计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):243−249 [doi: 10.15888/j.cnki.csa.007557]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# BiLSTM_DPCNN 模型在电力客服工单数据分类中的应用①

李　灿, 田秀霞, 赵　波

(上海电力大学 计算机科学与技术学院, 上海 200090)
通讯作者: 李　灿, E-mail: lcshmily@mail.shiep.edu.cn

**摘　要**: 电力客服工单数据以文本形式记录电力用户的需求信息, 合理的工单分类方法有利于准确定位用户需求, 提升电力系统的运行效率. 针对工单数据特征稀疏、依赖性强等问题, 本文对基于字符级嵌入的长短时记忆网络 (Bidirectional Long Short-Term Memory network, BiLSTM) 和卷积神经网络 (Convolution Neural Network, CNN) 组合的结构模型进行优化. 该模型首先对 Word2Vec 模型训练的词向量进行降噪处理, 得到文本的特征表示; 其次, 利用 BiLSTM 网络递归地学习文本的时序信息, 提取句子特征信息; 再输入到双通道池化的 CNN 网络中, 进行局部的特征提取. 通过在真实客服工单数据集上的测试实验, 验证了该模型在客服工单分类任务上的具有较好的精确性和鲁棒性.

**关键词**: 电力客服工单; 文本分类; BiLSTM; CNN; Word2Vec

引用格式: 李灿,田秀霞,赵波.BiLSTM_DPCNN 模型在电力客服工单数据分类中的应用.计算机系统应用,2021,30(2):243–249. http://www.c-s-a.org.cn/1003-3254/7557.html

## Application of BiLSTM_DPCNN Model in Work Order Data Classification for Power Customer Service

LI Can, TIAN Xiu-Xia, ZHAO Bo

(College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)

**Abstract**: The power customer service order data records the demand of power users in text. A reasonable work order classification method is helpful to accurately identify the demand of users and improve the operating efficiency of the power system. To solve the problems of sparse feature data and strong dependency of work order data, this study optimizes the structural model that combines character-level embedded Bidirectional Long-Short-Term Memory network (BiLSTM) and Convolution Neural Network (CNN). Firstly, this model obtains the feature representation of text by noise reduction on the term vectors trained by the Word2Vec model. Secondly, it uses the BiLSTM network to recursively learn the time sequence information of the text to extract the feature information of sentences. Finally, those obtained are input into the double-channel pooled CNN for the extraction of local features. The test experiments on the real work order data set of power customer service demonstrate that the model has good accuracy and robustness in the task of classifying work orders of power customer service.

**Key words**: work order of power customer service; text categorization; BiLSTM; CNN; Word2Vec

① 基金项目: 国家自然科学基金面上项目 (61772327); 国家自然科学基金重点项目 (61532021)
Foundation item: General Program of National Natural Science Foundation of China (61772327); Key Program of National Natural Science Foundation of China (61532021)
收稿时间: 2020-01-07; 修改时间: 2020-01-22; 采用时间: 2020-03-11; csa 在线出版时间: 2021-01-27

电力客服系统作为供电企业与电力客户交流的窗口, 不仅为电力客户提供了便捷的服务, 还直接客观地反映客户用电诉求[1,2], 其工单数据记录着电力客户对供电企业的诉求信息, 根据工单信息描述, 准确地定位用户所属类别, 有利于提升客户满意度. 目前对工单数据的分析方式, 主要是由调查人员通过对用户诉求数据的分析, 来判别用户需求信息所属的服务类型[3,4]. 这种方式缺乏有效的分析方法, 严重影响信息分析和解决问题的效率, 直接影响到电力系统的高效运行和发展[5]. 因此, 找到一种高效的工单分类方法来实现对工单数据进行自动、准确的分类, 是电力客服系统亟待解决的主要问题[1].

传统的工单分类, 主要是采用特征工程和机器学习方法相结合的方式[6]. 林溪桥等[7] 分析各种类型工单的出现规律, 结合主成分分析方法, 实现客服工单分类模型的优化. Sun 等[8] 利用中文数据挖掘的方法, 对停电工单进行分析, 并结合支持向量机 (Support Vector Machine, SVM) 构建了故障案例句子分类模型. 上述分类模型虽然有着结构简单, 训练速度快的优点, 但在分类过程中, 依赖于特征工程的选择, 模型分类效果表现不突出.

