文章编号：1003-0077(2018)12-0001-10

文本可读性的自动分析研究综述

吴思远[1,2]，蔡建永[2,3]，于 东[1]，江 新[2]

（1. 北京语言大学 信息科学学院，北京 100083；
2. 北京语言大学 对外汉语研究中心，北京 100083；
3. 北京语言大学 汉语速成学院，北京 100083）

摘 要：文本可读性问题最初由教育学家提出，初衷是辅助教师为语言学习者推荐适合其阅读水平的文本。随着计算机技术的发展及网页文本的涌现，对文本进行可读性分析有了更加丰富的技术手段和应用场景。该文对可读性自动分析的相关研究进行了梳理，将可读性自动分析的方法总结为公式法、分类法和排序法三类；然后进一步介绍了可读性自动分析中的两项重要内容：文本特征的选择和数据集的使用；最后对可读性研究的发展方向进行展望。

关键词：文本可读性；可读性分析；特征提取

中图分类号：TP391　　**文献标识码**：A

A Survey on the Automatic Text Readability Measures

WU Siyuan[1,2], CAI Jianyong[2,3], YU Dong[1], JIANG Xin[2]

(1. College of Information Science, Beijing Language and Culture University, Beijing 100083, China;
2. Center for Studies of Chinese as a Second Language, Beijing Language and Culture University, Beijing 100083, China;
3. College of Intensive Chinese Studies, Beijing Language and Culture University, Beijing 100083, China)

Abstract: The concept of readability is originally proposed by educators to assist the selection of suitable reading materials for learners. This paper surveys the existing works on automatic text readability measures, and summarized three types of methods: formula-based method, classification method and ranking method. This paper also outlines the databases and the extracted features in the literature. And finally, the future developments of the automatic readability research is provided.

Keywords: text readability; readability analysis; feature selection

0 引言

阅读是人类获取信息和知识的重要途径。难度适当的阅读材料不仅可以使阅读过程顺利进行，还可以提升读者的阅读能力。相应地，超出或低于读者水平的文本不仅会影响读者的阅读体验，还可能对基本文本信息的提取造成阻碍[1]。随之而来的问题是：是什么导致了文本之间的难度差距？影响文本难度的核心特征是什么？文本难度是否可以进行度量？是否可以借助计算机对文本难度进行自动分析？学者们从不同角度对文本难度问题进行了探讨，这些研究后来被统称为可读性（readability）研究[2]。

可读性研究是语言学和心理学领域的重要课题之一，对文本进行可读性分析是可读性研究的核心。可读性分析的任务是，给定一篇文本，通过对文本进行分析，给出该文本的难度值或判断该文本适合哪一水平的读者。最初的可读性分析主要是请有经验的专家或教师对文本难度进行主观评定，这种方法具有很强的主观性，评定者的标准不同，目的不同，评定结果也往往不同。

文本可读性的自动分析可以追溯到 20 世纪 20 年代[3]。所谓可读性的自动分析，就是对文本难度

收稿日期：2018-09-29　定稿日期：2018-10-29
基金项目：国家社会科学基金(17ZDA305)；中央高校基本科研业务费专项资金资助项目(17PT05)

进行定量、自动的评估与分析，是一种预测性的手段，具有客观性和经济性的优点。可读性的自动分析有很多应用场景。在教育领域，评估文本难度可以帮助教师为学习者选择合适的阅读材料[4]，为教材编写提供科学依据[5]，对阅读测试、课程规划有一定参考价值[6]。在自然语言处理领域，计算机科学家把可读性分析应用于智能改编[7]、作文自动评分[8]等任务；或借助可读性自动分析提炼和归纳源文档的主要内容，对自动文摘的质量进行评估[9]；或通过分析网页文本，对用户的阅读兴趣和搜索习惯进行预测和推荐[10]。

根据分析思路和关键技术的不同，我们将可读性的自动分析方法分为公式法、分类法、排序法三类。①公式法：通过建立线性方程的方式，把文本难度最相关的一些语言特征作为变量来预测文本的难度值，使用的特征一般为浅层的语言特征，如词长、句长等；②分类法：研究者把文本难度的预测作为分类任务，从不同等级的文本中学习一系列具有区别性的文本特征，构造分类模型，输入没有标签的新文本后，分类模型根据学习的结果估计文本的难度等级；③排序法：构建比较器或人工标注得到文本的两两相对难度，对文本进行排序，得到按难度排序的文本集合，缺点是不能给出具体的难度值或难度等级。

本文主要梳理已有的可读性研究，组织如下：第1节总结可读性自动分析的主要方法和基本技术；第2节对可读性分析中的重要环节——文本特征选择和现有数据资源进行梳理；第3节回顾汉语文本的可读性研究；最后一节对未来的可读性研究进行展望。

1 可读性自动分析的主要方法

1.1 基于可读性公式的方法

所谓可读性公式，就是针对某种阅读文本，将影响阅读难度的、可进行量化的文本因素综合起来，制定的一个评估文本难易程度的公式[11]。它通常给出数值结果作为文本难度分数。

