文章编号：2096-1472(2023)-03-58-05
DOI:10.19644/j.cnki.issn2096-1472.2023.003.012

融合Bert和BiLSTM的中文短文本分类研究

郝 婷[1]，王 薇[2]

(1.长春大学网络安全学院，吉林 长春 130022；
2.长春大学计算机科学技术学院，吉林 长春 130022)
✉997236440@qq.com; 20017008@qq.com

摘 要：中文短文本具有长度短及上下文依赖强的特点，针对新闻主题文本用词不够规范、语义模糊、特征稀疏等问题，提出首先在词向量表示阶段引入Bert(双向Transformers编码器)生成融合字、文本及位置的词向量作为训练文本的词表征进行文本语义增强，然后将得到的词向量输送到BiLSTM(双向长短期记忆网络)网络中提取上下文关系特征，最后使用Softmax分类器进行文本分类，模型准确率达0.9391。通过与其他主流方法进行对比和实验验证，实验结果表明，文章提出的方法在进行新闻短文本分类时有良好效果。

关键词：文本分类；BiLSTM；Bert；新闻分类
中图分类号：TP391.1 **文献标识码**：A

Research on Chinese Short Text Classification Model based on Bert and Bi-LSTM

HAO Ting[1], WANG Wei[2]

(1.*College of Cyberspace Security*, *Changchun University*, *Changchun* 130022, *China*;
2.*College of Computer Science and Technology*, *Changchun* 130022, *China*)
✉997236440@qq.com; 20017008@qq.com

Abstract: Chinese short texts have the characteristics of short length and strong context dependence. Aiming at the problems of the non-standard words, semantic ambiguity, sparse features and other problems in news subject texts, this paper proposes to introduce Bert (a two-way transformers encoder) to generate a word vector that combines words, text and location. The word vector is used as the word representation of the training text for text semantic enhancement. Then, the obtained word vector is transmitted to the BiLSTM (Bidirectional Long and Short-term Memory) Network to extract context features. Finally, the Softmax classifier is used for text classification, and the accuracy rate is 0.9391. Compared with other mainstream methods and verified by experiments, the proposed method has better effect in news short text classification.

Keywords: text classification; BiLSTM; Bert; news classification

1 引言(Introduction)

互联网的蓬勃发展产生了海量的数据信息，人们进入大数据时代，文本数据通过互联网快速增长，人们可以时刻接触和处理海量的文本信息，如新闻、微博和商品评价等。此类文本具有海量性、实时性和不规则性等特点且大多属于非结构化的短文本数据，使得短文本的语义发散，特征词难以提取。如何对短文本进行准确、高效的分类是目前的研究热点。传统机器学习算法如朴素贝叶斯[1]和支持向量机[2-3]等方法常用于文本分类，但是这些算法存在对文本深层语义和上下文关联信息挖掘方面的短板。

近年来，基于神经网络算法的文本分析被广泛应用[4]。区别于传统基于统计机器学习，深度学习模型有多层网络，每层包含多个可进行非线性变换的神经元，因此具有更强的非线性拟合能力，在数据量较大的情况下效果更好。2018 年，预训练模型开始兴起。PETERS等[5]构建的新型语言模型(Embeddings from Language Models，ELMo)生成的词向量可以随语境进行多义词动态变换。Google[6]提出的Bert(Bidirectional Encoder Representations from

Transformers)通过其双向结构能够学习到词的上下文表示，该模型横扫了多项自然语言处理任务的排行榜纪录，极大地推动其发展。

本文主要从传统词向量语义表达上存在问题和短文本由于特征稀疏导致重要特征较难提取两个方面入手，提出融合Bert和BiLSTM的复合网络模型Bert-BiLSTM。通过在本文所选中文数据集上进行实验，结果显示本文所提模型分类效果良好。

2 文本分类相关工作(Related work of text classification)

文本分类是自然语言处理的重要任务之一，其过程为使用机器按照规定的分类标准对需要进行分类的文本进行自动分类标记。目前，关于英文文本分类的研究较多，针对中文文本分类的研究相对较少。分析原因，一方面是相关的语料库较少，另一方面是中文文本表示比英文复杂，采用传统方法难以提取其特征。

