计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200020

基于 BERT 与 Bi-LSTM 融合注意力机制的中医病历文本的提取与自动分类

杜 琳[1] 曹 东[1] 林树元[2] 瞿溢谦[2] 叶 辉[1]
1 广州中医药大学医学信息工程学院 广州 510000
2 浙江中医药大学基础医学院 杭州 310000
(3051095449@qq.com)

摘 要 中医逐渐成为热点,中医病历文本中包含着巨大而宝贵的医疗信息。而在中医病历文本挖掘和利用方面,一直面临中医病历文本利用率低、抽取有效信息并对信息文本进行分类的难度大的问题。针对这一问题,研究一种对中医病历文本的提取与自动分类的方法具有很大的临床价值。文中尝试提出一种基于 BERT+Bi-LSTM+Attention 融合的病历短文本分类模型。使用 BERT 预处理获取短文本向量作为模型输入,对比 BERT 与 word2vec 模型的预训练效果,对比 Bi-LSTM+Attention 和 LSTM 模型的效果。实验结果表明,BERT+Bi-LSTM+Attention 融合模型在中医病历文本的提取和分类方面达到了最高的 AverageF1 值(即 89.52%)。通过对比发现,BERT 较 word2vec 模型的预训练效果有显著的提升,且 Bi-LSTM+Attention 模型较 LSTM 模型的效果有显著的提升,因此提出的 BERT+Bi-LSTM+Attention 融合模型在病历文本抽取与分类上有一定的医学价值。

关键词:BERT;Bi-LSTM;Attention;LSTM
中图法分类号 TP391.1;TP183

Extraction and Automatic Classification of TCM Medical Records Based on Attention Mechanism of BERT and Bi-LSTM

DU Lin[1],CAO Dong[1],LIN Shu-yuan[2],QU Yi-qian[2] and YE Hui[1]
1 School of Medical Information Engineering,Guangzhou University of Traditional Chinese Medicine,Guangzhou 510000,China
2 School of Basic Medical,Zhejiang Chinese Medical University,Hangzhou 310000,China

Abstract The development of traditional Chinese medicine has gradually become a hot topic,among which the medical records of traditional Chinese medicine contain huge and valuable medical information. However,in terms of the text mining and utilization of TCM medical records,it is always difficult to extract effective information and classify them. To solve this problem,it is of great clinical value to study a method of extracting and automatically classifying TCM medical records. This paper attempts to propose a short medical record classification model based on BERT+ Bi-LSTM +Attention fusion. BERT preprocessing is used to obtain the short text vector as the input of the model,to compare the pre-training effect of BERT and word2vec model,and to compare the effect of Bi-LSTM +Attention and LSTM model. The experimental results show that BERT+ Bi-LSTM +Attention fusion model achieves the highest Average F1 value of 89.52% in the extraction and classification of TCM medical records. Through comparison,it is found that the pre-training effect of BERT is significantly improved compared with that of word2vec model,and the effect of Bi-LSTM +Attention model is significantly improved compared with that of LSTM model. Therefore,the BERT+ Bi-LSTM +Attention fusion model proposed in this paper has certain medical value in the extraction and classification of medical records.

Keywords BERT,Bi-LSTM,Attention,LSTM

1 引言

医疗文本包含价值丰富的医疗信息,是进行疾病预测、个性化信息推荐、临床决策支持、用药模式挖掘等的重要资源[1]。随着中医地位的不断提升,中医逐渐成为热点,中医病历文本中包含巨大而宝贵的医疗信息,但是目前仍然存在中医病历利用率低、抽取分类信息难度大的问题。中医病历文本中的中医病历常常是将患者的病症、用药、诊断、基本信息等信息杂糅在一起记录,形成非结构化且无序的长文本中医病历文本,宝贵的中医病历文本资源不能得到充分挖掘和使

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:2017 国家重点科技计划(2017YFB1002302);2019 国家重点研发计划(2019YFC1710400)
This work was supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (2017YFB1002302) and National Key Research and Development Project (2019YFC1710400).

