DOI:10.13878/j.cnki.jnuist.2023.02.003

张骏强[1]　高尚兵[1]　苏睿[1]　李文婷[1]

融合多粒度动态语义表征的文本分类模型

摘要

在对化工领域类文本进行分类任务时,由于文本的专业性以及复杂多样性,仅仅依靠现有的词向量表征方式,很难对其中的专业术语以及其他化工领域内相关字词的语义进行充分表征,从而导致分类任务的准确率不高.本文提出一种融合多粒度动态语义表征的文本分类模型,首先在词嵌入层使用动态词向量表征语义信息并引入对抗扰动,使得词向量具有更好的表征能力,然后利用多头注意力机制进行词向量权重分配,获得带有关键语义信息的文本表示,最后使用提出的多尺度残差收缩深层金字塔形的卷积神经网络与混合注意力胶囊双向LSTM网络模型分别提取不同粒度的文本表示,融合后对得到的最终文本表示进行分类.实验结果表明,相比于现有模型,所提出的模型使用不同词向量表示时,在化工领域文本数据集上F1-Score最高可达84.62%,提升了0.38~5.58个百分点;在公开中文数据集THUCNews和谭松波酒店评论数据集ChnSentiCorp上进行模型泛化性能评估,模型也有较好表现.

关键词

文本分类;对抗扰动;多粒度;多头注意力机制;深度残差收缩;预训练语言模型

中图分类号 TP391;TQ072

文献标志码 A

收稿日期 2022-01-12

资助项目 国家重点研发计划(2018YFB1004904);江苏省"六大人才高峰"资助项目(XYDXXJS-011);江苏省"333工程"资助项目(BRA2016454);江苏省教育厅重大研究项目(18KJA520001);淮阴工学院研究生科技创新计划项目(HGYK202121)

作者简介

张骏强,男,硕士生,研究方向为数据挖掘与推荐系统.zhangjq0906@hyit.edu.cn

高尚兵(通信作者),男,博士,教授,主要研究方向为机器学习、数据挖掘与模式识别.gaoshangbing@hyit.edu.cn

1 淮阴工学院 计算机与软件工程学院/江苏省物联网移动互联技术工程实验室,淮安,223003

0 引言

化工业属于国民经济基础产业之一,它在中国近现代工业的发展中占据着极为重要的地位,其制造出来的各种产品渗透在人们生活的方方面面中.新冠疫情的出现使得本就不景气的传统工业经济效益下滑加剧[1],而互联网行业受其影响相对较小,各行各业的海量信息以文本、图像、音频等方式被呈现在其中.通过新兴的计算机技术对互联网上海量的资源加以分析,挖掘其蕴含的内在价值,从而反哺传统工业,具有重大的现实意义.

化工领域文本涉及到化学这一自然科学,相比于其他仅涉及人文科学的文本数据,化工领域文本数据有着更高的专业程度,这使得对该领域相关人员专业水平要求很高,化工文本理解学习成本也较大.

对于化工领域内的从业人员而言,可以依据其经验以及专业知识对领域内化工产品文本所属衍生领域进行分类.而对于计算机而言,采用自然语言处理的方式对化工产品文本进行区别分类具有更大的可行性与便捷性.

目前针对文本分类的算法,使用的词向量大多还是基于Word2Vec[2]等模型训练静态词向量,而现如今很多有隐含价值的文本越来越趋向于碎片化,其上下文之间往往不具备很紧密的逻辑关系,静态词向量并不能很好地根据字词的上下文去变化,语义表达能力较弱[3],这使得文本分类精度受到极大影响,预训练语言模型[4]的出现很好地缓解了这个问题.Google于2018年提出了一种基于Transformer结构的双向编码表示模型(Bidirectional Encoder Representation from Transformers,BERT)[5],该模型的出现使得词向量模型的泛化能力进一步增强,并在文本分类领域做出了巨大的贡献.Lan等[6]通过矩阵分解以及共享参数的方法在仅仅损失小部分模型性能的基础上,进一步地减少了BERT模型的参数量.Yang等[7]通过将单词随机打乱词序从而实现上下文双向编码,进一步提升了模型性能.虽然BERT等预训练语言模型性能表现优秀,但该模型并不是针对中文领域文本所设计的预训练模型,也没有对中文领域文本特点进行针对性优化.因此,Cui等[8]提出一种新的中文预训练语言模型(MLM as correction BERT,MacBERT),并在相关中文自然语言处理任务中取得了较好成绩.

针对化工领域产品文本这一种类特殊的文本数据,本文总结了以下几个特点:1)文本专业性强,文本中包含有大量化学专业术语名词,主流分词方式缺乏化学名词词库,而人工理解文本进行分类成本较高,要求进行分类的人员有较高的相关知识水平;2)文本类别较多,例如本文所统计的化工领域内产品文本就涵盖有有机原料、化工试剂、化工中间体、化学矿、无机化工、农业化工、涂料油漆、聚合物、染料、食品添加剂、生物化工等17个类别,这也增加了文本分类难度;3)文本规范性差,文本中的化学名词由数字、中文、英文、符号组合,是文本中的重要局部特征之一,总体是一种交替间隔出现的趋势;4)文本篇幅长且关键特征呈现碎片化分布,化工产品文本主要有CAS号、产品描述、形状特征以及包装方式等字段内容构成,通常文本篇幅较长,但是字段之间联系较少,逻辑性弱;5)文本含噪声比例高,化工产品文本中会存在部分文本携带同厂家生产的其他类型化工产品广告推广,但是这些广告文本内容与其类别内的其他产品关键特征相似度高,这使得噪声文本很难被常规清洗手段去除.

现有的文本分类方法大多还是针对通用领域文本,其中包含的领域专业字词往往较少,这导致分类方法在词向量建模阶段就不能很好地表达语义信息,进而在后续使用传统网络模型进行文本特征提取时,会产生诸多问题.一方面,传统卷积神经网络(CNN)只能提取局部特征,由于化工文本逻辑性不强,并且分布呈现碎片化,这使得CNN提取到的局部特征往往不够全面,并且无法很好利用化工长文本中蕴含的全局语义信息.另一方面,传统循环神经网络及其变体虽然能提取全局特征信息,但由于化工文本含噪声比例高,这会使得提取到的全局特征受到影响.因此,如今传统的特征提取方法已经无法很好地适应专业领域文本的分类任务,亟需针对化工领域产品特点设计一种专业性强的文本分类方法.

为了准确、高效地对化工产品文本特征进行表征,有效获取化工文本特征语义信息,针对现有文本分类方法应用在化工文本分类任务效果欠佳的问题,本文提出一种融合多粒度动态语义表征的文本分类模型.本文贡献可总结为以下三点:

1)由于化工文本具有较强专业性以及复杂多样性,仅仅依靠现有的词向量表征方式,很难对其中的专业术语以及其他化工领域内相关字词的语义进行充分表征,从而导致分类任务准确率不高.针对这一问题,本文提出了一种融合多粒度动态语义表征的文本分类模型.该模型受对抗训练思想启发,将对抗扰动引入动态词向量训练过程中,进一步提升化工词向量表征能力,使用多头自注意力更好地突出化工专业名词特征的权重,并针对下游分类任务提出了一种多尺度残差收缩深层金字塔形的卷积神经网络和混合注意力双向LSTM胶囊网络模型进行化工文本深度特征提取,有效提升了化工领域文本分类任务的准确性.

