Modern Linguistics 现代语言学, 2025, 13(4), 575-581
Published Online April 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.134385

Hans汉斯

人工智能中国特色话语汉英翻译质量研究

凌　颖

上海海事大学外国语学院，上海

收稿日期：2025年3月2日；录用日期：2025年4月14日；发布日期：2025年4月27日

摘　要

随着人工智能技术的快速发展，人工智能在中国特色话语中的应用逐渐成为研究热点。本文对比分析了GPT-4o与文心一言3.5在中国特色话语汉英翻译中的表现差异。本文使用了BLEU值、Flesch-Kincaid Grade Level和Gunning Fog Index等工具，以官方译文为标准进行了译文准确度与复杂度对比测试。初步分析表明，文心一言3.5在准确度与复杂度方面比GPT-4o更接近于官方译文。研究还发现，当前人工智能技术尚无法完全替代人工译者在该领域的作用。本文对于人机协作进行中国特色话语翻译和未来人工智能翻译工具优化方向提出了建议。

关键词

ChatGPT，文心一言，中国特色话语，译文质量评估，翻译学

Research on Chinese-English Translation Quality of Discourse with Chinese Characteristics by Artificial Intelligence

Ying Ling

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: Mar. 2nd, 2025; accepted: Apr. 14th, 2025; published: Apr. 27th, 2025

Abstract

With the rapid development of artificial intelligence technology, the application of artificial intelligence in discourse with Chinese characteristics has gradually become a research hotspot. This paper compared and analyzed the performance difference between GPT-4o and ERNIE Bot 3.5 in Chinese-English translation of discourse with Chinese characteristics. In this paper, tools such as BLEU value, Flesch-Kincaid Grade Level and Gunning Fog Index were used to conduct a comparative test of

文章引用: 凌颖. 人工智能中国特色话语汉英翻译质量研究[J]. 现代语言学, 2025, 13(4): 575-581.
DOI: 10.12677/ml.2025.134385

translation accuracy and complexity using the official translation as the standard. Preliminary analysis shows that ERNIE Bot 3.5 is closer to the official translation than GPT-4o in terms of accuracy and complexity. The study also found that current AI technologies are not yet able to completely replace the role of human translators in this field. This paper makes suggestions for human-computer collaboration for discourse with Chinese characteristics translation and future directions for optimizing AI translation tools.

Keywords

ChatGPT, ERNIE Bot, Discourse with Chinese Characteristics, Translation Quality Assessment, Translation Studies

1. 引言

近年来，随着人工智能技术的快速发展，机器翻译已经从传统的基于规则和统计的方法，逐步演变为基于神经网络的深度学习模型。这些工具凭借其强大的语言生成能力和自适应语境的特性，逐渐在技术和应用层面取得了显著突破。而中国特色话语因其专业性、正式性，要求译文既能忠实传达原文信息，又要符合目标语言的表达习惯，这对翻译工具提出了更高的要求。在实际翻译实践中，中国特色话语往往面临两大挑战：一是如何保持信息的准确传递和忠实表达；二是如何在译文中平衡语言的可读性与正式性。因此，选取中国特色话语为测试材料将更好地评估人工智能在进行中译英工作时的表现。

2024 年 5 月 14 日，OpenAI 推出了 GPT-4o，GPT-4o 在处理速度上提升了高达 200%，ChatGPT 语言模型的崛起，引领人工智能走向了新的发展阶段，为语言文化及翻译研究带来了巨大变革。2023 年 11 月，百度推出了文心一言 3.5，文心一言 3.5 在翻译准确性和语境理解上得到了显著提升，特别是在处理专业术语和复杂语句时，翻译质量有了大幅度改进。因此，本研究聚焦于这两款国内外前沿的人工智能翻译工具，分析它们在处理中国特色话语时的翻译效果。

为了全面评估翻译质量，本文采用 BLEU 值、Flesch-Kincaid Grade Level、Gunning Fog Index 翻译质量评估工具，涵盖翻译的准确性、流畅性和可读性等多维度指标，并进一步结合文本的特性，探讨两种翻译工具在文本处理和语言生成上的优劣势。

通过对 GPT-4o 与文心一言 3.5 在翻译效果上的对比分析，本文旨在为人工智能翻译技术在政治领域的应用提供实证依据，并为未来 AI 翻译技术的优化和发展提供参考。本研究不仅有助于评估现有 AI 翻译工具的实际应用效果，还可为政策传播和国际交流实践提供技术参考，促进人工智能技术更广泛地应用于语言服务领域，助力人工智能翻译技术的创新与发展。

