

[神经网络]

基于LSTM循环神经网络的电力故障挖掘研究

文/国网浙江义乌市供电有限公司 周 洋
国网浙江省电力有限公司 陈丹升
浙江华云信息科技有限公司 刘晓枫 汪娟玉

随着电力信息化程度不断加深，电力单位积累了包含海量信息的文本数据。电力故障抢修记录作为重要文本资源，对其进行深入挖掘，有利于合理制定电力检修计划、排查电力故障。研究表明，引入外部知识的词向量使深度学习模型拟合得更快，分类准确率也大幅提升。

电力行业的信息化过程中，产生了大量包含重要信息的文本数据。电力单位故障抢修的过程中，积累了丰富的故障记录。这些故障记录涵盖了用户报修、下派、故障问题描述、故障原因描述及现场检修记录等内容，是实现运检工作自动化和智能化的重要数据支撑。然而，文本数据的非结构化使得故障文本的分析和挖掘具有相当的复杂性，造成故障数据的利用率相对较低。因此，当前针对电力故障记录文本的研究工作是十分有必要的。

近年来，人工智能技术蓬勃发展，中文文本挖掘技术也有了较大的进步。深度学习与传统自然语言处理技术的结合，给中文分词、分布式表示和关键词抽取等文本挖掘技术注入了新的活力。同时给中文文本挖掘技术在各个专业领域的广泛使用带来机遇，目前文本挖掘技术已经在医疗、互联网和物流等多个领域快速发展。但是，在电力行业中，单系统的故障文本还达不到这个数量级，训练后输出向量的表达能力也达不到理想效果。

本文针对电力行业故障文本数据规模较小的问题，引入外部语料库，将故障文本数据和外部

语料库融合后进行词向量的训练。使用该词向量对故障文本的故障成因进行分类，将故障成因分为5个主要大类，并使用长短期记忆网络（Long Short-Term Memory，LSTM）作为分类模型。在试验中，从准确率等方面对分类模型进行了测试，表明引入外部知识的词向量带获得了更好的分类效果。

电力故障文本

电力系统中，每日产生海量的文本记录。电力服务热线95598系统会同步保存用户来电详情，持续关注有价值的用户来电事件。若用户来电事件与电力线路故障相关，会下派任务到相应的检修班组进行故障排查检修。在检修故障的整个过程中，将记录包括接报时间、故障发生时间、故障发生地点、现场抢修记录及抢修完成时间等数据。现场抢修记录完整地保存了抢修各个环节，是进行文本挖掘的重要资源。

现场抢修记录包含了故障发生原因、故障路线、故障设备和停电区域等重要数据。使用手动的方式进行录入，使得现场抢修记录存在各种信息偏差和错误。例如，在描述故障原因时，由于

基金项目：国家电网公司科技项目（重点领域大数据应用技术与模型研究SGRIJSKJ（2016）1104）

天气原因造成的线路故障可能会被描述成“大风吹断树枝压导线”“大雨大风吹倒树木”“大风天气树木压线”“大风引起树枝触碰电线”等；针对同一类型故障设备，在描述其名称时也可能存在偏差，如“漏电保护装置”“漏电保护器”“漏保”等都是在描述同类型的设备。通过传统的文本挖掘技术，很难将它们辨别和分类。本文将引入的外部文本库和电力故障抢修文本进行融合，制作成词向量训练语料库，使用基于词向量的深度学习模型对抢修故障文本进行精确分类。

电力行业文本的分布式表示

为了使用文本类型数据，计算机需通过编码的方式将文本转换为数值，即对文本进行向量化表示。自然语言分布式表示是一种通过深度学习算法得到文本表示向量的技术，表示向量又可称为嵌入向量。与one-hot等传统编码技术相比，嵌入向量不仅可以缩减向量维度，同时可以将语义信息嵌入到向量当中。单词级别的分布式表示方法称为词嵌入（Word Embedding），它将每个单词嵌入到一个连续的低维实值向量。

本文使用基于Skip-Gram结构的词向量模型进行训练。对于词序列 $w_1, w_2, w_3, \dots, w_t$ ，其优化目标可以表示为

$$J = \frac{1}{L} \sum_i^L \sum_{-c \leq j \leq c, j \neq 0} \lg p(w_{i+j} | w_i) \quad (1)$$