通信作者:叶辉(yehui@gzucm.edu.cn)

用。因此，尝试对中医病历文本进行提取和自动分类具有很高的医学价值。该文尝试抽取中医病历文本中的描述句子，形成病历短文本训练集，通过机器学习和自然语言处理的技术返回每一个句子的具体类别，示例如表 1 所列。

表 1　中医病历文本分类示例

Table 1　Text classification examples of TCM medical records

ID	Input (short medical record)	Output (category)
S1	无食汗	表征
S2	晨起鼻塞流涕	表征
S3	小便不频急	二便
S4	经前白带多	经带
S5	脉浮弦微滑	脉症
S6	苔薄白腻	舌症

Sun 在研究神经网络情感倾向性时，提出 ABGC 模型，将 Attention 机制加入到 Bi-LSTM 中，避免了 LSTM 的梯度消失问题，解决了 LSTM 忽略上下文寓意的问题，有效提高了文本分类的准确率[2]。Yang 等在 Bi-LSTM-CRF 融合模型的基础上引入了 Attention 机制，Attention 机制能够获得词在当前全文范围中的上下文表示，Yang 等将该模型应用在化学药物实体识别的任务上，在生物文本上预训练词向量和 LSTM 模型，最终获得了 90.77% 的 F1 值[3]。Wang 等在研究警情文本命名实体识别问题时使用了 BERT-Bi-LSTM-Attention-CRF 混合神经网络模型，在词向量训练阶段使用 BERT 模型代替传统的 Skip-gram 和 CBOW 模型，最终获得了 90.77% 的 F1 值。Wang 等在研究警情文本命名实体识别问题时使用了 BERT-Bi-LSTM-Attention-CRF 混合神经网络模型，在词向量训练阶段使用 BERT 模型代替传统的 Skip-gram 和 CBOW 的方式，提升了词向量的表证能力，同时解决了中文语料采用字向量训练时词语边界的划分问题[4]。Yang 等针对传统词向量表示无法表征字的多义性问题，提出了 BERT-BiGRU-CRF 模型，BERT 预训练语言模型通过双向 Transformer 结构动态生成字的上下文语义表示，比传统的词向量表示更能表征语句特征[5]。2018 年 Google 提出的一种新型语言模型 BERT[6]，在各类 NLP 任务上达到了目前最好的结果。Jiang 在面向中医方剂配伍的中药文本挖掘研究中采用双向 LSTM-CRF 进行一般药理作用识别，并与双向 LSTM-Softmax、基于词典、基于规则、基于词典与规则组合的方法进行比较，识别准确率分别达到了 0.933 8，0.929 2，0.744 7，0.695 6 和 0.789 2[7]。Yao 等研究了将中医临床记录分为 5 类主要疾病的问题，使用临床语料库对 BERT 语言模型进行微调，获得宏观 F1 值为 88.64% ± 0.40%[8]。

因此，结合医学病历文本表述精简、上下文关系密切等特点，该文尝试选择 BERT 模型作为获取中医文本词向量的训练方法，结合 Bi-LSTM + Attention 模型（以下简称 BiL-Att）进行模型训练，提出一种 BERT-Bi-LSTM + Attention（以下简称 BERT-BiL-Att）的融合模型，进行中医病历文本的提取与自动分类研究。

2　研究方法

2.1　中医病历特点

中医特色本身决定了中医病历的特殊性和规范性，中医病历的特点是在一般电子病历的基础上增加中医诊断和中医特色治疗的项目内容，即在一般电子病历中增加中医诊断和中医特色治疗的数据元。利用这些数据元就可以对中医特色的中医临床治疗、中医临床研究进行支持[9]。中医病历的内容包括：中医四诊、辨证、立法、处方以及西医检查和诊断，包括现代医学的诊疗信息，同时包括中医药学辨证论治的诊疗信息。中医病历病案首页也与西医不同，它具有西医的所有内容，另外又增加了中医的诊断信息，因此中医病历信息量相比西医病历信息具有量大、上下文联系更加紧密的特点。