2. 文献综述

(一) 人工智能翻译

近年来，人工智能翻译技术迅速发展，成为翻译领域的研究热点。大语言模型的应用改变了翻译主体和翻译文本的属性，引发了翻译伦理、译者主体性等问题，但人工翻译仍将继续存在，未来将是人机协作的模式[1]。尽管新一代大语言模型在语用能力上有所提升，但在处理复杂语用现象时仍存在不足，提示人工智能翻译在语用层面的局限性[2]。另外，人工智能翻译在中国特色话语中的应用发现其虽有一

定优势，但人工智能生成的译文在意识形态把控上存在风险，可能导致误导性信息的传播[3]，面临着文化差异挑战等问题[4]。最后，人工智能技术在提高翻译效率和降低成本方面具有优势，但也存在局限性，如准确性不足。

(二) 中国特色话语翻译

中国特色话语翻译作为翻译研究的重要领域，近年来受到广泛关注。它不仅是语言转换的过程，更是跨文化交流和国家形象塑造的重要手段。中国特色话语的英译在翻译中占据重要地位，是国际社会了解中国的关键窗口。对中国特色话语进行翻译研究，有助于提升跨文化传播效果，促进不同文化之间的理解与沟通[5]。同时，中国特色话语翻译应坚持“二元统一”取向，即在语言层面侧重目的语取向，确保译文的地道性和可接受性。在对中国特色话语翻译的原则和策略方面，翻译时需遵循释疑解惑、对接目标语文本修辞、彰显文化自信等原则，并通过语义明晰化、增补文化背景、调整句法结构等手段，提升译文的可理解性和接受度[6]。由于翻译任务较为繁多，且面临较大的时间和质量压力，因此有必要将辅助翻译软件作为翻译工作的有效工具[7]。另外，从国家对外翻译传播能力的角度出发，提出中国特色话语翻译不仅是语言转换，更是构建国家对外话语体系的重要环节，强调翻译实践应注重信息的构建与议程设置，结合目标受众的文化背景和认知习惯，灵活调整翻译策略[8]。当前我国在中国特色话语翻译管理能力方面较强，但在实践能力、传播能力等方面仍有提升空间[9]。

(三) 翻译评价

翻译译文质量评估已在众多领域成为一个热点话题[10]，随着语言服务行业的不断发展，对翻译质量的评估变得日益重要。传统的翻译质量评估模型，虽然在一定程度上推动了翻译质量的标准化评估，但各自存在着局限性[11]。在机器翻译(MT)和自然语言生成(NLG)系统的评估中，BLEU 指标作为一种广泛使用的度量标准，主要基于与标准参考文本的词汇重叠进行计算[12]，是目前常用的自动评估工具之一。然而，随着 AI 翻译技术的发展，自动化的翻译质量评估方法逐渐成为研究热点。其中，基于可读性指标的评估方法(如 Flesch-Kincaid Grade Level 和 Gunning Fog Index)被广泛应用于文本复杂度和阅读难度的量化分析[13] [14]。

3. 研究设计

(一) 研究问题

本研究旨在讨论以下三个问题：第一，ChatGPT-4o 和文心一言 3.5 作为人工智能翻译工具，其生成的翻译质量两者谁更优？第二，ChatGPT-4o 和文心一言 3.5 是否有潜力替代人工译者在中国特色话语翻译中的功能？第三，在人工智能发展时代，人机协作中国特色话语的未来发展方向如何？

(二) 文本选择

首先，中国特色话语具有显著的专业性和政策性，通常包含大量的政策术语、法律条款以及具体的政府措施。所选文本涵盖了农村公路建设的技术细节、资金安排以及对未来发展的展望，要求译者和翻译工具不仅要准确理解专业术语，还要有效传递政策背后的战略意图和实施方案。这对翻译的准确性和流畅性提出了较高要求，尤其是对于人工智能翻译工具而言，如何精确处理这些复杂的术语和结构，是衡量其能力的重要标准。

其次，中国特色话语具有强烈的正式性和规范性，语言表达通常严格、结构清晰。中国特色话语往往不允许有任何形式的失误，译文的准确性、忠实度和规范性尤为关键。在语言上追求简洁、直接，但又充满政策性和信息量。在这种情况下，翻译工具不仅要避免对原文的任何误解，还必须确保译文符合目标语言的正式写作规范。这种对翻译工具精准度和表达规范性的双重要求，恰好能够检验人工智能翻译工具在处理政策性文本时的适应能力。

此外，中国特色话语的翻译还具有一定的时效性和历史背景性。随着国家政策的变化和社会发展的进程，相关文件中的某些表述可能带有浓厚的时代印记。对于人工智能翻译工具来说，如何理解这些背景信息并在翻译过程中作出适当的调整，是检验其综合能力的重要标准。