近年来, 神经网络技术被广泛应用到文本分类任务中[9]. 谢季川等[10] 使用 Word2Vec 语言模型训练电力工单数据, 得到电力文本词向量, 最后构建多分类文本模型, 实现 95 598 电力工单分类任务. 刘梓权等[11] 通过分析电力设备缺陷记录, 构建了一种基于 CNN 的电力缺陷文本分类模型. 受以上启发, 我们将 CNN 网络用于工单数据的分类中. 然而, 分类的效果并不理想, 这是因为工单数据存在依赖性强、冗余度高的特征, 传统的 CNN 网络在处理这些数据时遇到以下问题:

(1) 文本表示: 由于数据的高依赖性, 在特征文本表示时, 要考虑前后词语间的语义关系;

(2) 特征提取: 在句子高级建模阶段, 使用 CNN 网络只能捕获局部的语义信息, 造成隐层语义信息丢失.

为了解决上述问题, 本文将 BiLSTM 与 CNN 混合的神经网络应用到工单分类的任务中, 该模型充分利用了 BiLSTM 递归序列模型学习句子中的全局语义信息的特点; CNN 结构可以通过卷积运算, 挖掘句子局部语义特征的优势. 本文从以下 3 个方向入手, 并进行创新:

(1) 词向量标准化: 对词向量作标准化处理, 去除噪音, 加快网络训练收敛速度, 提高分类精度;

(2) 稀疏特征提取: 由于数据稀疏性, 包含大量的边际信息, 为了提取边际有效信息, 利用 BiLSTM 代替 RNN;

(3) 特征融合: 本文在卷积网络上进行改进, 克服了 Max-Pooling 丢失特征信息的问题, 保留了强特征词信息, 捕获全局深层的隐层语义信息.

## 1 相关工作

文本分类是将待分类文本数据合理地划分到相应的类别中, 是有监督学习的过程. 工单文本也是文本分类中的经典问题, 通过分析文本特征, 本文所采用的文本分类算法是基于 BiLSTM 神经网络和 CNN 网络这两种模型. 以下分别对实验中所涉及的相关技术进行介绍.

### 1.1 向量模型

2013 年, Hinton 提出了 Word Embedding 概念, Word Embedding 方法将单词映射到向量空间, 不仅可以避免“维度灾害”问题, 还能从更深层学习词与词之间的语义信息[12]. 与此同时, Mikolov 等[13,14] 提出了 Word2Vec 框架, 该神经网络语言在语言训练时, 考虑上下词语义间的相关性. 基于 Word2Vec 模型解决了传统的文本表示中数据稀疏和语义鸿沟的问题[15,16]. 本文实验部分, 采用上述神经网络语言模型得到文本的向量表示.

### 1.2 BiLSTM 神经网络

循环神经网络 LSTM 网络是 RNN 网络的扩展, 它解决了传统 RNN 网络中梯度消失或爆炸问题[17]. 使用 LSTM 网络可以学习到当前文本过去的信息, 但无法编码从后到前的信息, 因此出现了 BiLSTM 神经网络可以更好的捕捉双向的语义依赖[18,19]. 网络的结构通常包括 3 部分: 输入层、隐藏层和输出层. BiLSTM 单元示意图, 如图 1 所示.

$$i_t = \sigma(W_i \cdot [h_{t-1}, w_t] + b_i) \quad (1)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, w_t] + b_f) \quad (2)$$

$$q_t = \tanh(W_q \cdot [h_{t-1}, w_t] + b_q) \quad (3)$$

$$o^t = \sigma(W_o \cdot [h_{t-1}, w_t] + b_o) \quad (4)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes q_t \quad (5)$$

$$h^t = o_t \otimes \tanh(c_t) \quad (6)$$

上述计算公式为经过 LSTM 三个门信息保留计算表达式, 其中, $W_j$ 是权重矩阵, $b_j$ 是偏差向量, $j \in \{i, f, q, o\}$ $\cdot \otimes$ 表示逐点相乘, $\sigma$ 是激活函数. $f_t$ 决定哪些信息需要从单元状态中丢弃, $i_t$ 决定哪些值需要更新, $o_t$ 决定模型的最终输出.

