基于预训练语言模型的汉语典雅度研究

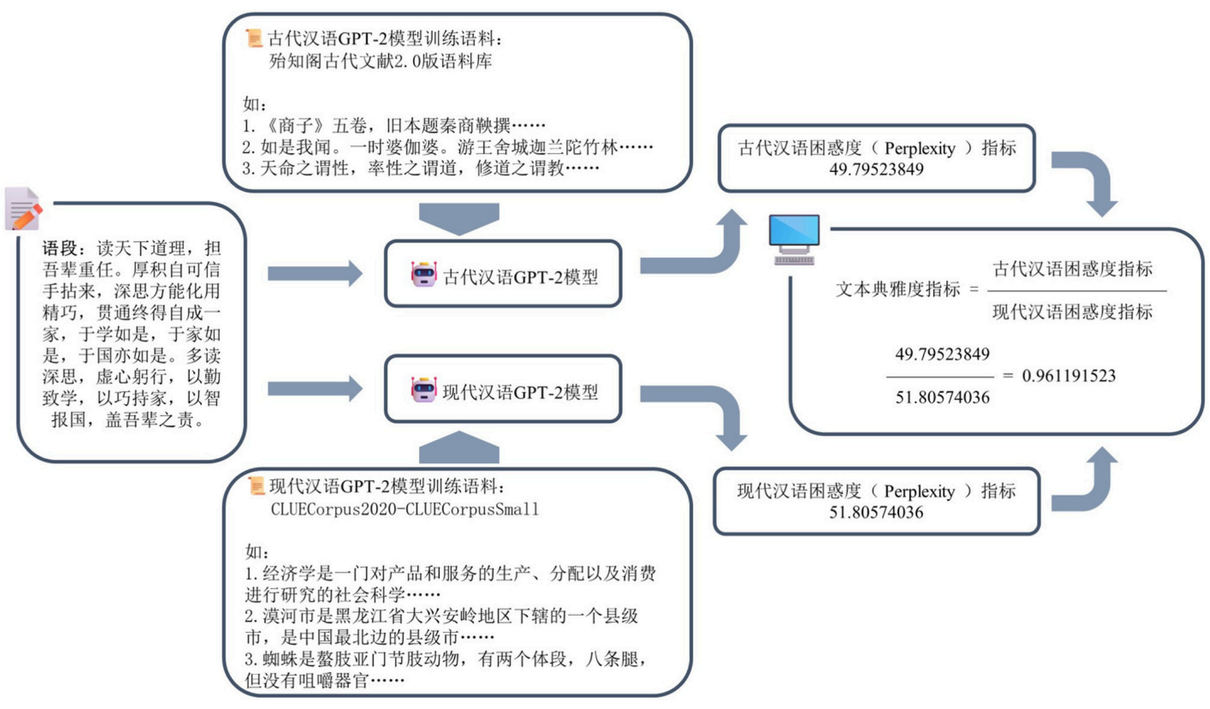
杨智惠 北京师范大学 202321091019@mail.bnu.edu.cn
莫凯洁 北京师范大学 mokaijie@mail.bnu.edu.cn
胡韧奋* 北京师范大学 irishu@bnu.edu.cn

一、背景

现代汉语中形成了口语体、一般书面语体和典雅语体三足鼎立的格局（李宇明，2023），借助量化分析方法探究这些语体的共时表征和历时演变，有助于更好地揭示汉语语言体系的发展规律。然而，现有的语体量化研究多集中于口语体和书面语体，对典雅体的关注不足。

自然语言处理目前进入了预训练模型范式时代（冯志伟等，2023），基于transformers的预训练语言模型可以自行建模词语及语段蕴含的语言信息，而无需人工预先标注词汇、语法特征。相比于传统机器学习方法，预训练语言模型用于判断语段的典雅程度时无需预先寻找其中的语言特征，例如古汉语常用词、文言句式等。

二、方法



为实现汉语文本的语体典雅度自动分析，本文提出了一种基于生成式预训练语言模型GPT的语体典雅度测量指标，并对1872年至1949年期间报刊语料展开分析，探究白话文运动前后的语体典雅度特征演变情况。

