

**基于Attention机制的CNN-BLSTM在人岗匹配中的应用**

**姓 名： 高珩 刘潇炜 钟天奕**

**田原 杨嘉玟 叶贾谊 方奕丹**

**学 院**： **理学院数学系**

**学科/专业名称：数学与应用数学**

**2021年07月**

# 摘 要

随着网络招聘的发展，互联网为就业市场提供了海量的岗位信息和人才信息，但是同时也增加了准确获取所需信息的复杂度，岗位推荐系统也应运而生。

本文基于人岗匹配的实际机制和数据的具体格式，将简历中的非结构化长文本转化为词向量，再运用深度学习中的相应算法对词向量进行训练和学习。本文首先介绍了基于Attention机制的BLSTM算法、TextCNN算法，分析了两种算法的优势与局限性后，本文将两种算法结合生成了基于Attention机制的CNN-BLSTM算法。

最后，本文利用新致软件公司提供的数据集验证了该模型的可行性和精准性，发现精确率和召回率都有小幅度的提升。

**关键词**：岗位推荐；TextCNN；BLSTM；Attention机制

# ABSTRACT

With the development of online recruitment, the Internet provides a large amount of job information and talent information for the job market, but at the same time, it also increases the complexity of accurately obtaining the required information, so the job recommendation system comes into being.

Based on the actual mechanism of person-post matching and the specific format of data, this paper converts the unstructured long text in the resume into word vectors, and then uses the corresponding algorithm in deep learning to train and learn the word vectors. This paper first introduced the BLSTM algorithm and TextCNN algorithm based on the Attention mechanism. After analyzing the advantages and limitations of the two algorithms, this paper combined the two algorithms to generate the CNN-BLSTM algorithm based on the Attention mechanism.

Finally, this paper uses the data set provided by Xinzhi Software Company to verify the feasibility and accuracy of the model, and it is found that the accuracy rate and recall rate have been improved slightly.

**Key words:** Job Recommendation Problem; TextCNN；BLSTM；Attention Mechanism

**目 录**

[**摘 要 2**](#_Toc77270372)

[**ABSTRACT 3**](#_Toc77270373)

[**第1章 绪论 5**](#_Toc77270374)

[**1.1 引言 5**](#_Toc77270375)

[**1.2 本文研究主要内容 6**](#_Toc77270376)

[**1.3 本文研究意义 6**](#_Toc77270377)

[**第2章 模型建立 7**](#_Toc77270378)

[**2.1 背景知识 7**](#_Toc77270379)

[**2.1.1 基于Attention机制的BLSTM模型 7**](#_Toc77270380)

[**2.1.1.1 研究背景 7**](#_Toc77270381)

[**2.1.1.2 模型介绍 7**](#_Toc77270382)

[**2.1.2 TextCNN模型 9**](#_Toc77270383)

[**2.1.2.1 研究背景 9**](#_Toc77270384)

[**2.1.2.2 TextCNN的原理简介 9**](#_Toc77270385)

[**2.1.2.3 TextCNN的具体流程 10**](#_Toc77270386)

[**2.1.2.4 TextCNN的优势分析及改善想法 12**](#_Toc77270387)

[**2.2 基于Attention机制的CNN-BLSTM模型 12**](#_Toc77270388)

[**第3章 数值实验 13**](#_Toc77270389)

[**3.1 TextCNN 13**](#_Toc77270390)

[**3.2 基于Attention的CNN-BLSTM 14**](#_Toc77270391)

[**第4章 总结 14**](#_Toc77270392)

[**4.1 主要结论 14**](#_Toc77270393)

[**参 考 文 献 15**](#_Toc77270394)

