

文章编号: 1003-0077(2023)02-0158-11

基于神经网络模型的汉语文本难度分级

李文彪¹, 吴云芳²

(1. 北京大学 软件与微电子学院, 北京 102600;

2. 北京大学 计算机学院, 北京 100871)

摘要: 文本难度分级是自然语言处理在教育领域的一个基础性研究课题, 用于自动判定一篇文章的阅读难度。该文基于深度神经网络模型对汉语文本阅读难度进行了探索, 提出了一种 CNN+LSTM 的难度分级模型, 并结合分级语料的特点采用了变长卷积层和块结构。在教材测试集和人工构建的测试集上进行了详细的实验分析, 该文的神经网络模型超越了传统机器学习方法和主流神经网络方法, 在根据学段划分的 5 级数据上分级系统的正确率达到了 75.4%。

关键词: 可读性评估; 特征提取; 深度学习

中图分类号: TP391

文献标识码: A

Chinese Readability Assessment Based on Deep Neural Networks

LI Wenbiao¹, WU Yunfang²

(1. School of Software and Microelectronics, Peking University, Beijing 102600, China;

2. School of Computer Science, Peking University, Beijing 100871, China)

Abstract: Readability assessment is to automatically determine the reading difficulty of a given document. Focusing on Chinese readability assessment, this paper proposes a CNN + LSTM difficulty classification model with the variable-length convolutional layer and block structure. Extensive experiments on school textbooks and a manual-constructed test set show that the proposed method achieves 75.4% accuracy on 5-level difficulty prediction, which is superior to the existing models.

Keywords: readability assessment; feature extraction; deep learning

0 引言

文本难度分级 (Text Difficulty Classification), 也称为可读性评估 (Readability Assessment), 是教育教学领域面临的一项基础研究课题。其任务是, 对给定文本进行词法、句法、语义等方面的分析, 判断其难度适合什么水平的读者。近年来, 阅读难度分级引起了产业界广泛的兴趣, 应用领域从早期的出版教育扩展到信息检索、个性化推荐、文本摘要、机器翻译等研究领域。目前, 西方国家已形成比较成熟的英文分级阅读体系, 例如, 著名的蓝思分级阅读测评体系 (Lexile Framework), 全球 180 多个国家、450 家以上的出版社, 数千种期刊均采用了蓝思难

度分级, 美国超过一半的州几乎所有重要的标准化考试都用蓝思分值来报告学生的阅读和写作分数。

汉语的复杂性大大提高了文本可读性评价的难度, 相关研究有限, 更缺少像蓝思那样广泛应用的评级体系。本文基于神经网络模型对汉语文本的难度分级进行了研究。在主流神经网络模型中, 卷积神经网络 CNN 能够抽取局部 N -gram 特征却无法捕捉长距离依赖关系; 循环神经网络 RNN 可以处理序列结构但在分类任务上会由于关键词的不同位置而影响实验结果。结合两种模型的优点, 本文提出了 CNN+LSTM 的难度分级模型, 通过逐渐增加特征提取的程度, 符合人在阅读过程中对文章从局部到整体的理解过程。本文搜集了教材语料库作为实

收稿日期: 2021-06-15 定稿日期: 2021-09-18

基金项目: 国家自然科学基金 (62076008, 61936012)

验数据,同时人工构建了一个汉语文本分级评测语料,基于两个数据集对不同神经网络模型进行了详细的实验。结果表明,本文模型在五分类和八分类数据上的性能均超过了传统的机器学习模型和主流神经网络模型,达到了最优性能。

1 相关工作

1.1 可读性公式

可读性公式是将影响阅读难度的可量化的因素(词语级、句法级、语义级和篇章级等)综合起来而制定的评价文本难易程度的公式。可读性公式在英语文本难度评测中得到了广泛应用,如 Flesch-Kincaid 公式、Dale-Chall 公式、Gunning Fog 公式、SMOG 公式等。蓝思 Lexile^① 分级阅读测评体系主要利用图书的语义难度(词汇)和句法的复杂程度(句子长度)来衡量一个出版物的难易程度。汉语方面,郭望皓^[1]运用 CRITIC 加权赋值法建立句子难度与汉字、词汇难度之间的数量关系,采用多元线性回归拟合汉语文本易读性公式。左虹和朱勇^[2]在对教师问卷调查和对欧美留学生完形填空测试的基础上,通过多元线性回归的方法建立了一个针对中级欧美留学生的可读性公式。王蕾^[3]以初、中级日韩学习者汉语记叙性短文的阅读理解成绩为因变量,建立了一个衡量记叙性文本难易度的可读性公式。对于可读性公式,虽然比较简单直观,但是通过一些浅层语言特征构建的线性模型很难拟合不同的文本难度等级。我们曾将上述英语文本的难度公式应用于汉语文本难度分级,学段预测的准确率仅有 26% 左右。

1.2 文本分类模型

文本难度分级常常被视为一个文本分类问题,包括特征提取和分类器的构建两个方面。Schwarm 和 Ostendorf 等利用支持向量机(Support Vector Machine, SVM)并结合传统阅读难度测量、统计语言模型等特征,提出了一种阅读难度评估的方法^[4]。Pitler 和 Nenkova 考虑了各种语言因素,将词汇、句法和语篇特征结合起来评价文本可读性^[5]。Wang 和 Andersen 利用语法模板(Grammatical Templates)对日语文本可读性进行评估^[6]。汉语方面,孙刚^[7]选取各种特征建立线性回归模

型进行可读性预测,蒋智威^[8]选取各种特征实验了多种模型来评价文本阅读难度,吴思远等人^[9]对文本可读性研究做了较为全面的综述,于东等人^[10]使用支持向量机训练汉语句子难度评估模型,程勇等人^[11]基于 BERT 预训练模型来提取文本中句子的特征表示,融合语言特征与 BERT 特征来进行难度预测。虽然上述方法在汉语阅读难度分级方面取得了一些进展,但神经网络模型强大的建模能力还未在汉语文本难度分级上进行深入探索和利用。

