文章编号：2095-3747（2023）-03-0018-03

基于层级特征和 DPCNN 的文本数据治理方法

丁行硕，　鞠通

（青岛远洋船员职业学院数字信息中心，山东 青岛 266427）

摘　要：大规模文本的数据划分是数据治理中的关键问题，而传统的中文文档建模方法容易忽视上下文语义关系和文档层级结构。针对以上问题提出一种基于层级特征和 DPCNN 的文本数据治理方法。该方法首先通过 BERT 模型抽取文本的层次特征信息，然后将结合全文信息的向量传入 DPCNN 模型中；经过金字塔型池化层后，最终通过全连接层进行数据划分。该方法能够有效提高特征稀疏文本数据的预测准确率。

关键词：数据治理；层级特征；BERT；DPCNN

中图分类号：TP391　　文献标识码：A

近年来，随着大数据和人工智能的快速发展，数据量呈现出爆发式增长，由此造成了数据冗余的现象[1]。目前比较流行的数据治理在高等院校、互联网公司和大型企业等有着广泛的关注。现阶段如何对冗余的数据进行有效治理，进而打造更大的平台为企业和个人服务是亟待解决的问题。自然语言处理是人工智能领域重要的研究方向。它因能有效理解和处理文本数据而广泛应用于舆情监测、信息提取和数据治理等方面。有效发现文本特征进而提高数据治理准确率非常重要。

1 文本数据治理方法研究概述

目前，有研究尝试使用机器学习的方法进行大规模文本的划分。例如苏金树等人[2]总结了基于机器学习的文本分类技术研究现状，从模型、算法和评测等方面对该研究进展进行了综述评价，讨论了SVM（支持向量机）、决策树和KNN（K近邻）等算法。这些算法因其模型简单，在文本划分中效率较高，但忽略了影响分类效果的词序和语义信息。

为进一步提高划分准确率，语言模型在文本特征抽取中得到广泛应用，神经网络依据出色的自适应和实时学习特点成为文本分类的常用方法。在语言模型中，Word2Vec 的提出证明了向量空间中单词表示的有效性；随之提出的 GloVe 和 ELMo 取得了很大成功，解决了 Word2Vec 只考虑词的局部特征问题。2018 年 Google 发布了基于双向 Transformer 的大规模预训练语言模型，在文本处理等任务中取得了很好的效果。在神经网络中，Kim Y[3]对 TextCNN（卷积神经网络）网络进行了系统的阐述，提出用 TextCNN 对文本进行标签提取。李洋等人[4]为进一步发现文本的长距离依赖关系，将双向 RNN（循环神经网络）与池化层融合形成 RCNN 模型，既使用具有提取局部特征优点的 CNN，又使用 BiLSTM（双向长短期记忆网络）发现上下文信息。除此之外，许多研究者还通过构建大规模语料库进行特征扩展，提高预测精度。

2 基于层级特征和 DPCNN 的文本数据治理方法

针对数据治理中大规模文本的数据划分容易

收稿日期：2023—03—06

第一作者简介：丁行硕，男，硕士，助理工程师

基金项目：青岛远洋船员职业学院科研项目：“高职院校航海类专业技能型人才工匠精神培育研究”（2023-R-004）；山东省职业教育教学改革研究项目：“AI 时代职业学校技能型人才工匠精神培育研究与实践”（2022286）

忽视上下文语义关系和文档层级结构的问题，本文提出了一种基于层级特征和 DPCNN 的文本数据治理方法。首先通过 BERT 模型抽取文本的层次特征信息，然后将多特征向量传入 DPCNN 网络，经过金字塔型池化层后得到数据划分结果。该过程如图 1 所示。

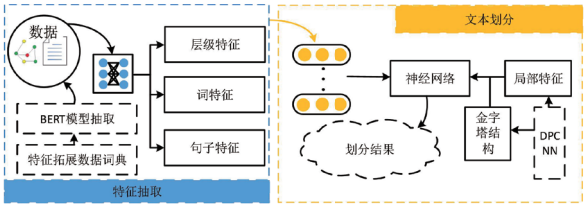

图 1 本文模型

2.1 BERT 模型

本文使用的 BERT 模型拥有 12 个 Transformer 模块层，768 个隐层，一共有 1.1 亿个参数，模型结构如图 2 所示。BERT 模型是基于双向 Transformer 编码器构建的语言模型，能够发现词语间的相互关系。Transformer 是由 6 个编码器和 6 个解码器堆叠而成。它接收向量序列，并输出处理后的数据序列，在经过 6 个编码器处理后分别传入 6 个解码器进行解码。而编码器和解码器的核心是注意力机制。它能够根据句子中的关键点去理解句子的整体含义。通过计算 Key 的注意力分布并整合到 Value 上，从而计算注意力价值。其原理如公式 1 所示。

$$Attention(Q,K,V)=softmax\left(\frac{Q\cdot K^{T}}{\sqrt{d_k}}\right) \quad (1)$$

