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 *

朱君辉[1,3]　刘　鑫[1,3]　杨麟儿[1,3]　王鸿滨[2]　杨尔弘[3]

（1. 北京语言大学信息科学学院　北京　100083；
2. 北京语言大学汉语国际教育研究院　北京　100083；
3. 北京语言大学国家语言资源监测与研究平面媒体中心　北京　100083）

[摘要] 汉语二语文本难度自动分级任务是国际中文教育与计算语言学领域中的一个重要主题。本文依据《国际中文教育中文水平等级标准》，提出了基于语法点多样性与复杂性的 25 个语法点特征并实现了相关特征的自动抽取与计算，在此基础上构建了自动分级模型。实验结果表明，融合语法点特征后多元逻辑回归算法的分级准确率为 86.40%，比基于现有语言特征的实验提升了 2.4%。进一步研究发现，六级语法点多样性、语法点难度等级均值是区别文章难度级别的关键特征。此外，本文将包含语法点特征在内的 207 项语言特征融入基于 BERT 的深度学习模型，取得了 87.6% 的准确率，超过了基于传统语言特征的方法和基于神经网络的方法。

[关键词] 语法点特征；汉语作为第二语言教学研究；文本难度；自动分级

[中图分类号] H087 [文献标识码] A [文章编号] 1003–5397（2022）03–0087–13

Chinese Grammatical Structure Features and its Application in Automatic Grading of L2 Texts

ZHU Junhui, LIU Xin, YANG Liner, WANG Hongbin, YANG Erhong

Abstract: Automatic Chinese L2 texts grading is an essential topic in international Chinese language education and computational linguistics. This paper proposes 25 grammatical structure features based on the diversity and complexity of grammatical structure, and implements the automatic extraction and calculation of the relevant features, based on which an automatic grading

[收稿日期] 2022–06–06

[作者简介] 朱君辉，北京语言大学硕士生，主要研究语言信息处理；刘鑫，北京语言大学硕士生，主要研究自然语言处理；杨麟儿（通讯作者），北京语言大学副教授，博士，主要研究计算语言学、计算机辅助语言学习；王鸿滨，北京语言大学教授，博士，主要研究分级阅读研究、国际中文教育标准研究；杨尔弘，北京语言大学教授，博士，主要研究语言信息处理、语言监测、语言资源建设。

* 本研究得到国家语委项目“智能辅助汉语应用文写作研究”（ZDI135-131）、国家社科基金重点项目“汉语作为第二语言学习者阅读素养评价标准的构建与测评研究”（20AYY016）和北京语言大学研究生创新基金项目“国际中文教育阅读文本自动分级”（22YCX114）资助。承蒙胡韧奋老师、编辑部和审稿专家提出宝贵的修改意见和建议，谨此一并致谢！

model is constructed. The results show that the accuracy of the model is 86.40%, which is 2.4% higher than that of the existing language features. It was further found that Grade 6 grammar structure diversity and grammar structure difficulty level means were the most relevant features to distinguish the difficulty level of the texts. In addition, the paper incorporated 207 linguistic features including grammar structure features into the BERT-based deep learning model and achieved an accuracy of 87.6%, outperforming both the traditional linguistic feature-based and neural network-based methods.

Keywords: grammatical structure features; L2 Chinese; text difficulty; automatic grading

一 引言

阅读是人类学习语言和获取知识的基本方式之一，也是语言学习的核心和重点（王钟华，1999）。过难或过易的阅读材料都会降低学习者的阅读效能感及兴趣，如何快速有效地为汉语学习者提供难度适宜的阅读材料成为了汉语教学领域关心的重要问题（蔡永强，2017）。为了降低人工评估的主观性与时间成本，文本难度自动分级研究应运而生。文本难度自动分级即通过将文本转化为一组可量化的、影响阅读难度的语言特征，利用这些特征训练语言模型，实现对新文本难度的自动预测（Vajjala，2015；Schwarm & Ostendorf，2005）。在汉语二语阅读材料需求量巨大，而汉语二语文本难度自动分级研究仍较为薄弱的背景下，建立应用性强的汉语二语文本自动分级模型十分必要。其中，探索影响汉语二语阅读难度的核心语言特征是构建高质量分级模型的重点及难点之一（王鸿滨，2020）。

近年来，越来越多的研究指出二语学习者的阅读过程与母语者存在差异，从二语习得的视角探索更为精细有效的语言特征尤为重要（王蕾，2022）。以往研究表明，英语二语学习者的语法知识与其二语阅读理解的关系最为密切（Jeon & Yamashita，2014）。由于二语学习者的语法知识有限，一个结构简单但含有超出学习范围语法点的句子亦会对其阅读理解造成困扰（Heilman et al.，2007）。作为语法知识与技能的呈现形式，“语法点”即各类语言教材、教学语法大纲中收录的语法项目，是学习者在语言课堂中接触到的最直接的内容。将语法点由易到难进行编排是语言教材以及分级读物编写的重要原则（吕文华，2002），也是二语文本自动分级任务的重要依据。Wang 等（2016）参考日语的语法点提出了 5 个语法特征，使用多元线性回归算法在五个级别的文本分类上准确率达到了 87.7%。王丽（2020）根据英语教师总结的英语语法点对英语文本进行语法知识的自动解析，在传统语言特征中加入语法特征，分类模型准确率达到 70% 左右。这些研究启示我们转换视角，遵循第二语言习得的认知规律，探究语法点特征在汉语文本难度自动分级中的作用。

