

基于源端易错词的大语言模型机器翻译优化方法研究

李扬韬⁺, 王康, 杜天译

苏州大学 计算机科学与技术学院, 苏州 215006

+ 通信作者 E-mail: 2524278328@qq.com

摘要:

机器翻译在大语言模型 (LLMs) 的推动下取得了显著进展, 但特定源端词汇仍与低质量翻译显著相关。该文提出一种基于源端易错词的翻译优化方法。首先, 通过大规模语料统计识别出显著影响翻译质量的易错词, 并发现这些词汇并非直接导致翻译错误, 而是复杂翻译挑战的指示器。在此基础上, 设计了结合 BGE 语义检索与无监督质量评估的检索增强生成 (RAG) 策略, 用于优化上下文学习示例选择, 在中英与阿英翻译任务中分别带来 0.76 和 1.53 的 COMET 提升。同时, 构建针对易错词的高质量训练数据集, 并采用 LoRA 微调基线模型, 使中英和阿英的 COMET 得分分别提升 1.85 和 5.78。结果表明, 该方法能够显著提高翻译质量, 为提升大语言模型机器翻译的稳定性与准确性提供了一种有效途径。

关键词: 易错词; 上下文学习; LoRA 微调

文献标志码: A **中图分类号:** TP391

Research on Large Language Model Machine Translation Optimization Method Based on Source-side Error-prone Words

LI Yangtao¹⁺, WANG Kang², DU Tianyi³

1. School of Computer Science & Technology, Soochow University, Suzhou 215006, China

2. School of Computer Science & Technology, Soochow University, Suzhou 215006, China

3. School of Computer Science & Technology, Soochow University, Suzhou 215006, China

Abstract: Large Language Models (LLMs) have greatly advanced machine translation, yet certain source words remain highly associated with low-quality outputs. This study presents a translation optimization approach centered on source-side error-prone words. Using large-scale corpus statistics, we identify words that strongly influence translation quality and find that they act as indicators of complex translation challenges rather than direct error sources. We then design a Retrieval-Augmented Generation (RAG) strategy integrating BGE-based semantic retrieval and unsupervised quality assessment to optimize in-context example selection, yielding COMET gains of 0.76 and 1.53 for Chinese - English and Arabic - English translation, respectively. Furthermore, we build high-quality datasets focused on these error-prone words and apply LoRA fine-tuning to baseline models, achieving additional COMET improvements of 1.85 and 5.78. The results confirm that the proposed method effectively enhances translation stability and accuracy in LLM-based machine translation.

Key words: Error-prone Words; In-Context Learning; LoRA Fine-tuning

基金项目: 国家自然科学基金 (62276179, 62261160648); 江苏高校优势学科建设工程资助项目

作者简介: 李扬韬(2001), 男, 硕士研究生, CCF 学生会员, 研究方向为机器翻译、大模型; 王康(2002), 男, 硕士研究生, 研究方向为大模型、自然语言处理; 杜天译(2001), 男, 硕士研究生, 研究方向为大模型、自然语言处理。

收稿日期: 2025-00-00 **修回日期:** 2025-00-00

近年来,大语言模型(LLMs)凭借卓越的语言理解与生成能力,推动了机器翻译技术的革新。与依赖平行语料的神经机器翻译(NMT)不同,LLMs通过预训练与上下文学习(In-Context Learning, ICL)能够在少样本甚至零样本条件下实现高质量翻译,显著提升流畅性与泛化能力,其性能已接近或超越传统NMT与商用系统水平^[1]。

虽然大模型机器翻译系统的经验分布较为可靠,可以在大多数情况下保持准确,但并不能保障其在所有源端词汇上都能准确翻译,术语的文化特异性、隐喻表达和复杂句式常导致翻译质量下降,从而导致一些低质量翻译。例如Llama在翻译“重实体,轻程序”时,直接直译“轻”,“重”两词——“Heavy substance and light procedure”。识别此类易错词具有重要意义:一方面,它们往往预示着潜在的翻译难点,是低质量译文的重要诱因;另一方面,针对这些词进行优化能够显著提升模型的稳定性与整体性能。然而,依赖人工识别代价高昂且难以扩展,因此需要自动化的识别与优化方法。

现阶段优化翻译质量往往从两个角度出发:一是在推理阶段,利用大语言模型的上下文学习能力^[2,3]学习提示里隐藏的模式,进行翻译。有研究显示,大语言模型的性能与任务描述、示例选择、示例顺序、提示模板格式等因素相关^[4-6],一些研究者通过选择高质量示例提高大语言模型翻译能力^[7,8]。二是在训练阶段,通过平行语料对大模型进行微调,如ALMA^[9]。它对单语数据进行微调开始,然后使用高质量的并行数据进一步优化。ALMA-R^[1]在此基础上,使用对比偏好优化(CPO)进一步进行LoRA微调。尽管已有研究尝试通过微调(Fine-Tuning)和提示工程(Prompt Engineering)改善翻译质量,但针对特定词汇或结构引发低质量翻译的系统性分析仍较为缺乏。因此,如何针对源端易错词开展有效的检测与优化,仍是一个有待解决的研究问题。

基于上述观察,本文提出了一种以源端易错词为核心线索的大模型机器翻译优化方法,从推理阶段与训练阶段两个层面协同提升翻译质量。具体而言,在推理阶段,我们利用易错词作为翻译难点的指示信号,引导基于检索增强生成(Retrieval-Augmented Generation, RAG)的示例选择策略,从而为模型提供更具针对性的上下文示例;在训练阶段,我们基于易错词构建靶向的高质量平行语料,并在相同数据规模与训练配置下对模型进行LoRA微调,以强化模型对复杂翻译模式的

学习能力。需要强调的是,本文并不将易错词视为词级纠错对象,而是将其作为复杂句法与语义现象的分布指示器,通过数据与示例层面的引导实现性能提升。

本文的主要贡献总结如下:

1) 通过大规模统计分析,构建了与低质量翻译显著相关的源端易错词集合,并系统分析其分布特性。

2) 提出一种基于易错词的推理阶段优化策略,利用RAG技术改进ICL示例选择,从而提升复杂句子的翻译质量。

3) 提出一种易错词驱动的训练数据构建方法,在相同数据规模条件下,通过LoRA微调显著提升模型在中英与阿英翻译任务上的整体性能。

1 相关工作

1.1 大模型机器翻译与提示工程

近年来,大语言模型(LLMs)在自然语言处理领域取得了突破性进展,并在机器翻译(MT)任务中展现出强大的潜力^[1,10]。与传统的基于Transformer^[11]的模型相比,拥有更丰富的语言知识和更强的泛化能力。它们能够执行零样本(Zero-shot)或少样本(Few-shot)翻译,即在没有任何或仅有少量平行语料微调的情况下生成高质量译文^[12]。这主要归功于LLMs强大的上下文理解、生成能力以及从预训练中习得的世界知识^[13]。