式中， c 是窗口尺寸； p 是概率函数， $p(w_{i+j} | w_i)$ 表示使用输入词预测前后词语。

模型使用梯度下降法SGD对模型进行求解。Word2vec是谷歌公司开源的一种词向量计算工具，本文使用基于Word2vec的改进模型进行词向量的训练。

电力行业的文本挖掘研究中，训练词向量时主要使用电力行业内文本数据作为输入数据。这些文本词数量相对较小，难以达到百万级别，甚至难以达到十万级别，使得词向量的训练效果较差，给下一步的建模工作带来不良影响。本文在已有研究的基础上，引入外部语料库，将其和电力故障文本组合起来，通过神经网络语言模型（NNLM）进行词向量的训练，其主要结构如图1所示。

本文使用的95598热线故障记录数据为6.1万条，进行分词和去停用词等预处理操作后，包含字数38.6万，词数1.3万，见表1。经过预处理后的

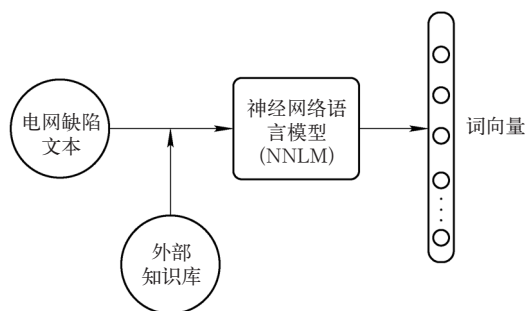


图1 电力文本的分布式表示

词组数量明显减少，以此为输入来训练词向量得到的效果较差。如图2所示，是使用95598数据训练的3维词向量，可以看到与单词“短路”相近的词组为“老化”“闲丰”“一线”“变夏山”。除“老化”外，其余的词组在语义上与“短路”相差甚远。而与“抢修班”相近的词组为“营业班”“抢修三班”“德军”“鲍天佳”，可以看出后两个词组与“抢修班”在语义上也相差甚远。这说明使用较小电力故障文本语料库训练的词向量表达能力有限。

表1 语料库规模

数据集	记录数/万	字数/万	词数/万
95598报修记录	6.1	38.6	1.3
Sougou新闻语料库	282.9	3025.6	50.3
95598 + Sougou	289.0	3064.3	50.9

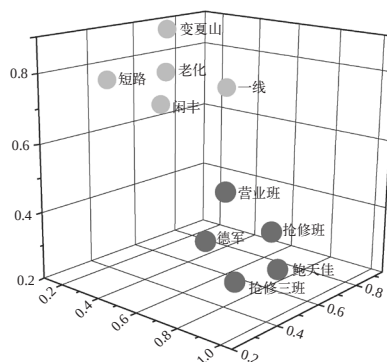


图2 未使用外部语料库训练的3D词向量

将Sougou新闻语料库加入到电力故障文本语料库中，重新训练词向量，得到的词向量如图3所示。可以看出，引入规模较大的外部语料库，词向量的表达能力明显增强，距离较近的词组之间在语义上也相对较近。

基于LSTM的电力故障文本分类模型

1. 长短期记忆网络LSTM

随着技术的不断改进，各类型神经网络被成功应用到文本分类任务中去。在众多模型中，循环神经网络首次将时间顺序的概念引入到神经

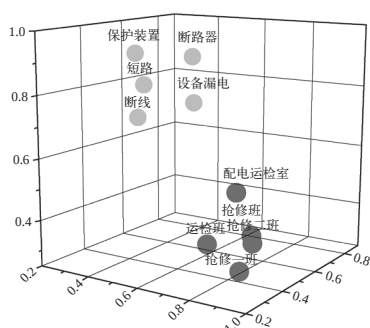


图3 引入外部语料库训练的词向量

网络当中，使其在具有上下文结构的文本数据中表现突出。但是，通过反向传播算法（Back Propagation, BP）训练的循环神经网络存在一定问题。主要问题是梯度消失和梯度爆炸，通过网络传递的梯度要么衰减严重，要么呈指数膨胀。一种可能的改进方法是利用高阶信息进行更复杂的算法优化。然而，这种方法使得计算成本大幅增加，导致这类神经网络对自然语言建模的吸引力降低。