现有的汉语二语文本难度自动分级研究对语法点特征的考虑十分有限。目前，在汉语二语文本难度特征构建的研究中，衡量语法难度的特征大多借鉴汉语母语的研究成果，主要通过平均句长、平均小句长、句法树深度、词性类别数量等度量句法复杂性的传统语言特征间接地衡量语法的难度（张宁志，2000；孙刚，2015；孙未未等，2018；程勇等，2020；吴思远，2020）。这些特征对母语的文本难度分级很有效，但在二语文本分级任务上表现不佳（Curto et al.，2015）。王蕾（2005）与杨金余（2008）关注到了汉语的语法点与文本难度之间的关系，参考《对外汉语教学初级阶段教学大纲》，将不同等级的语法点作为可读性公式构建的变量之一，研究结论证明了不同等级的语法点数量是预测汉语二语文本难度的有效特征。不足之处是王蕾与杨金余的研究均基于小样本语料进行人工统计，未能实现语

法点的自动抽取，从而无法在大规模语料上进行更复杂的语法点特征的自动计算。

以上研究为汉语二语文本自动分级研究做出了探索性贡献，但仍存在以下问题：首先，所提语言特征大多借鉴母语研究，忽略了从二语习得的视角考察分析语法点的作用。第二，一些研究对汉语的语法点与文本难度的关系进行了探讨，但由于无法实现语法点在文本中的自动抽取，实证研究往往基于小规模语料，依赖人工标注的方式，时间成本较高。第三，现有的汉语复杂度分析工具（如中文 Coh-Metrix[①]、CRIE[②]）未能从语法点的角度提供相关特征的自动计算，缺乏具体的特征构建和实验证据。

针对上述问题，本文提出 25 个汉语语法点特征并实现其自动计算，验证语法点特征在汉语二语文本难度自动分级任务中的有效性。本文构建了带等级标签的国际中文教材语料库，引入前人研究中提出的衡量文本难度不同层面的特征作为传统语言特征。在此基础上，对比语法点特征与传统语言特征在汉语二语文本分级任务上的预测准确度，并探索了深度学习方法在汉语二语文本分级任务上的应用。最后，本文考察了对汉语文本自动分级贡献最大的语法点特征，对实验结果进行了分析。

二　语法点的选取及其在文本中的自动抽取

（一）语法点的选取及难度标准

作为构建语法点特征的基础工作，确定语法点的抽取对象是实现语法点自动抽取的第一步。为了实现语法点特征的自动计算，本文提出了一种汉语语法点的自动抽取方法，并基于国家 2021 年新发布的《国际中文教育中文水平等级标准》（GF0025 – 2021）（以下简称新《等级标准》），实现了其《附录 A（规范性）语法等级大纲》（以下简称《语法等级大纲》）中的语法点在大规模语料中的自动抽取。《语法等级大纲》依据 3000 余册国际中文教材语法点频率统计，参考 11 部国际中文教学大纲与标准的研制成果，将语法点分别从形式和功能角度划分为三等九级，其中七至九级并为高等，共包含 572 条语法点。

《语法等级大纲》语法点的体系完整，在语法资源的建设中占据着无可质疑的基础框架地位，绝大多数对外汉语教材的编写都以其作为标准。由于教材语法点通常与课文文本有一定的相关性，因而《语法等级大纲》中的语法点具有最广泛的教材语料基础。另外，新《等级标准》将“语言范围”提炼优化为“语言量化特征”，形式化更强，具有更高的可信度和可操作性（王鸿滨，2021），便于计算机进行识别。

（二）语法点的重构

虽然《语法等级大纲》可以用来作为语法点自动抽取的基础框架，但与计算机可理解的实用的语法点检索式之间还存在着差距。本文根据便于检索的原则，使用以下三种方式对《语法等级大纲》进行重构，获得既可以反映语法大纲体系，又便于自动抽取的语法点：

对语法点进行归并整合。对语法点进行归并整合的工作包括两个方面。一是将同一级别、同一词性且具有相同句法结构关系的语法点进行合并。如介词“对”“往”“给”，同属于二级词汇的介词语法点，且同是“对 / 往 / 给 + 宾语 + 动词”的介宾短语修饰动词的结构，因此将这三个语法点合并为一个语法点。另一方面，《语法等级大纲》中不同分类下存在共同的语法点，如动态助词“了 $_1$”被划分到“动作的态”和“助词”两个类别下，对此我们同时保留语法描述信息，但构建检索式时只择其一。

对概括性较高的语法点进行具体化。具体化的目的是进一步明确检索的关键词或标签。例如，将“特指问句”具体化为以“谁、什么、哪儿、哪、哪些、哪里、怎么了、怎么样、

多会儿、多少、为什么”为疑问词的疑问句。

删除通过现有技术手段难以准确抽取的语法点。限于当前中文信息处理自动句法分析的水平，为了确保自动抽取的准确性，我们对以下三类语法点进行了删除处理。一是本身暂时无法形式化的语法点，如“主语、宾语的语义类型”“意念被动句”“表致使意义的把字句”。二是紧缩复句与无标记复句。三是语篇层面的语法点，即多重复句与句群，由于暂时无法通过技术手段准确识别，故予以删除。

通过重构《语法等级大纲》的语法点，最终我们确定 554 条语法点作为抽取对象。根据各等级语法点的内容，我们按语法单位的不同粒度重新划分为语素层语法点、词层语法点、短语层语法点以及句层语法点四个层面。动作的态、提问的方法、强调的方法与特殊表达法按内容分别划分到词和短语两个层面中。

（三）语法点检索式的构造

不同类别的语法点对应着不同的文本特征，所谓语法点检索式是用某种模式表达文本特征，从而匹配一类字符串的一种公式。利用语法点检索式对输入的文本进行相应的模式匹配，可找出其中蕴含的特定语法点。