2.1 文本向量化

文本表示是文本分类任务中非常重要的步骤，通过文本表示过程将其转化成计算机能够处理的数据信息，其好坏影响后续模型的表现，最重要的是如何选择合适的表示方法，并且应当尽可能地包含原本的信息，这是因为一旦在空间映射时丢失了信息，则在后续的模型处理中再也无法获取。良好的文本向量可以更好地在向量空间中有一个文本空间映射，从而使得文本可以计算。自然语言处理领域因其自身的特性而难以向量化，并且存在多种高级语法规则及其他特性，比如近义词、反义词、上下文之间的联系等。文本表示过程的实质是对文本特征进行建模。

2.1.1 One-Hot Encoding(独热编码)

传统文本表示方法中最基本的表示方法是One-Hot编码方式。One-Hot Encoding是最早的一种比较直观的词向量生成方式。这种映射方式通过汇总语料库里的所有词汇得到N个词汇，并将每个个体生成一个N维向量。这是一种较为简单的映射方式，仅利用了单词的相关位置信息，没有把单词的语义信息考虑在内，并且随着语料库的增加，会产生“维度灾难”问题。

2.1.2 Word Embedding(词嵌入)

词向量采取稠密向量对文本进行表示，使“维度灾难”问题得以解决，因此被广泛应用于各种自然语言处理任务中。钟桂凤等[7]使用Word2Vec(词嵌入)进行词向量的训练，并采用改进注意力机制的方法进行文本分类。Word2Vec根据预测方法提出了连续词袋模型(CBOW)和跳元模型(Skip-gram)两种模型结构。CBOW模型预测目标词语采取的方法为根据上下文进行预测；Skip-gram则是根据当前出现的词预测上下文的模型。FastText(快速文本分类)模型[8]是对Word2Vec模型的一种改进，用于预测中心词。方炯焜等[9]同时考虑了文本的局部信息与整体信息，采用全局词向量(Global Vectors，GloVe)模型，再利用GRU(门控循环单元)进行训练。下游文本分类任务效果的提升得益于Word2Vec、GloVe等模型训练得到的词向量特征表示，但本质上这些模型属于静态的预训练技术，即便是在不同的上下文中，同一词语可能会有相同的词向量，所以会出现一词多义的问题，这也导致在下游分类任务中的技术性能受限问题。

2.1.3 Bert词向量

2018年以来，基于Transformer的预训练模型相继被提出，并用于不同的下游任务。Bert模型可以捕捉更深层次的语义信息，基于Bert的文本分类模型是由预训练(Pre-Training)和预微调(Fine-Tuning)两个部分构成。预训练采用自监督训练，使用大量未经标注的文本语料完成训练，可以很好地学习到文本语义特征和深层次的文本向量表示；预微调的起点为预训练Bert模型，其拟合和收敛则需根据具体的分类任务完成。杨彬[10]提出在罪名和相关法律条文文本分类任务中使用Bert词向量结合Attenton-CNN模型，取得了比较好的分类效果。

2.2 循环神经网络

在自然语言处理领域发展迅猛的有循环神经网络(Recurrent Neural Network，RNN)，并在文本分类任务中得以广泛应用，循环神经网络是用于建模序列化数据的，并且可以捕获长距离输入依赖的一种深度学习模型。但是，循环神经网络在处理文本时可能会出现“梯度消失”或“梯度爆炸”问题，学习能力有限。张云翔等[11]采用长短期记忆网络进行文本分类，该网络降低了循环神经网络的学习难度，长短时记忆神经网络(Long Short-term Memory，LSTM)模型是对RNN的扩展，可以对有价值的信息进行长期记忆，解决了循环神经网络存在的“梯度消失”或“梯度爆炸”问题。与此同时，一些组合模型也相继被提出用于解决文本分类题，田园等[12]采用双向LSTM网络模型提取文本的上下文信息，并融合注意力机制以提高文本分类效果。吴小华等[13]对文本进行情感分析时，利用基于自注意力机制的双向长短时记忆网络可以得到更好的文本句法信息；XIAO等[14]提出了char-CRNN模型，首先进行卷积操作，然后用循环神经网络进行特征的提取。