2)针对化工长文本含噪比例高,从而会导致文本特征提取困难的问题,提出MSRS-DPCNN模型应用于文本分类任务下游.通过将深度残差收缩网络中的注意力机制与软阈值机制引入到DPCNN模型[9]的残差连接中,减少化工文本中噪声对特征提取的影响,增强模型对噪声的抑制能力,使得模型对于含噪声比例较高的化工领域文本样本具有较好的鲁棒性.实验表明该模型可以有效提取含噪化工文本中的长距离关键依赖信息.

3)考虑到对于逻辑性差、结构性弱的化工文本,其空间语义信息本就包含较少,而MSRS-DPCNN模型在池化的过程中又无法有效提取化工文本结构空间语义信息,从而导致模型分类效果差的问题,提出HAC-BiLSTM模型,引入胶囊网络增强空间语义信息提取能力,并通过去除原胶囊网络中卷积特征提取模块,改用可以更好保留化工长文本上下文语义特征的双向循环神经网络特征提取模块,使得HAC-BiLSTM保留空间语义信息能力得到进一步提升,最终实现对整个化工文本上下文空间语义信息的高效提取.

1 相关工作

深度学习技术的迅猛发展使得神经网络模型在自然语言处理任务的许多应用领域中都有极佳表现,因而逐渐受到研究学者们的关注,大量基于神经网络的算法被应用在文本分类等任务上.

Kim[10]提出TextCNN通过对文本表示进行一维卷积的形式来获取句子中的多尺度特征表示信息,只使用了一层卷积与一层最大池化,最后通过全连接层输出分类.尽管该模型对文本表示的浅层特征的提取性能很强,但由于隐藏层太浅,仍然不足以提取出更高层特征,并且也没有解决CNN模型的通病,即模型无法充分获取上下文语义信息.Zeng

等[11]使用深层卷积神经网络进行语义特征提取,充分利用卷积深度捕捉文本语义信息,该方法摒弃了传统特征抽取环节中对各种处理工具的依赖,从而带来了文本分类在准确性上的提升.Liu 等[12]提出了一个基于 RNN 的多任务结构,多任务结构由三个包含多层 LSTM(Long Short-Term Memory)的模型组成,克服了 CNN 由于感受野大小固定,很难完全采集到文本的所有信息的缺点,并且多个多层 LSTM 也能较好地提取深层语义特征.Yang 等[13]提出一种基于分层注意力的网络模型,在词级编码和句子级编码的过程中引入注意力机制,充分考虑到了文本之间的相关性,最终模型效果均超过 LSTM、TextCNN 等模型.

考虑到卷积结构抽取特征的过程中会丢失大量空间信息,且无法关注到语序结构对字词之间的影响,Sabour 等[14]提出采用胶囊网络大量保留空间要素信息.贾旭东等[15]将可以融合多通道特征的多头注意力机制引入到胶囊网络中进行文本分类,通过该机制编码文本中的字词间依赖关系,获取长距离词间关联信息,验证了多头注意力机制以及胶囊网络在文本分类任务上的可行性.林悦等[16]将胶囊网络引入到跨领域文本分类中,设计了额外的胶囊网络层辅助目标领域的适应,有效提高了跨领域情感分类任务精度.

然而上述这些算法大多还是基于传统静态词向量的文本分类方法,静态词向量无法很好适应语境变化带来的语义变化,语义表达过于死板,单纯的静态词向量表征方式已经无法满足文本分类的要求.因此基于动态词向量的分类方法逐渐受到研究者们的关注,Li 等[17]提出一种基于 BERT 和特征融合的文本自动分类方法.该方法通过 BERT 预训练模型生成具有更丰富语境信息的动态词向量,然后用特征融合的方法充分利用 CNN 提取局部特征以及 BiLSTM 利用内存进行链接的优势,来更好地表征文本的语义信息,从而提高中文文本分类任务的准确性.

对抗训练[18]最早于 2015 年被提出并应用在图像领域.研究发现,通过向图像样本[19]中添加微小扰动得到对抗样本,使得模型经过训练修复扰动产生的误差,从而可以使得模型鲁棒性有所提升.鉴于文本数据不同于图像数据,是一种离散型数据,Miyato 等[20]提出将对抗训练的思想应用在文本模型的词嵌入层上,实验结果表明,在多个任务中都使得模型的性能得到提升.受此启发,本文将对抗扰动同样加入到词嵌入层中,不同于传统静态词向量,而是加入到动态词向量中,文本表示可以始终随着模型训练而调优,使得文本表示的鲁棒性得到提高.

深度残差收缩网络[21]继承了残差收缩网络的优点,同时集成了注意力机制与软阈值化,被广泛应用于图像领域进行样本降噪.由于化工领域产品文本具有噪声比例高的特性,本文将深度残差收缩网络加入到下游卷积神经网络结构中,抑制噪声文本对于模型分类产生的不利影响.

综上,考虑到化工领域文本的特殊背景,单一种类的神经网络使用静态词向量进行文本表示的不能充分表征文本信息,这些网络结构无法很好地在化工领域文本分类任务中发挥作用.

提高化工领域文本分类任务精度的关键在于如何有效地考虑到因为其特殊领域背景而与常规领域文本之间产生的数据差异鸿沟.本文利用对抗扰动与动态词向量对文本信息的强表征能力,降低模型在预处理词向量建模过程中无法有效处理专业名词而带来的负面影响,同时构建深度模型结构提取长距离关键依赖关系,并应用深度残差结构抑制化工文本噪声,使用提出的混合注意力的双向 LSTM 结合动态路由胶囊网络结构提取保留全局空间语义信息,从而得到融合了长距离依赖局部关键信息和全局空间语义信息的多粒度特征表达,有效解决化工领域文本分类准确率低的问题.

2 融合多粒度动态语义表征的文本分类模型

本文提出的融合多粒度动态语义表征的文本分类模型主要由生成动态对抗词嵌入的 MacBERT、进行权重强化调整的多头自注意力模型、进行关键语义信息深度抽取的 MSRS-DPCNN 模型、构建全局空间语义要素的 HAC-BiLSTM 模型、特征融合层和输出层构成.其模型结构如图 1 所示,下面将对各层进行详细阐述.