语法点检索式构造的基本步骤为：针对每一条语法点，或通过正则表达式，或在依存句法关系解析的基础上制定规则，最终形成相应的检索式。构造过程中，我们使用 Penn Chinese TreeBank（以下简称“CTB”）（Xue et al., 2005）的词性标签与 SD（Stanford Dependency）（De Marnee & Manning, 2008）依存标签。检索式由基本项、操作符和复杂项组成，其基本构成单元包括汉字、词性标签、依存句法标签、操作符、量词等。详见表 1。

表 1　基本项、操作符和复杂项的构成形式及示例

项目		构成形式	示例
基本项	字符项	由字符串和基本项表达式操作符构成	[word= 时候]
	词性标签项	由词性标签和基本项表达式操作符构成	[tag=VV]
	命名实体项	由命名实体名称和基本项表达式操作符构成	[entity=PERCENT]
	依存项	由依存标签和带有方向的依存弧操作符构成	[]>nsubj[]
	词汇难度项	由词汇难度等级和基本项表达式操作符构成	[level= 三]
操作符和量词		基本表达式操作符、正则表达式操作符	//、[]
		逻辑操作符、依存弧操作符、量词	|、&、<、>、+、*
复杂项		多个基本项与操作符、量词构成	略

依存句法分析（Dependency Parsing）能够揭示句子中词与词之间的依存关系，获取句子层面的构造信息，广泛应用于中文信息处理领域。形成的依存句法树，是由中心词与从属词通过两者之间的二元非对称关系连接而成的树状结构图。理论上，一个词具有与其他词构成多种依存关系的能力，但在具体的语法点中，能够实现的依存关系是有限且确定的。凭借依存关系的约束能够更加准确地识别语法点，而不需要过多考虑字面特征。

根据语法点构成成分和结构形式的不同，本文主要通过以下两种方式构造检索式：

1. 使用正则表达式构造，包括使用有序的结构组合及标点符号等形式手段。此方法适用于在形式上边界较为清晰的语法点，我们将其分为有指定标志词与无指定标志词两种。对于有指定标志词的语法点，使用正则表达式的语法直接构造检索式，例如词形与词性确定的词、固定短语及口语格式等。对于无指定标志词的语法点，则引入外部词表资源，依据词前后的搭配现象构造检索式。例如三级语法点“离合词”在大纲中的示例有限，而新

《等级标准》的《词汇等级大纲》中标识了 522 个离合词，我们引入袁毓林（2018）在本体研究中总结的类型制定规则，对这些离合词进行抽取；再如句子成分中三级语法点“复合趋向补语的趋向意义用法”，则通过引入汉语复合趋向动词词表来构造检索式。

2. 使用依存约束规则配合正则表达式构造。组成成分之间并不紧邻的语法点仅通过正则表达式无法准确抽取，此类语法点需结合前文所提的依存句法关系构造检索式。即根据大规模句子实例的依存句法树分析词性与依存关系的特点，获取当前语法点句法语义层面的构造信息，总结出结构上的规律，最后在基于字面特征的规则中加入依存关系的约束规则。如虚词语法点、短语的结构类型、短语的功能类型、固定格式、句式、特殊句型、复句等，通常采用语法模式“[word= 关键词 & tag= 词性标签](> 依存标签)”表征。

同一词的不同词性或用法分布在不同难度等级的语法点普遍存在，对于这些语法点，除了使用“关键词 + 词性标签”的基本模式之外，还需要通过观察其相邻词、搭配或总结依存树库中句子实例的依存关系进行限定。例如：结构助词“的 $_1$”属于一级语法点，一般用于定语后、名词前，作为定语的标志；语气助词“的 $_2$”属于二级语法点，常用在句末，表示肯定或已然。检索式如表 2 中例（1）、例（2）所示。

短语的结构类型中如“数量短语”的检索式，可以通过约束前一个词的词性为数词“CD”，后一个词的词性为量词“M”，依存关系为表示数量修饰的“mark:clf”来表征，如表 2 中例（3）所示。

表 2　各层面语法点实例及其检索式

句子示例	语法点	语法点检索式
（1）他是一个很认真**的**学生。	结构助词“的 $_1$”	[tag=/N.*|PN|VA/] >case [word= 的 &tag=DEG]
（2）明天不会下雨**的**。	语气助词“的 $_2$”	[ ]>discourse[word= 的 &tag=SP][tag=PU]
（3）我吃了**一个**面包。	数量短语	[tag=CD]>mark:clf[tag=M&level= 一]
（4）学校请他做报告。	兼语句 1：表使令	[]<nsubj[word=/ 叫 | 派 | 请 | 让 /&tag=VV] >ccomp [tag=/V.*/]
（5）**除了**春节**以外**，**还**有别的传统节日。	除了……（以外），……还 / 也 / 都……	[word= 除了 &tag=P]<case[]>case[word=/ 外 | 以外 /&tag=LC][word!=’，’]*[word=/ 还 | 也 | 都 /&tag=AD]

汉语中丰富的特殊句型、特殊句式通常也可以根据这种方式构造检索式，如“把”字句、被动句、连动句、兼语句等。以兼语句为例，汉语中兼语句的常见结构为“S（主语）+ 使令动词 +O（宾语）+VP”，我们以使令动词为锚点，其支配的宾语同时做第二个动词短语的主语，依据其与两个成分间的依存关系构造检索式，如表 2 中例（4）所示。

此外，汉语中还存在大量的固定格式，也称作“框式结构”，即由前后两个不连贯的词语相互照应、相互依存，形成一个相对凝固的框架式结构（邵敬敏，2008），如语法点“除了……（以外），……还 / 也 / 都……”，其检索式由表 2 中例（5）所示。