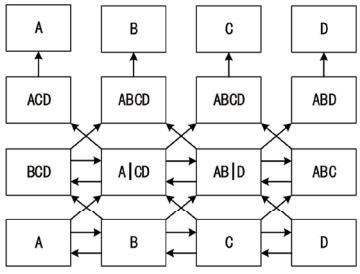

图 2 BERT 模型结构图

其中 Q,K,V 为输入的字向量矩阵，d_k 为输入向量的维度。在此计算过程中，Transformer 编码器将句子中任意两个单词通过一步计算直接联系起来，并将所有单词的表示进行加权求和，以此来缩短远距离依赖之间的距离，提高特征的有效利用率。

2.2 DPCNN 网络

CNN 是卷积层与池化层交错出现的前馈网络，能够有效利用词序，但 CNN 的并行处理友好性使其在大型训练数据中增加了计算复杂度。DPCNN（深层金字塔卷积神经网络）是一种简单的网络结构，它可以在增加网络层数的同时不会大幅度增加计算成本，而且能够获得最佳的准确率。DPCNN 模型结构如图 3 所示。

DPCNN 的结构使得每层的计算时间呈“金字塔形”下降。将离散文本转换为连续表示后，DPCNN 的结构仅仅是交替使用卷积层和下采样层，使得深度的神经网络降低了模型内部数据量以及每层神经网络的计算量，并且呈金字塔形减少。随着网络的不断深入，“金字塔”可以有效地发现文本中的长距离关联以及更多的全局信息，使得 DPCNN 可以比只能使用短距离关联的浅层卷积神经网络获得更好的准确率。每层计算量的减少过程如图 4 所示。

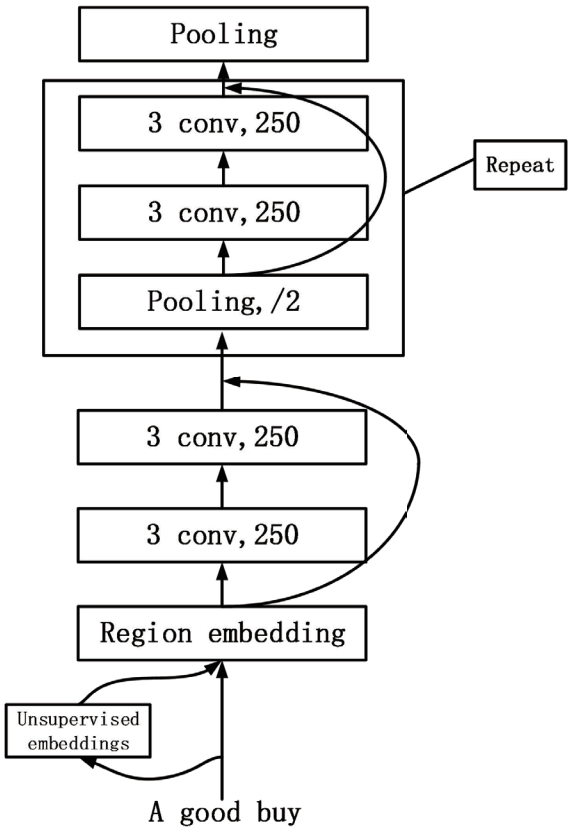

图 3 DPCNN 网络结构

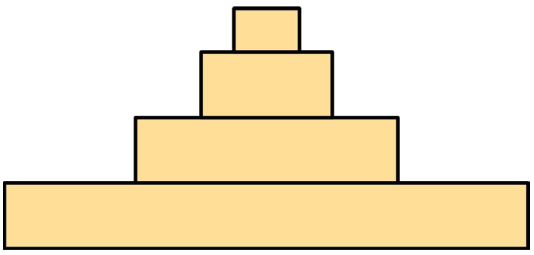

图 4 DPCNN 每层计算量变化

3 仿真实验及结果分析

本节研究本文方法的性能。首先，描述真实的数据集；然后描述评估指标；最后，对这个实验数据集进行实验结果分析。

3.1 实验数据

本文选取文本 Ente-pku 数据集进行实验。Ente-pku 数据集为定制的企业数据。首先利用网络爬虫爬取企业网站的文本信息，然后对爬取到的企业文本进行数据清洗，最后经过统计过滤生成 Ente-pku 数据集。Ente-pku 数据集包含 24 万余篇结构化数据。该数据为拥有企业经营范围和经营偏好的标签数据，共分为 9 个行业类别。将含有标签的数据集以 6：2：2 的比例分为训练集、验证集和测试集。使用本文提出的方法与传统模型进行比较，验证模型的优越性。

