Advances in Social Science, Education and Humanities Research, volume 551
Proceedings of the 6th International Conference on Education Reform and Modern Management (ERMM 2021)

Text Classification of Enterprise Technical Requirements Based on RCNN_ATT Model

Xingbing Liu, Bin Chai*, Yingying Wang, Yachao Zhai

College of computer and Information Engineering, Henan Normal University, Xinxiang 453007, China
**Corresponding author. Email: 1479868561@qq.com*

ABSTRACT
The text information of enterprise technical requirements is miscellaneous, which leads to the feature extraction is not prominent enough, and cannot be further accurately and effectively matched to the scientific research team of colleges and universities. In this paper, attention mechanism is added to the two way LSTM network to calculate the contribution score of the category to which the output vector belongs, and the word vector combined with attention matrix is connected to the maximum pooling layer, then RCNN_ATT model for enterprise technical requirement text is proposed, so that the technical requirements text can be automatically classified according to the industry. The experimental results show that, compared with other neural network models, this model performs better in technical requirement text classification, which can narrow the scope of supply and demand matching and improve the efficiency of matching calculation.

Keywords: *Text classification; Attention mechanism, Long and short term memory, Max pooling, Technical requirements.*

基于RCNN_ATT的企业技术需求文本分类

刘行兵, 柴斌*, 王英英, 翟亚超

河南师范大学，计算机与信息工程学院，新乡 453007，中国
**通讯作者. 邮箱: 1479868561@qq.com*

中文摘要
企业技术需求文本信息冗杂，导致提取的特征不够突出，无法进一步精准有效地匹配到高校的科研团队。本文在双向 LSTM 网络后加入注意力机制，计算输出向量所属类别的贡献分值，将结合注意力矩阵的词向量连接最大池化层，提出了针对企业技术需求文本的 RCNN_ATT 模型，使技术需求文本可以根据所属行业完成自动化归类。实验结果表明，与其他神经网络模型相比较，本文模型的技术需求文本分类表现更好，能够达到缩小供需匹配范围和提高匹配计算效率的目的。

关键词: *文本分类，注意力机制，长短时记忆，最大池化，技术需求*

1. 相关工作

中小企业在发展过程中会不断产生新的技术需求问题，但由于信息的不对称，科研团队并不能及时地被这些中小企业所发掘，高校的科研团队人才和科研实验室资源供给无法精准对接到企业的技术需求，这种现象的出现，是因为知识、技术与产业之间存在着巨大的鸿沟。尽管中小企业作为社会经济发展中非常重要的角色，但是在当今社会技术创新热潮的背景下，

依然面临着全面发展和技术升级的巨大考验[1]，大数据时代中的各种信息呈现出爆炸式增长的现象，中小企业技术需求文本信息也包含在其中，由于文本信息错综复杂，人工无法对其有效区分，挖掘与管理，而文本分类可以预先识别含义不明确的文本信息，有效判断其所属类别，是 NLP 中不可或缺的方法。

传统文本分类方法中，一般考虑将文本特征融入机器学习的模型内。文献[2]将 SVM 与 KNN 相结合，提出 SV-NN 特殊组合算法，对哈萨克语文本进行分类，该算法在保证 SVM 分类性能的同时，还解决了 KNN 算法中 k 值选定的问题。文献[3]针对不平衡的文本分类数据集，提出改进的随机森林算法，通过对训练样本的数据类型采取不同的采样处理方法，在 Spark 平台上多结点并行化运行，使算法的效率得到提升。Li 等人[4]使用自然语言处理中的 TF-IDF、Word2Vec 和 TextRank 关键词抽取算法对铁路运输投诉文本数据集进行特征提取，然后用朴素贝叶斯完成对投诉文本的分类，有助于相关部门加强管理，提高乘客的服务质量。