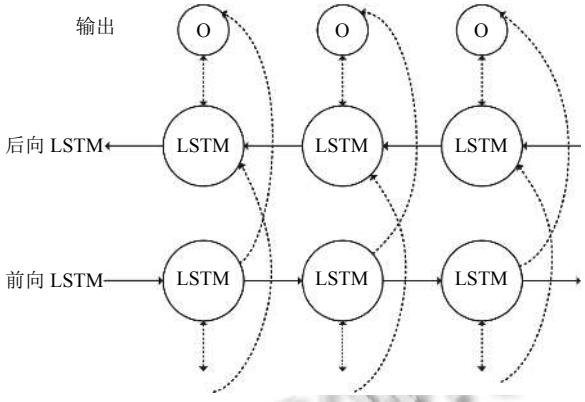


图 1 BiLSTM 单元示意图

本文实验中使用 BiLSTM 网络是由两个方向的 LSTM 组成, 使用前向和后向 LSTM 网络可以学习句子的上下文信息, 得到全局的句子语义信息.

### 1.3 卷积神经网络

TextCNN 是一种前馈神经网络, 最初应用于计算机视觉, 在图像处理中有突出的表现[20]. 随着深度学习的发展, CNN 被广泛应用自然语言处理的任务中, 如文本分类、情感分析等[21,22].

TextCNN 网络结构主要包括卷积和池化层, 实现过程如下: 首先, 将卷积层输出的特征向量分别输入最大池化层; 其次, 将池化的输出结果拼接表示最终的句子特征向量. 为了提取不同位置的局部信息, 实验中同时使用多个通道和不同卷积核大小进行卷积操作. 其结构图 2 所示.

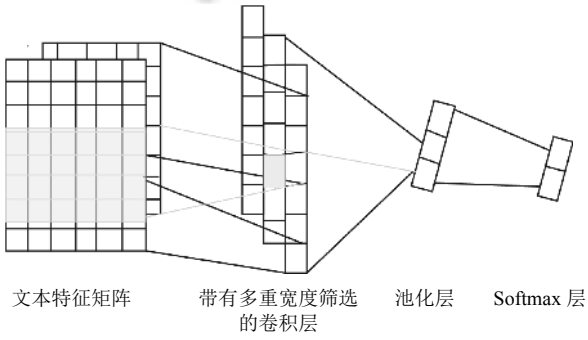


图 2 TextCNN 网络结构示意图

设给定矩阵 $W \in R^{M \times N}$, 卷积核 $F \in R^{M \times N}$, 且 $m \ll M$, $n \ll N$, 则卷积的表达式如 (7) 所示:

$$B = \sum_{u=1}^{m} \sum_{v=1}^{n} f_{uv} \cdot w_{i-u+1:j-v+1} \tag{7}$$

$B$ 表示卷积后的结果, 卷积后得到由句子局部信息构成的特征矩阵, 输入到 Max-Pooling 层提取整个矩阵中最大值, 作为当前通道的特征信息, 再与其他通道的特征信息融合, 得到多通道组合筛选后的句子特征向量.

## 2 模型架构与算法设计

考虑到传统 CNN 网络在工单分类中不足, 而 BiLSTM 神经网络在文本分类中可以很好地提取上下文信息, 在高级语义建模阶段, CNN 网络可以凭借多通道组合筛选, 对句子进行二次特征提取, 从而得到句子特征向量. 鉴于以上两种模型的优点, 本文在 BiLSTM+CNN 组合网络的基础上进行优化. 即在 TextCNN 网络层中采用双池化操作, 我们称之为双池化的卷积神经网络 (Double Pooling Convolution Neural Network, DPCNN), 并将优化后的组合模型应用到电力客服工单数据的分类中, 其模型结构如图 3 所示. 整个分类过程从左向右, 包括文本向量化表示和训练 BiLSTM_DPCNN 分类器两部分, 以下按模型的搭建过程逐一论述.

### 2.1 向量的优化表示

在以往的模型训练过程中, 都是直接将训练出的文本词向量输入到模型中, 这种方法导致模型难收敛, 分类准确率低. 究其原因, 主要是因为短文本数据的高稀疏性, 不同词的出现频率不同, 对于一些高特征词出现频率较高, 同时也存在边缘流特征信息. 这将使得到词向量的权值存在较大的偏差, 分布不均衡.