表 2 呈现了各层面语法点实例及其检索式。构造过程中，每条语法点对应 1~3 条检索式。由此，我们得到了 554 条语法点的 693 条检索式，各等级语法点检索式分布如表 3 所示。

表 3　各等级语法点检索式数量及比率

等级	一级	二级	三级	四级	五级	六级	高级	总计
数量	66	110	121	81	82	76	157	693
比例	9.52%	15.87%	17.46%	11.69%	11.83%	10.97%	22.66%	100%

（四）语法点自动抽取

文本中语法点的自动抽取流程主要分三步，如图 1 所示。

1. 调用自然语言处理工具 Stanford CoreNLP（Manning et al., 2014）对每篇文章进行预处理，获得分词、词性标注、依存分析等标注语料；

2. 将标注后的语料上传到自行搭建的检索平台中，创建索引以供检索；

3. 读入预先构建好的语法点检索式，通过云服务的方式调用已存检索式，逐一向检索平台发送对应语法点检索式的检索请求，检索平台检索并返回所有包含对应语法点的匹配结果。语法点抽取过程中，文本输入长度不受限制。

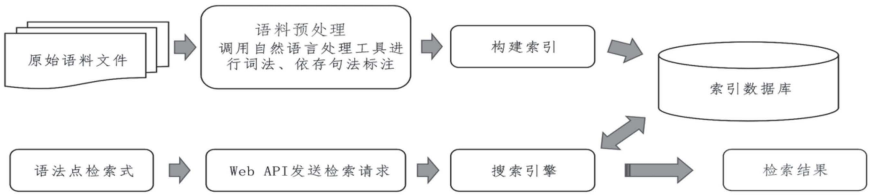

图 1 语法点自动抽取流程

通过上述方案，本文实现了构建语法点特征所需语法点的自动抽取，抽取实例如下：

（1）北京有座美丽的中山公园，公园里有个用五色土砌成的社稷坛。（《成功之路 · 成功篇》第 2 册《古迹今日》首句）

一级语法点及频数：结构助词“的$_1$”（2 次），方位名词“里”（1 次），量词“个”（2 次），“有字句”（2 次），陈述句（1 次）；

二级语法点及频数：偏正短语（3 次），名词性短语（3 次），方位短语（1 次）。

三 自动分级实验

上一节详细介绍了汉语语法点在文本中的自动抽取工作，本节将说明语法点的量化过程，并基于国际中文教育教材语料库，通过不同的算法来构建自动分级模型，验证所提语法点特征的有效性。另外，本节还将引入 BERT 模型提取文本的深层语义特征，融入外部语言特征，探究外部语言特征在模型中的作用。

（一）实验数据及评测指标

对于汉语二语文本难度自动分级来说，构建具有级别标注的文本数据集是进行研究的基础。为获取较大规模的数据，同时利用较为权威的级别信息来进行标注，我们选取国际中文教育领域广泛使用的 16 套 76 册教材的主课文，通过 OCR 和人工校对的方法，构建了国际中文教材语料库。所用教材有《标准中文》《博雅汉语》《成功之路》《发展汉语》《体验汉语》《新标准汉语》等。这 16 套教材包含了从面向汉语初学者到高级学习者所有级别的文本，完整文本共 2000 篇。我们收录了其所包含的初级到高级的教材文本。

剔除了诗歌、表单等特殊题材与重复出现的文本后，本文共使用 1929 篇文本（44731 句）作为实验语料。由于每套教材划分的级别数量不一，有些教材仅覆盖部分级别，我们通过人工标注的方法对文本的难度级别进行重新标注和调整。根据教材的制定标准与使用说明，我们对所有教材（册）依照难度进行排序，在基本还原文本所在教材原级别的原则下，对部分难度与原级别不相符的文本级别进行了调整。最终，将 1929 篇文本标注为初、中、高三个级别。本文对语料库中的所有文本进行了统计分析，各级别文本的统计信息如表 4 所示。从中可以看出，从初级到高级，其平均文章长度、平均句数都有较为明显的增加。

表 4　汉语二语教材分级语料库统计信息

文本级别	数量 / 篇	数量 / 句	占比 /%	平均文章长度 / 字	平均句子数量 / 句
初级	782	10580	40.54%	180.33	13.53
中级	664	14235	34.42%	547.34	21.44
高级	483	19556	25.04%	1399.91	40.57
所有级别	1929	44371	100%	611.62	23.01

为了保证评估结果的可靠性，实验采用五折交叉验证的方法对模型的内部效度进行考察。随机抽取 229 篇文本作为测试集，其余 1700 篇文本被分为 5 组，每次使用四组进行训练，一组作为开发集调试参数。使用准确率（Accuracy，ACC）和 F_1 值（F–Score，F1）对模型性能进行评估和比较，作为评价自动分级性能的评测指标。

（二）语言特征量化与特征抽取

1. 语法点特征

本研究从多样性和复杂性两个角度对语法点进行量化。语法点与词汇的习得过程在很大程度上存在相似性，词汇层多样性特征和复杂性特征的量化指标为我们提供了参考。

在多样性方面，最常用类符（Type）和形符（Token）之比（Type Token Ratio，简称为“TTR”）来衡量（Jarvis，2013）。我们通过类符数 / 开根形符数（Root Type Token Ratio，RTTR）（Torruella et al.，2013）计算每篇文章的语法点多样性，取 Root 值减轻文本长度对 TTR 的影响。语法点多样性包括各等级语法点的多样性和所有语法点的多样性，即对于单篇输入的文本分别抽取每个级别的语法点，计算每篇文本中各等级的语法点各自出现的次数与所有等级的语法点共出现的次数作为形符数，语法点的个数作为类符数。

