阅读分级相关研究综述

饶思敏　郑婳　李素建*
北京大学,计算语言学教育部重点实验室
北京市,100871
{raosimin,zhenghua,lisujian}@pku.edu.cn

摘要

阅读分级的概念在二十世纪早期就被教育工作者提出，随着人们对阅读变得越来越重视，阅读分级引起了越来越多的关注，自动阅读分级技术也得到了一定程度的发展。本文总结了近年来的阅读分级领域的研究进展，首先介绍了阅读分级现有的标准和随之而产生的各种体系和语料资源。在此基础之上整理了在自动阅读分级工作已经广泛应用的三类方法：公式法、传统的机器学习方法和最近热门的深度学习方法，并结合实验结果梳理了三类方法存在的弊利，以及可以改进的方向。最后本文还对阅读分级的未来发展方向以及可以应用的领域进行了总结和展望。

关键词： 阅读分级 ； 分级体系与资源 ； 文本难度 ； 深度学习

A Survey of Leveled Reading

Simin Rao, Hua Zheng, Sujian Li
Key Lab of Computational Linguistics (MOE), Peking University
Beijing，100871
{raosimin,zhenghua,lisujian}@pku.edu.cn

Abstract

The concept of Leveled Reading (LR) originates in the early twentieth century among educators. As people gradually emphasize more on reading, they pay more attention to LR, and thus fostering developments in automatic LR methods. This paper provides an overview of the recent LR developments. Firstly, we introduce the existing standards of LR, followed by the development of various systems and data resources. Then, we classify widely-applied automatic LR methods into three types:formula method, traditional machine learning method and the recently mainstream deep learning method. We explore the advantages and disadvantages of these methods through experiments and investigate possible performance improvements. Finally, we summarize and prospect the future development of LR and provide several fields than can benefit from LR.

Keywords: Leveled Reading , Leveled Reading Systems and Resources , Text Difficulty , Deep Learning

*通讯作者

第二十届中国计算语言学大会论文集，第689页-第702页，呼和浩特，中国，2021年8月13日至15日。

1 引言

阅读分级就是按照儿童不同年龄段的智力和心理发育程度，为儿童提供科学的阅读计划和适合的读物。随着当今社会对教育和阅读越来越看重，阅读分级对于作者、出版商、老师家长和学生都是极为重要的。不同难度的文本在用词造句、主题思想表达等方面有本质的区别，不同年龄段的读者对于题材的选择也有很大的不同，对于血腥、暴力等题材的阅读接受程度也会不同。阅读分级要在理解文本和掌握不同年龄段儿童阅读水平的基础上进行图书的分级，并不是一件容易的工作，需要语言学、心理学和教育学等各领域的研究工作者共同推动前进(李云飞和袁曦临, 2013)。

近年来阅读分级工作在很多国家和地区都得到了相关部门和研究工作者的重视(李云飞和袁曦临, 2013)，而阅读分级的核心问题是分级标准的建立。英文的阅读分级工作相比其他语言更加成熟，市面上也已经有很多广为使用的阅读分级标准。对于不同的语言，分级标准的建立天然有所不同，不可生搬硬套。随着科技的进步，自动阅读分级引起了不少研究工作者的关注(Martinc et al., 2019)，自动阅读分级指将文本输入模型就可以对文本的难度估计一个数值或其他形式的预测结果。目前已存在一些可用于自动阅读分级任务的语料和分级方法(Martinc et al., 2019)，大体分为公式法、机器学习方法和深度学习方法。至今在英文上已经有100多个阅读分级公式，这些公式多为相关领域学者的研究成果(罗德红和余婧, 2013)。而在非英文的低资源语言上，阅读分级的工作正处于起步阶段，尚不存在正式的阅读分级标准。以中文为例，目前的分级图书有限且多为引进的译本，而中文的分级标准的建立正处于分级建议的时期(罗德红和余婧, 2012)。这导致在中文的自动阅读分级研究缓慢很多，在语料库构建和分级方法的探索上也远远落后于英文。如何在英文等高资源的语言上取得更好的效果以及如何在中文等低资源语言上取得可以广泛应用的成果是我们需要思考和解决的问题。所以本文我们重点介绍的是自动阅读分级的背景资源和技术发展，为了今后的科研工作者进行阅读分级的研究提供一些参考和指导。

研究阅读分级的自动评估技术会给很多的应用场景提供便利。我们可以根据自动阅读分级模型来评估书籍文本的难度，方便找到适合阅读的阅读材料；可以为图书的写作提供科学依据，指导作者写出特定难度级别的阅读材料。在自然语言处理领域，研究自动阅读分级技术对于难度可控的文本生成、文本简化、自动文章评分等任务会有一定的促进作用。

本文的组织安排如下：

第二节主要介绍阅读分级的体系和资源；

第三节对现有的阅读分级方法进行全面而详细的介绍与总结，提出目前面临的挑战；

第四节展望阅读分级的未来研究方向；

第五节对阅读分级相关研究综述做出总结。

2 阅读分级体系与资源

阅读分级的概念最早于二十世纪早期在美国被提出，随后在相关研究工作者的科学分析下各家机构逐步形成了不同的阅读分级体系。研究自动阅读分级技术需要有高质量的大规模标记语料被用于训练和测试模型。由于英语国家拥有较早的读物分级意识，因此语料资源比较丰富(吴思远等, 2018)。在2.1节中，我们将描述阅读分级的相关概念和影响因素；在2.2节中，我们将介绍和比较市面上广泛使用的阅读分级体系；在2.3节中，我们将重点介绍并比较被广泛使用的英文阅读分级资源，也会简要介绍已知的其它语言资源。

2.1 阅读分级相关概念和因素

不同的论文中对于阅读分级的描述会采用到不同的术语，例如分级阅读、文本可读性，下文中若出现上述词汇，表示的是一样的概念。影响阅读分级的主要因素是文本难度，因此要用更系统科学的方法来分析与文本难度相关的主观和客观因素，以及如何支持读者去找到合适难度的文本正是阅读分级研究的内容。很多教育工作者对于影响阅读分级标准的因素给出了总结，大致分为三类。首先是人的层面，主要包括本身年龄、认知的发展、以及阅读的基础能力。第二是书籍，大致分为篇幅长短、篇章排版、题材类型、语法难度、词汇的难度、情节的难度等六个方面。第三是环境，大致包括个人兴趣专业、学习动机、教育和社会背景等方面。

对于自动阅读分级研究,核心是分级标准的构建，当我们输入一本图书的时候就可以根据分级标准给我们对应的级别。落后的分级标准的制定会导致一系列的问题，第一个问题就是其内

第二十届中国计算语言学大会论文集，第689页-第702页，呼和浩特，中国，2021年8月13日至15日。

容产出上，没有一套完善的阅读分级标准去规范，导致内容产出者如何去产出没有确定的评判指标。第二个问题就是一本书与一个读者之前的适配性是不确定的。以中文为例，中文的阅读分级一直没有确定下来的标准，语言上存在的差异也使得我们无法直接照搬已有的英文分级标准。

