基于TextCNN的医疗语义分诊技术

胡瑞文　李伟超　朱昔羽　徐硕
（华中科技大学,湖北　武汉　430074）

[摘　要]由于患者对大型三甲医院的信赖，普通医院诊疗服务较少等因素，各地大型三甲医院的门诊就诊量在近年来逐年上升，并大多突破一万人\天。当患者在门诊进行就诊时第一个环节便是候诊，通常来说候诊需要患者等待很长的时间，正是如此候诊从一定意义上使患者形成了焦虑、烦躁的情绪，很容易引发护患纠纷，进而导致门诊护理工作的工作量的下降。为了解决此问题，我们使用自然语言处理技术，采用能够对文本进行分类的TextCNN模型。在基于一定量的训练后，该模型能够对于病情描述进行语义分诊，并且达到了0.74的准确率，极大提高了病人就诊效率，缓解了高峰期时医院的压力。

[关键词]中文医疗对话；语义分类；TextCNN模型；Bi-LSTM模型
【DOI】10.12252/j.issn.2096-6288.2021.08.1094

一、引言

中文医疗对话语义分类技术广泛用于实体医疗体系，互联网医疗体系和健康咨询行业。在缺乏人工分诊台的环境下，利用语义分类技术可以很大程度上减少挂错号产生的时间和资源的消耗，或者是互联网医疗中错误分配科室所带来的时间和医疗资源的浪费。语义分类模型使患者被正确分配到正确科室的准确率对患者正确就医有重要意义。采用合适的语义分类模型对于提高患者就医准确率有重要意义。

近年来，自然语言处理在医疗语义分类模型上应用非常广法，现主要应用于医疗文本语义分类的自然语言处理方法主要包括循环神经网络(RNN)、长短期记忆神经网络(LSTM)、卷积神经网络(CNN)，Transformer等。RNN 能够充分学习文本数据,所以能处理大量的数据集,但RNN在优化过程中存在着梯度消失的问题。为了解决该问题,Liu等[1]建立LSTM文本分类模型，LSTM应用到文本分类中来,但该方法结果复杂因此计算需要大量的时间和空间。Kim[2]提出运用CNN 进行文本分类的一种模型。随着CNN在文本分类领域的不断发展,该方法能够很好地降低文本特征提取的难度。TextCNN模型结构简单，训练速度快，在处理小样本中文对话数据上有显著优势。本文主要是用了双向LSTM，LSTM，RNN和TextCNN去对中文医疗对话数据集进行语义分类，研究表明在处理该问题上使用TextCNN所得到的结果准确度最高。

二、研究现状分析

语义分诊是基于语言分类、语义分类、自然语言识别、自动文本分类等技术学科创新而来的医疗会诊创新技术。该系统通过对患者、医生所述状况进行初步分类，以此达到提高门诊精确度、提高分诊速度的目的。

传统方法中，文本分类任务作为一种特殊的专家系统而出现。具体而言,就是先由专家根据自己的知识，制定很多用于分类文本的规则。用这些规则去计算文本应该属于的类别。这样方法的缺点是显然的，首先在于规则的制定是非常困难和难以检验的,从某种意义上说，这甚至比让专家自己去分类文本代价还大，这很显然是无法满足现有需求的。另一种现在普遍使用的方法是学习的方法。在准备输入学习机器的向量时会结合到自然语言处理的方法，把文本表示成向量。一般而言，此方法的第一步是将文本表示为下一步分类计算所需要的向量形式，第二步则是对这些向量进行分类，这是一个典型的模式识别问题，故可以采用多种机器学习方法处理此问题。

虽然外国英文文本实体识别已经相对成熟，但国内中文文本实体识别依然处于起步阶段。这是因为汉语中缺乏字形变化信息，相对难以转变为可识别向量。由于中文人名、地名和机构名的内部组成规律和上下文知识不同，这种方案针对某一类名词的特点,提出了有效的识别方法。但是这种方案忽视了不同种类命名实体间的歧义问题[3-6]。