1. 绪论
   1. 引言

近些年，我国面临较大经济下行压力，就业形势严峻，产业结构需要转型升级，劳动力市场需求发生重大改变，人才供给与岗位需求出现不对称。虽然互联网为就业提供了海量的岗位需求数据，但是同时也增加了准确获取所需信息的时间和复杂度。目前就业部门、各大招聘网站( BOSS直聘、智联招聘、拉勾网) 等为广大求职者提供了海量岗位招聘信息，业务系统除了支持传统的岗位检索，也支持根据求职者的简历信息进行岗位推荐。现有岗位推荐系统通过分析求职者的历史搜索记录，挖掘其求职意愿、需求和求职倾向，为他们提供准确、快速、个性化的岗位推荐服务。

近年来很多学者推出岗位推荐系统，能够为用工岗位与求职者搭建匹配推送平台，努力实现就业岗位精准对接。随着互联网的迅速发展，将岗位推荐算法应用于移动互联网、电子商务、新闻媒体、网站、社交软件等对外发布招聘岗位信息的网络发布台，将为求职者找工作提供更便捷的通道，特别是在公共就业服务领域的职业介绍业务中得到有效的应用。例如，尹传城等[1]针对高校毕业生，提出了一种综合在校历史信息的高校毕业生互惠就业推荐方法，该方法具有较高的推荐准确率和用户满意度，可以为高校毕业生就业提供有效指导。在人力资源应用方面，古振威［2］将推荐技术应用于人力资源领域，提出了基于隐语义模型与深度森林的人力资源推荐算法，挖掘用户潜在兴趣并推荐用户感兴趣的岗位，从而更好地进行人力资源推荐。于海棠［3］将知识地图、人岗匹配理论与人力资源配置方法相结合，首先根据知识地图构建方法构建人员知识地图与岗位知识地图，然后利用二分图匹配算法进行匹配，从而实现一个基于知识地图的人岗匹配模型。欧俊［4］通过对乡镇领导干部人岗匹配现状进行分析，综合运用行为事件访谈法、统计分析法、问卷调查法等研究方法，构建了乡镇领导干部胜任力模型; 然后，以构建的乡镇领导干部胜任力模型为基准，把人岗匹配相关理论应用到知岗、知人、匹配和培训各环节工作中，为乡镇领导干部人岗匹配的实现提出有针对性的对策和建议。

* 1. 本文研究主要内容

出于对深度学习算法的学习，以及为了提高岗位匹配模型的计算速度和准确性，本文以深度学习为切入点，尝试解决岗位匹配中的处理非结构化文本（主要是指简历）的问题，以及对如何提高模型的准确性进行探讨。本小组先从推荐系统的角度来处理该问题，尝试使用了知识图谱dkn模型和决策树模型，但是在处理非结构化长文本上遇到了困难，转变方向为能够直接将长文本转化为向量的bert模型，然后在新致软件的工作人员的建议下，决定围绕长文本处理展开研究，最后决定使用基于Attention机制的BLSTM模型和同样能将长文本转化为向量而且计算速度更快、精度更高的TextCNN模型来解决问题。

* 1. 本文研究意义

随着经济和互联网的发展，传统的线下招聘模式逐渐被第三方在线招聘取代，一些大型的第三方招聘平台每年能受到数千万份的电子简历，在不同求职者的根据不同思路不同模板撰写的简历中，简历的书写习惯不能保持统一，对文本块的排列顺序也是不固定的。此外，求职者保存简历为各种文件类型，例如txt，pdf和docx文档等。与传统的招聘模式相比，处理大量的电子简历更具有挑战性，如何更快更有效的从这些形式内容各不相同的简历中提取出有效信息，并对海量的简历进行初步筛选，成为了一个值得关注的问题。