2.2 阅读分级体系

不同的人或者机构对阅读分级的标准的制定都有所不同，在不同语言上也有很大差异。在欧美国家阅读分级已成为一种落到实处的阅读教学和阅读出版社的指导工具，并产生了一系列得到广泛应用的阅读分级标准，从而形成了各种各样的阅读体系。英文上被广泛应用的阅读分级体系大致可以分为数字体系、字母表体系和年级体系。

数字体系的阅读分级方式最多，如Lexile[0]、恢复阅读能力（Reading Recovery，RR）[1]、阅读发展评价体系（Developmental Reading Assessment，DRA）[2]、阅读促进计划（Accelerated Reader Level, AR）[3]等，都采用数值对阅读量化计分，图书越难，分数越高，学生通过测评分数确定水平。平时Lexile的用处最为广泛，同一个分值的书籍很多，内容覆盖很广。Lexile分级法主要考察的是单词难度和句法难度等客观因素，进行精确量化而得到的分数，缺点是只涉及到语言难度，不考虑到图书的内容思想，毕竟内容和思想无法用简单的公式来进行量化。

字母表体系如市面上广泛使用的RAZ系列书籍，即按照字母顺序分级。国内外的很多学校都是RAZ的忠实粉丝，目前正在配合教学使用。RAZ 综合考虑了孩子的英文水平、心智发育程度、兴趣爱好等要素，为不同英文水平的孩子提供不同难度的英文读物。RAZ主要靠专业的工作人员评估，不仅仅分析了文本的量化指标，也分析了难以量化的因素，比如句式、句子复杂度、语义明晰度、主题熟悉度、思想内涵、插图信息、篇章排版等信息。内容分为科普和故事两类，覆盖人文、地理、童话、小说、科学和认知等方面。形式包括文本和插画这些，插画在低年级图书中占得比例尤为重要。在语言能力方面，在每本书中同样的句式会反复出现，帮助儿童在阅读的时候熟悉句式和适宜的单词集，但随着级别的变高文本会出现不再刻意重复、基础拼写模式变小、不规则拼写增多和专业词汇变多等特点。在认识能力方面，级别低的图书内容都是贴近生活的、熟悉的话题，并且很少有故事情节，级别越高就会出现抽象和陌生的概念、话题等等特征。RAZ相比Lexile有更加精细的测评，会考虑到语言难度、内容难度两个方面因素，对于阅读分级更具有解释性。

年级体系指按照年级区分应有的阅读水平，如Grade Equivalent Level (GEL)，在各国的教育体系中，教育部门出版的教学材料是一种形式的年级体系图书，这属于资源比较丰富的材料，但是目前作为语料用于自动阅读分级的研究工作中还比较少见，原因大致分为版权原因和处理为可用于训练的文本数据过于麻烦。

虽然各分级方式产生的目的、诉求、计算方法等不同，但是英文阅读分级的方式有很多的相通的地方，各分级标准可以相互参照，更利于满足读者的不同需求。在附录A11中我们对数字体系、字母体系和年级体系的几个有代表性的分级方式进行了对照，不同的方式中级别与级别之间均存在一定程度的交叉。针对这种边界性的问题，对于日常的图书选择并不是造成很大的干扰，例如针对这个问题，Lexile等机构提供了阅读评估体系评估阅读能力，方便根据实际的阅读水平去进行图书的选择，而不是依靠年级或者年龄，这样图书和阅读者之间会有一个更好的适配性。但是如果仅仅是从图书的角度去考虑，将不同体系的图书进行对齐用作自动阅读分级技术的语料资源，如何有效地处理这些边界性问题是一个需要考虑的问题。阅读分级体系的发展给教育领域带来了很多教育资源，也为很多阅读分级语料的形成提供了数据和参考依据。如第2.3节会提到的Newsela语料[4]就是参考Lexile来进行评判和分级的。

2.3 阅读分级语料库

阅读分级相关语料资源主要分为两类，第一类是研究工作者已经收集到的带有阅读级别标记的语料，目前已经用于训练和测试模型；第二类是由于版权等问题，目前还没有处理为可用

[0]https://lexile.com/
[1]https://readingrecovery.org/
[2]https://www.pearsonclinical.ca/en/products/product-master/item-597.html
[3]http://elkin.k12.nc.us/ees/ARLevel3Title.htm
[4]https://newsela.com

的语料或者是处于没有公开的状态。在本文的自动阅读分级研究中，我们只从书籍的层面出发去研究文本的难度，而不考虑个人的因素。同时根据目前语料库的局限性，我们也较少的关注书籍的图形层面，例如字体大小、颜色对比、篇章排版等。本节对阅读分级相关语料库进行介绍，以备开展后续的自动阅读分级研究工作。

年级	数量	单词数
2^{nd}	351	71.5K
3^{rd}	589	444K
4^{th}	766	927K
5^{th}	691	1M

Table 1: Weekly Reader语料

年龄	级别	数量
6∼7	L1	529
8∼9	L2	767
10∼11	L3	801
12∼13	L4	1288
14∼17	L5	845

Table 2: Weebit语料

年龄	Lexile	数量
5∼6	150-350	150
7∼9	550-750	10
10∼14	950-1150	10
15∼19	1350-1550	10
20+	1750-1950	10

Table 3: Lexile语料

年龄	级别	数量
3∼8	L1	189
8∼9	L2	189
9∼10	L3	189

Table 4: OneStopEnglish语料

年级	Slovenian数量	Slovenian块	Newsela数量
1^{st}	8	85	N/A
2^{nd}	7	181	224
3^{rd}	7	334	500
4^{th}	13	1258	1569
5^{th}	15	1480	1342
6^{th}	12	1196	1058
7^{th}	13	1837	1210
8^{th}	15	2304	1037
9^{th}	16	2689	750
10^{th}	11	2077	750
11^{th}	4	737	2
12^{th}	3	662	1853
13^{th}	1	56	N/A

