Journal of Computer Applications
计算机应用, 2020, 40(3): 651 - 657
ISSN 1001-9081
CODEN JYIIDU
2020-03-10
http://www.joca.cn

文章编号:1001-9081(2020)03-0651-07
DOI:10.11772/j.issn.1001-9081.2019071210

融合基于语言模型的词嵌入和多尺度卷积神经网络的情感分析

赵亚欧[1,2*],张家重[1],李贻斌[3],付宪瑞[1],生　伟[1]
(1. 浪潮集团 金融信息技术有限公司,济南 250101; 2. 济南大学 信息科学与工程学院,济南 250022;
3. 山东大学 控制科学与工程学院,济南 250061)
(*通信作者电子邮箱 zhaoyaou@inspur.com)

摘　要:针对Word2Vec、GloVe等词嵌入技术对多义词只能产生单一语义向量的问题,提出一种融合基于语言模型的词嵌入(ELMo)和多尺度卷积神经网络(MSCNN)的情感分析模型。首先,该模型利用ELMo学习预训练语料,生成上下文相关的词向量;相较于传统词嵌入技术,ELMo利用双向长短程记忆(LSTM)网络融合词语本身特征和词语上下文特征,能够精确表示多义词的多个不同语义;此外,该模型使用预训练的中文字符向量初始化ELMo的嵌入层,相对于随机初始化,该方法可加快模型的训练速度,提高训练精度;然后,该模型利用多尺度卷积神经网络,对词向量的特征进行二次抽取,并进行特征融合,生成句子的整体语义表示;最后,经过softmax激励函数实现文本情感倾向的分类。实验在公开的酒店评论和NLPCC2014 task2两个数据集上进行,实验结果表明,在酒店评论数据集上与基于注意力的双向LSTM模型相比,该模型正确率提升了1.08个百分点,在NLPCC2014 task2数据集上与LSTM和卷积神经网络(CNN)的混合模型相比,该模型正确率提升了2.16个百分点,证明了所提方法的有效性。

关键词:情感分析;自然语言处理;卷积神经网络;ELMo;字向量
中图分类号:TP183　　**文献标志码**:A

Sentiment analysis using embedding from language model and multi-scale convolutional neural network

ZHAO Ya'ou[1,2*], ZHANG Jiachong[1], LI Yibin[3], FU Xianrui[1], SHENG Wei[1]
(1. *Inspur Financial Information Technology Company Limited*, *Jinan Shandong* 250101, *China*;
2. *School of Information Science and Engineering*, *University of Jinan*, *Jinan Shandong* 250022, *China*;
3. *School of Control Science and Engineering*, *Shandong University*, *Jinan Shandong* 250061, *China*)

Abstract: Only one semantic vector can be generated by word-embedding technologies such as Word2vec or GloVe for polysemous word. In order to solve the problem, a sentiment analysis model based on ELMo (Embedding from Language Model) and Multi-Scale Convolutional Neural Network (MSCNN) was proposed. Firstly, ELMo model was used to learn the pre-training corpus and generate the context-related word vectors. Compared with the traditional word embedding technology, in ELMo model, word features and context features were combined by bidirectional LSTM (Long Short-Term Memory) network to accurately express different semantics of polysemous word. Besides, due to the number of Chinese characters is much more than English characters, ELMo model is difficult to train for Chinese corpus. So the pre-trained Chinese characters were used to initialize the embedding layer of ELMo model. Compared with random initialization, the model training was able to be faster and more accurate by this method. Then, the multi-scale convolutional neural network was applied to secondly extract and fuse the features of word vectors, and generate the semantic representation for the whole sentence. Experiments were carried out on the hotel review dataset and NLPCC2014 task2 dataset. The results show that compared with the attention based bidirectional LSTM model, the proposed model obtain 1.08 percentage points improvement of the accuracy on hotel review dataset, and on NLPCC2014 task2 dataset, the proposed model gain 2.16 percentage points improvement of the accuracy compared with the hybrid model based on LSTM and CNN.

Key words: sentiment analysis; Natural Language Processing (NLP); Convolutional Neural Network (CNN); Embedding from Language Model (ELMo); character embedding

0　引言

情感分析(Sentiment Analysis)是对带有情感色彩的主观性文本或句子进行分析、处理和抽取的技术[1]。随着互联网技术的发展,人们的活动越来越多地集中在网络上,人们通过网络进行学习、交流、购物、娱乐,同时对社会热点事件、热门

收稿日期:2019-07-11;**修回日期**:2019-09-07;**录用日期**:2019-09-09。
基金项目:国家重点研发计划云计算和大数据重点专项(2016YFB1001100, 2016YFB1001104); 国家自然科学基金青年项目(61702218)。
作者简介:赵亚欧(1982—),男,山东济南人,讲师,博士,CCF会员,主要研究方向:自然语言处理、人工智能; 张家重(1965—),男,山东日照人,教授,博士,主要研究方向:人工智能、数据挖掘; 李贻斌(1960—),男,山东聊城人,教授,博士,主要研究方向:机器人、人机交互; 付宪瑞(1986—),男,山东济南人,工程师,主要研究方向:软件架构、人工智能; 生伟(1983—),男,山东济南人,信息系统项目管理师(高级),主要研究方向:人工智能、金融自助终端。

商品和相关服务进行评论。通过这些评论挖掘数据背后人们的观点、倾向是十分重要的，利用这些数据，政府部门能够及时作出反应，进行舆论引导，避免重大舆情事件；商户能够掌握用户需求，进行个性化精准营销；产品制造商也可以了解产品优劣，及时对产品进行改进。

从评论中自动提取用户观点并不容易，现在主流的方法主要有有监督和无监督两类。有监督方法主要是利用机器学习技术，如支持向量机（Support Vector Machine，SVM）、最大熵方法和朴素贝叶斯方法等[2-4]，对文本进行学习，然后进行情感分类。无监督方法主要是分析文本中的情感词、语法和语义，通过抽取文本的统计特征实现情感分类。

