自然语言处理

实践报告

2022 年　1 月　10　日

一、 实验概述

本实验通过对缺血性卒中患者病历进行预处理、分析。分别构建了基于 TextCNN，BILSTM，BERT 的基于电子病历辅助诊断模型，并在数据集上对模型效果进行了对比。

二、模型简介

2.1 TextCNN

TextCNN 将卷积神经网络应用到文本分类上，利用多个不同 size 的 kernel 来提取句子中的关键信息，从而能够更好地捕捉局部相关性。模型框架如下图所示。

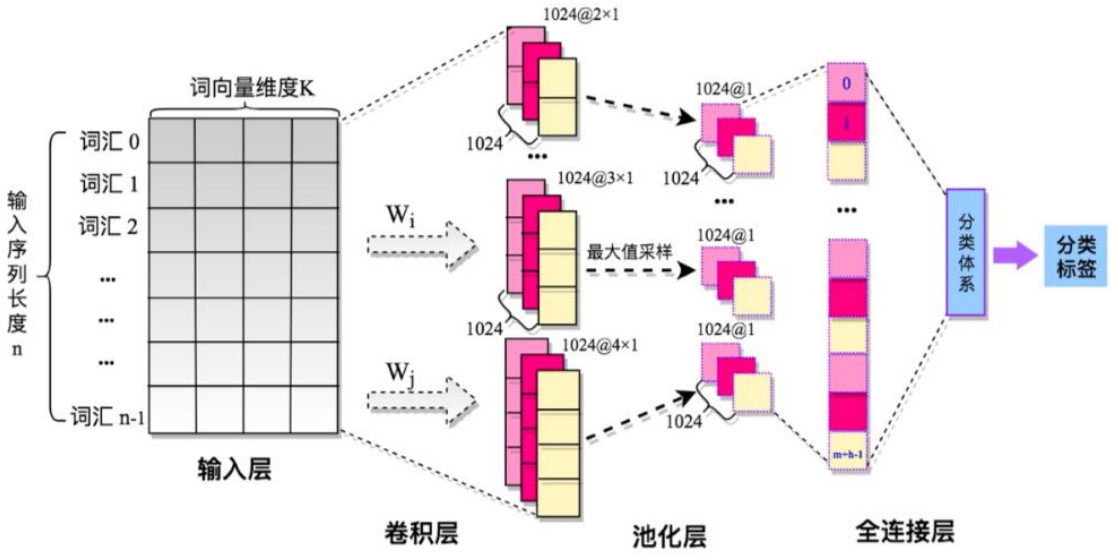


图 TextCNN 结构图

2.2 BiLSTM

长短期记忆网络通过对遗忘门的使用，它在适应了时序数据、具有序列性特征的数据，能够在提取与前文相关的后文特征的同时，能够适当的对前文信息进行记忆和遗忘，非常适合长序列文本的处理任务。而 BiLstm 作为 Lstm 的改进，同时由两个 Lstm 组成，这两个 Lstm 网络分别负责提取前向与后向的信息，提取之后对其进行组合，这样在提取特征的同时捕捉了上下文双向的语义依赖。

