现代汉语词汇语体属性探测模型研究 *

莫凯洁　胡韧奋

（北京师范大学　北京　100875）

[摘要] 本文立足于正式—非正式的语体维度，提出了基于机器学习方法的现代汉语词汇语体属性探测模型，旨在实现符合语体连续统特性的词语正式度测量。研究首先构建了现代汉语语体语料库，设计了语体分类特征，并基于《现代汉语词典》（第 7 版）中的〈书〉〈口〉标注数据训练语体属性自动分类模型。模型五折验证准确率达 87.26%。进一步的误例分析发现：词典中的语体标注存在部分缺漏、过时、不对称等问题，而基于语体语料库的语境特征能有效修正数据偏差。为了更好地服务词汇语体教学，本研究使用上述模型对《国际中文教育中文水平等级标准》词表和《义务教育常用词表（草案）》主表的共 25500 个词语进行了语体正式度测量，并分析了该方法在词典编纂和教学方面的应用。

[关键词] 语体特征；正式度；机器学习；语体词表

[中图分类号] H087 [文献标识码] A [文章编号] 1003–5397（2023）04–0118–14

DOI:10.16499/j.cnki.1003-5397.2023.04.004

An Automatic Detection Model of Register Attributes for Vocabulary in Mandarin Chinese

MO Kaijie, HU Renfen

Abstract: This paper introduces a model for detecting register attributes, particularly the degree of fomality, of the Mandarin Chinese vocabulary using machine learning algorithms. The study encompasses a multi-register Mandarin Chinese corpus, designing register classification features, and building an automatic register classification model. This model utilized the 〈written〉 and 〈spoken〉 tags from lexical entries in Modern Chinese Dictionary (7th edition). The experimental results showed that the model achieved an accuracy of 87.26% through five-fold cross-validation. Further analysis revealed issues such as missing, outdated, and asymmetric annotations in the dictionary data. To mitigate these issues, contextual features extracted from register-specific corpora were integrated, exhibiting substantial improvement in correcting data biases. Additionally, the model was employed to evaluate the formality degrees of 25,500 words

[收稿日期] 2023-08-10

[作者简介] 莫凯洁，北京师范大学国际中文教育学院硕士生，主要研究计算语言学；胡韧奋（通讯作者），北京师范大学国际中文教育学院教师，博士，主要研究计算语言学。

* 本研究得到教育部中外语言交流合作中心国际中文教育中外联合研究专项课题“基于新标准的智能化语言分析技术研究”（22YH04ZW）和中央高校基本科研业务费专项资金（北京师范大学优秀青年创新团队项目“基于数字人文的《说文》学跨学科研究”）资助。本研究得到“《语言文字应用》青年学者论学”第二期专家学者的指导，谨此致谢。

from the vocabulary list in the Chinese Proficiency Grading Standards for International Chinese Language Education and the List of Common Words in Compulsory Education (Draft, primary list), thereby facilitating register-focused lexical instruction. Finally, this paper discussed the application of the model in dictionary compilation and vocabulary teaching.

Keywords: register features; formality; machine learning; register-focused vocabulary list

一　引言

语体是不同交际领域中具有差异的语言体系，承载着人类语言的社会性，在人类交际中起着重要作用。学界对语体的定义较多，王德春（1987）认为，语体就是语言素材在各个交际环境中形成的系列特征集合。刘大为（1994）指出，语体是一种语言在交际过程中产生变异而形成的特征集合体。冯胜利、施春宏（2018）将语体界定为“实现人们在直接交际中具有元始属性的、用语言来表达或确定彼此之间关系和距离的一种语言机制”。总体而言，语体和交际息息相关，不同交际距离中的语体存在各自不同的语言特征。目前学界通常将语体划分为口语体和书面语体（邵敬敏，2016；胡裕树，2019），也有学者提出了更细化的区分方式，如李文明（1994）将语体划分为科学、艺术和应用三大类，每类下设书面语体、口语体；冯胜利（2010）提出，正式与非正式、典雅与便俗是构成语体的两对基本范畴。综上来看，正式与非正式是最基本的语体范畴。通常，交际者和交际对象之间的距离越远，所用的语体正式程度越高。

语体词指某类语体中常用且专用的词语（袁晖，2004），通常包括日常口头交际中使用的口语体词汇和在正式交际场合中使用的正式体词汇，以及不具有明显语体属性的通用体词汇（符淮青，2011；汪维辉，2014；冯胜利、王永娜，2017）。若一个语体词被使用在非所属语体的交际语境中，可能会导致所谓“不得体”情况。以下为一则汉语学习者的语用偏误示例（冯胜利，2015）：

（1）老师您媳妇儿在家吗？ *

（2）老师您妻子 / 夫人在家吗？

“媳妇儿”是口语体词。当面向交际距离较远，需以尊重态度去进行交际的对象“老师”时，交际者使用该词是不够得体的行为。当前研究表明，汉语学习者语体意识较为缺乏，对语体掌握存在困难（张翠吉，2018；马明艳，2017）。学界也意识到语体参与语言教学的重要性，李泉（2003）认为应改善“中性语体”在教学中占优势的局面，增加针对口语和书面语的教学，促进学习者的交际用语习得。冯胜利、王永娜（2017）指出在二语教学和本族语教学中，若要学生具有语体意识、学会得体表达，需进行语体分级教学研究。词汇正是语言教学中的关键点，汉语学习者学会区分和掌握汉语语体词十分重要。

然而，语体的边界模糊，给词汇语体属性的划定带来了挑战。从正式—非正式范畴来看，语体的过渡呈现为一个连续统（崔希亮，2020），而仅靠语感难以实现对连续统的精确描述。目前标注词语语体的方法主要有以下三种：

1. 人工标注：根据语言学理论及人的语感确定词语的正式程度或语体分类，如冯胜利和王永娜（2017）、黄国敬等（2022）。该方法以非连续统的方式进行分类标注，标注成本较高，且易受主观因素影响。

2. 从典型语料中提取高频词：选择较为典型的正式体语料和非正式体语料，统计两种语料中的高频词汇，从中提取语体词汇（潘先军，2021）。该方法的频度阈值较难确定，

且提取出的语体词易混入高频的通用词汇。

3. 对比词语在典型语料中的频度：选择典型语体语料库（如报刊、科技和对话语料），对比某词在不同语料库中的频度（张文贤等，2012；崔希亮，2020；黄劲怡、彭宣维，2022）。该方法的统计结果易受语料库规模影响，且其对语体的评估仅限于词频，刻画维度较为单一。