2 层级网络模型

本文采用层级网络模型来进行汉语文本难度分级,模型结构由图 1 的 N -gram 特征抽取和图 2 的篇章级特征提取两个部分组成。

2.1 N -gram 特征抽取

N -gram 特征抽取利用卷积神经网络 CNN 通过不断加深网络层数由词向量表示逐渐提取不同程度的 N -gram 特征表示。本文采用了变长卷积层和块结构(Block)。

2.1.1 变长卷积层

文本难度分级的语料多为中小学教材。不同于一般的文本长度服从正态分布的数据集,本数据集由于其天然的属性使得数据集中文本的长度相差很大,大部分低年级文本较短,而高年级文本可以很长。因此,本模型先对文本序列长度做压缩处理,采用变长卷积层对文本做特征提取,即随着序列的增加,相应的卷积核大小和步长也随之增加。设输入的文本序列为 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, $\mathbf{X} \in \mathbb{R}^{n \times \text{dim}}$, 其中 $\mathbf{x}_i$ 为第 i 个词语的词向量。具体计算如式(1)、式(2)所示。

$$f_i = \text{conv}([\mathbf{x}_{\lambda_i} : \mathbf{x}_{\lambda_{i+1}}]), i = 0, 1, 2, 3, 4 \quad (1)$$

$$F = \text{conv}(\text{concat}(f_i)), i = 0, 1, 2, 3, 4 \quad (2)$$

其中, λ_i 表示对文本序列切分的系数, $[\mathbf{x}_{\lambda_i} : \mathbf{x}_{\lambda_{i+1}}]$ 表示从 $\mathbf{X}$ 中截取 $(\mathbf{x}_{\lambda_i}, \mathbf{x}_{\lambda_i+1}, \dots, \mathbf{x}_{\lambda_{i+1}-1})$, conv 表示卷积操作, f_i 表示卷积之后得到的特征表示, concat 表示将 f_i 沿着高度这一维度进行拼接, F 表示经过整个变长卷积层的输出特征表示。