为了在不重新训练模型的情况下提升LLM的翻译性能,研究者们广泛探索了推理阶段的优化方法,其中上下文学习^[2]和提示工程是主要手段^[14]。ICL利用LLM理解和模仿示例的能力,通过在提示中提供少量高质量的翻译范例来引导模型生成更好的翻译^[12]。然而,ICL的效果高度依赖于提示的设计,如任务指令、示例选择与顺序、提示模板等因素^[4-6]。近期研究致力于自动化和优化示例选择过程。例如,一些工作利用语义相似度检索相关示例^[14]。

1.2 训练阶段优化:微调与数据策略

除了推理时优化,通过微调使LLM适应特定任务或领域也是提升性能的有效途径。全参数微调虽然效果显著,但计算成本高昂。因此,参数高效微调(Parameter-Efficient Fine-Tuning, PEFT)技术,特别是LoRA (Low-Rank Adaptation),在LLM微调中得到广泛应用。LoRA通过引入少量可训练的低秩矩阵,显著降低了微调的计算和存储需求,同时能达到接近全参数微调的效果^[15]。

在机器翻译领域,ALMA^[9]及其改进版本ALMA-R^[1]是利用LoRA微调LLM进行翻译的代

表性工作。它们证明了结合高质量平行数据进行 LoRA 微调能够显著提升 LLM 的翻译水平。ALMA-R 进一步引入对比偏好优化（如 CPO），利用人类反馈或模型偏好数据进行微调，以更好地对齐人类期望的翻译质量。本文的工作对比了基于源端易错词的采样数据和 ALMA 高质量平行数据在同等数据规模和实验设置下的微调翻译效果。

微调的效果很大程度上取决于训练数据的质量和相关性。已有研究表明，使用精心筛选或生成的、与目标任务/领域高度相关的语料进行微调，比使用通用或随机采样的数据更有效^[16, 17]。

1.3 特定词汇及其对翻译质量的影响

在传统的神机翻译（NMT）中，术语、隐喻表达以及文化特异性常常导致模型在部分词汇上的翻译错误。近年来，一些工作开始关注输入中局部敏感因素对输出质量的影响，但系统地分析源端“易错词”——对大模型翻译性能影响的研究仍较少。

已有研究表明，专业术语的准确翻译对于确保译文在领域内的准确性至关重要。例如，Ghazvininejad 等人的研究提出了基于词典提示的翻译方法，通过在提示中注入专业术语映射，有效改善了低频和专业词汇的翻译表现^[18]。Bogoychev

等人探讨了通过受限解码和大模型提示来实现术语约束，从而确保专业术语在译文中的正确呈现^[19]；而 Moslem 等人则研究了利用大模型生成合成的双语术语数据，并在后处理阶段整合预审批术语，显著提高了领域内翻译中术语的覆盖率和准确性^[20]。这些方法表明，通过精细化处理专业术语及生词，可以显著提升翻译系统（包括大模型）的表现。

我们的工作与上述工作不同的地方在于——我们发现了一些高频源端词汇也与低质量翻译显著相关，而不仅仅是低频词或术语。通过统计低质量翻译样例发现，含有特定易错词的句子在 COMET 评分上明显低于平均水平，这为后续基于易错词的干预策略提供了依据。

2 方法

本节详细介绍基于源端易错词的大语言模型机翻译优化方法。该方法包含三个核心组件：易错词识别与分析、基于 RAG 的上下文学习优化、以及针对性训练数据构建与 LoRA 微调。方法架构图见图 1

基于源端易错词的翻译优化方法

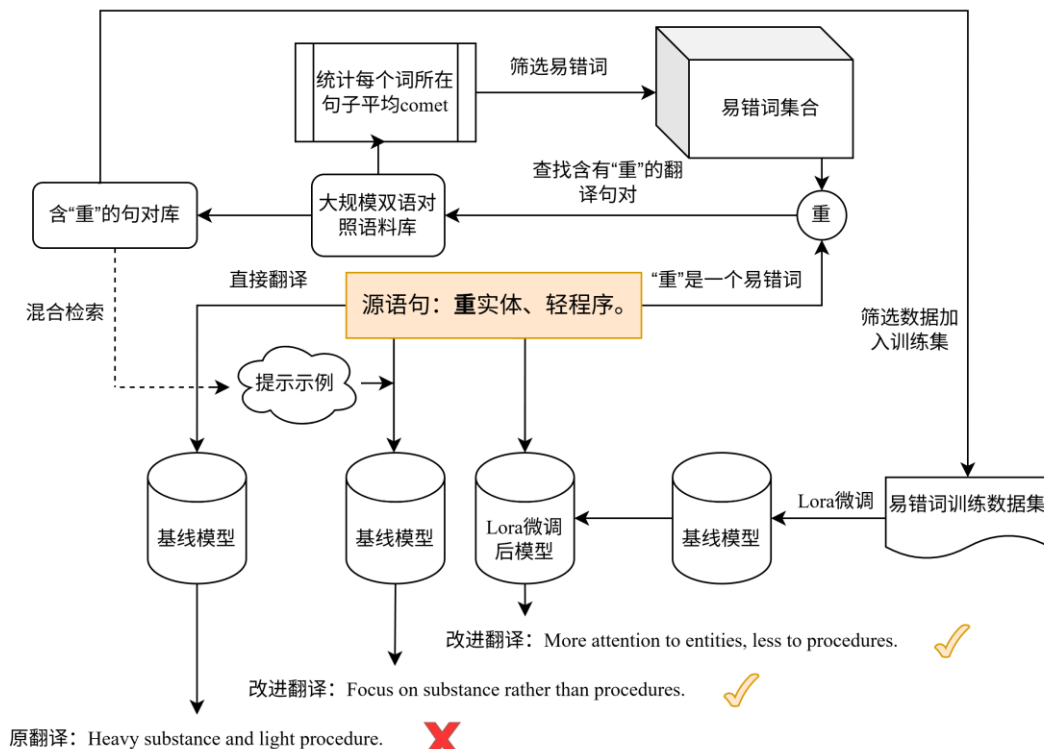


图1 方法架构图

Fig.1 Overall Architecture

2.1 易错词识别与分析

2.1.1 易错词定义

本文所称“易错词”并非指模型在翻译时必然出错的词汇,而是指那些在源端出现时显著提升句子整体翻译错误概率的词。换言之,它们更像是潜在翻译难点的“指示信号”,而非错误本身的直接来源。这些词汇的存在往往导致翻译系统产生低质量输出,其特征包括:

