doi:10.6043/j.issn.0438-0479.202412010

基于大语言模型的智能译后编辑系统构建与应用

赵泽龙[1,2],朱俊国[1,2*]

(1.昆明理工大学信息工程与自动化学院,云南 昆明 650600;2.昆明理工大学云南省人工智能重点实验室,云南 昆明 650600)

摘要:[**目的**]尽管神经机器翻译(NMT)已经成为当前机器翻译(MT)的主流方法,但在关键场景下,仍需要对 NMT 模型的输出进行译后编辑,以纠正错误并提高质量.[**方法**]提出了一种基于大语言模型 GPT-4o mini 自动译后编辑系统 SSAE(smart suggest AutoEdit),SSAE 系统通过引入术语约束以及多维翻译建议的自反馈机制,来确保翻译过程中术语使用的一致性和准确性,进而提高关键场景下翻译的整体质量.[**结果**]实验结果显示,该系统对翻译进行有意义且值得信赖的编辑,有助于提高其总体质量,并消除了翻译中的主要错误.在处理专业领域文本时,它的表现尤为出色,显著减少了术语错误和一致性问题.[**结论**]值得一提的是,本文提出的自动译后编辑系统在 WMT-23 术语翻译任务中提升了德语-英语、中文-英语和英语-捷克语的翻译性能.

关键词:神经机器翻译;大语言模型;自动译后编辑

中图分类号:TP 391.2 **文献标志码**:A **文章编号**:0438-0479(2025)06-0958-12

Construction and application of intelligent post-editing editing system based on large language model

ZHAO Zelong[1,2],ZHU Junguo[1,2*]

(1. Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China;
2. Yunnan Key Laboratory of Artificial Intelligence,Kunming University of Science and Technology,Kunming 650500,China)

Abstract:[**Objective**] Although neural machine translation (NMT) has become the mainstream method for current machine translation (MT),the need to post-edit outputs of NMT models in critical scenarios to correct errors and improve quality remains. [**Methods**] Herein we propose an automatic translation post-editing system called Smart Suggest AutoEdit (SSAE),based on the large language model GPT-4o mini. The SSAE system ensures consistency and accuracy in terminology usage during the translation process by introducing term constraints and a self-feedback mechanism for multidimensional translation suggestions, thereby improving the overall quality of the critical scenarios translation. [**Results**] Experimental results show that the system performs meaningful and reliable edits on translations, hence helping to enhance their overall quality and eliminating major errors in the translations. When handling texts in specialized fields,the system performs very satisfactorily,thus reducing terminology errors and consistency issues significantly. [**Conclusion**] Notably,our proposed automatic translation post-editing system has achieved state-of-the-art performance in the WMT-23 terminology translation tasks for Chinese-English,German-English,and English-Czech.

Keywords:neural machine translation;large language models;automatic post-editing

大语言模型(LLM),特别是基于生成式预训练 Transformer 架构的模型,如 OpenAI 团队开发的

收稿日期:2024-12-11 **录用日期**:2025-04-04

基金项目:国家自然科学基金地区基金(62166022);云南省"兴滇英才支持计划"(KKXX202403023)

* **通信作者**:zhujunguohit@gmail.com

引文格式:赵泽龙,朱俊国.基于大语言模型的智能译后编辑系统构建与应用[J].厦门大学学报(自然科学版),2025,64(6):958-969.

Citation:ZHAO Z L,ZHU J G. Construction and application of intelligent post-editing editing system based on large language model[J]. J Xiamen Univ Nat Sci,2025,64(6):958-969. (in Chinese)

GPT-3[1]和后续推出的 ChatGPT[2],已在自然语言处理(NLP)任务、复杂推理能力展示以及智能代理系统构建等方面展现出卓越的性能,且这些 LLM 能够整合多个 NLP 任务,并能为人类查询生成详细且全面的回应[3-5]. 除此之外,LLM 还能对后续问题做出恰当回应并在多轮对话中保持一致性. 然而,尽管 LLM 在 NLP 多个任务已取得突破,但在翻译领域,其会生成幻觉的问题仍是其完全替代传统有监督神经机器翻译模型的关键障碍[6]. 最近,研究发现 ChatGPT 等 LLM[7]开始超过商业机器翻译(MT)系统[8-11]. 然而,许多语言对的表现在监督模型中仍旧优于 LLM[12-13],LLM 的性能在模型、语言和翻译方向等方面表现参差不齐[14].

自动译后编辑(automatic post-editing,APE)的目标是通过修正系统性错误,以及适应特定领域的风格和词汇,从而改善任意 MT 系统的输出质量[15-16]. 虽然研究表明 APE 能有效地减少翻译错误并提升翻译质量,但关于加入术语约束的效果的研究仍相对匮乏. 术语约束指通过预定义术语表或规则,强制模型在翻译过程中遵循特定术语的拼写、格式或语境用法,以解决专业场景下的术语一致性问题. 该类系统在诸多情境下具有价值,比如,一些内容提供商为其特定领域精心制定了术语列表,而这些术语列表则指定了首选的技术性翻译. 对于跨语言信息检索,词汇约束的 APE 也大有裨益. 在展示检索出的文档片段时,如果原文中含有查询术语,那么它应该出现在翻译的输出中,因为这能向最终用户明确其相关性. 在这个情况下,查询本身即充当了术语.

先前机器翻译的研究中[17-19]已经指出,LLM 在处理专业术语时存在术语不一致的困难. 尽管已有多项研究建议通过提示词来鼓励 LLM 检查术语翻译的准确性,但结果往往不太理想. 即使是最先进的模型[20],在术语翻译的成功率方面也存在困难,这主要是由于上下文的细微差别和领域特定语言的复杂性. 因此,采用术语约束等译后编辑的额外措施是提高翻译准确性和一致性的必要手段.