2.1 动态对抗词嵌入生成

考虑到由于化工领域文本专业性较强、文本篇幅长且词间逻辑联系性较差的特点,传统静态词向量很难充分捕捉到化工领域字词间的各种语义联系,进而无法有效地表征化工领域文本的语义信息.因此,模型采用 MacBERT 模型生成动态词向量,并且在词向量动态训练过程中加入对抗扰动,进一步

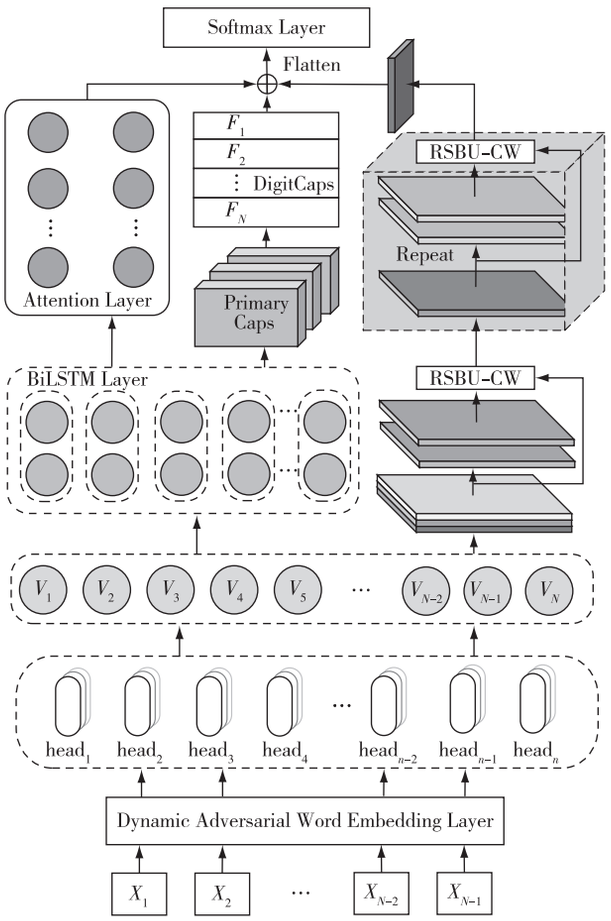

图 1　融合多粒度动态语义表征的文本分类模型

Fig. 1　Text classification model incorporating multi-granularity dynamic semantic representation

提升生成的化工文本词向量的鲁棒性以及表征能力,由此生成动态对抗词嵌入.MacBERT 模型是在 BERT 基础上提出的一种用于中文文本的预训练语言模型,该模型同样采用双向 Transformer 结构.为了提升动态词向量的表征能力,在词向量训练过程中加入对抗扰动[22],具体过程如下所示:

设输入文本序列矩阵为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{l(X)}\}$,$l(\boldsymbol{X})$ 为 $\boldsymbol{X}$ 中序列长度,输入预训练好的 MacBERT 进行向量化处理.模型对 $\boldsymbol{X}$ 进行 tokenization 分词并转化为向量,然后混合句子编码和位置编码输入到 transformer 中,在此过程中叠加对抗扰动进行计算.对抗扰动计算公式具体如下:

$$\Delta\boldsymbol{x}=\epsilon\cdot\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2},\tag{1}$$

$$\boldsymbol{x}=\boldsymbol{x}+\Delta\boldsymbol{x},\tag{2}$$

$$\boldsymbol{g}=\nabla_x L(\boldsymbol{x},y;\theta),\tag{3}$$

式(1)中 $\Delta\boldsymbol{x}$ 表示扰动值,$\|\cdot\|_2$ 表示计算 2 范数,$\boldsymbol{g}$ 表示求解的梯度,ϵ 表示权重参数,用于控制产生大对抗扰动的幅度;式(2)表示对抗样本的建立;式(3)中 L 表示预训练语言模型的损失,∇_x 表示对损失函数求偏导,$\boldsymbol{x}$ 表示添加过扰动后的迭代输入,y 表示真实标签,θ 表示模型参数.

最终经过对抗训练后的词向量序列 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n,\cdots,\boldsymbol{s}_{l(X)}\}$,$\boldsymbol{s}_n$ 是第 n 个文本的输出向量表示.

2.2　权重强化调整

注意力机制最早在机器翻译任务领域取得成功[23].为了进一步地优化所生成的词向量对化工领域文本的语义表征能力,词向量通过注意力机制对字词权重进行重新分配,从而获得化工文本字词在全局上的深层语义信息,缓解化工文本字段间联系性差、逻辑性弱的问题.多头注意力机制通过线性变换、分割操作、多头线性投影、子空间注意力计算以及最后的拼接五个操作,实现对不同子空间中提取的关键特征进行交互,更好地关注化工文本中更为重要的语义信息,模型结构如图 2 所示.

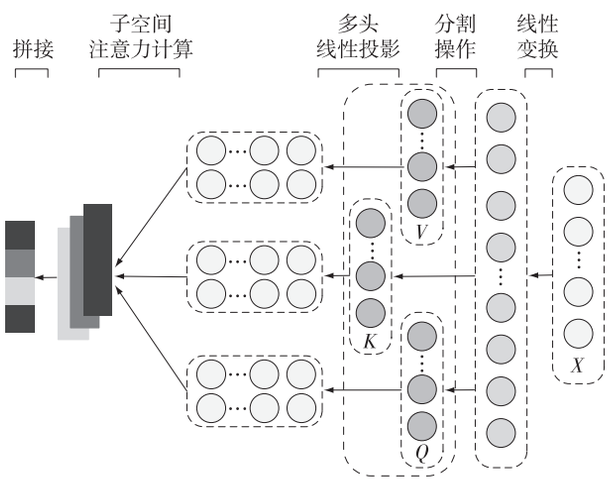

图 2　多头注意力模型

Fig. 2　Multi-headed attention model

文本序列经过训练语言模型向量化处理后生成的动态词向量,依旧可以进一步提取语义特征.为了使动态词向量在往后的模型训练过程中获取除去上游预训练语言模型以外的模型归纳偏置,继续使用多头注意力机制二次强化调整词向量之间的权重.

输出词向量 $\boldsymbol{M}_0$ 已经进一步加强了对化工文本中关键特征的权重,将其和动态词向量 $\boldsymbol{S}$ 进行残差连接得到最终输出的词向量序列 $\boldsymbol{E}$:

$$\boldsymbol{E}=\boldsymbol{M}_0+\boldsymbol{S}.\tag{4}$$

2.3 关键语义信息深度抽取

化工领域长文本含噪比例较高,仅仅通过一般的浅层卷积结构很难在充分摒除噪声影响的同时提取到长文本特征以及上下文语义间的联系.为此本文提出一种多尺度残差收缩深层金字塔形的卷积神经网络模型(Multi-Scale Residual Shrinkage Deep Pyramid Convolutional Neural Networks, MSRS-DPCNN),通过不断加深卷积网络深度,在抑制噪声的同时对化工词向量序列中的长距离依赖关键信息进行有效抽取,模型结构如图 3 所示,其中 k 为输入词向量维度.

MSRS-DPCNN 模型考虑到化工文本词间联系弱进而会导致语义连贯性差的问题,所以模型在初始进行卷积时,进行了不同尺度的卷积拼接操作,用以获得更多尺度的特征信息,增强词间语义联系,囊括更多语义信息.具体公式如下:

$$\boldsymbol{c}_i = f(\boldsymbol{W}_1 \cdot \boldsymbol{E} + \boldsymbol{b}_1), \tag{5}$$

$$\boldsymbol{C} = \mathrm{concat}(c_1, c_2, \cdots, c_i), \tag{6}$$

其中,$\boldsymbol{c}_i$ 表示第 i 个卷积操作的输出,$\boldsymbol{E}$ 表示输入向量序列,$\boldsymbol{C}$ 表示多种卷积尺度的拼接操作最终输出结果.

同时,为了增强模型对化工文本中噪声的抵抗能力,模型在残差连接之间使用了改进的残差收缩模块(Residual Shrinkage Building Unit with Channel-Wise thresholds, RSBU-CW)[21].RSBU-CW 模型结构如图 4 所示.

