基于 CNN+LSTMAttention 的营销新闻文本分类

刘高军,王小宾
(北方工业大学 信息学院,北京 100144)

摘 要:针对营销新闻分类识别任务,传统方法采用的长短期记忆神经网络 LSTM 和卷积神经网络 CNN 存在分类识别率不高的问题,因此提出一种融合 CNN 和引入注意力机制的长短时记忆(LSTMAttention)来提高营销新闻识别分类能力。首先通过 word2vec 获取营销新闻文本词向量形成的矩阵,分别输入到传统机器学习分类模型中,在此基础上使用模型融合技术融合单一模型中分类效果较好的模型,最后得到融合模型和单一模型的分类结果并进行对比。实验结果显示,在基础模型 LSTM 引入了注意力机制之后准确率,召回率和 F1 值分别达到 67.01%,66.07%,0.680,而 CNN 和 LSTMAttention 进行模型融合之后的准确率,召回率和 F1 值进一步达到了 68.29%,71.27%,0.692。表明基于 CNN 和 LSTMAttention 融合之后的神经网络模型相较于单一模型,最终分类效果更好,可以达到提高营销新闻文本分类识别效果的目的。

关键词:营销新闻;文本分类;卷积神经网络;注意力机制;长短期记忆神经网络
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)11-0059-05
doi:10.3969/j.issn.1673-629X.2020.11.011

Marketing News Text Classification Incorporating CNN+LSTMAttention

LIU Gao-jun, WANG Xiao-bin
(School of Information, North China University of Technology, Beijing 100144, China)

Abstract: For the classification and recognition task of marketing news, the long-term and short-term memory neural network LSTM and convolutional neural network CNN used by traditional methods have a low classification recognition rate. Therefore, we propose a long and short-term memory (LSTMAttention) that combines CNN and introduces attention mechanism to improve the ability to identify and classify marketing news. Firstly, the matrix formed by the word vector of the marketing news text is obtained by word2vec and input into the traditional machine learning classification model. Based on this, model fusion technology is used to fuse the model with better classification in a single model, and finally the classification results of fusion model and single model are obtained and compared. The experiment shows that after introduction of the attention mechanism in the basic model of LSTM, the accuracy, recall and F1 values reach 67.01%, 66.07% and 0.680 respectively, and the accuracy, recall and F1 value after model fusion of CNN and LSTMAttention further reach 68.29%, 71.27% and 0.692. It is shown that the neural network model based on the fusion of CNN and LSTMAttention has a better final classification effect than a single model, and can achieve the purpose of improving the classification and recognition effect of marketing news text.

Key words: marketing news; text classification; convolutional neural network; attention mechanism; long short-term memory neural network

0 引 言

近年来,许多公司宣传自己的产品使用了一种新型营销方式-营销新闻,营销媒介经过多年的演化之后形成了营销新闻。营销新闻同新闻一样从多方面多角度给用户灌输产品信息、产品理念等,通过传播产品资讯信息来引导消费者购买。但是随着信息技术和互联网的迅猛发展,千人千面的信息推荐方式给亿万网民的阅读带来了便利,但同时营销、低俗、标题党等低质量新闻的掺杂也给用户带来了不同程度上的困扰。为了给用户更好的阅读体验,需要识别出有营销意图的新闻,因此文本挖掘技术的研究变得越来越重要。文本分类作为高效的挖掘技术和信息检索,对文本数

收稿日期:2019-12-10 **修回日期**:2020-04-14
基金项目:国家自然科学基金(61672040);新闻出版业科技与标准重点实验室项目(4020548418X8)
作者简介:刘高军(1962-),男,硕士,教授,CCF 会员(78191M),研究方向为自然语言处理、软件工程及数据结构;王小宾(1992-),男,硕士,研究方向为自然语言处理、文本分类识别等。

据的管理有着重要的作用,目前已经广泛应用于垃圾邮件检测[1]、舆情分析[2]和新闻分类[3]推荐等场景。文本分类使用传统机器学习,自然语言处理和深度学习技术来有效地分类和发现不同类型的文本。如何从这些庞大的文本中提取有价值的信息并自动检索,分类和汇总,是文本挖掘的重要目标,因此利用文本分类技术对营销新闻进行分类识别近年来成为一个值得研究的课题。

