人工智能生成文本语言风格形成因素探究

王　玥[1]　孙瑞琼[2]
1.黑龙江外国语学院　黑龙江哈尔滨　150020
2.黑龙江大学　黑龙江哈尔滨　150000

摘　要：人工智能生成文本的语言风格是当前自然语言处理领域的重要研究议题。本文从语言风格的基本内涵出发，系统分析了人工智能文本生成的技术原理与风格特征，重点探讨了训练数据、模型架构、人类反馈三大核心要素对AI文本语言风格的塑造作用。研究发现，训练数据的来源与质量直接决定风格基调，模型架构与算法设计影响风格表达能力，人类反馈与提示工程则起到精准调控作用。在此基础上，本文提出了数据筛选与平衡、模型参数调优、人机协同引导等优化策略，为提升AI生成文本的语言风格质量提供参考。

关键词：人工智能；文本生成；语言风格；形成因素；优化策略

引言

人工智能文本生成技术发展迅速，正在改变人类语言生产方式，从早期规则模板生成发展到如今深度学习大语言模型，AI生成文本质量显著提高。但语言风格作为文本的灵魂，其形成机制远比表面上的语法正确、语义连贯复杂得多，语言风格包含词汇选择、句式结构、修辞手法、情感基调等各个方面，是训练数据、模型架构、人类干预等多种因素共同作用的结果，对AI生成文本语言风格形成因素进行深入探究，有利于改良模型表现，提升生成质量，扩大应用范围，具有重要的理论和实践意义。本文主要对影响AI文本语言风格的因素进行梳理，并提出相应的优化策略。

一、人工智能生成文本语言风格的内涵与特征

（一）语言风格的定义与表现维度

语言风格是语言使用中形成的个性特点，是说话人或作者在一定的语境中对语言要素的特殊选择和组合方式。从语言学的角度来说，风格包括词汇层面的用词偏爱，句法层面的结构喜好，修辞层面的表达手法，语篇层面的组织准则，AI文本生成语境中语言风格不只是表层的语言形式，它更深层次的表现在语义倾向以及价值取向上，从正式程度、情感色彩、简繁程度、客观性强弱等角度描述AI生成文本的风格，正式风格多用书面语、完整句式、抽象词汇；非正式风格则偏爱口语化表达，情感色彩的强弱体现在褒贬倾向和情绪浓度的差别，简繁倾向反映在句子长度，信息密度，修饰成分的多寡上，客观性关乎事实陈述和主观评价的比重，各方面交错，一起形成了AI产出文本丰富多彩的语言风格谱系。

（二）AI文本生成的技术机制与原理

人工智能文本生成的主要技术基础是深度学习神经网络，主要是以变换器架构的大语言模型为主。本类模型用自注意力机制来捕获文本中词语之间的长距离依赖关系，从而完成语言模式的深度学习。预训练阶段，模型在大量的文本语料上进行无监督学习，通过预测下一个词或者填补缺失的词等任务，自动提取语言的统计规律和语义关联。在微调阶段，模型针对某个任务或者领域做有监督训练，进一步提升生成效果。生成的过程本质上就是概率采样的过程，模型根据已经存在的上文计算出下一个词的条件概率分布，然后通过采样的方式选择输出词。温度参数、top-k采样、nucleus采样等技术手段可以调节生成文本的随机性和多样性，进而影响语言风格的呈现。基于统计学习、概率生成的机制使AI生成文本既能够保持语言的流畅连贯，又能在一定程度上模拟人类写作的风格多样性[1]。

基金项目：黑龙江省语委项目：人工智能生成文本语言风格及其少年语言的影响研究（项目编号：G2025Y025）

作者简介：
1.王玥（1980.07—），女，汉族，黑龙江外国语学院，教授，研究方向：汉语言文字学。
2.孙瑞琼（2001.03—），女，汉族，黑龙江大学，研究生，研究方向：国际中文教育。

（三）AI生成文本风格的独特性分析

相较于人类创作的文本而言，AI生成文本的语言风格具有与众不同的特征。一是统计规律性，AI模型本质上是对训练数据中的语言模式进行统计学习和概率重组，生成的文本多是高频词汇、常见句式的组合选择，风格趋于大众化、规范化。其次为风格的可塑性，通过对训练数据配比、模型参数或者提示词进行调整，可以较为灵活地改变生成文本的风格倾向，这种可塑性是人类作者所不具备的。最后是情感表达的表面化，虽然AI可以生成带有丰富情感色彩的文本，但是这些情感往往缺乏真实的情感体验作为支撑，更多的是模仿情感表达的模式。不过，AI生成的文本在创造性和个性化上存在不足，容易造成表达模式化、内容雷同化。独特性既体现出人工智能技术的优势以及潜在的发展空间，也表现出人工智能在风格生成上所遇到的障碍与困难[2]。