因此，该研究不仅有助于探讨人工智能翻译在处理政策性、法律性和专业性文本时的表现，也为评估人工智能翻译工具在跨文化传播中的应用提供了一个具体而重要的实践案例。通过这一文本的翻译评估，能够更全面地考察 GPT-4o 和文心一言 3.5 在复杂中国特色话语处理方面的能力。

(三) 检测工具的选取

1) BLEU 检测

BLEU 是 Bilingual Evaluation Understudy 的缩写，最早由 IBM 在 2002 年提出。BLEU 评分机制主要通过评估 n-gram 精度来衡量机器翻译的质量，即计算机器翻译中 n-gram 与参考译文中对应 n-gram 出现的频率，即机器翻译的结果越接近人工参考译文就认定它的质量越高。得分通常在 0 到 1 之间，分数越高，表示译文质量越好。该指标因其简便性和较高的计算效率而受到青睐。

2) Flesch-Kincaid Grade Level

Flesch-Kincaid Grade Level 是评估文本可读性的常用工具之一，主要通过计算句子长度和单词音节数来推测读者理解该文本所需的年级水平。得分通常对应着理解该文本所需的年级水平，得分越高，表示文本越复杂，适合较高年级的读者。该方法因其简便且广泛应用于英语文本评估中，成为衡量翻译可读性的有效工具。

3) Gunning Fog Index

Gunning Fog Index 是一种通过语句结构和词汇复杂度来评估可读性的经典方法，适用于各种类型的英语文本。该指数的得分反映了理解该文本所需的最低教育年级。较高的得分表示文本更难理解，而较低的得分则表明文本较为简明易懂。

4. 实验结果与分析

(一) BLEU

在 Python 27 的环境中运行可得到表 1。

Table 1. BLEU values of the translated texts of ChatGPT-4o and ERNIE Bot 3.5
表 1. GPT-4o 和文心一言 3.5 译后文本的 BLEU 值

	GPT-4o	文心一言 3.5
P1	0.55	0.54
P2	0.24	0.25
P3	0.13	0.14
P4	0.07	0.08
BP	0.68	0.98
BLEU	0.1256	0.1941

从 BLEU 值结果分析，GPT-4o 生成的译文得分为 0.1265，而文心一言 3.5 的得分为 0.1941。可以看出，文心一言 3.5 在 BLEU 值这一指标上优于 GPT-4o，这表明其在与参考译文的 n-gram 匹配中表现得更为出色，尤其是在翻译准确性和文本一致性方面具有一定优势。

从具体的 n-gram 精度(P1、P2、P3、P4)来看，文心一言 3.5 在所有的 n-gram 匹配率上均高于 GPT-

4o。例如，文心一言 3.5 的一元组匹配率(P1)为 0.54，二元组匹配率(P2)为 0.25，而 GPT-4o 分别为 0.55 和 0.24，尽管差距较小，但仍体现出文心一言 3.5 在更高阶的 n-gram 匹配中表现更稳定。而在三元组(P3)和四元组(P4)匹配率方面，文心一言 3.5 也占有一定优势，表明其生成译文在词汇组合和上下文流畅性上可能更贴近参考译文。此外，文心一言 3.5 的惩罚因子(BP)为 0.98，接近于 1，说明其译文长度与参考译文更为接近；相比之下，GPT-4o 的 BP 值为 0.68，表明其生成的译文长度较短，可能存在省略或内容不足的问题。

总体而言，在词汇匹配和长度控制方面，文心一言 3.5 在本次政治文本翻译任务中的表现优于 GPT-4o。然而，这两个系统在应对政治文本的复杂性和多样性方面均有提升空间。

(二) Flesch-Kincaid Grade Level、Gunning Fog Index

GPT-4o、文心一言 3.5 译后文本及官方译文的 Flesch-Kincaid Grade、Gunning Fog Index 值如表 2 所示。

Table 2. Values of Flesch-Kincaid Grade Level and Gunning Fog Index
表 2. Flesch-Kincaid Grade Level、Gunning Fog Index 值

	GPT-4o	文心一言 3.5	官方译文
Flesch-Kincaid Grade Level	14.9	18.3	16.2
Gunning Fog Index	11.45	13.4	12.64