语法点复杂性的测量主要通过考查语法点的变化和结构的复杂性。Wolfe–Quintero 等（1998）从频率（frequency）和比率（ratio）两个方面归纳了可以衡量语法点复杂性的若干测量指标。据此，我们将各等级语法点在文本中出现的数量与比率作为语法点复杂性的两种量化指标。现有研究在衡量词汇复杂性时主要根据频次设置相应阈值选出较为复杂的词语，进而计算文中复杂词占比（Advanced Guiraud index，AG）作为复杂性指标（Daller et al.，2003）。《语法等级大纲》本身反映了汉语学习者对语法点的接触顺序与接触频率，本研究将中等语法点与高等语法点视为复杂语法点，计算复杂语法点出现的比率。同时，参考 Sung 等（2015）对词汇难度的量化方法，提出语法点难度等级均值（Mean of Grammar Structures Difficulty，MGSD）和语法点难度等级方差（Variance of Grammar Structures Difficulty，VGSD）的概念。

据此，我们提出了 25 个语法点特征，包括 8 个语法点多样性特征与 17 个语法点复杂性特征，名称及其计算公式参见表 5。

2. 传统语言特征

除了前文所提的语法点特征之外，考虑到现有的汉语二语文本复杂度特征研究已取得较好的成果，本实验还引入了前人研究用来衡量字词、句法、搭配、依存、语法项目[③]五个层面的语言特征（以下合称“传统语言特征”）（Wang & Hu，2021）。不同的是，依存层面特征采用 CTB 的中文依存关系标签进行计算。各层面特征构成及数量如表 6 所示。

本文构建的语言特征体系共包括语法点特征在内的六个层面 207 项具体度量特征，采用自然语言处理技术对各个层面特征进行自动抽取与计算。本文调用 CoreNLP 自然语言处理工具包对文本进行预处理，得到文本分词、词性标注、命名实体识别、依存关系和短语

句法树等结果。分别对字词、句、依存、搭配、语法项目、语法点层面的语言特征进行提取，通过 Python 编程进一步计算语言特征值。

表 5 语法点特征描述及公式

类别	特征（数量）	公式
语法点多样性特征	各等级语法点的多样性（7）	$\mathrm{Type}_i/\sqrt{\mathrm{Num}_i}$
	所有语法点的多样性（1）	$\sum_{i=1}^{7}\mathrm{Type}_i/\sqrt{\sum_{i=1}^{7}\mathrm{Num}_i}$
语法点复杂性特征	各等级语法点数量（7）	Num_i
	各等级语法点比率（7）	$\mathrm{Num}_i/\sum_{i=1}^{7}\mathrm{Num}_i$
	复杂语法点占比（1）	$\sum_{i=4}^{7}\mathrm{Freq}_i/\sum_{i=1}^{7}\mathrm{Num}_i$
	语法点难度等级均值（1）	$\sum_{i=1}^{7} i*\mathrm{Type}_i/\sum_{i=1}^{7}\mathrm{Num}_i$
	语法点难度等级方差（1）	$\sum_{i=1}^{7}\mathrm{Num}_i*(i\ -\mathrm{MGSD}^2)/\sum_{i=1}^{6}\mathrm{Num}_i$

注：i 表示语法点等级，Type_i 表示第 i 级语法点在文章中出现的类符数，Num_i 表示第 i 级语法点在文章中出现的形符数

表 6 传统语言特征构成及数量

特征层面（数量）	特征名称
字词层面（4）	字数、词数、词汇多样性、词汇复杂性
句层面（7）	平均大句长、平均小句长、平均 T 单位长、平均小句数、平均 T 单位数、平均句法树深度、最大句法树深度
搭配层面（23）	搭配整体多样性、汉语特有搭配多样性、低频搭配比例、不同类型搭配多样性、不同类型搭配比例
依存层面（133）	依存多样性、依存比例、依存距离
语法项目层面（15）	语法项目多样性、语法项目密度、语法项目比例、低级语法项目密度、低级语法项目比例、高级语法项目密度、高级语法项目比例

（三）语法点特征的有效性验证

考虑到不同的机器学习算法在难度预测中的性能有异，为了全方位验证本文所提语法点特征的有效性，实验部分采用常用的五种不同的机器学习方法来构建模型。具体参数设置分别为：

• 多元线性回归：线性回归是指将一个多维线性函数尽可能地拟合到所有数据点的过程。模型通过建模多维线性函数拟合训练数据，使用默认参数设定。

• 随机森林模型：使用 bootstrap 抽样从训练集全集 D 中抽取 k 个样本分别建立决策树模型，将测试集数据输入 k 个决策树模型后，通过投票表决预测其最终分类。

• XGBoost：XGBoost 是梯度增强算法 GDBT 的一个改进版本。在实验中，最大树深度限制为 3，估计数限制为 300，学习率设置为 0.05，gamma 设置为 5。

• 逻辑回归：逻辑回归是一种用于二元分类的广义线性模型。实验分别使用有序逻辑回归与多元逻辑回归两种算法，最大迭代阈值设置为 1000，以确保算法尽可能收敛。

我们以 Wang 和 Hu（2021）基于 TF–IDF 的词袋向量作为输入构建的模型作为基线模型。分别使用 25 个语法点特征、82 个传统语言特征、207 个语法点特征与传统语言特征的融合特征为输入建立文本难度自动分级模型。由于特征繁多，为了避免特征间存在共线性导致过度拟合，我们首先对特征进行降维优化。每个模型的训练过程中均使用逐步线性回归方法选择最优特征集，只有当特征与文本等级有显著相关性，且特征之间不存在共线性依赖关系，即方差膨胀因子（Variance Inflation Factor，VIF）小于 10 时才能进入回归分析。