基于以上考虑，本文将从词典编纂、汉语教学实际需求出发，结合汉语语体特性及多维度语言特征，构建现代汉语词汇语体属性探测模型，以平衡客观数据与人类语感，促进解决语体标注模糊性难题。在理论层面上，本研究将分析不同语言特征对语体分类的影响，挖掘更多语体分类依据；在实践层面上，本研究利用语体语料库及词向量技术来抽取表层、深层语言特征，以构建现代汉语词汇语体属性探测模型，实现高效率的词汇语体属性量化分析；在应用层面上，本研究拟将词汇语体属性测量模型应用于二语教学、义务教育领域的词表自动标注，并探讨其在语言教学中的应用价值。

二　词汇语体属性探测模型构建

本文主要从正式—非正式的范畴出发，采用机器学习方法训练承载现代汉语语体知识的分类模型，进一步提取模型的预测概率，实现词语正式度的“连续统”测量。机器学习方法需从训练数据中提取特征，抽象出数据模型，从而对未知新数据进行预测和分析（李航，2019）。由于机器学习可以客观、定量地学习数据中的规律，并对大规模未知样本进行预测，该方法不仅成为自然语言处理、图像识别、语音识别等计算机应用学科的基础性方法，也在金融、医疗、教育等领域有广泛应用。在应用语言学方面，机器学习方法常用于文本分类、分级研究，如文本可读性分析（杜月明等，2022；吴思远等，2020）、作文自动评分（Attali & Burstein，2006；Wang & Hu，2021）等。接下来，本文将从数据、特征、模型三个层面介绍现代汉语词汇语体属性探测模型的构建方法。

（一）词汇数据

对于训练机器学习模型而言，词汇语体标注数据的质量十分重要。现有较为全面、系统的语体标注数据主要见于词典。其中，《现代汉语词典》（第 7 版）采用〈口〉〈方〉和〈书〉标记来分别标注口语词、方言词和书面文言词汇；冯胜利等（2020a）所编著的《汉语八百对单双音节对应词词典》使用通体、正式体、口语体、庄典体将单双音节对应词进行分类标注。此外，也有专门的语体词典，例如施光亨（2012）编著的《汉语口语词词典》、李行健（2022）主编的《现代汉语口语词典》。

本文旨在对现代汉语词汇的语体属性进行测量，需选择较为全面、系统且经多次校对的语体标注数据。《现代汉语词典》（第 7 版）（以下简称《现汉》）中标注了口语体词和书面文言词汇。其中，书面文言词汇属于书面语，是现代汉语中继承自文言文的书面语成分，属于现代汉语书面语的子集（孙德金，2012）。《现汉》中词汇语体标注数据来源较为权威，且质量高、数量多，因此本文选择《现汉》的语体标注数据为主要训练、测试数据。具体来说，从《现汉》中提取标注〈口〉的口语体词 838 个，标注〈书〉的书面语词 3002 个，并随机抽样 2500 个不带任何语体标注的通用体词汇。最终，从《现汉》中得到非正式语体词汇 3338 个，正式语体词汇 3002 个，其比例为 1.1:1，较为均衡。

（二）面向词汇语体分类的基础资源

本研究主要从正式—非正式语体范畴出发建立语体语料库，为抽取语体特征提供资源支撑。冯胜利（2010）指出，非正式语体和正式语体的区别在于一个是日常性的或亲密

随便的语言交际，另一个是非日常的或严肃庄重的语言交际。邹沁清、饶高琦（2021）对汉语文本进行了语体聚类，发现属于正式语体的文本有公文、学术文献、政论、新闻报道；属于非正式语体的文本有小说、散文、微博、歌词、谈话、问答。根据前人研究的分类依据，本文构建了较均衡的语体语料库，表 1 为语体语料库的具体情况。

表 1　语体语料库具体情况

语体	语料类型	容量	词数	词种	来源描述
正式语体语料	百度百科	4.2G	703M	10794K	Li 等（2018）发布的词向量训练数据 https://github.com/Embedding/Chinese-Word-Vectors
	学术文献	251M	44M	878K	Li 等（2022）发布的中文科学文献数据集 https://github.com/ydli-ai/csl
	新闻	1.8G	308M	1953K	人民日报新闻语料（2000~2017） http://data.people.com.cn/
	新闻	1.5G	264M	1398K	孙茂松等（2016）发布的新闻文本分类语料 http://thuctc.thunlp.org/
	新闻、学术文献	2.7M	0.464M	33K	北京外国语大学发布的 ToRCH 语料库 http://corpus.bfsu.edu.cn/info/1082/1782.htm
	正式场合发言	9M	1M	27K	北京理工大学 NLPIR 实验室发布的中国外交部例行记者会语料（2017~2020） http://www.nlpir.org/wordpress/2021/10/11/
非正式语体语料	百度问答	121M	21M	485K	百度发布的 QA 数据集 https://ai.baidu.com/broad/introduction
	社区问答	3.2G	626M	3433K	Xu（2019）发布的开源项目 https://github.com/brightmart/nlp_chinese_corpus
	小说	2.7M	0.50M	56K	北京外国语大学发布的 ToRCH 语料库 http://corpus.bfsu.edu.cn/info/1082/1782.htm
	小说、散文	1.3G	258M	2956K	Li 等（2018）发布的词向量训练数据 https://github.com/Embedding/Chinese-Word-Vectors
	微博	588M	104M	1430K	北京理工大学张华平博士发布的微博数据 http://www.nlpir.org/wordpress/2018/01/26/
	对话	583M	124M	1270K	GitHub 开源的个性化对话数据集 https://github.com/silverriver/PersonalDilaog
	对话	685M	148M	1231K	Wang 等（2020）发布的对话数据集 https://github.com/thu-coai/CDial-GPT
	对话	24M	4.9M	71K	Wang 等（2021）发布的对话数据集 https://ai.tencent.com/ailab/nlp/dialogue/#datasets
	对话	790M	160M	1420K	Codemayq（2018）开源的聊天数据集 https://github.com/codemayq/chinese_chatbot_corpus