RSBU-CW 利用注意力机制来生成软阈值函数所需的阈值,实现对化工文本中噪声的弱化乃至消除处理.逐通道阈值化使得其能更好关注不同通道中的重要特征,而软阈值化是信号降噪处理中的常用算法.通过软阈值化机制收缩输入的特征,当特征值低于注意力机制生成的阈值时,可以认为这部分特征即为噪声,对这部分特征进行置零消除,其他部分特征会得到保留,通过这种方式可以实现对输入特征向量的降噪处理,其公式如下:

$$\boldsymbol{y} = \begin{cases} \boldsymbol{x} - \tau, & \boldsymbol{x} > \tau, \\ 0, & -\tau \leqslant \boldsymbol{x} \leqslant \tau, \\ \boldsymbol{x} + \tau, & \boldsymbol{x} < -\tau, \end{cases} \tag{7}$$

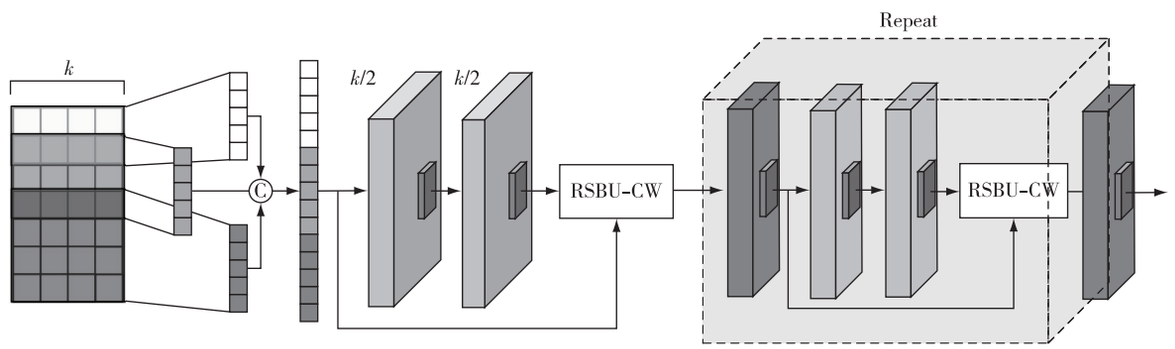

图 3 多尺度残差收缩深层金字塔形的卷积神经网络

Fig. 3 Multi-scale residual shrinkage deep pyramid convolutional neural networks

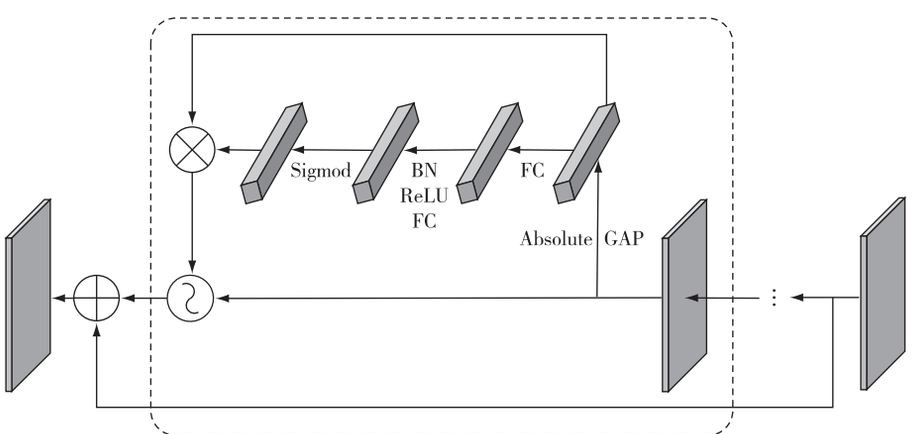

图 4 改进的残差模块

Fig. 4 Residual shrinkage building unit with channel-wise thresholds

式(7)中 $\boldsymbol{x}$ 表示输入特征向量,$\boldsymbol{y}$ 表示输出特征向量,τ 为不同特征向量下注意力机制产生的自适应阈值.

最后得到 MSRS-DPCNN 模型的输出向量 $\boldsymbol{M}_{\mathrm{DP}}$.

2.4 全局空间语义要素构建

在进行文本特征提取的过程中,考虑到上下文语义信息对于篇幅较长的化工文本尤为重要,依靠单一卷积结构只能关注到局部关键特征,并且在池化的过程中还会丢失大量的空间语义信息,反映到文本序列中就是词的上下文位置顺序等空间信息丢失.而化工文本本身蕴含的空间信息就少,因此如何有效捕捉这些信息对于提升化工文本分类精度就显得更为关键.而前人的工作中也验证了胶囊网络可以有效保留特征空间结构信息[24],因此,本文提出了一种混合注意力胶囊双向 LSTM 模型(Hybrid Attention Capsule Bidirectional LSTM network model,HAC-BiLSTM).其模型结构如图 5 所示.

通过 BiLSTM 与注意力机制捕获化工文本中隐含的全局语义信息并对关键信息权重进行加强,弥补卷积结构无法充分关注上下文信息的缺点.同时,由于上文构建的 MSRS-DPCNN 模型进行的卷积和池化操作会丢失了大量空间语序结构信息,因此在 HAC-BiLSTM 模型中构建了胶囊网络模型,用以保留并获取相关文本的空间要素信息.

2.4.1 全局语义信息构建

为了有效获取化工长文本中的上下文语义信息,模型选择 BiLSTM 对输入进行双向的特征计算,相比于传统 LSTM 结构,BiLSTM 很好地解决了序列化处理输入而无法有效地获取上下文信息的问题[25],然后将得到的正反双向隐层状态序列表示 $\overrightarrow{\boldsymbol{H}_i}=\{\boldsymbol{h}_{i0},\boldsymbol{h}_{i1},\cdots,\boldsymbol{h}_{i(n-1)}\}$,$\overleftarrow{\boldsymbol{H}'_i}=\{\boldsymbol{h}'_{i0},\boldsymbol{h}'_{i1},\cdots,\boldsymbol{h}'_{i(n-1)}\}$ 合并拼接得到 $\boldsymbol{h}_t$,随后送入激活函数中,得到输出特征向量.

2.4.2 全局语义注意力权重

由于 BiLSTM 在对化工文本特征提取过程中仍然会存在一定程度上的梯度弥散以及上下文语义不充分的问题,模型将对 BiLSTM 输出进行进一步地注意力加权操作,提高关键特征的权重,详细计算过程如下所示,最终得到输出的特征向量为 $\boldsymbol{V}_{\mathrm{att}}$.

$$\boldsymbol{h}'_n=\tanh(\boldsymbol{W}_2\boldsymbol{h}_t+\boldsymbol{b}_2),\tag{8}$$

$$a_n=\frac{\exp(\boldsymbol{h}'_n\boldsymbol{W}_3)}{\sum_{j=1}^{N}\boldsymbol{h}'_n\boldsymbol{W}_3},\tag{9}$$

$$\boldsymbol{V}_{\mathrm{att}}=\sum_{n=1}^{N}a_n\boldsymbol{h}_t,\tag{10}$$

其中,$\boldsymbol{h}_t$ 是 BiLSTM 的输出词向量,$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 分别是权重矩阵和偏置,$\boldsymbol{h}'_n$ 为经过 tanh 激活函数处理后的词向量,$\boldsymbol{W}_3$ 为权重矩阵,a_n 词注意力概率权重分布,即词的重要性信息,$\boldsymbol{V}_{\mathrm{att}}$ 表示经过词的加权平均后的词向量特征表示.