深度学习分类方法中，文献[5]中的 TextCNN 模型通过对句子序列做卷积化处理以完成对文本的分类。文献[6]使用 CNN 在短文本的词向量表示中提取抽象特征，结合 KNN 进行分类，解决了传统 KNN 分类算法中数据稀疏与特征维度太高的问题。由于卷积窗口大小的限制，文本的大部分特征信息不能够被CNN很好的挖掘与学习，且文本中缺乏长期依存关系，针对这种问题，LSTM[7]、BiLSTM[8]、GRU[9]等基于递归神经网络的文本分类模型逐渐被提出，对结合上下文信息文本的语义性进行挖掘，文献[10-11]使用注意力机制对现有的循环神经网络结构进行优化改进，通过对文本分类贡献较大词汇的关注，得到更优的特征表示向量。Lai 等人[12]将 CNN 中的最大池化层连接在双向 LSTM 层之后，提出 RCNN 文本分类模型，该模型的 BiLSTM 结构可以对上下文信息进行捕捉分析，充分解决了卷积神经网络中由于卷积窗口固定，导致词向量上下文信息受限的问题。

企业技术需求的文本分类中，文献[13]对 LDA 主题模型的特征选择方法加以改进，提取文本特征后，结合 SVM 完成企业技术需求文本分类器的设计。文献[14]将定性与定量两种分析法相结合，通过二次聚类构建企业技术创新情报产品需求的分类模型，为实现“精准”情报产品的有效供给奠定基础。目前有关企业技术需求文本分类的文献较少，且都使用的传统机器学习方法来构造分类器，其中提取的文本特征也依赖于人为设计，因此分类精度在一定程度上受限。通过研究，深度学习中的 RCNN 结合 BiLSTM 的结构和 CNN 的池化层，在捕获上下文信息的同时又能挖掘文本关键特征，而且注意力机制能关注对文本分类贡献大的词汇，因此本文在 RCNN 模型结构中加入注意力机制，充分结合三种结构的优势，利用 RCNN_ATT 文本分类模型完成中小企业技术需求文本的行业分类。

2. RCNN_ATT 模型

2.1. 企业技术需求文本向量化

本文在 Google 开源的 Word2Vec 工具包中，选择 Skip-gram 模型预训练向量，然后通过词嵌入来实现企业技术需求文本的向量化，数据集由技术需求文本的标题和内容组成，利用 padding 机制对技术需求的句子长度短填长切，以解决各样本数据的长度无法统一的问题。

文本向量化可以理解为把文本数据中的词语，句子或文章以数字化向量矩阵的形式呈现出来。文本向量化主要有两种方式，最初传统的 One-Hot 词表示方法比较简单，在处理海量的文本数据时，往往面临着维度爆炸的灾难和向量矩阵十分稀疏的问题，而且在这种方法中文本的语义性也没有被考虑到，另外是现在常用的 Word2vec 模型，Mikolov[15-16]在 Bengio[17]原有 NNLP 方法（神经网络概率语言模型）的基础上提出 Word2Vec 中的 CBOW 和 Skip-Gram 两种模型，通过两种模型构建出的词向量有低维稠密的特点，而且词汇的语义信息能够很好的被表示出来。

2.2. 注意力机制

在企业技术需求文本中,句子信息大多比较冗杂，词汇繁多且无用，很有可能造成了对分类贡献大的词语消融在了大量词汇中，以至于文本中重要的技术词语对分类的影响程度大大降低，而注意力机制却能够很好地改善这个问题，通过为词语分配不同的贡献分值，从而有效地降低了关键信息丢失的风险。注意力机制的原理是通过模仿人类大脑对眼前信息的处理机制——快速扫描图像或文字，将视觉焦点锁定在大脑当前感兴趣的区域。文献[18]最初在机器翻译的任务上使用了注意力机制，随后 RNN/CNN[19]等神经网络模型也与注意力机制相结合，注意力机制同时也在其他诸多学术领域中被广泛应用。

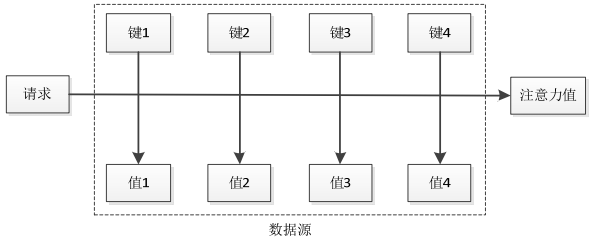

图 1 注意力机制模型

如上图 1 所示，注意力机制首先需要计算请求向量（Query）与每个数据特征的键向量（Key）之间的相似度，$Simi_i$ 表示请求向量 $Query$ 和键向量 Key_i 的相似度，计算两个向量之间相似度一般采用点积和 cosine 两种相似性计算公式，本文则采用公式(1)计算相似度。

$$Simi_i = Query \bullet Key_i \tag{1}$$

$$Simi_i = \frac{Query \bullet Key_i}{\| Query \| \bullet \| Key_i \|} \quad (2)$$

