Modern Linguistics 现代语言学, 2025, 13(10), 92-96
Published Online October 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.13101030

Hans汉斯

基于BLEU指标的机器翻译译文差异对比

——以*China Daily*新闻为例

李　杭[1]，湛　军[2]

[1]上海海事大学外国语学院，上海
[2]上海海事大学经济管理学院，上海

收稿日期：2025年8月25日；录用日期：2025年9月18日；发布日期：2025年9月30日

摘　要

随着全球化进程的加速，机器翻译技术在跨语言交流中扮演着至关重要的角色。本文以*China Daily*新闻为源文本，选用Google Translate、DeepL、有道翻译、微软翻译四款机器翻译工具以及新兴翻译工具ChatGPT对中文和英文新闻进行翻译，并采用BLEU指标对译文进行量化分析。研究重点探讨不同翻译工具在处理中文和英文新闻时的表现差异，研究这些工具在处理不同源语言新闻时是否存在显著差异。结果表明，尽管不同翻译工具在BLEU值上的差异不显著，ChatGPT在中译英任务中表现略优于其他机器翻译工具，但在英译中任务中并未展现明显的优势。本文通过量化分析不同翻译引擎在新闻翻译中的表现并为实际应用中翻译工具的选择提供了实证依据。

关键词

机器翻译，新闻翻译，量化分析

A BLEU-Based Comparative Study of Machine Translation Output Variations

—A Case Study of *China Daily*

Hang Li[1], Jun Zhan[2]

[1]School of Foreign Languages, Shanghai Maritime University, Shanghai
[2]School of Economics and Management, Shanghai Maritime University, Shanghai

Received: August 25, 2025; accepted: September 18, 2025; published: September 30, 2025

Abstract

With the acceleration of globalization, machine translation technology plays a crucial role in cross-

文章引用: 李杭, 湛军. 基于 BLEU 指标的机器翻译译文差异对比[J]. 现代语言学, 2025, 13(10): 92-96.
DOI: 10.12677/ml.2025.13101030

linguistic communication. This study takes *China Daily* news articles as the source texts and selects four mainstream machine translation tools—Google Translate, DeepL, Youdao, and Microsoft Translator—along with the emerging tool ChatGPT to translate Chinese and English news. The BLEU metric is applied to conduct a quantitative analysis of the outputs. The research focuses on the performance differences of these tools when handling Chinese and English news, examining whether significant variations exist depending on the source language. Results show that although the BLEU score differences among the tools are not substantial, ChatGPT performs slightly better than other machine translation systems in Chinese-to-English tasks, while it does not demonstrate a clear advantage in English-to-Chinese tasks. By quantitatively analyzing the performance of different translation engines in news translation, this paper provides empirical evidence and offers practical guidance for the selection of translation tools in real-world applications.

Keywords

Machine Translation, News Translation, Quantitative Analysis

1. 引言

自从统计机器翻译(SMT)到神经机器翻译(NMT)的转变，机器翻译的准确性和流畅性得到了极大的提高。目前，诸如 Google Translate、DeepL、Micorsoft Translate 等机器翻译软件已经被广泛应用于个人、学术和商业领域。尽管这些翻译工具在某些语言对上表现出色，但不同机器翻译软件在处理相同文本时，往往存在显著的质量差异。

本研究旨在通过比对不同机器翻译软件对相同原文的翻译结果，重点分析英译中和中译英两种翻译任务中的质量差异回答以下几个问题：1) 中国机器翻译引擎是否在英译中新闻翻译任务中较外国翻译引擎更有优势？2) 海外机器翻译引擎是否在中译英新闻翻译任务中较中国本土翻译引擎更有优势？3) ChatGPT 作为新兴大语言模型翻译工具是否较目前主流的机器翻译引擎在新闻翻译上更有优势？

2. 文献综述

机器翻译的质量评估方法经历了从单一参数到多维度考量、从静态评估到动态适应的过程。随着自然语言处理和人工智能技术的发展，研究者们提出了多种翻译质量评估方法，以确保译文能够满足不同用户的需求和应用场景。总体而言，翻译质量评估的研究路径大致经历了三个阶段：基于人工评价的方法、基于自动化指标的方法，以及融合人工与机器的综合评估方法。

在国内研究方面，何三宁[1]探讨了翻译质量评估中的多种参数，包括语言、实体和思维等维度，指出这些参数是评价译文质量的重要依据，奠定了多维度考察的理论基础。王华树和张彦希[2]则从技术角度出发，分析了计算机辅助翻译质量控制工具在翻译实践中的作用，强调了工具化、系统化评估方法的必要性。刘亚猛[3]回顾了翻译质量评估领域长期存在的问题，指出不同评估模式背后的前提假设存在冲突，其内在结构也有明显缺陷，这一批评推动了更加综合化的研究思路。

