基于《国际中文教育中文水平等级标准》的中文文本难度自动分级研究

——以 HSK 中高级阅读文本为例

丁安琪 张 杨 兰韵诗

摘 要: 文本难度自动分级是指通过一系列计算机自动计算,实现对文本难度级别自动识别。该研究依托 2021 年发布的《国际中文教育中文水平等级标准》,提取其基本汉字、词汇、语法特征,使用基于 BERT 的神经网络预训练模型来提取文本语义特征,融合以上两类特征构建多维特征集,并对比多种特征组合方式以及多种机器学习算法对于 HSK 中高级阅读文本的难度分级效果。实验结果表明:随机森林算法对 HSK 中高级阅读文本分级效果最优;单一维度中,语义特征对文本难度识别精确率最高,词汇次之;《国际中文教育中文水平等级标准》对 HSK 中高级阅读文本具有较强指导意义。

关键词: 文本难度分级;HSK;《国际中文教育中文水平等级标准》;BERT
中图分类号:H09 **文献标识码:**A **文章编号:**1004-9142(2023)06-0081-12

一、引言

文本难度自动分级是指根据文本易于理解程度(也称可读性、易读性),通过计算机自动计算,根据文本特征自动判断文本所属难度级别。文本难度分级研究主要服务于语言与阅读教学,在教材编纂与出版、分级阅读、语言水平测试等方面都发挥着重要作用。[①]目前进行文本难度分级主要依靠人工设计

作者简介:丁安琪,华东师范大学国际汉语文化学院教授,全球中文发展研究中心研究员;
张杨(通讯作者),华东师范大学国际汉语文化学院博士研究生;
兰韵诗,华东师范大学数据科学与工程学院副教授。

基金项目:本文系教育部中外语言交流合作中心 2021 年度国际中文教育创新项目“实境直播短期中文教学模式构建与实践”(21YH013CX1)、2022 年度《国际中文教育中文水平等级标准》教学资源建设重点项目“学术汉语语料库建设与应用”(YHJC22D067)、华东师范大学 2022 年度文化传承创新研究专项项目“国际中文中高级分级阅读系统开发与建设”(2022ECNU-WHCCYJ-29)、华东师范大学新中文教育专项课题(2022ECNU-WHCCYJ-31)阶段性成果。

① 王泉根:《新世纪中国分级阅读的思考与对策》,《中国图书评论》2009 年第 9 期。

的易读性公式。依靠人工对文本进行难度分级费时费力,随着语言信息处理技术的发展,利用计算机进行文本难度自动分级越来越受到关注。文本难度自动分级可以降低使用人工评估方式进行文本难度分级的主观性及时间成本,快速有效地为使用者提供难度适宜的文本材料。

在国际中文教育领域,利用计算机技术进行文本难度自动分级研究的成果尚不多见,目前仅有杨纯莉使用支持向量机(SVM)①和朴素贝叶斯(NB)②的算法,对对外汉语报刊进行文本难度分级实验;③朱君辉等以 16 套国际中文教育教材为阅读语料,使用多种机器学习算法考察汉语语法点对国际中文教育文本难度分级的影响;④杜月明等以 HSK 阅读文本为语料集,采用多种机器学习算法进行实验。⑤ 但杨文两种算法分级准确率只有 45%左右;朱文使用人工标注定级方式为训练语料定级,研究结果客观性不足;杜文将 HSK 一至三级文本纳入统计中,这些文本多以短句形式呈现,加大了字词因素对文本的影响,对整个模型整体预测效果也会产生一定影响。

本文拟以《国际中文教育中文水平等级标准》(GF0025—2021,以下简称《标准》)为基础,以 HSK 中高级阅读文本为参照,通过对已标注 HSK 等级的中高级阅读文本进行难度分析,探讨文本难度自动分级构建最佳路径。选择《标准》作为基础,是因为《标准》是首个面向外国中文学习者全面描绘、评价其中文语言技能和水平的规范标准,将外国学习者中文水平分为初、中、高三个等第,每个等第包含三个级别,并针对"三等九级"每一级别,从音节、汉字、词汇、语法四个维度,给出了明确说明。《标准》为国际中文教育领域文本难度分级研究提供了权威的国家级标杆。以 HSK 中高级阅读文本为参照,是因为这些材料来源于汉考国际官方公布正式样卷或正式出版的具有权威性的 HSK 教材,其等级已经明确,可以为难度自动分级结果提供检测依据。

具体来说,本文旨在探索以下问题:

1. 不同层面特征对 HSK 中高级阅读文本难度预测能力如何?哪种模型对 HSK 中高级阅读文本预测能力最佳?

2. 模型对 HSK 不同级别阅读文本预测能力如何?