① <https://lexile.com/>

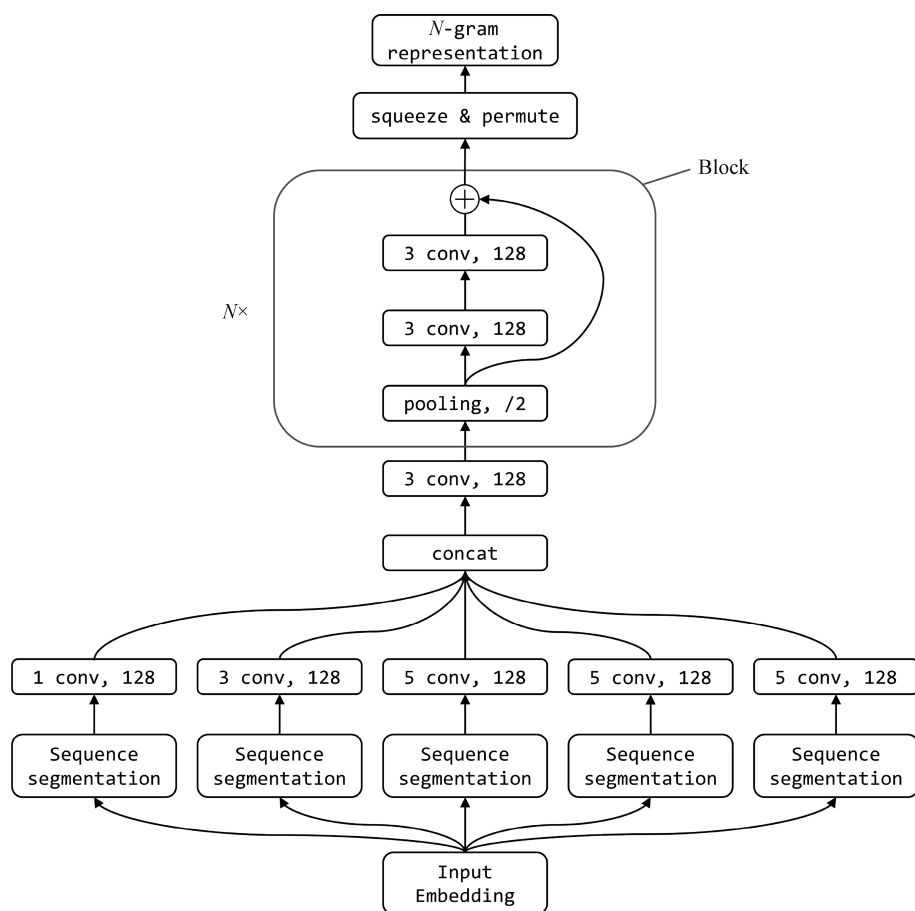


图 1 N-gram 特征抽取

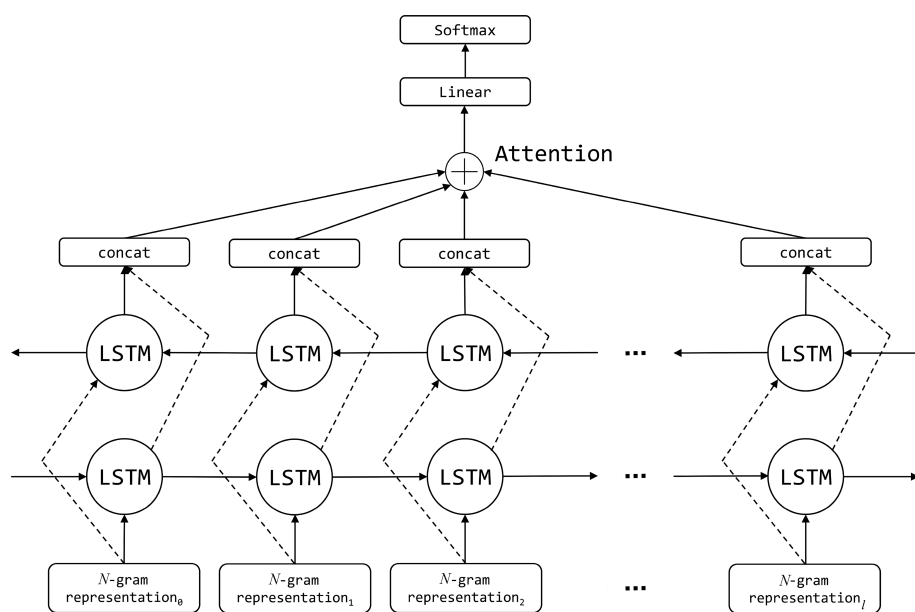


图 2 篇章级特征提取

2.1.2 块结构

块结构 Block^[12]对输入进行下采样(Downsampling),每做一次下采样,下一个卷积层的序列长度就会减半。经过这样不断的下采样,可以增加卷积核对原文的覆盖度。在 K 次下采样之后,卷积核相当于覆盖了 2^K 长度的原文,因此模型对全局特征的提取能力得到了大幅增强。同时,由于压缩了序列长度,有助于接下来双向 LSTM 对篇章级文本表示进行建模。具体计算如式(3)、式(4)所示。

$$px = \text{pooling}(F) \quad (3)$$

$$FM = \text{conv}(\text{conv}(px)) + px \quad (4)$$

其中, pooling 表示池化操作, conv 表示卷积操作, 此处为迭代结构, 迭代次数由序列长度和特定任务决定。

$$S = \text{permute}(FM) \quad (5)$$

其中, permute 表示将特征图 FM 的通道维度和高度维度进行交换, S 表示提取到的 N -gram 特征表示, $S \in \mathbb{R}^{l \times C}$, 其中, l 为特征图的高, C 为卷积核的通道数。

2.2 篇章级特征提取

篇章级特征提取的作用是由 N -gram 特征表示提取到篇章级文本表示, 采用了双向 LSTM 网络和注意力机制。具体计算如式(6)所示。

$$\vec{h}, \overleftarrow{h} = \text{BiLSTM}(S) \quad (6)$$

其中, $\vec{h} = (\vec{h}_1, \vec{h}_2, \dots, \vec{h}_l)$ 和 $\overleftarrow{h} = (\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_l)$ 分别表示正向和反向每个时刻的隐层向量。

本文实验了两种方法。一种方法是取正向 LSTM 和反向 LSTM 的最后一个时间步的隐层表示拼接作为最终的文本表示 T , 如式(7)所示。

$$T = \text{concat}(\vec{h}_l, \overleftarrow{h}_1) \quad (7)$$

另一种方法是, 将每个时刻的正反隐层表示拼接, 得到附带上下文表示的序列 h , 如式(8)所示。

$$h = \text{concat}(\vec{h}, \overleftarrow{h}) \quad (8)$$

其中, $h = (h_1, h_2, \dots, h_l)$, $h_i = (\vec{h}_i, \overleftarrow{h}_i)$, $i \in [1, l]$

再初始化一个可学习的上下文向量 u , $u \in \mathbb{R}^{2 \times \text{hidden}}$, hidden 为 LSTM 隐层向量维度, 计算出每个 h_i 的权重 α_i , 如式(9)所示。

$$\alpha_i = \frac{\exp(h_i u)}{\sum_{i=1}^l \exp(h_i u)} \quad (9)$$

最后对 h_i 进行加权求和, 得到最终文本表示 T , 具体计算如式(10)所示。

$$T = \sum_{i=1}^l \alpha_i h_i \quad (10)$$

2.3 标签平滑的损失函数

将式(7)或式(10)得到的文本表示 T , 经过线性变换得到文本属于第 i 类的概率 p_i , 如式(11)所示。

$$p_i = \text{softmax}(\text{linear}(T)) \quad (11)$$

本文模型使用交叉熵损失函数进行训练, 并且加入了标签平滑。主要考虑到有以下原因:

(1) 标签平滑训练可以抵抗噪声, 产生更好的校准网络, 从而更好地泛化, 最终对不可见的测试数据产生更准确的预测。

(2) 对于传统的文本分类任务, 例如新闻文本分类, 积极消极情感分析等, 不同标签之间是具有互斥性的, 而难度分级任务, 不同的标签之间是非完全互斥的, 级别与级别之间有交叉的部分, 尤其是邻接级别, 使用平滑后的标签能够更加真实地反映文本难度分布。

具体计算如式(12)~式(14)所示。

$$y_i = (1 - \text{eps}) \times y_{\text{onehot}} + \frac{\text{eps}}{N} \quad (12)$$

$$y_i = \begin{cases} 1 - \text{eps}, & i = \text{target} \\ \frac{\text{eps}}{N}, & i \neq \text{target} \end{cases} \quad (13)$$

$$\text{Loss} = \sum_{i=1}^N y_i \log(p_i) \quad (14)$$

其中, y_{onehot} 是真实标签的独热编码形式, eps 是平滑系数, N 是多分类的类别总个数, y_i 是平滑后的真实标签, p_i 表示样本属于第 i 类的概率。

3 实验设置

3.1 数据集

我们通过各种途径搜集了北师大版、人教版、苏教版、部编版等十多个出版社的小学、初中、高中教材, 从中删除了诗歌和文言文, 作为本文的实验数据集。依据 8:1:1 的比例划分训练集、开发集和测试集。

为了进行充分的实验, 作了两种不同的难度等级划分。第一种是五分类, 对应于《语文课程标准》^[13]的学段划分, 即第一学段(1~2 年级), 第二学段(3~4 年级), 第三学段(5~6 年级), 第四学段(7~9 年级)和高中。第二种是八分类, 采用年级进行等级划分, 包括小学 6 个年级、初中、高中, 由于初