在本项研究中,探讨了自我生成的多维翻译建议和术语约束在指导 LLM 精细化能力方面的作用,如图 1 所示. 可以看出,本文提出的 SSAE(smart suggest AutoEdit)无需在后期编辑前对翻译进行任何质量估计或错误检测. 本文的贡献主要体现在以下几个方面:

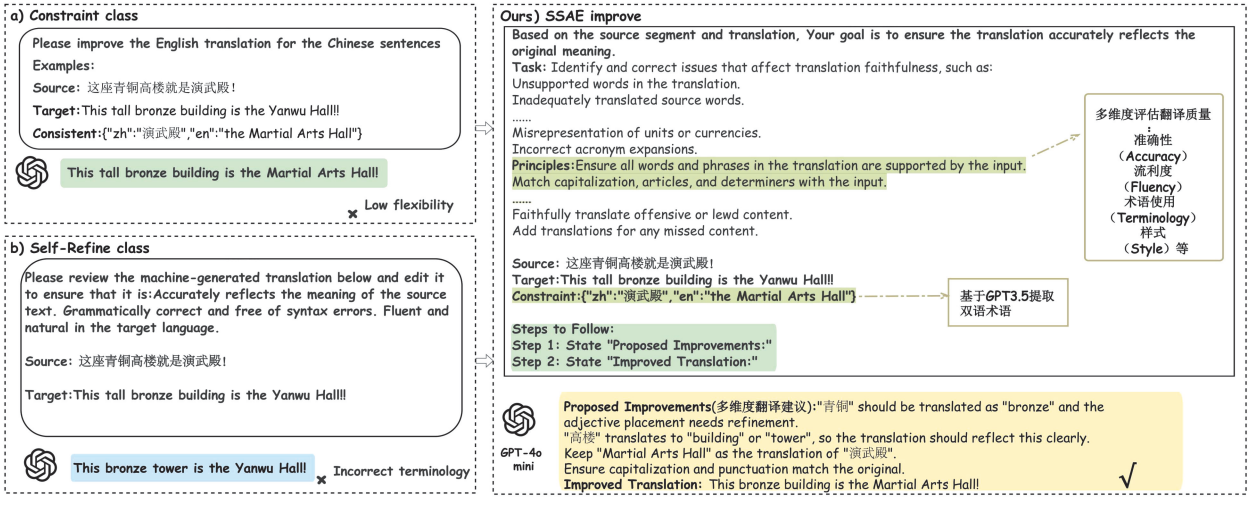

图 1　多维翻译建议和术语约束的自动译后编辑系统

Fig. 1　Multidimensional translation suggestions and terminology-constrained automatic post-editing system

(i) 创新的术语约束应用:通过强化术语的约束以及一致性,提升翻译质量,并增强专业翻译的精确度.

(ii) GPT-4o mini 大模型的有效利用:本文展示了 GPT-4o mini 在自动译后编辑中的高效性和可靠性. 该系统生成了符合人类判断的有意义的编辑,从而提升了翻译的可信度至最先进的水平.

(iii) 经济效益显著:该系统优化了译后编辑过程,降低了人工干预的需求,从而实现了最为经济和便捷的翻译质量提升.

(iv) 综合评估和实践验证:通过对各种商业翻译输出和各个语言对训练顶尖神经机器翻译(NMT)模型输出的实验验证,该系统在实际应用中的性能表现优良. 而且,该系统即插即用,既方便又有效.

我们已证明,在 WMT-23 术语翻译任务中[21],通过使用自反馈的多维翻译建议及术语约束为特定的译后编辑自动化系统,使中文-英语(Zh-En)、德语-英语(De-En)和英语-捷克语(En-Cs)翻译性能得以提高.

1 相关工作介绍

对于 APE 任务,使用神经编码器-解码器模型的多源变体是目前最流行的方法[22]. 其中,多源 Transformer(MST)在 2018 年取得了最先进的成果[23]. 基于注意力机制(AT)模型[20],MST 模型由两个 Transformer 编码器和一个解码器组成. 源句子和 MT 系统输出分别被输入到两个编码器,然后将其输出连接,再输入到解码器以执行后期编辑.

最近的研究已经探索了 APE 的替代架构. 例如,2019 年 APE 任务的获胜者[24]使用了基于 BERT 的编码器和解码器. Gu 等[25]提出了 LevT 模型,并展示了其在 APE 任务中的实用性.

APE 的任务是校正 MT 系统输出中的系统误差. APE 模型采用两个序列作为输入:要翻译的源语言句子和由 MT 系统将该句子翻译成目标语言的输出. 预期的输出是 MT 系统初始翻译的校正版本. 受约束的 APE 允许指定术语约束:源语言输入中的一个或多个短语的翻译可以被预先指定为附加输入. 在执行 APE 任务时,受约束的 APE 模型必须使用所提供的术语约束.

近年来,大多数研究依赖 LLM,减少了对后期编辑监督示例的需求. Chen 等[26]使用 GPT-3.5 Turbo 对多种系统和语言的 MT 输出进行了优化,结果显示,基于字符串的质量度量显著降低,而基于神经网络的评估指标(如 COMET-DA/QE)未显著波动. 在此 DA 和 QE 版本之间的区别在于 COMET-DA 需要源语句、候选译文和参考译文,而 COMET-QE 不需要参考译文,从而减少了 MT 输出中的“翻译腔”. Raunak 等[27]将译后编辑框定为思路链,进行译后编辑. Kojima 等[28]表明 GPT-4 显著提高了多个语言对(尤其是英语)的 MS 翻译输出的 COMET 分数. 然而,这也导致了翻译幻觉和经济效益不佳的问题. 最后,Xu 等[29]提出将迭代精化作为一种搜索策略,进一步提升翻译质量. 这些研究展示了将 LLM 应用于译后编辑的潜力,显著提高了翻译的质量和一致性. 在本项研究中,受多维质量指标模型(MQM)(附录 A(http://jxmu.xmu.edu.cn/upload/html/20250605)中)的启发,研究了使用 LLM 的 MT 精化,不同的是合并了类 MQM 的自反馈的翻译建议与术语约束作为反馈以精化 MT 输出.

2 实验设置

2.1 数据集

利用 WMT 23 术语翻译共享任务[21]的测试集① De-En、Zh-En 以及 En-Cs,由于 LLM 可能在其训练集中使用了这些数据集(如 WMT 21 术语翻译共享任务). 故在本实验中排除之前的数据集,以降低数据污染的潜在风险. 表 1 提供了相关测试集的统计数据.