Table 5: Slovenian和Newsela语料，N/A表示Newsela语料上没有对应的等级。

英语国家的语料资源较为丰富，对于缺乏成熟阅读分级标准的语言，语料集的主要来源是教材课本。三个使用最广泛的英文语料集包括WeeBit(Vajjala and Detmar , 2012), Newsela和OneStopEnglish(Vajjala and Ivana , 2018)。WeeBit语料集是基于两个阅读分级网站the Weekly Reader[5]和BBC-Bitesize[6]的文本建立的，共包含对应不同的年龄段的5个类别，覆盖了从科学到实事等多个领域，Xia et al. (2016)在原有2012年WeeBit语料集的基础上进行了数据的清洗，原版语料包含很多的断句和额外的例如版权标注、链接等网页信息，新版语料对这些进行了处理，得到了较为干净的数据，新版语料具体信息如表2所示；OneStopEnglish语料集(Vajjala and Ivana , 2018)是基于阅读分级网站OneStopEnglish[7]建立的，其数据文本主要面对英语的二语习得者的阅读训练，不同于WeeBit，OneStopEnglish对得到的文本进行了重写，分别重写为基础难度、中等难度和高等难度，因此语料总共包含189篇文章的3个难度版本，具体信息如表4所示；Newsela语料集(Xia et al., 2016)主要覆盖新闻类文本，和OneStopEnglish相似，也是将得到的文本进行重写，但是其数据量更大，并重写为5个难度版本，对应的年级是2年级到12年级，Newsela的分级标准是Lexile。由于Newsela目前没有公开的语料集，其统计数据取自论文，语料的具体信息如表5所示。

还有不少论文工作中也构建了语料集，例如Weekly Reader(Weekly Reader, 2004)是由Weekly Reader出版的针对2、3、4、5四个年级段的教育读报，主要包含科学、历史和实事的非小说类读物，具体信息如表1所示。Sun et al. (2020)中收集了200本Lexile的图书，按照年龄和Lexile分值分为五个等级，具体如表3所示，论文在数据处理问题上避免了边界性问题。例如350-550这段分值的书籍可能既适合5-6岁儿童，也适合7-9岁儿童，会有一定的干扰性，所

[5]http://www.weeklyreader.com/
[6]https://www.bbc.co.uk/bitesize
[7]https://www.onestopenglish.com/

以语料集在处理的时候特意避免了这个问题。也有一些通用的语料集也被用于阅读分级的研究中，例如Wikimedia (2018)中描述的simpleWiki语料集，分为简单和复杂两类的维基百科文章，简单版本由复杂版本改编而来，用词和语法都更加简单，语料集中每类都包含131459篇文章。

近年来非欧美国家的阅读分级研究也促进了一些语料集的产生，Glazkova et al. (2020)在俄语上进行阅读分级的研究，由于版权的限制，收集了5592篇儿童和成人图书的预览部分，大概占全书的5-10%。其中4492篇用于训练，1000篇用于测试。Martinc et al. (2019)提到的Slovenian语料中包含2年级到13年级的125本图书，论文对数据进行分割处理，按照20个句子的长度将书本划分为块，增加了数据量同时处理得到的短文本更便于处理，数据的具体信息如表5所示。

经过调研我们发现还有很多未开发为语料集的阅读分级资源，例如上文提到的RAZ系列图书。RAZ每个级别每个月都在增加读物，相信如果能用于自动阅读分级的研究，将会有很大的助益。目前RAZ还没有公开可用的语料，我们对RAZ的部分书籍进行了处理得到文本数据，每个类别各获取10篇文章，考虑到年龄范围和与其他语料集的对照关系，我们将D-P分为第一类，对应Weebit的L1级别，Q-W分为第二类，对应Weebit的L2级别，X-Z分为第三类，对应Weebit的L3级别。我们在此数据上展开了一些实验，考虑到版权问题，不会将数据进行公开，第三节中我们将对实验结果进行说明和分析。

3 阅读分级技术

李云飞和袁曦临 (2013)研究发现，传统阅读分级方法主要集中于对文本特征的提取和分析。这类方法的优势在于计算量小，但提取的特征较浅。近年来的自动阅读分级技术是基于大规模语料的训练方法，融合了传统特征、机器学习和深度学习训练出的抽象特征，并在多个语料资源的多项指标上取得了目前最佳的结果。为了评估阅读分级方法的有效性，我们一般采用准确率、精确率、召回率和F1值作为分级结果的评估指标。其中准确率指的是预测正类占总的比率，精确率指的是预测正类中的实际正类占全部预测正类的比率，召回率指的是预测正类中的实际正类占全部实际正类的比率，F1值结合了精确率和召回率两项指标。在本节中，在3.1节中介绍传统分级公式法，在3.2节中介绍传统机器学习方法，并在3.3节中介绍深度学习方法，其中我们会着重介绍并比较各类深度学习方法的模型与结果。

3.1 基于公式的阅读分级方法

阅读分级的公式方法从1920年代开始兴起，且大多数公式都是基于英语开发的，根据人工经验进行量化。早期的公式主要衡量两项内容，即较小语言单位（如词、短语等）的难易程度和句子的复杂程度。为了衡量每个词的难易程度，分级公式一般采用两种做法，第一种是利用词长或词的音节数多少来表示难易程度，第二种是建立词汇分级表，按照该词在词表中的等级来给予其特定的难易等级程度。相应的，句子的难易程度一般利用句长来表示。

大部分阅读分级公式旨在确定该文本目标读者的学龄，在表6中，我们总结了8个被广泛使用的英语文本阅读分级公式。在确定词和短语的难易程度时，Gunning Fog Index (GFI)(Gunning , 1952) , Flesch Reading Ease (FRE)(Kincaid et al., 1975), Flesch-Kincaid Grade Level (FKGL)(Kincaid et al., 1975), Automated Readability Index (ARI)(Smith and Senter egroup, 1967)和Simple Measure of Gobbledy-gook (SMOG)(Laughlin and Harry , 1969)都基于使用第一种方法，通过来衡量词句的难易程度来进行分级，即利用词长或词的音节数，这里“长词”表示长度大于7的词。Dale-Chall Readability Formula (DCRF) (Dale et al., 1948)采用了第二种方法构建词表来衡量词的难易程度，他们建立一个包含763个词的难词表，包含在该词表中的词被认定为“难词”，否则为一般词，这个词表后来被扩展到3,000词(Chall and Dale , 1995)；Fry Short Passage measure (FSP) (Fry, 1990) 在此基础之上采用了一个更大规模的词表来标定词的难易程度，包括了4300个词；Lexile Measure(Stenner et al., 1988) 通过计算词语在Carroll-Davies-Richman语料库(Carroll , 1972)中的出现频率来决定词的难易程度。

上述公式都是针对英文的阅读分级公式，由于各类语言差异较大，因此无法被直接应用于其它语言。例如，公式中出现的“长词”就无法作为中文的有效统计指标，因为中文的词汇大多为二字词，因此词长并不能反映词汇的难易程度。也有部分研究工作致力于发展针对特定语言的阅读分级公式。例如在中文上，王进等 (2017)提出的基于自然语言处理的图书分级模型主