文本特征融合可以学习到更好的特征表示，即最具差异性的信息能从融合过程中涉及的多个原始特征向量中获得。本文针对中文新闻文本进行分类模型研究，综合考虑了Bert模型在文本表示方面和BiLSTM在语言模型构建的特征优化方面表现的优点，提出了基于Bert的特征融合网络模型Bert-BiLSTM。本文采用的融合方式为特征层次融合[15]，首先使用神经网络将原始词向量转化成高维特征表达，然后针对提取到的高维特征进行融合。中文文本分类整体流程如图1所示。

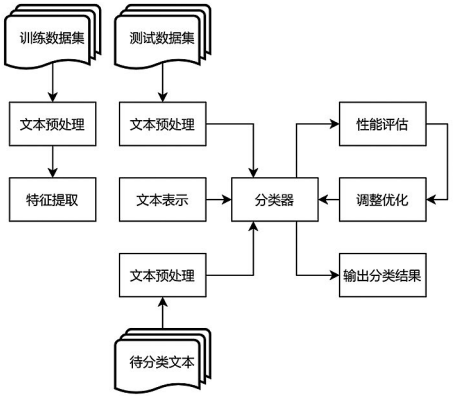

图1 中文文本分类流程图

Fig.1 Flow chart of Chinese text classification

3 相关理论与技术(Relevant theories and technologies)

Bert-BiLSTM模型结构图如图2所示。本模型在上游部分使用Bert生成的字符向量作为字符嵌入层，在下游部分将BiLSTM作为特征提取器进行建模，并使用Dropout降低过拟合风险，最后输入Softmax函数预测文本分类。Bert和BiLSTM的结合可以获得更复杂的语义特征，构建更准确的语义表达。

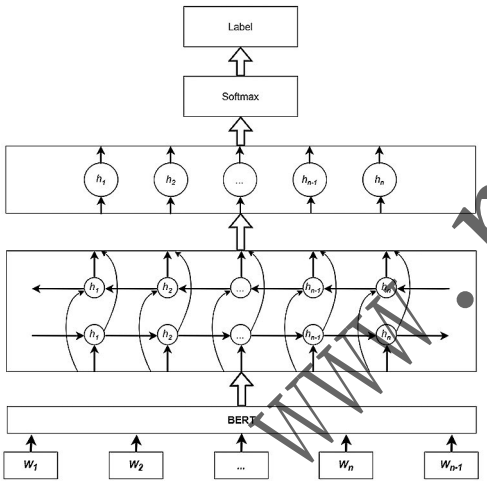

图2 模型总框架图

Fig.2 General framework of the model

3.1 Bert 词嵌入

Bert模型采用双向Transformer编码器获取文本的特征表示。多层双向Transformer编码器的输入为字符级的文本，训练过后输出为文本字符级特征。Bert词嵌入结构图如图3所示，由嵌入层、分割嵌入层及位置嵌入层构成。本文选用Bert做文本的词嵌入，将文本向量改变格式后输送到Bert中进行编码，便得到句子中每个字的向量表示。由于Bert使用更大规模的语料进行模型的训练，所以这也加强了词嵌入模型的泛化能力，使得文本序列中字符级、单词级、句子级及句与句间关系的上下文特征得到了更充分的描述。Bert的这一特点适用于新闻标题文本较短但含义丰富的特征，可以得到更好的词嵌入信息。

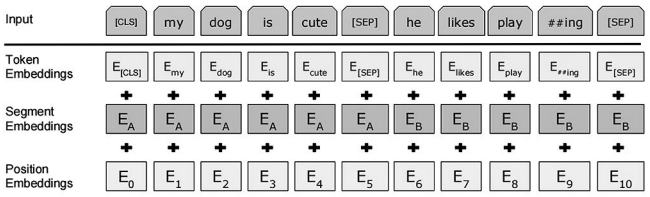

图3 Bert词嵌入结构图[6]