在求得请求向量和键向量的相似度之后，用 $Softmax$ 归一化处理，计算得出 $Simi_i$ 的概率分布，计算过程如公式(3)所示：

$$a_i = Softmax(Simi_i) = \frac{e^{Simi_i}}{\sum_{j=1}^{L} e^{Simi_j}} \quad (3)$$

求得的概率分布 a_i 就是值向量（Value）的权重分布，因此只需要对 Value 进行加权求和就能得到最终的“注意力”结果，其计算过程如公式(4)所示：

$$Attention(Q,K,V) = \sum_j a_j \bullet Value_j \quad (4)$$

2.3. 融合注意力机制的 RCNN 企业技术需求文本分类

本文将企业技术需求文本的数据集划分为训练集、验证集和测试集三部分，经过预训练向量的词嵌入后，得到词嵌入向量表示，再输入本文多结构融合的神经网络中进行训练、验证和测试，最终返回企业技术需求文本的分类结果，分类流程如下图 2 所示。

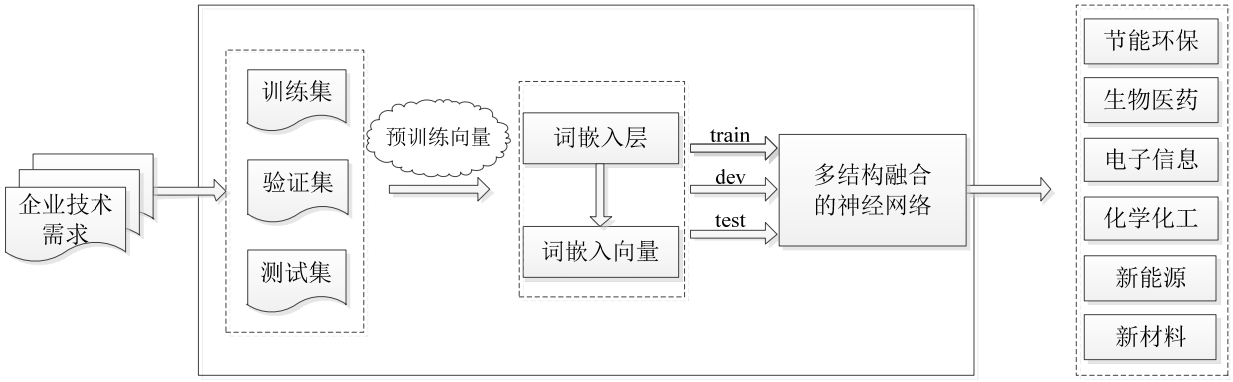

图 2 分类流程图

如下图 3 所示，本文的 RCNN_ATT 模型由 8 层结构组成，分别是 3 个并行的词输入层(左文本，文本，右文本)、词嵌入层、BiLSTM 层、注意力层、合并层、最大池化层、全连接层和输出层。