2.2　BERT 模型训练词向量

BERT 模型是近几年提出的一种新的语言表征模型，通过超大数据、巨大模型和极大的计算开销训练而成，在多项自然语言处理任务中取得了优异的效果。BERT 模型中的 Transformer 层采用双向编码器表示，可以通过一个额外的输出层进行微调（fine-tuning），能够通过联合调节所有层中的上下文来预先训练深度双向表示[8]。此外，BERT 模型为了增加对上下文的记忆，BERT 使用遮蔽语言模型来实现预训练的深度双向表示，在训练双向语言模型时以较小的概率把少量的词替成了 Mask 或者另一个随机的词。中医病历文本中上下文联系紧密，常常需要结合上下文才能获得精准的信息，BERT 上述的特性在中医病历文本中的应用对改善模型的效果具有明显的帮助。

BERT 模型的结果如图 1 所示。

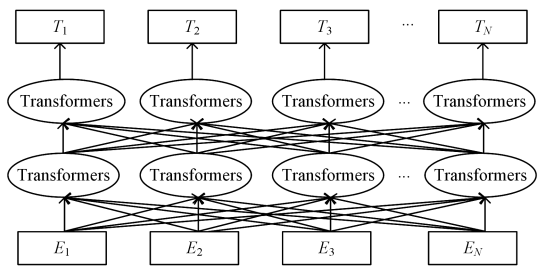

图 1　BERT 模型结构图

Fig. 1　Structure diagram of BERT model

文献[10]采用 BERT 预训练语言模型训练中文词向量，较完整地保存了医学文本语义信息，提升了模型的上下文双向特征抽取能力。使用句子注意力机制对文本语义信息进行编码使得传统的 Bi-LSTM 模型，对语义信息的利用更为充分，提升了模型对文本的识别率。

2.3　中医文本分类训练模型

2.3.1　双向长短记忆网络 Bi-LSTM 机制

Bi-LSTM 由前向 LSTM 与后向 LSTM 组合而成，从前后两个方向训练，最终连接到同一层输出[11]，从而具有记忆过去和将来信息的能力，它解决了传统的 LSTM 模型由于序列化处理问题无法捕捉上下文信息的困难，在理论上能够提升分类的准确率。Bi-LSTM 模型结构图如图 2 所示。

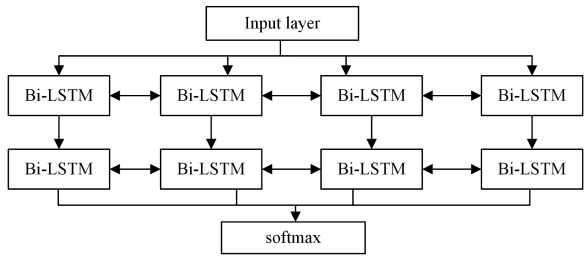

图 2　Bi-LSTM 模型结构图

Fig. 2　Structure diagram of Bi-LSTM model

在 Bi-LSTM 中，对一个句子“I loveApple”进行编码的过程如图 3 所示。前向的 $LSTM_L$ 依次输入“i”“love”“Apple”得到 3 个向量 $\{h_{L0}, h_{L1}, h_{L2}\}$。后向的 $LSTM_R$ 依次输入“i”，“love”，“Apple”得到 3 个向量 $\{h_{R0}, h_{R1}, h_{R2}\}$。最后将前向和后向的隐向量进行拼接得到 $\{[h_{L0}, h_{R2}], [h_{L1}, h_{R1}], [h_{L2}, h_{R0}]\}$，即 $\{h_0, h_1, h_2\}$。