深度学习是人工智能的一个热点，其方法也被广泛应用在情感分析领域。Socher等[5]于2011年提出使用递归自编码网络（Recurrent Auto-Encoder，RAE）对文本特征进行抽取和分类。随后，Qian等[6]提出使用动态卷积神经网络（Dynamic Convolutional Neural Network，DCNN）进行文本分类，DCNN采用动态K-Max池化操作，能够有效地捕捉词语之间的联系。Wang等[7]提出利用长短程记忆（Long Short-Term Memory，LSTM）网络对推特文本进行情感分类，该方法利用LSTM的门控结构，能够更好存储文本特征。LSTM随后被扩展为双向LSTM、叠层双向LSTM等一系列模型。此外，多种网络结构的融合方法也被提出，如Wang等[8]提出的卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent Neural Network，RNN）相结合的模型CNN-RNN、Guggilla等[9]提出的LSTM-CNN模型等。Akhtar等[10]于2017年提出利用多个神经网络集成进行情感分类，也取得了很好的效果。

注意力机制最早应用于图像领域，它可以聚焦图像的特定区域，抽取图像的有用特征。2016年，Bahdanau等[11]最先把注意力机制应用到自然语言处理任务中，构造了当时性能最好的机器翻译模型。随后，研究人员也把注意力机制应用到情感分析任务中。曾锋等[12]将注意力机制和循环神经网络相结合，通过双层注意力分别对单词层和句子层进行建模，捕获不同单词和不同句子的重要性。曾碧卿等[13]将注意力机制和卷积神经网络相结合，提出一种双注意力机制的卷积神经网络模型，用来确定情感倾向。韩萍等[14]提出了基于情感融合和多维自注意力机制的微博文本情感分析模型，实现了文本中词语间依赖关系的建立以及多角度情感语义信息的获取。此外，石磊等[15]将注意力机制与树形结构的LSTM网络相结合，提升情感分析的准确率。

使用深度学习技术对自然语言进行处理，另外一个重要的问题是如何将文字符号转化为数字特征。前人的方法大都利用Word2Vec、GloVe（Global Vectors）等词嵌入工具得到词语的嵌入向量（embedding）。但此类方法存在的主要问题是难以对多义词进行向量表示。比如苹果，由于上下文不同，可能表示的是一种水果，也可能是指苹果公司，甚至可能是指电影名称，然而，利用词嵌入技术，这三个不同的语义只会被表示为一个向量。为了解决这个问题，本文提出使用基于语言模型的词嵌入（Embedding from Language Model，ELMo）获取词向量。该模型获得的词向量不但包含词语本身信息，也包含其对应的上下文信息，能够表达多义词的不同语义。此外，针对中文环境下ELMo难以训练的问题，本文提出使用预训练获得的字符向量初始化ELMo的嵌入层，可加快模型的训练，提高训练精度。最后，针对情感分类，本文提出使用融合ELMo和多尺度卷积神经网络（Multi-Scale Convolutional Neural Network，MSCNN）作为分类器。该模型使用不同尺度的卷积核对文本施加卷积操作，并利用最大值池化操作对获得的特征进行筛选，最后将这些特征融合，作为最终文本的特征。该方法考虑了不同尺度的短语特征，相较于单纯采用词向量或者字向量，效果更好。

1 相关工作

自然语言处理的核心任务是抽取语言文字表象符号背后的隐藏含义，得到计算机可以理解的数据化表征。一个好的数字化表征对于后续自然语言处理任务，如情感分析、语义分析、机器翻译等十分重要，是自然语言处理研究的热点。方法从早期的独热编码（One-hot），到词袋（Bag-of-words）模型，再到后来的tf-idf（term frequency-inverse document frequency），不断涌现。

最近几年，随着深度学习技术的发展，越来越多的学者尝试使用神经网络方法抽取词语的特征：2003年，Bengio等[16]提出了NNLM（Neural Network Language Model），利用三层神经网络学习语言模型，这是神经网络在自然语言表征方面的首次尝试；随后在2013年，谷歌的Mikolov等[17]在CBow（Continuous Bag-of-words）和Skip-gram模型的基础上构建了Word2vec工具，该模型使用两层神经网络，去掉了隐含层，简化了神经网络结构，并且使用噪声对比估计（Noise-Contrastive Estimation，NCE）和层级Softmax（Hierarchical Softmax）技术，减小了算法复杂度，使神经网络的大规模应用成为可能；同一时期，斯坦福的自然语言处理小组的Pennington等[18]提出GloVe算法，也取得了不错的效果。

随着研究的不断深入，开始从研究词语表征转向研究句子表征。2015年Kiros等[19]提出Skip-thoughts方法。该方法首先利用循环神经网络（RNN）对句子进行编码，然后构造另外两个RNN对句子进行解码，通过编码-解码模型获取句子向量。2018年，Logeswaran等[20]在此基础上提出quick-thoughts方法，该方法去掉了解码过程，直接将句子的编码作为特征接入后续网络，简化了网络结构，提高了模型性能。

虽然句子表征取得了一定进展，但效果往往不佳。2017年，McCann等[21]提出了句子特征和词语特征的融合方法。首先在机器翻译语料上训练编码器-解码器（Encoder-Decoder）模型，然后抽取编码器的输出层和嵌入层（embedding），其中，编码器的结果作为句子特征，嵌入层作为句子中词语的特征，最后将两者融合作为最终特征。该方法的意义在于，对于每一个词语特征，都融入了该词语所在的句子特征，语义表示更加准确。