二、基于《标准》的 HSK 中高级阅读文本难度自动分级

从计算机角度来看,文本难度自动分级问题可以看作是一个分类问题,这是自然语言处理领域非常经典的问题。一般来说,包括以下三个步骤:第一步,构建文本难度数据集。数据集包括训练数据集、验证数据集以及测试数据集三个部分。其中,训练数据集用于训练分级模型,帮助模型调整自身参数以拟合数据集;验证数据集用来调整分级模型超参数,增强模型拟合能力;测试数据集用于测试模型分级效果。第二步,构建文本难度特征集。文本难度特征集是文本难度分级依据,在以往研究中,学者多从字、词、句、篇等维度对文本特征进行标记。⑥⑦⑧ 第三步,构建实验模型并分析实验结果。

本节将介绍数据集的构建、特征集的构建和抽取过程、实验模型的构建以及实验结果分析。具体实验流程见图 1。

---

① Cortes, C., & Vapnik, V., Support-vector Networks, *Machine Learning*, vol. 20, no. 3, 1995, pp. 273-297.

② Cover, T., & Hart, P., Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory*, vol. 13, no. 1, 1967, pp. 21-27。

③ 杨纯莉:《基于统计算法的对外汉语报刊文本易读性词汇因素分析》,华东师范大学硕士论文,2018。

④ 朱君辉、刘鑫、杨麟儿、王鸿滨、杨尔弘:《汉语语法点特征及其在二语文本难度自动分级研究中的应用》,《语言文字应用》2022 年第 3 期。

⑤ 杜月明、王亚敏、王蕾:《汉语水平考试(HSK)阅读文本可读性自动评估研究》,《语言文字应用》2022 年第 3 期。

⑥ 郭望皓:《对外汉语文本易读性公式研究》,上海交通大学硕士论文,2010。

⑦ 左虹、朱勇:《中级欧美留学生汉语文本可读性公式研究》,《世界汉语教学》2014 年第 2 期。

⑧ 吴思远、于东、江新:《汉语文本可读性特征体系构建和效度验证》,《世界汉语教学》2020 年第 1 期。

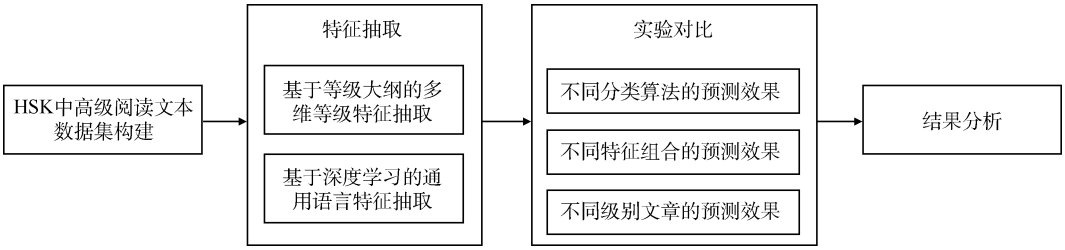


图 1　HSK 中高级阅读文本难度自动分级实验流程

1. HSK 中高级阅读文本数据集构建

本研究面向 HSK 中高级阅读文本进行难度分析,因此实验数据均为 HSK 相关官方阅读文本材料,包括来自北京语言大学出版社的《HSK 标准教程》①(4~6 级,各上下两册,共 6 册)、配套练习册《HSK 标准教程练习册》②(4~6 级,各上下两册,共 6 册)、孔子学院总部与国家汉办编制的《新汉语水平考试样卷》③(4~6 级,各 1 册,共 3 册)、《新汉语水平考试真题集》④(4~6 级,各 40 余册,共 129 册),共计 144 份材料。

首先,借助 CnOCR 文字识别(Optical Character Recognition,OCR)工具包⑤,我们从上述教材中抽取出每单元课文正文,从上述练习册、样卷、真题集中抽取部分阅读文本。考虑到阅读部分题型较为丰富多样,且部分题型特殊,我们对不同等级阅读部分的题型进行了筛选,如因四级阅读第一部分针对短对话的选词填空题,对话字数过少,未达到一篇文本的长度,我们将其舍弃。基于上述考虑,我们抽取的阅读文本范围为:四级阅读第三部分、五级阅读第一二三部分、六级阅读第三四部分。

然后,我们通过人工校对方式,对文字识别抽取的文本结果进行格式、标点、空格校对,构建了 HSK 中高级阅读文本数据集。数据库最终保留 2595 篇文本,详细信息如表 1 所示。

表 1　HSK 中高级阅读文本数据集统计信息

文本级别	文本数/篇	总字数/字	总句数/句	平均文本长度/字	平均文本句数/句	平均句长/字	字均笔画数/笔
四级	971	82789	2810	85.26	2.89	29.46	7.05
五级	1163	273628	8279	235.28	7.12	33.05	7.14
六级	461	267644	7220	580.57	15.66	37.07	7.29
所有级别	2595	624061	18309	240.49	7.06	34.08	7.13

2. HSK 中高级阅读文本特征集构建

阅读文本难度等级与其所在标准下的汉字、词汇、语法等语言基本要素等级大纲有着密切关系。判断阅读文本难度,往往需要考察该文本中汉字、词汇与语法的难度。而对于中高级阅读文本来说,随着文本难度升高,其所包含语义信息也更为丰富,语义信息也会较大程度影响一篇中高级阅读文本整体难度。

我们根据《标准》中汉字、词汇、语法等级大纲,构建基于《标准》的多维等级特征以衡量一篇文本的局部难度;同时,基于深度学习模型构建语义特征以衡量一篇文本的全局难度。由以上两类特征,共同

① 姜丽萍、李琳、于淼:《HSK 标准教程》,北京语言大学出版社,2015 年。
② 姜丽萍:《HSK 标准教程练习册》,北京语言大学出版社,2019 年。
③ 数据来源:https://www.chinesetest.cn/gosign.do?id=1&lid=0#,新汉语水平考试样卷。
④ 国家汉办、孔子学院总部:《新汉语水平考试真题集 HSK(六级)》,商务印书馆,2012 年。
⑤ CnOCR:CnOCR 是用来做中文 OCR 的 Python 3 包。CnOCR 自带训练好的识别模型,安装后即可直接使用。下载链接:https://gitee.com/cyahua/cnocr.