- 包含该词汇的句子平均翻译质量显著低于整体平均水平
- 该词汇在语料库中具有一定的出现频次
- 该词汇对翻译质量的负面影响具有统计显著性

2.1.2 易错词评分机制

为了量化易错词的影响程度,定义易错词评分公式:

$$\text{score}(w) = (\mu_{\text{global}} - \mu_w) \times \text{count}_w \quad (1)$$

算法 1 易错词识别算法

输入: 大规模双语语料库 $D = \{(s_i, t_i)\}_{i=1}^N$, 翻译模型 M ,

分词器 T

输出: 易错词集合 W_{bad}

```

1: 初始化词汇统计字典 word_stats = {}
2: for 每个句对  $(s_i, t_i) \in D$  do
3:    $\hat{t}_i = M(s_i)$  //生成翻译
4:    $\text{score}_i = \text{COMET}(\hat{t}_i, t_i, s_i)$  //计算 COMET 分数
5:   words =  $T(s_i)$  //对源句分词
6:   for 每个词  $w \in \text{words}$  do
7:     更新 word_stats[w] 的分数列表和出现次数
8:   end for
9: end for
10: 计算全局平均 COMET 分数  $\mu_{\text{global}}$ 
11: for 每个词  $w$  在 word_stats 中 do
12:   计算  $w$  的平均分数  $\mu_w$  和出现次数  $\text{count}_w$ 
13:   if  $\text{count}_w \geq \theta_{\text{count}}$  AND  $\mu_w < \mu_{\text{global}} - \theta_{\text{diff}}$  then
14:      $W_{\text{bad}} = W_{\text{bad}} \cup \{w\}$ 
15:   end if
16: end for
17: return  $W_{\text{bad}}$ 

```

2.2 易错词机理探究

2.2.1 易错词因果判断

在筛选出易错词之后,我们希望弄明白低质量翻译是否是易错词导致的。换言之,一个含有易错词的句子翻译质量差,是因为这个易错词本身被直接翻译错了,还是该词的存在预示着句子中存在其他更深层次的翻译挑战?为了解答这个问题,我们设计了一套基于对齐和错误定位的因果判断框架。

其中, μ_{global} 为全局平均 COMET 分数, μ_w 为包含词汇 w 的句子平均 COMET 分数, count_w 为词汇 w 的出现次数。该评分综合考虑了质量差异和出现频次,分数越高表示该词汇对翻译质量的负面影响越大。

2.1.3 易错词识别算法

易错词识别过程如算法 1 所示:

其中, θ_{count} 为最小出现次数阈值, θ_{diff} 为质量差异阈值。我们对不同阈值组合 ($\theta_{\text{count}} \in \{20, 50, 100\}$, $\theta_{\text{diff}} \in \{0.03, 0.05, 0.07\}$) 进行了对比分析,以考察参数对易错词识别结果的影响。实验发现,随着 θ_{diff} 的增大,筛选出的易错词数量逐渐减少,但其与低质量翻译样本的相关性显著增强。

为兼顾易错词的覆盖率与其与翻译质量之间的相关性,我们最终选择 $\theta_{\text{count}}=50$ 、 $\theta_{\text{diff}}=0.05$ 作为平衡方案。在该参数组合下,易错词集与整体翻译质量的皮尔逊相关系数达到-0.26,表现出较好的相关性 with 覆盖率。

考虑到分析成本与代表性,我们聚焦于其中对翻译质量负面影响最显著的子集,即“严重易错词” ($\text{score}(w) < -50$),用于后续的分析。对于每一个严重易错词,提取 30 条包含该词的翻译句对样本。

我们的判断方法结合了词对齐工具 awesome-align 和翻译质量评估模型 XCOMET,具体流程如下:

1. 对齐短语提取:对于每个样本句对中的源端易错词,我们使用 awesome-align 分别在机器翻译

(MT) 和参考翻译 (Reference) 中寻找其对应的目标端短语, 记为 qt 和 qr 。

2. 错误定位: 同时, 我们利用 XCOMET 对机器翻译进行错误分析, 它能够精准定位译文中的错误片段, 并输出其具体位置和内容, 我们记为 wi 。

3. 归因逻辑判断: 根据前两步提取的信息, 我们设计了以下规则进行判断:

• **错翻**: 如果易错词在机器翻译中有对应的短语 (即 qt 不为空), 我们计算 qt 与模型检出的错误短语 wi 之间的文本重合度。若存在任何一个 wi 与 qt 高度重合, 我们则判定该易错词在此句中被错翻。

• **漏翻**: 如果易错词在机器翻译中没有对应的短语 (qt 为空), 我们会进一步检查其在参考翻译中的对应 qr 。若参考翻译中存在对应短语 (qr 不为空), 这表明该词在当前语境下是需要翻译的, 因此我们判定其被漏翻。

通过上述流程, 我们可以为每个严重易错词统计出其在样本中的错翻率和漏翻率, 从而定量地判断它多大程度上是低质量翻译的直接原因。

重合度计算为了量化对齐短语与错误片段的相似程度, 设两段待比较文本经统一小写与必要规范化后分别为 s 与 t 。记

$$B = \{(i_k, j_k, \ell_k)\}_{k=1}^K \quad (2)$$

为按字符级获得的匹配块集合, 其中 ℓ_k 为第 k 个匹配块的长度。定义匹配字符总数

$$M = \sum_{k=1}^K \ell_k \quad (3)$$

则两段文本的重合度 (相似度) 定义为

$$\text{overlap}(x, t) = \frac{2M}{|s| + |t|} \in [0, 1] \quad (4)$$

$|s|$ 为 s 的长度, $|t|$ 为 t 的长度。在实验中, 当 $\text{overlap}(s, t) \geq \tau$ (如 $\tau=0.5$) 时, 判定两段文本高度重合。