相对于词语，句子的复杂度更大，需要大规模语料作为训练集，但大多数自然语言处理任务的语料规模都很小。为了解决这个问题，人们开始探索预训练的方法。2018年Cer等[22]提出在大规模语料中预训练句子向量，然后通过迁移学习技术应用到小规模任务中，但这些方法仅考虑了句子特征，而没有考虑句中的词语特征。

2 ELMo

Peters等[23]于2018年提出了ELMo。该方法结合了上述模型的优点，既采用预训练的方式，又注意融合词语和句子特征。算法主要分为两步：第一步是构建基于LSTM的双向语

言模型(bilateral Language Model, biLM),并在大规模语料上进行训练,获取模型参数;第二步是将文本输入biLM,抽取biLM的输入层和隐含层,将其进行加权组合,获得文本的ELMo特征向量。

2.1　基于LSTM的双向语言模型

和Word2vec的思想类似,ELMo也是通过构造语言模型获取词向量的。一个含有N个词的句子$S=\{t_1,t_2,\cdots,t_N\}$,其出现概率$P(t_1,t_2,\cdots,t_N)$可以通过计算每个词t_k的出现概率$P(t_k)$得到,而$P(t_k)$只与t_k前面出现的词语$t_1,t_2,\cdots,t_{k-1}$有关,因此,$P(t_1,t_2,\cdots,t_N)$可以使用式(1)计算:

$$P(t_1,t_2,\cdots,t_N)=\prod_{k=1}^{N}P(t_k|t_1,t_2,\cdots,t_{k-1}) \tag{1}$$

如果使用LSTM对语言模型进行建模,则t_k对应LSTM隐状态$\boldsymbol{h}_k$。如果LSTM存在L层细胞单元(Cell),则对应隐状态集合为$\{\boldsymbol{h}_k^1,\boldsymbol{h}_k^2,\cdots,\boldsymbol{h}_k^L\}$。将最后一层的隐状态$\boldsymbol{h}_k^L$输入softmax层获得输出$\boldsymbol{o}^k$,$\boldsymbol{o}^k$代表语言模型中$t_k$出现概率$P(t_k|t_1,t_2,\cdots,t_{k-1})$。

t_k出现概率不但与其前面的词有关,也可能与其后面的词有关,因此需要构建后向模型。后向模型和前向模型一样,也采用一个L层的LSTM,将$t_N,t_{N-1},\cdots,t_k$依次输入网络,得到隐状态集合$\{\boldsymbol{h}_k'^1,\boldsymbol{h}_k'^2,\cdots,\boldsymbol{h}_k'^L\}$。将隐状态$\boldsymbol{h}_k'^L$输入softmax层得到输出$\boldsymbol{o}'^k$、$\boldsymbol{o}'^k$表示后向模型中$t_k$出现概率$P(t_k|t_{k+1},t_{k+2},\cdots,t_N)$。

连接前向LSTM和后向LSTM,抽取最后一层的隐状态$\boldsymbol{h}_k^L$、$\boldsymbol{h}_k'^L$,形成向量$\boldsymbol{H}_k^L=[\boldsymbol{h}_k^L,\boldsymbol{h}_k'^L]$,统一送入softmax层得出输出$\boldsymbol{O}^k$。$\boldsymbol{O}^k$为$t_k$的上下文条件概率,即$P(t_k|t_1,t_2,\cdots,t_{k-1},t_{k+1},\cdots,t_N)$。

ELMo所采用的biLM模型如图1所示,从图中可以看出,biLM的核心是两个LSTM网络,两个网络都由多层cell组成(一般采用二层结构),一个网络负责前向语言模型的建模,一个负责后向语言模型的建模。为了保持网络训练稳定,在两层cell之间加入残差连接。最终层的隐状态融合前后两个网络输出,通过softmax计算上下文条件概率。

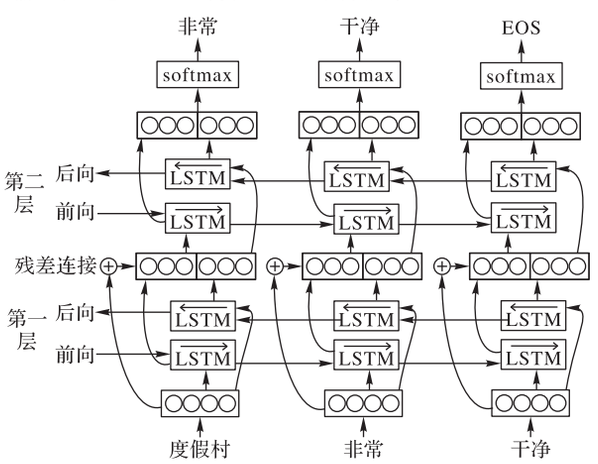

图1　基于LSTM的双向语言模型(biLM)架构

Fig. 1　Architecture of bilateral language model based on LSTM

模型采用的损失函数为句子中所有词t_k对应概率乘积的似然,即:

$$l=\sum_{k=1}^{N}\ln P(t_k|t_1,t_2,\cdots,t_{k-1};\ \boldsymbol{\theta})+\ln P(t_k|t_{k+1},t_{k+2},\cdots,t_N;\ \boldsymbol{\theta}') \tag{2}$$

其中:$\boldsymbol{\theta}$、$\boldsymbol{\theta}'$分别对应前向LSTM模型和后向LSTM模型的待优化参数。

2.2　ELMo特征向量

将后续分类任务中的文本输入biLM模型,获取biLM每一层隐状态$\boldsymbol{H}_k^1,\boldsymbol{H}_k^2,\cdots,\boldsymbol{H}_k^L$。将$t_k$对应的词嵌入向量$\boldsymbol{x}_k$与每一层隐状态进行线性组合,获得$t_k$的ELMo表示,其计算公式为:

$$\boldsymbol{ELMo}_k=\gamma\left(s_0\boldsymbol{x}_k+\sum_{j=1}^{L}s_j\boldsymbol{H}_k^j\right) \tag{3}$$