2.4.3 全局语义空间要素

考虑到 MSRS-DPCNN 模型在使用卷积模块对化工文本进行深度特征提取时会丢失大量空间信息,本文引入改进的胶囊网络缓解这一问题.本文模型丢弃了原胶囊网络中的卷积层转而使用 BiLSTM 进行底层特征抽取,BiLSTM 可以充分建模上下文全局语义信息,即可以关注到某个字词在整句话中的位置语序关系.胶囊网络最先被应用于图像领域,局部关键信息相对来说更为重要,因此会选择使用卷积对文本建模提取特征.这使得胶囊网络在被应用于自然语言处理领域时,只能关注到某个字词在局部一段话中的相对特征信息,很难获取全局语义信息,而在文本特征中上下文语序信息(全局信息)是十分重要的,因此本文模型局部关键信息由上文的 MSRS-DPCNN 模型抽取,使用 BiLSTM 代替胶囊网络中的卷积进行特征提取,从而使得文本空间语序要素可以进一步被保留.

胶囊网络由主胶囊层与数字胶囊层构成,层间通过动态路由算法进行联系.设抽取到的特征为 $\boldsymbol{B}_i$,则胶囊网络的输出为 $\boldsymbol{V}_{\mathrm{cap}}$.具体计算过程如下:

首先,胶囊网络为了更好地保留空间要素,选择

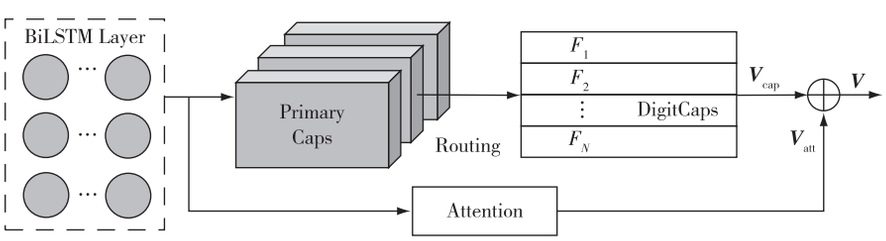

图 5 HAC-BiLSTM 层结构

Fig. 5 HAC-BiLSTM layer structure

使用矢量输出代替传统卷积操作中的标量输出.主胶囊层的计算如式(11)所示,$\boldsymbol{u}_i$ 表示第 i 个通过卷积操作生成的胶囊向量,实现将 $\boldsymbol{B}_i$ 特征映射到 $\boldsymbol{u}_i$ 的过程.

$$\boldsymbol{u}_i = \text{squash}(\boldsymbol{W}_3 \cdot \boldsymbol{B}_i + \boldsymbol{b}_3). \tag{11}$$

其次,为了获得分类运算所需的概率预测向量,胶囊网络通过一个 squash 挤压函数实现对向量的压缩,由此即开始动态路由的计算,详细计算过程如式(12)至式(16)所示:

$$\boldsymbol{v}_j = \text{squash}(\boldsymbol{s}_j) = \frac{\| \boldsymbol{s}_j \|^2}{1 + \| \boldsymbol{s}_j \|^2} \frac{\boldsymbol{s}_j}{\| \boldsymbol{s}_j \|}, \tag{12}$$

式(11)与式(12)中 squash 表示挤压函数.在式(11)中利用挤压函数对向量 $\boldsymbol{s}_j$ 进行压缩,规范其长度,使得 $\boldsymbol{v}_j$ 始终保持在(0,1)之间,从而可以根据 $\boldsymbol{v}_j$ 进行概率判断.

$$\boldsymbol{s}_j = \sum_i \boldsymbol{c}_{ij} \hat{\boldsymbol{u}}_{j|i}, \tag{13}$$

$\boldsymbol{s}_j$ 的计算需要对预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 进行加权求和,并计算胶囊层 i 总输入 $\boldsymbol{s}_j$,耦合系数 $\boldsymbol{c}_{ij}$ 与预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 的计算方法为

$$\boldsymbol{c}_{ij} = \frac{\exp(\boldsymbol{b}_{ij})}{\sum_k \exp(\boldsymbol{b}_{ik})}, \tag{14}$$

$$\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_j \boldsymbol{u}_i, \tag{15}$$

式(15)中预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 由主胶囊层的输出 $\boldsymbol{u}_i$ 经过权重矩阵 $\boldsymbol{W}_j$ 加权计算得到;式(14)中待更新权重 $\boldsymbol{b}_{ij}$ 的计算公式为

$$\boldsymbol{b}_{ij} = \boldsymbol{b}_{ij} + \hat{\boldsymbol{u}}_{j|i} \boldsymbol{v}_j, \tag{16}$$

$\boldsymbol{b}_{ij}$ 经过预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 与输出向量 $\boldsymbol{v}_j$ 一致性计算迭代更新.

2.5 特征融合

利用集成学习的方式,将 MSRS-DPCNN 模型提取局部关键特征以及部分长距离依赖特征与 HAC-BiLSTM 模型提取的全局上下文语义关系特征进行特征融合,如式(17)与(18)所示:

$$\boldsymbol{V}_{\text{HAC}} = \boldsymbol{V}_{\text{att}} + \boldsymbol{V}_{\text{cap}}, \tag{17}$$

$$\boldsymbol{G} = \text{concat}(\boldsymbol{M}_{\text{DP}}, \boldsymbol{V}_{\text{HAC}}), \tag{18}$$

其中:$\boldsymbol{V}_{\text{cap}}$ 与 $\boldsymbol{V}_{\text{att}}$ 为 HAC-BiLSTM 模型提取的两种全局语义信息,融合后得到 $\boldsymbol{V}_{\text{HAC}}$ 为 HAC-BiLSTM 模型输出的向量;$\boldsymbol{M}_{\text{DP}}$ 为 MS-DPCNN 模型输出的特征向量,与 $\boldsymbol{V}_{\text{HAC}}$ 拼接后得到特征融合层的输出向量 $\boldsymbol{G}$.

2.6 输出

将前面通过特征融合得到的特征向量输入全连接层进行调整得到 $\boldsymbol{H}$:

$$\boldsymbol{H} = \text{liner}(\boldsymbol{G}), \tag{19}$$

随后传入 softmax 层进行分类,得到最终分类结果.

3 实验过程与分析

3.1 实验环境与数据

本文模型基于 Pytorch 1.6 实现,运行环境为 Ubuntu 18.04.3,GPU 为 1 块 Tesla V100(16 GB),编程语言为 Python 3.7.

化工领域产品数据收集自中国化工制造网(http://www.chemmade.com)、化工产品网(http://www.chemcp.com)以及盖德化工网(https://china.guidechem.com)等国内几家较大的化工化学类交易平台的化工产品信息,共包含有 221 216 条带有标签的化工领域产品文本数据,平均文本长度 261.43 字,标签种类分为 17 种,样本类别之间数量比例分布不均衡,最高达到 130:1.文本数据涵盖了主要化工产品分布领域,包括有机原料、化工试剂、化工中间体、化学矿、无机化工、农业化工、涂料油漆、聚合物、染料、食品添加剂、生物化工、香精、胶粘剂、日用化工、催化剂以及植物提取物.