Fig.3 Structure diagram of Bert word embedding

3.2 BiLSTM模型

LSTM即长短期记忆网络，是RNN(循环神经网络)的一种变体，其解决了RNN存在的长期依赖问题。LSTM具有遗忘门、输入门和输出门，其结构如图4所示。

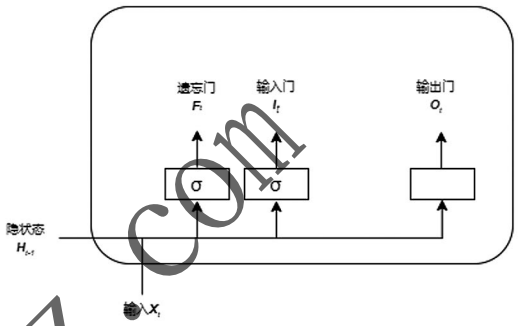

图4 LSTM结构

Fig.4 LSTM structure

双向循环网络由1 个正向LSTM和1 个反向LSTM构成。单向LSTM根据前一时刻的信息预测当前时刻的输出。BiLSTM与LSTM一样，具有门控状态，可以捕捉更长距离的信息，使循环神经网络长期依赖的问题得以有效解决。BiLSTM模型可以将各个字符以句子的形式进行表达，并且考虑字符之间的依赖关系。因此，本文选择使用BiLSTM捕捉每个单词的上下文语义信息，其结构如图5所示。

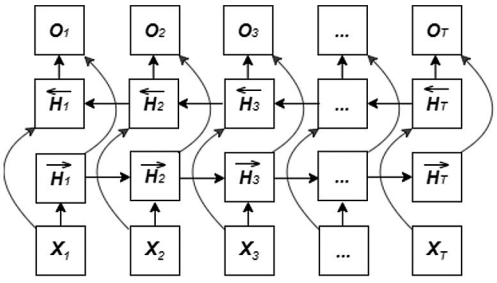

图5 BiLSTM结构

Fig.5 BiLSTM structure

4 实验(Experiment)

4.1 实验数据

本文使用的数据来自THUCNews新闻数据集[THUCNews是根据新浪新闻RSS订阅频道2005—2011 年的历史数据筛选过滤生成，包含74万篇新闻文档(2.19 GB)，均为UTF-8纯文本格式]。在开源网站搜集到的THUCNews的新闻标题文本数据集，共20万条数据，包含财经、房产、股票等10 个类别，其中训练集数量为16万条，测试集数量为2万条，验证集数量为2万条。

4.2 实验环境

本文实验环境为操作系统Windows 10，显卡型号为GTX2060，开发语言为Python 3.10，搭建深度学习模型使用框架为Pytorch。

4.3 评价指标

本文对分类结果进行评估的指标为*Precision*(精确率)、*Recall*(召回率)、*F*1值和*Accuracy*(准确率)。精确率是指分类正确的正样本个数占模型判定为正样本的样本个数的比例。召回率是指分类正确的正样本个数占真正的正样本的样本个数的比例。只有当精确率与召回率的数值同为1时，*F*1值才能达到最大。*F*1-*score*是*Precision*与*Recall*两个指标的结合，可以更加全面地反映分类性能。用*F*1值评估模型性能时，模型性能越好，*F*1值越接近于1，是衡量分类效果的重要评价指标。准确率是指分类正确的样本占总样本个数的比例。相关计算如式(1)—式(4)所示。*TP*表示实际正样本预测为正，*TN*表示负样本预测为负，*FP*表示负样本预测为正，*FN*表示正样本预测为负。

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$P=\frac{TP}{TP+FP} \tag{2}$$

$$R=\frac{TP}{TP+FN} \tag{3}$$

$$F1=\frac{2TP}{2TP+FP+FN} \tag{4}$$

4.4 对比实验设置

为了验证本文所提模型对网络新闻主题分类的有效性，选择以下被广泛应用于新闻分类的模型进行对比，具体对比情况如下。

BiLSTM：词向量由Word2Vec训练所得，并且作为词嵌入层输入BiLSTM层中进行特征提取。

AttentionBiLSTM：由BiLSTM和Attention组合的复合网络模型。

FastText：JOULIN等[9]提出的快速文本分类方法，其训练速度较基于CNN和RNN的模型要快得多。

Bert-RCNN：输入为Bert学习到的词向量，然后通过RCNN[16]进行进一步的学习。该网络由循环神经网络学习文本的上下文表示，文本中的关键信息再用最大池化层捕获。实验结果如表1所示。