Flesch-Kincaid Grade Level 用于评估文本对读者的教育水平要求，得分越高，表示文本中单词的音节数更多，句子的平均长度更长，需要较高的教育背景才能理解。对于本研究中的中国特色话语，具有语言正式、结构严谨、术语密集等特点。官方译文的 Flesch-Kincaid Grade Level 为 16.2，表明该译文的复杂度较高，适合具备较高教育背景的读者群体。一方面，GPT-4o 的得分为 14.9，较官方译文有所降低，表明其译文在一定程度上简化了语言，降低了理解难度，从而适合具有高中或本科教育背景的读者。另一方面，文心一言 3.5 的得分为 18.3，与官方译文的得分差值大于 ChatGPT-4o，表明其译文在保留原文本复杂性的同时，使用了更多的专业术语和复杂句式，适合具有更高教育背景和更深层次专业知识的读者群体，对理解要求较高。

其次，Gunning Fog Index 是衡量文本复杂度的重要指标，得分越高意味着文本包含更多复杂词汇和长句，阅读难度较大。GPT-4o 的得分 11.45 低于官方译文 12.64，表明其译文使用了较短的句子和较为通俗的词汇，降低了中国特色话语原有的复杂性，使其更适合普通读者理解。而文心一言 3.5 的得分较高，表明其在翻译过程中保留了更多复杂句式和高级词汇。尽管这种处理方法在一定程度上保证了文本的专业性，但也可能增加了其阅读难度。因此，文心一言 3.5 的译文更适合对精确性要求较高的专业读者，而 GPT-4o 则更侧重于易读性和普适性。但文心一言 3.5 与官方译文的 Gunning Fog Index 得分差值小于 GPT-4o，因此文本复杂度来看，文心一言 3.5 的译文更接近于官方。

综上，GPT-4o 的译文通过简化语言和结构，提高了文本的可读性，适合较为广泛的读者群体，尤其是对于普通读者而言，更加易于理解。相比之下，文心一言 3.5 的译文则在较大程度上保持了原文的复杂性和正式性，适合具有较高教育背景和专业知识的读者。官方译文在这两者之间，试图在保持文本严谨性的同时，也兼顾了可读性和理解性。因此，官方译文的风格在一定程度上平衡了复杂性和易懂性，适合更广泛的受众群体。

综上所述，从 Flesch-Kincaid Grade Level 和 Gunning Fog Index 值来看，文心一言 3.5 的译文虽然复杂度更高，但更加接近官方译文的风格，而 GPT-4o 的译文则过于简单。

5. 人机协作翻译中国特色话语未来发展方向建议

尽管人工智能翻译技术在多个领域得到广泛研究，但中国特色话语翻译仍较少受到关注。大部分研究集中在文学文本翻译中。相较于文学文本，中国特色话语的翻译更注重语境和文化背景的传达，对人工智能而言更能体现其翻译水平。另外，对于人工智能译文质量研究，大部分都集中于使用 BLEU 值作为质量评判标准，但 BLEU 指标的有效性在学术界一直存在争议[15]，因此本文另外选择了 Flesch-Kincaid Grade Level 和 Gunning Fog Index，以官方译文为标准，除准确度之外，对译文复杂度和可读性也进行了研究，揭示了人工智能在中国特色话语领域翻译中的潜力与局限性。

结合上述两个指标和 BLEU 值，可以看出无论是精确度还是复杂度，文心一言 3.5 都比 GPT-4o 更加接近于官方译文。GPT-4o 通过降低文本的复杂性和语言难度，使得译文更具可读性和普适性，但同时也由于过于高度提取内容而省略太多，从而可能导致出现内容可能无法完整传达的情况。而文心一言 3.5 则通过更为正式、专业化的语言表达，保持了中国特色话语原有的严肃性和权威性，译文也更加接近于官方。因此，GPT-4o 可被用于需要快速传播和广泛传播的场景中，如当下热门的短视频推送，受众读者多为受教育水平较低或更偏好快速阅读。而文心一言 3.5 则更适合对译文要求高度忠实于原文语言特点的场景，如官方平台或会议，受众读者多为领域专家或具有相关知识基础的人。两者在翻译中国特色话语时的不同表现，反映了其在模型设计和优化目标上的不同取向，也为研究如何在中国特色话语翻译中平衡可读性和忠实性提供了重要启示。

另外，关于人机协作翻译，虽然一些研究提出了人工智能与人工译者的合作模式，但传统的研究往往侧重于技术层面的提升，忽略了人工译者在翻译中的独特作用，且具体的协作方式并未明确界定。依据上述实验，本文提出一种具体的人机协作方法，即人工智能可以根据使用场景来处理大批量标准化的政治文本，而人工译者则应专注于复杂的政治语境和意识形态的准确传达。这一方法不仅为人工智能翻译技术提供了新的应用场景，也为未来的翻译实践提供了理论依据。