构成适用于 HSK 中高级阅读文本难度分级的语言特征体系。《标准》汉字、词汇、语法等级大纲统计信息见表 2。下面分别对各项特征进行具体说明。

表 2　《标准》汉字、词汇、语法等级大纲统计信息

级别	汉字	词汇	语法
一级	300	500	48
二级	300	772	81
三级	300	973	81
四级	300	1000	76
五级	300	1071	71
六级	300	1140	67
七—九级	1200	5636	148
总计	3000	11092	572

(1)基于《标准》的多维等级特征

基于《标准》的多维等级特征由三个部分构成:汉字等级特征、词汇等级特征、语法等级特征。为探究各类型特征对于文本难度分级的有效性,针对每一个特征,我们对 HSK 中高级阅读文本数据集中所有文本分别进行相应的特征难度等级统计。通过分析各等级特征所对应的统计图,认定各等级特征有助于区分文本难度等级,并基于该发现将各等级特征首次纳入 HSK 阅读文本难度分级语言特征体系。

① 汉字等级特征

汉字等级特征,指各等级汉字数量在文本总汉字数中占比分布情况,共 7 维,每一维对应一个汉字等级。

对文本中每一个汉字,我们均根据《标准》中的汉字等级大纲,查找其对应的难度等级。遍历文本中所有汉字后,通过计算各等级汉字占比,获得汉字等级特征。我们对 HSK 中高级阅读文本数据集进行上述处理并统计汉字难度等级,结果如图 2 所示。

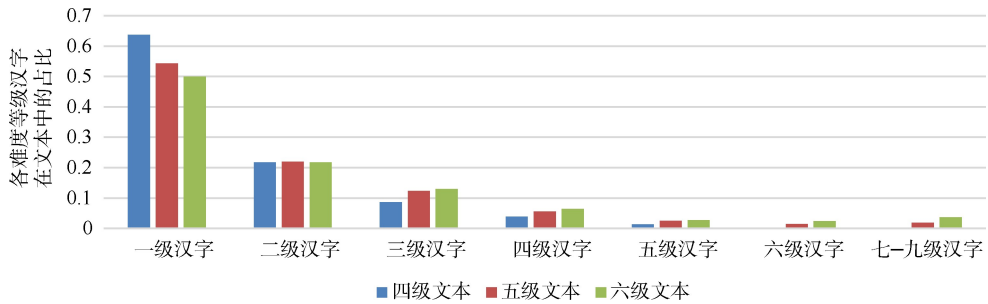


图 2　HSK 中高级阅读文本汉字难度等级占比统计

由图 2 可知,随着文本难度等级升高,高难度汉字使用占比也逐渐增加。因此我们认为文本汉字难度等级分布情况能够在一定程度上表征文本难度,汉字等级特征有助于区分文本难度等级。

② 词汇等级特征

词汇等级特征,指各等级词汇数量在文本总词汇数中的占比分布情况,共 9 维。我们以《标准》中词汇等级大纲为依据,前 7 维分别对应一个词汇等级,第 8 维为未收录词语,第 9 维为专有名词和特殊词汇。

借助中文分词工具,我们首先对文本执行分词操作,①继而根据《标准》中的词汇等级大纲,分别对文本中每一个词语,查找其相对应的难度等级。针对大纲没有的词语,增加一个“未收录词语”维度和一个“专有名词和特殊词汇”维度。“专有名词”收录特定的人、地名、机构名称等;“特殊词汇”收录专业术语、特殊的俗语、成语、诗词等。“专有名词”和“特殊词汇”统计信息和样例见表 3。《标准》提供的词汇难度等级大纲共 7 个难度等级,增加上述两个维度后,共计 9 维。

表 3　“专有名词表”和“特殊词汇表”统计信息和样例表

词表	数量	样例
专有名词表	1426	阿尔卑斯、北京、嫦娥、成龙、关羽、哈尼族、康熙、李世民、李小龙、纽约大学、蒙古
特殊词汇表	1394	不管三七二十一、初生牛犊不怕虎、独钓寒江雪、胆固醇、地心引力、胡萝卜素、硫酸盐、脑溢血、内分泌