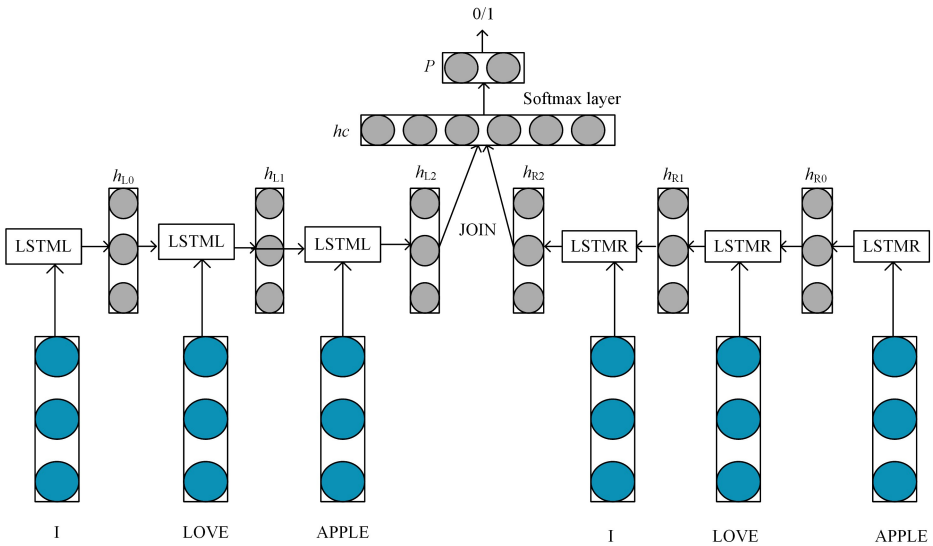

图 3 Bi-LSTM 模型编码示意图

Fig. 3 Schematic diagram of Bi-LSTM model coding

2.3.2 Attention 注意力机制

Attention 机制近年来在从机器翻译到各种各样的任务中显示出了巨大的成功。Attention 功能可以描述为映射查询和输出的键值对集，如图 4 所示，计算 Attention 值的主要 3 个步骤是：

(1)对序列和键值进行相似度计算得到权重；

(2)使用 Softmax 函数对(1)计算的权重进行归一化；

(3)将权重和相应的键值进行加权求和计算最终的 Attention。

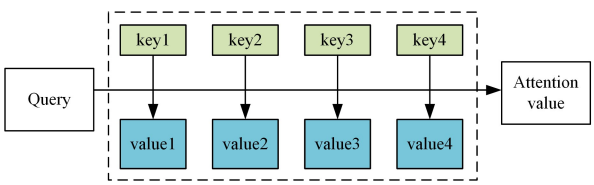

图 4 Attention 机制功能结构图

Fig. 4 Function structure of Attention mechanism

因为 Bi-LSTM 中存在 LSTM 的梯度消失问题和忽略上下文寓意的问题，为了解决该问题，本文增设 Attention 机制，通过区分不同特征的重要程度，忽略不重要的特征，将注意力放在重要的特征上，提高了分类准确率。Attention 模型对 Bi-LSTM 存在的问题的解决步骤分为以下 3 步：1)保留 Bi-LSTM 编码器对输入序列的中间输出结果；2)训练一个选择性学习的模型，将 1)中的结果作为输入；3)在 Attention 输出时将输出序列与 2)中的模型进行关联。

2.3.3 BERT-BiL-Att 融合模型机制

在采用 BERT 模型进行词向量训练的基础上，结合 BiL-Att 模型进行中医医学文本抽取与分类。Bi-LSTM 与 Attention 融合模型就是在 Bi-LSTM 的模型上添加 Attention 层，在 Bi-LSTM 中用最后一个时序的输出向量作为特征向量，选择 Softmax 函数进行分类。Attention 模型是先计算每个时序的权重，然后将所有时序的向量进行加权，并将结果作为特征向量，再选择 Softmax 函数进行分类[12]。结合 BERT 获得的短文本向量，作为 Bi-LSTM 与 Attention 融合模型的输入，构建完整的 BERT-BiL-Att 融合模型机制进行训练与预测。