可读性公式的构建主要包括两方面的内容：①与可读性级别密切相关的文本因素；②各因素与可读性级别之间的函数关系。可读性公式以学生的阅读理解成绩作为文本难度，在客观数据的基础上，利用相关性分析确定影响文本难度的主要因素，根据因变量（文本可读性）与自变量（文本各因素）之间的关系，拟合文本可读性公式。

可读性公式假设因变量与自变量线性相关，其模型被定义为式(1)。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

在20世纪20年代，Vogel等[3]首次使用回归方程的方式，将多个文本特征纳入可读性公式，该研究方法对后来的可读性公式研究影响深远。20世纪50年代之后可读性公式的构建逐渐兴盛，到80年代，超过200个可读性公式被构建出来并广泛应用于出版社、研究所、医疗说明、法律、保险等行业[12]。美国教育部和国防部也建立了以可读性公式为中心的可读性分析体系，用来对教育体系中使用的教材、国家政策中使用的文件进行评估和定级。英文中几个较为权威的可读性公式如表1所示。

表1 具有代表性的英文可读性公式

可读性公式	备注
The Dale-Chall Formula(1948年)[13] RL=(0.049×SL)+(0.1579×DW)+3.636 5	以3 000常用词表为基础
The Gunning FOG Index(1952年)[14] RL= 0.4(SL+HW)	适用于小学高年级和中学
SMOG Grading (1969年)[15] $RL=1.0430\sqrt{WL\geqslant 3\times\frac{30}{sent}+3}$	可以通过拟合曲线转化为相应等级
The FORCAST Formula(1973年)[16] RL=20−((WL=1)/(150 words)/10)	美国陆军技术手册可读性测量公式
Flesch-Kincaid Formula(1975年)[17] RL=0.39×SL+11.8×WL−15.59	美国国防部可读性测量公式

注：RL(Reading Level)：可读性级别；SL：平均句长，即平均每个句子的平均单词数；DW：不在3 000常用词表的非常用词的数量；HW指文本中难词的比例；WL：平均单词长度；sent：句子数；150 words：在150词表里的词数。

使用可读性公式评估文本的难易程度具有客观性、简便性和经济性等特点。使用公式可以快速地获得文本难度的分析结果，比较实用。但是，影响文本难度的因素很多，可读性公式只能考虑有限的可计量的文本特征，无法把所有影响文本可读性的变量如语法语义、句法、篇章等考虑在内[10,18]，因此可读性公式的效度一直颇受争议[19]。不可否认的是，可读性公式法，是研究者试图针对特定阅读人群，通过量化手段客观地评估文本阅读难度的方法。可读性公式的构建是传统性公式的重要内容，也为后来的可读性研究奠定了基础。

1.2　基于分类的方法

在机器学习中，分类被定义为：给定一组训练实例 $X_1, X_2, \cdots, X_n$，每个训练实例有类别标签。通过学习有标签的训练实例，训练模型 $f(X \rightarrow Y)$ 从而对新的实例给出类别预测[20]。基于分类的可读性分析方法把可读性评估任务当成分类任务，通过学习一系列具有区别性的语言特征，训练分类模型，以确定未知文本的可读性级别不同可读性级别的语料中学习一系列具有区别性的语言特征，构建分类模型，分类模型通过对未知文本特征进行分析，判别该文本是否属于某一难度级别。

大量研究表明，除了浅层的句长、词长等，基于分类方法的可读性自动分析能考虑更多的语言特征，如词汇熟悉度、句法复杂度等，评估结果比可读性公式准确，而且在区分高难度文本上有显著优势[21-23]。研究常使用的分类模型有 N 元词串隶属度模型和支持向量机。

1.2.1　N 元词串隶属度模型

N 元词串隶属度模型是一种基于词概率的统计语言模型。该方法把文本当成一连串的字符序列，并假定文本的可读性级别和文本的用词有关且文本的可读性级别互相独立。在训练阶段，该方法首先根据训练样本数据，统计每个 N 元词串隶属于每个级别的概率模型。在预测阶段，对于一个未知级别文本 T，计算其属于所有级别的隶属度，取隶属度最大的为与文本相匹配的难度等级，如式(2)所示。

$$L(G_i \mid T) = \arg\max_{G} \sum_{w \in T} C(w) \log P(w \mid G_i) \tag{2}$$

给定某一级别的概率模型 G_i，w 为文本 T 的用词，$C(w)$ 为词汇 w 在 T 中出现的频次。

不同难度的文本词汇的使用和分布不同，文本词汇信息能有效预测文本的难易程度[22]。Si 等[24]首次在文本可读性分析上使用一元词串隶属度模型。该研究在 3 个等级共 91 篇文本的数据集上训练了一元模型，并和句长一起进行文本可读性预测，模型准确率为 75.4%，而 Flesch-Kincaid 公式[17]的准确率仅为 21.3%。实验表明，使用该模型预测文本难易度比仅使用句长、词长特征的可读性公式表现更好。Collins-Thompson 等[25]收集了 12 个难度等级共 550 篇网页文本来训练概率模型，该研究通过相邻等级文本之间的关系，使用 Good-Turing 平滑算法对预测文本出现在某一等级的概率进行估计，模型的预测结果与原等级的相关性最高为 0.93。

通过文本的词汇信息判断文本难度等级的统计语言模型比可读性公式的准确率更高。其次，N 元词串隶属度模型在网页文本和短文本上表现较好，而可读性公式一般要求文本长度大于 200 词。