HSK 中高级阅读文本数据集词汇难度等级占比统计结果如图 3 所示。

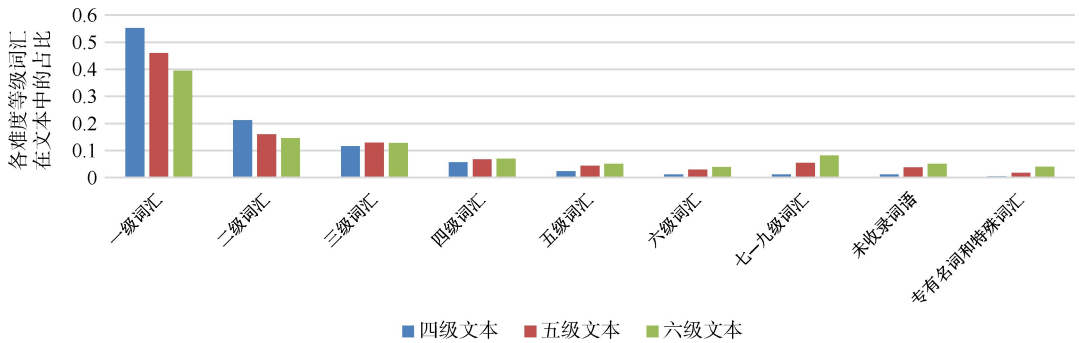


图 3　HSK 中高级阅读文本词汇难度等级占比统计

从图 3 中可以看出,随着文本难度升高,高难度词汇使用占比逐渐升高,低难度词汇使用占比逐渐降低;文本难度越高,未收录词语占比越高;随着文本难度提升,文本中专有名词和特殊词汇占比也逐级升高。因此,可以认为文本词汇难度等级分布情况能够在一定程度上代表文本难度,词汇等级特征有助于区分文本难度等级。

③ 语法等级特征

语法等级特征,指各等级语法数量在文本总语法数中的占比分布情况,共 7 维,每一维对应一个语法等级。

我们通过正则表达式匹配找到文本中包含的所有语法点,并根据《标准》中的语法等级大纲,找到各语法点对应的难度等级,以此得到每篇文本中包含的各等级语法点数量。如“只要你认真学习,就一定能取得好成绩”,通过正则表达式匹配得到语法点“只要……,就……”,查找语法等级大纲,可知其语法等级为“二级”。继而将各等级语法点数量除以该文本中总语法点数量,计算得到文本中各等级语法点频率,将各级别频率拼接成 7 维向量,作为该文本语法等级特征。HSK 中高级阅读文本数据集语法难度等级统计结果见图 4。

由图 4 可知,随着文本难度等级的升高,五、六、七级语法点使用占比逐渐升高,而低难度等级语法点使用占比则有所下降。说明文本难度在一定程度上与文本语法点等级分布情况有关,语法等级特征有助于区分文本难度等级。

① 分词是自然语言处理的基本操作之一,目的是将连续文本分割成一个个独立的词元。分词工具来自,https://github.com/fxsjy/jieba.

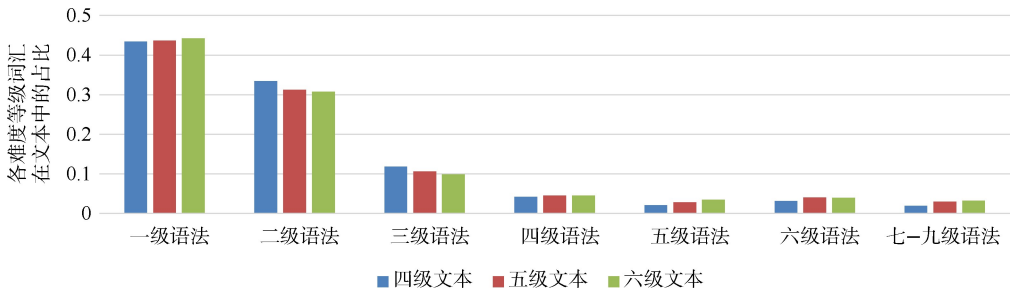


图 4　HSK 中高级阅读文本语法难度等级占比统计

(2)基于深度学习的通用语言特征

一般来说,难度等级高的文本会表达更复杂的语义,而难度等级较低的文本可能包含更多简单句,传递更简单的语义。这是区分文本难度的重要特征。因此,除了基于《标准》的多维等级特征之外,我们还使用了基于深度学习的通用语言特征来衡量文本语义信息。① 该信息通过深度学习模型对文本进行整体编码,语义特征共两项,其中一项为 BERT② 语义特征,共 768 维;另一项为 DeBERTa③ 语义特征,共 768 维。

以上两个深度学习模型旨在通过将离散的字符映射到连续的向量空间表示字符语义信息。两者都使用多层 Transformer④ 结构实现,并在大规模语料上通过一系列预训练任务进行训练。训练得到的模型参数学习了自然语言表达模式,使模型具有优秀的编码文本语义特征能力,从而使模型对输入文本能够在一定程度上表征其深层语义信息。一般认为,模型输出层第一个字符"[CLS]"的 768 维向量,能够在一定程度上表示该输入文本的语义信息,我们使用这 768 维向量来表示文本中所包含的语义信息。

3. 特征抽取过程

(1)基于等级大纲的多维等级特征抽取

汉字等级特征部分,我们使用 Python 编程工具清除文本中非中文字符后,遍历文本中所有汉字,根据《标准》中的汉字等级大纲,对文本中每一个汉字,查找其相对应的难度等级,从而获得每一等级汉字频数。将各等级汉字频数除以总汉字数后,各级频数转换为频率,各级频率拼接形成汉字难度等级分布特征,该特征共 7 维。