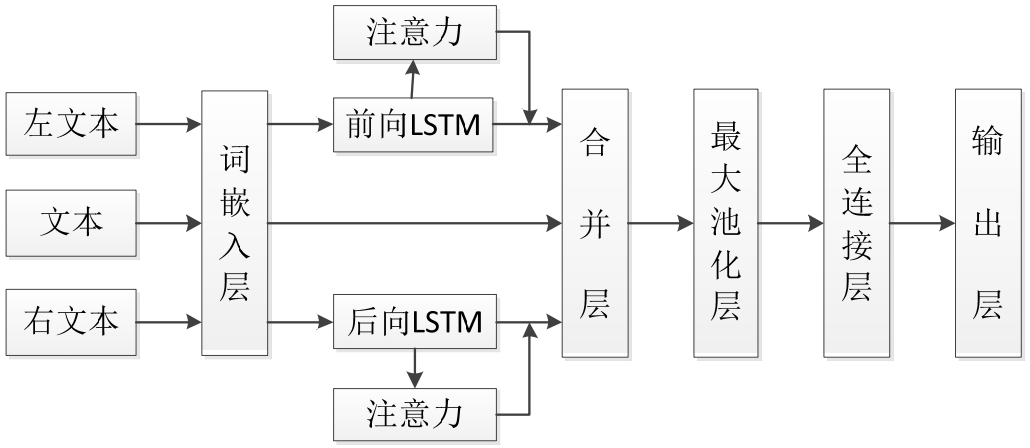

图 3 RCNN_ATT 模型

词输入层由左文本、文本和右文本组成，左右文本以文本为中心，索引值相对应产生减 1 和加 1 的变化。本文将 Skip-Gram 模型预训练的词向量应用于词嵌入层，可以把技术需求文本中的词汇表示成向量，左右文本分别经过词嵌入层，BiLSTM 层和注意力层，获得左右文本表示向量，然后和文本的词嵌入向量在合并层拼接成组合表示向量，再输入到下一层。

BiLSTM 层处理一句话的过程如下图 4 所示，使用 LSTM 网络结构，不仅使传统 RNN 结构中曾出现过的梯度消失以及梯度爆炸问题得到解决，而且让网络中的长时依赖问题得到一定缓解，相比较于单向的 LSTM，BiLSTM 结构更能够提取到上下文信息中更深层的语义向量，比如在“你不要有思想包袱”这句话中，可能提取到特征词“包袱”，但这里的“包袱”是一个多义词，并不能明确其具体含义，但是利用 BiLSTM 网络可推断此处的“包袱”意思为“负担”。

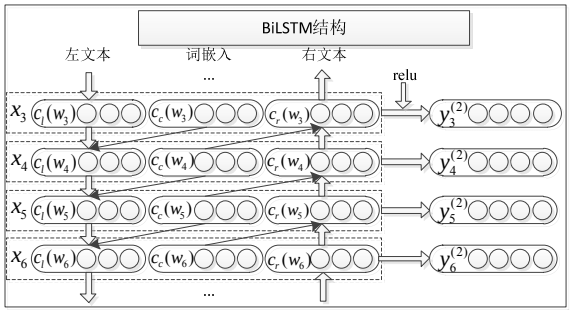

图 4 BiLSTM 模型

词汇 w_i 左侧和右侧的上下文信息表示的计算公式如下(5)和(6)所示：

$$c_l(w_i) = f(W^{(l)}c_l(w_{i-1}) + W^{(sl)}e(w_{i-1})) \tag{5}$$

$$c_r(w_i) = f(W^{(r)}c_r(w_{i+1}) + W^{(sr)}e(w_{i+1})) \tag{6}$$

左右文本的上下文信息表示分别输入到注意力层，通过 softmax 计算得出注意力矩阵 ATM_l 和 ATM_r。注意力矩阵会重新返回到 BiLSTM 层，与原输出向量加权相乘，得到注意力加权后的信息表示 $c_{la}(w_i)$ 和 $c_{ra}(w_i)$。

$$ATM_l = Softmax(W^{hl}c_l(w_i) + b^{hl}) \tag{7}$$

$$c_{la}(w_i) = ATM_l \bullet c_l(w_i) \tag{8}$$

$$ATM_r = Softmax(W^{hr}c_r(w_i) + b^{hr}) \tag{9}$$

$$c_{ra}(w_i) = ATM_r \bullet c_r(w_i) \tag{10}$$

词嵌入层输出的文本词嵌入向量 $c_c(w_i)$，与输入到合并层中的左文本上下文信息表示 $c_{la}(w_i)$ 和右文本上下文信息 $c_{ra}(w_i)$ 合并为三元向量组，以向量 x_i 表示。

$$x_i = [c_{la}(w_i); c_c(w_i); c_{ra}(w_i)] \tag{11}$$

在得到一条文本内包含上下文信息的第 i 个词 w_i 对应的特征表示向量 x_i 后，使用线性变换公式(12)转换每个 x_i，并附加偏置向量 $b^{(2)}$，再用激活函数 $relu$ 输出。