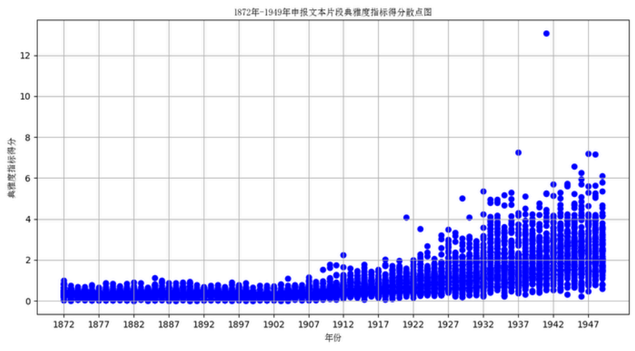
本文首先采用古代汉语GPT-2模型和现代汉语GPT-2模型对语段的成立概率进行估计，分别得到古代汉语和现代汉语困惑度（Perplexity）指标。困惑度反映了模型的“语感”，如模型判定该语段成立概率大，则困惑度指标低。进一步地，本文以古代汉语和现代汉语困惑度比值作为衡量文本典雅度的指标，指标数值越低，表示文本越接近古汉语文言表达，而越远离于现代汉语白话文表达。

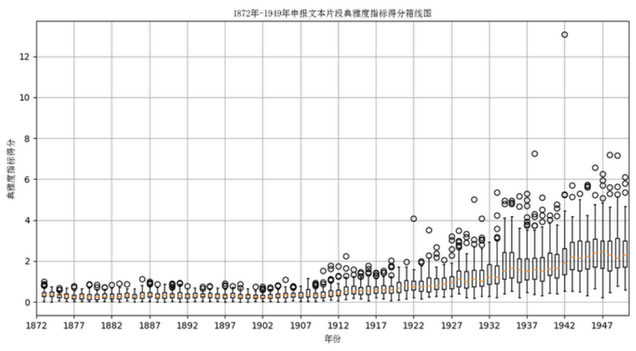
三、实验

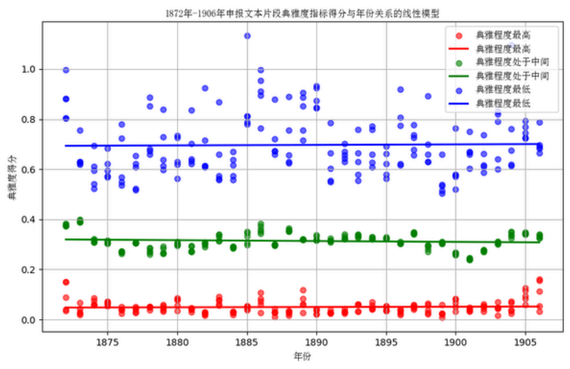
为验证指标的有效性，本文采集了文学、新闻、相声等不同类型文本，构建了语段级语料库，并从中随机选取300对、共600条典雅程度不同的文本片段作为测试集。如果典雅度指标测量结果与文本标签相符，则认为测量正确。在测试集上，该指标取得了95.67%的准确率，证明其对文本典雅程度的判断能较好地反映人类语感。

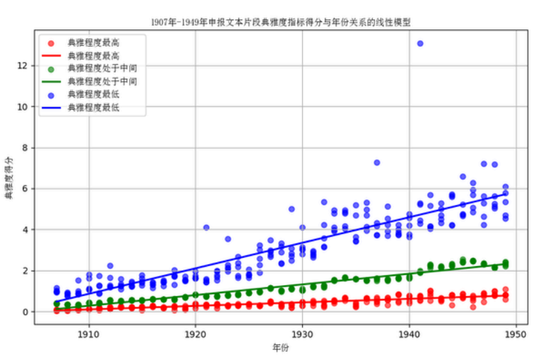
四、应用

本文将该测量方法应用于历史语体演变研究。我们采集了1872年至1949年期间的《申报》语料，共计116万字符，对文本的典雅度特征展开历时分析。测量结果显示，1919年白话文运动后语言的典雅度出现大幅下降。进一步分析发现：（1）白话文运动前，生活、商业领域的书面表达已呈现出接近现代汉语的语言特点。（2）白话文运动后，文学书面表达呈现出激进的俗白化语体转向，国际新闻、科普、商业主题的书面表达逐渐向俗白转变，而国内政论文章仍含有较多的文言成分，典雅度维持在较高水平。









五、总结

本文提出的基于生成式预训练语言模型GPT的语体典雅度测量指标能较好地反映人类语感，实现汉语语体典雅度的自动量化。本文在1872年至1949年期间的《申报》语料上应用语体典雅度测量指标，展现了白话文运动前后汉语语言系统内部语体演变的复杂过程，并为典雅语体研究提供了重要的数据和方法支撑。