表 1 测试数据集大小、每行平均单词数和每行平均术语数

Tab. 1 The size of test dataset, average number of words per line, and average number of terms per paragraph

$X\rightarrow Y$	Count	X/Y Words	Terms
De→En	2 963	22.2/22.6	3.8
Zh→En	2 640	9.7/36.9	1.1
En→Cs	3 005	25.6/21.6	3.6

2.2 LLM 和基线

考虑到性能、可靠性和多语言处理上的优势,本研究选取 GPT-4o mini② 和 GPT-3.5 Turbo③ 进行实验. 这两种模型是目前公开可用的 LLM 模型中,在能力与成本效益方面具有代表性的选择[30]. 本文根据不同的约束条件设计了不同的提示词,将 LLM 的系统角色设定为机器翻译译后编辑者. 已有研究表明在思维链(CoT)[27]提示策略的引导下,LLM 能够先对给定的初始翻译进行分析,进而生成经过改进的最终译文. 对于译后编辑,在 4 种设置下进行实验:(i) 利用合成数据去改善开源基础模型,再进行自动译后编辑. (ii) 使用 GPT 系列模型预测,再进行自动译后编辑. (iii) 利用 LLM 生成翻译,再进行自动译后编辑. (iv) 利用商业翻译服务生成翻译,再进行自动译后编辑.

为了在 WMT-23 测试集上生成企业级初始翻译,使用以下开源神经机器翻译模型和商业翻译服务作为基线系统.

(1) 开源基础 NMT 模型:

OPUS-MT-de-en④, 德语到英语翻译模型;

① 数据集:https://wmt-terminology-task.github.io/

② GPT-4o mini:https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

③ GPT-3.5 Turbo:https://openai.com/

④ opus-mt-de-en:https://huggingface.co/Helsinki-NLP/opus-mt-de-en

OPUS-MT-zh-en①,中文到英语翻译模型;OPUS-MT-en-cs②,英语到捷克语翻译模型.

(2) GPT 系列模型

GPT-3.5 Turbo,在"无约束"和"有约束"两种提示策略下使用;GPT-4o mini,作为性能与成本均衡的优选;GPT-4o-mini-agent③,由吴恩达教授基于 GPT-4o mini 开源的智能体翻译项目,用于探索自主翻译工作流.

(3) 代表性 LLM

NLLB (no language left behind)④: 一个旨在提供高质量翻译的多语言模型, 支持多达 200 种语言的翻译任务.

mBART⑤: 一个预训练的多语言序列到序列生成模型, 擅长生成流畅且上下文相关的文本.

Gemini⑥: 专注于高效生成自然对话的模型, 优化了对话的上下文理解和生成能力.

DeepSeek-Chat⑦: 具有深度语境理解的聊天生成模型, 能够生成自然且相关的对话内容.

(4) 商业翻译服务

谷歌翻译⑧、微软翻译⑨和百度翻译⑩作为高性能基线,代表了当前业界领先的商用翻译水平.

为了探索不同的模型特性及其对翻译质量和性能的影响,方便比较研究译后编辑系统的效果,我们从 WMT-2023 术语翻译任务获取了相关参与者提交的系统. 有关 WMT-2023 术语翻译任务参与者的详细信息,请参见附录 B(http://jxmu.xmu.edu.cn/upload/html/20250605).

2.3 指　标

在 sacreBLEU[31] 工具包⑪中实现的传统 BLEU 指标[32]:使用指数平滑来计算翻译质量.

翻译编辑率(TER)[33]:衡量为了使翻译完全匹配参考文献所需的编辑数量.

现代神经指标(基于参考的 COMET-DA 评分)[34]:COMET-DA 评分通过参考文献的平均长度和现代神经网络方法进行归一化.

根据 MT 评估的最近趋势[35],使用 ChrF:用于一般翻译质量评估. 在这种情况下,还关注由 ChrF 捕获的 n-gram 的精确匹配.

术语成功率[21]:用于评估术语翻译的成功率. 参考 WMT23 组织提出的方法,具体而言,对 MT 中出现的术语与字典中的对应术语进行统计. 鉴于我们拥有每个句子的参考术语翻译,在翻译输出中应用了这些术语的子串搜索. 为实现这一点,使用正则表达式,并计算匹配的次数. 具体的计算公式如下:

$$\text{normalized\_match\_count}=\frac{\text{match\_count}}{\text{total\_terms}},$$

其中:match_count 表示机器翻译中出现的术语集合,total_terms 表示参考中的参考术语集合.

所有这些指标的分数均在 0 到 1 的范围内.

3　系统设置方法与步骤

我们采用广泛使用的 GPT-4o mini 以及 GPT-3.5 Turbo 来设计并实现术语约束与自反馈的多维翻译建议的译后编辑自动化系统—SSAE,如图 1 所示.

SSAE 系统模块设计:

(i) 读取模块:负责读取源文本、对应的待改善翻译文本以及 GPT-3.5 Turbo 提取的术语约束文本.

(ii) 自反馈模块:受到 MQM 框架的启发并引入了类 MQM 来增强 LLM 识别翻译错误的能力. 与 MQM 错误标注的不同之处在于,这个模块强调翻译过程中的逐步改进,而非单纯地记录错误,详细比较见附录 A (http://jxmu.xmu.edu.cn/upload/html/20250605). 为保证目标语言翻译的忠实度和流畅度,本文设立了提示词(附录 C(http://jxmu.xmu.edu.cn/upload/html/20250605)),这些提示词假设其用户为源语言和目标语语言的母语者以及翻译专家,其目标在于强调翻译要忠实于源文本,包括注意到拼写错误和占位符的处理. 将提示词设置为 12 个常见问题,如不支持的单词、未充分翻译的单词、标点符号的差异等,并要求它们被逐一修复. 另外,建立了 12 个原则,如禁止插入不支持的内容、

① opus-mt-zh-en:https://huggingface.co/Helsinki-NLP/opus-mt-zh-en
② opus-mt-en-cs:https://huggingface.co/Helsinki-NLP/opus-mt-en-cs
③ gpt4-o-mini-agent:https://github.com/andrewyng/translation-agent
④ NLLB:https://huggingface.co/facebook/nllb-200-distilled-600M
⑤ mBart:https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
⑥ Gemini:https://deepmind.google/technologies/gemini/flash/
⑦ DeepSeek-Chat:https://huggingface.co/deepseek-ai/DeepSeek-V2-Chat
⑧ 谷歌翻译:https://translate.google.com/
⑨ 微软翻译:https://translator.microsoft.com/
⑩ 百度翻译:https://fanyi.baidu.com/mtpe-individual/multimodal#/
⑪ https://github.com/mjpost/sacrebleu

需严格遵循输入的大写方式、确保包含适当的冠词和限定词、不允许遗漏输入文本的任何部分等,这些都为确保翻译的质量和忠实度服务. 在这样的设计引导之下,译后编辑过程严格遵循质量标准,从而提高了翻译的准确性和流畅度. 此模块还会提供有关翻译的建议反馈. 该模块的内部机制如图 2 所示.

