

基于ERNIE-TextCNN模型的医学文本分类方法

卓飞豹

(上海森亿医疗科技有限公司 上海 201203)

【摘要】目前医院信息系统中还包括大量的非结构化数据,从中挖掘信息存在较大的难度。本研究采用ERNIE-TextCNN模型,对医学内镜报告进行文本分类。实验显示,相比单纯的ERNIE模型和TextCNN模型,ERNIE-TextCNN模型的精确率、召回率和F1分数上都有所提升,说明该模型能够有效地应用到医学文本分类领域。

【关键词】医学文本;文本分类;ERNIE-TextCNN模型

一、相关背景

近年来随着人工智能的发展,自然语言处理(Natural Language Processing, NLP)在医疗行业得到了广泛的应用,通过实体识别和文本分类技术,实现诊断归一、病历质控、疾病筛查等功能。医院信息系统中,还存在着大量非结构化的文本,比如住院病历、超声报告、内镜报告等,如果对这些文本实现有效的分类,不仅能够提高医生的诊疗效率,还能为将来的医学研究提供便利。

医学文本分类方面,传统的方法主要集中于机器学习和深度学习算法,赵汉青等^[1]采用LibSVM和LibLinear方法对中医学术派文本分类研究。Chen等提出一种带自注意力的双向长短期记忆模型,对门诊的文本内容进行分类。潘劲等采用支持向量机算法,对ICD-0-3中的肿瘤病理形态学自动化编码进行分类,提高了专业人员的编码效率,避免不同编码人员主观上的判断偏差。袁野等融合CNN模型和LSTM模型,构成双通道神经网络,增强了句子的整体序列特征以及局部词序特征,实现疾病文本的分类。

自然语言的预训练方法,如GPT和BERT等模型,在很多领域被证明可以有效提高实体识别和文本分类的准确性。杨飞洪等采用BERT-CNN模型,对临床试验短文本进行分类,证明了BERT与CNN两种模型的融合,相比两种单纯的模型准确率都有所提升。但是,BERT预训练模型只关注了自身的文本信息,不能全面地获取潜在的文本特征。

本研究基于ERNIE预训练模型,融合TextCNN模型,提出ERNIE-TextCNN模型,并应用到医疗领域的非结构化文本分类上,可以提高医学术语的标准化程度,并提升临床的工作效率。

二、模型构建

TextCNN源于CNN神经网络,对关键词比较敏感,能够有效地捕捉文本序列中的n-gram信息;而ERNIE模型是在BERT模型基础上优化的,对中文分类效果显著。本研究融合TextCNN和ERNIE的优点,构建ERNIE-TextCNN的自然语言预训练模型,并应用于医学文本分类。如图1所示,该模型主要分为输入层、ERNIE层和TextCNN层。

(一) 输入层

输入层将原始的文本进行预处理,形成输入文本。

输入层将原始的词句转换成词嵌入表示,并指定句子的最大序列的长度。设定最大序列长度为padding_size,每批次训练的样本量为batch_size,那么文本短句转为(batch_size, padding_size)的词向量。

(二) ERNIE层

ERNIE层主要构建ERNIE(Enhanced Representation through Knowledge Integration)模型,该模型是百度公司提出的知识增强语义表示模型,基于BERT模型做了一些优化。BERT模型的建模对象主要聚焦在原始语言信号上,没有充分利用训练数据中的句子结构和语义信息。模型预训练过程中,BERT模型随机掩盖掉15%的词进行预测,这种随机掩盖的方式,会导致很多词组被拆散了。

考虑到BERT模型在掩码方面的不足,ERNIE对掩码做了适当的改进,提出了基于短语和实体的掩码策略。同一个短语或实体内的文字,被当作一个单元。训练过程中,被掩盖的文字来自同一个单元,这样模型就可以学习到潜在的知识依赖,甚至更长