在定量研究方面，王金铨[4]通过构建汉英自动评分模型发现，基于 N 元组匹配的方法在衡量译文“信度”方面具有较好效果；而流利度的评价则需要结合词法、句法等形式特征进行综合分析。这一研究为 BLEU 等自动化指标的使用提供了支持。随后，王均松[5]介绍了由翻译自动化用户协会(TAUS)提出

的 DQF 动态质量评估框架，推动了翻译质量评估由静态向动态转变，更加贴近实际翻译应用场景。王金铨和牛永一[6]提出了一种结合语言形式特征和语义内容特征的计算机辅助翻译质量评估方法，进一步提升了评估的全面性与科学性。王金铨和何泊稼[7]则对国内外翻译质量评价研究进行了系统梳理，从质化评估与量化评估两个维度总结了研究成果与不足，指出未来的发展趋势在于融合多种方法，形成跨学科的综合评估体系。

除了 BLEU 之外，学界还提出了多种自动化指标。例如，NIST 指标在 BLEU 的基础上更加注重信息量的权重分布；METEOR 指标考虑了词形变化、同义词和词序等因素，在评估译文流畅性方面具有一定优势；TER (Translation Edit Rate)则通过计算将机器译文改为人工译文所需的编辑操作数量来衡量翻译质量。这些方法的出现，表明翻译质量评估正逐步从单一指标向多维度、多层次方向发展。

尽管如此，BLEU 指标因其计算简便、可操作性强、自动化程度高，仍然是机器翻译研究中最常用的评估工具，并在学术界与工业界得到广泛应用。本研究将以 BLEU 指标为主要分析工具，对不同翻译引擎的译文效果进行量化分析，探讨其在中英互译任务中的表现差异，从而为机器翻译在新闻翻译场景中的应用提供实证依据。

3. 研究方法

3.1. 数据选择

为了进行机器翻译质量分析，本研究选择了 5 篇中文新闻和 5 篇英文新闻作为源文本。本文选用来自 *China Daily* 的新闻作为数据来源。*China Daily* 是中国知名的英文新闻平台，涵盖广泛的新闻领域，包括国内外时事、政治、经济、文化、科技等内容。选取 *China Daily* 的新闻，可以确保数据的多样性与代表性，并为后续分析提供可靠的文本基础。这些新闻内容在国内外均有一定的影响力，能够反映出一定的社会趋势与公众关注的热点问题。

每篇新闻的字数大致在 1000 至 2500 字之间，总字数约为 20,000 至 25,000 字。这一规模的文本能够充分展现机器翻译在处理较长专业文本时的表现，且不会造成计算过度负担。选取的文献类型和字数也能帮助我们观察翻译系统在不同主题文献中是否存在不同的表现。

3.2. 机器翻译软件选择

本文将选用 Google Translate、DeepL、有道翻译、微软翻译四款机器翻译对译文进行处理，并对译文进行比较。此外，ChatGPT 虽然不属于机器翻译的范畴，但其已被广泛应用于翻译实践中，出于实践目的，本文也会将其译文与上述译文进行比较。在使用 ChatGPT 进行翻译时，仅给出新闻中英或英中翻译的指令，译文统一采用未经后续修缮的第一版。

3.3. 译文质量评价方法

本文采用 BLEU (Bilingual Evaluation Understudy)指标对机器翻译质量进行量化评估。BLEU 指标通过计算译文与参考译文之间的 n-gram 重合度来衡量翻译的精确度，是目前评价机器翻译质量的主流方法之一。

通过计算得到相应指标之后，本文将对这些指标进行比较，分析国内外机器翻译引擎是否在翻译本土新闻时更具优势，同时也会探讨 ChatGPT 是否较目前流行的机器翻译更具优势。

3.4. 实验流程与数据处理

翻译前对源文本进行清理与规范化处理，例如去除特殊字符、图片、多余段落等。在获取数据后，

将数据汇总至 excel 表格，制成可视化表格进行数据对比分析。本文将采用方差分析(ANOVA)来判断不同翻译工具的表现是否有显著差异。

4. 结果与分析

各组译文 BLEU 值见表 1 和表 2。

Table 1. The BLEU score of MT and ChatGPT's C-E translation
表 1. 机器翻译及 ChatGPT 中翻英译文的 BLEU 值