文本分类是数据挖掘的重要组成部分,它对大量文本的分类和挖掘具有重要的理论和应用价值,它的主要任务是将给定的文本集划分为几种已知的类型集合,例如将新闻文本分为带有营销意图的垃圾新闻和普通的新闻,又比如分辨一个文本是人类作者还是机器自动生成的等等。目前文本分类任务已经应用到了许多领域,如主题分类、垃圾邮件检测、情感分析等。要解决这些问题,研究者需要对数据进行获取分析、挖掘、归类,帮助人们提高信息检索的效率。该文的主要工作是融合 CNN 和引入注意力机制的 LSTM 模型解决营销新闻文本识别分类的问题。

1 相关工作

文本分类技术是自然语言处理领域中的重要应用,它将文本分类为预定义的类别,在当今的网络时代,文本数据已成为最常见的数据形式之一,例如:用户评论[4]、新闻、电子邮件等。文本分类的基本过程通常包括:文本预处理[5],特征提取[6],文本表示[7]和分类器训练[8]。传统的特征表示方法通常会忽略上下文。对于捕获单词的语义,文本中信息的顺序仍然不能令人满意。同时,特征提取的方法存在数据稀疏、维度爆炸等问题,降低了训练模型的泛化能力。

深度学习在图像处理和语音识别等研究领域取得了令人瞩目的成就,在自然语言处理中也发展迅猛。深度学习技术已逐渐取代传统的机器学习方法,并已成为文本分类中的主流技术。深度学习可以更准确地表达对象,并且可以从海量数据中自动获取对象的特征。基于此类功能属性的深度学习模型包括卷积神经网络(CNN)[9]和递归神经网络[10]。

如何对这些海量文本数据进行有效的分类受到了极大的关注。2013 年崔建明等人提出基于支持向量机的文本分类[11],使用支持向量机(SVM)算法优化文本分类器的参数,从而提高了文本分类器的分类精度。2017 年武永亮等人提出基于 TF-IDF 和余弦相似度的文本分类方法[12],传统的机器学习方法被用于分类。TF-IDF 模型用于提取类别关键字,并通过这些类别关键字和需要分类的文本关键字进行余弦相似度计算。2011 年姚全珠等人提出基于 LDA 的文本分类研究[13],提到基于隐式 Dirichlet 分布的 LDA 模型和 SVM 算法的文本分类,但是在大量的短文本中,有短文本长度和过多的噪声,分类效果不好。2017 年夏从零等人提出了基于事件卷积特征的新闻文本分类[14],通过卷积神经网络从新闻文本中提取特征文本,以对文本进行分类,但是,尽管这种方法可以很好地提取特征,但是通常会忽略上下文并且文本语义不够准确。

基于以上考虑,该文拟结合卷积神经网络(CNN)和引入注意力机制的长短时记忆网络结构(LSTMAttention)来解决营销新闻文本分类的问题,提出一种新的新闻营销文本分类的训练模型 CNN-LSTMAttention,以提高新闻营销文本分类的准确率。实验是使用搜狐新闻给出的数据集进行训练和测试的,并使用单个基模型进行了比较。结果表明,与其他单个模型相比,该融合方法对营销新闻分类具有一定的提升。

2 模 型

本节将描述神经网络体系结构的组件(层),并且逐层介绍使用地神经网络中的神经层。

2.1 CNN 用于字符级表示

之前 Santos 和 Zadroznw,Chiu 和 Nichols 的研究表明[15],CNN 是一种有效的方法,可以从单词的字符中提取形态信息,并将其编码为神经表示。图 1 显示了用于提取单词的字符级表示的 CNN。CNN 与 Chiu and Nichols 中的 CNN 相似,不同之处在于该文使用了字符嵌入作为 CNN 的输入,而没有字符类型特征。在将字符嵌入输入到 CNN 之前添加 dropout layer。