方法	年级	公式
GFI(Gunning , 1952)	6^{th}-17^{th}	$0.4(\frac{\text{词汇数量}}{\text{句子数量}} + \text{长词数量}\%)$
ARI(Smith and Senter egroup, 1967)	0^{th}-17^{th}	$4.71(\frac{\text{字符数量}}{\text{词汇数量}}) + 0.5(\frac{\text{词汇数量}}{\text{句子数量}}) - 21.43$
SMOG(Laughlin and Harry , 1969)	5^{th}-18^{th}	$1.043\sqrt{\frac{\text{多音节词数量} * 30}{\text{句子数量}}} + 3.1291$
FRE(Kincaid et al., 1975)	5^{th}-18^{th}	$206.835 - 1.015(\frac{\text{词汇数量}}{\text{句子数量}}) - 84.6(\frac{\text{音节数量}}{\text{单词数量}})$
FKGL(Kincaid et al., 1975)	4^{th}-18^{th}	$0.39(\frac{\text{词汇数量}}{\text{句子数量}}) + 11.8(\frac{\text{音节数量}}{\text{词汇数量}}) - 15.59$
DCRF(Dale et al., 1948)	4^{th}-15^{th}	$0.1579(\frac{\text{难词数量}}{\text{词汇数量}} * 100) + 0.0496(\frac{\text{词汇数量}}{\text{句子数量}})$
FSP(Fry, 1990)	1^{th}-17^{th}	$\frac{\text{词汇难度} + \text{句子难度}}{2}$
Lexile(Stenner et al., 1988)	2^{th}-12^{th}	$9.82\log(\frac{\text{词汇数量}}{\text{句子数量}}) - \frac{2.15}{\log(\text{词汇出现频率})}$

Table 6: 针对英文的阅读分级公式。在年级体系中，0表示的是幼儿园年级；1-6年级是小学；7-9年级是中学；10-12年级是高中；13-18是大学及以上级别。

要研究字、词、语义、语法等因素对图书阅读理解的影响。就图书字难度而言，依据国务院于2013年公开发布了《通用规范汉字表》[8]，作者将字根据使用频率的高低分为五个等级，建立起几个字难度依次递增的字库。计算图书字难度时，文本每个字符到五个字符库中依次匹配，在某级字库中匹配到则乘以该字库难度系数。就图书句子难度而言，难度高的句子存在句子长度过长、包含多个陌生词组和句子语法冗余等特征。作者针对句子难度的研究主要从句子长度、包含词组数和短语数以及词组和短语属于几级词汇来计算评估，不涉及到语法特征和语义特征。在测试图书句子难度时，分别计算每个句子的难度值取平均来表示这篇文章的句子难度值。图书句子难度等于句子长度和词组因素的和。最后实验结果表明，当测试字数为2500-3000的白话文类型的图书时，测试结果可以接近人工测评的水平。

公式	准确率	F1值
GFI	0.39	0.42
ARI	0.34	0.40
FRE	0.14	0.40
FKGL	0.21	0.28
DCRF	0.73	0.61

Table 7: 公式法在Weebit语料集上的结果

公式	准确率	F1值
GFI	0.21	0.17
ARI	0.53	0.55
FRE	0.13	0.03
FKGL	0.24	0.22
DCRF	0.13	0.03

Table 8: 公式法在RAZ语料集上的结果

3.2 基于机器学习的阅读分级方法

传统的分级方法大多采用文本分类的形式，而这些方法大多是传统的机器学习分类器、统计语言模型或者特征提取。最早的阅读分级方法之一是由Schwarm and Ostendorf (2005)提出的使用支持向量机（SVM）作为分类器在WeeklyReader语料集上进行训练,模型将n-grams语言模型、统计可读性公式和分析树都作为分类器的特征。 由于阅读分级公式法的计算效率极高，因此在阅读分级领域中被长期采纳为主流方法。然而，面对现今大规模的线上语料文本，公式法有几项严重的局限。其一，公式法在统计阅读文本中的词汇与句子的难易程度时，是将其作为无结构且无序的离散变量来统计的，且假定输入的大规模文本中几乎没有噪音其次，公式法只能提取浅层的语言特征，而无法将深层的重要语言特征纳入考量，例如句法歧义、

[8]http://www.gov.cn/gzdt/att/att/site1/20130819/tygfhzb.pdf

Grade	Flesch-Kincaid	Lexile	SVM(2005)	SVM(2009)
2	78%	33%	5.5%	0%
3	67%	27%	3.3%	3%
4	74%	26%	13%	13%
5	59%	24%	21%	9%

Table 9: Weekly Reader语料集在Flesch-Kincaid, Lexile和SVM上的分类错误率。

语言连贯和主题适度等。本文中我们分别以应用广泛的Weebit语料和RAZ语料为代表来共同进行公式法的测试以便来探究公式的性能。公式法在Weebit语料的结果如表7所示，其中有三个指标无法参与计算：SMOG无法计算总句子数<30的文本，FSP和Lexile未找到公开词表。结果显示，除DCRF外，指标结果均比较低。但是DCRF的指标结果高是由于DCRF将大部分都预测为L5，而L5的在Weebit中占比较高。RAZ的结果如表8所示，ARI的指标结果表现最好，虽然FRE和DCRF两项指标结果较低，但是混淆矩阵结果显示各个公式的预测结果都更加准确。总体来说公式法在Weebit语料上和RAZ语料上的效果都不太好。随着大规模标记语料的产生和技术的发展，为了解决上述局限，研究者们开始采用机器学习方法。

Petersen and Ostendorf (2009)拓展和提升了这个模型，通过引入未标注的负训练数据来解决分类器泛化以处理可能包含其他级别的新数据的问题；探索了句法特征在多大程度上优于传统词汇特征；探索了在训练数据有限的情况下回归模型作为二元模型框架的可用性。最后实验结果的分类错误率如表9所示。我们将这两篇论文提出的方法和公式法Flesch-Kincaid作比较，我们可以看出来SVM的效果比公式法Flesch-Kincaid和Lexile好很多，分类错误率小很多。Feng et al. (2010)对Weekly Reader语料进行预处理之后保留下了1433篇合格的文章作为语料集，在Schwarm and Ostendorf (2005)提出的特征的基础上进一步探究了实体关系等语篇特征、语言建模特征、句法解析特征、基于词性标注（POS）的特征对于文本阅读分级的影响。实验发现POS特征相比其他特征更加有用；动词相比其他的词性与文本难度的关系更大；语篇特征并不是很有用，可能是因为低年级的文本存在相对低的复杂度，在未来的工作中我们可以探索一下高年级的阅读分级中语篇关系是否可以发挥明显的作用。

Vajjala and Detmar (2012)基于他们创建的分级语料库WeeBit训练了一个多层感知机分类器，作者一共手动设计了46个涉及到词汇特征、句法特征和句长等定量特征在内的三类特征。最终在测试集上得到了93.3%的准确率。Xia et al. (2016)使用了与Vajjala and Detmar (2012)相似的文本特征，不同的是添加了语言模型和篇章特征，最终使用SVM分类器在五倍交叉验证的Weebit语料集上取得80.3%的准确率。并且作者在构建的CEFR语料库Xia et al. (2016)上测试了训练过的模型取得了23.3%的准确率，并没有超过大多数的分类器基线。主要原因是语料库类型的不同，CEFR分级语料库是为英语学习者量身定做的语料库，主要是针对具有不同英文理解力的外国成年人，而WeeBit是针对不同年龄段的儿童，模型在可转移性这一步的尝试中不太成功。Vajjala and Ivana (2018)针对OneStopEnglish语料手动提取了155个包含n-grams、词性、心理语言学、句法、话语和传统特征在内的六类特征，最终利用SMO带有内核的分类器实现了78.13%的准确率。