表1 各模型实验结果

Tab.1 Results of each model experiment

模型	精确率	召回率	*F*1	准确率	*Loss*
BiLSTM	0.7240	0.7237	0.7233	0.7237	0.85
Attention-BiLSTM	0.7239	0.7196	0.7197	0.7196	0.85
FastText	0.9207	0.9199	0.9200	0.9199	0.26
Bert-RCNN	0.9224	0.9217	0.9218	0.9217	0.26
Bert-BiLSTM	0.9396	0.9391	0.9391	0.9391	0.21

4.5 实验结果与分析

Bert-BiLSTM模型在测试集上对每一种分类进行测试，实验结果如表2所示。与不同模型的实验对比结果如图6所示，实验结果证明Bert词嵌入模型与BiLSTM模型融合后的分类效果更好。

分析表2中的数据可知，使用Word2Vec的BiLSTM、Attention-BiLSTM的分类效果比Bert-BiLSTM差，证明预训练模型在提取句子语义特征表示方面优于Word2Vec。为了进一步证明本文使用的BiLSTM模型对特征提取的有效性，本文选择Bert-RCNN进行实验对比。从表2中可以看出，本文使用的Bert-BiLSTM组合模型的分类效果最佳。Bert-BiLSTM模型相较于Bert-RCNN模型，其准确率提升了0.0174。所提模型在分类时已经达到较高的精度。使用预训练词向量的模型和使用Word2Vec词向量的模型相比，使用了预训练词向量的模型准确率明显提升。

通过分析以上实验结果可得，本文构建的基于Bert-BiLSTM新闻短文本分类模型具有比其他基线模型更强的特征提取与特征组合能力，适用于处理新闻短文本分类任务，相比其他模型具有更出色的表现和效果。综上所述，本文所提Bert-BiLSTM模型在进行短文本分类时，获得的分类效果较好。

表2 Bert-BiLSTM模型实验结果

Tab.2 Bert-BiLSTM model experimental results

类别	精确率	召回率	*F*1
财经	0.9400	0.9240	0.9319
房产	0.9555	0.9450	0.9502
股票	0.8972	0.8990	0.8981
教育	0.9395	0.9780	0.9584
科技	0.8930	0.9260	0.9092
社会	0.9115	0.9480	0.9294
时政	0.9462	0.8970	0.9209
体育	0.9800	0.9820	0.9810
游戏	0.9792	0.9400	0.9592
娱乐	0.9539	0.9520	0.9530

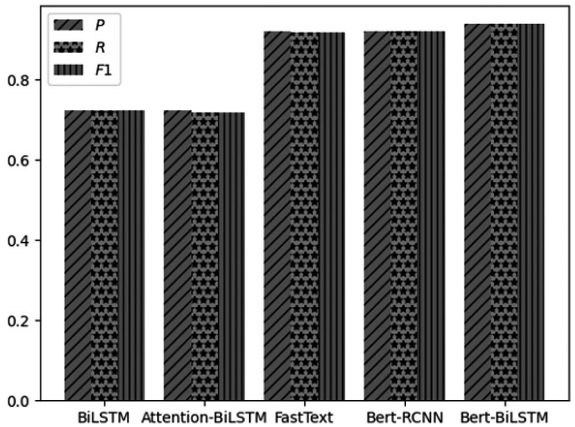

图6 各模型效果对比

Fig.6 Comparison of effects of various models

4.6 错误样本分析

从THUCNews新闻测试集中选取4 条预测错误的例子进行解释，具体分析结果如表3所示。

表3中，新闻样本(1)的实际类别为娱乐，预测类别为社会，该分类相对合理，这条新闻同时具有娱乐新闻和社会新闻两条属性。新闻样本(2)的实际类别为时政，但是被分类为娱乐，分析原因可能是“戛纳”一词偏娱乐属性。新闻样本(3)的实际类别为教育，但是被分类为社会，分析原因可能是“防身手册”偏社会属性。新闻样本(4)的实际类别为教育，预测类别为财经，分析原因可能是这条新闻可以认为是教育问题也可以认为是财经问题。