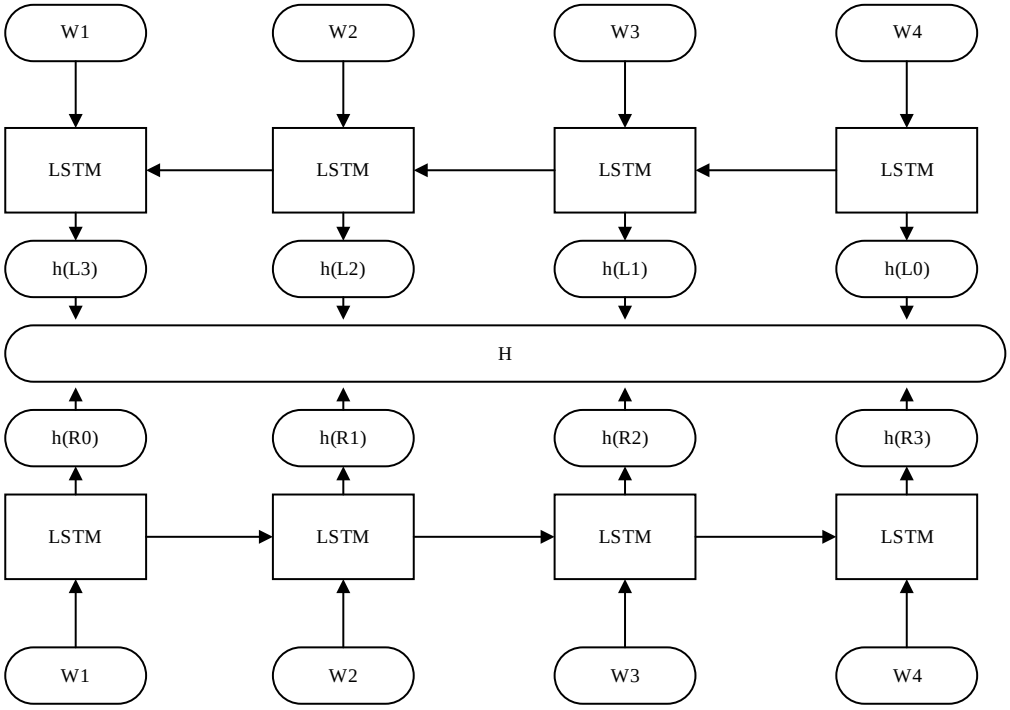


图 BI-LSTM 结构图

本设计采用的 BiLstm 原理如图所示，设其所接受的文本向量 $W = [w_1, w_2, .., w_n]$

其中 w_n 是文本中第 n 个元素的嵌入，输入模型后，前向与后向 LSTM 分别进行计算，则有 t 时刻前向 LSTM 输出为，后向 LSTM 输出为：

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(\omega_t, \overrightarrow{h_{t-1}})$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(\omega_t, \overleftarrow{h_{t+1}})$$

BiLstm 在 t 时刻输出为 $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$，而对于文本向量 W 的模型输出则为集合 $H = [h_1, h_2, \ldots, h_n]$，设模型隐藏节点个数为 u，则输出维数为 $n \times 2u$。

2.3 BERT

BERT 是由多层编码器组成的预训练模型，它可以从单词的两边考虑上下文。作为一种预训练模型，BERT 在大量数据集上进行了预训练，在预训练好的 BERT 模型后面根据特定任务加上相应的网络，可以完成 NLP 的下游任务，比如文本分类、机器翻译等