词汇等级特征部分,我们使用 Python 编程工具,借助 Jieba 工具⑤,对文本进行分词处理。为了提升 Jieba 分词效果,使其分词结果更加适合当前任务,我们在 Jieba 工具自定义词典中,导入了《标准》"词汇大纲"以及我们构建的"专有名词表"和"特殊词汇表"。

经过上述分词处理后,文本被转换为词汇列表。根据"专有名词表"和"特殊词汇表",我们遍历查找出专有名词和特殊词汇,统计其数量归入"专有名词和特殊词汇"维度,并从词汇列表中剔除。

继而遍历词汇列表中所有剩余词汇,根据《标准》中的词汇等级大纲,分别对文本中每一个词语,查找其相对应的难度等级,从而获得每一等级词语频数,同时将未在大纲中找到的词语暂时归入"未收录词语"维度中。对"未收录词语"进行进一步统计分析,我们发现其中很多能够继续进行拆分,因此我们

① Zha, J., Li, Z., Wei, Y., & Zhang, Y., Disentangling Task Relations for Few-shot Text Classification via Self-Supervised Hierarchical Task Clustering, 2022, arXiv Preprint arXiv:2211, p. 08588.

② Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018, arXiv preprint arXiv:1810, p. 04805.

③ He, P., Liu, X., Gao, J., & Chen, W. Deberta, Decoding-enhanced Bert with Disentangled Attention, 2020, arXiv preprint arXiv:2006, p. 03654.

④ Vaswani, A., Shazeer, N., Parmar, N, et al., Attention is All You Need, *Advances in Neural Information Processing Systems*, 2017, p. 30.

⑤ Jieba 工具是用来实现分词操作的工具。Jieba 工具来自:https://github.com/fxsjy/jieba.

制定了 38 项规则,以进一步拆分“未收录词语”。部分拆分规则见表 4。

表 4　部分“未收录词语”拆分规则内容与举例

规则序号	规则内容	举例
规则一	“第+数词+量词”结构 “数词+量词”结构	第|一|次 一|个
规则二	“这/那/哪+数+量+名”结构 “这/那/哪+数+量”	这|一|次 那|一|匹|马 哪|两|个
规则四	“形容词+的”结构	小|小|的 可爱|的

根据《标准》提供的词汇等级大纲,分别对文本中每一个词语,查找其相对应的难度等级,从而获得每一等级词语频数,同时将未在大纲中找到的词汇归入“未收录词语”维度中。将各级词语频数除以总词汇数后,各级频数转换为频率,各级频率拼接形成词汇难度等级分布特征,该特征共 9 维。

语法等级特征部分,《标准》中的语法等级大纲中共有 572 个语法点,分属于 12 个语法类型:词类、短语、句子成分、句子的类型、动作的态、特殊表达法、提问的方法、语素、强调的方法、口语格式、句群和固定格式。我们对语法点数量占比最高的“词类”“短语”“句子的类型”“口语格式”“固定格式”等五类进行了正则表达式的大规模匹配;其余语法点数量比较少的语法类型,“句子成分”“动作的态”“特殊表达法”“提问的方法”“语素”“强调的方法”“句群”,则暂不在本工作中考虑。最终我们通过构造正则表达式,完成 474 个语法点的匹配,匹配完成度达到 82.87%。正则表达式匹配语法点示例见表 5。

表 5　正则表达式匹配语法点示例

语法类型	语法点	语法点正则表达式	句子示例	语法难度等级
词类	人称代词——我、你、您、他、她、我们、你们、他们、她们	"^([我你您他她]|我们|你们|他们|她们)$"	你好,我要两个本子。	一级
固定格式	固定格式——除了……(以外),……还/也/都……	"除了.+(以外)*,.+(还|也|都).+"	除了上课,我还要参加各种活动。	三级
句子类型	转折复句——尽管……,但是/可是……	"尽管.+,(但是|可是).+"	尽管这次考试很难,但是很多人都通过了。	五级

我们使用 Python 编程工具,借助正则表达式对文本中语法点进行匹配,根据《标准》中的语法等级大纲,对匹配后语法点查找其对应难度等级,从而获得每一等级语法点频数。将频数除以该文本中所有语法点个数后,频数转换为频率,将各级别频率拼接,形成语法难度等级分布特征,共 7 维。

(2)基于深度学习的通用语言特征抽取

基于深度学习的通用语言特征,用于衡量文本语义信息,共有两个特征,各 768 维,共 1536 维。我们借助 BERT 和 DeBERTa 两个深度学习预训练模型,提取文本语义信息。具体做法为:下载预训练好的模型参数(其中 BERT 预训练模型来自 Hugging Face/bert-base-chinese①;DeBERTa 预训练模型来自

① BERT 预训练模型是一种基于 Transformer 的 encoder 层堆积模型,具有判断句子关系的能力,通常用来提取文本的语义信息。BERT 预训练模型来自:https://huggingface.co/bert-base-chinese.