实验结果我们在阿拉伯语 → 英语 (Ar → En) 和汉语 → 英语 (Zh → En) 两个翻译方向上, 针对 Llama2 和 Llama3 两个模型进行了实验。表 1 汇总了错翻率或漏翻率超过 30% 的严重易错词数量及其在各自集合中的占比; 各语向与模型的错翻/漏翻率分布见图 2。

表 1 易错词因果判断统计结果

Table 1 Statistical results of error-prone words and causal judgment

语向	模型	错翻率 > 30%	漏翻率 > 30%
Ar → En	Llama2	8.4%	10.0%
	Llama3	2.0%	2.4%
Zh → En	Llama2	18.4%	2.5%
	Llama3	12.6%	2.2%

易错词错译与漏译率分布
(基于 awesome-align 词对齐)

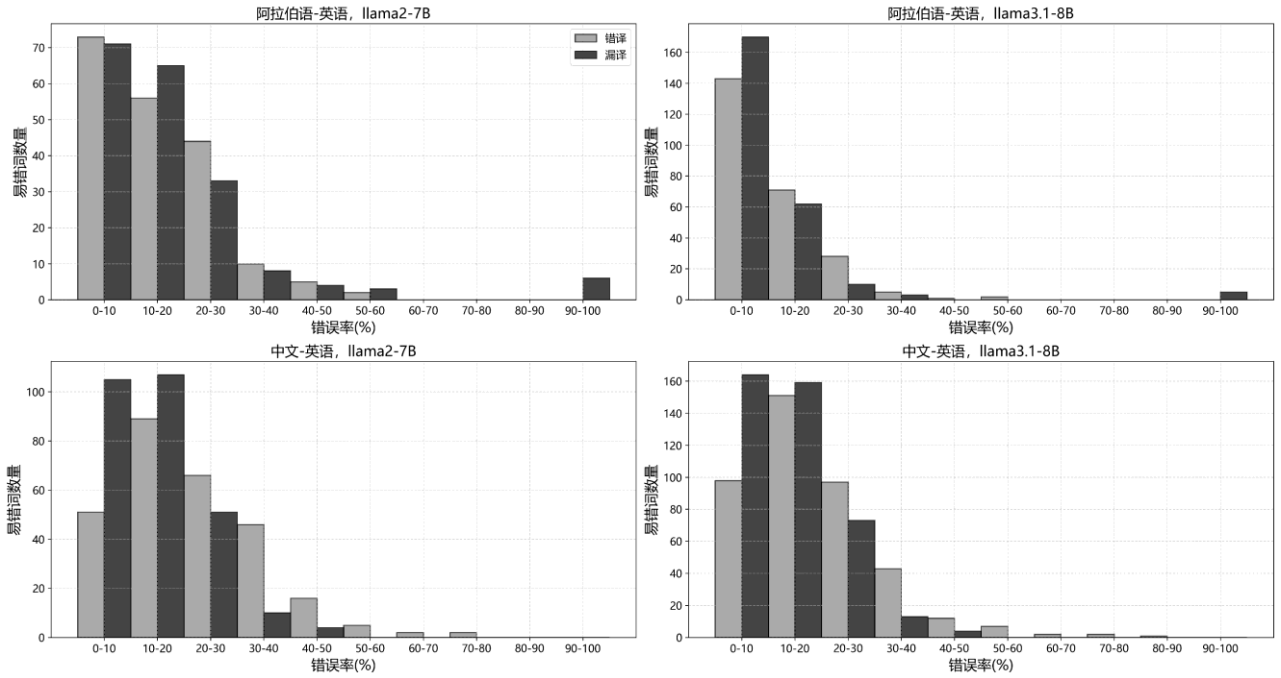


图2 易错词错翻率和漏翻率分布

Fig 2 Distribution of error rate and omission rate of easily misunderstood words

从表1可见,在所有设置下,绝大多数严重易错词的直接错翻率与漏翻率均低于30%。这些结果表明:易错词虽然与低质量翻译有显著相关性,但直接由该词引起的错翻/漏翻占比总体有限,更可能作为提醒句子存在更复杂翻译挑战的“指标”。

2.2.2 句法复杂度与语义歧义分析

为进一步揭示易错词背后的语言学机制,我们从句法复杂度与语义歧义度两个维度对含易错词句子进行了统计分析。前文的实验表明,易错词往往并非翻译错误的直接来源,而可能是复杂翻译挑战的“指示器”。为验证“指示器”假设,本节分

别量化比较含易错词与非易错词句子的句法结构复杂度,以及词汇层面的语义多义性特征。

(1) 句法复杂度分析

我们从中英(Zh→En)与阿英(Ar→En)两个语向中,随机抽取等量的含易错词句子与不含易错词句子(各2000条),利用依存句法分析器计算四项指标:平均依存深度(Avg Dependency Depth)、平均句长(Mean Sentence Length)、平均依存弧跨度(Avg Dependency Distance)、以及从句比例(Subordinate Clause Ratio)。结果如图3所示。

易错词句法复杂度四指标分组柱状图 (zh-en, ar-en)