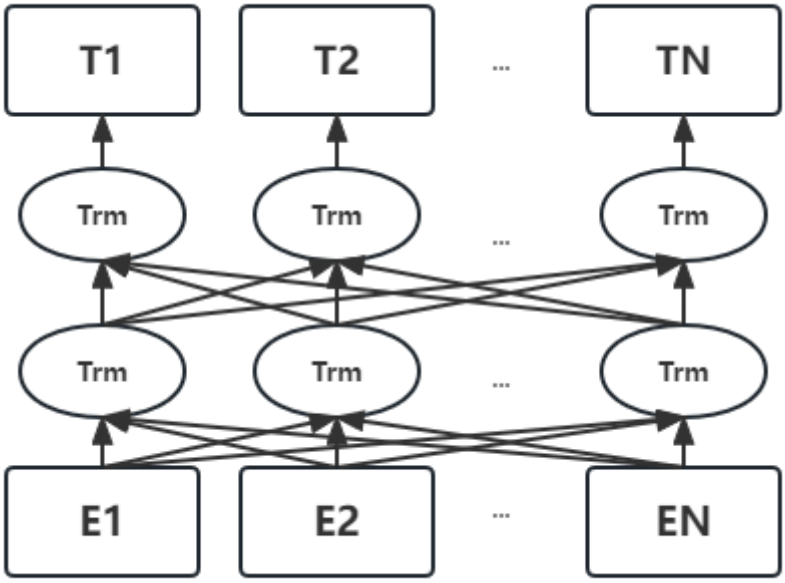


图 Bert 结构

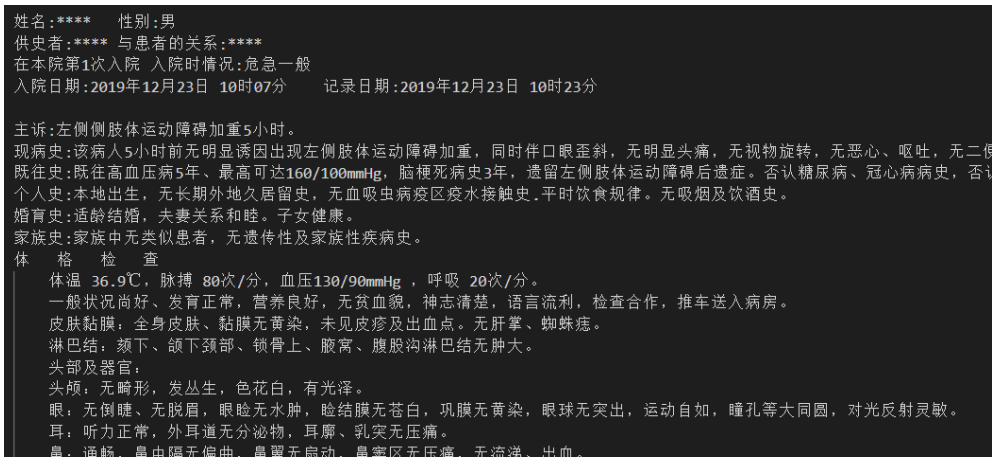
其中一个 transformer 的 encoder 单元由一个 multi-head-Attention、Layer Normalization 、feedforword 、 Layer Normalization 叠加产生，BERT 的每一层由一个这样的 encoder 单元构成。

三、实验过程

3.1 数据集构建

本实验的病历文本主要为高度结构化的、长度不定的、特征极为明确的、包含信息相似的语料，在前人的研究中，通过对结构化输入方式与首尾拼接输入方式的效果进行对比，发现二者性能差距仅为 1%，因此算法采用了首尾拼接的文本处理方法，将医生输入的医疗信息处理为字符串语料。

其中，针对主诉、现病史、既往史、个人史等进行研究，将原始数据中其他部分进行剔除，并进行规范化，最终得到规范化数据集。


姓名:****　性别:男
供史者:**** 与患者的关系:****
在本院第1次入院 入院时情况:危急一般
入院日期:2019年12月23日 10时07分　　记录日期:2019年12月23日 10时23分

主诉:左侧侧肢体运动障碍加重5小时。
现病史:该病人5小时前无明显诱因出现左侧肢体运动障碍加重，同时伴口眼歪斜，无明显头痛，无视物旋转，无恶心、呕吐，无二
既往史:既往高血压病5年、最高可达160/100mmHg，脑梗死病史3年，遗留左侧肢体运动障碍后遗症。否认糖尿病、冠心病病史，否
个人史:本地出生，无长期外地久居留史，无血吸虫病疫区疫水接触史.平时饮食规律。无吸烟及饮酒史。
婚育史:适龄结婚，夫妻关系和睦。子女健康。
家族史:家族中无类似患者，无遗传性及家族性疾病史。
体　格　检　查
体温 36.9℃，脉搏 80次/分，血压130/90mmHg ，呼吸 20次/分。
一般状况尚好、发育正常，营养良好，无贫血貌，神志清楚，语言流利，检查合作，推车送入病房。
皮肤黏膜：全身皮肤、黏膜无黄染，未见皮疹及出血点。无肝掌、蜘蛛痣。
淋巴结：颏下、颌下颈部、锁骨上、腋窝、腹股沟淋巴结无肿大。
头部及器官：
头颅：无畸形，发丛生，色花白，有光泽。
眼：无倒睫、无脱眉，眼睑无水肿，睑结膜无苍白，巩膜无黄染，眼球无突出，运动自如，瞳孔等大同圆，对光反射灵敏。
耳：听力正常，外耳道无分泌物，耳廓、乳突无压痛。
鼻：通畅，鼻中隔无偏曲，鼻翼无扇动，鼻窦区无压痛，无流涕、出血。

图 医疗文本示例

1912748	主诉：头晕头痛1天 现病史:该病人1天前无明显诱因出现头晕头痛，恶心、无呕吐，无肢体麻木，无肢体运动障碍，无饮	0
1912883	主诉:左侧侧肢体运动障碍加重5小时。 现病史:该病人5小时前无明显诱因出现左侧肢体运动障碍加重，同时伴口眼歪斜	0
1912897	主诉:左侧肢体运动障碍加重3天。 现病史:该病人3天前无明显诱因出现左侧肢体运动障碍加重，无头晕头痛，无视物旋	0
1912911	主诉:头晕、头痛伴心悸气短3天。 现病史:该病人3天前无明显诱因出现头晕、头痛，同时伴心悸气短，恶心，无呕吐，	0
1912920	主诉:左侧肢体无力、麻木感5小时。 现病史:该患于5小时前晨起时发现左侧肢体无力、麻木感，言语略笨拙，头晕，恶	0
1912971	主诉:间断头晕、头痛10余天。 现病史:该患10余天前无明显诱因出现头晕、头痛，头痛位于枕部，呈阵发性，每次持续	0
1912987	主诉:阵发性言语障碍及右侧肢体活动不灵6月余。 现病史:该患6月前无明显诱因出现阵发性言语障碍及右侧肢体活动不	0
1912991	主诉:右侧肢体麻木伴运动障碍2天。 现病史:该病人2天前因"右侧肢体麻木伴运动障碍2小时"来我院查头部磁共振平扫，	0