$$y_i^{(2)} = relu[W^{(2)}x_i + b^{(2)}] \tag{12}$$

对输出的特征向量 $y_i^{(2)}$ 进行最大池化，即对句子序列做一维全局最大池化，保留其中一个特征。式(13)中的 max 表示元素最大化函数，对 $y_i^{(2)}$ 做最大池化后输出特征 $y^{(3)}$，然后将 $y^{(3)}$ 输入到全接神经网络（14）中。

$$y^{(3)} = \max_{i=1}^{n} y_i^{(2)} \tag{13}$$

$$y = Wy^{(3)} + b \tag{14}$$

数字输出 y 可以运行 Softmax 函数，经过公式(15)转换成概率输出 p_i，输出的预测类别即为概率分布最大值对应的类别，然后通过 argmax 函数将概率值映射为输入的企业技术需求文本所属的行业标签值 l^*。

$$p_i = \frac{\exp(y)}{\sum_{k=1}^{n}\exp(y_k)} \tag{15}$$

$$l^* = \arg\max_{l} \sum_{k=1}^{n} p_i \tag{16}$$

3. 实验结果与分析

3.1. 实验数据集

本文的数据集，除了在科学家在线网（www.scientistin.com）、技需网（www.topposer.com）、技 E 网（www.ctex.cn）和科易网（www.1633.com）爬取外，还采用了全国各省市科学技术信息研究所和高校科技处近三年内发布的企业技术需求项目文件。

企业技术需求文本属于科技文本范畴，科技文本资源一般包含有大量特有的领域内专用词汇，因此科技文本区别于新闻，娱乐等生活中的普通文本，在训练词向量时，基于互联网百科等中文语料无法完全对应需求文本中包含的专有技术词汇，不能满足对科技文本的分类，所以需要选择相关的科技文本对百科语料集进行补充，除原有已获取的 1.89G 维基百科中文语料外，本文从 Soopat 专利数据库网爬取相关领域专利文本约 20 万条，共同作为语料集训练词嵌入向量。

对获得的原始数据去除一些重复无用的词汇和非必要特殊字符后，最终整理筛选出 7 类代表行业的企业技术需求文本共 70000 条，训练集，测试集和验证集之间的比例为 18：1：1。每一类文本的训练集数量为 9000 条，验证集和测试集各为 500 条，因此训练集总数为 63000 条，验证集和测试集总数各为 3500 条，不同类别的示例样本如表 1 所示。

表 1. 各类数据的样本

类别	企业技术需求文本示例
节能环保	寻求水处理新技术方面的合作，开发新型高效节能水处理设施，解决自控净水技术等水处理工艺课题…
生物医药	寻求红曲霉的研究与应用的技术需求，加大对红曲霉的研究与应用，应用于实际生产以提高浓香型白酒质量和产量…
机械制造	寻求生产汽车涡轮增压器中的涡轮（镍合金）的一种精密铸造技术…
电子信息	开发一套全新的智能无人值守自动呼叫的输液监测系统，实现医院输液的全自动可视化监测报警显示。围绕医院输液安全性与智能化的问题，重点突破…
化学化工	对常规湿法炼锌过程中锌精矿带入的硒铊元素的分布规律和行为规律进行究，以及硒铊对析出锌的影响机理，最终找出硒铊的脱除方法…
新能源	设计与开发基于建筑一体化的太阳能集热与光伏器件，构建超低能耗建筑太阳能热-电-冷联产复合能量系统研究技术体系…
新材料	寻求一种新型的 TVOC 吸附分解材料：要求其在吸附甲醛、TVOC 的同时，又能及时分解…

3.2. 实验设置

实验在 Torch 深度学习框架下进行，CPU 为 Intel Core i7-8700 3.2GHz，内存大小为 32GB，实验编程语言为 python 3.7，网络的搭建和程序的运行在 64 位 Win 10 系统的 PyCharm 软件上实现。

对语料集分词、去除停用词等预处理后，选择 Word2Vec 中的 Skip-Gram 模型训练词嵌入向量，设置训练窗口为 5，最小词频为 5。本文选择生物医药，机械制造和新能源三大领域中常用的三个词汇——“药物”，“金属”和“燃料”，加载训练好的 Word2Vec 模型，分别计算得出与其语义相近词汇的 TOP4 及其对应的相似值，词汇相似度如表 2 所示。