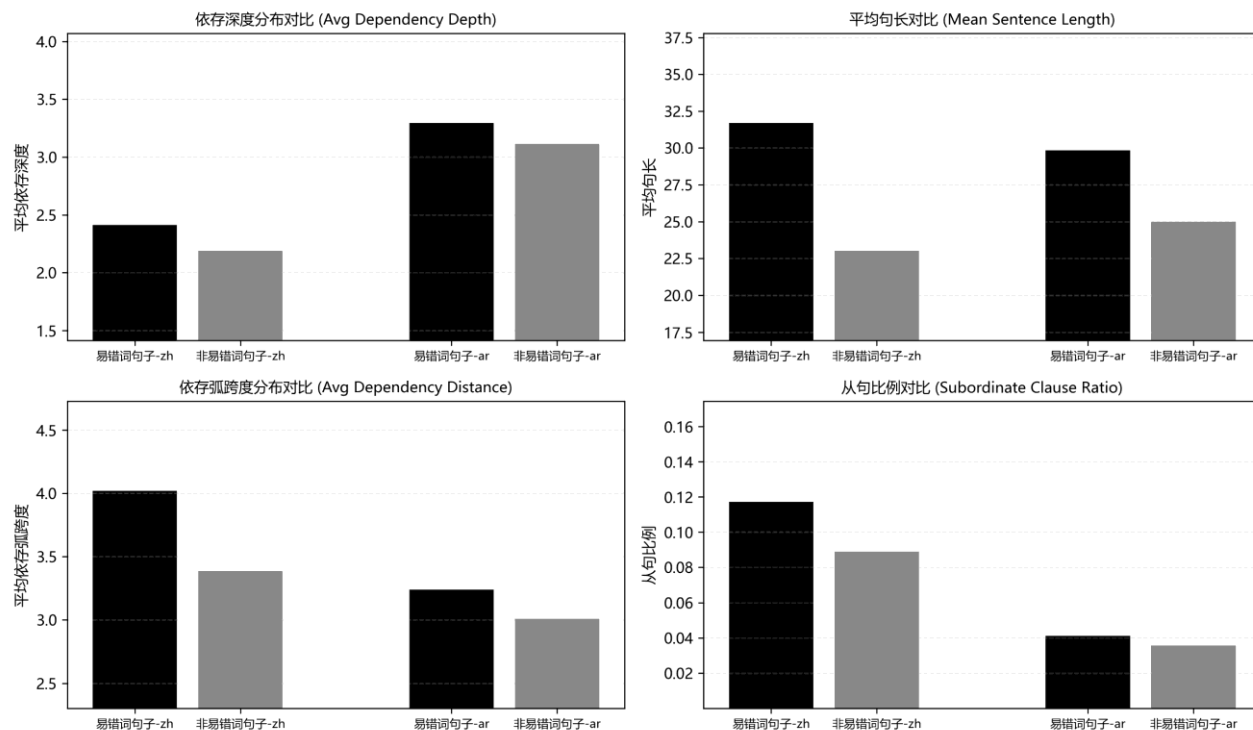


图3 易错词句法复杂度四指标分组柱状图

Fig 3 Bar Chart of Four Indicators of Syntactic Complexity for Error-prone Words

从图中可以看出,四项指标在两种语言对上均呈现一致趋势:

含易错词的句子在依存深度上普遍更高 (Zh→En: 2.41 vs 2.19; Ar→En: 3.29 vs 3.11), 表明这些句子结构嵌套更复杂;

句长也显著增加 (Zh→En: 31.7 vs 23.0; Ar→En: 29.8 vs 25.0), 反映出信息密度更高;

依存弧跨度更大, 说明句法依赖跨距更远, 模型在捕捉长距离关系时更易出错;

从句比例整体提升 (如中译英中 0.12 vs 0.09), 表明易错词句子更可能包含多个从属结构。

这些结果说明, 易错词倾向于出现在句法层级更深、结构更复杂的句子中。换言之, 易错词并非随机分布, 而是句法复杂度的显著信号。这验证了前文的假设——易错词为复杂翻译挑战指示器。

(2) 易错词语义歧义分析

在语义层面, 我们基于词库查询的方式利用 OpenHowNet (用于中文) 与 WordNet (用于阿拉伯文) 对易错词及对应的非易错词进行了义项数统计。非易错词为从同语料中随机抽取频次相近但未标记为易错的词。具体而言, 我们通过词库检索各词条的义项列表, 以义项数量作为语义多义性的度量指标。结果如表 2 所示。

表2 易错词与非易错词义项数量统计

Table 2 Statistical Analysis of Error-Prone and Non-Error-Prone Lexical Items

语向	模型	易错词义项数	非易错词义项数
Ar → En	Llama2	1.24	1.15
	Llama3	1.28	1.19
Zh → En	Llama2	2.72	2.38
	Llama3	3.07	2.34

可以看到, 易错词的平均义项数量在所有语向和模型中均高于非易错词, 其中中译英方向差异尤为显著。这表明易错词往往具有更高的语义歧义度和上下文依赖性。当模型在上下文建模或语义消歧能力不足时, 这类词更容易引发翻译偏差。

2.2.3 易错词词性分析

前两节的结果表明, 大多数易错词并非翻译错误的直接源头, 而是反映句子中深层翻译挑战的“指示器”。为进一步探究这些指示器的语言学共性, 我们对严重易错词开展词性分析, 尝试揭示不同词类与翻译错误模式之间的潜在关联。

词性分布不仅有助于识别哪些词类更易成为翻译“陷阱”, 也能为优化策略提供针对性参考。例如, 若语气词与助词占比高, 说明模型在文体风格与语篇流畅度处理上存在不足; 若多义动词或名

词占比高,则反映其语义消歧能力受限。

为实现自动化标注,我们使用 GLM-4.5 模型,并设计了两层词性分类体系:

第一层(优先类别):包括语气词、术语、符号、非源端词、多义词等任务相关类别,用于捕捉与翻译困难直接相关的语言现象;

第二层(标准类别):当词不属于任何优先类别时,再赋予常规词性标签(如名词、动词、形容词等)