以上方法都是从文本分类的角度出发的，Ma et al. (2012)没有采用本文分类的方法做而是排序的方法，在细粒度的阅读分级任务上，排序比分类表现得更好一些。作者处理了36篇Founts And Pinnell(Fountas and Pinnell, 1996) 的36篇书籍token，级别从A到L，对应的年级从幼儿园到二年级。在处理数据的时候还利用工具获取了文章的排版、注释、插画等视觉特征。论文首先将书籍排序，然后根据排序的结果预测书籍的级别。第一步是从每本书中提取特征，然后训练一个排名模型来尽可能减少每对书籍之间的误差。在第二步测试阶段使用这个SVMrank模型(Cai and Cao, 2017)对所有的书籍包括训练书籍和需要测试的书籍一起进行排序。根据排序的结果来计算距离这个书上下三本的书籍的难度，因为难度接近。实验中作者设计了两类特征，第一类是文本特征，第二类是视觉特征。也分别训练了排名、分类和回归模型，结果表明在排名模型上普遍表现比较好。并且单单使用文本特征或者单单使用视觉特征时，视觉特征要比文本特征表现得好很多。这说明对于针对偏小年龄的儿童的文学读物的分级，视觉上的信息是很重要的，是我们不能忽视的重要信息。

在中文上也有一些利用传统机器学习方法的工作，Lee et al. (2017)等人提出了一个测试中

文文本的阅读难度的网络系统。该系统对输入的文本进行分割、词性标注和依存分析，然后确定词汇和语法结构的难度。另外，系统还会高亮那些为了满足特定难度要求而必须改写或者简化的词汇或者短语。

3.3 基于深度学习的阅读分级方法

随着深度学习的火爆和优良的效果，深度学习在很多领域都得到了广泛的应用并取得了很好的效果，在阅读分级领域也做出了不少的工作。

Sun et al. (2020)构建了表3所示的Lexile语料集，在语料集上分别用卷积神经网络（CNN）(Krizhevsky et al., 2012)、递归神经网络（RNN）(Jia et al., 2014)、门控循环单元（GRU）(Cho et al., 2014)、循环神经网络（BiLSTM）(Hochreiter and Schmidhuber , 1997)、双向GRU（BiGRU）、BiGRU + 注意力机制（Attention）(Bahdanau et al., 2014)等深度学习模型进行训练和测试，其中BiGRU+Attention的效果最好，在测试集上达到了90.7%的准确率，并且Attention机制的添加可以高亮重点字词，这样有助于学生和老师的教学。但是该实验存在不少漏洞，主要表现在Lexile语料集的样本分布不均衡，并且语料集过小。

Martinc et al. (2019)在2.3节提到的WeeBit、OneStopEnglish、Newsela和Slovenian语料上训练了BiLSTM(Kim , 2014)、HAN(Yang et al., 2016)和BERT(Devlin et al., 2018)三种神经网络模型。在WeeBit语料上BERT实现了最好的性能，HAN模型的性能表现得最差。但是在OneStopEnglish和Newsela语料中BERT表现最差，很可能是因为这两个语料集在不同类之间存在着语义相似性。因为BERT是预训练模型，在分类过程中往往更依赖于语义而不是结构，所以在具有明显语义差异的语料集上表现得更好，例如Weebit和Slovenian。而在语义差异不明显的OneStopEnglish和Newsela语料上表现的不好。在OneStopEnglish语料上表现最好的是HAN分类器，准确率达到了78.1%。在Newsela语料上表现最好的也是HAN分类器，准确率到达了80.4%。在Slovenian语料上，所有分类器的准确率都不高，但结合混淆矩阵等其他因素，BiLSTM的表现好于其他，准确率为51.27%，这里的低准确率可能是因为有十三个等级。并且Slovenian语料库是包含了16个不同学科的教科书，从文学、数学到音乐，主题跨度非常的宽泛，这也会给分类带来很大的难度，但是也能给我们提示，我们需要将主题维度考虑到阅读分级的范畴。

Deutsch et al. (2020)探究了结合语言特征的深度学习模型能否进一步提高阅读分级的分类模型的性能。作者将深度学习模型如Transformer(Vaswani et al., 2017)、HAN和CNN的输出作为特征本身与语言特征结合，进一步输入到其他模型中，如SVM分类器，即SVM+HAN/HAN/CNN+linguistic模型。在Newsela语料库上的训练结果显示，HAN分类模型表现最好，虽然语言特征能够提高某些模型的性能，但是结合语言特征的模型没有HAN单独的模型表现得好。在Weebit语料中，Transformer表现得最好，结合语言特征的模型没有单独的Transformer表现得好。与Newsela语料库不同，WeeBit语料库的主题与难度之间存在很强的相关性，提取该主题和语义内容被认为是transformer的一个特殊优势(Martinc et al., 2019)，从而改进了该语料库的结果。在本文使用的方法中可以看出语言特征并不能提供实质性的帮助。但是语言特征是否能从其他角度加以利用提供自动文本分级的任务帮助还需要我们共同去研究。最后如表10所示，我们对以上提到的针对Weebit、Newsela和OnestopEnglish三个语料集利用深度学习的几种方法的实验结果做了一个总结。

同时，基于深度学习的阅读分级也在其他语言上有一定的进展。以俄语为例，Glazkova et al. (2020)从小说文本出发，选取了可读性公式、情感、词汇、语法、故事复杂性和发布属性等文本特征，比较了不同类型的特征在基于年龄的小说文本分类任务中的有效性。该论文使用TF-IDF向量(Shi et al., 2009)作为基准，将文本的TF-IDF向量与某种类型的特征对应的向量连接。由于书籍语料是很长的文本，而模型之一的RuBERT[9]只能处理有限长度的文本，所以论文中用256个token的片段来训练神经网络和TF-IFD向量。作者分别使用结合不同的特征的Random Forest(Breiman , 2001)等分类器和深度学习模型RuBRET、CNN进行实验来对文本进行训练，最后的结果显示使用标注属性（摘要和年龄分级）能显著提高分类的性能。在中文上，Liu et al. (2017)构建了一个包含初等难度79篇、中等难度122篇和高等难度144篇的中文教学材料语料集，使用CNN和LSTM以及CNN和LSTM结合的方法在该语料集上进行训练，实验表明CNN和LSTM结合的效果最好，CNN的输出作为LSTM的输入，既可以获取短距离的特