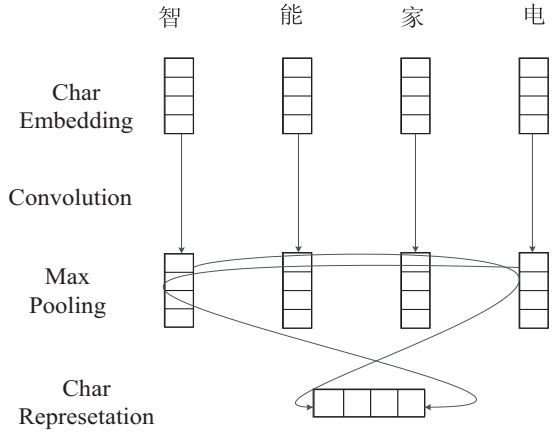

图 1 用于提取单词字符级表示的卷积神经网络

2.2 引入注意力机制的 LSTM

2.2.1 LSTM 单元

Hochreiter 和 Schmidhuber[16] 在 1997 年提出了 LSTM 的网络结构,引入 CEC 单元解决了 RNN 的梯度爆炸和梯度消失的问题。LSTM 由输入层、隐藏层和输出层构成。LSTM 区别于 RNN 的地方在于,它在

算法中加入了一个判断信息有用与否的“处理器”——cell,一个 cell 当中被放置了三扇门,分别叫做:输入门 i_t 、遗忘门 f_t 、输出门 o_t ,LSTM 还有一个记忆单元 c_t ,这些结合起来能够提高 LSTM 处理长序列数据的能力。图 2 给出了 LSTM 单元的基本构图。

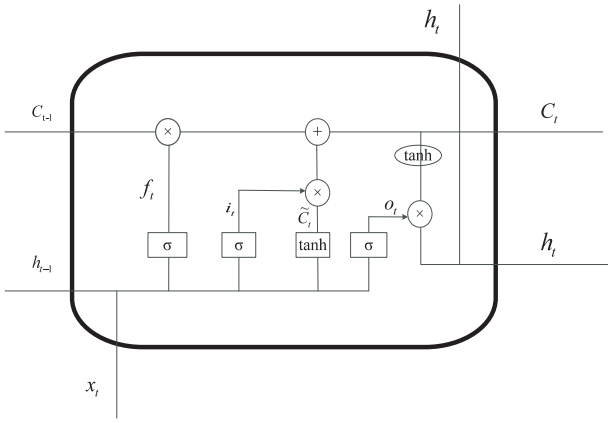

图 2　LSTM 单元示意图

在时间 t 更新的时候,LSTM 单元的公式为:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$m_t = \tanh(W_m \cdot [h_{t-1}, x_t] + b_m)$$
$$c_t = f_t * c_{t-1} + i_t * m_t$$
$$h_t = o_t * \tanh(c_t)$$

其中, i_t 、o_t 和 f_t 分别为输入门、输出门和遗忘门, W 和 b 是模型的参数, σ 表示 sigmoid 函数,输出在 0,1 之间,tanh 表示双曲正切函数,输出在 -1,1 之间, m_t 是当前节点的输入数据, C_{t-1} 是上一个 LSTM 节点的输出信息。

2.2.2　注意力层

注意力机制(attention mechanism)源于对人类视觉的研究。因此注意力机制最早是应用在图像领域的。在自然语言处理领域,注意力机制最成功的应用是机器翻译。基于神经网络的机器翻译模型也叫神经机器翻译(neural machine translation,NMT)。一般的神经机器翻译模型采用“编码 - 解码(encoder to decoder)”的方式进行序列到序列的转换。通过注意力机制直接从源语言的信息中选择相关的信息作为辅助,可以有效地解决编码向量的容量瓶颈问题和长距离依赖问题,无需让所有源语言信息都通过编码向量进行传递,在解码的每一步都可以直接访问源语言的所有位置上的信息,同时源语言的信息可以直接传递到解码过程中的每一步,缩短了信息传递的距离。因此在传统机器学习模型中引入注意力机制会有效地提升模型的实用效果。该文对 LSTM 模型进行改进,将注意力机制引入 LSTM 模型当中。