1.2.2　支持向量机

支持向量机是 Cortes 等[26]提出的基于结构风险最小化原理的统计学习理论，主要应用于分类问题。

Schwarm[27]使用支持向量机进行可读性分析。训练过程中使用了从 N 元模型中学习到的文本特征，以及一些词法、句法特征。该模型评估结果的准确率在 79%到 94.5%之间，而传统的 Flesch-Kincaid 可读性公式的准确率则在 21%到 41%之间。可见，支持向量机分类器的方法要明显优于传统的评估方法。该研究在低年级、短文本的分类中显示出了良好的性能，但对较高等级的文本却难以得到令人满意的区分结果。Petersen 等[27]在 Schwarm 的基础上，选取了相同的语言特征，通过在训练集中加入负样本的方法，提升了分类器的准确率。实验结果显示，加入负样本的支持向量机分类器在高等级文本的区分上有明显进步。支持向量机的训练要求求解计算复杂度极高的二次规划问题，为了缓解训练样本数越多、实际任务中的开销越大的问题，Aluisio 等[28]在训练支持向量机时使用了序列最小优化算法，高效优化了分类器的训练过程。

鉴于支持向量机在可读性评估上的优异表现，后来的研究者尝试在支持向量机的基础上对整个评估流程进行改进。或使用质量更高的训练语料[29]，或对语言特征进行进一步筛选整合[30-31]。Chen[32]借助从 E-HowNet 中学习的词汇关系为中学课文构建了词汇链，并结合词频-逆文件频率(Term Frequency-Inverse Document Frequency，TF-IDF)所筛选的词作为特征，支持向量机分类器在低年级的最好分类准确率为 96%，在中级的最好分类结果为 85%。Cha 等[33]在预测文本的可读性时使用 Word2Vec 和 FastText 两种方法构建词向量和段落向量，然后分别使用布朗聚类(Brown clustering)和 K 近邻进行聚类，支持向量机通过自主学习的特征对文本的难度进行预测，预测结果与原等级的相关性超过 80%。

1.3 基于排序的方法

良好的分类模型需要带有文本难度标注的语料库。英文的可读性研究起步较早，资源较多，其他语言中分级文本语料库较少且难以获取，如果使用标注准确度很高的教材课文文本，又可能涉及版权问题。因此，如何在缺乏带有标注的大规模语料库的情况下对文本的可读性进行评估，是可读性分析面临的问题之一。

在缺乏带难度等级标签数据的情况下，Tanaka-Ishii 等[34]使用基于排序的方法对文本的难度进行测定。假定文本存在难易值，对于任意两个文本，它们的难易关系有三种：

$$\gamma(x)>\gamma(y);\gamma(x)<\gamma(y);\gamma(x)=\gamma(y) \quad (3)$$

如果可以从数据中学习一个难度比较器，就可以对语料库中的文本进行排序。对于排序好的文本集 C 中的任意两个文本都满足 $\gamma(C_i)\leqslant\gamma(C_{i+1})$。该研究首先利用只有难易两个类标注的文本训练比较器，然后使用二分插排算法对经过比较的文本进行排序，如此循环直到数据集中的所有文本全部被比较，即可得到排序好的文本集 C。

该研究开发了基于排序方法的 Terrace 网页分析器，如图 1 所示。网页分析器每天收集 CNN 的新闻文本，文本经过支持向量机比较器后，所有新闻文本在后台以有序状态排列。当用户上传文本后，分析器会给出分析文本在后台语料库中的难度位置，并向用户推荐语料库中与待分析文本可读性距离最近的文章。

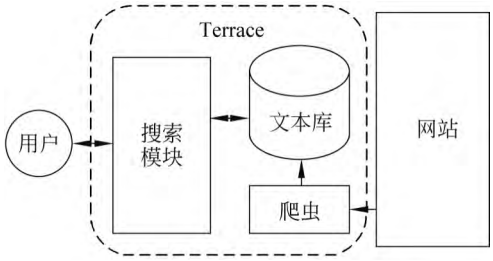

图 1 Terrace 网页分析器

在理想的情况下，比较器可以对两篇文本进行准确的难易判断，但事实是，比较器总存在一定的误差，从而影响比较器的判断。佐藤理史[35]在对文本进行难度排序时考虑了 ρ 误差的存在，把比较器修改为式(4)。

$$\begin{cases}0 & \text{if } \gamma(x)=\gamma(y)\\ +1 & \text{if } \gamma(x)>\gamma(y)+\rho\\ -1 & \text{if } \gamma(x)+\rho<\gamma(y)\end{cases} \quad (4)$$

除了构建比较器外，Schumacher 等[36]使用人工标注的方法得到排序文本，要求众包平台上的评估者阅读两篇文本，并判断这两篇文本的相对难度。研究者得到两两比较的数据，通过使用评分排序算法如 Trueskill[37]得到最终的排序文本集。

排序法的优势在于：第一，文本的相对难度更符合实际认知，人们不能给出文本绝对的难度值，但对于给定的两篇文本，人们可以判定哪篇文本更难。同时，文本的难度值是一个连续统，文本与文本之间有难度的连续关系。第二，排序算法对标注信息要求不高，在缺乏多等级标注语料库的情况下不失为一种好的选择。三种方法的对比如表 2 所示。