通过提示工程引导 GLM-4.5 严格遵循上述两层逻辑,我们获得了结构化的严重易错词词性标签。图 4 展示了其总体分布情况。

易错词词性分布

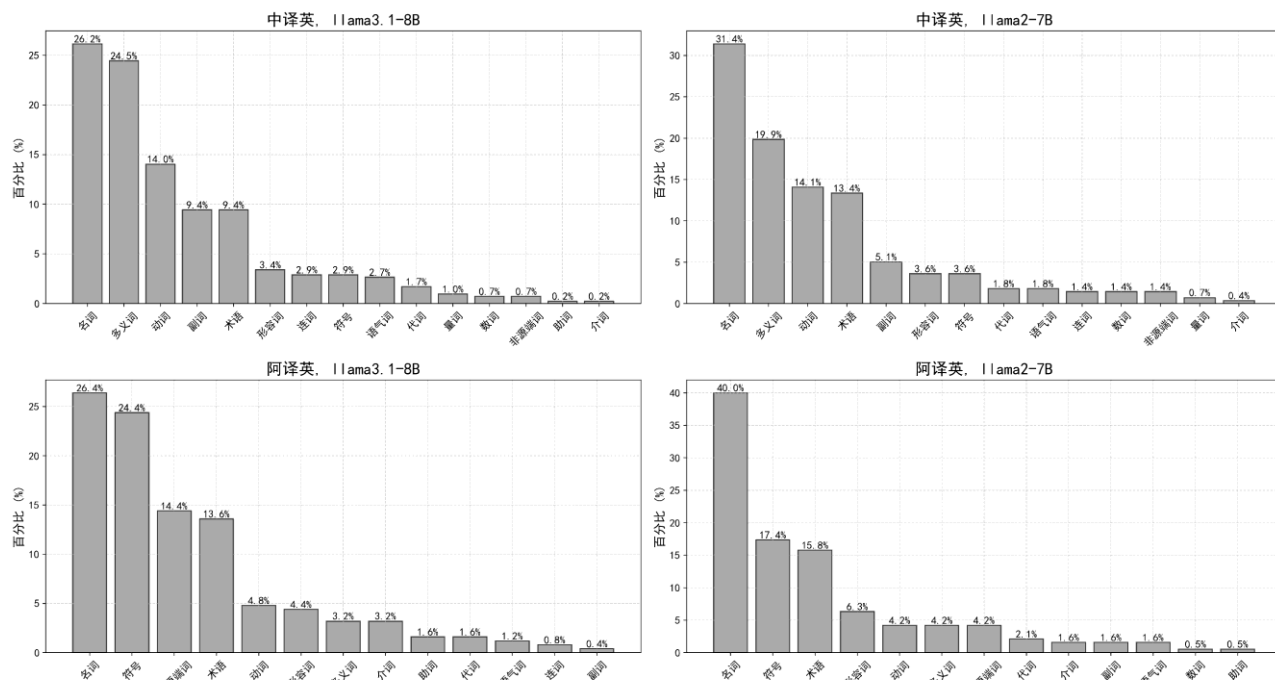


图 4 易错词词性分布

Fig 4 Distribution of error-prone words by part of speech

如图 4 所示,对严重易错词的词性分布进行分析后,我们观察到几个显著的现象。首先,在所有语向和模型中,名词占比最高。然而,这一现象未必意味着名词本身更易被错误翻译。考虑到名词在自然语言中频率最高,其高占比更可能反映出整体词频分布的影响,而非特定翻译机制的缺陷。其次,不同语向展现出独特的错误模式:在中译英任务中,多义词是仅次于名词的第二大错误来源,这凸显了词义消歧是该语向的核心瓶颈;而在阿译英任务中,符号与非源端词的错误占比较高,后者的占比较高,反映了模型在处理源文本中的语码转换或外来词夹杂现象时的鲁棒性不足。

2.2.4 易错词样本分析

虽然因果分析表明多数易错词并非直接错误源头,而是复杂翻译挑战的“指示器”,但这些指示器具体预示着哪些类型的翻译问题,仍需系统性的实证分析。为此,我们设计了基于大规模样本的错误类型分布分析,旨在建立“特定词性易错词”与“特定翻译错误模式”之间的量化关联,从而为针对性优化策略提供精细化指导。

分析方法针对中英翻译方向上 Llama2-7B 和

Llama3.1-8B 两个模型,我们对所有已识别的严重易错词进行了系统性的错误类型分析。具体流程如下:

1. 样本提取:对于每个严重易错词,我们从相应的翻译语料库中提取最多 30 条包含该词的句对样本,确保分析的代表性和统计可靠性。

2. 错误分类标注:采用 GLM-4.5 模型作为自动标注器,参照多维度质量评估 (Multidimensional Quality Metrics, MQM) 错误分类体系^[21],对每个句对中的机器翻译进行错误分析。模型被要求识别并分类句子中最严重的翻译错误,包括但不限于:

- 准确性错误:误译、漏译、增译、未翻译
- 流畅性错误:语法错误、词序错误、字符编码

- 术语错误:术语误译
- 风格错误:语域不当
- 本地化错误:日期/时间/数字/货币格式

3. 分布统计:基于标注结果,我们计算每个易错词在其所有样本中各错误类型的比例分布,形成该词的“错误类型画像”。

4. 聚合分析:将个体易错词按词性分类进行聚合,计算不同词性类别与各错误类型之间的关联强

度,采用对数优势比(logOR)作为量化指标。

关联强度分析为揭示词性与错误类型之间的系统性关联模式,我们构建了基于 2×2 列联表的统计分析框架。对于给定词性 c 和错误类型 e ,设:

a : 该词性出现该错误类型的加权频次 b : 该词性出现其他错误类型的加权频次 c' : 其他词性出现该错误类型的加权频次,对数优势比计算为:

$$\log OR(c,e) = \ln \left(\frac{(a+0.5)(d+0.5)}{(b+0.5)(c'+0.5)} \right) \quad (5)$$

其中 $\log OR > 0$ 表示正关联, $\log OR < 0$ 表示负关联, $|\log OR|$ 的大小反映关联强度。

实验结果通过对413个严重易错词(Llama3.1-8B)和277个严重易错词(Llama2-7B)的全面分析,我们获得了词性-错误类型关联强度热力图。图5、图6显示了若干显著的关联模式:

词性-错误类型关联强度热力图 (LOG_OR) (*表示显著性)