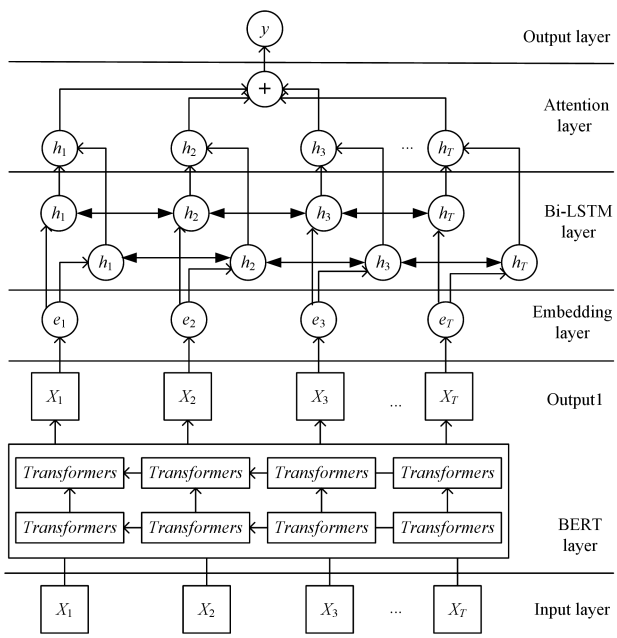

图 5 BERT-BiL-Att 融合模型结构图

Fig. 5 Structure diagram of BERT-BiL-Att fusion model

3 实验步骤

3.1 数据预处理

本文使用的中医病历文本数据来自于某中医名医工作室数据，数据量共计 6118 条，由多位中医教授共同标注中医病历文本的标签名称即所属类别，累计设计了 13 个类别。该文选择所有实验数据的 30%作为测试数据，70%的数据作为训练数据，即选择 1835 条中医文本数据作为测试数据，4283 条作为训练数据，对标注数据进行整理和清洗之后得到的中医

病历短文本具体数据结构如表 2 所列。

表 2　中医病历文本数据结构示例

Table 2　Example of TCM records text data structure

Number	Text content	Category
S1	汗较前多(以前无汗)	表征
S2	出现早晨、晚上和睡醒咳嗽	表征
S3	偶有滴沥不尽大便 1 日 1～2 次	二便
S4	经期小腹疼凉、腰酸疼凉、手足凉不麻	经带
S5	脉左沉细无力右脉近无	脉症
S6	舌红嫩苔白腻水润边齿痕中裂痕	舌症

3.2　BERT 机制获取文本向量

本文采用 BERT 模型进行数据预处理，将自然语言字词转化为计算机可以识别运算的向量，具体的方法描述为，将文档中的一行抽取为一个句子，对每个句子进行分词，获得每个词的词向量，通过对词向量进行加权，计算该句子的句向量，同理可以计算整个文档文本向量。因此一个普通的自然语言文档经过 BERT 处理后生成一个 768 维的向量，这个向量就是模型的输入。

例如，选取训练集中的表征类 TXT 文档，未处理的文档内容如图 6 所示。该文档经过预处理后得到的向量如图 7 所示。

夜眠觉热
出汗不多
左手微潮微凉
无鼻塞流清涕

图 6　表征类文档原文示例

Fig. 6　Example of representation class document text

```
[[-0.26787433  0.11554431  -0.88197297  …  -0.06789639
  -0.5923385  0.2913299]
 [0.23357767  0.30197334  -0.19711249  …  0.05476131
  -0.62948644  0.4216648]
 [-0.12248445  0.09298227  -0.1357589  …  0.49766797
  0.11956039  -0.01948904]
 [0.07720596  0.12524968  -0.36508483  …  0.27214274
  -0.4024575  0.29295844]]
```