表 2 公式法、分类法、排序法的主要思路及优缺点

方法	主要思路	分析结果	优 点	缺 点
公式法	通过可计量的特征，建立多元线性回归方程，预测文本的难度值	数值或等级	构建容易，使用简单	特征少，预测能力差
分类法	有监督的机器学习方法，通过有等级标注的语料库，学习文本特征，构建分类模型	难度等级	纳入的特征多，模型的预测效度好	需要带难度标注的语料库，构建相对复杂
排序法	通过构建比较器或人工标注得到文本的两两相对难度，利用相对难度通过排序算法得到按难度排序的文本集合	相对位置	更符合实际认知，可在缺乏标注语料情况下使用	不能反映文本的绝对难度，模型计算时间长

2 特征与数据集

2.1 文本特征及特征选择

目前可读性研究主要集中在对文本特征的分析及效度验证上[38]，本文把英文可读性研究中使用的特征分为四个一级特征，在此基础上，将该范畴下所涉及的子特征细分为二级特征，将具体可度量的文本特征作为三级特征，从而构建一个层级化的可读性特征体系，如表 3 所示。

表 3　英文可读性特征体系表

一级特征	二级特征	三级特征举例
公式特征	词汇难度	词长\音节数\常用词比
	句子难度	平均句长
	公式分数	FK 分数\ Lexile 分数等
词汇特征	多样性	类符-形符比(TTR)
	独特性	常用词词表\领域词典
	复杂性	语言模型的困惑度
句法特征	词性层	名词代词\介词数等
	短语层	名词短语\动词短语等
	从句层	同位语从句\修饰语等
	句子层	句法树高\复句占比等
篇章特征	概念密度	实体词占比\命题数等
	实体网格	实体转移序列的概率
	词汇链	词汇链的链长、链距
	指代消解	指代链的链长、链距

大多数可读性公式把词汇难度和句子难度作为衡量文本难度的标准。平均句长是仅有的衡量句子难度的特征。一些公式使用词长、音节数和字母数作为衡量词汇复杂度的指标，但 Dale-Chall 可读性公式[39]在衡量词汇复杂度时使用了常用词表来计算文本中常用词的占比：文本使用的常用词越多，读者对词汇越熟悉，文本越简单。Petersen 等[27]在四个语料库上分别训练了一元、二元和三元语言模型，把这 12 个语言模型的困惑度(perplexity)作为词汇难度的指标。Feng[23]使用了四种文本序列表示方法，即词序列、词性序列、词＋词性序列、信息增益(information gain)选择后的词＋词性序列来表示四个训练集，也分别训练了三个语言模型，把 48 个困惑度作为文本词汇特征。Schwarm 等[21]把机器学习的方法应用于文本可读性分析，可以纳入更多的特征，如衡量词汇多样性常用的类符/形符比(the type-token ratio，TTR)和文本的句法特征，包括句法树的高度、从句及复杂从句的数量和长度、实词和功能词的数量、动词短语和名词短语的数量等。

Graesser 等[40]认为，文本的衔接和连贯影响文本的可读性，在对文本可读性进行分析时，不仅要考虑词汇句法特征，还要加入深层篇章语义特征。为了更好地分析文本，该团队开发了一个文本分析工具 Coh-Metrix。Coh-Metrix 作为一个基于网络的文本分析工具，融合了计算语言学和语料库语言学的多种技术，可以对文本的 106 个词汇语法和篇章特征进行自动抽取。Feng[18]从实体词密度、词汇链、指代推理和实体网格(entity grid)[41]三种范畴出发抽取了共 20 个篇章有关的特征来评估文本可读性。Lin 等[42]在评估文本可读性时使用了语义网和 WordNet 的词汇关系。其做法是，对于给定的名词，根据其在 WordNet 中的位置，找出其上位词和下位词，将阅读者最容易理解的概念定义为基础词，基于由基础词构成的短语频率和上下位词的长度差异，利用目标文本中基础词的比例来估计文本可读性等级。

表 4 对比了相同数据集下使用不同特征及其组合进行预测时模型的准确率。

从数量上看，一般情况下，特征的数量与模型的效度成正比，特征越多，模型的预测能力越好。Feng 等[18]的研究中，经过扩充的特征集(8→21)使得模型准确率从 50.91%提升到 57.79%。研究也同时显示，加入所有特征的模型表现最好，但在相同的范畴下，与所有特征相比(72.21%)，经过筛选的 28 个特征也有不错的预测能力(70.06%)。