为了解决上述问题, 本文实验中采用 Word2Vec 模型训练以字为单位的词向量的前提下, 并对词向量进行标准化处理, 达到降噪效果. 首先, 遍历词汇表, 统计每个单词出现的次数, 计算其频次; 其次, 计算所有词的权重误差; 最后, 采用标准化处理, 得到均值表示的词向量. 再把处理后的词向量作为训练模型的输入, 具体计算公式如下:

$$E(v) = \sum_{j=1}^{k} f_j v_j \tag{8}$$

$$Var(v)=\sum_{j=1}^{k} f_j(v_j-E(v))^2 \tag{9}$$

$$\overline{W_k}=\frac{W_k-E(v)}{\sqrt{Var(v)}} \tag{10}$$

其中, $f_j$ 表示单词出现的次数, $k$ 表示词的个数, $v_j$ 表示对应词的权重. 经标准化后的词向量, 其范围固定在一个固定值之间, 弱化某些词向量的权重值过大对模型的影响. 加快网络的收敛速度, 提高模型的分类精度.

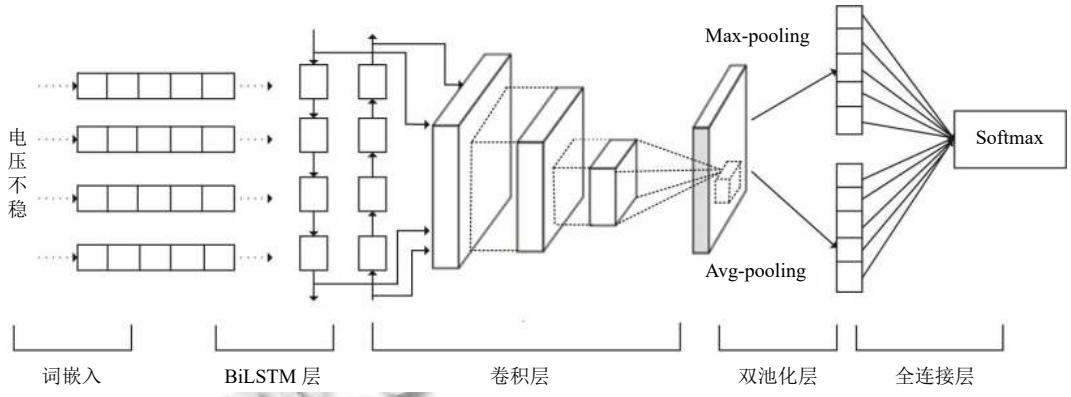


图 3 基于 BiLSTM_DPCNN 混合神经网络模型

### 2.2 BiLSTM_DPCNN 分类模型

实验中所采用的 BiLSTM_DPCNN 组合模型的思想, 充分利用了各个模块的优点. 首先, 待分类文本经过数据预处理后通过 Embedding Layer 层把单词表示成模型可以识别的文本向量. 再利用 BiLSTM 网络提取句子特征语义信息, 由于文本在处理时, 调用 Keras 库提供的 pad_sequence 方法将文本 pad 为固定长度, 因此在 BiLSTM 输出时乘以 MASK 矩阵来减小 pad 带来的影响. 最后, 将 BiLSTM 的输出作为改进 CNN 网络的输入, 进行二次特征提取, 最后实现分类.

BiLSTM_DPCNN 模型由输入层、BiLSTM 层、卷积层、双池化层、分类输出层五部分组成. 以下详细介绍每一层的实现过程.

首先, 采用 BiLSTM 网络学习每个字的向量语义, BiLSTM 因其双向计算的特点, 可以获取目标序列的“左序列”和“右序列”的文本信息. 具体计算公式如下.

$$C_l(x_i)=g(W_lC_l(x_{i-1})+W_{sl}E(x_{i-1})) \tag{11}$$

$$C_r(x_i)=g(W_rC_r(x_{i+1})+W_{sr}E(x_{i+1})) \tag{12}$$

其中, $C_l(x_i)$表示左边的文本向量, $W_l$是一个从当前隐藏层到下一隐藏层转换的$W_{sl}$矩阵表示连接当词的左文本的语义信息. $C_l(x_{i-1})$表示前一个词的左边文本信息, $E(x_{i-1})$表示前一个词的字向量. 同理, 可计算$x_i$的右文本$C_r(x_i)$.