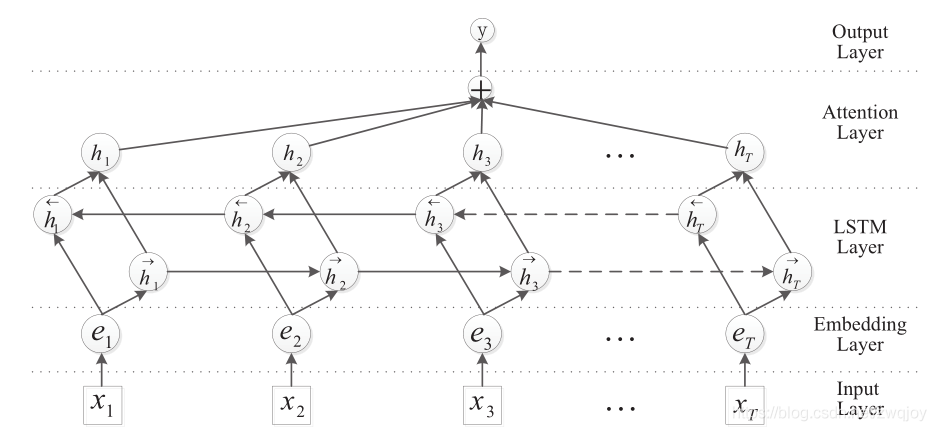
本文着重分析了这一问题，并提出了一些解决方式，使用深度学习的方法，将已有的“简历-岗位”数据进行了匹配，学习并选择了一些适合该问题的深度学习模型，然后让模型对这些已有的数据进行学习，最后得到了一些在处理该问题上效果比较好的模型，并对这些模型进行了一些改进。希望能对日后的岗位匹配问题提供一些思路和帮助。

1. 模型建立
   1. 背景知识
      1. 基于Attention机制的BLSTM模型
         1. 研究背景

关系抽取是自然语言处理中一个重要的任务，也即从自然语言文本中提取两个实体之间的语义关系。关系抽取属于信息抽取的一个部分。信息激增的时代，快速、准确获取关键信息的需求也日益激增，相比于传统的信息检索，信息抽取能够快速、准确提取出海量非结构化信息中的结构化知识，它也逐渐成为搜索引擎发展的方向。而关系抽取同命名实体识别、事件抽取等任务一起，都是信息抽取的一部分或者中间过程，可应用于结构化知识抽取、知识图谱构建、自动问答系统构建等。

关系抽取从本质上看是一个多分类问题，对于这样一个问题来说最重要的工作无非特征的提取和分类模型的选择。传统的方法中，大多数研究依赖于一些现有的词汇资源、NLP系统或一些手工提取的特征。这样的方法可能导致计算复杂度的增加，并且特征提取工作本身会耗费大量的时间和精力，特征提取质量的对于实验的结果也有很大的影响。因此，这篇论文从这一角度出发，提出一个基于Attention机制的双向LSTM神经网络模型进行关系抽取研究，Attention机制能够自动发现那些对于分类起到关键作用的词，使得这个模型可以从每个句子中捕获最重要的语义信息，它不依赖于任何外部的知识或者NLP系统。

* + - 1. 模型介绍



1. **输入层：将句子输入到模型中**

输入层输入的是以句子为单位的样本。

1. **Embedding层：将每个词映射到低维空间**

对于一个给定的包含T个词的句子。每一个词都是转换为一个实数向量。使用这个词向量矩阵，可以将每个词转化为其词向量的表示：, 其中，是一个大小为 的向量，在下表为处为1，其他位置为0。于是，句子将被转化为一个实数矩阵：，并传递给模型的下一层。

1. **LSTM层：使用双向LSTM从Embedding层获取高级特征**

为了解决循环神经网络中的梯度消失问题。主要思想是引入门机制，从而能够控制每一个LSTM单元保留的历史信息的程度以及记忆当前输入的信息，保留重要特征，丢弃不重要的特征，将上一个细胞状态同时引入到输入门、遗忘门以及新信息的计算当中。对于序列建模的任务来说，每一个时刻的未来信息和历史信息同等重要，标准的LSTM模型按其顺序并不能捕获未来的信息。