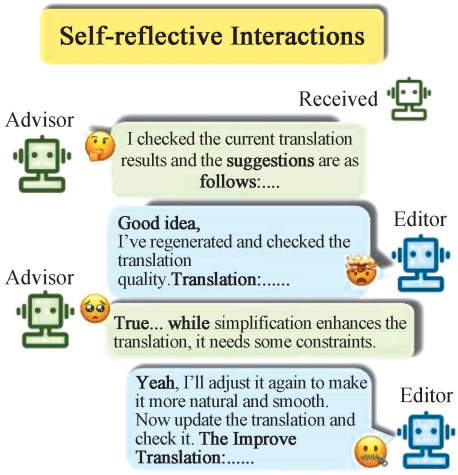

图 2 自反馈模块

Fig. 2 Self-feedback module

(iii) 修正模块:根据是否有术语约束,模型执行不同策略. 若有术语约束,模型结合术语约束和自反馈模块的输出改善初始翻译. 若无术语约束,仅依据自反馈模块的输出改善翻译.

4 实验结果及分析

在以下几个方面对 SSAE 系统进行了检验:1) 开源基础模型生成的翻译;2) LLM 生成的翻译;3) GPT 系列模型生成的翻译;4) 商业翻译服务生成的翻译. 从多维角度验证了我们系统的有效性. 最后,将其与参加 WMT-2023 术语翻译任务的相关参与者提交的系统进行比较,以证明该系统的优越性.

4.1 开源基础模型生成的翻译

开源基础模型如 3.2 节所述,其数据处理流程遵循 Moslem 等[36]的方法,首先使用 GPT-3.5 Turbo 模型,并依据 WMT23 术语翻译组织者提供的术语表,生成双语术语对. 具体而言,对于给定的每个目标术语,模型需要生成多个翻译对,包括源(例如德语)和目标(例如英语),并为每个术语生成 5 个例句. 由于生成的句子数量有限,为了过滤生成的数据,首先根据源和目标从整个数据集中删除重复的句子,然后使用 fastText① 和 pycld2② 库对数据两侧进行语言检测,以确保生成的句子是我们所需的语言. 排除了分数低于特定阈值的任何句子,即 FastText 设置为 0.6,pycld2 设置为 60. 合成数据的总体数目和使用数目如表 2.

表 2 ChatGPT 生成的基于术语的双语数据,用于微调 OPUS 模型

Tab. 2 Terminology-based bilingual data generated by ChatGPT for fine-tuning the OPUS model

语言对	合成数据总量	过滤数据量	合成数据使用量
De-En	14 965	4 435	10 530
Zh-En	4 473	1 244	3 229
En-Cs	13 998	4 792	9 206

先使用 OPUS 系列模型直接翻译得到结果,再使用合成数据进行微调来检验术语翻译的质量. 结果如表 3 所示.

表 3 未微调 OPUS 模型的模型结果与微调 OPUS 模型的模型结果

Tab. 3 The model results of the OPUS model without fine-tuning and the model results of the OPUS model with fine-tuning

语言对	模型	BLEU↑/%	ChrF↑/%	TER↓/%	COMET-DA↑	术语成功率↑/%
De-En	OPUS-MT-de-en	21.73	53.47	76.82	0.793 5	36.80
	OPUS-MT-de-en 微调	23.54	54.46	72.64	0.731 0	46.20
Zh-En	OPUS-MT-zh-en	6.89	29.74	85.07	0.637 6	8.10
	OPUS-MT-zh-en 微调	8.96	34.31	84.09	0.662 3	61.80
En-Cs	OPUS-MT-en-cs	29.22	56.27	61.99	0.813 2	36.50
	OPUS-MT-en-cs 微调	34.09	61.38	56.92	0.882 2	48.50

注:指标右侧↑表示数值越大表现越好,↓表示数值越小表现越好.

① https://fasttext.cc/docs/en/language-identification.html

② https://github.com/aboSamoor/pycld2

虽然合成数据对于术语翻译的效果有帮助,但与顶尖模型的表现还是有一定的差距.

在模型-OPUS 系列模型得到翻译结果后去应用 SSAE 系统以及将 SSAE 系统应用于微调之后的 OPUS 系列模型.所有的结果如表 4.通过对比译后编辑前后的性能指标来评估改进效果.所用的性能指标包括 BLEU、ChrF、TER、COMET-DA 和术语成功率.结果显示,译后编辑在所有指标上均显著提升了翻译质量.

表 4　OPUS 模型的初始翻译和微调后应用 SSAE 系统的译后编辑质量

Tab. 4　The post-editing quality of applying the SSAE system after generating the initial translation with the OPUS model and fine-tuning

语言对	模型	BLEU↑/%	ChrF↑/%	TER↓/%	COMET-DA↑	术语成功率↑/%
De-En	OPUS-MT-de-en+SSAE	35.67(+13.94)	67.04(+13.57)	63.63(−13.19)	0.810 0(+0.016 5)	87.65(+50.85)
	OPUS-MT-de-en-微调+SSAE	33.00(+9.46)	64.32(+9.86)	65.53(−7.11)	0.802 5(+0.011 5)	80.05(+33.85)
Zh-En	OPUS-MT-zh-en+SSAE	13.14(+6.25)	39.40(+9.66)	77.73(−7.34)	0.735 6(+0.098 0)	79.25(+71.15)
	OPUS-MT-zh-en-微调+SSAE	13.69(+4.73)	40.07(+5.76)	77.89(−6.20)	0.747 8(+0.085 5)	77.43(+15.63)
En-Cs	OPUS-MT-en-cs+SSAE	49.51(+20.29)	72.71(+16.44)	43.46(−18.53)	0.882 2(+0.069 0)	85.94(+49.44)
	OPUS-MT-en-cs-微调+SSAE	47.99(+13.90)	72.10(+10.72)	44.28(−12.64)	0.882 2(+0.033 0)	84.09(+35.59)

注:指标右侧的()表示与表 3 模型的初始翻译质量相比使用 SSAE 系统提升的质量.