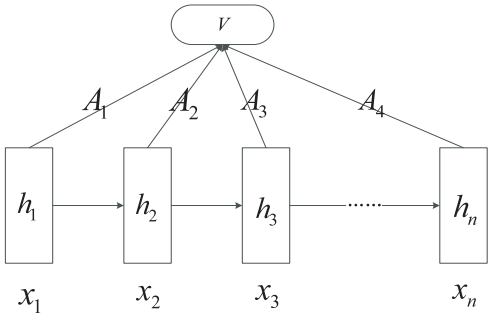

图 3　引入注意力机制示意图

图 3 所示输入层中的词向量 $X_1, X_2, \cdots, X_n$,通过输入到上层的 LSTM 单元层中,在隐藏层输出的过程中,引入注意力机制,分配各个输入的注意力概率的分布值 $A_0, A_1, \cdots, A_n$,之后加和求平均到 V 。

分配权重公式为:

$$A_i = \frac{\exp(\text{score}(\bar{h}, h_i))}{\sum_j \exp(\text{score}(\bar{h}, h_j))}$$

其中, h_i 为第 i 个时刻隐藏层的输出状态, $\bar{h}$ 是一种文本表示向量。

2.3　ATTLSTM-CNN

最后,文本通过将引入注意力机制的 LSTM 加权求平均之后,输出给 softmax 层,从而进行全连接操作,最后得到预测分类结果 Y ,从而构建整个神经网络模型。图 4 为神经网络主要架构。

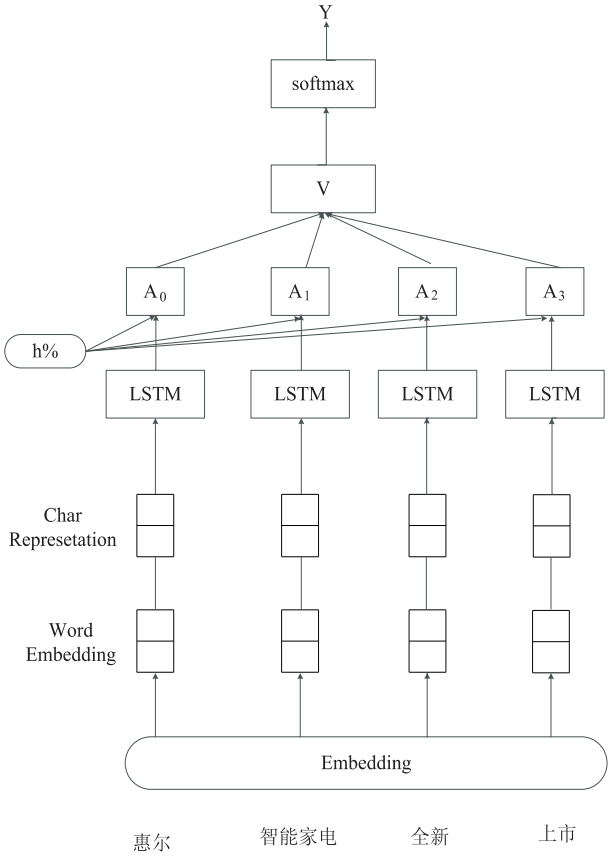

图 4　神经网络主要架构

对于每个单词,有字符嵌入作为输入,由图 1 中的 CNN 计算出字符级表示。然后将字符级表示向量与单词嵌入向量连接起来,以输入到 LSTM 层,在 LSTM 的隐藏层上引入不同权重的注意力机制分配,最后加权求平均,经过 softmax 全连接得到分类结果。

softmax 得到预测分类结果 Y,公式为:

$$Y = \text{softmax}(W_v v + b_v)$$

3 实 验

3.1 数据集

本次实验是在 Linux 系统下进行的,使用 GPU 是 Cuda -v 9.0,实验编程语言为 Python3.6,使用到的深度学习框架为 PyTorch 0.4.0。实验数据集的来源为搜狐新闻数据,从中提取出用于训练词向量的新闻文本语料,大约 40 000 条左右,根据有标注的训练集,分为 0:无营销新闻,1:部分营销新闻,2:全营销新闻,利用没有经过训练的测试集文本数据去评估模型的分类效果。

3.2 实验评价指标

本次实验使用三个评价指标来衡量模型的效果,包括准确率(precision)、召回率(recall)和 F1 值:

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

3.3 实验设置及结果分析

本次实验针对新闻语料数据集,使用 CNN 模型、LSTM 模型、LSTMAttention 模型和 CNN + LSTMAttention 融合模型对新闻进行分类。