因而这篇论文采用了双向LSTM模型，在原有的正向LSTM网络层上增加一层反向的LSTM层

1. **Attention层：生成一个权重向量，通过与这个权重向量相乘，使每一次迭代中的词汇级的特征合并为句子级的特征**

由于LSTM获得每个时间点的输出信息之间的“影响程度”都是一样的，而在关系分类中，为了能够突出部分输出结果对分类的重要性，引入加权的思想，注意力机制本质上就是加权求和。

将LSTM层输入的向量集合表示为。其Attention层得到的权重矩阵由下面的方式得到

1. **输出层：将句子级的特征向量用于关系分类**

该方法巧妙地在双向LSTM模型中加入Attention机制，用于关系抽取任务，避免了传统的任务中复杂的特征工程，大大简化了实验过程并得到相当不错的结果，也为相关的研究提供了可操作性的思路。

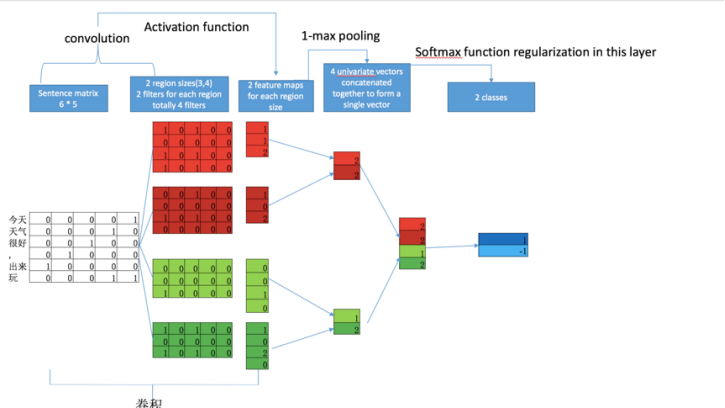
* + 1. TextCNN模型
       1. 研究背景

对于文本分类问题，首先会想到抽取文本的特征并将其转化为机器可以识别的固定维度的特征向量，然后基于抽取的特征向量直接训练出一个分类器，然而，当数据量很庞大时时，直接用特征向量进行训练的方式虽然理论可行，但是在训练速度上会有所欠缺，因此考虑寻找更加高效的方式对文本数据进行分类处理。

我们都知道CNN（卷积神经网络）是在传统神经网络的基础上多加入了卷积层和池化层来近似当作特征的提取与简化，它在图像处理领域尤其是模式分类方面由于网络结构的简易以及训练的高效性从而得到了广泛应用。基于CNN网络的这些优势，科学家们转而考虑改变CNN的输入方式，保留网络的整体结构从而使得CNN能够应用于除计算机视觉外的其他领域，由此诞生了能够对文本进行高效分类的TextCNN深度学习方法。

* + - 1. TextCNN的原理简介

前文提及TextCNN模型是基于CNN的网络架构而建立的，因此两者在隐藏层和输出层中所做的工作一致——输入数据经过卷积层和池化层的特征提取和简化后进入全连接层进行分类，流程图如下：