图 规范化文本数据

综上，最终集如表所示

数据集表

疾病种类	训练集	测试集
缺血性卒中	1760	440
非缺血性卒中	1780	500
总计	3540	940

3.2 模型构建

3.2.1 实验流程

实验流程如图所示，首先构建数据集（3.1 节），然后使用构建的三种模型分别进行实验，并对实验过程进行对比分析。

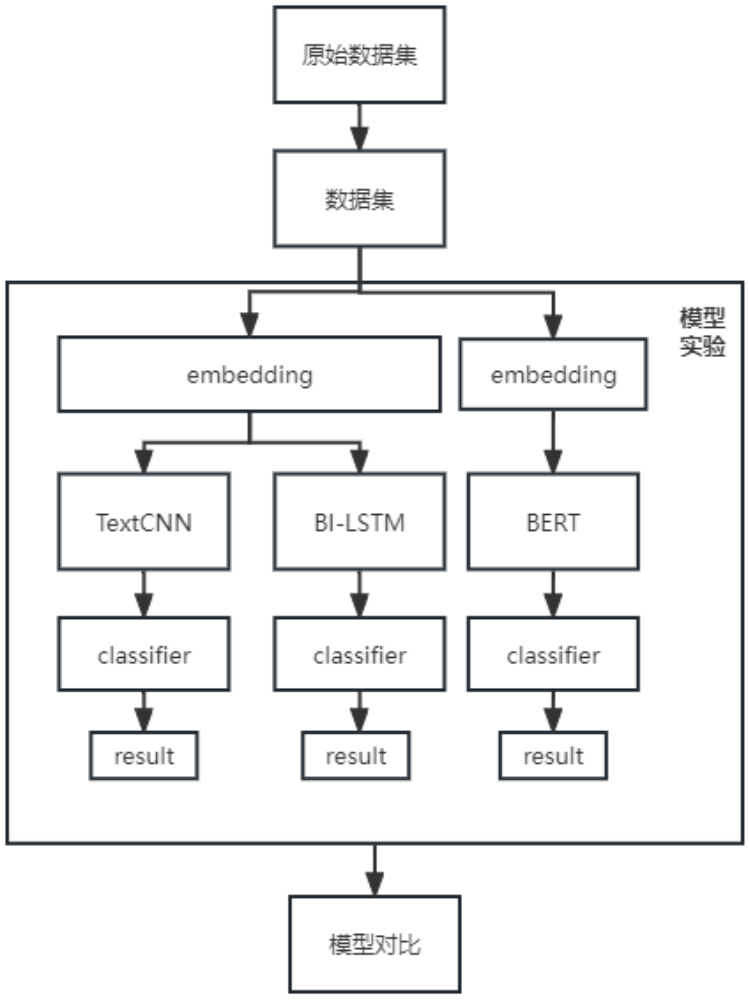


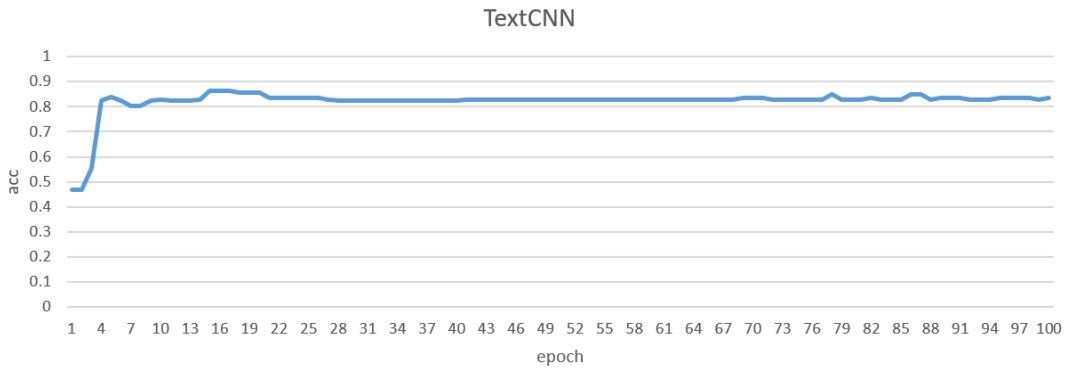
图 实验流程图

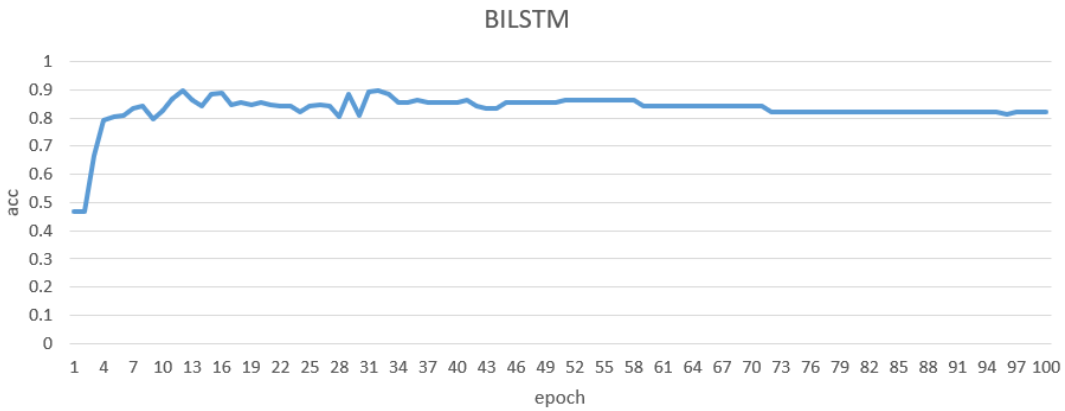
3.2.2 实验结果

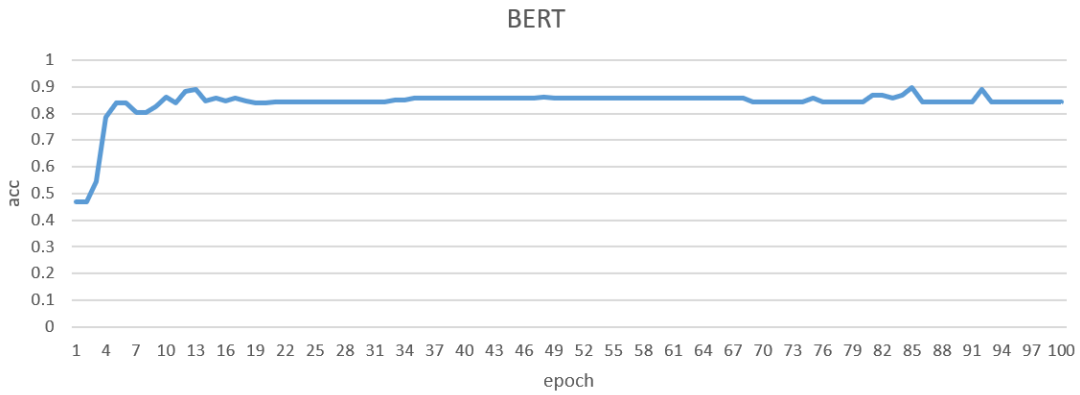
模型参数如表所示：

表 模型参数表

	参数名	参数值	参数意义
TextCNN	num_epochs	100	epoch 数
	batch_size	128	mini-batch 大小
	learning_rate	0.001	卷积核尺寸
	filter_sizes	(2, 3, 4)	每层隐藏节点数
	num_filters	256	卷积核数量(channels 数)
	参数名	参数值	参数意义
BISTM	num_epochs	100	epoch 数
	batch_size	128	mini-batch 大小
	learning_rate	0.001	学习率
	hidden_size	128	lstm 隐藏层
	num_layers	2	lstm 层数
	参数名	参数值	参数意义
BERT	num_epochs	100	epoch 数
	batch_size	64	mini-batch 大小
	learning_rate	0.00005	学习率
	hidden_size	768	每层隐藏节点数







如上图所示，所有模型均迭代 100epoch，基于 TextCnn 的模型最大 ACC 出现在第 17 轮，值为 0.8617；基于 BILSTM 的最大 ACC 出现在第 15 轮，值为 0.8882；基于 BERT 的模型最大 ACC 出现在第 85 轮，值为 0.8982。另外 TextCNN，BILSTM 模型均出现了较为严重的过拟合。而 Bert 模型在过拟合方面表现较好。