如上表所示，本研究立足于正式体和非正式体的分类框架，建立了规模较大且分布均衡的语体语料库。其中，正式语体语料收录百科文本、学术文献和新闻报道等，共计 24 亿字符；非正式语体语料涵盖小说、散文、微博、对话和问答等，共计 20 亿字符。所有库中数据均经清洗，并基于 Python 程序库 Pyltp 工具实现了分词及词性标注，以便后续分析使用。

（三）词汇语体分类特征

语体特征是机器学习模型的分类依据。在文本语体定量研究中，已有学者使用多样的语体特征对文本进行语体属性评估（冯胜利等，2008；李果、王长林，2021；王用源、陈宇豪，2023）。本文也从多个角度展开了词汇语体特征设计，包括①词本身特征：就词语本身而言，其构词特征常表现出一定语体倾向，如音节数、词缀等，此外，词语在不同类型语料中的频率分布也能较好地反映其语体特点；②语料库语境特征：语体不同则语法各异（冯胜利，2012），不同语体的语法会对词汇的使用产生影响，因此，语料库中词语上下文的词汇选择、句法结构等特征会具有一定的语体差异；③词向量语义特征：词汇的语体分类依据既存在于语法形式之中，也作为所谓“语体色彩”存在于词汇的语义之中，人们在交际时会根据不同的交际场合选择具有不同语体色彩的词语。

综合考虑上述因素，本文设计了三类特征作为词汇语体的分类依据。图 1 以“考妣”为例，展示了各维度特征的提取思路及结果。接下来，本文将对三类特征分别进行介绍。

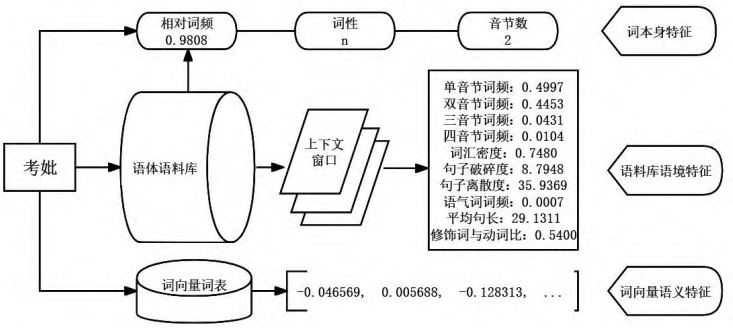


图 1 词汇语体分类特征概览

1. 词本身特征

主要统计了音节数、词性以及词语在语体语料库中的相对词频。相对词频的计算公式为：

$$R = \log\frac{F}{I}$$

其中，F 为词语在正式体语料库中的词频，I 为该词在非正式体语料库中的词频。例如，“妻室”在语料库中的相对词频为 0.4915，“老婆”在语料库中的相对词频为−1.5603，二者的语体正式度通过相对词频有所体现。

2. 语料库语境特征

除了词语本身的音节数、词性和相对词频等特征，我们还能通过该词在语料库中的上下文信息来测量其语体属性。受 Harris（1954）分布式语义假说的启发，本文以目标词为基点，从语料库中提取多个上下文窗口来表示该词的语境信息，多个窗口拼接而成的文本可以反映该词常出现的语言环境（后文简称为“语境文本”）。提取语境特征时，需要对上下文窗口大小、数量进行设置。本研究提取目标词的前 400 字和后 400 字作为上下文窗口，获得目标词的所有上下文窗口后进行随机抽样，使每个目标词上下文窗口数量保持在 3000 个以内，最后将所有上下文窗口所取文本合并成语境文本。参考学界的语体研究成果，本文进一步从语境文本中统计了如下特征：

a. 词音节信息：分别统计语境文本中单音节、双音节、三音节、四音节词所占比例。韵律是语体分化的重要体现，在单双音节对应词中，单音节词通常偏口语化，而双音节词

更正式（冯胜利等，2020b）。此外，“音节 + 音步”的形式构成悬差律偏向口语体特征，“音步 + 音步”的形式构成平衡律具有正式语体特征（王丽娟，2018）。

b. 修饰词与动词比（Modifier and Verb Rate，MVR）：语境文本中修饰性词语数量与动词数量之比。该值可反映文本的描写性和修饰性，数值较高表明文本语言生动，具有文学性；反之，则文本说明性强，话语较平淡（郜沁清、饶高琦，2021）。

c. 词汇密度：指语境文本中的实词比例。文本中包含实词数量越多，文本表达的实际意义更多，书面化程度越高（Ure，1971）。

d. 句子破碎度：统计语境文本中总字数除以句中因断句而停顿的次数。句中标点符号多代表停顿更多，文本口语性强；反之，文本流畅通达，正式性强（郜沁清、饶高琦，2021）。

e. 句子离散度：统计语境文本中每个句子的句长偏离平均句长的程度，即句长的总体标准差。离散度小则表明句子富有韵律性，平稳有秩；反之则跌宕起伏、富于变化（郜沁清、饶高琦，2021）。

f. 语气词词频：指语境文本中的语气词频率（语气词数量 / 总词数）。语气词主要在非正式语体中出现，在正式语体里较少使用（崔希亮，2020）。

g. 平均句长：统计语境文本中句子长度的平均值。相较于非正式语体而言，正式语体中句子通常较长（崔希亮，2020）。

3. 基于词向量的语义特征

除了利用语料库中的上下文窗口文本提取显式语境特征外，我们还可以通过词向量获取隐式语境特征。词向量模型可通过分析大规模文本语料库，捕捉词汇之间的上下文关系，并将其编码而得到向量表示（Mikolov et al.，2013）。在语义维度的研究中，中文词向量可以帮助研究者们更深入地探索词汇语义、汉字部件表义等信息（梁诗尘等，2021；胡楠等，2022）。本文旨在充分挖掘词语的现代汉语语体属性，需在覆盖面广、时效性强的语料中捕捉词汇的语体特征，因此选择了 Li 等（2018）在中文混合语料库（涵盖新闻、百科、问答、微博等语料，共 15G）上训练的词向量作为语义特征，该词向量为 300 维，词表大小为 129 万词，训练时额外引入汉字、n-gram 等特征，中文语义表示效果较强。