Hugging Face/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli①),借助 Python 编程工具和 Pytorch 深度学习工具,将文本转化为字符编码序列,输入模型中,取模型输出层第一个字符"[CLS]"的 768 维向量,将其作为所输入文本语义特征。

4. 实验模型构建及结果分析

(1)实验数据划分

我们按照 8∶1∶1 的比例将 HSK 中高级阅读文本数据集随机划分为训练集、验证集和测试集。训练集用于训练模型,帮助模型调整自身参数以拟合数据集,共 2076 个文本;验证集用于调整模型超参数,增强模型拟合能力,共 259 个文本;测试集用于计算模型预测准确率,判断模型效果,共 260 个文本。

我们将 HSK 中高级阅读文本难度自动分级任务抽象成分类任务,抽取汉字、词汇、语法、语义四个层面特征作为文本特征。使用支持向量机(SVM)②、K 近邻(KNN)③、朴素贝叶斯(NB)④、决策树(DT)⑤、Adaboost (Freund, 1999)和随机森林(RF)⑥算法作为模型分类算法。

(2)实验评价指标

我们将 HSK 中高级阅读文本难度自动分级任务抽象成分类任务,故而采用分类模型评价指标准确率(Accuracy)、精确率(Precision)、召回率(Recall)、调和值(F1),来衡量模型在该任务上的有效性。

(3)实验结果

为探究最适合 HSK 中高级阅读文本难度分级任务的特征组合与分类算法,我们对不同分类算法的预测效果进行了实验,也对多维等级特征的预测效果进行了讨论,对不同特征及其组合的整体预测效果和对不同级别文本的预测准确率进行了比较。

① 不同分类算法的预测能力

为了探究不同分类算法在 HSK 中高级阅读文本难度自动分级任务上的精度,我们选择"汉字+词汇+语法+语义"特征组合,比较不同分类算法在该特征组合下,在 HSK 中高级阅读文本数据集上的预测效果,统计结果见表 6。

表 6　不同分类算法下的预测效果统计表

分类算法	准确率	精确率	召回率	F1 值
支持向量机(SVM)	0.8423	0.8420	0.8423	0.8413
K 近邻(KNN)	0.7769	0.7837	0.7769	0.7744
朴素贝叶斯(NB)	0.5962	0.6501	0.5962	0.5432
决策树(DT)	0.8000	0.7990	0.8000	0.7985
Adaboost	0.7077	0.7286	0.7077	0.7053
随机森林(RF)	**0.8462**	**0.8469**	**0.8462**	**0.8460**

由表 6 可知,在当前特征组合下,随机森林算法在四项指标上均获得了最佳效果,支持向量机算法次之。原因可能是由于随机森林是一种集成式算法,通过构造多个决策树对预测结果进行投票,具有更

① DeBERTa 预训练模型在 BERT 预训练模型的基础之上增加了注意力解耦机制和增强的掩码解码器。DeBERTa 预训练模型来自:https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli.

② Cortes, C., & Vapnik, V., Support-vector Networks, *Machine Learning*, vol. 20, no. 3, 1995, pp. 273-297.

③ Cover, T., & Hart, P., Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory*, vol. 13, no. 1, 1967, pp. 21-27.

④ Lewis, D. D., Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, *In European Conference on Machine Learning*, Springer, Berlin, Heidelberg, 1998, pp. 4-15.

⑤ Quinlan, J. R., Induction of Decision Trees, *Machine Learning*, vol. 1, no. 1, 1986, pp. 81-106.

⑥ Breiman, L., Random Forests, *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32.

好的模型泛化性。因此我们决定选用随机森林算法进行后续实验。

② 单一层面特征以及多维特征组合的预测能力

为探究不同特征对 HSK 中高级阅读文本难度自动分级的影响，我们以预测准确率最佳的随机森林算法为分类模型，比较汉字、词汇、语法和语义特征及其组合在 HSK 中高级阅读文本数据集上的预测效果。不同特征组合下模型准确率、精确率、召回率、F1 值见表 7。

表 7　汉字、词汇、语法和语义层面特征及其组合预测效果统计表

特征组合	特征维度	准确率	精确率	召回率	F1 值
汉字	7	0.7269	0.7231	0.7269	0.7245
词汇	9	0.7692	0.7699	0.7692	0.7632
语法	7	0.5885	0.5869	0.5885	0.5869
语义(BERT)	768	0.8077	0.8094	0.8077	0.8051
语义(DeBERTa)	768	0.7846	0.7859	0.7846	0.7848
汉字+词汇	16	0.7769	0.7773	0.7769	0.7706
汉字+词汇+语法	23	0.8000	0.8055	0.8000	0.7963
汉字+词汇+语法+语义(BERT)	791	**0.8577**	**0.8589**	**0.8577**	**0.8528**
汉字+词汇+语法+语义(DeBERTa)	791	0.8154	0.8154	0.8154	0.8154
汉字+词汇+语法+语义(BERT+DeBERTa)	1559	0.8462	0.8469	0.8462	0.8460

从单一层面来看，语义(BERT)对文本难度预测准确率最高。除此之外，从语言要素——汉字、词汇、语法角度来看，单一层面词汇预测效果最佳。

从多维特征组合层面来看，可以发现“汉字+词汇+语法+语义(BERT)”特征组合在准确率、精确率、召回率、F1 值上均获得了最佳效果，说明汉字、词汇、语法、语义层面特征对于文本难度衡量均具有重要作用。值得一提的是，“汉字+词汇+语法”这一组合方式准确率、精准率、召回率均达到了 80%以上，F1 值为 79.63%。该结果揭示了《标准》对 HSK 中高级阅读文本具有一定的指导意义。