中和高中每个年级的文本非常少而且年级之间文本难度界限比较模糊,本文将初中 3 个年级、高中 3 个年级合并为初中和高中两个级别。实验用的数据集分布见表 1、表 2。虽然我们经过各种努力来搜集分级阅读的文章,但是数据集规模依然很小,这对训练一个高性能的分类模型有很大挑战。

此外,本文构建了一个汉语文本难度等级的人

工测试集。从不同来源的文本(阅读材料、教辅材料等)搜集了 200 篇文章,邀请了三位小学教师进行双盲人工标注,对每一篇文章标注唯一的难度等级标签。欣喜的是,三位教师的标注一致性非常高,标注结果的肯德尔和谐系数(Kendall coefficient of concordance)达到了 0.946,保证了一个高水平的测试集。人工测试集的数据分布如表 1、表 2 所示。

表 1 五分类数据集划分(文章篇数)

标签	训练集	验证集	教材测试集	平均字数	平均词数	人工测试集	平均字数	平均词数
1	507	59	69	269	180	65	189	128
2	630	83	86	684	446	85	674	443
3	647	82	87	1 151	745	50	1 968	1 289
4	335	43	43	2 154	1 394	0	0	0
5	251	31	32	3 168	2 028	0	0	0
合计	2 370	298	317	1 192	771	200	840	552

表 2 八分类数据集划分(文章篇数)

标签	训练集	验证集	教材测试集	平均字数	平均词数	人工测试集	平均字数	平均词数
1	209	23	27	175	119	28	106	72
2	295	40	42	333	221	37	251	170
3	314	41	45	558	367	52	483	319
4	316	39	43	810	524	33	974	639
5	341	46	46	1 034	671	25	923	596
6	300	40	40	1 290	834	25	3 013	1 982
7	336	43	43	2 163	1 399	0	0	0
8	250	33	31	3 168	2 028	0	0	0
合计	2 361	305	317	1 194	773	200	840	552

3.2 参数设置

在神经网络模型中,序列长度设置为 4 500,可以覆盖 99%以上的文本。自训练词向量使用 Word2Vec 算法的 Skip-gram 模型,窗口大小设置为 10,最小词频设置为 5,词向量维度设置为 128。卷积层激活函数使用 ReLU。Block 结构迭代次数除八分类 VBCNN+BiLSTM+Att 模型设置为 3 外,其他模型均设置为 5。Dropout 添加在最后的线性层之前,设置为 0.5。优化器使用 Adam。标签平滑损失函数的系数(公式(12)中的 ϵ)五分类设置为 0.8,八分类设置为 0.2。学习率除八分类 VBCNN+BiLSTM+Att 模型设置为 0.000 1

外,其他模型均设置为 0.000 01。batch-size 五分类设置为 16,八分类设置为 64。剩余模型相关参数如表 3 所示。

表 3 模型相关参数

参数名称	参数值
序列切分系数 λ_i	0, 300, 900, 2 100, 3 300, 4 500
变长卷积层卷积核大小	1, 3, 5, 5, 5
变长卷积层卷积核步长	1, 2, 4, 4, 4
变长卷积层卷积核通道数	128, 128, 128, 128, 128
其它卷积核大小、步长、通道数	3, 1, 128

续表

参数名称	参数值
池化核大小、步长	3, 2
双向 LSTM 隐层维度	64×2

3.3 评价方法

实验采用准确率 P 、召回率 R 、 F 值, 正确率 (Accuracy) 作为模型的评估指标。同时, 考虑到文本难度等级是一个连续系统, 本文也采用了邻接正确率 (Adjacency Accuracy) 作为指标, 即将模型预测类别在真实标签两侧 (± 1) 的情况视为正确。

3.4 机器学习基线模型

为了验证模型的效果, 本文选择了支持向量机 SVM 作为传统机器学习基线模型, 选取了丰富的语言特征 (表 4), 利用 MinMax 方法对不同特征进行了归一化处理。

表 4 SVM 难度分类器的语言特征

特征	名称
字	1.总字数;2.字种数;3.字 TTR;4.平均笔画数;5.加权平均笔画数;6.少笔画字(笔画数 ≤ 10)比例;7.中笔画字($10 < \text{笔画数} \leq 20$)比例;8.多笔画字(笔画数 ≥ 20)比例;9.平均字频;10.单次字数量;11.单次字比例;12.常用字比例;13.未登录字比例;14.一级字数量;15.一级字比例;16.二级字数量;17.二级字比例;18.三级字数量;19.三级字比例;20.四级字数量;21.四级字比例;22.平均汉字等级
	23.总词数;24.词种数;25.词 TTR;26.平均词语长度;27.平均词频;28.单字词比例;29.二字词比例;30.三字词比例;31.四字词比例;32.多字词比例;33.未登录词比例;34.单次词数量;35.单次词比例;36.成语数量;37.成语比例;38.甲级词数量;39.甲级词比例;40.乙级词数量;41.乙级词比例;42.丙级词数量;43.丙级词比例;44.丁级词数量;45.丁级词比例;46.平均词语等级;47.名词比例;48.动词比例;49.数词比例;50.形容词比例;51.副词比例;52.地名比例;53.人名比例
句	54.句子数;55.句子平均汉字数;56.句子平均词语数;57.句子最大汉字数;58.句子最大词语数;59.分句数;60.分句平均汉字数;61.分句平均词语数;62.分句最大汉字数;63.分句最大词语数;
段	64.总段数;65.段落平均汉字数;66.段落平均词语数;67.段落最大汉字数;68.段落最大词语数
主题	69~100.文章主题概率分布