根据式 (11)、式 (12), 得到当前字$x_i$的向量表示. 计算公式如下:

$$x_i=[C_l(x_i);E(x_i);C_r(x_i)] \tag{13}$$

其次, 把 BiLSTM 网络层输出的句子特征矩阵输入到卷积层中进行卷积操作, 提取深层语义信息. 经卷积操作后, 使用双通道池化进行特征筛选, 进而提取句子特征矩阵中相邻文字的关联特征, 其中池化部分的具体计算公式如下:

$$c_{\max}=\max_{i=1}^{n}\{c_i\} \tag{14}$$

$$c_{\text{avg}}=\frac{1}{h}\sum_{i=1}^{h} c_i \tag{15}$$

$$z_i=c_{\max}\oplus c_{\text{avg}} \tag{16}$$

其中, $c_{\max}$ 表示最大池化的输出, $c_{\text{avg}}$ 表示平均池化的输出, $\oplus$表示拼接运算, $h$ 表示滑动窗口大小. 该方法弥补了由于每次最大池化操作只能取一次最大值, 从而丢失强特征词信息的缺点. 将两个池化操作的结果进行特征融合, 保证了文本特征信息的完整性, 得到的更全面的、深层次的句子特征.

最后, 由双池化层提取到句子的特征表示作为 Softmax 层的输入, 分类过程的具体计算公式所示:

$$p(y|Z,W_s,b_s)=\underset{y}{softmax}(W_s\cdot Z+b_s) \tag{17}$$

其中, $y\in\{0,1\}$, $W_s\in R^{|Z|}$, $b_s$ 代偏置项, $Z$ 表示句子特征矩阵.

### 2.3 模型算法流程

BiLSTM_DPCNN 组合模型的算法过程如算法 1.

算法 1. BiLSTM_DPCNN 模型训练过程

1) 初始化模型参数配置, 设置每批训练量 batch_size 和总迭代次数 epochs;
2) 将由标准化后的字向量表示的句子信息输入到 BiLSTM 网络层中, 提取全局句子语义信息;
3) 将第 2) 步的输出句子矩阵输入到 DPCNN 网络中, 通过卷积操作, 进一步捕获局部语义信息;
4) 将第 3) 步中卷积后的结果分别输入到双通道池化层进行降维操作, 并将特征融合;
5) 将第 4) 步融合后文本向量经过矩阵的 concat 和 reshape 之后送入 Softmax 分类器, 输出类别标签;
6) 模型训练过程中, 采用 mini-batch 的梯度下降法进行训练, 训练过程中保存最优的模型, 减少再训练过程中的开销;
7) 重复 2)~6) 步, 设置 epochs=50, 若训练集的精度不在上升, 则提前结束训练.

实验中 batch 的大小设置为 30, 能够获得较好的效果, 则所有参数$\theta$的计算公式如下所示.

$$\theta=\theta-\lambda\frac{\partial(\theta)}{\partial\theta} \tag{18}$$

其中, $\lambda$表示学习率.

## 3 实验设计及分析

### 3.1 实验环境

本实验基于 Python 编程语言和 Tensorflow 1.8.0 深度学习框架展开, 主要参数配置 CPU: Intel Core i9-9900 K; 内存: 32 GB; 操作系统: Windows 10.

### 3.2 数据来源和预处理

在本次实验中采用来自电网公司客服工单记录的真实数据, 数据主要记录客户对电力公司的用电反馈信息. 其中每条工单数据以短文本的形式记录着工单类别及相应的信息反馈, 共分为停电、安全隐患、停电未送电、电压不稳、缺相、供电故障、用户资产故障 7 个工单类别, 下述分别用 $C_1$~$C_7$ 命名. 所有数据都经过数据清洗和停用词过滤, 实验中将数据集划分为训练集、验证集和测试集, 其比例为 3:1:1.

### 3.3 模型超参数设置

模型中主要参数包括字向量的维度 $d$, 滤波器的个数 $m$ 等, 另外 BiLSTM 层中隐藏层层数为 2 层, 神经元个数 256 个. 模型最优参数设置如表 1 所示.

### 3.4 评估指标

由于工单数据是多分类问题, 工单类别分为 7 类, 采用查准率 ($P$), 召回率 ($R$), $F1$ 值 ($F1$) 和宏平均 ($Macro\_F1$) 等 4 个指标, 对分类的准确度进行评估, 其计算公式如下:

$$P=\frac{TP}{TP+FP} \tag{19}$$

$$Macro\_P=\frac{1}{V}\sum_{i=1}^{V}P \tag{20}$$

$$R=\frac{TP}{TP+FN} \tag{21}$$

$$Macro\_R=\frac{1}{V}\sum_{i=1}^{V}R \tag{22}$$

$$F1=\frac{2P*R}{P+R} \tag{23}$$

$$Macro\_F1=\frac{2*Macro\_P*Macro\_R}{Macro\_P+Macro\_R} \tag{24}$$

其中, $V$ 表示工单类别个数, $TP$ 表示正确的标记为正, $FP$ 错误的标记为正, $FN$ 错误的标记为负, $TN$ 正确的标记为负.