其中:γ为缩放因子,s_j为归一化的系数,表示每个特征的占比,这些参数都需要在后续任务中进行二次训练获得。

2.3　基于字符的卷积神经网络

在实际使用过程中,ELMo的输入既可以是词向量,也可以是字符向量。如果输入的是字符向量,则需要额外添加一个基于字符的卷积神经网络(character Convolutional Neural Network, char-CNN),用其生成字符对应单词的嵌入向量。其对应的结构如图2所示。

这样做的优点是:1)能够避免词典外词语无法表示的问题(Out Of Vocabulary, OOV);2)无需存储词典中词语的词嵌入向量,只需存储char-CNN的模型参数即可,减小了存储空间占用。

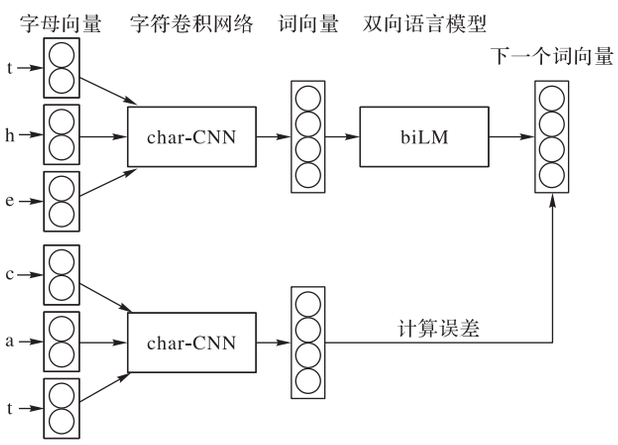

图2　char-CNN模型结构

Fig. 2　Architecture of char-CNN model

3　ELMo和多尺度卷积神经网络融合模型

3.1　基于预训练字向量的ELMo

虽然采用基于字符的卷积神经网络有诸多好处,但如果直接迁移到中文语料中,还存在一定的问题。主要是因为英文字符的初始向量一般采用随机初始化的方式产生,但英文仅有26个字母,加上特殊符号,也不会超过256个。然而,中文系统中汉字数量一般都在5 000以上,如果也采用和英文相同的随机初始化方式,则向量的不确定性太大,后续char-CNN网络将难以训练。

为了解决这个问题,本文提出使用预训练的中文字符向量来初始化ELMo。该方法首先在大规模语料中预训练中文字符向量,然后将训练得到的字符向量作为char-CNN模型的初始向量。这种方式预先加入了汉字的语义信息,与随机初始化相比,不确定程度大大降低,有助于加快模型的训练,提高训练精度。

该算法分为两步,第一步利用Word2vec工具对汉字字符进行预训练,获取字符向量,其模型结构如图3所示。

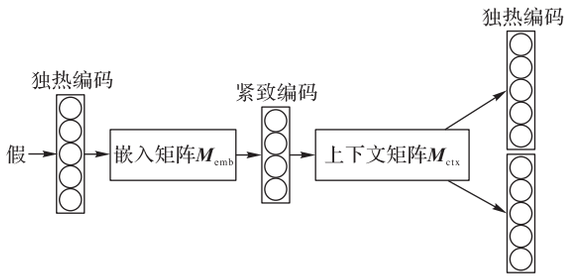

图3 汉字字符向量预训练模型

Fig. 3 Pre-training model for Chinese characters

具体做法是，对于第i个汉字，首先获取其独热编码$\boldsymbol{b}_i$，编码维数为汉字字典的大小N^{char}，然后与嵌入矩阵$\boldsymbol{M}_{\text{emb}}$相乘（$\boldsymbol{M}_{\text{emb}} \in \mathbf{R}^{N^{\text{char}} \times D^{\text{char}}}$），得到维度为$D^{\text{char}}$的紧致编码$\boldsymbol{c}_i$，再将其与上下文矩阵$\boldsymbol{M}_{\text{ctx}}$（$\boldsymbol{M}_{\text{ctx}} \in \mathbf{R}^{D^{\text{char}} \times N^{\text{char}}}$）相乘，经过softmax激励，获得输出$\boldsymbol{o}_i$（$\boldsymbol{o}_i \in \mathbf{R}^{N^{\text{char}}}$）。最后，根据$\boldsymbol{o}_i$计算似然损失函数，并利用梯度下降法调整嵌入矩阵$\boldsymbol{M}_{\text{emb}}$和上下文矩阵$\boldsymbol{M}_{\text{ctx}}$。

第二步，使用预训练的嵌入矩阵$\boldsymbol{M}_{\text{emb}}$初始化char-CNN的嵌入层，并进一步精调参数。

假设句子中第k个词为t_k，将t_k中每一个汉字的独热编码与矩阵$\boldsymbol{M}_{\text{emb}}$相乘，得到汉字对应的字向量，然后输入char-CNN网络，进行卷积、池化操作，得到对应的词向量$\boldsymbol{v}_k$。对于t_k的前驱词t_{k-1}和后继词t_{k+1}，按照同样操作得到对应词向量$\boldsymbol{v}_{k-1}$和$\boldsymbol{v}_{k+1}$。

将$\boldsymbol{v}_k$输入biLM模型，生成其前驱词向量$\boldsymbol{v}'_{k-1}$和后继向量$\boldsymbol{v}'_{k+1}$。比对$\boldsymbol{v}'_{k-1}$、$\boldsymbol{v}'_{k+1}$与$\boldsymbol{v}_{k-1}$、$\boldsymbol{v}_{k+1}$，计算似然损失，据此调整ELMo对应的模型参数，包括嵌入矩阵$\boldsymbol{M}_{\text{emb}}$，char-CNN和biLM的参数。其模型结构如图4所示。