CNN 模型:输入卷积神经网络的是经过 word2vec 训练后的词向量形成的矩阵,从而进行分类。

LSTM 模型:输入长短期记忆神经网络的是经过 word2vec 训练后的词向量形成的矩阵,从而进行分类。

LSTMAttention 模型:经过 word2vec 后得到的词向量矩阵输入到引入注意力机制的 LSTM 模型中。

CNN+LSTMAttention 融合模型:使用 word2vec 训练后的词向量矩阵作为输入,CNN 计算出字符级表示。然后将字符级表示向量与单词嵌入向量连接起来,以输入到 LSTM 层。

通过以上三个评价指标对本次实验的分类结果进行衡量。为了说明本次实验当中引入注意力机制的 LSTM 模型,再融入 CNN 模型中对分类结果的影响,使用对比实验的方式将 CNN+LSTMAttention 融合模型与 CNN 模型,经典的 LSTM 模型,LSTMAttention 模型的分类结果进行对比。通过相同数据集上对比实验结果说明本次融合实验的优势,实验结果如表 1 所示。

表 1 不同分类模型的分类结果比较

模型	precision/%	recall/%	F1
CNN	65.60	68.10	0.668
LSTM	66.09	69.32	0.677
LSTMAttention	67.01	66.07	0.680
CNN+LSTMAttention	68.29	1.17	0.692

从表 1 可以看出,在相同数据集的基础上,CNN 模型的分类效果略次于 LSTM 模型,而通过引入注意力机制的 LSTMAttention 模型要表现的比 LSTM 好,因为在特征提取的过程中,LSTMAttention 模型会更多地关注句子中更加重要的词语信息,进而减少那些不重要的信息带来的干扰。该文提出的 CNN + LSTMAttention 将 CNN 和 LSTMAttention 相结合,能够更好地完成分类任务。

此外,本次还设计了显著性检验实验,在搜狐新闻数据集上使用准确率、召回率、F1 值进行评价,从而进一步验证 CNN+LSTMAttention 模型与其他神经网络模型 CNN、LSTM 以及引入注意力机制的 LSTMAttention 模型的比较结果,如图 5 ~ 图 7 所示。