表 1 模型参数配置

| 参数 | 含义 | 最优值 |
|---|---|---|
| *batch_size* | 每批次使用的样本数 | 128 |
| *num_epochs* | 训练轮数 | 50 |
| *hidden_dim* | 隐藏层层数 | 128 |
| *dropout* | 丢失率 | 0.5 |
| $\lambda$ | 学习率 | 0.001 |
| $d$ | 字向量维度 | 300 |
| $m$ | 滤波器个数 | 128 |
| $h$ | 滑动窗口大小 | [2, 3, 4] |

### 3.5 对比实验设置

实验中设置以下几个分类模型来评估本文模型, 通过对比模型验证 BiLSTM_DPCNN 模型的分类性能:

(1) TextCNN 模型. CNN 模型采用文献 [23] 中的网络结构, 以字符粒度建模, 使用卷积提取文本特征图后, 输入到最大池化层, 提取矩阵实现对文本的分类.

(2) BiLSTM 模型. BiLSTM 模型是由两个不同方向的 LSTM 组合, 它可以同时捕获上下文语义信息, 解决了分类过程中长距离依赖问题.

(3) RCNN 模型. RCNN 模型采用文献 [24] 中的结构, 使用循环网络学习单词的上下文信息, 得到当前词的向量表示, 再通过一维卷积, 提取句子特征, 经池化后输入到 Softmax 分类器中获得分类结果, 可以简单看成 LSTM 和 CNN 的混合网络.

为了更好地评估模型, 以上对比模型输入的词向量均是随机生成, 不做标准化处理.

### 3.6 实验结果及分析

如表 2 所示, 给出各个模型在电力客服工单数据集上的测试结果, 其中 Dev-Macro_F 和 Test-Macro_F 分别表示在验证集和测试集上的宏平均分数.

表 2 实验中各个模型分类精度

| 模型 | Dev-Macro_F | Test-Macro_F |
|---|---|---|
| TextCNN | 0.950 | 0.958 |
| BiLSTM | 0.962 | 0.960 |
| RCNN | 0.946 | 0.950 |
| 本文模型 | 0.974 | 0.978 |

通过表 2 中的实验结果可看出,BiLSTM 网络在分类结果上与 TextCNN 网络相比而言, 其在测试级上的精度, 达到 96%, 明显高于 TextCNN 分类模型. 这是因为 BiLSTM 网络具有记忆单元, 选择性记忆和遗忘信息, 具有递归学习信息的优势, 此外, 又能捕获前后两个方向上的时序信息, 在电力短文本数据集中表现突出. 但是在一般的文本分类任务中, 较多采用 TextCNN 结构, 因为训练速度较快, 模型易收敛. RCNN 网络模型由于其结构较为简单, 其分类指标低于其他模型. 而本文提出的将 BiLSTM 和 DPCNN 网络组合的模型, 无论是在测试级和验证集, 模型的分类准确率都表现最优. 其在测试集上的 *Macro_F*1 值较 TextCNN 模型提高了 2.0%. 主要原因是因为在文本实验中, 不仅在网络模型上进行组合优化, 还在词向量上做进一步处理. 具体而言, 首先以字符级嵌入词向量, 从细粒度层次分析字与字之间的语义关系, 并对词向量进行归一化处理, 减少数据噪音; 其次, 将 BiLSTM_DPCNN 模型应用到客服工单数据的任务中. 该模型首先, 利用 BiLSTM 网络具有“门控”结构, 对句子特征信息选择性保留, 又可以同时捕获上下文信息, 在文本分类任务中具有较高的准确率; 其次, 考虑到对 TextCNN 网络而言, 不仅对相邻语义信息的捕获能力较强, 又能通过设置不同滑窗大小来提取不同位置的局部信息, 最后融合多层卷积特征, 得到较好的分类效果. 但模型存在一个最大的弊端, 即采用最大池化操作, 只能取一次最大值, 从而丢失强特征词信息. 因此, 本文实验中的在原有网络的基础上进行改进, 添加平均池化层, 把卷积层提取的句子特征, 分别输入到最大池化和平均池化层, 最终把池化输出的特征进行融合, 通过全连接的方式和 Softmax 分类器, 实现分类输出. 因此, 在分类性能上要高于单一网络结构.

