基于BERT融合多特征的临床试验筛选短文本分类

报告人：丁泽源

2019/11/24

目录

任务介绍

任务描述

临床试验是指通过人体志愿者进行的科学研究，筛选标准是临床试验负责人拟定的鉴定受试者是否满足某项临床试验的主要指标。通过自然语言处理和机器学习的方法对临床试验筛选标准自动解析，将一系列中文临床试验筛选标准的描述句子分到事先定义好的44种筛选标准类别。

ID	输入(筛选标准)	输出(类别)
S1	年龄>80岁	Age
S2	近期颅内或椎管内手术史	Therapy or Surgery
S3	血糖<2.7mmol/L	Laboratory Examinations

任务介绍

评价指标

本任务的评价指标包括宏观准确率(Macro Precision)，宏观召回率(Macro Recall)，Average F1值。最终排名以Average F1值为基准。假设我们有n个类别，C1, … …, Ci, … …, Cn。

准确率Pi = 正确预测为类别Ci的样本个数 / 预测为Ci类的样本个数

召回率Ri = 正确预测为类别Ci的样本个数 / 真实的Ci类的样本个数

$$\text{Average F1} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \frac{2*Pi*Ri}{Pi+Ri}$$

目录

数据预处理

- **数据**
  - 训练集数据22,962条，验证集数据7,682条，测试集数据7,697条
- **预处理**
  - 去掉停用词与标点符号
  - 去除特殊符号
  - 去除长串数字

设计特征

- **句法特征**
- **本次评测所用的依存句法分析器为HanLP的基于神经网络的高性能依存句法分析器。主要用于提取句子的主谓宾。**
  - 短文本长度很短，不能提供足够的信息和知识依赖的问题.

    原句： “过去个月内出现呼吸暂停的患儿”

    提取的主谓宾： “出现呼吸暂停患儿”，

    处理后的句子： “过去个月内出现呼吸暂停的患儿；出现呼吸暂停患儿”。

设计特征

- **句法特征**
- **本次评测所用的依存句法分析器为HanLP的基于神经网络的高性能依存句法分析器。主要用于提取句子的主谓宾。**
  - 某些类别句子很少的问题。

    原句：”宫颈脱落上皮细胞样本不符合被考核试剂和对比试剂的样本要求”，

    类别为：Receptor Status

    提取的主谓宾：“宫颈脱落细胞样本符合对比试剂要求”，作为类别Receptor Status的新句子

设计特征

- **词性特征**
- **本次评测中使用的词性标注工具是HanLP的感知机词性标注，将标注的词取BIEOS的标记策略。**

不	能	理	解	量	表	内	容	无
S-d	S-v	B-v	E-v	B-n	E-n	B-n	E-n	B-v
法	进	行	疗	效	评	价	者	；
E-v	B-v	E-v	B-n	E-n	B-n	I-n	E-n	S-w

设计特征

- **关键词特征**
- **本次评测使用的关键词提取工具是HanLP用TextRank算法实现的关键词提取工具包。**
  - 短文本长度很短，不能提供足够的信息和知识依赖的问题.

    原句：“3、存在脑外伤、中毒等因素；”

    提取关键词：“因素、存在、脑外伤”，

    处理后的句子：“3、存在脑外伤、中毒等因素；因素、存在、脑外伤”