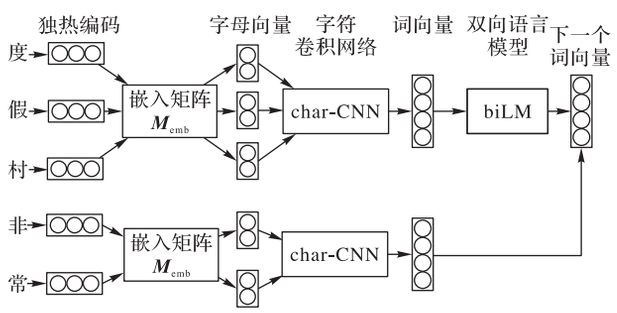

图4 基于预训练汉字向量的ELMo

Fig. 4 ELMo based on pre-trained vectors of Chinese characters

3.2 多尺度卷积神经网络（MSCNN）

ELMo仅仅对文本特征进行了编码，进行情感分类还需要一个分类器。常用的分类器有支持向量机（SVM）、卷积神经网络（CNN）和循环神经网络（RNN）等。

句子的含义通过其组成词语来体现，然而词语并不直接组成句子，而是首先组成短语，然后通过短语组成句子。与char-CNN的思想类似，如果对词语进行卷积，则可以得出短语对应的语义向量，如果采用不同尺度大小的卷积核，则可以获得不同尺度的短语语义向量，将这些短语向量与词向量融合，作为句子最终的特征，其语义信息比单纯使用词向量要丰富许多。基于上述思想，本文构造了多尺度卷积神经网络分类器。

假设长度为N的句子所对应的词向量分别为$\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N$，将$N$个向量进行连接，构成矩阵$\boldsymbol{V}$，即$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N]$，$\boldsymbol{V} \in \mathbf{R}^{d \times N}$，其中$d$为词向量的维数。假设采用的卷积核为$\boldsymbol{K}$，$\boldsymbol{K} \in \mathbf{R}^{d \times w}$，$w$表示卷积宽度，则$\boldsymbol{V}$通过$\boldsymbol{K}$进行卷积运算的公式为：

$$f_i = \tanh(\langle \boldsymbol{K}, \boldsymbol{V} \rangle + \boldsymbol{b}) \tag{4}$$

其中：$\langle \cdot \rangle$代表卷积操作，f_i为卷积之后的特征，$f_i \in \mathbf{R}$，$i \in \{1, 2, \cdots, N - w + 1\}$。

将得到的$N - w + 1$个卷积特征输入池化层，按照式（5）进行最大值池化运算，其公式为：

$$y = \max_i f_i \tag{5}$$

其中$y \in \mathbf{R}$。

对卷积核$\boldsymbol{K}$，矩阵$\boldsymbol{V}$经过卷积、池化操作后可得到输出y。假设存在不同尺度的m个卷积核，如$\boldsymbol{K}_1 \in \mathbf{R}^{d \times w_1}$，$\boldsymbol{K}_2 \in \mathbf{R}^{d \times w_2}$，$\cdots$，$\boldsymbol{K}_m \in \mathbf{R}^{d \times w_m}$，则经过卷积操作会得到$m$个输出$y_1, y_2, \cdots, y_m$。将其连接成一个向量，得到$\boldsymbol{Y} = [y_1, y_2, \cdots, y_m]$，$\boldsymbol{Y} \in \mathbf{R}^m$，$\boldsymbol{Y}$为不同尺度短语的融合特征。最后，将$\boldsymbol{Y}$输入到两层全连接神经网络中，并利用softmax激励实现分类。多尺度卷积神经网络（MSCNN）模型结构图5所示。

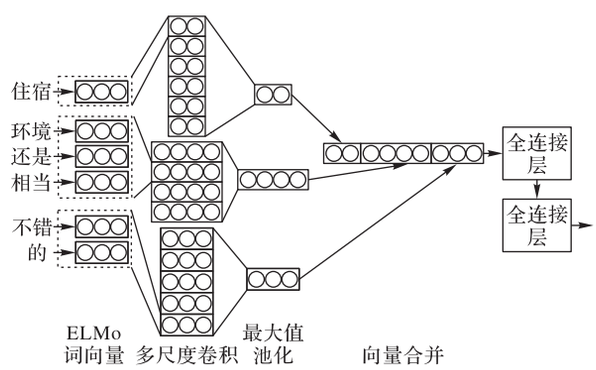

图5 多尺度卷积神经网络模型结构

Fig. 5 Model architecture of MSCNN

3.3 基于ELMo和多尺度卷积神经网络的融合模型

基于上述模型，本文提出了基于ELMo和多尺度卷积神经网络的融合模型用于情感分析。该模型主要由两部分组成：第一部分是利用ELMo学习预训练数据，生成句子上下文相关的词向量；第二部分是利用多尺度卷积神经网络，对词向量的特征进行二次抽取，并进行特征融合，生成句子的整体语义表示。最后，经过softmax激励函数实现文本情感倾向的分类。

该模型的优势在：1）ELMo是预训练模型，通常在大规模语料上进行训练，生成的特征其泛化能力更强；2）ELMo学习的词向量不但可以准确表示多义词的多个不同语义，而且还融入了词语所在句子的语义；3）该模型既利用了循环神经网络结构（ELMo中的双向LSTM网络结构），也利用卷积神经网络结构（MSCNN），相对于采用单一网络结构的方法，提取的特征更加丰富多样。

4 实验

4.1 数据集

实验在两个数据集上进行（表1），一是Tan等[24]收集的酒店评论数据，该数据集由10 000个文件组成，每个文件存储一条酒店评论信息，其中7 000条为正面评论，3 000条为负面评论。该数据集按照不同规模被划分为4个数据集，分别是htl-2000、htl-