在语义层面特征中，相比 DeBERTa，BERT 语义特征在难度等级分级任务上表现更好，这可能是因为两者在对每个词语的编码方式上存在一定差异。BERT 在输入层中每个词语用内容嵌入和位置嵌入拼接向量表示，而 DeBERTa 采用注意力解耦机制，每个词用内容嵌入和位置嵌入两个向量表示。在文本难度评级任务中，由于文本中每个词语义内容与位置关系相对较为紧密，每个词更适合用内容嵌入和位置嵌入的拼接向量表示，因此，与 DeBERTa 相比，BERT 语义特征更适合本任务。

语义特征的加入，使模型预测准确率得到了较大幅度提升。从实验结果来看，BERT 模型提取到的深度语义特征能够在一定程度上代表文本中所表达的语义信息。

③ 单一层面特征及多维特征组合在不同级别文本上的预测能力

以表现最优的随机森林算法作为分类模型，比较汉字、词汇、语法和语义特征及其组合在 HSK 中高级阅读文本数据集上对不同级别文本的预测效果。我们用精确率这一指标，来衡量模型在不同级别的预测精度。不同特征组合下模型在四级文本、五级文本、六级文本上的预测精确率见表 8。

表 8　不同特征组合对不同级别文本预测精确率统计表

特征组合	特征维度	四级	五级	六级
汉字	7	0.8111	0.7143	0.5909
词汇	9	0.8780	0.7241	0.6970

续表

特征组合	特征维度	四级	五级	六级
语法	7	0.6087	0.5873	0.5476
语义(BERT)	768	0.8605	0.7737	0.8108
语义(DeBERTa)	768	0.8272	0.7634	0.7708
汉字+词汇	16	0.8506	0.7376	0.7500
汉字+词汇+语法	23	0.8488	0.7589	0.8485
汉字+词汇+语法+语义(BERT)	791	**0.9205**	0.8188	**0.8529**
汉字+词汇+语法+语义(DeBERTa)	791	0.8706	0.8032	0.8500
汉字+词汇+语法+语义(BERT+DeBERTa)	1559	0.8810	**0.8244**	0.8444

从表8中可以发现,就单一层面特征来看,对HSK四级阅读文本来说,词汇预测精确率最高(0.8780),语义次之(0.8605);对HSK五级和六级阅读文本来说,语义维度预测精确率依然最高,与表7一致。

就多维特征组合层面来看,“汉字+词汇+语法+语义(BERT)”特征组合方式在四级文本和六级文本上的预测精确率均达到了最佳水平,且远高于其他特征组合方式。在五级文本预测精确率上,“汉字+词汇+语法+语义(BERT+DeBERTa)”特征组合方式取得了精确率为0.8244的最佳效果,“汉字+词汇+语法+语义(BERT)”特征组合方式次之,精确率为0.8188。因此,综合来看,“汉字+词汇+语法+语义(BERT)”特征组合方式预测能力最佳。

三、结论与讨论

1. 各层面特征及多维特征组合模型预测能力

本研究结果显示,加入语义特征的多维特征组合模型,就语言要素特征组合而言,具有更好的预测能力。与此同时,就单一层面特征来说,基于语义特征的模型预测准确率最高,即使是基于汉字、词汇、语法三种特征组合的模型,其预测能力也略逊色于单一语义特征。语义特征是一种更具有普适意义的特征,可适用于不同应用场景(如二语文本和母语文本),不同文本类型(如科技文本和文学文本),具有较强泛化能力。因此,对于中文文本难度分级的研究,关注语义特征是十分必要的。

单一层面特征中,基于词汇特征的模型预测准确率位于第二。词汇特征是在众多研究中被广泛证明的有效特征。①②③④ 本研究以《标准》中的词汇大纲为原型,构建了适用于国际中文教育领域的词汇特征,在面向HSK中高级阅读文本的难度分级上,呈现了较好的预测能力。

基于汉字特征的模型预测准确率位于第三。不同于英文文本,汉字是汉语的文字载体,因此,汉字

① Feng L., Jansche M., Huenerfauth M., et al., A Comparison of Features for Automatic Readability Assessment, 23*rd International Conference on Computational Linguistics*, Posters Volume, 23-27 August 2010, Beijing, China, Association for Computational Linguistics, 2010.