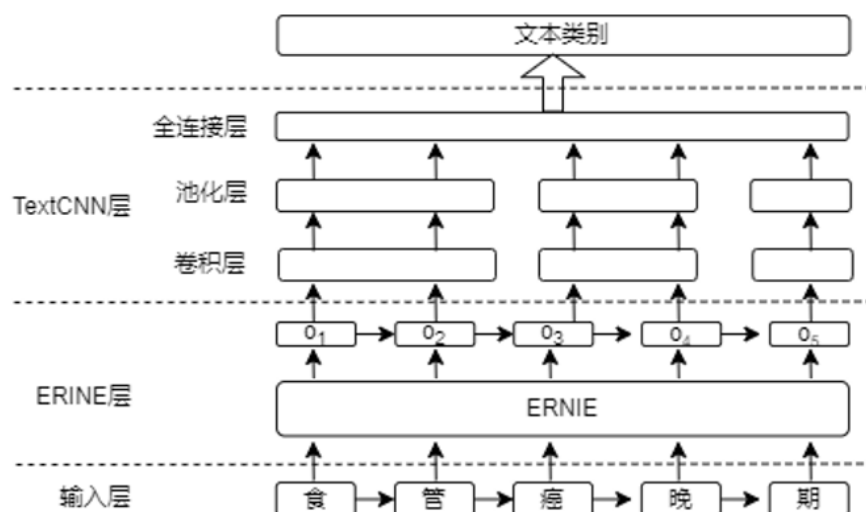


图1 ERNIE-TextCNN 模型结构

的语义关系，最终更具泛化性。如图2所示，词语“溃疡”在Bert模型的掩码中可能被拆散，但是在ERNIE模型中会被统一掩码，通过前面的“十二指肠”可以推断此处被掩盖的词语是“溃疡”。

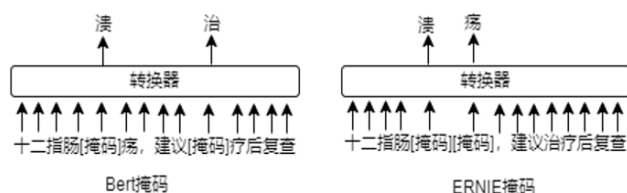


图2 Bert模型和ERNIE模型的掩码

除了以上的知识掩码外，ERNIE还使用不同来

源的语料库对模型进行训练，通过这些改进以保证模型在字词、语句和语义方面更深入地学习到语言知识。当ERNIE通过这些预训练任务学习之后，就会变成一个更懂语言知识的预训练模型，接着在不同的下游任务进行微调，提高分类的准确率。

(三) TextCNN 层

TextCNN层主要构建TextCNN模型，该模型是卷积神经网络(convolutional neural network, CNN)在文本分类领域的一个变体模型，能够高效地提取文档或句子的重要特征，而这些特征是分类的重要依据。该模型结构类似CNN，通常包括输入层、