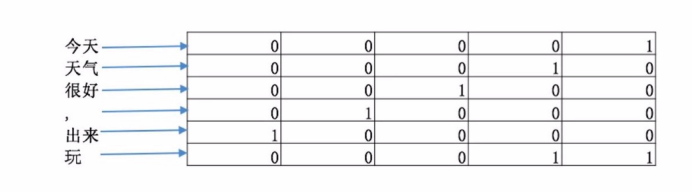


CNN的核心思想就是添加了卷积和池化操作进行局部特征的捕捉，TextCNN也主要沿用了这一思想，但因为输入数据的差别，不同于图像处理中以图像本身的像素点网格作为特征，对于文本来说，由于其为一维数据，即使经过word-embedding生成了二维向量，但是在卷积层中对词向量做从左到右滑动来进行卷积没有意义，假设流程图中"今天" 对应的向量[0,0,0,0,1], 若按窗口大小为 1\* 2 从左到右滑动得到[0,0],[0,0],[0,0],[0,1]这四个向量, 对应的都是"今天"这个词汇, 这种滑动对于句子本身的特征没有意义，因此对于文本数据，卷积窗口只进行从上往下的滑动且窗口的大小也要对应词向量的维度以保证能够覆盖单词的素所有信息，所以文本的局部特征可以视作由若干单词所组成的滑动窗口，这也是TextCNN与CNN的主要差别。剩余操作中两者的步骤完全一致甚至TextCNN的卷积层数可以更少使得网络训练速度更快。

* + - 1. TextCNN的具体流程

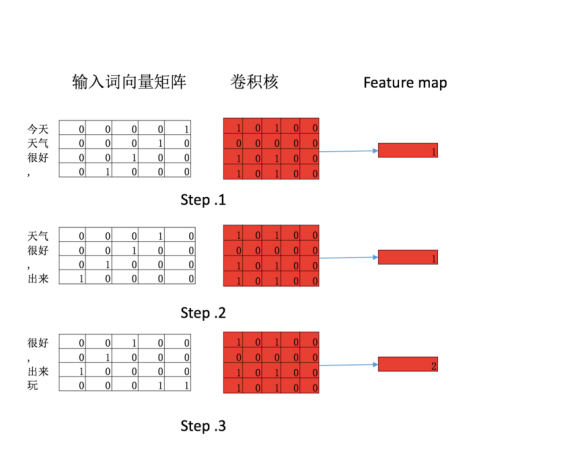
**1. Word-embedding 构建词向量**

相比CNN中直接将图像作为输入数据，由于TextCNN是针对文本做处理，计算机并不能识别出自然语言，因此需要在输入层中添加Word-embedding操作将文本数值化，比如可以将文本进行分词后，通过word2vec或者glove等embedding方式将每个词成映射成一个词向量,最后将词向量按行组合后形成矩阵作为能够代表该文本的输入数据，如：



**2. 卷积**

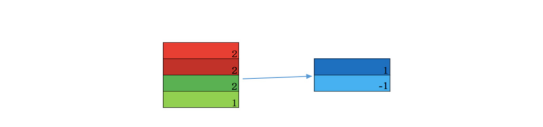
与CNN中卷积层的功能一致，都是对输入数据进行特征提取，其内部包含多个卷积核，组成卷积核的每个元素都对应一个权重系数和一个偏差量，类似于一个前馈神经网络的神经元，其中每个卷积单元的参数都是通过反向传播算法优化得到的。本质上，卷积核是一个个小的矩阵，卷积核会在原数据上移动，每次移动，做一次点乘，得到一个数字。该步骤会一直进行直至遍历完完整的原数据，在原理中我们也提及TextCNN中的卷积窗口只能够上下滑动，因此假设有4\*5的卷积核，对流程1中经过Word-embedding的文本作卷积操作我们有：



**3. 池化**

在TextCNN的池化层中多选择max-pooling操作来降低经过卷积层所提取的特征个数，只提取主要特征，这进一步地降低了数据和参数的量使得训练速度加快，在卷积层中我们对”今天天气很好,出来玩”进行了特征提取得到了的特征向量，再经过max-pooling操作后得到该文本的主要特征2。

**4. Softmax 分类**



如上图所示，假设左侧为四个文本在经过卷积池化操作后所得到的各个特征值的拼接，将其送入到softmax当中,得到属于各个类别的概率。

如果是训练模型的话，便会根据预测label以及实际label来计算损失函数, 计算出softmax 函数,max-pooling 函数, 激活函数以及卷积核函数四个函数当中参数需要更新的梯度,来依次更新这四个函数中的参数，完成一轮训练。