（四）探测模型

研究将正式—非正式维度的词汇语体属性分析问题建模为二分类任务，使用多种机器学习模型构建分类器开展实验，包括支持向量机（Support Vector Machine，SVM）、岭回归分类（Ridge Regression Classifier，RRC）、随机森林（Random Forest，RF）和逻辑回归（Logistic Regression，LR）。实验中，首先针对上文介绍的词典标注数据提取词汇语体分类特征，将其输入模型进行训练，让模型学习从特征到标签（正式、非正式）的映射；然后，在测试集上进行评测，获得可靠分类效果后，便能够对任意词语进行预测。进一步地，提取模型预测词语为正式标签的置信度作为该词的正式度。与以往词汇语体标注研究不同的是，本文首次为词语标注符合语体连续统特性的正式度值，该标签能更精确地揭示词语的语体属性差异。

三　实验与分析

（一）实验设置

首先，对来自《现汉》的标注数据进行过滤，剔除在语料库、词向量词表中未收录的词，共计得到 5422 词。包括 2433 个正式语体词、2989 个非正式语体词。然后，依据上文方法抽取词语特征，并按照 8:2 分割训练集和测试集。在构建机器学习模型时，将 SVM

的 probability 参数调整为 True，RRC 的 alpha 参数设置为 0.3，LR 的 penalty 设置为“l1”，solver 设置为“liblinear”，RF 的 max_depth 参数设置为 4，random_state 参数设置为 0。除以上参数外，其他模型参数均保持默认值。

（二）实验结果及分析

实验对比不同特征和模型的组合效果，各组模型均在完整数据集上采用五折交叉验证来计算平均准确率，结果如表 2 所示。相较于单独使用词本身特征、语料库语境特征，大部分情况下，组合使用这两类特征时，模型分类预测效果会更好，说明这两类特征为有效且互补的词汇语体分类依据。单独使用词向量特征训练出来的模型准确率处于较高水平，在 SVM 模型中达到了 87.26%，表明词向量特征能够较好地捕捉词汇的语体特点。

表 2　模型的五折验证准确率（%）

特征 \ 模型	SVM	RRC	LR	RF
词本身特征	65.97	63.42	63.03	66.30
语料库语境特征	59.72	66.53	66.78	69.10
词向量特征	87.26	86.52	86.30	83.29
词本身特征 + 语料库语境特征	60.18	68.90	68.51	71.89
词本身特征 + 词向量特征	87.07	87.12	86.20	82.92
语料库语境特征 + 词向量特征	84.45	86.51	86.37	83.06
全部特征	84.34	87.02	86.32	82.57

如前文所述，我们可以提取模型预测正式体标签的置信度，用于表示词语的正式程度。表 3 为使用基于词向量特征训练的 SVM 模型在单独划分的测试集（非正式体词 598 个，正式体词 487 个）上进行标注的示例，测试集的正式度阈值区间为 [0.0006，0.9999]。可以看出，不同正式度区间的词语呈现出明显的语体差异，正式度偏高的词语更正式、庄重，正式度偏低的词语更通俗、随意，模型也能区别出具有细微语体差别的词，例如“赧然（0.9942）”和“羞赧（0.7957）”。

表 3　SVM 模型在测试集上的预测示例

正式度区间	示例
正式度高 （0.8~1）	鼎革 _v: 0.9875；勋业 _n: 0.9904；天阙 _n: 0.9871 赧然 _a：0.9942；弱冠 _n: 0.9656；阿附 _v: 0.9633
正式度较高 （0.6~0.8）	归心 _v: 0.7860；敌忾 _v: 0.7370；潇潇 _a: 0.7781 羞赧 _a：0.7957；危亡 _v: 0.7516；冶艳 _a: 0.7319
正式度一般 （0.4~0.6）	劳逸 _n: 0.4761；养气 _v: 0.5714；心潮 _n: 0.4069 独 _a: 0.4802；偏颇 _a: 0.5584；一瞬 _nt: 0.5473
正式度较低 （0.2~0.4）	延展 _v: 0.3172；真确 _a: 0.3337；见外 _a: 0.2410 不合 _v: 0.3898；牛饮 _v: 0.3179；丧气 _a: 0.3424
正式度低 （0~0.2）	脆生 _a: 0.0423；撒尿 _v: 0.0107；歪 _v: 0.0068； 姑子 _n: 0.0488；两口儿 _n: 0.0318；前儿 _n: 0.0114

为探测词本身特征和语料库语境特征的重要性，实验使用特征排列法计算了各个特征的重要性分数。特征排列法指观察单个特征值被多次随机打乱后模型分数下降的均值，均值越高说明特征越重要（Breiman，2001）。图 2 为 RF 模型（词本身特征 + 语料库语境

特征训练）使用特征排列法后得到的各个特征重要性。

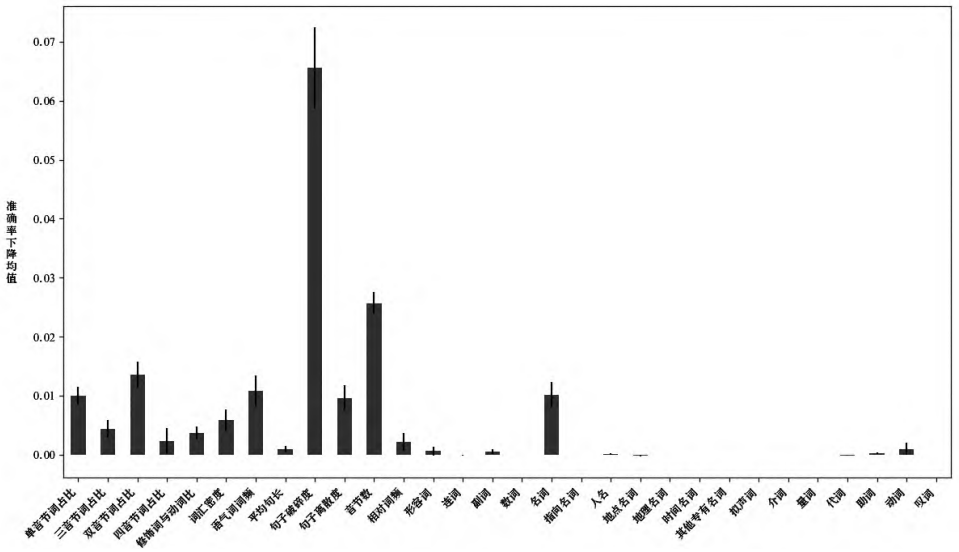


图 2　基于 RF 模型得到的特征重要性

在语料库语境特征中，主要起作用的是句子破碎度，同时语气词词频、句子离散度、单双音节词占比也为语体区分做出一定贡献。这说明在语境里，句中的停顿多少、句子齐整与否会影响语体的正式程度，语气词、单音节词、双音节词的出现也会影响模型对语体的判断。在词本身特征中，主要起效果的是音节数，以及是否为名词。这首先体现了语体语法在韵律角度的影响。音节、音步的组合情况是模型判断词语是否正式的依据之一。

