.

徐琳宏(1979—)，硕士，讲师，主要研究领域为文本情感计算。
E-mail：qingniao1203@163.com

林鸿飞(1962—)，博士，教授，主要研究领域为文本挖掘和信息检索。
E-mail：hflin@dlut.edu.cn

祁瑞华(1974—)，博士，教授，主要研究领域为自然语言处理。
E-mail：rhqi@dlufl.edu.cn