表 2. 词汇相似度

药物	相似度	金属	相似度	燃料	相似度
抗生素	0.8635	合金	0.7738	燃油	0.7905
中药	0.8534	铝	0.7406	核燃料	0.7469
制剂	0.8441	非金属	0.7310	推进剂	0.7420
口服	0.8207	机械	0.7168	煤油	0.7293

同时构建了 TextCNN、BiLSTM、BiLSTM_ATT、和 RCNN 作为参照模型与本文模型的分类性能进行对比。各模型所使用卷积神经网络和长短时记忆神经网络的基本实验参数设置详情如表 3 和表 4 所示。

表 3. CNN 网络参数

参数	值	参数	值
滤波器数量	256	Batch_size	128
卷积核尺寸	[2，3，4]	Epoch	10
Dropout	0.5	Pad_Size	32
学习率	1e-3		

表 4. LSTM 网络参数

参数	值	参数	值
神经元数目	256	Batch_size	128
隐藏层层数	2	Epoch	10
激活函数	relu	Pad_Size	32
学习率	1e-3		

3.3. 结果分析

各模型在不同词向量维度下分类准确率的表现变化情况如表 5 和图 3 所示。从表 5 中可以看出，本文的 RCNN_ATT 模型分类性能表现最好。

在图 5 中，通过五个维度的词向量对应的分类准确率变化趋势对比可以看出，各模型都在维度值为 250 时达到最佳，且本文模型的准确率相比于其他模型最高，单一网络结构的模型中，比 TextCNN 模型的分类准确率提升了 1.40%，比 BiLSTM 模型的分类准确率提升了 1.69%，多网络结构的融合模型中，比 BiLSTM_ATT 模型的分类准确率提升了 0.57%，比 RCNN 模型的分类准确率提升了 0.51%。

表 5. 各模型在不同维度下的准确率（%）

维度	TextCNN	BiLSTM	BiLSTM_ATT	RCNN	RCNN_ATT
100	88.69	89.17	90.18	91.55	92.03
150	90.47	91.23	92.29	91.85	93.05
200	91.75	91.18	92.37	92.65	92.75
250	**91.85**	**91.56**	**92.68**	**92.74**	**93.25**
300	91.36	91.12	92.08	92.57	92.85

图 6 和图 7 分别为各模型训练集损失值和训练集准确率随时间变化曲线，从图 4 可以看出本文模型加入注意力机制，提取每个词汇被聚焦的概率值，筛选出应该需要重点关注的词汇，使算法收敛速度更快，同时也极大程度减少了震荡，从而实现快速精确的分类。

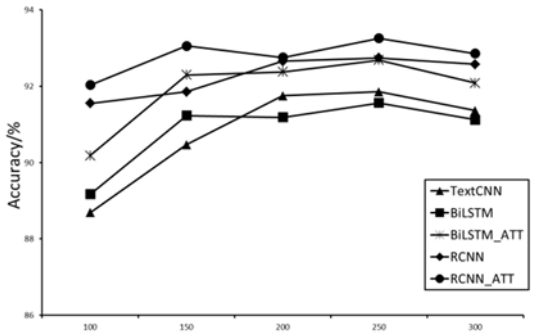

图 5 词向量维度对分类性能的影响

在图 5 中，通过与其他四种分类模型作对比，可以看出本文模型的分类准确率更高，且到达较高准确

率的时间明显缩短。实验证明本文 RCNN_ATT 模型的分类效果更佳，池化层考虑到对关键特征信息的挖掘，双向 LSTM 考虑到上下文语义联系，而本文又考虑到注意力机制能够聚焦对文本分类贡献大的词汇，综合各模型的优点，使网络的结构获得进一步的优化。