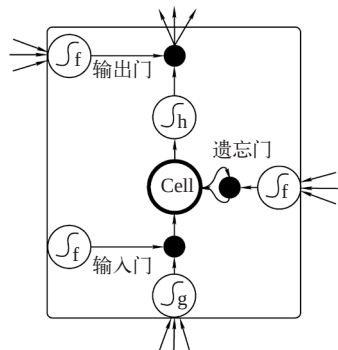


图4 LSTM记忆单元模型

长短期记忆网络LSTM是另一种具有时序结构的神经网络，它对传统RNN网络的结构进行了改进，减少了梯度消失现象，保持了训练算法的稳定性。LSTM模型定义了一个记忆单元，用来替代传统RNN模型中的隐层节点，这种记忆单元的结构如图4所示。添加输入门、输出门和遗忘门，导致网络权重可以不断变化。在模型参数固定的情况下，这种改动使得不同时刻的积分尺度动态变化，降低了梯度消失或者梯度膨胀的出现频率。具体而言，LSTM的前向传播计算方法可表示为

$$f_t = \delta(w_f[h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \delta(w_i[h_{t-1}, x_t] + b_i) \quad (3)$$

$$o_t = \delta(w_o[h_{t-1}, x_t] + b_o) \quad (4)$$

$$\bar{c}_t = \tan h(w_c[h_{t-1}, x_t] + b_c) \quad (5)$$

$$c_t = f_t c_{t-1} + i_t \bar{c}_t \quad (6)$$

$$h_t = o_t \tan h(c_t) \quad (7)$$

遗忘门、输入门和输出门的计算公式分别对应式（2）和式（4）。在 t 时刻，门结构接受上一时刻记忆单元的输出 h_{t-1} 和当前时刻记忆单元的输入 x_t ，与对应的权重向量 w 相乘，再加上偏置向量，通过 Sigmoid函数将输入映射到0~1之间，对信息进行筛选。式（5）~式（7）是对Cell状态进行更新。其中，式（7）计算记忆单元最终的输出，记忆单元通过 $\tan h$ 函数对当前时刻的状态进行计算，使模型变为非线性的，最后由输出门决定将哪些信息输出。

2. 基于LSTM的电力故障文本分类

本文使用LSTM作为模型的后端分类器。模型主要由6层构成，分别是：输入层、Embedding层、LSTM层、Softmax层、Dropout层和输出层，其中，第3层为主要的LSTM层，它对之前的预处理数据进行学习，利用记忆单元的时序特性学习文本数据的上下文信息。Embedding层的主要作用是使用神经网络将输入文本向量化。该层可以直接将预训练向量引入到模型中，也可以通过再训练的方式同步学习词向量。由于输入文本的词向量已经在预处理部分进行了预训练，此处的Embedding层不再参加模型的训练调试，仅仅起到数据格式转换的作用。Dropout层主要作用是利用一种降采样技术来防止网络过拟合。

本文的故障分类任务是一个多分类任务，因此在网络结构中加入了Softmax层。它实际上是一个使用神经网络实现的Softmax回归模型，是logistic回归的一种推广算法，主要用于多分类任务。Softmax回归具有较好的数学性质，不仅可以预测对应的类别，还能输出各类别对应的概率值，方便进一步的处理。一般来说，Softmax层主要用于模型的最后一层，进行最后输出处理。

在试验时，首先使用电力故障文本语料库训练词向量，得到词向量字典 a ；将电力故障文本语料库和搜狗新闻语料库组合成一个大的语料库后，训练得到词向量字典 b 。在构建LSTM模型时，分别使用词向量字典 a 和词向量字典 b 训练数据，最后对比使用不同词向量后得到的模型的效果。具体的算法流程图如5所示。

试验结果和分析

1. 试验准备

试验数据为杭州市2014—2018年，95598服务热线的接报记录数据。这些数据记录了用户报

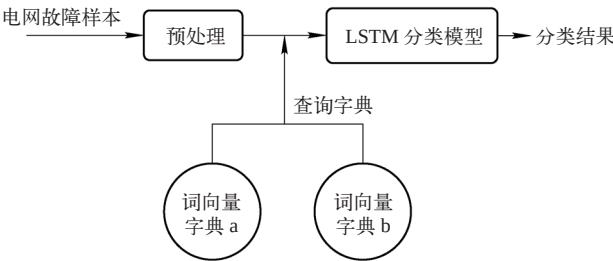


图5 基于LSTM的故障文本分类模型

修、任务下派、故障问题描述、故障原因描述和现场检修记录等内容，为运检工作的自动化和智能化提供了重要的数据支撑。本试验首先对数据进行了初步筛选，去除错误和缺失的数据，并根据文本内容对数据进行人工类别标注。将数据根据引起故障的原因分为5大类，建立电力故障文本语料库，具体类别标签见表2。建立的电力故障文本语料库共有5大类，约61 068条故障文本记录。训练模型时，随机选择48 854条样本作为训练集，其余样本作为测试集。