② 宋曜廷、陈茹玲、李宜宪、查日苏、曾厚强、林维骏、张道行、张国恩:《中文文本可读性探讨:指标选取,模型建立与效度验证》,《中华心理学刊》2013年第1期。

③ 吴思远、于东、江新《汉语文本可读性特征体系构建和效度验证》,《世界汉语教学》2020年第1期。

④ 杜月明、王亚敏、王蕾:《汉语水平考试(HSK)阅读文本可读性自动评估研究》,《语言文字应用》2022年第3期。

对文本难度的影响也不容忽视。众多学者对于文本难度分级的研究都将汉字因素纳入其特征体系。①②③④ 在本研究中,同样证明了汉字对于文本难度分级预测的重要作用。

基于语法特征的模型预测准确率最低,仅在58%左右。说明仅依据语法难度对文本进行难度分级效果较差。然而,在汉字特征与词汇特征基础之上,加入语法特征,其预测准确率达到了80%左右。由此可见,语法特征的加入可显著提高模型预测能力。该结论在朱君辉关于汉语语法点特征的研究中亦得到证明。⑤

就多维特征组合层面而言,融合BERT语义特征的多维特征组合模型预测准确率最高,可达到85%以上,说明本研究基于《标准》设计的汉字、词汇、语法特征在融合语义特征之后,能够较好地完成对HSK中高级阅读文本的难度自动分级任务。

2. 最优模型对不同级别的预测能力

在本研究中,基于汉字+词汇+语法+语义多维特征组合的随机森林模型是预测HSK中高级阅读文本难度的最优模型。在表8中,我们可以看到最优模型对四级、五级、六级文本预测能力差异较大,模型对四级文本和六级文本预测能力显著高于五级文本。因此,我们通过构建混淆矩阵的方法,尝试对测试集中的分级情况进行进一步分析,结果如图5所示。

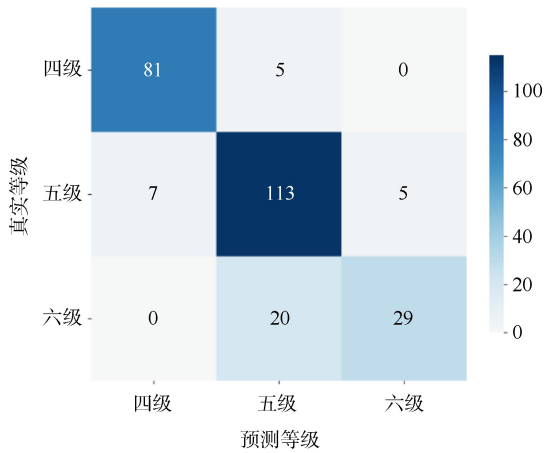


图5　随机森林模型预测结果的混淆矩阵

由图5可知,在预测等级为四级和六级的文本中,均有少量文本真实等级是五级;而在预测等级为五级的文本中,有一部分文本真实等级是六级,少量文本真实等级是四级。由此可见,模型可以较好地区分四级和六级文本,展示了对越级文本的显著区分能力,对于位于中间的五级文本的区分能力,尚有进步空间。

众所周知,处于中间级别的文本具有模糊性,即使对专家而言,确定中间级别也是困难的,进一步挖掘五级文本和六级文本之间的特征差异,或可提高模型对五级文本的预测能力。

① 郭望皓:《对外汉语文本易读性公式研究》,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》,《世界汉语教学》2014年第2期。

③ 王蕾:《初中级日韩学习者汉语文本可读性公式研究》,《语言教学与研究》2017年第5期。

④ 刘苗苗、李燕、王欣萌、甘琳琳、李虹:《分级阅读初探:基于小学教材的汉语可读性公式研究》,《语言文字应用》2021年第2期。

⑤ 朱君辉、刘鑫、杨麟儿、王鸿滨、杨尔弘:《汉语语法点特征及其在二语文本难度自动分级研究中的应用》,《语言文字应用》2022年第3期。

四、结语

本研究将《标准》中汉字、词汇、语法等级大纲作为特征集的主要特征，并融合基于深度学习的BERT语义特征，构成本研究的多维特征集。通过对比六种常见机器学习算法在HSK中高级阅读文本中的应用，实现了基于多维特征的随机森林（RF）算法模型对HSK中高级阅读文本的难度自动分级。研究结果表明：第一，单一特征维度语义特征对HSK中高级阅读文本难度自动分级精确率最高，加入语义特征的多维特征模型对HSK中高级阅读文本难度自动分级效果最佳。因此，面向文本难度分级的研究应关注语义特征对文本难度的影响。第二，语言要素中，词汇特征对HSK中高级阅读文本难度预测能力最强，汉字特征次之，加入语法特征之后，模型预测能力能够得到显著提升。第三，融合BERT语义特征的多维特征组合随机森林算法模型对HSK中高级阅读文本预测效果最佳，达到了85%左右。该模型在HSK四级、五级、六级阅读文本中预测精准率分别达到了92%、81%、85%，模型对四级和六级文本区分度较高，对五级和六级文本区分度较低，尚有进一步优化空间。

自《标准》发布以来，由于其与大家熟悉的HSK大纲不完全一致，不断有人质疑两者之间关系。本研究基于“汉字+词汇+语法”特征组合方式的随机森林算法模型，对HSK中高级阅读文本预测能力可达到80%左右，对HSK四级和六级阅读文本预测精准率甚至达到了85%左右，对HSK五级阅读文本预测精准率在82%左右。这也从一个侧面证明了《标准》作为国际中文教育领域唯一国家级标准，与HSK中高级文本关联度较高，可以为HSK优化和改革奠定坚实基础。

（本文得到谭可人的协助，其所做工作包括实验设计、实验实施和结果分析，特此致谢！）

（责任编辑：子聿）