上述数据按照 6:2:2的比例切分为训练集、验证集以及测试集,数据集的数据格式如表 1 所示.

表 1 化工产品数据格式

Table 1 Chemical product data format

文本内容	类别
15318-45-31 * 25 振华甲砜霉素,硫霉素,甲砜氯霉素 thiamphenicol C12H14Cl2N2O7S 分子量 Mr:401.22 上游产品:氨基乙酸、对甲砜基甲苯、二氯醋酸甲酯、溴素、乙醇 主要用于治疗呼吸、泌尿、肝胆、伤寒等肠道外科、妇产科和五官科感染等症,特别对中轻度感染作用尤其明显	生物化工

为了对模型进行泛化性评估,额外在 THUCNews 和 ChnSentiCorp 两个中文公开数据集上进行实验.三个数据集详细信息如表 2 所示.

表 2 数据集详情

Table 2 Dataset details

数据集	类别	数据量/条	文本平均长度/字	类间样本数量最大比例
Che_products	17	221 216	261.43	130:1
THUCNews	10	200 000	22.34	1:1
ChnSentiCorp	2	7 765	128.52	2.2:1

THUCNews(http://thuctc.thunlp.org)随机抽取20万条数据,涉及财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐共计10个类别,每个类别2万条,平均数据长度22.34字,属于短文本数据集;ChnSentiCorp(https://github.com/SophonPlus/ChineseNlpCorpus)是酒店评论数据集,一共分为正面和负面2个评价类别7 765条数据,其中,正面评价5 322条,负面评价2 443条,平均数据长度128.52字,属于长文本数据集.

3.2 数据预处理

具体的数据预处理主要包括以下几个步骤:

1)数据集清洗.此步骤主要包括去除重复出现的无意义字词(例如:啊、呃、呢、用途、性状、外观等)、去除多余空白、去除回车换行符和制表符以及繁简体的统一.

2)选择性中文分词.将数据集进行分词用于生成静态词向量,对所清洗好的中文领域数据集利用jieba(https://github.com/fxsjy/jieba)分词工具进行中文分词,此处选用的停用词表为哈尔滨工业大学停用词表.

3.3 实验参数设置

具体参数设置如表3所示.

表3 模型参数设置

Table 3 Model parameter setting

超参数名称	参数值
文本最大长度/字	128
batch_size	8
学习率	5e-5
dropout	0.5
注意力头数/个	12
卷积核1大小	{5,6,7}
卷积核1数量/个	128
卷积核2大小	3×3
卷积核2数量/个	128
BiLSTM隐藏节点/个	300
胶囊网络迭代次数/次	3

3.4 评价指标

本文分别采用精确率(Precision,P)、准确率(Accuracy,A)以及F1值作为评价指标用以对模型的分类效果进行评价.

精确率指的是在所有预测为正例的样本中,预测正确的样本所占的比例,主要用于验证特征提取效果和计算F1值,计算公式为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{20}$$

准确率指模型预测正确样本数占样本总数的比例,计算公式为

$$A=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{21}$$

召回率(Recall,R)指在所有真实为正例的样本中预测正确的样本所占的比例,计算公式为

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{22}$$

F1值用于结合精确率和召回率,对模型效果进行综合评价,计算公式为

$$\mathrm{F1}=2\times\frac{P\times R}{P+R}. \tag{23}$$

其中:TP为真正例,表示实际为正例且预测为正例;FP为假正例,表示实际为负例但预测为正例;TN为真负例,表示实际为负例且预测为负例;FN表示假负例,表示实际为正例但预测为负例.

3.5 实验结果分析

3.5.1 模型有效性评估

为了验证提出的模型在化工领域产品数据集上的有效性,本文将模型与TextCNN[10]、DPCNN[9]、BiLSTM[12]、Capsule Network[26]四个基线模型以及三个多阶段模型进行了实验对比,实验结果如表4所示.

表4 化工领域文本实验结果对比

Table 4 Comparison of text classification results for chemical field %

词向量模型	分类模型	准确率	F1-Score
Word2Vec[2]	TextCNN	68.43	67.46
	DPCNN	68.64	68.19
	BiLSTM	69.61	69.13
	Capsule	68.24	67.90
	Ours	**75.06**	**74.71**
GloVe[27]	TextCNN	69.53	68.65
	DPCNN	69.01	68.50
	BiLSTM	70.85	70.47
	Capsule	67.57	67.09
	Ours	**74.27**	**74.20**
MacBERT	MacBERT	84.24	84.24
	MacBERT+BiLSTM+Attention	82.62	82.41
	MacBERT+Capsule+BiLSTM+Attention	83.44	83.36
	MacBERT+DPCNN+BiLSTM+Attention	83.85	83.73
	Ours	**84.79**	**84.62**

从表4可以看出,针对化工领域类文本数据,使用动态词向量能更好地提取文本表示,从而有效提高模型性能.本文模型在使用动态词向量的情况下相比于仅使用静态词向量,F1-Score值分别上升了9.91和10.42个百分点,和仅使用MacBERT中文预训练语言模型相比F1-Score上升了0.38个百分点.在多阶段模型中加入胶囊网络可以一定程度上提升模型性能,可能原因是化工类文本逻辑性较弱且碎片化分布,单靠BiLSTM提取全局语义信息,无法兼顾到局部碎片化文本中的语序信息,而加入胶囊网络可以有效弥补这一点.相较于原始MacBERT模型,三种多阶段基线模型性能均出现不同程度下降,并且下接网络越简单,模型性能下降越显著.可能原因是预训练语言模型参数量过多,而下接的网络由于参数量较小并且仅仅是简单的模型拼接,并未考虑到不同下接模型特征提取方式的优缺点,以及特殊领域数据背景对模型性能产生的影响[28],故对接后很难充分发挥前者的优异性能,甚至会产生干扰,导致模型性能下降.

同时,从表4数据可知,相较于几组基线模型,本文构建的模型在评价指标上均达到最优,在使用静态词向量时,本文模型较单阶段基线模型中最优模型F1-Score分别提升了5.58个百分点和3.73个百分点.在使用动态词向量时,本文模型较多阶段基线模型中最优模型提升了0.89个百分点.可以看出本文构建的模型可以更好地适应化工领域文本分类任务,并提升分类任务精度.

为了更有效地说明模型各部分的作用,进行了模型消融实验,实验结果如表5所示.其中α代表词消融对抗扰动机制,β代表消融多头注意力机制,γ代表消融MSRS-DPCNN模型,δ代表消融HAC-BiLSTM模型,δ^*代表在消融HAC-BiLSTM模型基础上继续对降噪模块RSBU-CW进行消融的模型.