其中, 字(词)TTR 指字(词)种数与总字(词)数之比。特征 14~22 的汉字等级、特征 38~46 的词语等级使用了《汉语水平词汇与汉字等级大纲》^[14] 中的甲乙丙丁四级, 计算时将未登录字(词)看作等级为 0, 甲级为 1, 乙级为 2, 依此类推。主题特征 69~100 用于抽取文章的内容信息, 设置主题数为 32, 以无监督的方式训练得到 LDA 主题模型, 对数据集中的每一篇文章, 使用训练好的 LDA 模型进行推断, 获得每篇文章的主题概率分布, 作为文章的主题特征。

3.5 神经网络基线模型

TextCNN^[15]: 使用了卷积核大小为 2、3、4, 通道数均为 128 的 3 个卷积层分别对文本做特征提取, 然后将其拼接作为文本表示。

TextRNN^[16]: 使用双向 LSTM 的最后时刻正向隐层和反向隐层表示拼接作为文本表示。

FastText^[17]: 使用词向量的平均作为文本的表示。

Transformer^[18]: 使用了 3 层 Transformer 的 encoder 结构, 前馈神经网络中间层大小为 256。

BERT^[19]: 取文本长度为 5 000 (字级别) 的文本, 将其分割成长度为 50 的序列 100 个, 每个序列使用 CLS 符号对应的输出作为其表示, 然后最大池化得到文本表示。

同时, 对本文模型, 实验了 3 种不同结构以对比结果:

VBCNN+BiLSTM: 取双向 LSTM 最后时刻的正、反隐层表示拼接作为最终文本表示。

VBCNN+BiLSTM+Att: 对双向 LSTM 每个时刻的正、反隐层表示拼接, 然后进行加权作为最终文本表示。

VBCNN+BiLSTM+Att+LS: 损失函数加入标签平滑, 这是本文的完整模型。

4 结果分析

4.1 主要实验结果

表 5 呈现了五分类的实验结果, 表 6 呈现了八分类的实验结果。

表 5 五分类难度等级的实验结果

模型	教材测试集					人工测试集					总体	
	<i>P</i>	<i>R</i>	<i>F</i> ₁	Acc	Acc(A)	<i>P</i>	<i>R</i>	<i>F</i> ₁	Acc	Acc(A)	Acc	Acc(A)
SVM	0.677	0.669	0.667	0.669	0.965	0.774	0.715	0.737	0.715	0.965	0.687	0.965
TextCNN	0.663	0.634	0.630	0.634	0.921	0.703	0.675	0.687	0.675	0.995	0.650	0.950
TextRNN	0.564	0.568	0.541	0.568	0.953	0.677	0.655	0.665	0.655	0.955	0.602	0.954
FastText	0.736	0.710	0.712	0.710	0.972	0.800	0.735	0.755	0.735	0.980	0.720	0.975
Transformer	0.667	0.656	0.651	0.656	0.953	0.780	0.715	0.732	0.715	0.950	0.679	0.952
BERT	0.702	0.685	0.683	0.685	0.981	0.728	0.705	0.709	0.705	0.995	0.692	0.986
VBCNN+BiLSTM	0.735	0.719	0.721	0.719	0.984	0.814	0.755	0.782	0.755	0.965	0.733	0.977
VBCNN+BiLSTM+Att	0.748	0.726	0.730	0.726	0.978	0.813	0.765	0.787	0.765	0.980	0.741	0.979
VBCNN+BiLSTM+Att+LS	0.756	0.741	0.745	0.741	0.987	0.819	0.775	0.793	0.775	0.995	0.754	0.990

表 6 八分类难度等级的实验结果

模型	教材测试集					人工测试集					总体	
	<i>P</i>	<i>R</i>	<i>F</i> ₁	Acc	Acc(A)	<i>P</i>	<i>R</i>	<i>F</i> ₁	Acc	Acc(A)	Acc	Acc(A)
SVM	0.517	0.502	0.504	0.502	0.861	0.526	0.495	0.504	0.495	0.880	0.499	0.868
TextCNN	0.502	0.505	0.500	0.505	0.839	0.505	0.445	0.461	0.445	0.840	0.482	0.839
TextRNN	0.539	0.527	0.530	0.527	0.896	0.545	0.515	0.527	0.515	0.870	0.522	0.886
FastText	0.571	0.558	0.552	0.558	0.886	0.585	0.520	0.527	0.520	0.890	0.544	0.888
Transformer	0.499	0.502	0.488	0.502	0.874	0.618	0.545	0.561	0.545	0.885	0.518	0.878
BERT	0.484	0.476	0.475	0.476	0.826	0.496	0.475	0.480	0.475	0.905	0.476	0.857
VBCNN+BiLSTM	0.562	0.552	0.552	0.552	0.902	0.651	0.590	0.615	0.590	0.880	0.567	0.894
VBCNN+BiLSTM+Att	0.558	0.558	0.555	0.558	0.899	0.643	0.585	0.604	0.585	0.895	0.569	0.897
VBCNN+BiLSTM+Att+LS	0.549	0.543	0.539	0.543	0.896	0.640	0.590	0.606	0.590	0.915	0.561	0.903

在五分类数据上,本文最终模型 VBCNN + BiLSTM + Att + LS 在两个数据集上、在各项指标上均取得了最佳性能,在教材测试集上正确率达到了 0.741,超过 SVM 模型 7 个点,在人工测试集上正确率为 0.775,超过 SVM 模型 6 个百分点。由于人工设计了大量语言特征,SVM 模型也取得了较为满意的效果。而常用的神经网络模型并没有取得理想性能,CNN、RNN、Transformer 的性能都不如 SVM 分类器,尤以 RNN 效果最差。比较意外的是,简单作词向量平均的 FastText 性能超过了 SVM。在大规模语料上做预训练的 BERT 取得了与 SVM 相媲美的结果,但其简单易行,不需要人工来设计林林总总的语言特征。