图 7　对图 6 所示文本获取的文本向量

Fig. 7　Text vector obtained from text in Fig. 6

3.3　BERT-BiL-Att 融合模型实现医学文本分类

实验过程中 BERT-BiL-Att 模型构建的具体步骤如图 8 所示。

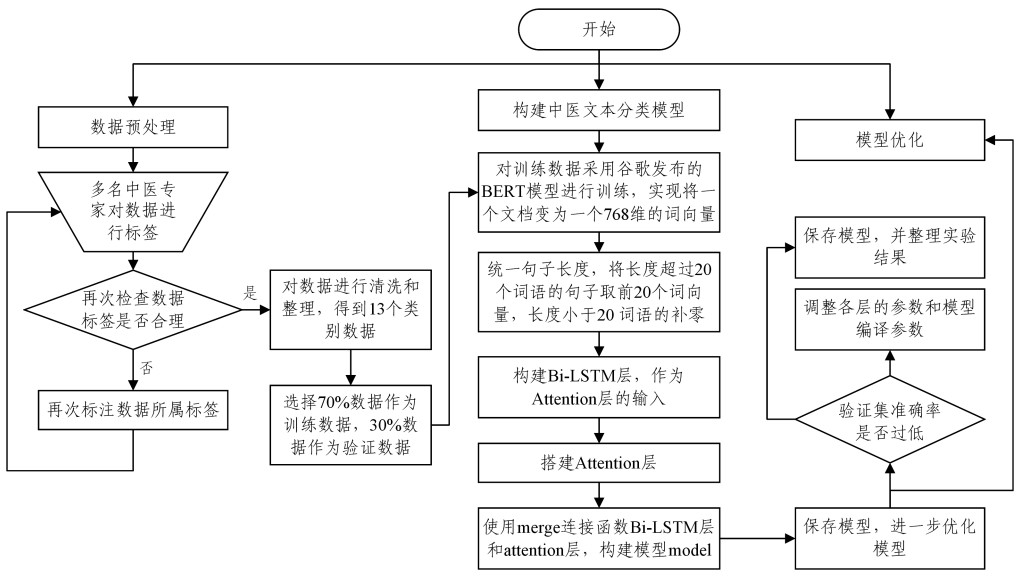

图 8　BERT-BiL-Att 融合模型进行文本分类的步骤

Fig. 8　Flow chart of text classification steps in BERT-BiL-Att fusion model

BERT-BiL-Att 融合模型指在采取 BERT 作为词向量与训练的基础上，使用 Bi-LSTM 的模型，并添加 Attention 层。在 Bi-LSTM 中用最后一个时序的输出向量作为特征向量，选择 Softmax 函数进行分类。Attention 层的作用是先计算每个时序的权重，然后将所有时序的向量进行加权，并将结果作为特征向量，再选择 Softmax 函数进行分类。

4　结果分析

该文采用的数据集共有 13 个类别，分别表示为 C_1，…，C_i，…，C_{13}，每个类别都有各自计算的准确率、召回率和 F1 值，该文采用的评价指标为宏观准确率 P 值、宏观召回率 R 值和 AverageF1 值，它们的计算公式定义如下。

每个类的准确率和召回率的计算公式表示为：

$$\text{准确率 } P_i=\frac{\text{正确预测为类别 } C_i \text{ 的样本个数}}{\text{预测为 } C_i \text{ 类的样本个数}} \tag{1}$$

$$\text{召回率 } R_i=\frac{\text{正确预测为类别 } C_i \text{ 的样本个数}}{\text{真实的 } C_i \text{ 类的样本个数}} \tag{2}$$

宏观准确率 P 值的计算表达式为：

$$P=\left(\frac{1}{n}\right)\sum_{i=1}^{n}P_i \tag{3}$$

宏观召回率 R 值的计算表达式为：

$$R=\left(\frac{1}{n}\right)\sum_{i=1}^{n}R_i \tag{4}$$

AverageF1 值的计算表达式为：

$$Average\cdot F1=\left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{2*P_i*R_i}{P_i+R_i} \tag{5}$$

表 3　4 种模型的评价指标比较

Table 3　Comparison of evaluation indexes of 4 models

Model	P	R	$Average.F1$
word2ve-LSTM	0.7192	0.6767	0.6719
BERT-LSTM	0.8543	0.8564	0.8515
word2vec-BiL-Att	0.7645	0.7660	0.7630
BERT-BiL-Att	0.8964	0.8984	0.8952

在使用词向量模型获取句向量的情况下，LSTM 模型和 BiL-Att 模型训练的结果存在明显差异，BiL-Att 融合模型的效果有显著提升。在使用相同训练模型的情况下，采用 BERT 预训练和 Word2vec 预训练的结果存在明显差异，BERT 模型的效果有显著提升。由表 3 可知，BERT-BiL-Att 融合模型的效果最佳得到的 AverageF1 为 89.52%。说明 BERT-BiL-Att 融合模型对于医学短文本的分类效果较优。

5 结论

本文通过构建句子 BERT 词向量计算文本词向量，将结果作为模型输入，进行医药短文本分类，尝试对比 BERT 和 word2vec 两种词向量模型，对比 BiL-Att 和 LSTM 两个模型。研究发现 BERT 获取的词向量进行中医病历文本分类有较好的效果，BERT 模型中的 Transformer 层采用双向编码器表示能够很好地增强文本上下文记忆。研究发现 BiL-Att 模型的分类效果最好。BiL-Att 模型中融合 Attention 机制，Attention 机制通过计算时序向量进行加权，将和权重作为特征向量的方法，解决了 Bi-LSTM 模型梯度消失问题和忽略上下文寓意的问题。BERT-BiL-Att 模型能够返回每一条筛选标准的具体类别，为临床实验筛选标准自动解析问题提供了研究基础，具有很大的临床价值。