（三）误例分析

使用基于词向量特征训练的 SVM 模型，输出其在测试集上预测错误且置信度 0.7 以上的例子。分析后发现，虽然模型预测结果和《现汉》标注不同，但并不代表模型一定预测错误，而可能是《现汉》标注数据存在一定问题，因此模型实际的预测准确率应高于 87.26%。以下对误例进行分类讨论。

1. 标注遗漏

此类情况较多，如“毋庸”“拨冗”等，《现汉》未标记其语体，但是模型预测为正式体。从这些词语使用的交际场景来看，其常见于偏正式书面语的文本，模型预测较为准确。因此，《现汉》可能遗漏了一些书面语体词的标注。

2. 标注过时

伴随时代发展，《现汉》中标记〈书〉〈口〉的部分词语的语体属性也在发生变化。如“奔命”在《现汉》中标注为口语体词。经查询BCC语料库，它在古汉语子库中出现3082次，在现代汉语多领域语料中仅有 201 条记录，且大多来自报刊和文学作品等，说明现代汉语非正式语境中已很少有人这样表述。

3. 标注不对称

孙德金（2012）指出，《现汉》需要标注与标〈口〉词形成对照关系的书面语词，而不是“书面上的文言词语”。模型在训练中学习到了标注数据不对称的倾向，例如“岚烟”“效颦”“沉吟”在《现汉》中无语体标注，但被模型认为是正式体词汇。这些词语的文言特点突出，使用这些词会让人觉得交际对象很典雅、有文化，而不是正式程度高，属于语体三分理论中的典雅体范畴。

（四）模型误差校正

针对标注体系不对称的问题，本文展开了进一步分析。在 13.2 亿词次的现代汉语正式体语料库里，2433 个带有〈书〉标记的词中有 50% 的词语频次低于 69 次，说明训练数据中的正式体词收录了大量在现代汉语书面语中并不常见的文言成分。那么，基于该数据训练得到的模型便容易存在“文言偏见”，即“一个词语越接近文言表达，正式度越高”。进一步对比发现，对《现汉》数据拟合越好的模型越容易学习到这一偏见。如表 4 所示，对于“纪传体”“等离子态”“祖国”等非文言成分且偏书面语体的现代汉语词汇，分类效果最佳的 SVM 模型倾向于给出较低的正式度，而对“蠲除”“胪陈”“翛然”等现代书面语中少见但文言特点突出的词语，该模型所预测的正式度值很高。

表 4　不同模型的词语正式度预测值

词语 \ 模型	词向量特征（SVM）	语料库语境特征（RF）	词本身特征 + 语料库语境特征（RF）
纪传体 _n	0.0107	0.5830	0.5163
会计师 _n	0.1333	0.2741	0.4097
祖国 _n	0.1345	0.3621	0.4690
蠲除 _v	0.9939	0.5996	0.7168
胪陈 _v	0.9781	0.7053	0.7725
翛然 _a	0.9999	0.6956	0.7747

相较于使用词向量特征训练的 SVM 模型，引入语料库语境特征的模型基于正式、非正式维度对词汇进行刻画，因而对文言特征不敏感，能更好地修正数据中的“文言偏见”。从表 4 可以看出，语料库语境特征参与训练的模型正式度预测值和人类语体属性判断更为相符。我们进一步分析了 RF 模型（词本身特征 + 语料库语境特征）的分类错误情况，如表 5 所示。该模型的分类五折准确率为 71.89%，虽然并不突出，但具体分析测试集中的误例，发现它对口语词和无语体标注词的预测更准确，大部分错误来自对标〈书〉词语的预测，而这些被错误预测的词语多为在现代汉语书面语中使用较少的文言成分（如“弁言”“畏葸”），以及在非正式语料库中更常见的词（如“鸡肋”“慵懒”）。此外，《现汉》的无语体标注数据中有 35.28% 的词在正式体语料库中的频次达到该词在口语体语料库频次的 2 倍以上，这意味着这些词更容易被使用于正式语体。该模型将 11.56% 的无语体标注词预测为正式体词，也在一定程度上挖掘出了《现汉》未标注的偏正式体色彩的词汇（如“岚烟”“群英”）。因此，可以引入 RF 模型（词本身特征 + 语料库语境特征）来矫正训练数据中的“文言偏见”。

表 5　RF 模型（词本身特征 + 语料库语境特征）在测试集上的错误分布

	标〈书〉词	无语体标注词	标〈口〉词
测试集分布	489	476	120
模型错误分布	112	54	5
模型错误率	22.90%	11.56%	4.16%

四　词表语体属性预测

根据实验结果分析，词汇语体属性探测模型能够为词典中的语体属性标注提供一定参考。在汉语教学领域的词表中，大部分词语均未被《现汉》标记语体属性，如果能通过

模型对其语体正式程度进行量化判定，将有助于教材编写者和教师将语体信息融入词汇教学，加强学生的语体意识。本文应用模型对《国际中文教育中文水平等级标准》词表（以下简称“《标准》词表”）、《义务教育常用词表（草案）》主表（以下简称“《常用词表》”）进行语体正式度预测，以服务该领域的教学和研究。

（一）《国际中文教育中文水平等级标准》词表语体属性预测

《标准》词表收录 11092 个词条。因词表未提供词性标注，本文使用《现汉》对词表中的词条进行词性标注。经拆分词组、排除词缀和短语、词性标注后，进一步检查词条是否被语料库和词向量词表收录，最终得到 10028 词。

本文使用两种模型进行加权组合预测，模型一为基于词向量特征的 SVM 模型，该模型对《现汉》的拟合效果最佳，模型二为使用词本身特征 + 语料库语境特征训练的 RF 模型，可以修正数据中存在的“文言偏见”。经尝试不同权重设定，发现 RF 模型加权分数为 60% 时，得到的正式度分值对不同词汇有较好的语体区分效果，且与人的语感较为相符。