从范畴上看，公式特征、词汇特征和句法特征是被广泛使用的文本特征，三者的效度得到了相关研究的证实[28,32]。三个范畴特征的组合使得 Vajjala 等[29]模型的准确率达到 91.3%。从单个特征来说，词汇特征的预测效度最好。Collins-Thompson 等[25]研究显示，以词汇特征为基础建立的语言模型在预测 1～12 等级的网页文本时表现更好。Flor 等[43]基于回归模型考察了词汇紧密度与文本复杂度之间的关系。结果显示，词汇紧密度都和文本复杂度密切相关；文本等级越高，词汇的紧密度就越小，预测能力越小；文本中的词汇紧密度与文本复杂度的关系受文本类型的影响。词汇特征的贡献率大于语法特征，但二者结合起来的模型预测能力更好。虽然词汇特征的效度高于句法特征，但句法特征在面向二语者的文本可读性预测任务时表现更为突出[44-45]。篇章特征的效度还有待验证。一些研究者认为，篇章特征与阅读时的认知过程有关，是重要的评估文本难度的特征[40]。Pitler 等[46]从六个角度(词汇特征、句法特征、指代特征、实体词和篇章特征)对比了文本难度相关的特征，发现每句中动词短语的数量、词数、词汇似然度、篇章似然度与文本难度等级密切相关。但在另一些研究中，加入了篇章特征的模型，其性能并没有明显的提高[23,47]。

表4 Weekly Reader下不同特征的效度对比

相关研究	特征数	特征范畴	准确率/%
Schwarm和Ostendorf[21]	4	句法	50.91
	25	词汇+句法	63.18
Feng[18]和Huenerfauth	8	公式	57.79
	80	词汇	68.38
	21	句法	57.79
	28	TOP 28	70.06
	273	公式+词汇+句法+篇章	72.21
Vajjala和Meurers[30]	19	词汇	84.10
	25	句法	64.30
	46	公式+词汇+句法	91.30

2.2 现有数据集

文本可读性的自动分析多是有监督任务，需要带有难度标记的数据集来训练预测模型。英语国家拥有较早的读物分级意识，数据资源比较丰富。带难度标注的数据集主要有各州共同核心标准(Common Core State Standards，CCSS)中附属的文本、the Weekly Reader分级杂志、The Weebit Corpus等。CCSS由美国教育部官方制定推广，旨在为数学、艺术、文学领域的教育提供统一、具体的教育标准。该标准对美国各年级(从幼儿园到初中)学生的学习目标和阅读能力进行了明确的划分，并给出了具体的符合各年级能力的阅读文本范例。除了等级的划分，该语料还标注了文本类型，如故事、诗词、说明文、戏剧等。The Weekly Reader(WR)分级杂志①是针对青少年发行的在线教育类周刊。Vajjala等[29]综合了The Weekly Reader分级杂志和the BBC-Bitesize网站②的文本，建立了规模更大的语料库——The Weebit Corpus。三个语料库的对比如表5所示。

为了抽取文本的篇章特征，Pitler等[46]在可读性研究中使用了宾州语篇树库(The Penn Discourse Treebank，PDTB)[48]语料库。宾州语篇树库是Prasad等于2004年建立的大规模语料库。宾州语篇树库标注了文本的局部篇章关系，没有难度标注。该研究随机选取了PDTB的30篇文本，从宾州树库中抽取篇章关系作为文本特征，同时对文本可读性进行了人工标注，请大学生限时阅读文本并按照一定规则对文本进行1～5分的难易度评价，把每篇文本得分的均值作为文本的可读性级别。

表5 CCSS、WeeklyReader、WeeBit语料库对比

	CCSS		WR		WeeBit	
年龄	等级	文章	等级	文章	等级	文章
3～6	K—1	61				
7～8	K2～3	56	L2	629	L1	629
8～9			L3	801	L2	801
9～10	K4～5	38	L4	814	L3	814
11～12	K6～8	38	L5	1 325	L4	644
13～14	K9～10	63				
15+	K11+	73			L5	3 500
总计	6	329	4	3 569	5	6 388

对于缺乏成熟数据集的语言，如日语、汉语等，研究者们选择自己构建语料库，语料来源一般为教材课文文本[42,49]。众包平台的成熟使得部分研究者选择利用众包构建语料库[32,36,47]。Clercq等[50]的研究中，要求标注者阅读两个段落并对比它们的相对难易度，把标注者的标注结果与专家的标注结果进行对比，发现二者并没有显著差别。

3 汉语文本可读性研究

英语文本的可读性研究发展较早，且成果丰富。与英语不同，汉语文本可读性研究仍处于起步阶段，多集中在可读性公式的研制上。

汉语可读性公式的构建大致遵循了英语可读性公式的研究范式，但在特征选择和应用领域上具有自己的特点。特征选择的不同是由汉、英各自的语言特点决定的。汉语的文字载体是汉字，从形体上来说，汉字是由笔画构成的方块字；从性质上来说，汉字是语素音节文字，一个汉字通常表示汉语里的一个词或一个语素，具有音形义相统一的特点。杨孝滨[51]从字词句三个粒度选取了笔画数、完全对称字率、单音词率、成语比例等23个语言特征对中文报刊文本的可读性进行了相关性分析。Hong[52]应用趋势分析法，从词、语义、句法、连贯四个层面选取了32个特征进行对比分析。

在应用上，汉语文本可读性研究的成果主要集中在教学领域。在汉语作为母语的教学领域，张必隐等[53]利用初中二年级学生的完形填空成绩对20

① http://classroommagazines.scholastic.com/

② https://www.bbc.com/education

篇字数在 250 字左右的段落进行了可读性公式的拟合。荆溪昱[54]以年级作为因变量，对台湾 1～12 年级的语文中国课本进行了难度的量化分析，并比较了每篇课本实际年级与实际难度的偏差。