本文实现的基础 VBCNN + BiLSTM 模型就超越了 SVM 分类器和所有主流的神经网络模型,

在教材测试数据上正确率达到了 0.719。在其上添加注意力机制,正确率提升了 0.07,再利用标签平滑损失函数,正确率进一步提升 0.15。在人工测试集上呈现同样的上升趋势。

而在八分类数据上,由于数据集太小,各个模型的性能急剧下降。总体上,本文模型取得了最佳性能。

观察邻接正确率可以发现,不同模型都取得了较为满意的结果。在五分类数据上,我们的模型 VBCNN + BiLSTM + Att + LS 在教材测试集和人工测试集上分别达到了 0.987 和 0.995;在八分类数据上,邻接正确率也达到了 0.9 左右。这说明,虽然计算机给每篇文章赋予一个准确的类别有一定的难度,但是预测结果都围绕在真实标签附近。

另外,本文还实验了百度百科和人民日报的预

训练词向量,但是总体效果不如表 4、表 5 所汇报的仅基于本文训练数据训练出的词向量。虽然预训练词向量所用的语料巨大,包含的信息也更为丰富,但是宽泛通用的语料可能并不适用中小学教材文本,而且处理本文数据时存在缺词现象,训练语料中词频大于等于 5 的词语在百度百科和人民日报预训练词向量的词表中分别缺失 8.03%和 8.32%。

4.2 平滑系数分析

为了探究损失函数中的平滑系数(即式(12)中的 ϵ)对模型性能的影响,本文对不同平滑系数下模型的表现进行了实验,具体结果如图 3 所示。

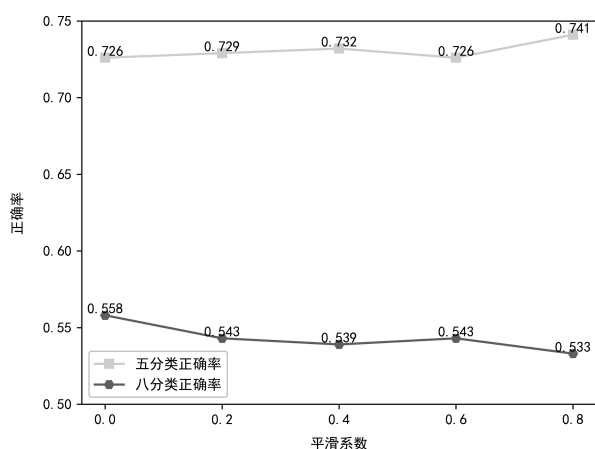


图 3 不同平滑系数下模型的正确率

可以看到,在五分类任务上,随着平滑系数的增加,模型的性能呈上升趋势;而在八分类任务上,随着平滑系数的增加,模型的性能呈下降趋势。本文认为这是由于五分类划分精度较为粗糙,不同难度级别之间的交叉情况相比八分类要严重,所以使用平滑后的标签更能代表其真实分布;而八分类任务本

身划分要精细一些,不同难度级别之间的交叉情况比五分类要轻微,所以使用原始的独热编码标签效果要好一些。

4.3 模型对比分析

为了探究神经网络模型结构对性能的影响,本文对模型结构拆解进行了实验。

VBCNN 表示只使用 N -gram 特征提取,将整篇文本看成是由词语组成的长序列,重复迭代 Block 结构至序列长度为 1 作为最终文本表示;

BiLSTM+Att 表示将词向量特征直接输入 LSTM 特征提取器,然后将附带上下文信息的词语序列加权作为最终文本表示;

VBCNN+BiLSTM+Att 是本文采用的层级神经网络结构,用 CNN 提取 N -gram 特征进而用 LSTM 提取篇章特征。

实验结果如表 7 所示。不同模型结构中,VBCNN 将序列长度大幅压缩,对长文本(多为难度较高的文本)比较有益;BiLSTM+Att 由于其本身建模的序列长度不能太长,直接使用时对短文本(多为难度较低的文本)比较有益。实验结果也验证了这一点:由于人工测试集中只有小学阶段的文本,其中短文本占比较高,而教材测试集里面还包括了初高中文本,高中文本中短文本占比要稍微低一些,所以在教材测试集上 VBCNN 的表现比 BiLSTM+Att 要好,尤其是对于粗划分的五分类任务,效果更为明显;而在人工测试集上面,BiLSTM+Att 的表现比 VBCNN 要好,同样在粗划分的五分类任务上效果更为明显。可以看到,本文模型较好地融合了两种结构的优点,取得了最好的性能。

表 7 不同模型结构的正确率

模型结构	五分类正确率			八分类正确率		
	教材测试集	人工测试集	总体	教材测试集	人工测试集	总体
VBCNN	0.722	0.690	0.710	0.517	0.540	0.526
BiLSTM+Att	0.612	0.770	0.673	0.508	0.575	0.534
VBCNN+BiLSTM+Att	0.726	0.765	0.741	0.558	0.585	0.569

4.4 混淆矩阵分析

在教材测试集上,分别根据 SVM 分类器和表现最好的神经网络模型的预测结果构建了混淆矩阵,其对角线上的值为每个类别的召回率,结果如

图 4~图 7 所示。

在五分类数据上,与 SVM 分类器相比,本文提出的神经网络模型明显提升了 3 级别(第四学段)和 4 级别的召回率。在八分类数据上,本文模型明显提升了 2、4、6 级别的召回率。总体上,本文神经网络