表 6 展示了单独使用两种模型和加权组合的结果，可以看出，SVM 模型因拟合了《现汉》数据的“文言偏见”，对该词表的正式度预测均值过低（0.0945），RF 模型使用词本身特征 + 语料库语境特征预测，正式度均值较高（0.5333）。从加权后的分数来看，《标准》词表正式度均值偏低（0.3577），超过 75% 的词语正式度都低于 0.5。不同正式度区间的词语示例可参见表 7。

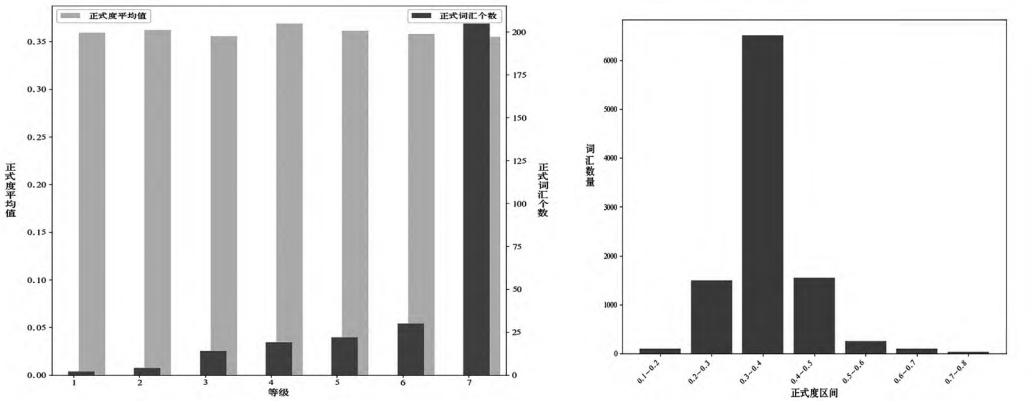
表 6　模型在《标准》词表上的预测结果

模型	非正式词汇个数	正式词汇个数	正式度平均值	正式度标准差	25% 分位点	50% 分位点	75% 分位点	最大值	最小值
SVM	9732	296	0.0945	0.1328	0.0219	0.0492	0.1043	0.9999	0.0007
RF	2421	7610	0.5333	0.0787	0.5062	0.5386	0.5894	0.7288	0.2494
40%SVM+60%RF	9732	296	0.3577	0.0721	0.3192	0.3523	0.3878	0.7968	0.1608

表 7　《标准》词表正式度预测示例（基于 SVM+RF 加权模型）

正式度区间	示例
正式度较高（0.6~0.8）	晚年 _nt: 0.6110；携手 _v: 0.6561；屡 _d: 0.7405 诸位 _r: 0.6078；圣贤 _n: 0.6630；勿 _d: 0.6576
正式度一般（0.4~0.6）	西北 _nl: 0.5086；诚实 _a: 0.4169；考验 _v: 0.4070 心酸 _a: 0.4471；低下 _a: 0.4138；生效 _v: 0.4071
正式度较低（0.2~0.4）	免不了 _v: 0.2321；笑话 _n: 0.3309；糊涂 _a: 0.3180 吃 _v: 0.3593；烂 _a: 0.3496；笼统 _a: 0.3217
正式度低（0~0.2）	两口子 _n: 0.1646；一会儿 _d: 0.1947；半边天 _n: 0.1987 大腕儿 _n: 0.1860；小伙子 _n: 0.1972；马后炮 _n: 0.1862

进一步统计各等级正式体词数和平均正式度信息，结果如图 3a 所示。随着词汇等级的逐步上升，正式体词语数量呈增加趋势，且绝大部分正式体词集中于高级水平（七至九级），这意味着随着水平提升，学生能接触到更多的正式语体表达。然而，各等级的平均正式度差异较小，从图 3b 也可以观察到，《标准》词表中大部分词语的正式度偏低且集中于 0.3~0.4 区间内。这表明汉语学习者在学习过程中主要接触正式度低的词汇，即使到高级学习阶段接触到的书面正式体词汇也较少。词表的语体不均衡现象可能导致学习者对词汇的语体区别意识较弱。在教学中，需要关注学生对不同语体的理解和运用，让其能准确、恰当地运用词语，避免出现语体偏误。因此，可考虑在补充词表或教学中适当增加不同语体的词汇，以培养学生的语体意识。



（a）各等级词语统计结果　　（b）词语正式度整体分布情况

图 3　《标准》词表各等级统计结果及正式度整体分布（基于 SMV+RF 加权模型）

（二）《义务教育常用词表（草案）》主表语体属性预测

《常用词表》是面向中小学语文教学的词表，分为四级，对应义务教育中的四个学段，主要收录通用性高、有一定书面语色彩的词语，其主表原有 15114 词。本文将多义词词条按词性进行拆分，并筛除语料库、词向量词表未收录的词语，得到 15472 词。同样使用预测《标准》词表的两个模型，设置 RF 模型的加权比例为 60%，以修正数据中存在的“文言偏见”。

如表 8 所示，从组合模型的不同分位点来看，《常用词表》中的词汇正式度中等偏低，总体符合义务教育阶段的学习需求。不同正式度区间的词语示例如表 9 所示。

表 8　模型在《常用词表》上的预测结果

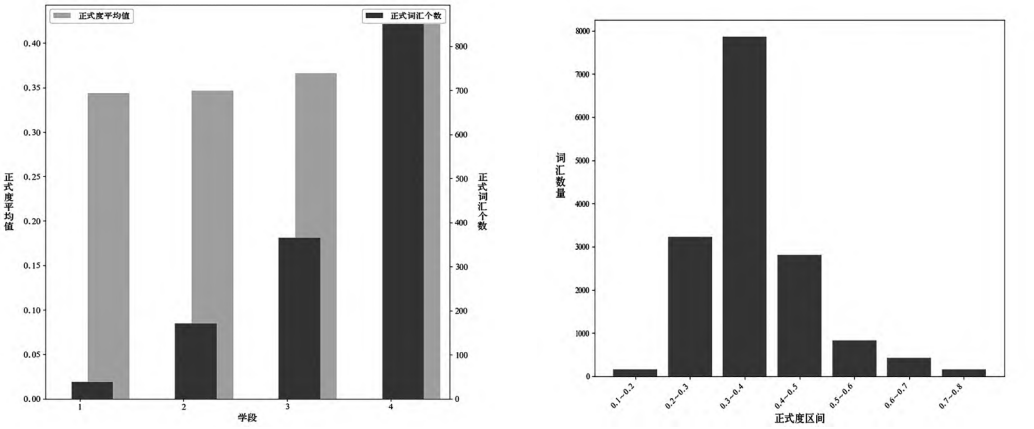
模型	非正式词汇个数	正式词汇个数	正式度平均值	正式度标准差	25%分位点	50%分位点	75%分位点	最大值	最小值
SVM	14047	1425	0.1560	0.2022	0.0291	0.0715	0.1902	0.9999	0.0009
RF	6303	9169	0.5114	0.0963	0.4303	0.5217	0.5944	0.7086	0.2705
40%SVM+60%RF	14047	1425	0.3692	0.0973	0.3078	0.3604	0.4052	0.7888	0.1647