6. 结语

本研究通过对比 GPT-4o 与文心一言 3.5 在英译中的表现，结合 BLEU 值、Flesch-Kincaid Grade Level 和 Gunning Fog Index 多维度评价指标，对两种人工智能翻译工具的翻译质量进行了深入分析。研究结果表明，人工智能翻译工具在中国特色话语翻译中表现出一定的潜力，但在面对专业术语、政策表达及文化背景的准确传达方面，仍存在一定局限性。中国特色话语的复杂性和严谨性对人工智能翻译提出了更高的要求，当前技术尚无法完全替代人工译者在该领域的作用。

人工智能翻译工具在未来的发展中应进一步改进其模型，以提高中国特色话语中特定术语的识别和处理能力，增强术语库的建设，确保译文的准确性。同时，需要在可读性与忠实度之间取得平衡，开发适应不同文本类型的翻译策略，满足多样化的受众需求。此外，加强对上下文的理解能力，减少内容遗漏，确保译文完整性也是优化方向之一。在中国特色话语翻译中，人工智能工具可以作为辅助，与人工译者形成互补，提升翻译效率，减少重复性劳动，先由人工智能进行初步翻译，再由人工润色审核，将有助于确保译文质量。未来的翻译策略应结合目标受众的文化背景，调整语言风格，使其更加符合受众的阅读习惯。人工智能翻译技术的发展趋势将朝着更高的精度、减少误译和遗漏的方向迈进，同时深度融合自然语言处理、语义分析等技术，以进一步提升翻译质量。本研究为人工智能翻译在政治文本领域的应用提供了实证参考，同时也为未来人工智能翻译工具的优化方向提供了有益的启示。

本研究为人工智能翻译在中国特色话语领域的应用提供了实证参考，同时也为未来人工智能翻译工具的优化方向提供了有益的启示。在中国特色话语翻译的实际应用中，仍需谨慎评估工具的能力，并结合人工干预，确保翻译质量满足预期需求。另外，本研究主要聚焦于特定中国特色话语的翻译质量评估，

未能全面涵盖其他类型中国特色话语的翻译特点，使用的大语言模型数量也有限，未来研究可进一步扩大样本范围，并引入更多维度的评估标准，以提高研究的全面性和实用性。

参考文献

[1] 胡开宝, 李晓倩. 大语言模型背景下翻译研究的发展: 问题与前景[J]. 中国翻译, 2023, 44(6): 64-73.

[2] 刘海涛, 亓达. 大语言模型的语用能力探索——从整体评估到反语分析[J]. 现代外语, 2024, 47(4): 439-451.

[3] 王赟, 张政. ChatGPT 人工智能翻译的隐忧与纾解[J]. 中国翻译, 2024, 45(2): 95-102.

[4] 文旭, 田亚灵. ChatGPT 应用于中国特色话语翻译的有效性研究[J]. 上海翻译, 2024(2): 27-34.

[5] 周兴华, 王传英. 人工智能技术在计算机辅助翻译软件中的应用与评价[J]. 中国翻译, 2020, 41(5): 121-129.

[6] 邱大平. 论政治话语外宣翻译取向的二元统一[J]. 中南大学学报(社会科学版), 2018, 24(6): 205-212.

[7] 雷鹏飞, 张浮凌. 基于机器翻译软件的外宣文本翻译质量评估研究[J]. 未来与发展, 2024, 48(6): 41-48.

[8] 任文, 赵田园. 国家对外翻译传播能力研究: 理论建构与实践应用[J]. 上海翻译, 2023(2): 1-7.

[9] 胡开宝. 国家外宣翻译能力: 构成、现状与未来[J]. 上海翻译, 2023(4): 1-7+95.

[10] House, J. (2014) Translation Quality Assessment: Past and Present. In: House, J., Ed., *Translation*: *A Multidisciplinary Approach*, Palgrave Macmillan, 241-264. https://doi.org/10.1057/9781137025487_13

[11] 田朋. 翻译多维质量标准 MQM 模型介评与启示[J]. 东方翻译, 2020(3): 23-30.

[12] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 7-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[13] Kincaid, J.P., Fishburne, R.P., Rogers, R.L. and Chissom, B.S. (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research Branch Report 8-75, Naval Air Station Memphis.

[14] Gunning, R. (1952) The Technique of Clear Writing. McGraw-Hill.

[15] Callison-Burch, C., Osborne, M. and Koehn, P. (2006) Re-Evaluating the Role of BLEU in Machine Translation Re-search. *Proceedings of EACL*, Trento, 3-7 April 2006, 249-256.