三、数据集

中文医疗对话数据集https://github.com/Toyhom/Chinese-medical-dialogue-data/其数据格式如下图：

department	title	ask	answer
心血管科	高血压患者能吃党参吗?	我有高血压这两天女婿来的时候给我拿了些党参泡水喝，您好高血压可以吃党参吗?	高血压病人可以口服党参的。党参有降血脂，降血压的作用，可以彻底消除血液中的垃圾，从而对冠心病以及心血管疾病的患者都有一定的稳定预防工作作用，因此平时口服党参能远离三高的危害。另外党参除了益气养血，降低中枢神经作用，调整消化系统功能，健脾补肺的功能。感谢您的进行咨询，期望我的解释对你有所帮助。
消化科	哪家医院能治胃反流	烧心，打隔，咳嗽低烧，以有4年多	建议你用奥美拉唑同时，加用吗丁啉或莫沙必利或援生力维，另外还可以加用达喜片

图1 中文医疗对话数据集数据格式

我们只保留其中的 department和ask项（department为科室，ask为患者询问的问题描述），使用 pandas进行数据处理，选取其中数目较多的前10项科室，打乱数据后，划分成训练集和测试集，比例为4：1。由于采用 word2vec模型，故需要进行分词处理，先完成训练集和测试集的预处理，去除空项和无意义的项目，使用 jieba 库（Python中文分词组件Jieba）对 ask 进行分词，并储存在数据文件夹中，方便训练时使用。

四、实验方法

（一）TextCNN和Bi-LSTM

1.word2vec。自然语言中词是表义的基本单元。用词向量来表示词，也可被认为是词的特征向量或表征，把词映射为实数域向量的技术也叫词嵌入（word2vec）。word2vec有连续词袋模型(Continuous bag-of-words，CBOW)和Skip—Gram两种模型。word2vec 能够将文本词语转化为向量空间中的向量，而向量的相似度可以表示文本语义的相似度[7]。

2.TextCNN。TextCNN即是由Yoon Kim提出的一种用于处理文本分类问题 的卷积神经网络模型。首先进行清洗数据工作，然后生成词向量，然后送入TextCNN网络中进行训练[2]。

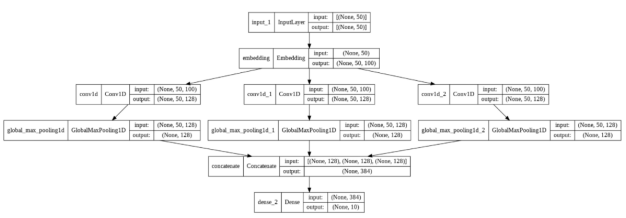

图2 TextCNN网络结构图

3. Bi-LSTM。连续词袋模型（CBOW）：连续词袋模型假设基于某中心词在文本序列前后的背景词来生成该中心词，给定一个长度为 T 的文本序列，设时间步 t 的词为 $w(t)$，背景窗口大小为 m 连续词袋模型的似然函数是由背景词生成任一中心词的概率：

$$\prod_{t=1}^{T} P\left(w^{(t)} \middle| w^{(t-m)}, \ldots w^{(t-1)}, w^{(t+1)}, w^{(t+m)}\right)$$

等价于最小化以下损失函数：

$$-\sum_{t=1}^{T} \sum_{-m \le j \le m, j \ne 0} \log P\left(w^{(t+j)} \middle| w^{(t)}\right)$$

双向 LSTM（Bi-LSTM）不仅能利用到过去的信息，还能捕捉到后续的信息，比如在词性标注问题中，一个词的词性由上下文的词所决定，那么用双向LSTM 就可以利用好上下文的信息。双向 LSTM 由两个信息传递相反的LSTM循环层构成，其中第一层按时间顺序传递信息，第二层按时间逆序传递信息。双向卷积神经网络的隐藏层要保存两个值，A 参与正向计算，A' 参与反向计算[8]。最终的输出值 y 取决于 A 和 A'：