由于本文以字为单位作为模型的输入, 所以在这里研究了字向量的特征维度对模型分类效果的影响. 图 4 表示字向量的特征维度从 50 维变换到 400 维的过程中, 宏平均值 (*Macro_F*1) 的变化情况. 由图 4 可知, 随着字向量维度的逐渐增加, 整体的分类性能不断上升, 但当维度为超过 300 时, 分类性能趋于下降状态, 原因是字符级别过大导致计算成本增大. 本次实验选 300 作为字向量的表示维度, 此时模型的分类效果最佳.

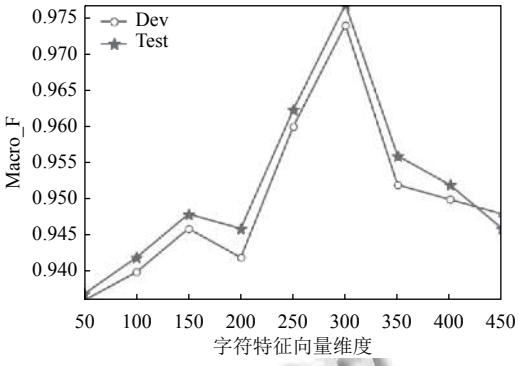


图 4 分类性能与字向量特征维度的关系

表 3 给出不同模型在各个分类中的 *P*、*R*、*F*1 值, 通过表 3 可以看出 BiLSTM 模型在 $C_1$、$C_3$、$C_6$ 类别上的 *F*1 值高于 TextCNN 模型. RCNN 模型分别在 $C_2$、$C_5$、$C_7$ 类别上的 *F*1 值高于 BiLSTM 模型, 尤其是在 $C_5$ 类别中 *F*1 值相比 BiLSTM 模型, 提高了 5%. 但从整体分类效果看, 本文模型表现优于其他模型, 其中最好的区分类别是 $C_1$ 和 $C_5$, *F*1 值达到 99% 和 97%, 在其他类别中的 *F*1 值也均有提升.

表 3 不同神经网络模型的实验对比

| 类别 | TextCNN模型 | | | BiLSTM模型 | | | RCNN模型 | | | 本文模型 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *P* | *R* | *F*1 | *P* | *R* | *F*1 | *P* | *R* | *F*1 | *P* | *R* | *F*1 |
| $C_1$ | 0.98 | 0.98 | 0.96 | 0.97 | 0.99 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 |
| $C_2$ | 0.79 | 0.89 | 0.84 | 0.82 | 0.79 | 0.80 | 0.81 | 0.84 | 0.83 | 0.84 | 0.90 | 0.87 |
| $C_3$ | 0.84 | 0.86 | 0.62 | 0.83 | 0.80 | 0.84 | 0.62 | 0.67 | 0.64 | 0.96 | 0.92 | 0.94 |
| $C_4$ | 0.93 | 0.94 | 0.94 | 0.94 | 0.93 | 0.93 | 0.94 | 0.93 | 0.93 | 0.95 | 0.95 | 0.95 |
| $C_5$ | 0.94 | 0.99 | 0.96 | 0.95 | 0.87 | 0.91 | 0.99 | 0.96 | 0.96 | 0.98 | 0.96 | 0.97 |
| $C_6$ | 0.96 | 0.90 | 0.93 | 0.97 | 0.95 | 0.96 | 0.88 | 0.92 | 0.90 | 0.94 | 0.97 | 0.95 |
| $C_7$ | 0.89 | 0.91 | 0.90 | 0.87 | 0.88 | 0.88 | 0.89 | 0.88 | 0.89 | 0.92 | 0.90 | 0.91 |

表 4 给出不同模型的训练和测试时间对比, 从表中数据可以看出, 就训练时间而言, TextCNN 模型的训练速度快比 BiLSTM 快很多, 主要是因为该模型适合并行计算; 本文实验中的 BiLSTM_DPCNN 混合模型, 时间复杂度要高些.