4000、htl-6000和htl-10000。另一个是中文信息学会举办的自然语言处理会议公布的深度学习情感分类评测数据集(NLPCC2014 task2),包含训练数据集和测试数据集,其中训练数据集含有1万条中文产品评论数据;测试数据集包含2 500条中文产品评论数据,1 250条正面评论,1 250条负面评论。

表1　实验数据集

Tab. 1　Datasets used in the experiment

数据集	正面评论数	负面评论数
htl-2000	1 000	1 000
htl-4000	2 000	2 000
htl-6000	3 000	3 000
htl-10000	7 000	3 000
NLPCC2014 task2 train	5 000	5 000
NLPCC2014 task2 test	1 250	1 250

4.2　评价标准

本文使用正确率(Accuracy)对实验结果进行评价,其计算公式为:

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN} \tag{6}$$

其中:TP(True Positive)代表真阳性,TN(True Negative)代表真阴性,FP(False Positive)代表假阳性,FN(False Negative)代表假阴性。

4.3　预训练

在进行情感分类之前,需要对ELMo进行预训练,以获得上下文相关的词向量信息。预训练在百科类问答数据集上进行,该数据集通过爬取社区问答数据获得,含有150万个预先过滤的问题和答案,数据大小为1.48 GB。

原始数据为json格式,预处理首先抽取每一条数据的content部分,去除内容中的空行、特殊符号,过滤词语少于5的句子。然后,利用北京大学提供的通用切词工具pkuseg,对句子进行切词。处理后得到9 205 479个句子,每个句子占一行,句子词语之间用空格分隔。样例如图6。

```
第一 切记 ， 人民币 理财 产品 不同于 储蓄 、 国债 等 产品 ， 并非 没有 风险 ， 预期 收益率 也 并非 实际 收益率
不论 哪个 银行 ， 一般 都 不 会 承诺 保本 保息 之类 的 最低 收益 或 固定 收益 ， 打 出 的 底牌 往往 是 预期 收益率
无论 是 哪 一 种 理财 产品 ， 它 的 主要 投资 对象 是 国债 、 央行 票据 和 金融 债券 等
因此 基金 存在 的 风险 ， 理财 产品 照样 存在
第二 切记 ， 个人 理财 产品 的 高 收益 通常 伴随 着 高风险
```

图6　预处理后的训练语料样例

Fig. 6　Sample of preprocessed training corpus

ELMo的训练分为二步:第一步采用Word2vec工具的skip-gram模型,学习字符预训练语料,获取字符向量,实验中字符向量的维度设为128,上下文窗口大小设为5;第二步训练ELMo,利用上述学习到的字符向量初始化char-CNN的嵌入层,然后学习分词语料,获得模型参数,实验中char-CNN的输出维数为256,biLM模型层数设为2,隐状态的维数为512,展开深度设为30。

实验对比了采用预训练字符向量初始化和采用随机向量初始化两种方法,ELMo的训练误差曲线如图7所示。从图中可以看出,对于采用预训练向量初始化的ELMo,其训练收敛速度更快,训练精度更高。

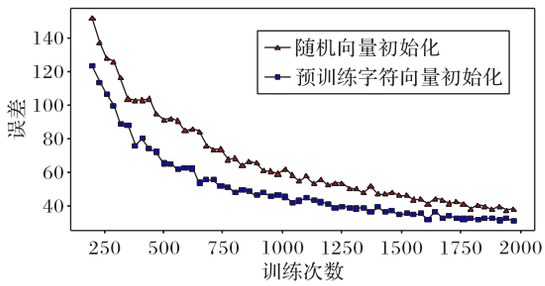

图7　采用不同初始化方法的ELMo训练误差曲线

Fig. 7　Training error curves of ELMo with different initialization strategies

4.4　多尺度卷积神经网络进行情感分类

利用上述方法对酒店评论数据进行预处理,将处理后的句子输入改进ELMo,获取对应的ELMo特征向量,最后将特征输入MSCNN模型,进行分类。

为了获得MSCNN的最优卷积尺度,分别在酒店评论和NLPCC2014 task2两个数据集上进行测试,实验结果如图8所示。

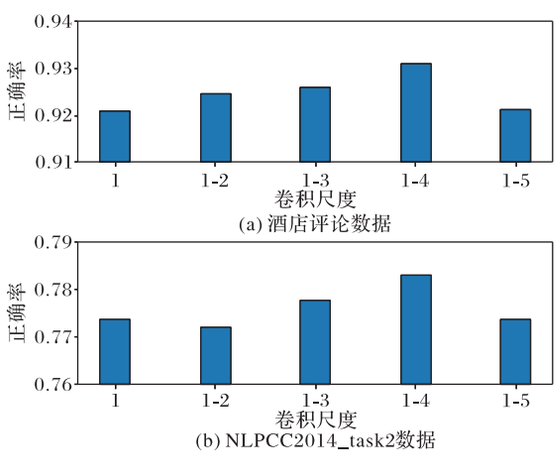

图8　不同卷积尺度下MSCNN的分类结果

Fig. 8　Classification results of MSCNN on different convolution scales

图中卷积尺度1–n表示同时使用卷积尺度为1,2,…,n的卷积核。卷积尺度为卷积核的宽度,代表卷积核所覆盖的词语数目。卷积核的实际大小为512×卷积核的宽度,其中512为ELMo的特征向量维数。卷积核的尺度越大,对应卷积核的数目也应该越多。实验中,对于尺度为1的卷积核,卷积核数目设为64;尺度为2、3的卷积核,数目设为128;尺度为4、5的卷积核,数目设为256。

从图8可以看出,分类正确率随着卷积尺度的增大而逐渐增加,这说明使用多尺度特征确实有助于提高分类精度。当选用卷积尺度为1-4时,效果最好。当选用卷积尺度为1-5时,效果不如1-4,这有可能是由于汉语中距离超过4的词语之间的语义联系较弱的缘故。