* + - 1. TextCNN的优势分析及改善想法

TextCNN的主要优势就在于参数数目少，计算量少，训练速度快，这也得益于其网络结构的简单，但这一优势仅仅证明了这一网络结构是能够更快更好处理NLP问题的一个方式，TextCNN最主要的成功是通过引入了训练后的词向量后依然能够达到很好的结果，即在分类的准确性上有显著的效果，但这绝非取决于网络结构的优势，最终我们依然是通过文本的特征来进行分类，TextCNN的网络结构只是将特征的提取以及参数的获取进行了简化，在保证不丢失主要特征的情况下达到更好的训练速度，而分类的效果很大程度上取决于初始输入层中对文本Word-embedding 的方式。事实上，我们最开始引入TextCNN方法的原因就是对文本词向量直接训练分类器的速度过慢，分类效果的实质还是要回归到对自然语言的特征转化上，这也是处理NLP问题的一个重要条件，因此，对于运用TextCNN模型去进行文本分类，我们考虑改变输入层中对文本转词向量的方式来进一步提升最终结果。

* 1. 基于Attention机制的CNN-BLSTM模型

我们的CNN-BLSTM模型用预训练好的Glove模型对简历和岗位要求的文本进行词嵌入，然后利用带有Attention层的双向LSTM进一步提取语义信息，最后通过全连接层进行二分类得到最终输出（0或1）。不过，由于循环神经网络对于序列信息的“遗忘”将会导致序列数据不能有效在模型内部对信息依赖进行建模，进而导致模型性能瓶颈的出现。那么，通过在循环神经网络前增加CNN（卷积神经网络）可以有效减缓循环神经网络对上下文信息的依赖性，从而对模型的性能进行小幅度的优化。