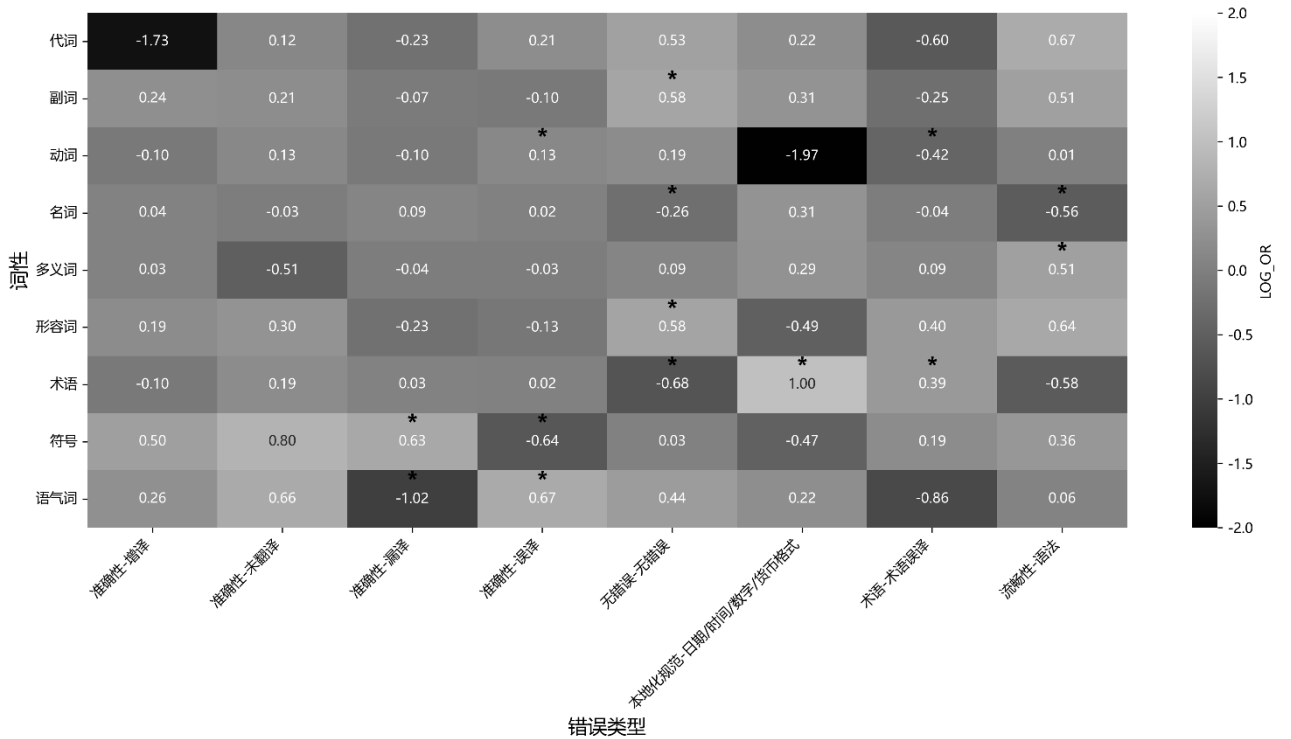


图5 Zh→En, Llama2 词性-错误类型关联强度热力图

Fig 5 Zh→En, Llama2 Heat map of association strength between part-of-speech and error type

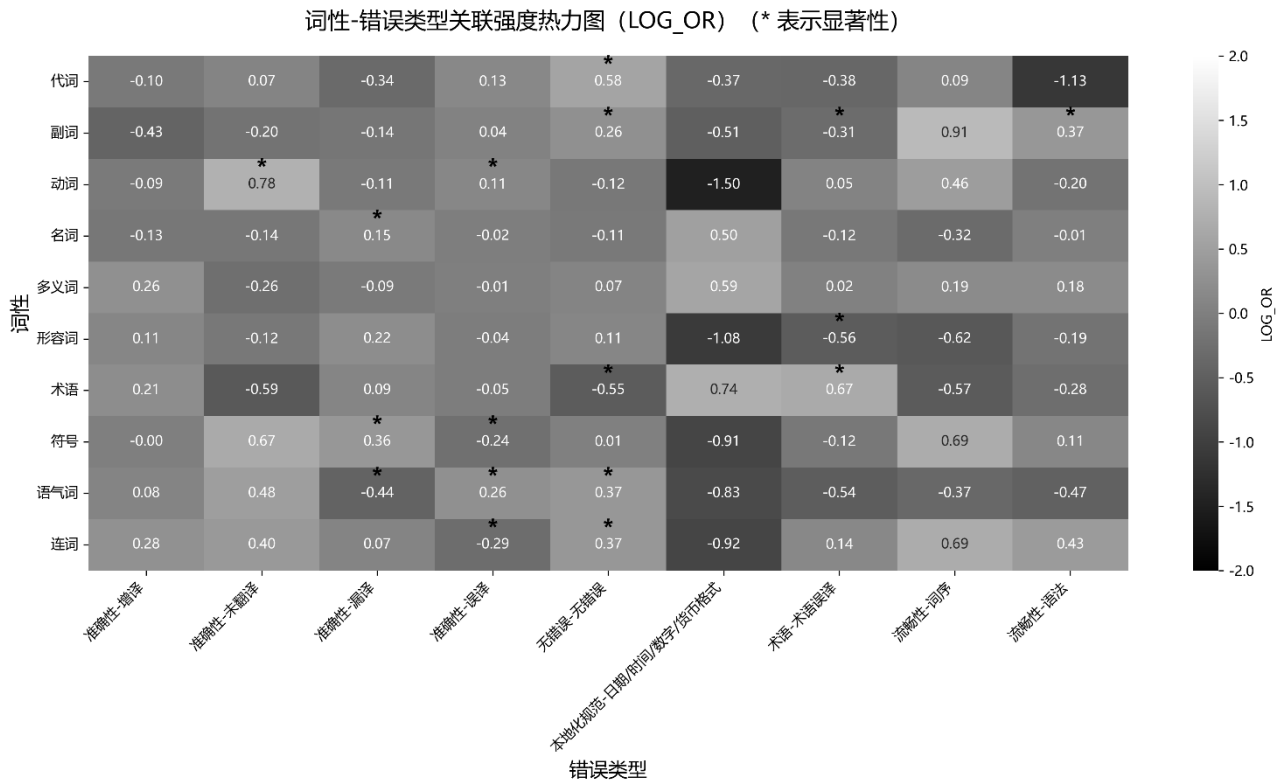


图6 Zh→En, Llama3 词性-错误类型关联强度热力图

Fig 6 Zh→En, Llama3 Heat map of association strength between part-of-speech and error type

•术语类词汇：与术语误译呈显著正关联，同时与“无错误”呈显著负关联，确认了术语处理是模型的系统性短板。

•符号类词汇：易出现漏译或整句未翻译，但较少出现误译，且在两个模型中表现出相同的倾向。反映出模型对符号的“要么不译，要么正确”的处理模式。

•动词类词汇：两模型均显示动词与误译存在一定正关联，说明一些动词的语义理解是两个模型共同的挑战。但 Llama3.1 暴露出一个更严重且独特的倾向：动词易被“未翻译”，其关联强度远超其他错误类型，这在 Llama2 中并非显著模式。这表明 Llama3 在处理某些动词时可能存在触发机制上的缺陷，导致其直接放弃翻译。

从词性分布来看，动词类易错词相比名词等词类更容易触发完全未翻译，这一现象可能与动词在句法结构与语义建模中的核心作用有关。作为句子的谓词中心，动词决定了论元结构、语义关系及整体生成路径，一旦模型在动词层面的跨语言映射或语义解析失败，往往会导致句法规划阶段的不稳定，从而使模型偏离翻译任务本身，转而生成与源句相同的文本。相比之下，名词或术语的错误通常局限于局部词汇层面，较少引发整体翻译脱靶。对于以生成式目标函数训练的大语言模型而言，这种“谓词级失败”更可能被放大为任务级失败，从而表现为完全未翻译的输出。上述分析为本文在实验中观察到的动词易错词与脱靶翻译之间的关联提供了