	有道	微软	Google	DeepL	ChatGPT
机器人	0.761	0.757	0.755	0.747	0.738
奥运	0.561	0.545	0.579	0.549	0.635
文化	0.515	0.562	0.497	0.418	0.585
政策	0.660	0.644	0.656	0.658	0.693
睡眠	0.651	0.653	0.633	0.620	0.653
均值	0.630	0.632	0.624	0.598	0.661

Table 2. The BLEU score of MT and ChatGPT's E-C translation
表 2. 机器翻译及 ChatGPT 英翻中译文的 BLEU 值

	有道	微软	Google	DeepL	ChatGPT
世界自然	0.621	0.564	0.655	0.895	0.463
脊椎	0.394	0.372	0.336	0.365	0.336
短视频	0.388	0.313	0.371	0.350	0.399
大蒜	0.584	0.562	0.608	0.545	0.578
炸鸡	0.582	0.547	0.590	0.537	0.528
均值	0.514	0.472	0.512	0.538	0.461

中译英第一组有道均值为 0.630 第二组微软均值为 0.632，第三组 Google Translate 均值为 0.624，第四组 DeepL 均值为 0.598，第五组 ChatGPT 均值为 0.661，总体均值为 0.6377，组间平方和为 0.036565，组内平方和为 0.306303，F 值为 0.5974。

在显著性水平 0.05 下，查得的临界 F 值大约为 2.866。F 值 0.5974 小于临界值 2.866，这意味着不同翻译软件之间的差异不显著。换句话说，从统计学角度来看，不同机器翻译软件在处理中译英文本时，他们之间的 BLEU 值没有表现出显著差异。不过就均值而言，ChatGPT 在译文的表现上会略优于其他机器翻译引擎。

英译中第一组有道均值为 0.514 第二组微软均值为 0.472，第三组 Google Translate 均值为 0.512，第四组 DeepL 均值为 0.538，第五组 ChatGPT 均值为 0.461，总体均值为 0.4994，组间平方和为 0.036565，组内平方和为 0.42446，F 值为 0.248。

在显著性水平 0.05 下，查得的临界 F 值大约为 2.866。F 值 0.248 小于临界值 2.866，这意味着不同翻译软件之间的差异不显著。就均值而言，ChatGPT 在英译中上就机器翻译而言没有优势。

5. 总结

本文通过对四款机器翻译工具和 ChatGPT 的译文质量进行评估，发现不同翻译工具译文在 BLEU 值上的差异在统计学角度上不显著。因此，不同源语言的机器翻译引擎在将新闻翻译成当地语言上均没有显著优势。在没有特殊指令下的 ChatGPT 在新闻中译英任务中表现略优于其他机器翻译，在新闻英译中任务中则表现平平。

虽然当前的研究结果未能显示出翻译工具之间显著的差异，但数据表明翻译任务的语言方向(如中译英和英译中)可能会影响翻译工具的表现。未来的研究可以探索更多的翻译质量评估指标，如流利度、语境适应性等，结合人工评估与自动化评估，进一步完善机器翻译工具的选择和优化。此外，本研究的样本量和翻译任务的单一性可能限制了结果的广泛性。未来研究可以扩展文本的种类与规模，分析不同翻译工具在其他领域(如法律、医疗等)的表现差异，为翻译实践提供更多参考，以更全面地评估机器翻译之间的差异，为译员的翻译提供更加准确的选择。

基金项目

2024 年上海高校一流本科课程建设(沪教委高(2024)2 号)；上海高校重点课程建设项目(沪教委高(2021)34 号)；2022 年上海海事大学“国家双万课程建设”项目(沪海大(2021)200 号)前期成果。

参考文献

[1] 何三宁. 再探翻译质量评估参数[J]. 中国翻译, 2012, 33(2): 27-31.

[2] 王华树, 张彦希. 技术视角下的翻译质量控制研究[J]. 语文学刊(外语教育教学), 2015(4): 1-5.

[3] 刘亚猛. 翻译质量评估的理想与现实[J]. 中国翻译, 2018, 39(2): 8-16, 128.

[4] 王金铨, 万昕, 董子云. 翻译质量评价方法及其在计算机翻译评价系统中的应用[J]. 中国翻译, 2018, 39(4): 73-78.

[5] 王均松. 翻译质量评估新方向: DQF 动态质量评估框架[J]. 中国科技翻译, 2019, 32(3): 27-29.

[6] 王金铨, 牛永一. 计算机辅助翻译评价系统中的翻译质量评估[J]. 上海翻译, 2023(6): 52-57.

[7] 王金铨, 何泊稼. 基于神经网络模型的翻译语义质量量化评价[J]. 中国外语, 2024, 21(1): 92-101.