正如实验结果所看到的,我们提出的 SSAE 系统,在几乎所有的模型翻译效果上均有较大的提升,值得注意的是:虽然经过微调的 OPUS 基础通用翻译模型会产生比原始更优的翻译效果,但在译后编辑之后整体的质量不如原始 OPUS 模型直接译后编辑产生的效果,正如先前的研究中所描述[37-38],可能减少对神经机器翻译系统的硬性约束可以带来显著的性能提升.实验数据显示,在无微调去增加约束的环境下,模型展示了更大的灵活性.这种灵活性使模型能够更好地适应复杂的语言结构,提高翻译的准确性和流畅性.无约束模型的译后编辑翻译在多个指标上表现出色,尤其是在 BLEU 和 ChrF 分数上,较有约束模型的提升尤为显著.这验证了灵活性对提高翻译质量的重要性.

4.2　LLM 生成的翻译

在本研究中,对选取的 4 种 LLM 进行实验,以评估在不同初始翻译下 SSAE 系统的性能.各模型在其特定的任务场景中展示了不同的优势.其中 NLLB、生成 3 种语言对的初始翻译,再进行译后编辑的结果如表 5 所示,其余 3 种模型 mBART、Gemini、DeepSeek-Chat 的结果列在附录 D(http://jxmu.xmu.edu.cn/upload/html/20250605)中.

表 5　NLLB 模型的初始翻译和微调之后应用 SSAE 系统的译后编辑质量

Tab. 5　The post-editing quality of translations generated by the NLLB model after initial translation and fine-tuning with the SSAE system

语言对	模型	BLEU↑/%	ChrF↑/%	TER↓/%	COMET-DA↑	术语成功率↑/%
De-En	NLLB	19.11	50.34	77.69	0.776 6	31.64
	NLLB+SSAE	33.57	64.36	64.02	0.823 1	77.22
Zh-En	NLLB	6.31	28.52	90.16	0.610 6	4.47
	NLLB+SSAE	14.30	40.58	77.00	0.751 6	79.83
En-Cs	NLLB	25.37	53.71	66.87	0.820 7	31.67
	NLLB+SSAE	44.13	68.84	49.18	0.882 3	75.84

实验结果显示,SSAE 系统显著提升了 NLLB 模型在多个语言对(De-En,Zh-En,En-Cs)上的翻译表现. 具体而言,SSAE 使 BLEU、ChrF 和 COMET-DA 分数大幅提升,同时 TER 分数显著下降,表明翻译质量有了明显改进. 此外,术语成功率也有显著提高,尤其是在 Zh-En 语言对中,术语翻译成功率从 4.47%跃升至 79.83%. 整体来看,SSAE 在提高翻译准确性和术语处理方面展现出了卓越的效果.

4.3 GPT 系列模型生成的翻译

为了探索不同的模型特性及其对翻译质量和性能的影响. 在本研究和实验中采用各种翻译配置来全面评估模型在不同翻译任务上的性能. GPT-3.5 Turbo 提供了自然语言生成能力,适用于那些需要灵活性和创新性的翻译任务. 在严格规则下进行翻译,保证精确符合特定规范和标准. GPT-4o mini 是一个高效简化的模型,适用于资源有限的环境,并为我们提供了上佳的翻译性能. 利用两个 GPT 系列的模型生成初始翻译,在 SSAE 系统改善下,结果如表 6 所示.

对于 De-En 的翻译任务,在应用 SSAE 后,GPT-3.5 Turbo 的 BLEU 提升为 13.70 个百分点,ChrF 提升为 12.94 个百分点,而 GPT-4o mini 则分别为 13.10 个百分点和 12.59 个百分点. 术语成功率方面,GPT-3.5 Turbo 经 SSAE 后,提高了 49.72 个百分点,略高于 GPT-4o mini 的 48.70 个百分点. 这一结果表明,译后编辑在增强这两个模型的翻译流畅性和术语翻译准确性方面发挥了明显的作用.

表 6 GPT 系列模型的初始翻译以及微调之后应用 SSAE 系统的译后编辑质量

Tab.6 The initial translation generated by the GPT series models and the post-editing quality after applying the SSAE system

语言对	基线模型/改进模型	BLEU↑/Improve BLEU/%	ChrF↑/Improve ChrF/%	TER↓/Improve TER/%	COMET-DA↑/Improve COMET-DA	术语成功率↑/Improve 术语成功率/%
De-En	GPT-3.5 Turbo/GPT-3.5 Turbo+SSAE	21.20/**34.90**	53.79/**66.73**	77.33/**64.31**	0.797 2/**0.808 7**	38.00/**87.72**
	GPT-4o mini/GPT-4o mini+SSAE	21.77/**34.87**	54.11/**66.70**	76.61/**64.52**	0.800 0/**0.808 6**	38.76/**87.46**
Zh-En	GPT-3.5 Turbo/GPT-3.5 Turbo+SSAE	10.92/**13.63**	34.64/**39.75**	85.94/**79.41**	0.689 6/**0.744 8**	28.00/**78.70**
	GPT-4o mini/GPT-4o mini+SSAE	14.50/**15.98**	41.13/**42.51**	77.73/**76.89**	0.758 2/**0.758 6**	46.44/**79.56**
En-Cs	GPT-3.5 Turbo/GPT-3.5 Turbo+SSAE	28.87/**47.55**	57.70/**71.93**	63.23/**45.46**	0.856 7/**0.887 2**	40.30/**85.11**
	GPT-4o mini/GPT-4o mini+SSAE	32.31/**50.29**	59.04/**73.49**	60.54/**42.87**	0.855 2/**0.890 2**	41.77/**85.95**

注:表中每个指标的右侧为应用 SSAE 系统之后的翻译质量.

在 Zh-En 翻译中,GPT-3.5 Turbo 在应用 SSAE 后虽然显示出较高的 BLEU(2.71 个百分点)和 ChrF(5.11 个百分点)提升,而 GPT-4o mini 的提升分别为 1.48 个百分点和 1.38 个百分点. 但是 GPT-4o mini 的初始翻译质量相比之下要高,这反映出 SSAE 系统在复杂语法结构处理上的优势.