表5 消融实验结果

Table 5 Results of ablation experiments %

组别	α	β	γ	δ/δ^*	F1-Score
对照	✓	✓	✓	✓	84.62
1		✓	✓	✓	83.98
2	✓		✓	✓	84.17
3	✓	✓		✓	83.94
4	✓	✓	✓		83.61/83.32

通过表5实验数据可知,第1组消融对抗扰动机制使得模型F1-Score下降0.64个百分点,模型性能出现较大幅度下降,主要原因可能是该机制可以提高模型鲁棒性,降低过拟合风险,而本文构建的化工领域文本分类模型模块较多,参数量较大,因此消融了可以提升模型鲁棒性的对抗扰动机制会让模型性能下降较大.第2组消融多头注意力机制使得模型F1-Score下降0.45个百分点,多头注意力机制可以进一步优化上一层生成的词向量对化工领域文本的语义表征能力,因此消融该部分同样会对模型性能产生影响.第3组消融了MSRS-DPCNN模型使得模型F1-Score下降0.68个百分点.由于该模型负责对化工长文本进行深度特征提取,属于重要的特征提取模块,因此消融该部分同样对模型整体性能产生较大影响.第4组先是对HAC-BiLSTM模型进行消融实验,模型F1-Score下降1.01个百分点,在此基础上继续对MSRS-DPCNN模型中的降噪模块进行消融,模型F1-Score继续下降0.29个百分点,模型整体性能出现大幅下降,这表示该部分模型提取的上下文语义信息以及构建的空间语序等结构信息,对进行化工领域这类特殊背景的文本分类有着至关重要的地位,同时降噪模块也在一定程度上起到了抑制文本中噪声干扰的能力.

综上所述,本文提出的融合多粒度动态语义表征的文本分类模型对于化工领域文本分类任务有较好的性能表现,通过抽取关键语义信息、全局语义信息以及空间要素这些不同粒度的语义表征可以有效提升分类任务精度.

3.5.2 模型泛化性评估

为了验证模型在中文文本分类任务上的泛化性能,本文在THUCNews和ChnSentiCorp两个中文公开数据集上进行实验,实验结果分别如表6、表7所示.

表6 THUCNews数据集实验结果

Table 6 Model experimental results on THUCNews dataset %

数据集	模型	准确率	F1-Score
THUCNews	TextCNN	90.56	90.56
	DPCNN	91.15	91.12
	BiLSTM	90.26	90.31
	Capsule	90.64	90.64
	Ours-Word2Vec	**91.47**	**91.46**
	MacBERT+BiLSTM+Attention	92.92	92.92
	MacBERT+Capsule+BiLSTM+Attention	93.05	93.03
	MacBERT+DPCNN+BiLSTM+Attention	93.88	93.88
	MacBERT	94.57	94.56
	Ours	**94.21**	**94.20**

从表6中数据可知,本文提出的模型在THUCNews数据集上性能略低于MacBERT模型,主要原因是本文模型针对化工领域文本进行了针对性设计.THUCNews数据集与化工领域文本特性相差过大,属于短文本,所含关键特征较少,并且本文所构建的MacBERT模型下接结构较为复杂,对较短的文本会产生语义过度解读,同时其中的降噪机制亦会对短文本中特征的提取有一定抑制,因此本文提出模型相比较于单纯使用预训练语言模型性能有一定下降.而在消融预训练语言模型仅使用静态词向量时,本文提出的下接结构可以有效提升模型性能,相比于最优基线模型提升了0.34个百分点,这表明模型在使用静态词向量时的下接结构在短文本数据集上可以拥有的良好泛化性能.

而对于ChnSentiCorp数据集,从表7可以看出,本文构建的模型即使是在使用静态词向量的情况下,准确率和F1-Score分别提升1.35和1.31个百分点.与其他三个多阶段模型相比,准确率和F1-Score分别提升0.45和0.60个百分点,模型性能提升较明显,主要是由于ChnSentiCorp数据集与化工领域数据集都属于长文本数据集,而本文构建的网络可以很好地提取长文本中的特征,因而模型性能表现较好.

表7 ChnSentiCorp数据集实验结果

Table 7 Model experimental results on ChnSentiCorp dataset

%

数据集	模型	准确率	F1-Score
ChnSentiCorp	TextCNN	84.74	84.57
	DPCNN	84.93	84.83
	BiLSTM	84.22	84.02
	Capsule	85.00	84.75
	Ours-Word2Vec	**86.35**	**86.06**
	MacBERT+BiLSTM+Attention	89.89	89.94
	MacBERT+Capsule+BiLSTM+Attention	90.21	90.24
	MacBERT+DPCNN+BiLSTM+Attention	90.41	90.26
	MacBERT	90.60	90.60
	Ours	**90.86**	**90.86**

从三个多阶段模型在两个公开数据集上的实验结果可以看出,胶囊网络在长文本数据集上可以发挥出更好的优势.加入胶囊网络的多阶段模型在ChnSentiCorp数据集上的F1-Score指标比THUCNews数据集提升0.19个百分点,可能是因为长文本中字词的空间结构(语序)信息更丰富,从而使得效果提升更为明显.

综合在两个数据集上以及与七个基线模型的实验对比结果,本文提出的模型在与化工领域数据集相似数据特点的长文本数据集上具有较好泛化性能,在短文本数据集上使用静态词向量时也拥有较好表现.

3.5.3 不同文本长度对模型性能影响

为了探究不同化工文本长度对模型性能的影响,本文进行了实验对比,实验结果如图6所示.

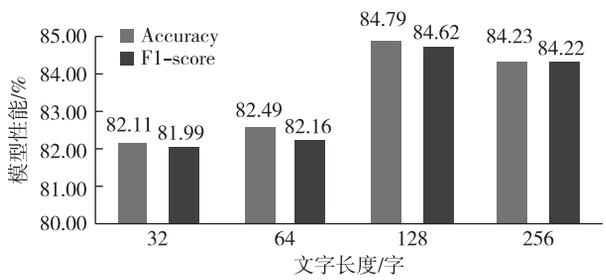

图6 文本长度对模型性能的影响

Fig. 6 Impact of text length on model performance

从图6可以看出,不同文本长度对模型性能有较大的影响.数据集的平均长度为261.43字,实验结果表明当数据长度在128字时模型性能最佳.经过对数据集分析后发现,化工产品数据集文本长度中位数为148字,文本长度最长为1 946字,最短为27字,文本长度在区间[1,128]与[129,256]之间的比例达到2.02∶1.因此,文本长度超过148字时会使得大量的短文本数据被过度填充,低于148字时会使得文本数据过度截断,因而在文本长度选择位于中位数148字附近的128字时模型性能最佳准确率达到84.79%,F1-Score达到84.62%.

4 结束语

本文描述了融合多粒度动态语义表征的文本分类模型研究,针对化工领域产品文本这类特定领域的文本数据,将MacBERT预训练语言模型作用在分类任务上游用以获取句子的动态词向量,并在其中引入对抗训练思想,增加文本表征的鲁棒性.借助多头注意力机制对文本表征二次权重调整,在任务下游利用带有抑制噪声文本数据能力的MSRS-DPCNN模型以及可以有效提取全局语义信息和空间要素的HAC-BiLSTM模型对预训练模型输出的词向量进行深度特征提取,输入分类器进行分类.将本文提出的模型与其他几种神经网络分类算法进行比较,实验

结果表明,在两个公开数据集中本方法对长文本分类任务有较好表现,较深的神经网络使得模型具有提取长距离语义依赖能力,但对于短文本,较深的网络会导致性能过剩,反而效果不佳;在化工领域的中文化工产品数据集中,本方法优于几个基线模型,提高了分类的准确性.