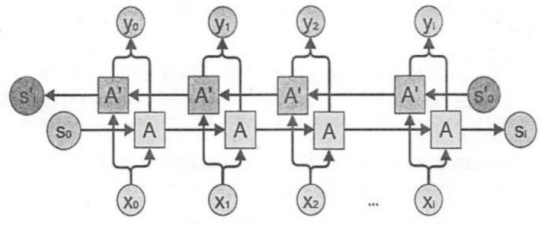

图3 Bi-LSTM网络结构图

（二）训练过程对比

1. Bi-LSTM与LSTM、RNN训练过程对比

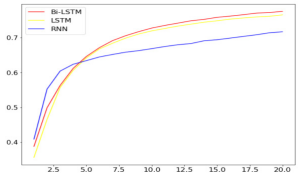

图4 Accuracy曲线

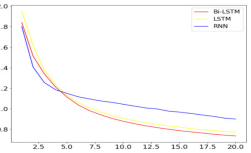

图5 Loss曲线

2. Bi-LSTM与TextCNN训练过程对比

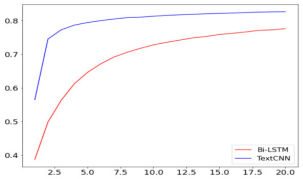

图6 Accuracy曲线

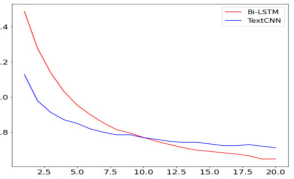

图7 Loss曲线

五、实验结果

表1 实验结果

	Bi-LSTM	LSTM	RNN	TextCNN
测试集准确率	0.70	0.71	0.61	0.74

综合图表来看，各模型的优劣：TextCNN > 双向 LSTM >= LSTM > RNN，TextCNN 虽然没有 RNN 这种序列依赖的结构，但是通过一维卷积和池化操作也可以捕捉到文本的局部特征。由于样本数量较少，简单的 TextCNN 模型反而效果最好。

六、讨论

（一）EDA分析

1. 用户问题长度分析。将问题内容长度绘制成直方图和箱形图，可看出大部分问题的长度都在 200 个字符以内。

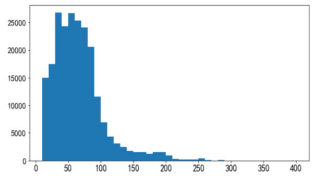

图8 问题长度-频数直方图

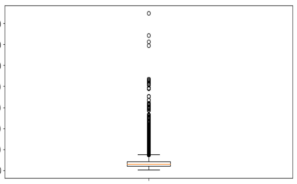

图9 问题长度箱线图

2. 类别分布分析。对数据集类别分布进行统计，即统计每类科室的样本个数，发现其分布不均匀。

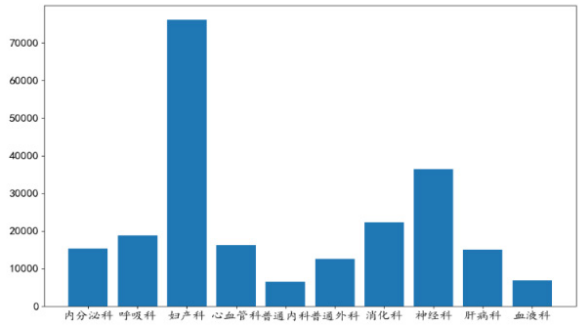

图10 科室样本个数直方图

（二）Bi-LSTM与TextCNN讨论

Bi-LSTM在中文医疗对话数据集上的效果不及TextCNN，可能因为样本数据量较少、样本分布不均匀以及样本本身存在的错误，同时Bi-LSTM中存在一定的资源浪费，而TextCNN本身网络结构简单，计算快，在小样本数据集上可能产生更好的效果。

参考文献：

[1]Liu P,Qiu X ,Huang X .Recurrent Neural Network for Text Classification with Multi-Task Learning[J]. AAAI Press, 2016.

[2]Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv,2014.

[3]王懿.基于自然语言处理和机器学习的文本分类及其应用研究[D].中国科学院研究生院（成都计算机应用研究所）,2006.

[4]史海峰.基于CRF的中文命名实体识别研究[D].苏州大学,2010.

[5]蔡慧苹.基于卷积神经网络的短文本分类方法研究[D].西南大学,2016.

[6]陈振宁.现代汉语量范畴语义模型初探[D].四川师范大学,2006.

[7]赵明,杜会芳,董翠翠,等.基于word2vec和LSTM的饮食健康文本分类研究[J]. 农业机械学报,2017,48(10):7.

[8]J.Chen and J. Lin, The Application of Bi-LSTM in Computerized Adaptive Test Research, 2020 International Conference on Big Data and Social Sciences (ICBDSS), 2020, pp. 33-36, doi: 10.1109/ICBDSS51270.2020.00015.