在 En-Cs 翻译任务中,GPT-3.5 Turbo 和 GPT-4o mini 在应用 SSAE 后的 BLEU 提升分别为 18.68 个百分点和 17.98 个百分点. TER 差值的减少表明译后编辑有效减少了翻译错误,提升了译文的准确性.

4.4 商业翻译平台生成的服务

在这一研究中,将微软、百度和谷歌翻译作为初始翻译平台. 这 3 个平台是全球商业翻译服务中最流行的,代表了当前机器翻译技术的普遍水平. 它们采用了不同的翻译技术和方法,为译后编辑的技术验证提供了多维度的视角. 此外,它们支持多语言对和翻译任务,使我们可以在不同语言和场景下评估译后编辑的效果. 研究这些平台的翻译结果可以直接反映实际应用中的翻译质量,验证译后编辑在真实世界中提高翻译准确性和流畅性的可能性. 因此,

这些平台使我们的研究具有较高的实际应用价值和通用性. 实验结果见表 7.

表 7　商业翻译模型初始翻译及应用 SSAE 系统的译后编辑质量

Tab. 7　Quality comparison of the initial translation of the commercial translation model after applying the SSAE system

语言对	模型	BLEU↑/Improve BLEU/%	ChrF↑/Improve ChrF/%	TER↓/Improve TER/%	COMET-DA↑/Improve COMET-DA	术语成功率↑/Improve 术语成功率/%
De-En	百度翻译	21.53/**35.80**	53.12/**66.83**	75.54/63.17	0.788 1/0.808 5	36.50/**86.80**
	微软翻译	22.13/35.54	54.07/66.52	76.53/64.42	0.796 8/0.804 1	38.60/86.48
	谷歌翻译	**22.13**/35.57	**54.07**/66.41	**76.53/63.10**	**0.796 8/0.820 8**	**38.60**/85.67
Zh-En	百度翻译	13.20/15.07	39.44/41.51	79.10/77.50	0.741 4/0.751 2	47.30/81.55
	微软翻译	13.19/15.40	39.36/**41.86**	79.69/77.23	0.730 4/0.753 1	45.70/80.90
	谷歌翻译	**14.81/15.47**	**40.97**/41.76	**77.64/76.95**	**0.753 8/0.753 3**	**49.50/82.14**
En-Cs	百度翻译	Null	Null	Null	Null	Null
	微软翻译	37.14/**52.61**	**62.98/74.62**	**54.67**/40.90	**0.873 9/0.892 6**	**47.10/86.62**
	谷歌翻译	**37.20**/52.49	62.54/74.54	54.74/**40.79**	0.868 7/0.892 6	46.40/86.62

注:商业翻译模型(谷歌/微软/百度)通过 API 调用生成初始翻译,SSAE 系统直接对其输出进行译后编辑,表中每个指标的右侧为应用 SSAE 系统之后的翻译质量,粗体为每个语言对表现最好. Null 表示对于百度翻译中需要企业级应用服务商才能够使用 En-Cs 语言对的翻译,无法测量评估,故标记为 Null.

在 De-En 的翻译中,所有模型应用 SSAE 后在流畅性和准确性上均有显著提升. 百度翻译的 BLEU 和 ChrF 差值最高,分别为 14.27 个百分点和 13.71 个百分点,术语成功率也达到了50.30%. 谷歌翻译和微软翻译在这些指标上表现相近,但谷歌翻译应用 SSAE 在语义质量的提升(COMET-DA 差值为 0.024 0)上略优于其他两者.

在 Zh-En 的翻译中,微软翻译应用 SSAE 在流畅性(BLEU 差值 2.21 个百分点,ChrF 差值 2.50 个百分点)和术语成功率(35.20%)方面表现最佳. 百度翻译紧随其后,显示出适度的改进. 谷歌翻译的表现略逊于其他模型,但术语成功率提升相对较低,为 32.64%.

在 En-Cs 的翻译任务中,微软翻译应用 SSAE 后的 BLEU 差值为 15.47 个百分点,显示出显著的翻译质量提升. 谷歌翻译在术语成功率上表现最佳,为 39.22%,并在语义质量上也有所提升(COMET-DA 差值为 0.023 9).

总体而言,SSAE 译后编辑系统即使应用在最先进的商业翻译上,也显著改善了所有模型的翻译表现,尤其在术语翻译和语义质量方面. 百度翻译在 De-En 翻译中表现突出,微软翻译在 Zh-En 和 En-Cs 翻译中表现优越,而谷歌翻译在某些语义和术语处理方面显示出优势. SSAE 译后编辑系统在不同语言对上的一致性提升,证明了其在提高翻译模型性能中的重要作用.

4.5　与其他 APE 系统进行比较

利用谷歌翻译作为初步翻译,与以下几种主流 APE 系统进行比较,以评估 SSAE 在术语数据集翻译的效果(表 8):

1) GPT-3.5 Turbo(constraint):此系统基于 GPT-3.5 Turbo 进行初始翻译,并加入术语约束进行译后检查.

2) self_correct_mt[39]:此框架基于 LLM 自我提炼,首先使用 LLM 生成翻译,然后进行 MQM 标注. 基于这些错误注释进行译后编辑改善.

3) GPT-4o mini-APE:使用 GPT-4 模型标注微软翻译中的错误,随后进行译后编辑.

4) GPT-4o mini-agent 翻译:此方法使用 LLM 翻译文本,并要求 LLM 对其翻译进行反思,提出建设性建议,据此进行编辑.

5) AdaptTerm-APE[36]:源于 WMT23 参与者,此系统使用术语数据微调通用模型以生成初始翻译,随后用 LLM 修正漏掉的术语.

6) OPUS-CAT-APE[40]:同样源于 WMT23,OPUS-CAT 是一个基于 OPUS 项目的开源计算机辅助翻译工具. OPUS 项目提供的是一个大型开源平行

语料库,其模型帮助翻译人员进行自动和辅助翻译.

7) UEDIN-LLM-APE[41]:此系统是 WMT23 参与者中使用的系统,采用 2-shot 解码和源语言术语提示实施术语约束. 如术语约束无效,系统将使用基于提示的方法识别并调整目标语言中的错误术语,从而引导解码器生成正确的词汇.