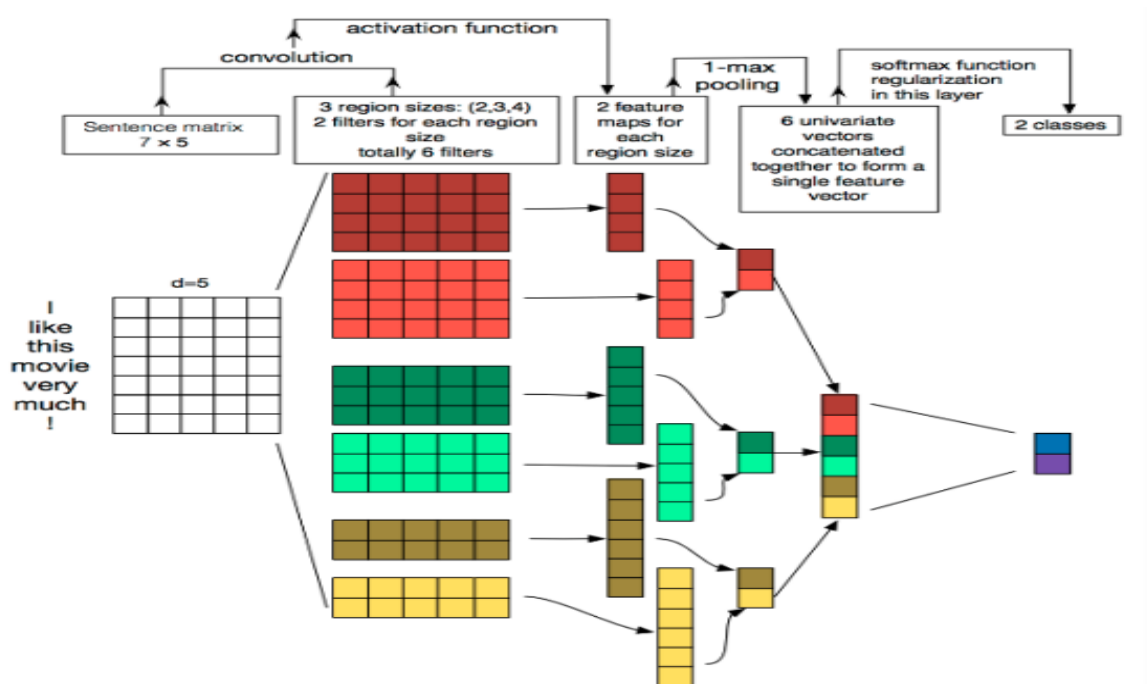


图3 TextCNN 模型结构

卷积层、池化层和全连接层，常规的网络结构如图3所示。

输入层：即词嵌入层，因为本研究的TextCNN模型是建立在ERNIE模型之上的，因此将ERNIE最后一层的输出作为这里的输入。

卷积层：主要是通过卷积提取文本的局部特征。假设输入的词向量大小为 m ，那么卷积核的宽也是 m 。卷积核的高度 n 是一个超参数，值的设定范围一般在 $2 \sim 8$ 之间。与图像卷积不同的是，这里的卷积核只进行一维的滑动，没有横向移动，只是从上向下移动。不同尺寸的卷积核能够提取不同的特征，因此模型中一般有多种不同尺寸的卷积核。假设模型有三种尺寸的卷积核，高度 n 分为2、3、4，每种卷积核有128组，那么卷积层会输出 3×128 个一维向量。

池化层：用于压缩数据和参数的量，同时让模型更加关注某些特征，提高模型的泛化能力。TextCNN采取最大池化(1-max-pooling)，操作，获取每个向量的最大值。若卷积层输出是 3×128 个一维向量，经过最大池化后，得到 3×128 个数值，再将这些数值进行拼接，得到一个 $[384, 1]$ 的向量，即池化层最终的特征向量。

全连接层：相当于分类器。通过Softmax函数，将模型学习到的各个类别标签的分值映射到 $0 \sim 1$ 之间，分值最大的标签就是预测的文本类别。为了防止过拟合，这里加上了dropout。

三、实验与分析

(一) 数据情况及评价指标

本研究采集某三甲医院的700份内镜报告，涵盖了5种不同的内镜技术，即文本类别。样本长度在10~40之间。样本类别的分布如表1所示。将其中450份作为训练集，150份作为验证集，剩余100份作为测试集。

表1 数据类别情况

内镜类别	数量
圈套器切除术	111
内镜下粘膜切除术	217
内镜下粘膜下剥脱术	231
分步内镜下黏膜切除术	75

表2 三个模型的超参数设置

模型	批样本量	学习率	卷积核大小	卷积核数量	序列长度	dropout
ERNIE	3	$1e-5$	-	-	32	-
TextCNN	3	$1e-5$	2, 3, 4	128×3	32	0.5
ERNIE-TextCNN	3	$1e-5$	2, 3, 4	128×3	32	0.5

氩离子凝固术	66
--------	----

为了评估ERNIE-TextCNN在内镜报告分类上的有效性，采用精确率(Precision)、召回率(Recall)及F1分数(F1 score)三个指标对模型性能进行评价。三个指标的计算公式如下：设数据集有 t 个类别($C_1, C_2, \dots, C_i, \dots, C_t$)，

$$P_i = \frac{\text{正确预测为 } C_i \text{ 的数量}}{\text{预测为 } C_i \text{ 的数量}} \quad (1)$$

$$R_i = \frac{\text{正确预测为 } C_i \text{ 的数量}}{\text{真实的 } C_i \text{ 的数量}} \quad (2)$$

$$F1 = \frac{1}{t} \sum_{i=1}^t \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} * 100\% \quad (3)$$

F1分数是精确率和召回率的调和平均数。3个指标值范围均在 $0 \sim 1$ 之间，结果越接近1，说明模型的性能越好。

(二) 实验环境及超参数设置

本次实验使用Windows10操作系统、Pycharm平台、Pytorch深度学习框架进行模型训练，主要开发语言为Python。

为了验证ERNIE-TextCNN模型的可行性，我们将其与单纯的ERNIE模型和TextCNN模型进行对比实验。两个对比模型的设置如下。

ERNIE模型，使用百度发布的预训练好的“ERNIE1.0 Base 中文”模型，接着配套全连接层，最后在输出层输出文本类别。

TextCNN模型，在文本词向量化后，直接进入TextCNN层训练，最后在输出层输出文本类别。

三种模型的超参数尽量保持一致，若是由于模型差异导致的，则采取最优的参数设置。实验中对对比模型的参数设置见表2。另外，TextCNN模型和ERNIE-TextCNN模型优化算法选择Adam优化器。

(三) 实验结果与分析

三个模型在测试集的结果如表3所示，可以看出，ERNIE-TextCNN模型的精确率、召回率、F1分数分别为85.62%、87.71%、86.68%，均高于其它两个模型。由此可见，ERNIE-TextCNN模型用于医学领域的文本分类是有效的。