实验基于 Skicit–Learn 机器学习框架进行。具体结果如表 7 所示。

表 7　各层面语言特征在机器学习方法上的结果对比

特征（数量）	线性回归		随机森林		XGBoost		有序逻辑回归		多元逻辑回归	
	ACC	F_1	ACC	F_1	ACC	F_1	ACC	F_1	ACC	F_1
TF-IDF (6000)	70.9%	43.1%	81.1%	80.6%	75.0%	73.9%	77.5%	76.1%	80.2%	78.5%
语法点特征 (25)	72.1%	44.0%	81.9%	81.8%	73.4%	72.6%	82.9%	81.9%	83.9%	83.2%
传统语言特征 (182)	72.4%	44.2%	82.3%	81.7%	81.6%	79.9%	83.4%	83.3%	84.0%	83.6%
传统语言特征及语法点特征 (207)	73.7%	44.8%	83.3%	83.7%	82.7%	80.1%	83.8%	84.1%	86.4%	85.7%

由表 7 可知，在五种机器学习方法中，随机森林、有序逻辑回归和多元逻辑回归在各个层面的特征上两项评测指标均达到 80% 以上，而线性回归与 XGBoost 模型的性能较弱，准确率与 F_1 值均未超过 80%。其中，多元逻辑回归算法在综合评价指标上均略高于有序逻辑回归，相较于其他算法则有明显优势，取得最高的预测准确率为 86.4%，F_1 值为 85.7%。

单独使用语法点特征与传统语言特征构建的模型预测准确率都在基线模型以上。最优结果出现于基于多元逻辑回归构建模型，仅使用25个语法点特征达到了83.9% 的准确率，而使用 182 个传统语言特征达到了 84.0% 的准确率，只高出单独使用语法点特征 0.1%。由上表可见，无论使用哪种机器学习模型，加入语法点特征后的模型表现均为最佳，准确率与 F_1 值均高于未加入之前模型的预测准确率。表现最好的模型是依旧是基于多元逻辑回归，预测准确率为 86.4%，相较于未加入语法点特征之前提升了 2.4%。可见语法点特征的融合能够提升文本分级模型的性能。这也有力地表明，基于《语法等级大纲》构建的语法点特征对汉语二语文本自动分级的准确率有明显贡献，验证了本文所提语法点特征在汉语二语文本自动分级任务上的有效性。

（四）融合深层语义特征的汉语二语文本自动分级

本文将二语文本难度自动分级视为文本分类任务。前人研究表明，BERT 在多种分类任务上表现优异（Sun et al., 2019）。BERT 模型是基于 Transformer 编码器对大规模语料进行训练的预训练语言模型（Devlin, 2018），在自然语言处理领域及其众多下游任务中应用广泛。经过大规模语料的训练，BERT 具有优秀的编码文本语义特征的能力，能够在一定程度上表征文本的深层语义信息。那么，BERT 模型在汉语二语文本自动分级任务上表现如何？在 BERT 深层语义特征的基础上加入外部语言特征能否提升分级模型的准确率？

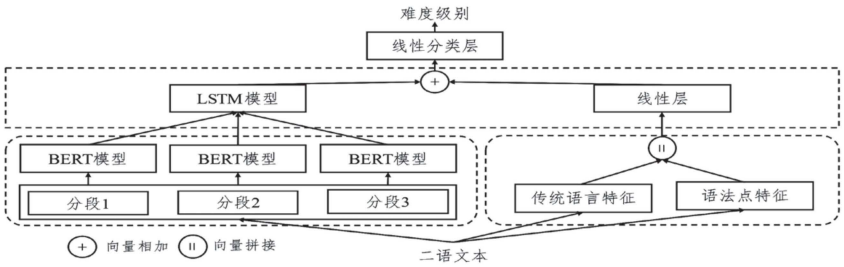

图 2　BERT-LING 模型结构

为考察以上问题，本文使用 BERT 中文模型作为自动分级模型的底层构架展开对比试验。实验使用的模型结构如图 2 所示，包括 BERT 语义特征抽取、外部语言特征抽取（包括传统语言特征抽取与语法点特征抽取）、特征融合及分类器四个部分，下文称之为 BERT-LING 模型。首先通过 BERT 语义特征抽取得到深度语义特征，与外部语言特征进

行融合后输入到线性分类器中，获得难度级别。在 BERT 语义特征抽取部分，先对文本进行分段编码后再使用 LSTM 融合各段文本的语义特征，获得整个二语定级文本的深度语义特征编码向量。需要注意的是，我们以向量拼接的方式融合了传统语言特征和语法点特征。特征融合部分利用 LSTM 模型融合多个分段文本的特征，再使用向量加和的方式与经过线性变换的语言特征进行融合。本文以前文中多元逻辑回归的实验结果作为基线，将此模型的预测准确率与 F_1 值对比分析。实验结果如表 8 所示。

表 8　融合外部语言特征和 BERT 语义特征的分级实验结果

模型	语言特征	ACC	F_1
多元逻辑回归	外部语言特征	86.4%	85.7%
BERT	BERT 深度语义特征	86.8%	85.2%
BERT-LING	BERT 深度语义特征及外部语言特征	87.6%	86.4%