[9]https://github.com/vlarine/ruberta

方法	测评方式	Weebit	Newsela	OneStopEnglish
BERT	准确率	83.9%	58.1%	59.0%
BERT	精确率	84.6%	58.0%	60.4%
BERT	召回率	83.9%	58.1%	59.0%
BERT	F1值	**84.0%**	57.6%	57.7%
HAN	准确率	77.0%	80.5%	79.0%
HAN	精确率	77.6%	80.7%	81.2%
HAN	召回率	77.0%	80.5%	79.0%
HAN	F1值	77.0%	**80.4%**	**79.0%**
BiLSTM	准确率	78.2	69.4%	72.1%
BiLSTM	精确率	78.7%	71.6%	75.3%
BiLSTM	召回率	78.2%	69.4%	72.1%
BiLSTM	F1值	78.2%	70.2%	72.0%
Vec2Read	准确率	N/A	52.7%	N/A
Transformer	F1值	83.9%	54.4%	N/A
CNN	F1值	78.6%	33.8%	N/A
SVM +HAN+linguistic	F1值	83.4%	80.1%	N/A
SVM +transformer+linguistic	F1值	76.8%	80.1%	N/A
SVM +CNN+linguistic	F1值	79.2%	72.3%	N/A

Table 10: WeeBit, Newsela和OneStopEnglish在多种深度学习模型上的结果比较。N/A表示论文没有提供在语料集上的测试结果，黑体表示不同语料在F1测评方式上的最高值。

征，又可以获取长距离的依赖，准确率达到了87.1%。

同时在多语言阅读分级工作也有一些进展。Azpiazu et al. (2019)引入了Vec2Read，一个针对多语言阅读分级的多注意力循环神经网络结构，利用多注意力机制允许系统考虑单词和对阅读水平影响最大的句子文本,对于分级有一定的可解释性。该论文在包含英语、荷兰语等七种语言的语料集上展开实验，如Newsela语料、SimpleWiki语料等，相较于公式方法和当时的方法分级效果有一定的提升。

4 未来研究方向

阅读分级虽然在欧美国家拥有了相对成熟和应用广泛的阅读分级体系，但是在分级标准和分级资源上仍存在着不少仍需完善和改进的地方，在其他国家，分级标准和分级资源还都处于初级阶段，甚至还未引起足够的重视；在科研工作上，虽然现有的工作已经迈出了重要的一步，对自动阅读分级技术有了一定的探索，但是方法和语料资源上还远远满足不了相关工作者的需求，这都有待我们进一步去开发和研究。在这里，我们作为计算机相关研究人员，希望从科研的角度出发，针对阅读分级的未来的发展方向给出我们的一些建议和思考。

4.1 低资源阅读分级方法

首先我们将目光聚焦于跨语言阅读分级上，如何用拥有高资源的英语语言的阅读分级成果去指导低资源语言阅读分级，这是我们可以着手解决的问题。跨语言学习近年来一直是比较火热的话题，也取得了不小的进展。跨语言学习旨在抽取语言不变性特征来帮助从高资源语言到低资源语言的跨语言学习(Litschko et al., 2018; Kondratyuk and Straka , 2019)，实现知识的迁移。我们假设这些语言不变特征在阅读分级任务中也存在，虽然是不同的语言，但是同一难度等级的文本极有可能描述的事件和主旨思想是相同的，用词造句方面也会有异曲同工之处。为了验证我们的假设，为了将英语阅读分级的知识迁移到低资源语言中，我们目前正在探索现有的跨语言的方法，例如跨语言预训练模型(Lample and Conneau , 2019; Conneau et al., 2019)和对抗模型(Chen et al., 2016)这些，以指导中文等低资源语言的阅读分级。多任务学习(Crawshaw , 2020)和知识蒸馏(Wu et al., 2020)等方法都有被用于低资源跨语言任务的学习，这也是我们可以尝试的方向。

4.2 阅读分级的语料构建

总体上来说现阶段用于阅读分级研究的语料的发展是不够完善的，现存在训练和测试数据不够充分、语料本身质量不过关、语料与语料之间不够匹配等问题。在这里我们提出几点相关的建议。

第一，提高语料库的质量，主要指语料应该具有信服力的分级依据。另外也应该扩充数据，可以利用现在火热的数据增强技术(Wei and Zou , 2019)。同时需要多收集一些额外的不同种类的阅读分级材料用于测试，这样才能凸显自动阅读分级模型的泛化性。

第二，我们从细粒度的角度出发，关注影响阅读分级的不同层面的因素。在情感分析任务中有一类针对多个维度的进行分析的任务，比如在商品评价语料中关注商品的质量与服务(Lin et al., 2021; Lin et al., 2021)。那么在阅读分级任务的时候我们也可以将语言难度、主题难度和思想难度等层面分开来进行研究，这样也会更加具有可解释性，随之而来的是在语料库的构建上提出更高的要求。

第三，现有的语料集普遍没有考虑到文本的排版信息，研究中使用到的语料基本上是经过预处理的文本数据，我们是否可以考虑构建语料的时候利用排版信息和插图信息，相信对于指导阅读分级的研究有所助益的，不过构建上会付出相对大的成本。

4.3 自动阅读分级方法研究

随着技术的发展，我们如何更好的利用现有的资源和方法应用到阅读分级任务，以及如何去设计更符合阅读分级任务的模型是我们需要不断去思考和行动的。

我们可以结合主题模型去做阅读分级任务，可以假设虽然语言不同，但是相同难度的阅读读物之间的主题分布存在一定的一致性。在其他任务中，主题模型也经常被融合去辅助主任务，例如Peinelt et al. (2020)提出的针对特定领域结合主题模型和BERT模型去解决语义相似度计算问题。另外考虑到不少阅读书籍存在文本过长的问题，我们可以尝试先对文本进行话语分割再进行分级，或者直接抽取摘要之后再进行文本的分级，或者我们分别从句子、段落、篇章的角度去进行多任务学习，这些方法是否实用存疑。再者我们可以利用知识增强，知识增强运用在了自然语言处理的很多领域例如文本生成(Yu et al., 2020)，并取得了不错的效果,知识增强有助于模型理解背景知识和篇章信息，对于文本分级应该会起到一定的作用。

4.4 阅读分级的可延伸性研究

对于教育图书来说很重要的一点是新词、新句式的掌握，我们希望在分级的同时能标注出这些新词、新句式，以帮助进行学习，同时在阅读分级任务上实现更高的可解释性。我们需要去找到那些适合该阶段的学生学习的新字词、新句式，并且标注出来提供更好的可解释性。原则上可以当做一个新的任务和阅读分级任务结合，也可以用现在流行的attention方法给予这些词、句更大的权重以此作为评判依据。另外关于难度分级问题我们也可以扩展为阅读分级的基础上进行可控难度文本生成、可控难度问题生成等任务，为适宜阶段的阅读者创造更多的读物和习题。