母语教学领域的工作给汉语作为二语的教学领域提供了可借鉴的经验。对外汉语教学领域教材多样，但多套教材在同一水平上重复，缺乏科学的语言点设置和对外汉语教材评估体系[55-56]。基于此状，张宁志[57]借鉴母语教材的评估经验，使用每百字的句子数、平均句子长度、非常用词数对常用的 16 本中高级教材进行了难度评估，具有开创性价值。类似研究还有李燕[58]、罗素华[59]等。郭望皓[60]对外汉语文本难度进行了探究，该研究首先通过问卷调查的方法，对影响对外汉语文本难度的因素进行了调查和筛选，筛选后的文本通过 CRITIC 加权赋值法计算了各因素的权重系数，最后拟合出对外汉语文本的可读性公式，如式(5)所示。

$$Y = -11.946 + 0.123x_1 + 0.198x_2 + 0.811x_3 \tag{5}$$

其中 x_1 为平均句长，x_2 为汉字难度，x_3 为词汇难度，该公式的拟合优度为 0.917。

左虹等[61]在教师问卷调查和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[62]以 90 名初中级水平日本及韩国留学生在记叙性短文上的完形填空成绩作为因变量，从字词句篇四个方面筛选了 17 个特征对 20 篇短文的难度进行量化，构建了专门针对初中级日韩汉语学习者的可读性公式。这两项研究明确了所建立可读性公式的适用范围，对教学来说有一定的针对性和实用价值。

除了教学领域外，邹红建等[63]对对外汉语教学中常用的报刊文本进行了可读性研究。研究先假设报刊文本的难易度与文本长度和常用词的比例有关，然后通过比较文本位置偏移累加和人工标注结果的方法确定二者的最佳系数。作者也指出，由于语料长度的限制，该系数并不是普遍适用的。宋曜廷等[64]对影响汉语文本可读性的因素进行了探究，并借鉴英文文本分析工具 Coh-metrix[40]，构建了适用于中文的文本分析工具 CRIE(the Chinese Readability Index Explorer)，该工具主要关注中文文本的衔接性和连贯性，可以分析的指标包括词性、词频、衔接性、词汇信息、连词、句子结构等。孙刚[65]选取表面特征、词汇特征、语法特征和信息熵特征建立线性回归模型进行可读性预测，重点探讨了特征选择工程对最终模型性能的影响。曾厚强等[66]结合 FastText 词向量表示与深度学习模型(卷积神经网络)对文本可读性进行分类预测。

汉语文本可读性的自动分析研究虽然取得了一些成果，但仍具有以下不足：

(1) 汉语文本可读性研究在研究对象、数量、方法和应用领域等方面都还比较有限，大部分是针对某个特定群体的学生进行的教材分析和教学研究工作。从总体上看，面向二语者的可读性研究成果丰富，面向广泛母语人群的可读性研究有广阔的发展空间。

(2) 影响或预测汉语文本可读性的指标还有待扩充和验证[64]。一方面，影响或预测拼音文本可读性的语言特征不一定适用于汉语文本可读性研究；另一方面，现有可读性研究工作中使用的各项特征在范畴归属和特征效度上存在冲突，还有待系统地梳理和验证。

(3) 主要以线性模型为主，自然语言处理技术在中文可读性的自动分析研究上应用不足。

(4) 公开的文本难度标注语料库构建不足。由于缺乏公开的训练和测试数据，研究者只能自己构建教材课文语料库，在模型评价时只能采用自评的办法，缺少研究的横向对比。

4　总结与展望

本文对近年来文本可读性的自动分析研究进行了综述。随着网络文本的大量涌现，文本分析日益成为热点，文本可读性分析是文本分析的重要内容，涉及计算机科学、语言学、教育学和心理学多个学科。从最初的可读性公式的研制，到近期的可读性自动分析工具[40,64]和模型的建立，自然语言处理技术的进步为可读性的自动分析提供了多种思路和方法。文本可读性研究作为一项有着丰富应用场景的课题，今后的发展呈现以下趋势：

(1) 知识信息的加入，包括篇章连接关系、推理知识和读者知识背景等。知识信息的加入有助于区分难度较高的文本，需要分析和抽取文本篇章信息，或结合读者的知识背景等个体差异。

(2) 探究文本类型对文本难度的影响。人们阅读不同类型的文本时会采用不同的理解和加工策略[19]。可读性公式无法区分由文本类型带来的文本难度的差距，文本难度分类模型会产生类型偏差(genre bias)，模型倾向于把文学文本(literary

texts)划分为更高的难度级别,把信息文本(informational text)划分为更低的难度级别[67],现有的研究仅有部分注意到了文本类型的影响[68],却没有进行更深入的分析。

(3)使用深度学习模型和新的文本表示方法,如神经网络模型和基于词向量的文本表示[33,66]。近年来随着表示学习方法技术的蓬勃发展,训练可读性模型所需要的特征可以不需要仰赖专家知识,这使得可读性自动分析的发展有了一个崭新的研究方向。

参考文献

[1] Michael B W Wolfe, et al. Learning from text: Matching readers and texts by latent semantic analysis[J]. Discourse Processes, 1998, 25(2-3):309-336.