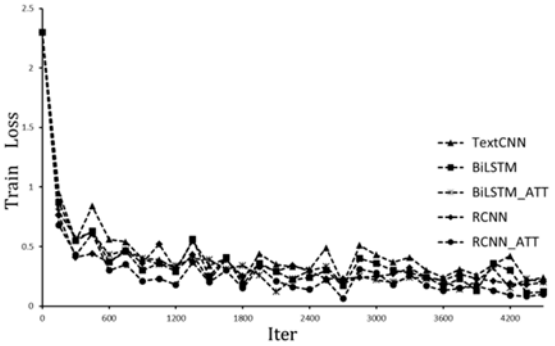

图 6 训练集损失值随时间变化曲线

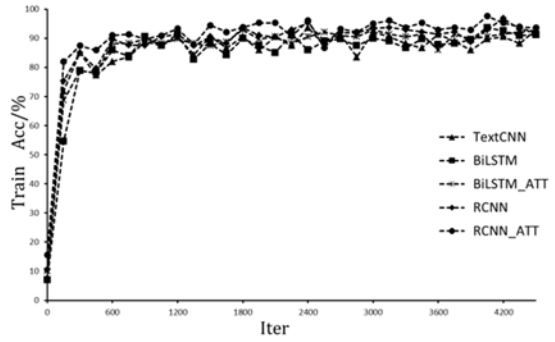

图 7 训练集准确率随时间变化曲线

4. 结束语

本文结合双向 LSTM、注意力机制和最大池化，充分考虑各网络结构的优点，使模型的分类性能获得提升。通过 RCNN_ATT 模型对企业技术需求文本进行行业的自动分类，为科技协同服务平台的管理者在整理科技文本资源过程中提供便利，也与接下来与高校科研团队研究方向文本的有效匹配奠定了基础，进一步促进了校企合作，推动产学研深度融合与发展。在接下来的实验中，会考虑继续结合更好的网络结构，尝试与其他模型提取的特征表示向量进行加权融合计算，以寻求更优特征表示向量，使文本的特征信息获得更加充分有效的表达。

REFERENCES

[1] Yurong Zeng, Lin Wang, Xianhao Xu. An integrated model to select an ERP system for Chinese small and medium-sized enterprise under uncertainty[J]. Technological and Economic Development of Economy, 2017, 23(1):38-58.

[2] Alimjan GULNAZ, Jumahun HURXIDA, Tieli Sun, et al. The nearest neighbor text classification method based on support vector[J]. CAAI transactions on intelligent systems, 2018, 13(5): 799–807.

[3] Zheng Peng, Lingjiao Wang, Hua Guo. Parallel Text Categorization of Random Forest[J]. Computer Science, 2018, 45(12), 148-152.

[4] Lifeng Li, Wenxing Li, Daqing Gong. Naive Bayesian Automatic Classification of Railway Service Complaint Text Based on Eigenvalue Extraction[J]. Tehnički vjesnik, 2019, 26(3):778-785.

[5] Kim Y. Convolutional neural networks for sentence classification[C]. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014.1746–1751.

[6] Yabo Yin, Wenzhong Yang, Huiting Yang, et al. Research on short text classification algorithm based on convolutional neural network and KNN[J]. Computer Engineering, 2018, 44(7):193-198.

[7] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.

[8] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing. 1997, 45(11):2673-2681.

[9] Bengong Yu, Peihang Zhang. WPOS-GRU patent classification method based on two-channel feature fusion[J]. Application Research of Computer. 2020, 37(03), 655-658.

[10] Du C, Huang L. Text classification research with attention-based recurrent neural networks[J]. International Journal of Computers Communications & Control, 2018, 13(1):50-61.

[11] Bin Feng, Youwen Zhang, Xin Tang, et al. Power Equipment Defect Record Text Mining Based on BiLSTM-Attention Neural Network[J]. Proceedings of the CSEE, 2020:0258-8013.

[12] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015:2267-2273.

[13] Liping Zhang. Design and implementation of text classifier for enterprise technology requirements [D]. Southeast University, 2017.

[14] Wenyi Rui, Zhenyan Yuan, Ming Yin, et al. Research on classification of enterprise technological innovation information product demand based on innovation chain[J]. Information engineering. 2018, 4 (06), 75-86.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector

space[C]. Proceedings of the International Conference on Learning Representations. arXiv:1301.3781, 2013.

[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Proceedings of the Advances in Neural Information Processing Systems, 2013:3111-3119.

[17] Bengio Y, Schwenk H, Senecal J S, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6):1137-1155.

[18] Bahdanau D, Cho K, Bengio Y. Nerual machine translation by jointly learning to align and translate[C]. International Conference on Learning Representations. arXiv:1409.0473, 2015.

[19] Qing Shao, Huiping Ma. Text classification model based on convolutional neural network with self-attention mechanism[J]. Minicomputer system, 2019, 40(6):1137-1141.