5 总结

阅读分级是教育领域广受关注的话题，教育需要科技为之赋能是社会一直在倡导的话题，用科技的手段去实现阅读分级也是教育界和学界一直想做的事情。随着技术的发展，自动阅读分级技术研究越来越受到相关研究者的关注，本文充分总结了阅读分级相关的体系、资源和技术，重点探讨了自动阅读分级技术的现状和未来发展方向。自动阅读分级技术的发展主要面临两个方面的挑战：第一是语料资源的缺乏与不健全，特别体现在非英文的语言中；第二是分级方法亟待改进，公式法很难考虑到主旨思想层面，机器学习方法和深度学习方法是可以改进的方向，在此基础上分级的可解释性是需要延伸的部分。

致谢

感谢各位盲审评审人对于本篇论文的认可和意见，这对我们的改进有很大的助益。另外感谢阅读分级的相关研究工作者的支持，每当发邮件询问相关数据集和论文工作时，都得到了积极的回应。最后感谢李老师对于我们的指导，在最初想法的提出、调研以及之后研究探索的过程中都给予我们很多的帮助和耐心。

参考文献

Sarah E.Schwarm, and Mari Ostendorf. 2005. *Reading level assessment using support vector machines and statistical language models. Association for Computational Linguistics*,pages 523–530.

Sarah E.Petersen and Mari Ostendorf. 2009. *A machine learning approach to reading level assessment. Computer speech language*,23(1):89–106.

Regina Barzilay and Noemie Elhadad. 2003. *Sentence alignment for monolingual comparable corpora. In EMNLP '03 Proceedings of the 2003 conference on Empirical methods in natural language processing*,pages 25-32.

Rohit Kate ,Xiaoqiang Luo ,Siddharth Patwardhan ,Martin Franz ,Radu Franz ,Raymond Mooney ,Salim Roukos and Chris Welty. 2010. *Learning to Predict Readability using Diverse Linguistic Features. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).*

Lijun Feng and Martin Jansche and Huener fauth, Matt and Elhadad. 2010. *A Comparison of Features for Automatic Readability Assessment. Coling 2010: Posters*,pages 276-284.

Sowmya Vajjala and Meurers Detmar. 2012. *On improving the accuracy of readability classification using in sights from second language acquisition. Association for Computational Linguistics*,pages 163–173.

Menglin Xia, Kochmar Ekaterina and Briscoe Ted. 2016. *Text readability assessment for second language learners. Association for Computational Linguistics*,pages 12–22.

Sowmya Vajjala and Lucic Ivana. 2018. *Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. Association for Computational Linguistics*,pages 297–304.

Yi Ma, Fosler-Lussier Eric, and Lofthus Robert. 2012. *Ranking-Based Readability Assessment for Early Primary Childrens Literature. Association for Computational Linguistics.*

John Lee, Meichun Liu, Chun Yin Lam, Tak Lau, Bing Li and Keying Li. 2017. *Automatic Difficulty Assessment for Chinese Texts. The Companion Volume of the IJCNLP 2017 Proceedings:System Demonstrations*,pages 45–48.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. *Supervised and Unsupervised Neural Approaches toText Readability. ArXiv:1907.11779 [Cs].*

Nicole Peinelt and Dong Nguyen and Maria Liakata. 2020. *tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Yuxuan Sun ,Keying Chen ,Lin Sun and Chenlu Hu. 2020. *Attention-based Deep Learning Model for Text Readability Evaluation. T2020 International Joint Conference on Neural Networks (IJCNN)*,pages 1-8.

Federico Bianchi ,Silvia Terragni ,Dirk Hovy ,Debora Nozza and Elisabetta Fersini. 2020. *Cross-lingual Contextualized Topic Models with Zero-shot Learning. arXiv e-prints.*

Hao Liu, Si Li, Jianbo Zhao, Zuyi Bao, Xiaopeng Bai 2017. *Chinese teaching material readability assessment with contextual information. 2017 International Conference on Asian Language Processing (IALP). IEEE*,pages 66–69.

Tovly Deutsch and Masoud Jasbi and Stuart Shieber. 2020. *Linguistic Features for Readability Assessment. Association for Computational Linguistics*,Seattle, WA, USA.

Wenhao Yu and Chenguang Zhu and Zaitang Liand Zhiting Hu and Qingyun Wang and Heng Jiand Meng Jiang. 2020. *A Survey of Knowledge-Enhanced Text Generation. arXiv e-prints.*

Madrazo Azpiazu ,Ion Madrazo and Maria SoledadPera. 2020. *Is Cross-lingual Readability Assessment Possible. Journal of the Association for Information Science and Technology*,71(6): 644–656.

Madrazo Azpiazu ,Ion Madrazo and Maria SoledadPera. 2019. *Multiattentive Recurrent Neural Network Architecture for Multilingual Readability Assessment. Transactions of the Association for Computational Linguistics*,7:421–436.

Yoon Kim. 2014. *Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.*

Alexis Conneau ,Douwe Kiela ,Holger Schwenk ,Loic Barrault and Antoine Bordes. 2017. *Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.*

Zichao Yang ,Diyi Yang ,Chris Dyera ,Xiaodong He ,Alex Smola ,Eduard Hovy. 2016. *Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,pages 1480–1489.

Jacob Devlin ,Ming-Wei Chang ,Kenton Lee and Kristina Toutanova. 2018. *Bert:Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Ashish Vaswani ,Noam Shazeer ,Niki Parmar ,Jakob Uszkoreit ,Lion Jones ,Aidan N.Gomez ,Lukasz Kaiser and lllia Polosukhin. 2017. *Attention Is All You Need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing .*

Anna Glazkova,Yury Egorov and Maksim Glazkov. 2020. *A Comparative Study of Feature Types for Age-Based Text Classification. eprint arXiv:2009.11898.*

Robert Litschko, Goran Glava˘s, Simone Paolo Ponzetto and Ivan Vuli . 2020. *Unsupervised cross-lingual information retrieval using monolingual data only. In The 41st International ACM SIGIR Con-ference on Research Development in Information Retrieval*,pages 1253–1256.

Dan Kondratyuk and Milan Straka. 2019. *75 Languages, 1 model: Parsing universal dependencies universally. 2019 Conference on Empirical Methods in Natural Language Processing.*

Alex Krizhevsky ,llya Sutskever and Geoffrey E.Hinton. 2012. *ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM.*

Guillaume Lample and Alexis Conneau. 2019. *Cross-lingual Language Model Pretraining. arXiv:1901.0729.*

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2019. *Unsupervised Cross-lingual Representation Learning at Scal. arXiv:1911.02116.*

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Weinberger, Claire Cardie. 2016. *Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. arXiv:1606.01614.*

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Jian-Guang Lou, Biqing Huang. 2020. *Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. arXiv:2004.12440.*

Xianggao Cai,Shujin Cao. 2020. *A Keyword Extraction Method Based on Learning to Rank[C]. International Conference on Semantics*,2017:194-197.