[2] 王蕾. 可读性公式的内涵及研究范式——兼议对外汉语可读性公式的研究任务[J]. 语言教学与研究, 2008,(6):46-53.

[3] Vogel M, Washburne C. An objective method of determining grade placement of children's reading material[J]. Elementary School Journal, 1928, 28(5):373-381.

[4] Sheehan K M, Kostin I, Napolitano D, et al. The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment [J]. Elementary School Journal, 2014, 115(2):184-209.

[5] 郭曙纶. 试论对外汉语教材中的超纲词[J]. 宁夏大学学报(人文社会科学版), 2008, 30(4):25-29.

[6] Sato S. Automatic assessment of Japanese text readability based on a textbook corpus[J]. Proc of Lrec08 Marrakech Morocco, 2008, 24(1):654-660.

[7] Jin T, Lu X. A data-driven approach to text adaptation in teaching material preparation: Design, implementation, and teacher professional development[J]. Tesol Quarterly, 2017,52(2):457-467.

[8] McNamara D S, et al. A hierarchical classification approach to automated essay scoring [J]. Assessing Writing, 2015(23):35-59.

[9] Nandhini K, Balasundaram S R. Improving readability through individualized summary extraction, using interactive genetic algorithm[J]. Applied Artificial Intelligence, 2016, 30(7):635-661.

[10] Jin Y K, et al. Characterizing web content, user interests, and search behavior by reading level and topic [C]//ACM International Conference on Web Search and Data Mining. ACM, 2012:213-222.

[11] 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.

[12] Dubay W H. The principles of readability[J]. Online Submission, 2004, 102(1):631-3309.

[13] Dale E, Chall J S. A formula for predicting readability[J]. Educational Research Bulletin, 1948, 27(1): 37-54.

[14] Gunning R. The technique of clear writing [M]. McGraw-Hill, 1952:36-37.

[15] Laughlin G H M. SMOG Grading-A new readability formula[J]. Journal of Reading, 1969, 12(8):639-646.

[16] Caylor John S, et al. Methodologies for determining reading requirements of military occupational specialties[J]. Adult Literacy, 1973:81.

[17] Kincaid J P, Fishburn R P, Chisson B S. Derivation of new readability formulas for navy enlisted personnel[J]. Adult Basic Education, 1975:49.

[18] Feng L, Huenerfauth M. Cognitively motivated features for readability assessment [C]//Proceedingsof Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009:229-237.

[19] Danielle S McNamara, Walter Kintsch. Learning from texts: Effects of prior knowledge and text coherence[J]. Discourse Processes, 1996, 22(3):247-288.

[20] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社, 2008.

[21] Schwarm S E, Ostendorf M. Reading level assessment using support vector machines and statistical language models [C]//Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005:523-530.

[22] Heilman M, Collins-Thompson K, Eskenazi M. An analysis of statistical models and features for reading difficulty prediction[C]//Proceedings of the Workshop on Innovative Use of Nlp for Building Educational Applications, 2018: 71-79.

[23] Feng J. Automatic readability assessment[J]. Dissertations & Theses-Gradworks, 2010, (93):84-91.

[24] Luo S, Callan J. A statistical model for scientific readability [C]//Proceedings of 10th International Conference on Information and Knowledge Management. ACM, 2001:574-576.

[25] Collins-Thompson K, Callan J P. A language modeling approach to predicting reading difficulty[C]//Human Language Technologies: the 2004 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2004:193-200.

[26] Cortes C, Vapnik V. Support-vector networks[J].

Machine Learning，1995，20(3)：273-297.

[27] Petersen S E，Ostendorf M. A machine learning approach to reading level assessment[J]. Computer Speech & Language，2009，23(1)：89-106.

[28] Aluisio S，et al. Readability assessment for text simplification[C]//NAACL Hlt 2010 15th Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics，2010：1-9.

[29] Vajjala S，Meurers D. On improving the accuracy of readability classification using insights from second language acquisition[C]//Proceedings of the Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics，2012：163-173.

[30] Shen W，et al. A language-independent approach to automatic text difficulty assessment for second-language learners[J]. 2013：30-38.

[31] Kate R J，et al. Learning to predict readability using diverse linguistic features[C]//Proceedings of Coling 2010 -23rd International Conference on Computational Linguistics，Proceedings of the Conference. COLING，2010：546-554.

[32] Chen Y T，Chen Y H，Cheng Y C. Assessing Chinese readability using term frequency and lexical chain[J]. **中文计算语言学期刊**，2013，18(2)：1-17.

[33] Cha M，Gwon Y，Kung H T. Language modeling by clustering with word embeddings for text readability assessment[C]//ACM，2017：2003-2006.

[34] Tanaka-Ishii K，Tezuka S，Terada H. Sorting texts by readability[M]. MIT Press，2010.

[35] **佐藤理史. 均衡コーパスを規範とするテキスト難易度測定**[J]. **情報処理学会論文誌**，2011，52(4)：1777-1789.

[36] Schumacher E，et al. Predicting the relative difficulty of single sentences with and without surrounding context[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. 2016：1871-1881.

[37] Schölkopf，B，Platt，J，Hofmann，T. TrueSkill™：A Bayesian Skill Rating System[M]//Advances in Neural Information Processing Systems 19：Proceedings of the 2006 Conference. MIT Press，2007：569-576.