二、人工智能生成文本语言风格的形成因素

（一）训练数据的来源、规模与质量

训练数据是AI语言模型的基础，直接影响生成文本的风格基调和边界。数据来源的多样性影响风格的丰富程度，若训练语料集中于学术论文，模型将倾向于正式严谨的学术风格；若语料来自社交媒体，则会呈现口语化与情绪化特征。数据量影响模型学习语言模式的程度，语料量大，可以涵盖更多的风格类型，使模型有较强的风格迁移能力，但是会造成风格趋同于主流表达。数据质量涉及语料的准确性、规范性和代表性，低质量的数据中存在的语法错误、表达不当或者偏见倾向会被模型吸收并复制到生成的文本中，造成风格缺陷。另外，数据的时代性、地域性也会在模型风格上留下印记，不同时期、不同地域的语言使用习惯差异会在生成文本的用词、句式、表达习惯上体现出来。因此训练数据的精心挑选、合理配比是形成理想的语言风格的前提条件，在此过程中要兼顾数据多样性以及质量控制[3]。

（二）模型架构与算法设计的影响

模型架构和算法设计是AI文本风格形成的内部机制。Transformer架构的自注意力机制使模型具备了捕捉长距离依赖的能力，使生成文本在篇章层面上具有风格的一致性，而不仅仅是在句子层面上的局部连贯。模型层数以及参数的多少决定了其表达能力的上限，大模型可以学习更复杂的语言模式以及细微的风格差别，但也会出现过拟合和计算成本较高的问题。注意力机制的设计变体，例如局部注意力、稀疏注意力等等，会改变模型对于各种语言特征的敏感度，从而影响到生成风格的精细程度。解码策略的选择同样重要，贪婪解码倾向于生成保守、高概率的文本，束搜索或者采样的解码方式可以增加输出的多样性和创造性，但是也可能造成不连贯或者不合理的表达。位置编码方式、归一化方法、激活函数等技术细节会在微观层面影响模型的语言生成偏好。因此，模型架构与算法的精心设计是达到某种风格目标的技术保障，在模型复杂度、计算效率和生成质量之间要进行权衡优化[4]。

（三）人类反馈与提示工程的调控作用

人类反馈和提示工程属于引导AI文本风格的方法，可以克服纯数据驱动方法的缺陷。人类反馈学习通过收集人类评价者对于模型输出的偏好判断，使模型学习符合人类价值观和审美标准的表达方式，可以纠正模型在风格把握上的失误，使生成文本更得体、自然，更符合场景需求。提示工程就是通过精心设计的输入指令，在不改变模型参数的前提下，引导模型生成某种风格的文本。有风格指令、合适的例子、结构化的输入格式都会对输出风格产生影响，这种轻量的干预方式最大的优点就是灵活且成本较低。另外，思维链提示、角色扮演提示等高级技巧可以激活模型潜在的能力，从而实现风格控制更加精细与复杂的目的。但是人类反馈的主观性以及提示工程的技巧性也带来了挑战，不同的评价者审美会有差异，提示词的细微变化都会影响风格输出。因此，建立科学的反馈机制和系统化的提示方法论，是充分发挥人类智慧引导作用的关键所在[5]。

三、人工智能生成文本语言风格的优化策略

（一）数据筛选与平衡策略

优化AI文本语言风格的第一种方法就是实行科学的数据筛选与平衡。数据筛选要搭建起多维度的质量评定体系，从语法规范性，内容准确性，表达得体性这些角度去考察候选语料，剔除掉低质量的，带有偏见或者不当内容的数据，从而保证训练语料有较好的质量。筛选之后还要考虑数据风格是否均衡，避免出现某种风格过于突出造成模型风格偏向。按照正式和非正式、客观和主观、简洁和详尽等标准对语料进行标记分类，按应用需求确定各种风格的合理比例，用重采样或加权训练的方法来实现风格的平衡学习。对某个领域的应用要增大目标领域高质量语料的比例，并保留一部分通用语料来保持模型的泛化能力。数据清洗过程中还要保护语言多样性，不能因为过度规范化抹杀了方言、俚语等有文化