3.2 评价指标

为了验证模型有效性，本文使用精确率（Precision）、召回率（recall）和 F1-score 来评价模型效果。精确率是预测正确的数据量与被预测为正确的数据量之比（公式 2）。召回率是预测正确的数据量与实际为正的数据量之比（公式 3）；F1-score 是精确率和召回率的调和平均数（公式 4）。

$$Precision = \frac{Tp}{T_p+F_p} \quad (2)$$

$$Recall = \frac{T_p}{T_p+F_n} \quad (3)$$

$$F1-score = \frac{2Precision \times Recall}{Precision+Recall} \quad (4)$$

其中 Tp 为实际值和预测值均为正时的数据数量。Fp 为实际值为负、预测值为正时的数据数量。Fn 为实际值为正、预测值为负时的数据数量。

3.3 实验结果与分析

表 1 对比实验结果

算法	精确率 /%	召回率 /%	F1-score
KNN	80.22	80.22	79.83
SVM	86.17	86.23	86.12
Transformer	86.85	87.15	86.88
TextCNN	87.53	87.53	87.37
BiLSTM	89.20	89.32	89.24
RCNN	88.74	88.99	88.84
本文模型	90.87	90.87	90.83

选择 7 个算法进行对比实验，其中包含机器学习算法和神经网络算法。机器学习包括 KNN 和 SVM，神经网络算法包括 Transform、TextCNN、BiLSTM、RCNN 和本文模型。实现结果如表 1 所示。通过 7 个算法结果比较，充分说明改进后的模型要优于其它模型。

由表 1 可以看出，神经网络算法在评价指标上要明显高于机器学习算法，其中思想简单的 KNN 方法精确率最低，而使用双向网络的 BiLSTM 在对比方法中表现最好。单向的 Transformer 在实验中的精确率较低，因为其无法发现文本的层次特征，导致预测精度不佳。而拥有双向 Transform 的 BERT 模型能够有效发现文本的词特征、文本特征和层次特征，提高特征稀疏文本的预测精度。本文提出的模型在三个评价指标上都优于其它机器学习模型和神经网络模型，F1-score 值达到 90.83%。相较于其它模型分别在精确率、召回率、F1-score 上提高了 1.67%、1.55%、1.59% 以上，充分说明了本文模型的优越性。

4 结论与展望

本文针对文本数据治理的划分问题展开研究，提出了一种基于层级特征和 DPCNN 的文本数据治理方法。该方法首先通过 BERT 模型抽取文本的层次特征信息，然后将结合全文信息的向量传入 DPCNN 模型中，经过金字塔型池化层后，最后通过全连接层进行数据划分。根据真实数据集上的实验结果，该方法能够有效发现文本的层次特征，进而提高预测精度。

未来本模型研究重点将集中在以下两个方面：（1）进一步优化算法，使用蒸馏的方法提高 BERT 模型抽取效率。（2）在数据治理方面引入知识图谱技术对数据进行可视化表达。

参考文献：

[1] 丁行硕，李翔，谢乾．基于标签分层延深建模的企业画像构建方法[J]. 计算机应用，2022,42(04):1170-1177.

[2] 苏金树，张博锋，徐昕．基于机器学习的文本分类技术研究进展[J]. 软件学报，2006,(09):1848-1859.

[3]Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[4] 李洋，董红斌．基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用，2018,38(11):3075-3080.

（下转第 53 页）

Exploration and Practice of Curriculum Ideology and Politics in Marine Engineering in Higher Vocational Colleges —Take the Course "Ship Main Propulsion Power Plant" as an Example

SUN Hua-dong, TONG Yong-chen, FANG Feng, LI Yan

(Vocational Education Division, Qingdao Ocean Shipping Mariners College, Qingdao 266427, China)

Abstract: Summarizes the necessity of Ideology and Politics construction in Marine professional course in higher vocational college, analyzes the main problems in Ideology and Politics construction in marine engineering. With "ship main propulsion power device" course, for example, from the promotion team teachers' ideological cognition and course education teaching ability, clear course ideological goals, deep mining the ideological resources in the course, Optimize the course assessment, multi-mode integration teaching method helps the Ideology and Politics are discussed in the construction path of Curriculum Ideology and Politics

Keywords: Curriculum Ideology and Politics, higher vocational college, marine engineering, ship main propulsion power plant

(上接第 20 页)

Text Data Governance Method based on Hierarchical Feature and DPCNN

DING Xing—shuo, JU Tong

(The Center of Data&Information, Qingdao Ocean Shipping Mariners College, Qingdao266427, China)

Abstract: The data division of large-scale text is a key problem in data governance, but the traditional Chinese document modeling method is easy to ignore the contextual semantic relationship and the hierarchical structure of the document. To solve the above problems, a text data governance method based on hierarchical characteristics and DPCNN is proposed. Firstly, the hierarchical feature information of text is extracted by BERT model. Then the vector combined with the full text information is passed into DPCNN model, after passing through the pyramid pooling layer; Finally, the data is divided through the full connection layer. This method can effectively improve the prediction accuracy of sparse feature text data.

Keywords: data governance; hierarchical characteristics; BERT; DPCNN.