表3 错误样本分析表

Tab.3 Error sample analysis table

新闻样本	实际类别	预测类别
(1)去新西兰体验舌尖上的饕餮之旅(组图)	娱乐	社会
(2)戴安娜王妃车祸纪录片受英国抵制将在戛纳上映	时政	娱乐
(3)防身手册：女生留学必注意的18件事教育社会	教育	社会
(4)留学金融机构大全：汇丰银行(中国)	教育	财经

通过以上分析可以看出，模型在对比较有深意的文本进行分类时，效果较差，并且分类效果也受语料影响。但是，从新闻样本(1)和新闻样本(4)的分类结果可以看出，模型的分类结果具有合理性，并且能精准地识别类别。

5 结论(Conclusions)

互联网的快速发展产生了大量短文本，短文本不但有内容特征稀疏的特点，而且存在上下文依赖较强的问题。近年来，基于词向量的双向循环神经网络优势显著，成为文本分类任务的主流。本文针对文本表示模型中的词向量在不同语境下的词语多义问题，综合考虑了Bert模型在文本表示和BiLSTM在语言模型构建的特征优化方面的优势，提出基于Bert的特征融合网络模型(Bert-BiLSTM)，使用Bert模型获取文本的特征表示，将得到的特征表示输入BiLSTM网络中进行进一步的特征提取。通过实验证明，本文所提方法在进行新闻短文本分类时获得了良好的分类效果。

参考文献(References)

[1] 杨晓花,高海云.基于改进贝叶斯的书目自动分类算法[J].计算机科学,2018,45(08):203-207.

[2] 庄婷婷,李冬梅,檀稳,等.基于分层支持向量机的微博用户自杀倾向预测与分析[J].哈尔滨工程大学学报,2019,40(11):1890-1895.

[3] 周跃.基于SVM的文本分类算法研究[D].安徽:合肥工业大学,2021.

[4] 刘礼文,俞弦.循环神经网络(RNN)及应用研究[J].科技视界,2019(32):54-55.

[5] PETERS M, NEUMANN M, LYYER M, et al. Deep contextualized word representations[J]. Proceedings of NAACL, 2018, 9(32):2227-2237.

[6] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. Computation and Language, 2018, 23(2):3-19.

[7] 钟桂凤,庞雄文,隋栋.基于Word2Vec和改进注意力机制AlexNet-2的文本分类方法[J].计算机科学,2022,49(04):288-293.

[8] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of Tricks for Efficient Text Classification[EB/OL]. (2016-08-09)[2020-10-13]. https://arxiv.org/pdf/1607.01759v3.pdf.

[9] 方炯焜,陈平华,廖文雄.结合GloVe和GRU的文本分类模型[J].计算机工程与应用,2020,56(20):98-103.

[10] 杨彬.基于BERT词向量和Attention-CNN的智能司法研究[D].大连:大连理工大学,2019.

[11] 张云翔,饶竹一.基于LSTM神经网络的电网文本分类方法[J].现代计算机,2020(02):8-11.

[12] 田园,马文.基于Attention-BiLSTM的电网设备故障文本分类[J].计算机应用,2020,40(S2):24-29.

[13] 吴小华,陈莉,魏甜甜,等.基于Self-Attention和Bi-LSTM的中文短文本情感分析[J].中文信息学报,2019,33(06):100-107.

[14] XIAO Y, CHO K. Efficient Character-level Document Classification by Combining Convolution and Recurrent Layers[EB/OL]. (2016-02-01)[2021-05-25]. https://arxiv.org/pdf/1602. 00367.pdf.

[15] ATREY P K, HOSSAIN M A, SADDIK A E, et al. Multimodal Fusion for Multimedia Analysis: A Survey[J]. Multimedia Systems, 2010,16(6):345-379.

[16] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// AAAI. National Conference on Artificial Intelligence. Texas: AAAI Press, 2015:2267-2273.

作者简介：

郝 婷(1993-)，女，硕士生.研究领域：自然语言处理.

王 薇(1975-)，女，硕士，教授.研究领域：机器学习与智能数据分析.本文通信作者.