下转第142页

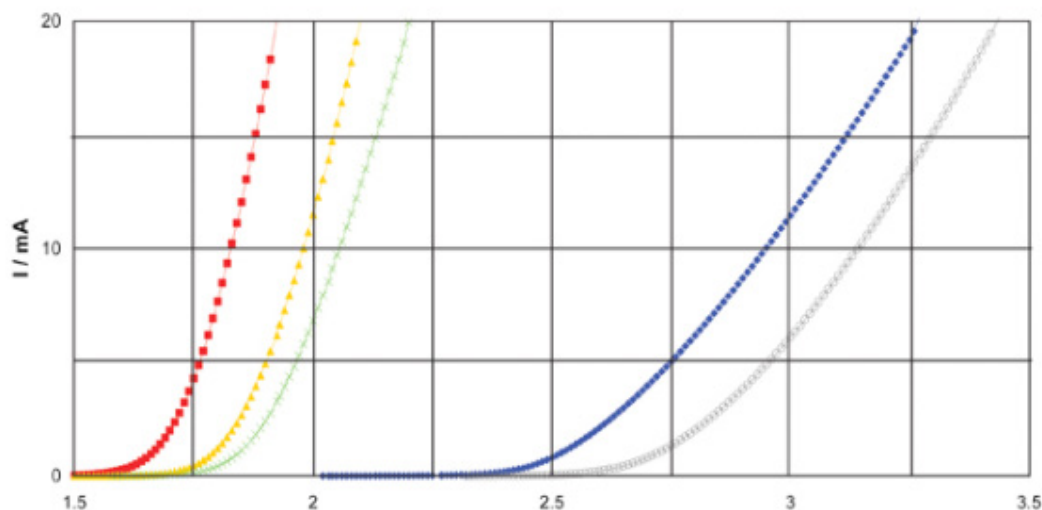


图2 LED灯电流电压控制原理图

经验，克服了传统气囊点爆高低温试验能力与行人保护高低温试验能力之间的差异，使两者之间相互协作，在降低工作成本的基础上，增加试验准确性和效果。在本项目建成之后，弥补了企业过去气囊点爆试验能力领域的空白和试验项目能力的缺失，为企业今后的稳定发展稳固了夯实的地基，成为国内少数掌握此能力的国家级检测机构。

5. 结束语

总之，本文对汽车安全气囊系统概况做了简要分析，进一步介绍了独立点爆区域设计方法、实验室气囊电阻准确性、设备触发与气囊展开零时刻点

的对应措施。针对试验中存在的问题给出了优化和应对措施，保证试验目的的顺利达成。

参考文献：

- [1] 东莞市升微机电设备科技有限公司, 常州检验检测标准认证研究院. 一种适用于汽车安全气囊的点爆试验装置: CN202021056948.6[P]. 2021-02-23.
- [2] 魏玉钊, 曹军帅, 任传波, 孙志钊, 吴凯伟, 颜廷浩. 发动机罩气囊在人车碰撞事故中对行人的保护作用[J]. 山东理工大学学报(自然科学版), 2021, 35(01): 52-55, 62.
- [3] 蔚来汽车有限公司. 气囊点爆测试装置和方法: CN201810871007.9[P]. 2020-02-14.

上接第139页

表3 三种模型的分类结果

模型	指标		
	精确率	召回率	F1分数
ERNIE	84.92%	86.27%	85.62%
TextCNN	81.03%	82.77%	81.92%
ERNIE-TextCNN	85.62%	87.71%	86.68%

四、结论

本研究构建的ERNIE-TextCNN模型，融合了ERNIE和TextCNN自身的优点，并应用到医学文本分类，相比单纯使用ERNIE模型和TextCNN模型，性能都有所提升，证明了该模型对中文的医学文本分类具有良好的效果。准确的分类不仅能为应用系统增加提示功能，而且可以提高医学领域专业术语的标准化程度，从而提升临床的工作效率，甚至为将来的医学研究带来诸多便利。但是，本研究还存在一些不足之处，未来模型的重点是融

合各类医学文本的特征，提高模型在医疗领域的泛化能力。

参考文献：

- [1] 杨飞洪, 王序文, 李姣. 基于BERT-TextCNN模型的临床试验筛选短文本分类方法[J]. 中华医学图书情报杂志, 2021, 30(1): 54-59.
- [2] Yu Sun, Shuohuan Wang, Yukun Li, et al. ERNIE: Enhanced Representation through Knowledge Integration [EB/OL]. (2019-04-19)[2020-08-23]. <https://arxiv.org/pdf/1904.09223.pdf>.
- [3] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. Proceedings of the 2014 Conference Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, October 24-29, 2011: 1746-1751.
- [4] Zhang Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification[J]. Computer Science, 2015.