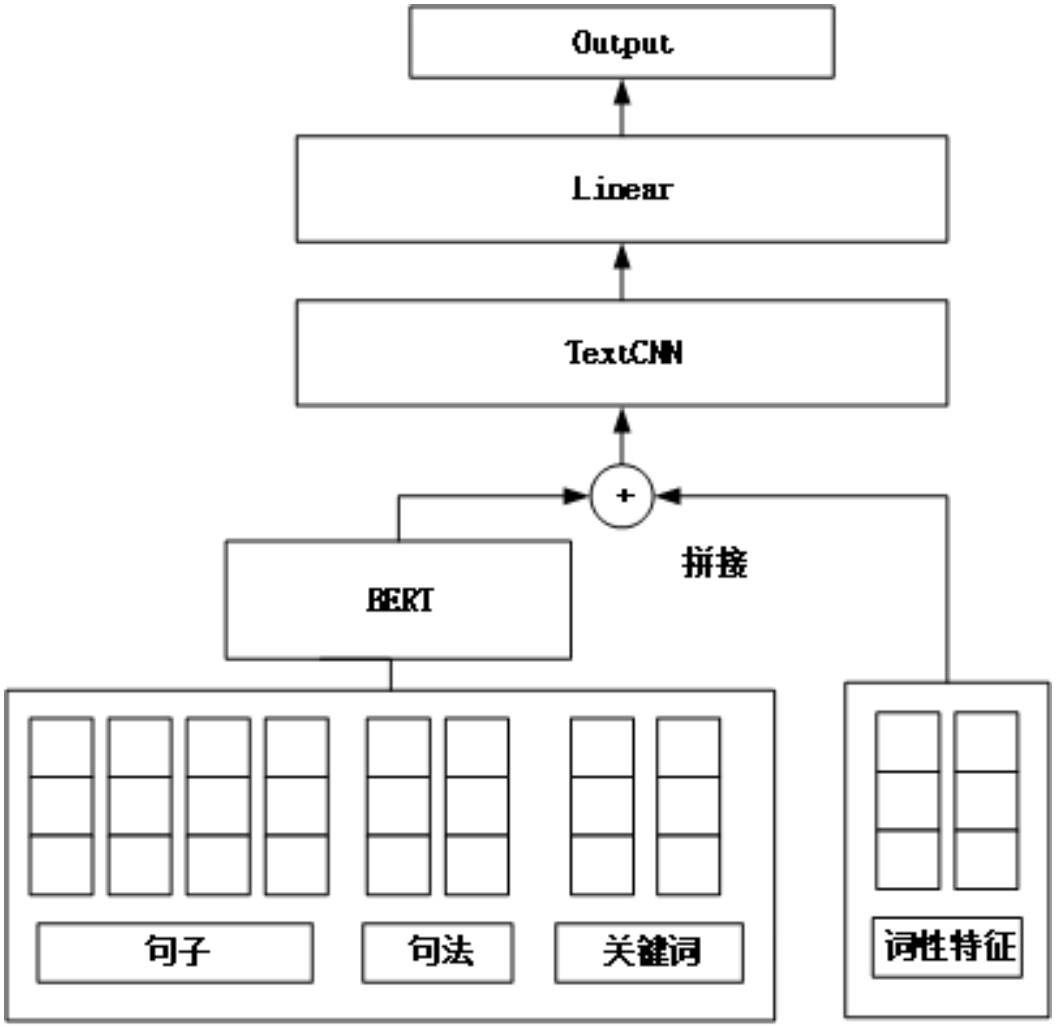
目录

模型方法

- **模型**
  - BERT+CNN
  - ERNIE+CNN、ERNIE+Attention
  - RoBERTa +CNN、RoBERTa + Attention
- **实验思路**
  - 重新划分数据集为4:1
  - 在上述模型上不断的尝试不同的特征

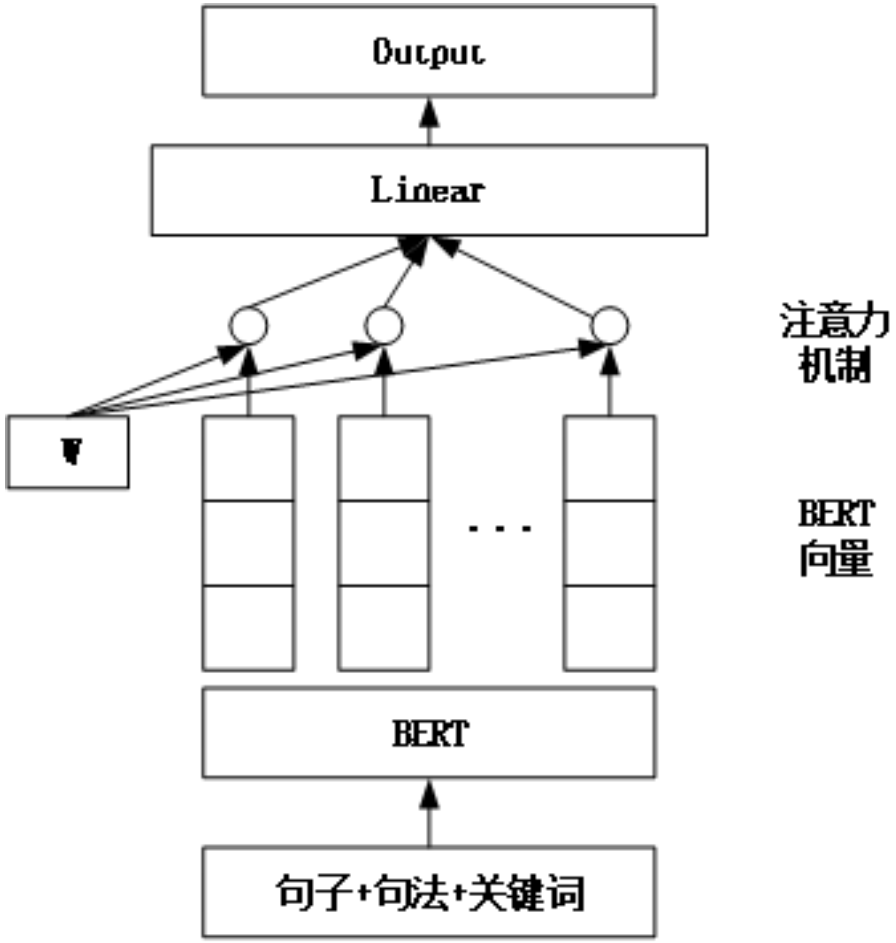
模型方法

- **BERT+CNN**



模型方法

- **BERT+Attention**



目录

实验结果

- **实验**
  - 数据集
    - **重新划分数据集，训练集：验证集 = 4 : 1**
  - Bert+TextCNN超参设置

超参	描述	值	取值范围
Sen_len	句子的长度	100	[50, 75, 100, 150]
Num_filter	卷积核数量	600	[150, 300, 600]
Filter_size	卷积核大小	3、4、5	\
Dropout	Dropout概率	0.5	[0.2, 0.5, 0.8]
Learning_rate	学习率	5e-5	[1e-5, 3e-5, 5e-5]

Results

- **模型结果对比**

模型	准确率(%)	召回率(%)	F1值(%)
ERNIE+CNN	79.58	80.29	79.00
ERNIE+CNN+句法特征	80.66	80.26	79.88
ERNIE+CNN+关键词	79.28	81.68	79.65
RoBERTa+ATT	78.62	80.73	79.06
RoBERTa+ATT+句法特征	81.66	80.04	79.82
RoBERTa+ATT+关键词	81.38	78.43	79.61
模型集成	**81.28**	**81.20**	**80.75**

总结

- 本模型在CHIP 2019临床试验筛选标准短文本分类评测中取得了80.75%排名第三的成绩。
- 我们通过对结果的分析，发现由于语料的规模限制，导致数量很少、特征不明显的类别的句子识别仍然不是很好.
- 我们未来的工作将尝试一下小样本学习和零样本学习，侧重于如何在给定训练集某个类别数量很少的情况下，仍然能很精确的将句子分类到正确的类别

谢谢大家!