从表中可见，未加入外部语言特征之前，基于 BERT 的模型与基于外部语言特征训练的多元逻辑回归在分类准确率上相差无几。在 BERT 语义特征的基础上融合外部语言特征后，相比于多元逻辑回归模型与 BERT 模型，BERT–LING 在准确率和 F_1 值两方面都有所提升，准确率达到 87.6%，F_1 值达到 86.4%。

实验结果表明，BERT 模型提取到的深度语义特征在一定程度上代表了文本的语义信息，在融入外部语言特征信息之后，模型在理解文本语义信息之外还学习到了语言特征知识，从而在自动分级时将外部语言特征对文本难度的贡献考虑在内，达到最佳的分级效果。

四　讨论与分析

通过前文分析，我们发现语法点特征的加入能够明显提升汉语二语文本难度自动分级模型的准确率。那么，各个等级的语法点在国际中文教材语料库中的分布如何？哪些语法点特征对文本难度的解释能力更强？为了探寻以上问题，本文做了进一步的实验分析。

本文首先对九个等级的语法点在数据集中的分布进行了统计，表 9 直观地展示了一到九级的语法点在三个等级的文本中出现的数量及比率。可以看到，三个等级文本中各等级语法点的分布有着清晰的界限。随着本文难度级别的升高，文本所涉及的语法点频数整体呈递增趋势；除了一级语法点占比降低，二级到高级的占比均呈现升高的趋势。初级文本中一级语法点的比例高于中、高级文本，这意味着从初级到中级的难度增加来自于更高级别的语法点使用的增加。即在越难的文本中，中、高级语法点出现的频次越多，文本的阅读难度越高，反之，中、高级语法点出现的频次越少，文本的阅读难度越低。

同样明显的是，无论该文本属于汉语水平等级中的哪一位置，初等语法点都是教材文本语法点中占比最高的一部分，而随着教材等级的提高，初等语法点的比例略有降低，但仍占到教材文本语法点总量的 90% 以上，说明初等的语法点基本上可以构成大多数汉语的句子。即使在高级的文本中，占比最高的仍是一级语法点，即大多数的句子都是由基础的语法点构成的，而高等的语法点只是偶尔出现。

为了更进一步研究语法点特征与二语文本难度之间的相关性，探寻哪些语法点特征对文本难度的解释能力更强，本文对单独使用语法点特征进行多元逻辑回归实验中的逐步线性回归的结果进行分析。经检验，共 16 个指标满足进入回归分析的条件，其余因与其他指标存在共线性关系而被剔除。其逐步回归分析结果详见本文附录。16 组回归模型的系数均达到显著（p<0.001），回归模型的残差经直方图、P–P 图检验符合正态分布。由此可见，

表 9　各等级语法点在数据集中的分布

语法点等级		初级文本			中级文本			高级文本		
		个数	比例		个数	比例		个数	比例	
初等	一级	62,007	58.79%	96.66%	121,743	51.30%	94.60%	185,415	49.06%	93.00%
	二级	26,357	24.99%		65,724	27.70%		108,369	28.67%	
	三级	13,585	12.88%		37,035	15.61%		57,715	15.27%	
中等	四级	2,305	2.19%	3.32%	7,973	3.36%	5.24%	15,461	4.09%	6.69%
	五级	468	0.44%		1,720	0.72%		2,961	0.78%	
	六级	730	0.69%		2,735	1.15%		6,879	1.82%	
高等		19	0.02%	0.02%	370	0.16%	0.16%	1,172	0.31%	0.31%

该回归模型是有效的，语法点特征一共可以解释文本难度变异的66.5%，R^2达到大效应量。其中，六级语法点多样性 SIX_RTTR 单个指标可以解释文本难度的 48.4%，是对文本难度预测贡献最大的指标。由此可以得知，对文本难度区分性最好的是六级语法点多样性，可以作为衡量文本难度水平的最佳指标。

五　总结与展望

本文以汉语二语文本难度自动分级为目标，以国家新发布的《国际中文教育中文水平等级标准》（2021）为依据，基本实现了汉语语法点的自动识别，并在此基础上提取了反映汉语语言特点的 25 个衡量语法难度的语法点特征，构建了基于语法点特征的汉语二语文本自动分级模型。实验结果验证了本文提出的语法点特征在汉语二语文本自动分级任务上的有效性。此外，本文将深度学习方法应用于汉语二语文本自动分级，通过融合文本语言特征与基于 BERT 的深度语义特征取得了较好的效果，从而证明了融合语言特征能提升模型的难度表征能力。相比以往研究，本文的方案达到了目前同任务上较高的准确率87.6%，也证实了从汉语的语言特点与第二语言习得的视角出发构建语法点特征的重要性。

在未来的工作中，我们将继续对语法点特征进行细化，进一步探究哪些类型的语法点在二语文本自动分级中发挥更为重要的作用。另外，我们将推出汉语二语文本难度分级平台，辅助二语教师进行教材编写的同时，根据汉语二语学习者的语言水平推荐适合其阅读的课内外辅助材料，以期为国际中文教育领域的发展提供参考和建议。

［附　注］

① 中文 Coh-Metrix 网址参见 http://141.225.61.35/cohmetrix2017。

② CRIE 网址参见 http://www.chinesereadability.net/CRIE/。

③ 参考 2009 版《国际汉语教学通用课程大纲》中的“常用汉语语法项目分级表”。

［参考文献］

［1］蔡永强 . 汉语读写教学一体化研究［M］. 北京语言文化大学出版社，2017.

［2］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，（4）.

［3］吕文华 . 对外汉语教材语法项目排序的原则及策略［J］. 世界汉语教学，2002，（4）.