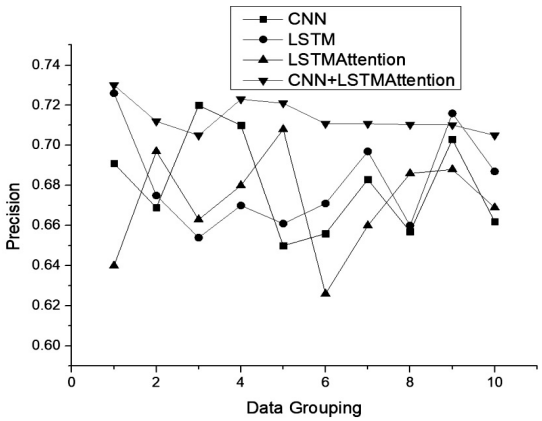

图 5 模型的准确率评估对比

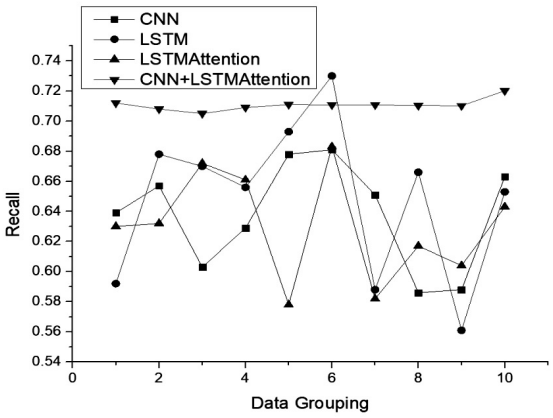

图 6 模型的召回率评估对比

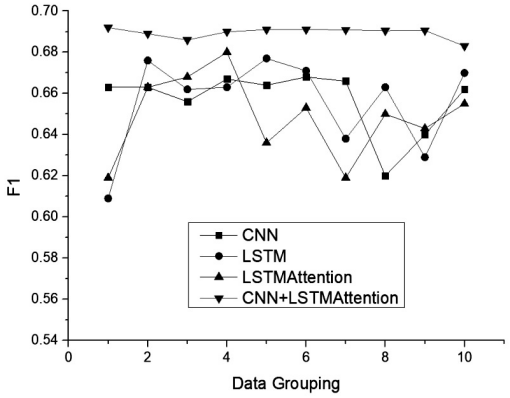

图 7　模型的 F1 值评估对比

从图 5 ~ 图 7 可以看出,在准确率、召回率和 F1 值的评估中,对于相同的搜狐新闻语料数据库,CNN+LSTMAttention 模型在三个评估指标大多分布在其他三个模型之上,评估指标准确率基本在 0.72 左右,F1 值也明显高于其他三个模型。因此该文提出的 CNN+LSTMAttention 融合模型的分类效果要优于 CNN 模型,LSTM 模型和引入注意力机制的 LSTM 模型。

4　结束语

提出了一种融合 CNN 和 LSTMAttention 的营销新闻分类,首先将大量的营销新闻语料使用 word2vec 来训练,然后再从这些训练的新闻文本信息中提取出新闻的文本特征表达-词向量,与其他用于提取特征值的深度学习方法相比较,word2vec 可以高效地对大流量本文数据进行分析,将文本 word 映射到数值空间,进行特征提取,再用提取好的新闻特征向量输入到预先准备的分类模型中,最后使用 softmax 分离器构建分类模型。根据实验,可以看出提出的 CNN+LSTMAttention 模型对于营销新闻的分类效果相对于 CNN 模型和引入注意力机制的 LSTM 模型都有了一定的提升,但是,实验只用了中文新闻语料,没有使用英文新闻语料。下一步的研究工作可以尝试使用该模型对英文新闻语料进行分类,而且可以尝试不用 Pytorch 深度学习框架改用 tensorflow 深度学习框架进行自然语言处理。

参考文献:

[1] 胡小娟,刘　磊,邱宁佳. 基于主动学习和否定选择的垃圾邮件分类算法[J]. 电子学报,2018,46(1):203-209.

[2] 柳　源. 基于数据挖掘技术的舆情分析系统的设计[J]. 电脑知识与技术,2019,15(20):9-10.

[3] 毕曦文,纪明宇,吴　鹏,等. 个性化高校新闻分类推荐的应用研究[J]. 计算机应用与软件,2019,36(7):218-223.

[4] 唐　利. 网络电影评论的情感倾向性分类研究[J]. 遵义师范学院学报,2018,20(6):160-164.

[5] 王永昌,朱立谷. 面向 Twitter 情感分析的文本预处理方法研究[J]. 中国传媒大学学报:自然科学版,2019,26(2):31-38.

[6] YU Yajie, CAO Hui, YAN Xingyu, et al. Defect identification of wind turbine blades based on defect semantic features with transfer feature extractor[J]. Neurocomputing, 2019, 19(3):56-62.

[7] 崔　莹. 深度学习在文本表示及分类中的应用研究[J]. 电脑知识与技术,2019,15(16):174-177.

[8] 李炳聪. 用正则的方法在正样本和无标签样本上训练二分类器[J]. 信息与电脑:理论版,2019(5):67-68.

[9] 蓝雯飞,徐　蔚,王　涛. 基于卷积神经网络的中文新闻文本分类[J]. 中南民族大学学报:自然科学版,2018,37(1):138-143.

[10] 黄　磊,杜昌顺. 基于递归神经网络的文本分类研究[J]. 北京化工大学学报:自然科学版,2017,44(1):98-104.

[11] 崔建明,刘建明,廖周宇. 基于 SVM 算法的文本分类研究[J]. 计算机仿真,2013,30(2):299-302.

[12] 武永亮,赵书良,李长镜,等. 基于 TF-IDF 和余弦相似度的文本分类方法[J]. 中文信息学报,2017,31(5):138-145.

[13] 姚全珠,宋志理,彭　程. 基于 LDA 模型的文本分类研究[J]. 计算机工程与应用,2011,47(13):150-153.

[14] 夏从零,钱　涛,姬东鸿. 基于事件卷积特征的新闻文本分类[J]. 计算机应用研究,2017,34(4):991-994.

[15] HUANG W, WANG J. Character-level convolutional network for text classification applied to chinese corpus[J]. arXiv:1611.04358,2016,23(6):102-109.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.