具体的模型框架如下图所示。

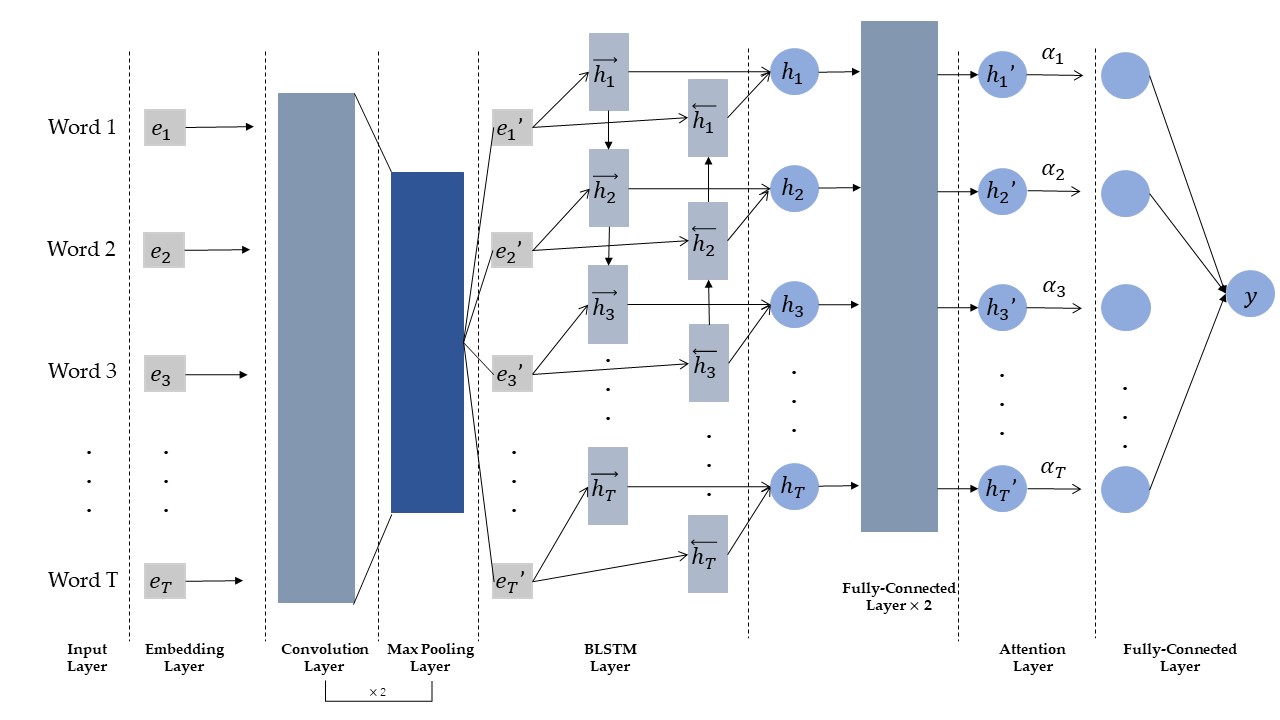


图 1 基于Attention机制的CNN-BLSTM神经网络原理示意图

1. 数值实验
   1. TextCNN

基于上述理论，本文在Keras框架下构建了3层卷积神经网络测试TextCNN模型性能。经数据组处理后，共获得4573条匹配数据进行二分类训练，其中0表示不通过初审，1表示通过，并以8:2为比例划分训练集（包含了验证集）和测试集。

本文基于百度百科的中文词向量库（300d）[[1]](#footnote-1)，在选取不同文本最大序列长度（神经网络输入序列）的情况下，使用CPU进行训练1轮、2轮后分别输出在训练集、测试机下的精确率和召回率。TextCNN的所有训练均在Intel Xeon Gold 6132（28C 56T）、64GB RAM的平台上进行。

表 1 TextCNN在不同最大序列长度下的精度

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 最大序列  长度 | 训练轮数 | 精确率  (验证集) | 召回率  (验证集) | 精确率  (测试集) | 召回率  (测试集) |
|  | 2 | 85.18% | 85.18% | 83.59% | 83.59% |
|  | 2 | 84.13% | 84.13% | 83.26% | 83.26% |

* 1. 基于Attention的CNN-BLSTM

在3.1节工作的基础上，我们在卷积层和池化层后又添加了双向LSTM层和Attention层，最终训练的结果如下表所示。

表 2 基于Attention的CNN-BLSTM在不同最大序列长度下的精度

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 最大序列  长度 | 训练轮数 | 精确率  (验证集) | 召回率  (验证集) | 精确率  (测试集) | 召回率  (测试集) |
|  | 2 | 85.34% | 85.34% | 83.96% | 83.96% |
|  | 2 | 84.35% | 84.35% | 84.20% | 84.20% |

1. 总结
   1. 主要结论

可以看出，在加上双向LSTM层和Attention层后，模型的分类精度和召回率都有了一定的提升。验证集上的精确率和召回率均上升约0.2%；测试集上的精确率和召回率也有超过0.4%的提升。

同时可以发现随着最大序列长度的增加，虽然验证集上的精确率和召回率浮动不大，但测试集上的精确率和召回率是有相对大的上升的。

整体来看，精确率和召回率提升的程度非常小，猜想可能是由于数据集的大小和数据质量方面的问题。所以，在项目后期，本项目组会进一步考虑对数据集进行进一步地处理，以提升数据的质量。

# 参 考 文 献

[1] 尹传城,王洪国,丁艳辉.一种基于在校历史信息的就业推荐算法[J].计算机与数字工程,2015,43(10):1742-1745.

[2] 古振威.基于隐语义模型与深度森林的人力资源推荐算法[D].广州:华南理工大学,2018．

[3] 于海棠.基于知识管理的高校人力资源配置[J].研究与发展管理,2006(1):122-127,135．

[4] 欧俊.基于胜任力的山区县乡镇领导干部人岗匹配研究[D].广州:华南理工大学,2018．

1. https://github.com/Embedding/Chinese-Word-Vectors [↑](#footnote-ref-1)