表 9　《常用词表》正式度预测示例（基于 SVM+RF 加权模型）

正式度区间	示例
正式度较高（0.6~0.8）	何尝 _d: 0.6612；国事 _n: 0.6208；且 _d: 0.6790 可鄙 _a: 0.6096；窈窕 _a: 0.7403；愠色 _n: 0.6684
正式度适中（0.4~0.6）	蓬勃 _a: 0.4469；感触 _n: 0.4371；伶俐 _a: 0.4889 齐整 _a: 0.4346；艰巨 _a: 0.4106；脱落 _v: 0.4014
正式度较低（0.2~0.4）	实惠 _a: 0.3759；排练 _v: 0.2711；马路 _n: 0.2261 勾结 _v: 0.3945；摊 _v: 0.3779；顺畅 _a: 0.3823
正式度低（0~0.2）	挑刺儿 _v: 0.1842；有点儿 _d: 0.1822；老天爷 _n: 0.1837 吃里爬外 _i: 0.1905；叽里咕噜 _o: 0.1802；纳闷儿 _v: 0.1884

通过对比表 6 与表 8 中的组合模型结果，可以看出，两表词语正式度均值接近，但面向义务教育的《常用词表》所收录的正式体词数量明显更多。经统计各等级词语信息，可以进一步看出两表收词差异。如图 4a 所示，《常用词表》各学段均有一定比例的正式体词，且正式体词语数量和平均正式度值也随学段上升而增加，这与《标准》词表在高等（七至九级）集中收录正式体词的情况存在差别。对比图 3b 和图 4b，两表正式度在 0.3~0.4 区间的词语均占据了最大比重，但《常用词表》收录的高正式度词语明显更多。《常用词表》

中提到，进入小学后，儿童的学习内容已由口语扩展到书面语。在这一语体转换的关键时期，词表收录遵循词汇习得规律中的“晓”层面，即准确知晓词汇意义，并理解该词使用的特定语境。引入语体标注数据后，《常用词表》可为语文词汇教学、教材和教辅编纂提供更为高效的语体维度支持。值得一提的是，本文开源了两个词表的语体属性测量结果，使用者可根据具体需求提取不同模型或权重的词语正式度值进行应用①。



（a）各学段词语统计结果　　（b）词语正式度整体分布情况

图 4　《常用词表》各学段统计结果及正式度整体分布（基于 SMV+RF 加权模型）

五　结语

本文尝试从词典编纂、汉语教学的需求出发，采用机器学习方法训练词汇语体属性分类模型，以促进解决语体属性标注的模糊性难题。在理论层面上，挖掘语体分类依据，设计多维度语言特征，首次实现了词汇语体正式度的连续统式标注；在应用层面上，研究成果发现了当前《现汉》语体标注的部分问题，可为其未来修订提供参考，还利用基于词汇特征、语料库语境特征的模型在一定程度上修正了训练数据中的“文言偏见”，并开放了汉语教学领域的语体词表资源，探讨语体知识引入语言教学中的应用价值。

本次研究尚存在一些待改进之处，如训练数据须进一步优化，应在《现汉》标记的基础上补充校对语体信息，添加成语类标注数据，并引入词义消歧方法，以实现更加细粒度的语体属性测量。此外，研究采用《现汉》中书面语、口语二分的语体标注数据，而事实上，标注的书面语体词汇中混杂典雅体词汇。李宇明（2023）在《中国语言生活状况报告》序中指出，在网络语言、领域语言迅速发展，传统文化也被愈发重视的当下，普通话形成了口语体、一般书面语体和典雅语体三足鼎立的格局。未来的词典编纂或语体研究有必要开启语体三分的研究路线。本研究仅从正式—非正式维度对词汇进行语体属性测量，未来还需尝试对典雅—通俗维度进行探索，挖掘更多语体相关的语言特征，探寻更完善、准确的语体标注体系。

［附　注］

① 本研究开源了《国际中文教育中文水平等级标准》词表和《义务教育常用词表（草案）》主表的语体属性预测结果，地址：https://github.com/QSJDAMTX/chinese-vocabulary-of-register。

［参考文献］

［1］崔希亮 . 正式语体和非正式语体的分野［J］. 汉语学报，2020，（2）.