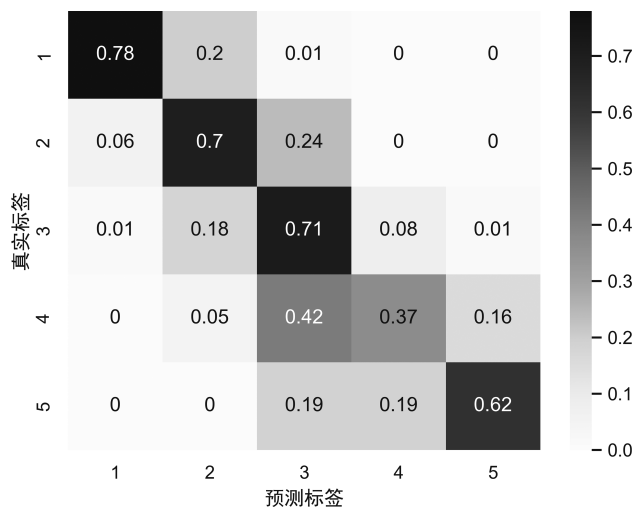


图4 五分类 SVM 模型混淆矩阵

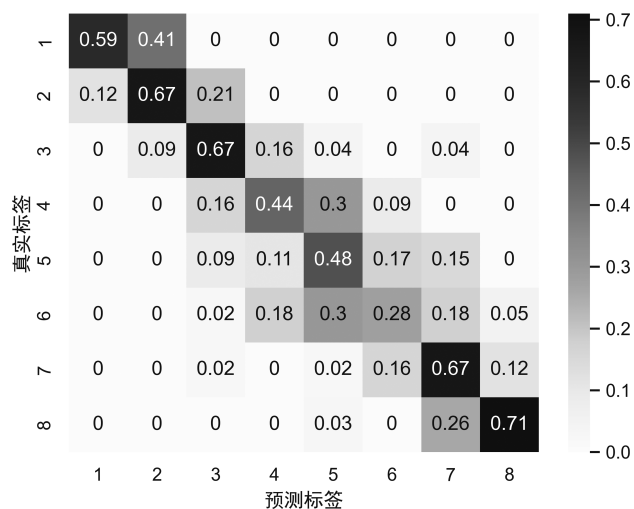


图7 八分类神经网络模型混淆矩阵

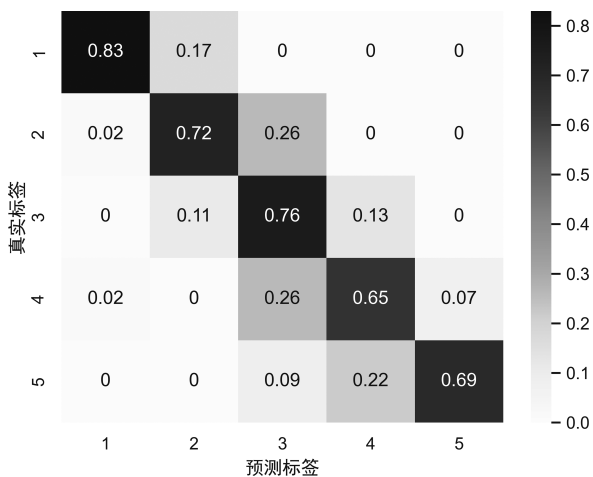


图5 五分类神经网络模型混淆矩阵

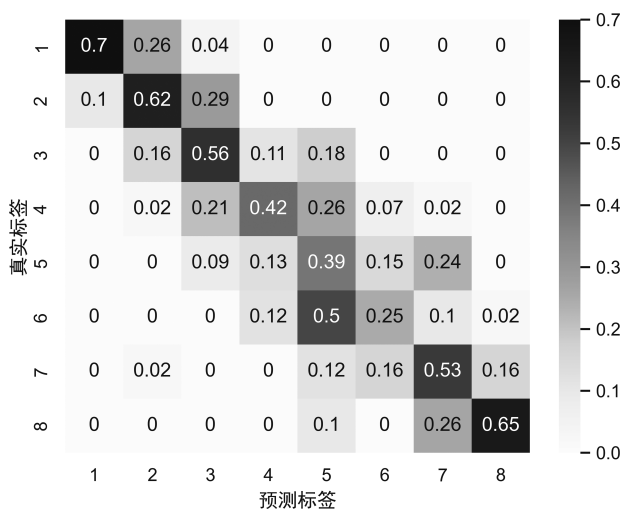


图6 八分类 SVM 模型混淆矩阵

络模型使预测类别更向真实标签靠近。

4.5 年级难度分析

进一步,在教材测试集上,根据表现最好的神经网络模型的输出结果,表8和表9汇报了不同级别的分类性能。

表8 五分类不同级别的分类性能

标签	P	R	F_1
1	0.950	0.826	0.884
2	0.738	0.721	0.729
3	0.647	0.759	0.698
4	0.609	0.651	0.629
5	0.880	0.688	0.772
weighted avg	0.756	0.741	0.745

表9 八分类不同级别的分类性能

标签	P	R	F_1
1	0.762	0.593	0.667
2	0.651	0.667	0.659
3	0.577	0.667	0.619
4	0.500	0.442	0.469
5	0.431	0.478	0.454
6	0.367	0.275	0.314
7	0.547	0.674	0.604
8	0.759	0.710	0.733
weighted avg	0.558	0.558	0.555

五分类和八分类数据都表现出了一个共同的趋势,即低年级(小学一、二年级)和高年级(高中)分类效果较好,而中间年级分类效果不理想。小学一、二

年级文本最简单,字、词、句特征明显;高中阶段文本最复杂,相应的语言特征也明显,分级系统的两端呈现出了较好的预测效果。而中间阶段的四、五、六年级文本易读性难以把握,语言特征不够鲜明,分级结果不够理想。

4.6 与人类专家的比较分析

为了进一步探究本文设计的神经网络模型对文章难度的评定性能,将本文模型在人工测试集上的预测结果也看成是一位专家(标示为 M)的标注结果,与三位人类专家(标示为 R1, R2, R3)的标注结果分别计算正确率和斯皮尔曼相关系数,其中正确率的计算方式是将一位专家的标注结果作为预测结果,另外三位专家的标注结果的平均值作为正确标签。评价结果如表 10、表 11 所示。