表 8 在谷歌翻译的基础上比较其他 APE 系统与 SSAE 系统的译后质量

Tab. 8 Comparing the post-translation quality of other APE systems with the SSAE system based on Google translate

语言对	模型	BLEU↑/%	ChrF↑/%	TER↓/%	COMET-DA↑	术语成功率↑/%
De-En	谷歌翻译	22.13	54.07	76.53	0.796 8	38.60
	self_correct_mt	18.39	50.43	86.95	0.742 7	35.62
	GPT-3.5 Turbo(constraint)	30.77	62.06	68.98	0.794 5	71.47
	GPT-4o mini -APE	17.20	48.46	84.06	0.733 6	32.04
	GPT-4o mini -agent 翻译	14.76	48.57	86.52	0.790 9	33.23
	AdaptTerm-APE	38.11	62.39	**54.40**	0.838 1	67.36
	OPUS-CAT-APE	**38.64**	62.88	60.27	**0.844 6**	60.88
	UEDIN-LLM-APE	35.79	61.34	59.09	0.840 9	67.82
	谷歌翻译+SSAE(Ours)	35.57	**66.41**	63.10	0.820 8	**85.67**
Zh-En	谷歌翻译	14.81	40.97	77.64	**0.753 8**	49.50
	self_correct_mt	13.16	38.70	82.05	0.715 4	46.20
	GPT-3.5 Turbo(constraint)	14.57	41.00	80.96	0.730 5	76.43
	GPT-4o mini -APE	13.09	38.73	80.08	0.725 8	44.90
	GPT-4o mini -agent 翻译	10.90	37.04	82.95	0.740 9	45.03
	AdaptTerm-APE	Error	Error	Error	Error	Error
	OPUS-CAT-APE	9.82	32.08	81.13	0.627 0	37.51
	UEDIN-LLM-APE	Null	Null	Null	Null	Null
	谷歌翻译+SSAE(Ours)	**15.47**	**41.76**	**76.95**	0.753 3	**82.14**
En-Cs	谷歌翻译	37.20	62.54	54.74	0.868 7	46.40
	self_correct_mt	31.28	57.90	65.33	0.811 2	41.42
	GPT-3.5 Turbo(constraint)	44.67	69.35	48.40	0.877 2	76.73
	GPT-4o mini -APE	33.00	60.67	60.20	0.833 7	43.69
	GPT-4o mini -agent 翻译	Null	Null	Null	Null	Null
	AdaptTerm-APE	25.87	55.84	56.88	0.856 5	33.27
	OPUS-CAT-APE	39.39	67.54	42.56	**0.920 3**	51.44
	UEDIN-LLM-APE	27.64	58.22	55.55	0.887 4	32.73
	谷歌翻译+SSAE(Ours)	**52.49**	**76.54**	**40.79**	0.892 6	**86.62**

注:加粗字体表示每个语言对中表现最好的 APE 系统,Null 表示没有获取其系统的表现分数,Error 表示在 WMT 官网获取的参赛者作品经检查有误,因此不做比较.

通过对比不同的译后编辑系统和谷歌翻译的基线结果,可以看到谷歌翻译+SSAE 在所有语言对上的翻译质量都有显著提升. 特别是在术语成功率方面,这种改进尤为突出,说明该方法在提高翻译的准确性和流畅性方面表现优异. 其他系统,如 GPT-3.5 Turbo(constraint)和 OPUS-CAT-APE,虽然在某些指标上表现良好,但在整体性能和术语成功率方面未能超越谷歌翻译+SSAE. 比较我们的系统与 OPUS-CAT,OPUS-CAT 使用适应语境的软约束,使模型能够在集成指定术语的同时保持翻译质量. 这可以使翻

译更符合上下文,尤其是在技术或专业内容中. OPUS-CAT 与 CAT 工具无缝集成,可提高生产力并保持翻译之间的一致性. 这种集成有助于有效地管理翻译,并可以对质量指标产生积极影响.

整体而言,SSAE 系统在提升术语成功率的同时,也显示出了在整体翻译质量上的显著改进. 我们的系统通过将 SSAE 与谷歌翻译的基础模型结合,实现了更高的翻译准确性和一致性. 这种方法不仅提高了翻译质量,还增强了对专业术语的处理能力. 尽管 SSAE 在各项指标上表现优异,但仍存在一些局限性. 例如,当前的方法可能在处理极为专业或高度领域化的术语时仍有一定挑战. 此外,系统的实际应用场景和用户反馈可能会影响翻译质量的稳定性和一致性,这需要进一步的研究和测试.

5　结　论

在本研究中,构建了一个基于 LLM 的术语约束与多维翻译建议的自我反馈系统,可以为翻译提供专家级别的建议和修正. 根据实验结果,SSAE 在评估得分和经济方面都优于当前常用的译后编辑方法. 展现了在处理多维度初始翻译时的系统能力. 对译后编辑在处理专业领域的效果进行了深入探究,并查明了术语约束的重要影响. 更重要的是,分析了我们系统的经济效益(附录 E(http://jxmu.xmu.edu.cn/upload/html/20250605)). 与顶级模型对比,SSAE 系统在经济成本上减少了 99%,不到人工成本比的百分之一. 另值得一提的是,SSAE 系统具备即插即用的特性,可以兼容任何现有的翻译系统,既方便又高效.

6　局限性

在 MT 领域,LLM 在译后编辑中展现了巨大潜力,但也面临挑战. 首先,我们的系统使用了 GPT-4o mini,缺乏对最新开源 LLM 的对比分析,可能限制了对先进模型性能的全面了解. 其次,LLM 的性能依赖于训练数据的质量和范围,在特定领域或专业词汇上可能效果有限,需要深入的术语研究. 此外,LLM 的运行和优化需要大量计算资源,限制了实际应用. LLM 在理解长篇上下文和处理复杂反馈方面也有局限,对非标准化内容的处理不够完善. 尽管我们的系统成功生成了专家级翻译建议,增强了可解释性,但其“黑盒”特性仍影响了对预测能力的评估. 同时,伦理和公平性问题也需高度关注.

因此,虽然 LLM 在翻译领域具有显著优势,但需要在后续研究中解决这些挑战,包括引入更多的开源模型进行对比,以推动技术进步和应用.

参考文献:

[1] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[2] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[3] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[4] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2303.08774.