[2] 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 [J]. 语言文字应用，2022，（3）.
[3] 冯胜利，王　洁，黄　梅 . 汉语书面语体庄雅度的自动测量 [J]. 语言科学，2008，（2）.
[4] 冯胜利 . 论语体的机制及其语法属性 [J]. 中国语文，2010，（5）.
[5] 冯胜利 . 语体语法："形式—功能对应律"的语言探索 [J]. 当代修辞学，2012，（6）.
[6] 冯胜利 . 语体语法的逻辑体系及语体特征的鉴定 [J]. 汉语应用语言学研究，2015，（0）.
[7] 冯胜利，王永娜 . 语体标注对语体语法和叙事、论说体的考察与发现 [J]. 汉语应用语言学研究，2017，（0）.
[8] 冯胜利，施春宏 . 论语体语法的基本原理、单位层级和语体系统 [J]. 世界汉语教学，2018，（3）.
[9] 冯胜利，王永娜，王丽娟 . 汉语八百对单双音节对应词词典 [Z]. 芝加哥：Phoenix Tree Publishing Inc，2020a.
[10] 冯胜利，刘丽媛 . 语体语法的生物原理与生成机制 [J]. 民俗典籍文字研究，2020b，（2）.
[11] 符淮青 . 现代汉语词汇 [M]. 北京：商务印书馆，2011.
[12] 胡　楠，张文强，胡韧奋 . 基于跨语言对齐词向量的汉日词汇意义比较研究 [J]. 语言文字应用，2022，（2）.
[13] 胡裕树 . 现代汉语（重订本）[M]. 上海：上海教育出版社，2019.
[14] 黄国敬，周立炜，饶高琦，臧娇娇 . 基于《同义词词林》的中文语体分类资源构建 [A]. 第二十一届中国计算语言学大会论文集 [C]，2022.
[15] 黄劲怡，彭宣维 . 正式性的测量方法和描写路径 [J]. 外语研究，2022，（2）.
[16] 李　果，王长林 . 论古白话正式体的体原子——以《祖堂集》《景德传灯录》"弘忍、惠能"篇为例 [J]. 历史语言学研究，2021，（2）.
[17] 李　航 . 统计学习方法（第 2 版）[M]. 北京：清华大学出版社，2019.
[18] 李　泉 . 基于语体的对外汉语教学语法体系构建 [J]. 汉语学习，2003，（3）.
[19] 李文明 . 语体是言语的风格类型——兼与刘大为先生商榷 [J]. 修辞学习，1994，（6）.
[20] 李行健 . 现代汉语口语词典 [Z]. 北京：华语教学出版社，2022.
[21] 李宇明，郭　熙 . 中国语言生活状况报告（2023）[M]. 北京：商务印书馆，2023.
[22] 梁诗尘，唐雪梅，胡韧奋等 . 基于分布式表示的汉字部件表义能力测量与应用 [J]. 中文信息学报，2021，（5）.
[23] 刘大为 . 语体是言语行为的类型 [J]. 修辞学习，1994，（3）.
[24] 马明艳 . 汉语学习者书面语作文"口语化"倾向的语体表征 [J]. 汉语学习，2017，（1）.
[25] 潘先军 . 话语标记的语体特征与对外汉语话语标记教学 [J]. 对外汉语研究，2021，（1）.
[26] 邵敬敏 . 现代汉语通论（第 3 版）[M]. 上海：上海教育出版社，2016.
[27] 施光亨 . 汉语口语词词典 [Z]. 北京：商务印书馆，2012.
[28] 苏新春 . 义务教育常用词表（草案）[M]. 北京：商务印书馆，2019.
[29] 孙德金 . 略论《现代汉语词典》第 5 版标〈书〉词及其改进建议 [J]. 辞书研究，2012，（4）.
[30] 孙茂松，李景阳，郭志芃，赵　宇，郑亚斌，司宪策，刘知远 . THUCTC：一个高效的中文文本分类工具包 [DB/OL]. 2023-10-20.http：//thuctc.thunlp.org/.
[31] 郜沁清，饶高琦 . 汉语语体特征的计量与分类研究 [A]. 第二十届中国计算语言学大会论文集 [C]. 2021.
[32] 王德春 . 语体略论 [M]. 福州：福建教育出版社，1987.
[33] 王丽娟 . 汉语旁格述宾结构的语体鉴定及其语法机制 [J]. 语言教学与研究，2018，（6）.

［34］王用源，陈宇豪 . 语体语法视角下语体正式度的自动测量［J］. 语言教学与研究，2023，（4）.
［35］汪维辉 . 现代汉语“语体词汇”刍论［J］. 长江学术，2014，（1）.
［36］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）.
［37］荀恩东，饶高琦，肖晓悦等 . 大数据背景下 BCC 语料库的研制［J］. 语料库语言学，2016，（1）.
［38］袁　晖 . 论语体词［J］. 修辞学习，2004，（3）.
［39］张翠吉 . 汉语中介语词汇、短语语体的偏误分析——基于 HSK 动态作文语料库［J］. 现代语文，2018，（9）.
［40］张文贤，邱立坤，宋作艳等 . 基于语料库的汉语同义词语体差异定量分析［J］. 汉语学习，2012，（3）.
［41］中国社会科学院语言研究所词典编辑室 . 现代汉语词典（第 7 版）［Z］. 北京：商务印书馆，2016.
［42］中华人民共和国教育部 . 国际中文教育中文水平等级标准（GF0025-2021）［S/OL］. 2023-12-29. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/t20210329_523304.html.
［43］Attali，Y & Burstein，J. Automated essay scoring with e-rater® V. 2［J］. *The Journal of Technology，Learning and Assessment*，2006，（3）.
［44］Breiman，L. Random forests［J］. *Machine learning*，2001.
［45］Che，W.，Feng，Y.，Qin，L.，et al. N-LTP：An open-source neural language technology platform for Chinese［A］. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing：System Demonstrations*［C］，2021.
［46］Codemayq. Chinese Chatbot Corpus［DB/OL］. 2023-12-29. https://github.com/codemayq/chinese_chatbot_corpus.
［47］Harris，Z.S. Distributional structure［J］. *Word*，1954，（2~3）.
［48］Li，S.，Zhao，Z.，Hu，R.，et al. Analogical reasoning on Chinese morphological and semantic relations［A］. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2：Short Papers)*［C］，2018.
［49］Li，Y.，Zhang，Y.，Zhao，Z.，et al. CSL：A large-scale Chinese scientific literature dataset［A］. *Proceedings of the 29th International Conference on Computational Linguistics*［C］，2022.
［50］Mikolov，T.，Sutskever，I.，Chen，K.，et al. Distributed representations of words and phrases and their compositionality［J］. *Advances in neural information processing systems*，2013.
［51］Ure，J. Lexical density and register differentiation［J］. *Applications of linguistics*，1971.
［52］Wang，Y. & Hu，R. A prompt-independent and interpretable automated essay scoring method for Chinese second language writing［A］. *China National Conference on Chinese Computational Linguistics*［C］. Cham：Springer International Publishing，2021.
［53］Wang，Y.，Ke，P.，Zheng，Y.，et al. A large-scale Chinese short-text conversation dataset［A］. *Natural Language Processing and Chinese Computing: 9th CCF International Conference*［C］. Springer International Publishing，2020.
［54］Wang，X.，Li，C.，Zhao，J.，et al. Naturalconv: A Chinese dialogue dataset towards multi-turn topic-driven conversation［A］. *Proceedings of the AAAI Conference on Artificial Intelligence*［C］，2021，（16）.
［55］Xu，B. NLP Chinese Corpus：Large Scale Chinese Corpus for NLP［DB/OL］. 2023-8-20. https://github.com/brightmart/nlp_chinese_corpus.

（责任编辑　刘　琪）