Michael Crawshaw. 2020. *Multi-Task Learning with Deep Neural Networks: A Survey. arXiv:2009.09796.*

Y Lin，C Wang，H Song，Y Li. 2021. *Multi-head Self-attention Transformation Networks for Aspect-based Sentiment Analysis. IEEE Access PP.99(2021):1-1.*

P Bai，Y Xia，Y Xia . 2021. *Fusing Knowledge and Aspect Sentiment for Explainable Recommendation. 10.1109/ACCESS.2020.3012347.*

Jason Wei，K Zou. 2019. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[J].*

Kyunghyun Cho,Bart van Merriënboer,Caglar Gulcehre, Dzmitry Bahdanau,Fethi Bougares,Holger Schwenk, Yoshua Bengio. 2014. *Learning phrase representations using rnn encoder-decoder for statistical machine translation. Association for Computational Linguistics.*

L Breiman. 2001. *Random forest[J]n. Machine Learning*,45:5-32.

Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of TGE International Conference on Learning Representations*.

YuKang Jia, Zhicheng Wu,Yanyan Xu, Dengfeng Ke and Kaile Su. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *computer science*.

CY Shi，XU Chao-Jun，XJ Yang . 2009. Study of TFIDF algorithm[J]. *Journal of Computer Applications*.

Sepp Hochreiter and Jurgen Schmidhuber 1997. Long Short-Term Memory. *Neural Computation*9.8(1997):1735-1780.

RJ Senter and Edgar A Smith. 1967. *Automated readability index. CINCINNATI UNIV OH*.

AJ Stenner ,Ivan Horabin ,Dean R Smith and Malbert Smith. 1988. *The lexile framework. Durham, NC: MetaMetrics*.

John B Carroll. 1972. *new word frequency book.Elementary English*，49(7):1070–1074.

JPeter Kincaid ,Robert P Fishburne Jr,Richard LRogers, and Brad S Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Institute for Simulation and Training*.

Robert Gunning. 1952. *The technique of clear writing. McGraw-Hill, New York*.

EdgarA Smith and R.J. Senter. 1967. *Automated readability index. AMRL-TR. Aerospace MedicalResearch Laboratories(US)*,pages 1–14.

Edgar Dale and Jeanne S Chall. 1948. *A formula for predicting readability. Educational research bulletin*,pages 37–54.

Mc Laughlin and G Harry. 1990. *A readability formula for short passages. J. of Reading*,May 1990, 594-597.

E Fry. 1990. *A formula for predicting readability. J. of Reading*,May 1990, 594-597.

W.A Gale and G Sampson. 1995. *Good-Turing frequency estimation without tears. J. of Quant. Linguistics*,v. 2, 217-237.

Irene C.Fountas and Gay Su Pinnell. 1996. *Guided Reading: Good First Teaching for All Children. Heinemann, 361 Hanover Street, Portsmouth*,NH 03801-3912 (32.50).

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Read-ability revisited: The new Dale-Chall readability for-mula. Brookline Books*.

Weekly Reader. 2004. *http://www.weeklyreader.com. Accessed July, 2004*.

Wikimedia. 2018. *Simplification guidelines for simple wikipedia. https://simple.wikipedia.org /wiki/Wikipedia:Simple English Wikipedia*.

吴思远, 蔡建永, 于东等. 2018. 文本可读性的自动分析研究综述*[J]*. 中文信息学报,32(12):1-10.

罗德红,余婧. 2012. 科学与价值:我国儿童汉语分级阅读研究的问题与展望*[J]*. 出版广角,(09):62-64.

罗德红,余婧. 2013. 儿童分级阅读研究的中美对比分析*[J]*. 图书馆,2013(02):34-37.

李云飞, 袁曦临. 2019. 国外儿童分级阅读研究现状述评*[J]*. 图书馆杂志,38(03):12-21.

王进, 周慧, 罗国峰等. 2017. 王进. 计算机时代,(8).

附录A.阅读分级体系对照

这里我们对数字体系、字母体系和年级体系的几个代表分级方式进行对照，对照表如表11所示。不同的体系中级别与级别之间存在一定的交叉，比如850L的Lexile难度的图书既适合四年级的同学，也适合五年级的同学；比如RAZ的Q等级和S等级都适合三年级的儿童，但是Q适合7-9岁，R适合8-9岁。这是由于各个级别大多是针对文本的难度进行计算，而参照年级的话，会考虑到了儿童阅读水平的发展特点，同一年级的儿童不一定阅读能力相同。另外虽然RR和DRA都属于数字体系且拥有相似的分值，但是相同的RAZ等级的图书在RR和DRA都不一定对应相同的分值，因为RR和DRA具有不同的分级标准。这些问题是不同体系之间进行参照的时候无法避免的问题，对照表只是提供一个参考，并不具备绝对的一致性。

RAZ	**年龄**	**年级**	**Lexile**	**RR**	**DRA**	**AR**
aa-A	4-6	K	BR-70	1	A-1	.1-1.5
B	4-6	K	BR-70	2	2	.1-1.5
C	4-6	K	BR-70	3-4	3-4	.1-1.5
D	4-7	1	80-450	5-6	6	1.6-3.3
E	6-7	1	70-450	7-8	8	1.6-3.3
F	6-7	1	80-450	9-10	10	1.6-3.3
G	6-7	1	80-450	11-12	12	1.6-3.3
H	6-7	1	80-450	13-14	14	1.6-3.3
I	6-7	1	80-450	15-16	16	1.6-3.3
J	6-8	2	451-500	17	18	2.8-4.2
K	7-8	2	451-550	17	18	2.8-4.2
L	7-8	2	501-550	18	20	2.8-4.2
M	7-8	2	551-600	19	24	2.8-4.2
N	7-8	2	551-650	20	28	2.8-4.2
O	7-8	2	601-650	20	28	2.8-4.2
P	7-8	2	601-650	28	28	2.8-4.2
Q	7-9	3	651-690	30	30	3.9-5.1
R	8-9	3	651-730	30	30	3.9-5.1
S	8-9	3	691-770	34	34	3.9-5.1
T	8-9	3	731-770	38	38	3.9-5.1
U	8-11	4	771-800	40	40	5.0-6.1
V	9-11	4	771-830	40	40	5.0-6.1
W	9-11	4	801-860	40	40	5.0-6.1
X	9-11	5	831-860	40	40	6.0-7.0
Y	9-11	5	861-890	40	40	6.0-7.0
Z	9-11	5	891-980	N/A	50	6.0-7.0
Z1	9-11+	5+	920-1070	N/A	60	7.0-8.0
Z2	9-11+	5+	980-1140	N/A	70+	8.0-9.0

Table 11: 各阅读体系之间的对照表，K表示幼儿园,N/A表示没有对应的级别。