最终确定最优卷积尺度为1-4,在该尺度下对酒店评论数据集进行实验,效果如表2所示。

从表中可以看出,对不同规模的数据集,本文方法的分类正确率始终维持在93%以上,即使是针对不平衡数据集htl-10000,正确率也达到了93.3%,和平衡数据集相比,几乎没有下降,这说明本文方法有很好的鲁棒性,并且能够处理不均衡样本集。

表2 不同酒店评论数据集上的实验结果

Tab. 2 Experimental results on different hotel review datasets

10折交叉验证	正确率(Accuracy)			
	htl-2000	htl-4000	htl-6000	htl-10000
验证0	0.960	0.925	0.935	0.910
验证1	0.950	0.910	0.928	0.918
验证2	0.920	0.955	0.930	0.932
验证3	0.945	0.955	0.930	0.931
验证4	0.905	0.930	0.945	0.938
验证5	0.945	0.932	0.920	0.941
验证6	0.925	0.932	0.923	0.937
验证7	0.950	0.960	0.937	0.937
验证8	0.945	0.950	0.935	0.939
验证9	0.925	0.932	0.927	0.943
平均	**0.937**	**0.938**	**0.931**	**0.933**

论文同样对比了其他方法,如支持向量机(SVM)、朴素贝叶斯(Naïve Bayes, NB)、融合字、词的双向LSTM模型(Character, Word and Part-of-Speech Attention model based on Bi-LSTM, CWPAT-Bi-LSTM)和卷积神经网络(CNN),结果如表3所示。

表3 不同方法在酒店评论数据集上的正确率指标比较

Tab. 3 Classification accuracy of different methods on hotel review datasets

模型	正确率(Accuracy)			
	htl-2000	htl-4000	htl-6000	htl-10000
SVM[25]	0.733	0.826	0.830	0.848
NB[25]	0.678	0.684	0.696	0.684
WCTAT-Bi-LSTM[25]	0.935	0.923	0.915	0.923
CNN[26]	0.872	0.884	0.913	0.928
本文方法	**0.937**	**0.938**	**0.931**	**0.933**

从表3中可以看出,本文方法明显优于SVM和NB方法,与SVM方法相比正确率平均提升12.56个百分点,与NB方法相比正确率平均提升24.93个百分点。这说明由于神经网络方法具有特征自动抽取的能力,能够获得更有效的语义特征,从而得到了更高的分类正确率。对比神经网络方法,本文方法相对卷积神经网络模型,正确率平均提升了3.56个百分点。即使对比当前最好方法,基于注意力机制的循环神经网络模型,正确率也平均提升了1.08个百分点。主要原因是本文方法实际上是循环神经网络模型和卷积神经网络的混合模型,在词向量生成阶段,使用双向LSTM构造上下文相关词向量,在分类器阶段,利用多尺度卷积进行词向量的融合,进一步抽取可用的语义特征。该混合模型综合了两种模型的优势,所以效果更好。

为进一步验证本文模型的性能,实验还在NLPCC2014 task2数据集进行了测试,其结果如表4所示。

表4 不同方法在NLPCC2014_task2数据集上的实验结果

Tab. 4 Experimental results of different methods on dataset NLPCC2014_task2

模型	正确率(Accuracy)
LSTM[27]	0.6664
GRU[27]	0.7392
CNN[27]	0.7336
LSTM+CNN[27]	0.7616
本文方法	**0.7832**

从表中可以看出,与经典的循环神经网络模型LSTM和门控循环单元网络(Gated Recurrent Unit, GRU)相比,本文方法的正确率提高了11.68个百分点和4.4个百分点,与卷积神经网络相比,正确率提高了4.96个百分点。该结果和酒店评论数据集上的结果是一致的,说明本文模型迁移到新数据集上同样有效。与LSTM和CNN的混合方法相比,正确率也提高了2.16个百分点。虽然本文模型从表面上看也是LSTM和CNN的混合模型,但本文的LSTM是蕴含在ELMo中的,而ELMo采用的是预训练的方式,是从大语料而非任务语料中训练语义向量,得到的词语表征其泛化程度更高,这是基于任务的训练模型无法比拟的。

5 结语

不同于传统的将词嵌入向量作为神经网络的输入,本文提出将ELMo获得的词语向量作为网络输入,该向量融合了词语本身和词语所在上下文的语义特征,可以很好地表示多义词的不同语义。此外,针对中文语料,本文提出采用预训练的汉字字符向量初始化ELMo的嵌入层,可加快ELMo的训练,提高训练精度。最后,分类器采用多尺度的卷积神经网络(MSCNN),该分类器能够融合不同尺度的短语特征,有利于后续分类。实验结果表明,本文提出的方法能有效提高情感分类的正确率。

最近几年,基于多头自注意力的网络,如Transformer模型在自然语言处理领域崭露头角,Transformer能够有侧重地进行语义向量的融合,其本质更像是针对自然语言的卷积操作,下一步也尝试将ELMo和Transformer模型混合,分别利用ELMo和Transformer各自的优势,构造更适用于情感分析的模型。

在句子语法层面,中、英文是有显著的区别的。例如,中文句子往往是由很多小的短句构成(大量逗号分隔的短句),而英文一般只会有1~2个从句,而且从句大都不以子句的形式存在(中间没有逗号分隔)。因此,如何改进ELMo结构,使其更适用于中文语法结构,也是下一步的研究方向。

参考文献 (References)

[1] ZHU Z, DONG S, YU C, et al. A text hybrid clustering algorithm based on HowNet semantics[J]. Key Engineering Materials, 2011, 474/476:2071-2078.

[2] PANG B, LEE L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1/2): 1-135.