特色的表达方式。动态的数据更新也很重要，定时给模型加新的语料，让它学会时代在变，语言也在变。通过筛选数据、平衡数据可以源头上给AI模型提供优质的、多样化的素材，为生成高质量、风格合适的文本打下良好的基础。

（二）模型参数调优与风格控制技术

模型层的优化主要是参数的调优、风格控制技术的应用。参数调优是调整模型超参数，温度参数控制生成的随机性，温度低时生成文本保守、风格一致，温度高时生成文本更具创造性但连贯性降低，top-k和top-p采样参数限定候选词的范围，在多样性和合理性之间取得平衡。按照风格要求可以采用风格迁移微调的方法，在预训练模型的基础上用特定风格的语料做有监督微调，使模型学会目标风格的表达方式。条件生成机制属于更灵活的风格控制方式，把风格标签或者控制向量嵌入到输入当中，让模型按照指定的风格生成文本，这样就可以使单个模型支持多种风格的切换。近几年出现的可控生成技术，如PPLM、CTRL等，通过增加判别器或者控制码的方式来对生成过程进行实时控制，从而实现风格调节。模型架构上使用风格编码器、风格感知注意力等可以更好地对风格特征进行显式的建模。集成学习方法可以用不同的子模型去专门训练不同的风格，通过模型的融合得到更加丰富、精确的风格表达。这些技术手段的综合运用，能够显著提升AI模型的风格控制精度与生成质量。

（三）人机协同的风格引导机制

人机协同的风格引导机制是提高AI文本风格质量的有效途径。这种机制体现的是人类智慧与机器能力的融合共生，扬长避短。在提示工程上搭建系统的提示词库和最佳实践指南，包含各种风格需求的提示模板，示例和技术手段，降低使用的门槛，提升引导效果的稳定性。交互式生成模式给用户在生成过程中提供即时反馈和调整的机会，模型会根据用户的即时评价来修改输出，这种迭代式的合作可以快速地达到用户所希望的风格目标。建立专业的评价团队，从风格一致性、表达得体性、情感适配度等维度对模型生成的文本进行评价，评价结果作为强化学习训练的依据，使模型不断学习人类的风格偏好。风格诊断和解释工具的开发同样重要，用可视化技术来展示模型在风格维度上表现，让使用者了解模型的行为从而提出有针对性的改进意见。从应用角度而言，可以设计出风格定制化服务，用户上传样本文本或者设置风格参数，系统就会为用户定制出专属的风格模型。另外要建立伦理审查机制，保证风格引导不会导致有害内容的产生或者加重社会偏见。构建起完整的一个人机协同生态系统之后，就可以对AI文本风格进行精准控制并不断优化，从而让技术更好地服务于不同类型的“人”的表达需求。

结束语

人工智能生成文本的语言风格形成是训练数据、模型结构、人类干预等各方面的因素相互作用的结果。探究出这些因素起作用的机制对提高AI文本生成质量、拓宽其应用场景有重大意义。本文系统梳理了AI文本风格的内涵特征和形成机制，提出了数据优化、模型调控、人机协同等优化策略。随着技术不断发展，AI文本生成的风格多样性、个性化表达、文化适应性等都会有新的发展。但是，我们也要防范技术发展所带来的风险，保证AI生成文本的风格魅力不被消磨，遵守伦理，服务于人。经过数据优化、技术创新、人机协同这三个方面入手之后，AI文本生成的风格控制水平将会得到提高。随着技术不断发展，人工智能对人类语言丰富性、多样性的理解与表现会越来越好，从而实现个性化、场景化的智能表达。

参考文献

[1]薛德军，孔祥煜，耿崇，等.人工智能生成中文学术论文文本检测研究[J].北京电子科技学院学报，2024，32（3）：104-112.

[2]彭展，杨红阳.人工智能生成文本对传统文学创作范式的挑战与启示[J].美化生活，2025，（22）：0118-0120.

[3]王雨欣，刘柯飞，李雪莲，等.一种识别和检测人工智能生成文本的算法[J].电讯技术，2025，65（3）：378-384.

[4]沈锡宾，王立磊.人工智能生成学术期刊文本的检测研究[J].科技与出版，2023，（8）：56-62.