尽管本文提出的模型在准确性上优于其他分类方法,但由于领域类文本相对专业且往往文本数据构成复杂,这使得构建的模型通用性不强,只能针对某一领域的特定任务.未来可以通过在领域类中文本预处理的过程中引入领域专业术语库对文本进行规范化从而整体提升数据集质量,以及通过领域知识迁移等方式降低数据对模型的要求,使得模型的通用性和泛化性得到提升,从而可以应用到更多领域中.

参考文献

References

[1] 李海洋,赵国伟.2020 年中国石油和化学工业经济运行报告[J].现代化工,2021,41(3):251-253
LI Haiyang, ZHAO Guowei. China petroleum and chemical industry economic operation report 2020[J]. Modern Chemical Industry,2021,41(03):251-253

[2] Mikolov T,Chen K,Corrado G,et al.Efficient estimation of word representations in vector space[J].arXiv e-print,2013,arXiv:1301.3781

[3] Liu W K,Xiao J E,Hong M.Comparison on feature selection methods for text classification[C]//Proceedings of the 2020 4th International Conference on Management Engineering,Software Engineering and Service Sciences,2020:82-86

[4] 陈德光,马金林,马自萍,等.自然语言处理预训练技术综述[J].计算机科学与探索,2021,15(8):1359-1389
CHEN Deguang,MA Jinlin,MA Ziping,et al.Review of pre-training techniques for natural language processing[J].Journal of Frontiers of Computer Science & Technology,2021,15(8):1359-1389

[5] Devlin J,Chang M W,Lee K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019:4171-4186

[6] Lan Z Z,Chen M D,Goodman S,et al.ALBERT:a lite BERT for self-supervised learning of language representations[J].arXiv e-print,2019,arXiv:1909.11942

[7] Yang Z L,Dai Z H,Yang Y M,et al.XlNet:generalized auto-regressive pretraining for language understanding[C]//Advances in Neural Information Processing Systems,2019:5754-5764

[8] Cui Y M,Che W X,Liu T,et al.Revisiting pre-trained models for Chinese natural language processing[C]//Findings of the Association for Computational Linguistics,2020:657-668

[9] Johnson R,Zhang T.Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers),2017:562-570

[10] Kim Y.Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),2014:1746-1751

[11] Zeng D J,Liu K,Lai S W,et al.Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014:2335-2344

[12] Liu P F,Qiu X P,Huang X J.Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence,2016:2873-2879

[13] Yang Z C,Yang D Y,Dyer C,et al.Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2016:1480-1489

[14] Sabour S,Frosst N,Hinton G E.Dynamic routing between capsules[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems,2017:3856-3866

[15] 贾旭东,王莉.基于多头注意力胶囊网络的文本分类模型[J].清华大学学报(自然科学版),2020,60(5):415-421
JIA Xudong, WANG Li. Text classification model based on multi-head attention capsule neworks[J]. Journal of Tsinghua University (Science and Technology),2020,60(5):415-421

[16] 林悦,钱铁云.基于胶囊网络的跨领域情感分类方法[J].南京信息工程大学学报(自然科学版),2019,11(3):286-294
LIN Yue,QIAN Tieyun.Cross-domain sentiment classification by capsule network[J].Journal of Nanjing University of Information Science & Technology (Natural Science Edition),2019,11(3):286-294

[17] Li W T,Gao S B,Zhou H,et al.The automatic text classification method based on BERT and feature union[C]//2019 IEEE 25th International Conference on Parallel and Distributed Systems.December 4-6,2019,Tianjin,China.IEEE,2019:774-777

[18] Goodfellow I J,Shlens J,Szegedy C.Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings,2015

[19] 黄菲,高飞,朱静洁,等.基于生成对抗网络的异质人脸图像合成:进展与挑战[J].南京信息工程大学学报(自然科学版),2019,11(6):660-681
HUANG Fei,GAO Fei,ZHU Jingjie,et al.Heterogeneous face synthesis via generative adversarial networks: progresses and challenges[J].Journal of Nanjing University

of Information Science & Technology (Natural Science Edition),2019,11(6):660-681
[20] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification[J].arXiv e-print,2016,arXiv:1605.07725
[21] Zhao M H, Zhong S S, Fu X Y, et al. Deep residual shrinkage networks for fault diagnosis[J].IEEE Transactions on Industrial Informatics,2020,16(7):4681-4690
[22] Huang S,Papernot N,Goodfellow I,et al.Adversarial attacks on neural network policies[J].arXiv e-print,2017,arXiv:1702.02284
[23] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J].arXiv e-print,2014,arXiv:1409.0473
[24] 倪斌,陆晓蕾,童逸琦,等.胶囊神经网络在期刊文本分类中的应用[J].南京大学学报(自然科学),2021,57(5):750-756
NI Bin,LU Xiaolei,TONG Yiqi,et al.Automated journal text classification based on capsule neural network[J].Journal of Nanjing University (Natural Science),2021,57(5):750-756
[25] Sachan D S,Zaheer M,Salakhutdinov R.Revisiting LSTM networks for semi-supervised text classification via mixed objective function[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33:6940-6948
[26] Yang M,Zhao W,Ye J B,et al.Investigating capsule networks with dynamic routing for text classification[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics,2018:3110-3119
[27] 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 (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014:1532-1543
[28] 范红杰,李雪冬,叶松涛.面向电子病历语义解析的疾病辅助诊断方法[J].计算机科学,2022,49(1):153-158
FAN Hongjie,LI Xuedong,YE Songtao.Aided disease diagnosis method for EMR semantic analysis[J].Computer Science,2022,49(1):153-158

Text classification model incorporating multi-granularity dynamic semantic representation

ZHANG Junqiang[1] GAO Shangbing[1] SU Rui[1] LI Wenting[1]
1 School of Computer and Software Engineering/Jiangsu Internet of Things Mobile Interconnection Technology Engineering Laboratory,Huaiyin Institute of Technology,Huaian 223003

Abstract The widely used word vector representation is incapable of fully representing the specialized texts and phrases in sphere of highly specialized chemical industry,which were quite professional and complex,resulting in the low accuracy of classification. Here, we propose a text classification model incorporating multi-granularity dynamic semantic representation.First,the adversarial perturbation was introduced into the word embedding layer of the model to enhance the ability of dynamic word vectors to represent the semantics.Then the word vector weights were redistributed by a multi-headed attention mechanism to obtain a better textual representation of key semantic information.Finally,text representations of different granularities were extracted through the proposed multi-scale residual shrinkage deep pyramidal convolutional neural network (MSRS-DPCNN) and hybrid attention capsule bidirectional LSTM (HAC-BiLSTM) network model,which were then fused for classification.The experimental results showed that the proposed model achieved an F1-score up to 84.62% on the chemical domain text dataset when using different word vector representations,an improvement of 0.38−5.58 percentage points compared with existing models.The model also had pretty good generalization performance on the publicly available Chinese dataset THUCNews and the Tan Songbo hotel review dataset ChnSentiCorp.

Key words text classification;adversarial perturbation;multi-granularity;multi-head attention mechanism;deep residual shrinkage;pre-trained language models