[38] **陈茹玲，蔡鑫廷，宋曜廷，等. 文本适读性分级架构之建立研究**[J]. **层级分析法**，2015，60(1)：001-032.

[39] Chall J S，Dale E. Readability revisited：the new Dale-Chall readability formula[J]. Brookline Books，1995：149.

[40] Graesser A C，Mcnamara D S，Kulikowich J M. Coh-Metrix：Providing multilevel analyses of text characteristics[J]. Educational Researcher，2015，40(5)：223-234.

[41] Barzilay R，Lapata M. Modeling local coherence：An entity-based approach[C]//Proceedings of Meeting on Association for Computational Linguistics. Association for Computational Linguistics，2005：141-148.

[42] Lin S Y，et al. Assessing text readability using hierarchical lexical relations retrieved from WordNet[J]. **中文计算语言学期刊**，2009，14(1)：45-83.

[43] Flor M，Klebanov B B，Sheehan K M. Lexical tightness and text complexity[C]//The Workshop on Natural Language Processing for Improving Textual Accessibility，2013：29-38.

[44] Lu X. Automatic analysis of syntactic complexity in second language writing[J]. International Journal of Corpus Linguistics，2010，15(4)：474-496.

[45] Heilman M，Collins-Thompson K，Callan J，et al. Combining lexical and grammatical features to improve readability measures for first and second language texts[C]//Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics，Proceedings，April 22-27，2007，Rochester，New York，USA. DBLP，2007：460-467.

[46] Pitler E，Nenkova A. Revisiting readability：A unified framework for predicting text quality[C]//The Conference on Empirical Methods in Natural Language Processing，2008：186-195.

[47] Vajjala S. Automated assessment of non-native learner essays：Investigating the role of linguistic features[J]. International Journal of Artificial Intelligence in Education，2016，28(1)：1-27.

[48] Miltsakaki E，Prasad R，Joshi A，et al. The Penn Discourse Treebank[J]. Proceedings of Lrec，2004，24(1)：2961-2968.

[49] Sung Y T，Chen J L，Cha J H，et al. Constructing and validating readability models：The method of integrating multilevel linguistic features with machine learning[J]. Behavior Research Methods，2015，47(2)：340-354.

[50] Clercq O D，et al. Using the crowd for readability prediction[J]. Natural Language Engineering，2014，20(3)：293-325.

[51] **杨孝滢. 实用中文报纸可读性公式**[J]. **新闻学研究**，1974，13：37-62.

[52] Hong J F，Sung Y T，Tseng H C，et al. A multilevel analysis of the linguistic features affecting Chinese text readability[J]. **台湾华语教学研究**，2016，(13)：95-126.

[53] **张必隐，孙汉银. 中文易懂性公式**[C]. **中美教育问题研讨会论文集**，1992：246-249.

[54] 荆溪昱. 中文国文教材的适读性研究：适读年级值的推估[J]. 教育研究资讯，1995，3(3)：113-127.
[55] 赵金铭. 论对外汉语教材评估[J]. 语言教学与研究，1998，(3)：4-19.
[56] 朱勇. 汉语分级读物的现状与研发对策[J]. 国际汉语教学研究，2015，(2)：15-17.
[57] 张宁志. 汉语教材语料难度的定量分析[J]. 世界汉语教学，2000，(3)：83-88.
[58] 李燕，张英伟.《博雅汉语》教材语料难度的定量分析——兼谈影响教材语言难度的因素和题材的选择[J]. 云南师范大学学报(对外汉语教学与研究版)，2010，8(1)：39-43.
[59] 罗素华. 汉语中级泛读教材难度定量分析——以三部中级汉语泛读教材为例[D]. 长沙：湖南师范大学硕士学位论文，2015.
[60] 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.
[61] 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，(2)：263-276.
[62] 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，2017，(5)：15-25.
[63] 邹红建，杨尔弘. 面向对外汉语报刊教学的文本难易度分类[C]//学生计算语言学研讨会，2006：363-367.
[64] Sung Y T，Chang T H，Lin W C，et al. CRIE：An automated analyzer for Chinese texts[J]. Behavior Research Methods，2015，48(4)：1-14.
[65] 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.
[66] 曾厚强，陈柏琳，宋曜廷. 探究使用基于类神经网路之特征于文本可读性分类[J]. 中文计算语言学期刊，2017，22(2)：31-45.
[67] Kucan L，Beck I L. Thinking aloud and reading comprehension research：inquiry，instruction，and social interaction[J]. Review of Educational Research，1997，67(3)：271-299.
[68] Sheehan K M，et al. Generating Automated Text Complexity Classifications That Are Aligned with Targeted Text Complexity Standards[J]. ETS Research Report Series，2010，10(2)：i-44.

吴思远(1998—)，硕士研究生，主要研究领域为第二语言习得、自然语言处理。
E-mail：wusiyuan2401@163.com

于东(1982—)，通信作者，博士，副教授，主要研究领域为自然语言处理。
E-mail：yudong_bluc@126.com

蔡建永(1985—)，博士研究生，讲师，主要研究领域为汉语作为第二语言的习得和认知加工研究。
E-mail：caijianyong@blcu.edu.cn