[5] QIN C, ZHANG A, ZHANG Z, et al. Is ChatGPT a general-purpose natural language processing task solver? [EB/OL]. [2024-12-01]. https://arxiv.org/abs/2302.06476.

[6] GUERREIRO N M, ALVES D M, WALDENDORF J, et al. Hallucinations in large multilingual translation models [J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1500-1517.

[7] ZHONG Q, WANG K, XU Z, et al. Achieving>97% on GSM8K: deeply understanding the problems makes LLMs perfect reasoners[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2404.14963.

[8] LU Q, QIU B, DING L, et al. Error analysis prompting enables human-like translation evaluation in large language models[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2303.13809.

[9] VILAR D, FREITAG M, CHERRY C, et al. Prompting PALM for translation: assessing strategies and performance[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2211.09102.

[10] HENDY A, ABDELREHIM M, SHARAF A, et al. How good are GPT models at machine translation? A comprehensive evaluation [EB/OL]. [2024-12-01]. https://arxiv.org/abs/2302.09210.

[11] JIAO W, WANG W, HUANG J, et al. Is ChatGPT a good translator? Yes with GPT-4 as the engine[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2301.

08745.

[12] ZHU W, LIU H, DONG Q, et al. Multilingual machine translation with large language models: empirical results and analysis[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2304.04675.

[13] KOCMI T, AVRAMIDIS E, BAWDEN R, et al. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet[C]// Proceedings of the Eighth Conference on Machine Translation. Singapore: ACL, 2023: 1-42.

[14] BAWDEN R, YVON F. Investigating the translation performance of a large multilingual language model: the case of BLOOM[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2303.01911.

[15] SIMARD M, GOUTTE C, ISABELLE P. Statistical phrase-based post-editing[C]// Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics. Rochester: ACL, 2007: 508-515.

[16] NEGRI M, TURCHI M, CHATTERJEE R, et al. ESCAPE: a large-scale synthetic corpus for automatic post-editing[EB/OL]. [2024-12-01]. https://arxiv.org/abs/1803.07274.

[17] RIOS M. Instruction-tuned large language models for machine translation in the medical domain[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2408.16440.

[18] ZHENG J, HONG H, LIU F, et al. Fine-tuning large language models for domain-specific machine translation [EB/OL]. [2024-12-01]. https://arxiv.org/abs/2402.15061.

[19] MOSLEM Y, ROMANI G, MOLAEI M, et al. Domain terminology integration into machine translation: Leveraging large language models[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2310.14451.

[20] VASWANI A. Attention is all you need[C]// In Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: NeurIPS, 2017: 5998-6008.

[21] SEMENOV K, ZOUHAR V, KOCMI T, et al. Findings of the WMT 2023 shared task on machine translation with terminologies[C]// Proceedings of the Eighth Conference on Machine Translation. Singapore: ACL, 2023: 663-671.

[22] BOJAR O, CHATTERJEE R, FEDERMANN C, et al. Findings of the 2017 conference on machine translation (WMT17)[C]// Proceedings of the Second Conference on Machine Translation. Copenhagen: ACL, 2017: 169-214.

[23] PAL S, HERBIG N, KRÜGER A, et al. A transformer-based multi-source automatic post-editing system[C]// Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Belgium: ACL, 2018: 827-835.

[24] LOPES A V, FARAJIAN M A, CORREIA G M, et al. Unbabel's submission to the WMT2019 APE shared task: BERT-based encoder-decoder for automatic post-editing[EB/OL]. [2024-12-01]. https://arxiv.org/abs/1905.13068.

[25] GU J, WANG C, ZHAO J. Levenshtein transformer[C]// Advances in Neural Information Processing Systems. Vancouver: NeurIPS, 2019: 32.

[26] CHEN P, GUO Z, HADDOW B, et al. Iterative translation refinement with large language models[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2306.0385.

[27] RAUNAK V, SHARAF A, WANG Y, et al. Leveraging GPT-4 for automatic translation post-editing[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2305.14878.

[28] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[J]. Advances in Neural Information Processing Systems, 2022, 35: 22199-22213.

[29] XU W D, DEUTSCH D, FINKELSTEIN M, et al. LLMRefine: pinpointing and refining large language models *via* fine-grained actionable feedback[C]// Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City: ACL, 2024: 1429-1445.

[30] LIANG P, BOMMASANI R, LEE T, et al. Holistic evaluation of language models[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2211.09110.

[31] POST M. A call for clarity in reporting BLEU scores [EB/OL]. [2024-12-01]. https://arxiv.org/abs/1804.08771.

[32] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proceedings of the 40th annual meeting of the Association for Computational Linguistics. New York: ACL, 2002: 311-318.

[33] SNOVER M, DORR B, SCHWARTZ R, et al. A study of translation edit rate with targeted human annotation [C]// Proceedings of the 7th Conference of the

Association for Machine Translation in the Americas: Technical Papers. New York: AMT, 2006: 223-231.

[34] REI R, STEWART C, FARINHA A C, et al. COMET: a neural framework for MT evaluation[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2009.09025.

[35] KOCMI T, FEDERMANN C, GRUNDKIEWICZ R, et al. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2107.10821.

[36] MOSLEM Y, ROMANI G, MOLAEI M, et al. Domain terminology integration into machine translation: leveraging large language models[C]// Proceedings of the Eighth Conference on Machine Translation. Singapore: ACL, 2023: 902-911.

[37] KOEHN P, KNOWLES R. Six challenges for neural machine translation[EB/OL]. [2024-12-01]. https://arxiv.org/abs/1706.03872.

[38] STAHLBERG F, BYRNE B. On NMT search errors and model errors: cat got your tongue? [C]// In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference. Hong Kong: EMNLP, 2019: 3356-3362.

[39] FENG Z, ZHANG Y, LI H, et al. Improving LLM-based machine translation with systematic self-correction[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2402.16379.

[40] NIEMINEN T. OPUS-CAT terminology systems for the WMT23 terminology shared task[C]// Proceedings of the Eighth Conference on Machine Translation. Singapore: ACL, 2023: 912-918.

[41] BOGOYCHEV N, CHEN P. Terminology-aware translation with constrained decoding and large language model prompting[EB/OL]. [2024-12-01]. https://arxiv.org/abs/2310.05824.

(责任编辑:汪　军)