表 4 各模型时间复杂度对比 (单位: s)

| 模型 | 训练时间 | 测试时间 |
|---|---|---|
| TextCNN | 76 | 19 |
| BiLSTM | 125 | 14 |
| RCNN | 168 | 15 |
| 本文模型 | 196 | 16 |

## 4 结论与展望

本文通过分析电力客服工单数据特征, 基于字符级嵌入的 BiLSTM_DPCNN 分类算法对工单文本进行分类. 在模型训练过程中, 首先对词向量进行优化表示; 其次, 使用 BiLSTM 循环结构学习上下文信息, 获取全局语义信息; 最后, 采用卷积双池化方法提取全局最优的语义特征值. 通过与其他分类算法对比, 验证了该模型分类效果的优越性. 但在大量客服工单数据中仍存在可用样本较少, 不足以构成训练集类别的问题. 因此, 下一步需对模型进行完善使其同样适用于样本不均衡分布的客服工单数据, 这对促进电网智能化发展有着重要意义.

## 参考文献


1 邹云峰, 何维民, 赵洪莹, 等. 文本挖掘技术在电力工单数据分析中的应用. 现代电子技术, 2016, 39(17): 149–152.

2 顾斌, 彭涛, 车伟. 基于词典扩充的电力客服工单情感倾向性分析. 现代电子技术, 2017, 40(11): 163–166, 171.

3 陈俐冰, 何容, 邱林, 等. 电力客服中心用户行为分析研究与实现. 计算机技术与发展, 2017, 27(2): 116–119, 124.

4 丁麒, 庄志画, 刘东丹. 基于文本数据挖掘技术的 95598 业务工单主题分析应用. 电力需求侧管理, 2016, 18(S1): 55–57.

5 何薇, 张剑, 于雪霞, 等. 基于文本挖掘的电网客户服务满意度评价模型. 电子世界, 2017, (7): 81, 83.

6 Prusa JD, Khoshgoftaar TM. Designing a better data representation for deep neural networks and text classification. 2016 IEEE 17th International Conference on information Reuse and Integration (IRI). Pittsburgh, PA, USA. 2016. 411–416.

7 林溪桥, 严旭, 黄蔚. 基于主成分分析法的 95598 客户服务工单分类优化. 广西电力, 2017, 40(4): 10–12, 30. [doi: 10.3969/j.issn.1671-8380.2017.04.003]

8 Sun HF, Wang ZY, Wang JH, *et al*. Data-driven power outage detection by social sensors. IEEE Transactions on Smart Grid, 2016, 7(5): 2516–2524. [doi: 10.1109/TSG.2016.2546181]

9 Song J, Qin SJ, Zhang PZ. Chinese text categorization based on deep belief networks. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). Okayama, Japan. 2016. 1–5.

10 谢季川, 宗振国, 刘宏国, 等. 基于词向量模型的 95598 工单文本挖掘. 电子世界, 2017, (23): 176, 178.

11 刘梓权, 王慧芳, 曹靖, 等. 基于卷积神经网络的电力设备缺陷文本分类模型研究. 电网技术, 2018, 42(2): 644–650.

12 Hinton GE. Learning distributed representations of concepts. Proceedings of the eighth annual Conference of the Cognitive Science Society. Amherst, MA, USA. 1986. 12.

13 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2013. 3111–3119.

14 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013.

15 牛雪莹, 赵恩莹. 基于 Word2Vec 的微博文本分类研究. 计算机系统应用, 2019, 28(8): 256–261. [doi: 10.15888/j.cnki.csa.007030]

16 汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究. 计算机系统应用, 2018, 27(5): 209–215. [doi: 10.15888/j.cnki.csa.006325]

17 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

18 Jiang W, Jin Z. Integrating bidirectional LSTM with inception for text classification. 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). Nanjing, China. 2017. 870–875.

19 Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada. 2013. 6645–6649.

20 Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv: 1408.5882, 2014.

21 黑富郁, 王景中, 赵林浩. 基于 CNN 和 LSTM 的异构数据舆情分类方法. 计算机系统应用, 2019, 28(6): 141–147. [doi: 10.15888/j.cnki.csa.006900]

22 陈巧红, 王磊, 孙麒, 等. 卷积神经网络的短文本分类方法. 计算机系统应用, 2019, 28(5): 137–142. [doi: 10.15888/j.cnki.csa.006887]

23 Zhang X, Zhao JB, LeCun Y. Character-level convolutional networks for text classification. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2015. 649–657.

24 Lai SW, Xu LH, Liu K, *et al*. Recurrent convolutional neural networks for text classification. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, TX, USA. 2015. 2267–2273.