[3] MORAES R, VALIATI J F, NETO W P G. Document-level sentiment classification: an empirical comparison between SVM and ANN [J]. Expert Systems with Applications, 2013, 40(2): 621-633.

[4] LIU B. Sentiment Analysis: Mining Opinions, Sentiments, And Emotions [M]. New York: Cambridge University Press, 2015: 47-68.

[5] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions [C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 151-161.

[6] QIAN Q, TIAN B, HUANG M, et al. Learning tag embeddings and tag-specific composition functions in recursive neural network [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Compu-

tational Linguistics, 2015: 1365-1374.

[7] WANG X, LIU Y, SUN C, et al. Predicting polarities of tweets by composing word embeddings with long short-term memory[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1343-1353.

[8] WANG X, JIANG W, LUO Z. Combination of convolutional and recurrent neural network for sentiment analysis of short texts[C]// Proceedings of the 26th International Conference on Computational Linguistics. Osaka: The COLING 2016 Organizing Committee, 2016: 2428-2437.

[9] GUGGILLA C, MILLER T, GUREVYCH I. CNN-and LSTM-based claim classification in online user comments[C]// Proceedings of the 26th International Conference on Computational Linguistics. Osaka: The COLING 2016 Organizing Committee, 2016: 2740-2751.

[10] AKHTAR S, KUMAR A, GHOSAL D, et al. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 540-546.

[11] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-03-12]. https://arxiv. org/pdf/1409. 0473. pdf.

[12] 曾锋,曾碧卿,韩旭丽,等. 基于双层注意力循环神经网络的方面级情感分析[J]. 中文信息学报, 2019, 33(6): 108-115. (ZENG F, ZENG B Q, HAN X L, et al. Double attention neural network for aspect-based sentiment analysis[J]. Journal of Chinese Information Processing, 2019, 33(6): 108-115.)

[13] 曾碧卿,韩旭丽,王盛玉,等. 基于双注意力卷积神经网络模型的情感分析研究[J]. 广东工业大学学报, 2019, 36(4): 10-17. (ZENG B Q, HAN X L, WANG S Y, et al. Sentiment classification based on double attention convolutional neural network model [J]. Journal of Guangdong University of Technology, 2019, 36(4): 10-17.)

[14] 韩萍,孙佳慧,方澄,等. 基于情感融合和多维注意力机制的微博文本情感分析[J]. 计算机应用, 2019, 39(S1): 75-78. (HAN P, SUN J H, FANG C, et al. Micro-blog sentiment analysis based on emotional fusion and multi-dimensional self-attention mechanism [J]. Journal of Computer Applications, 2019, 39(S1): 75-78.)

[15] 石磊,张鑫倩,陶永才,等. 结合自注意力机制和Tree-LSTM的情感分析模型[J]. 小型微型计算机系统, 2019, 40(7): 1486-1490. (SHI L, ZHANG X Q, TAO Y C, et al. Sentiment analysis model with the combination of self-attention and tree-LSTM[J]. Journal of Chinese Computer Systems, 2019, 40(7): 1486-1490.)

[16] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2019-03-072]. https://arxiv. xilesou. top/pdf/1301. 3781. pdf.

[18] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[19] KIROS R, ZHU Y, SALAKHUTDINOV R, et al. Skip-thought vectors[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 3294-3302.

[20] LOGESWARAN L, LEE H. An efficient framework for learning sentence representations[EB/OL]. [2019-03-12]. https://arxiv. org/pdf/1803. 02893. pdf.

[21] MCCANN B, BRADBURY J, XIONG C, et al. Learned in translation: Contextualized word vectors[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6297-6308.

[22] CER D, YANG Y, KONG S, et al. Universal sentence encoder [EB/OL]. [2019-03-11]. https://arxiv. org/pdf/1803. 11175. pdf.

[23] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 2227-2237.

[24] TAN S, ZHANG J. An empirical study of sentiment analysis for Chinese documents[J]. Expert Systems with Applications, 2008, 34(4): 2622-2629.

[25] 赵富,杨洋,蒋瑞,等. 融合词性的双注意力Bi-LSTM情感分析[J]. 计算机应用, 2018, 38(S2): 103-106, 147. (ZHAO F, YANG Y, JIANG R, et al. Sentiment analysis based on double-attention Bi-LSTM using part-of-speech[J]. Journal of Computer Applications, 2018, 38(S2): 103-106, 147.)

[26] WANG X, LI J, YANG X, et al. Chinese text sentiment analysis using bilinear character-word convolutional neural networks[C]// Proceedings of the 2017 International of Conference on Computer Science and Application Engineering. Lancaster, PA: DEStech Publications Inc., 2017: 36-43.

[27] 杜永萍,赵晓铮,裴兵兵. 基于CNN-LSTM模型的短文本情感分类[J]. 北京工业大学学报, 2019, 45(7): 48-56. (DU Y P, ZHAO X Z, PEI B B. Short text sentiment classification based on CNN-LSTM model[J]. Journal of Beijing University of Technology, 2019, 45(7): 48-56.)

This work is partially supported by the Key Special Project of Cloud Computing and Big Data of the National Key Research and Development Program of China (2016YFB1001100, 2016YFB1001104), the Youth Program of the National Natural Science Foundation of China (61702218).

ZHAO Ya'ou, born in 1982, Ph. D., lecturer. His research interests include natural language processing, artificial intelligence.

ZHANG Jiachong, born in 1965, Ph. D., professor. His research interests include artificial intelligence, data mining.

LI Yibin, born in 1960, Ph. D., professor. His research interests include robot, human-machine interaction.

FU Xianrui, born in 1986, engineer. His research interests include software architecture, artificial intelligence.

SHENG Wei, born in 1983, information technology project manager. His research interests include artificial intelligence, financial self-service terminal.