表 10 不同专家的预测正确率

评定者	R1	R2	R3	M
Acc(五分类)	0.865	0.880	0.790	0.775

表 11 五分类下不同专家的斯皮尔曼相关系数

评定者	R1	R2	R3	M
R1	1.000			
R2	0.922**	1.000		
R3	0.829**	0.833**	1.000	
M	0.862**	0.868**	0.815**	1.000

注: **表示在 0.01 水平下显著。

可以看到,在正确率指标下,模型的五分类表现已经接近人类专家 R3。在相关系数指标下,模型的预测结果都与专家的评定结果具有很高的一致性,斯皮尔曼相关系数显著相关。当以 R1 或 R2 为参照标准的时候,模型的表现甚至超过了人类专家 R3。

5 结语

本文围绕汉语分级阅读任务展开,提出了一种难度分级模型,在教材测试集和人工标注的测试集上对五分类和八分类任务进行了实验,取得了较为理想的效果,超越了传统的机器学习方法 SVM 分类器和主流神经网络模型。基于此模型,开发了汉语分级阅读网站^①供研究者和教育工作者免费使用。

汉语文本难度分级任重而道远,在将来的研究中,我们将构建更多、更权威的汉语分级语料库,将探索更多汉语自身的语言特点融入神经网络模型。

参考文献

- [1] 郭望皓. 对外汉语文本易读性公式研究[D]. 上海: 上海交通大学硕士学位论文, 2010.
- [2] 左虹, 朱勇. 中级欧美留学生汉语文本可读性公式研究[J]. 世界汉语教学, 2014, 028(002): 263-276.
- [3] 王蕾. 初中级日韩学习者汉语文本可读性公式研究[J]. 语言教学与研究, 187(05): 15-25.
- [4] SCHWARM S E, OSTENDORF M. Reading level assessment using support vector machines and statistical language models[C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, University of Michigan, USA. Association for Computational Linguistics, 2005.
- [5] PITLER E, NENKOVA A. Revisiting readability: A unified framework for predicting text quality[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.
- [6] WANG S, ANDERSEN E. Grammatical templates: improving text difficulty evaluation for language learners[C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 1692-1702.
- [7] 孙刚. 基于线性回归的中文文本可读性预测方法研究[D]. 南京: 南京大学硕士学位论文, 2015.
- [8] 蒋智威. 面向可读性评估的文本表示技术研究[D]. 南京: 南京大学博士学位论文, 2018.
- [9] 吴思远, 蔡建永, 于东, 等. 文本可读性的自动分析研究综述[J]. 中文信息学报, 2018, 32(12): 1-10.
- [10] 于东, 吴思远, 耿朝阳, 等. 基于众包标注的语文教材句子难易度评估研究[J]. 中文信息学报, 2020, 34(2): 16-26.
- [11] 程勇, 徐德宽, 董军. 基于多元语言特征与深度特征融合的中文文本阅读难度自动分级研究[J]. 中文信息学报, 2020, 34(4): 101-110.
- [12] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 562-570.
- [13] 中华人民共和国教育部. 义务教育语文课程标准: 2011 年版[M]. 北京: 北京师范大学出版社, 2012.
- [14] 国家对外汉语水平领导小组办公室汉语水平考试部. 汉语水平词汇与汉字等级大纲[S]. 北京: 经济科学出版社, 2001.
- [15] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the Conference on

① <http://www.chinese-pku.com/>

- Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1746-1751.
- [16] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016: 2873-2879.
- [17] JOULIN A, GRAVE É, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017: 427-431.
- [18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the NIPS, 2017.
- [19] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.



李文彪(1996—), 硕士研究生, 主要研究领域为自然语言处理。
E-mail: 2001210322@stu.pku.edu.cn



吴云芳(1973—), 博士, 副教授, 主要研究领域为文本生成、分级阅读、情感计算、语料库建设。
E-mail: wuyf@pku.edu.cn

欢迎订阅《中文信息学报》

《中文信息学报》(Journal of Chinese Information Processing) 是全国一级学会—社团法人中国中文信息学会和中国科学院软件研究所联合主办的学术性刊物, 创刊于 1986 年 10 月, 现为月刊。

《中文信息学报》是我国计算机、计算技术类中文核心期刊。主要刊登中文信息处理基础理论与应用技术方面的高水平学术论文, 内容涵盖计算语言学(包括语音与音位、词法、句法、语义、语用等各个层面上的计算), 语言资源建设(包括计算词汇学、术语学、电子词典、语料库、知识本体等), 机器翻译或机器辅助翻译, 汉语和少数民族语言文字输入输出及其智能处理, 中文语音识别及文语转换, 信息检索, 信息抽取与过滤, 文本分类、中文搜索引擎, 以自然语言为枢纽的多模态检索, 与语言处理相关的数据挖掘、机器学习、知识获取、知识工程、人工智能研究, 与语言计算相关的语言学研究等。也刊登相关综述、研究报告、成果简介、书刊评论、专题讨论、国内外学术动态等稿件。

读者对象主要是从事中文信息处理的研究人员、工程技术人员和大专院校师生等。

《中文信息学报》(国内统一刊号: CN11-2325/N; 国际统一刊号: ISSN 1003-0077) 国内外公开发行, 国内定价每期 50 元, 全年 600 元。

国内发行处: 《中文信息学报》编辑部

国外发行处: 中国图书进出口总公司 100020 北京 88-E 信箱

1. 支付宝转账: (请注明期刊征订)

账号: cips_pay@163.com

户名: 中国中文信息学会

2. 银行转账

开户银行: 工商行北京市分行海淀西区支行

户 名: 中国中文信息学会

账 号: 0200004509014415619

《中文信息学报》编辑部

地 址: 北京海淀区中关村南四街 4 号 7 号楼 201 房间

电 话: 010-62562916

电子信箱: jcip@iscas.ac.cn