表2 故障文本分类

类标号	含义	内容
0	电力设备本身问题	令克跌落、设备短路等
1	天气原因	大风、雷雨和大雪等
2	外力破坏	挖断电缆、车辆撞断电杆和动物触电等
3	用户内部原因	用户家中用电器短路等
4	其他	原因不明、未记录故障原因等

同时，为了获得表达能力更强的词向量，本试验引入了外部文本数据——Sougou新闻语料库。Sougou新闻语料库来自搜狐新闻2012年6月~7月期间国内、国际、体育、社会和娱乐等18个频道的新闻数据，提供URL和正文信息。对该语料库提取正文内容，进行预处理后约有282.9万条新闻文本记录。该语料库主要用来和电力故障文本语料库组合训练词向量，引入外部语义信息到电力故障文本中，不参与LSTM的模型训练。

试验分类模型基于开源深度框架Keras进行搭建。模型LSTM层神经元个数为50，激活函数使用 $\tan h$ ，采用交叉熵（Cross Entropy）作为损失函数；Dropout层参数设置为0.5。模型采用Adam算法进行优化。试验使用的计算机CPU为 Intel Core i7-6700HQ；内存为16GB；GPU型号为 NVIDIA GeForce GTX950M；操作系统为 64 位的 Ubuntu 14.04 LTS。

2. 试验结果和分析

为保证试验结果的稳定性，开始前将所有样

本随机打乱。进行 10 折交叉验证，取 10 次结果的平均值作为模型性能的估计值。试验采用5类测试样本的平均准确率作为评价指标。首先，为了对比引入外部语料库后词向量对分类模型的影响，使用电力故障文本语料库训练词向量，并根据该词向量将样本向量化，输入到LSTM分类模型进行文本分类。然后，将电力故障文本语料库和搜狗新闻语料库融合，使用融合语料库训练新的词向量，并根据新的词向量重新训练分类模型。

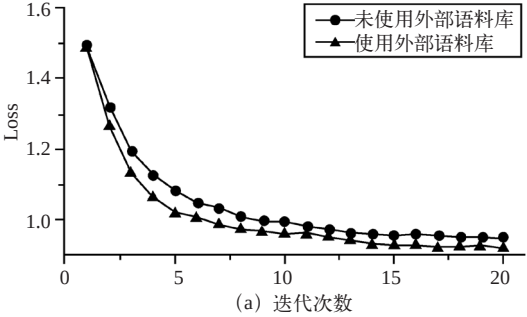
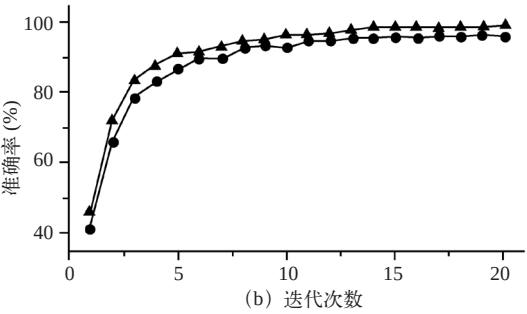


图6 模型效果对比图

在深度学习模型训练过程中，模型的损失Loss曲线可以很好地反映出其拟合能力和泛化能力的变化情况。如图6a所示为使用不同语料库产生的词向量时，LSTM模型在训练过程中损失Loss的变化情况。可以看到，随着迭代次数的不断增加，Loss曲线的值不断减小。但是加入外部语料库训练词向量后，Loss明显减小，且Loss下降的速率明显增大。这充分说明，使用外部语料和电力故障文本语料融合的方法训练词向量后，引入的外部语义知识对电力故障文本的分类起到了积极作用。图6b为使用不同语料库训练的词向量时，LSTM模型在测试集上的准确率变化情况。可以看到，融合外部语料库训练词向量后，准确率明显高于未使用外部语料库时的情况，且准确率上升的速度显著增大。这从另一个方面反映出，引入外部语义知识，有助于加快模型拟合速度，提高模型的泛化能力。EA

（收稿日期：2020.10.13）