结束语 该文针对中医病历挖掘和利用方面面临的利用率低、抽取分类信息难度大的问题，尝试提出一种基于 BERT＋Bi-LSTM＋Attention 融合的病历短文本分类模型。首先对中医病历数据进行预处理，采用多位中医专家对中医病历短文本数据进行人工标注，累计设计 13 个类别，选择所有实验数据中的 30%作为测试数据，70%数据作为训练数据。对标注数据进行整理和清洗，采用 BERT 词向量模型获取中医病历短文本向量，将短文本词向量作为模型输入。构建 Bi-LSTM＋Attention 文本分类模型，进行模型训练并且调整参数。实验结果表明，BERT＋Bi-LSTM＋Attention 模型的分类效果达到了最高，Average F1 值为 89.52%。通过对比 word2vec 模型发现，BERT 模型的预训练效果有显著的提升，通过对比 LSTM 模型发现，Bi-LSTM＋Attention 模型的分类效果有显著的提升。因此认为本文提出的 BERT＋Bi-LSTM＋Attention 融合模型在病历文本抽取与分类上有一定的医学价值。

本文实验现存待改进之处为，本文所采用的中医病历数据所使用的标签来自多位中医专家共同标注，由于标注数据量较大且有主观因素，因此可能存在数据所属类别有一定的争议性和不准确性的问题。

该文未来的工作可以尝试解决多个类别之间存在的数据不平衡的问题。

参 考 文 献

[1] ZHOU Y. Research on medical text analysis mining technology based on machine learning [D]. Beijing: Beijing Jiaotong University, 2019.

[2] SUN C A, DING Y, TIAN G. Analysis of emotional tendency of neural network based on GLU-CNN and attention-bilstm [J]. Shandong University of Science and Technology, 2019, 40(7): 62-66.

[3] YANG P, YANG Z H, LUO L, et al. Recognition of chemical drug named entity based on attention mechanism [J]. Computer Research and Development, 2008, 55(7): 1548-1556.

[4] WANG Y, WANG M X, ZHANG S, et al. BERT based alert text named entity recognition [J/OL]. Computer application: 1-7 [2019-11-21]. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2020&filename=JSJY202002040&v=p5eBltYjMg96L0PeNbfh3J3eRmFBqc7Bb8ovNkpxL0WtdeRGGLxEJhgL4xxvx4DQ.

[5] YANG P, DONG W Y. Recognition method of Chinese named entities based on BERT embedding [J/OL]. Computer Engineering: 1-7. [2019-11-21]. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2020&filename=JSJC202004006&v=fpDLQvPDFGf6wfMKb3vnBnPPBJfHHDXcSaF%25mmd2Bu59DcVlTutrMRBjr1z9Ri0PG2Gqa.

[6] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of DeepBidirectional Transformers for Language Understanding [J]. CS. CL, 2018: 4171-4186.

[7] JIANG Y. Research on Chinese medicine text mining for TCM prescription compatibility [D]. University of Electronic Science and Technology of China, 2019.

[8] YAO L, JIN Z, MAO C S, et al. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora[J]. Journal of the American Medical Informatics Association: JAMIA, 2019, 26(12).

[9] GUO X P. On constructing electronic medical records conforming to the characteristicsof traditional Chinese medicine [J]. Journal of Traditional Chinese Medicine Management, 2009, 17(5): 469-470.

[10] BIN Y. Intelligent Judicial Research Based on BERT Sentence Embedding and Multi-Level Attention CNNs[C]// International Informatization and Engineering Associations: Computer Science and Electronic Technology International Society. 2019: 7.

[11] ULLAHA, AHMAD J, MUHAMMAD K, et al. Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features[J]. IEEE Access, 2017, PP(99): 1-1.

[12] HU T T, FENG Y Q, SHEN L J, et al. Selection of main features of LSTM speech emotion based on attention mechanism [J]. Acoustic Technology, 2019, 38(4): 414-421.

DU Lin, born in 1998, undergraduate. Her main research interests include artificial intelligence and natural language processing.

YE Hui, born in 1978, postgraduate, is a member of China Computer Federation. His main research interests include medical natural language processing.