［4］邵敬敏 . “连 A 也 / 都 B” 框式结构及其框式化特点 ［J］. 语言科学 , 2008,（4）.
［5］孙　刚 . 基于线性回归的中文文本可读性预测方法研究 ［D］. 南京大学硕士学位论文 , 2015.
［6］孙未未 , 夏　菁 , 曾致中 . 基于回归模型的对外汉语阅读材料的可读性自动评估研究 ［J］. 中国教育信息化 , 2018,（15）.
［7］王鸿滨 . 汉语国际教育汉语文本分级及难度测查对比研究 ［J］. 云南师范大学学报（对外汉语教学与研究版）, 2020,（6）.
［8］王鸿滨 .《国际中文教育中文水平等级标准》中语法等级大纲的研制路径及语法分级资源库的开发 ［J］. 国际汉语教学研究 , 2021,（3）.
［9］王　蕾 . 初中级日韩留学生文本可读性公式初探 ［D］. 北京语言大学硕士学位论文 , 2005.
［10］王　蕾 . 文本可读性公式研究发展阶段及特点 ［J］. 语言教学与研究 , 2022,（2）.
［11］王　丽 . 基于语法知识的英文文本分级和读物推荐系统 ［D］. 西安电子科技大学硕士学位论文 , 2020.
［12］王钟华 . 初级阶段汉语教学四题 ［J］. 语言教学与研究 , 1999,（3）.
［13］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 ［J］. 世界汉语教学 , 2020,（1）.
［14］杨金余 . 高级汉语精读教材语言难度测定研究 ［D］. 北京大学硕士学位论文 , 2008.
［15］袁毓林 . 从形式转喻看离合词分开使用的句法性质 ［J］. 当代语言学 , 2018,（4）.
［16］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）.
［17］Curto P, Mamede N J, Baptista J. Automatic text difficulty classifier - assisting the selection of adequate reading materials for european portuguese teaching［C］. In Proceedings of CSEDU, 2015.
［18］Daller H, van Hout R, Treffers-Daller J. Lexical richness in the spontaneous speech of bilinguals［J］. *Applied Linguistics*, 2003.
［19］De Marnee M C, Manning C D. The Stanford Typed Dependencies Representation［C］. In Proceedings of COLING, 2008.
［20］Heilman M, Collins-Thompson K, Callan J. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts［C］. In Proceedings of NAACL, 2007.
［21］Jarvis S. Capturing the diversity in lexical diversity［J］. *Language Learning*, 2013.
［22］Jeon E hee, Yamashita J. L2 reading comprehension and its correlates: A Meta-Analysis［J］. *Language Learning*, 2014.
［23］Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit［C］. In Proceedings of ACL, 2014.
［24］Schwarm S E, Ostendorf M. Reading level assessment using support vector machines and statistical language models［C］. In Proceedings of ACL, 2005.
［25］Sung Y T, Lin W C, Dyson S B. Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR［J］. *The Modern Language Journal*, 2015,（2）.
［26］Torruella J I, Capsada R. Lexical statistics and tipological structures: A measure of lexical richness［J］. *Procedia - Social and Behavioral Sciences*, 2013.
［27］Vajjala Balakrishna S. Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications［D］. Universität Tübingen, 2015.
［28］Wang S, Andersen E. Grammatical templates: Improving text difficulty evaluation for language learners［C］. In Proceedings of COLING, 2016.

［29］Wang Y, Hu R. A prompt-independent and interpretable automated essay scoring method for chinese second language writing［C］. In Proceedings of CCL, 2021.

［30］Wolfe-Quintero K, Inagaki S, Kim H. *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*［M］. Honolulu: Second Language Teaching & Curriculum Center, The University of Hawai'i., 1998.

［31］Xue N, Xia F, Chiou F D. The penn chinese treebank: Phrase structure annotation of a large corpus［J］. *Natural language engineering*, 2005,（2）.

附录 1　单独使用语法点特征进行逐步线性回归的结果

模型	新增特征	特征名称	R	R^2	调整后 R^2	更改统计	
						R^2 变化量	F 变化量
1	SIX_RTTR	六级语法点多样性	.696[a]	0.484	0.484	0.484	1805.682
2	MEAN	语法点难度等级均值	.747[b]	0.558	0.558	0.074	324.414
3	ONE_RTTR	一级语法点多样性	.779[c]	0.607	0.606	0.049	237.786
4	SEVEN_RTTR	七级语法点多样性	.789[d]	0.622	0.622	0.015	77.962
5	FOUR_RATIO	四级语法点占比	.793[e]	0.629	0.628	0.007	36.248
6	FIVE_RTTR	五级语法点多样性	.797[f]	0.635	0.634	0.006	29.686
7	TWO_RATIO	二级语法点占比	.801[g]	0.641	0.640	0.006	34.341
8	TWO_RTTR	二级语法点多样性	.805[h]	0.647	0.646	0.006	32.781
9	TOTAL_RTTR	所有语法点多样性	.808[i]	0.653	0.652	0.006	32.538
10	FIVE_NUM	五级语法点数量	.811[j]	0.657	0.656	0.004	22.907
11	TWO_NUM	二级语法点数量	.812[k]	0.660	0.658	0.002	14.016
12	THREE_RTTR	三级语法多样性	.814[l]	0.662	0.660	0.002	12.164
13	SEVEN_NUM	七级语法点数量	.814[m]	0.663	0.661	0.001	5.233
14	SEVEN_RATIO	七级语法点占比	.816[n]	0.666	0.663	0.003	14.553
15	VARIANCE	语法点难度等级方差	.817[o]	0.667	0.664	0.001	6.844
16	THREE_NUM	三级语法点数量	.817[p]	0.668	0.665	0.001	5.763

（责任编辑　常文斐）