一种合理解释，但其内在机制仍有待进一步研究。

•语气词类词汇：两模型均显示语气词与误译显著正关联。进一步的样本分析揭示，许多包含语气词的句子也包含了成语（如“时不我待”）或俗语（如“将你的军”、“插不上手”）。这些非字面表达是导致误译的真正原因，而语气词仅仅是作为这些复杂语言现象的常见组成部分，从而在统计上呈现出与误译的强关联性。

值得注意的是，占比最高的名词类和多义词类与最常见的误译错误之间的关联强度接近零（ $\log OR \approx 0$ ），说明虽然出现次数多，但两者无显著的关联。

结论与启示基于样本的错误类型分析进一步证实了易错词作为翻译质量“指示器”的作用机理。不同词性类别的易错词系统地预示着特定类型的翻译挑战：术语类预示术语处理困难，符号类预示翻译不完整，动词类预示误译或脱靶翻译，语气词类预示复杂语言现象的翻译难点。这一发现为构建词性感知的翻译优化策略提供了实证基础，使我们能够针对不同类型的易错词采用相应的干预措施：术语约束、符号识别增强、句法结构训练、风格适应调优等。同时，模型间的差异化错误模式也为个性化优化路径的制定提供了指导。

2.2.5 小结

综上所述，易错词在翻译错误中多表现为“复杂翻译环境的指示器”而非直接错误源。基于这一认识，我们设计了两种相对轻量的优化策略：

(1) 在提示中引入含易错词的代表性示例, 引导模型关注相关结构;

(2) 利用包含易错词的句对进行参数高效微调 (LoRA), 借助其更丰富的结构与语义特征, 以增强模型翻译复杂句子的能力。

此外, 不同词性的易错词, 有其显著关联的错误类型。这一发现为更具针对性的翻译优化提供了方向。未来工作将进一步探索面向风格、结构与术语一致性的全句级优化建模。

2.3 基于 RAG 的上下文学习优化

为了在推理阶段提升包含易错词句子的翻译质量, 我们设计了基于检索增强生成 (RAG) 的上下文学习优化策略。该策略的核心思想是: 从已有的高质量翻译语料中检索出与待翻译句子最相关且质量最高的示例, 将其作为上下文提示来引导模型生成更好的翻译。

2.3.1 检索策略设计与实现

我们设计并对比了多种检索策略, 以找到最优的示例选择方法:

基线方法

- 直接翻译: 不使用任何示例, 让模型直接进行翻译

- 随机示例: 从语料库中随机选择翻译示例

- 含易错词示例: 随机选择包含相同易错词的翻译示例

检索优化方法

- BM25 词汇检索: 利用经典的 BM25 算法计算查询句子与候选示例的词汇相似度:

$$BM25(q, d) = \sum_{i=1}^n IDF(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{avgdl}\right)} \quad (6)$$

其中, q 为查询句子, d 为候选示例, $f(q_i, d)$ 为词汇 q_i 在文档 d 中的频次, k_1 和 b 为算法参数。

- BGE 语义检索: 采用 BGE 预训练模型计算句子级语义相似度:

$$\text{sim}_{BGE}(s_1, s_2) = \cos(E_{BGE}(s_1), E_{BGE}(s_2)) \quad (7)$$

其中, E_{BGE} 表示 BGE 模型的句子编码函数。

- 质量导向选择: 基于 COMET 无参考质量评分, 从包含易错词的候选集中选择翻译质量最高的示例。

- 混合检索策略: 结合质量分数和语义相似度的加权组合:

$$\text{score}_{\text{hybrid}}(\text{example}) = \alpha \cdot \text{COMET}_{\text{ref_free}}(\text{example}) + \beta \cdot \text{sim}_{BGE}(\text{query}, \text{example}) \quad (8)$$

其中, α 和 β 为权重参数, 通过网格搜索确定最优值。

2.3.1 提示构建

对于选定的示例, 我们采用统一的提示模板进行上下文学习:

```
Translate Chinese to English:
Chinese: {example_source}
English: {example_target}
Translate Chinese to English:
Chinese: {query_source}
English:
```

2.4 针对性训练数据构建与 LoRA 微调

除了推理时优化, 我们还探索了通过微调来提升模型对易错词及相关翻译难点的处理能力。关键在于构建高质量的、针对易错词的训练数据集。

2.4.1 训练数据构建策略

我们对比了以下几种数据构建策略:

随机采样从原始训练集中随机抽取固定数量的句对, 作为基线对照。

高质量采样 (High COMET) 基于 COMET 质量评分, 选择得分最高的句对, 确保训练数据的整体质量。

易错词加权采样基于易错词的影响程度进行加权采样, 具体步骤如下:

1. 计算每个易错词的采样权重:

$$\text{weight}(w) = \frac{\log(\text{score}(w) + 1)}{\sum_{w' \in W_{\text{bad}}} \log(\text{score}(w') + 1)} \quad (9)$$

其中, $\text{score}(w)$ 为易错词评分。

2. 分配采样数量:

$$n_w = N_{\text{total}} \times \text{weight}(w) \quad (10)$$

其中, N_{total} 为总采样句子数。

3. 质量排序与选择: 对于每个易错词 w , 从包含该词的所有句子中选择质量最高的 n 个, 质量评分综合考虑翻译质量、困惑度和句子长度:

$$\text{quality}(\text{sent}) = 0.8 \times \text{norm}(\text{COMET}_{\text{ref_free}}) + 0.1 \times \text{norm}(\text{perplexity}^{-1}) + 0.1 \times \text{norm}(\text{length}^{-1}) \quad (11)$$

参考数据 (ALMA) 使用 ALMA 项目^[9]提供的高质量人工筛选数据作为上界参考。

2.4.2 LoRA 微调方法

采用低秩适应 (LoRA) 进行参数高效微调, 在保持预训练模型大部分参数不变的情况下, 仅更新少量附加参数。对于权重矩阵 $W_0 \in \mathbb{R}^{d \times k}$, LoRA 将其更新分解为:

$$W = W_0 + BA \quad (12)$$

其中, $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, 秩 $r < \min(d, k)$ 。

3 实验

3.1 实验设置

3.1.1 数据集

我们在三个翻译方向上进行了实验评估：

中英翻译(Zh→En)：训练集包含 119 万句对，来自 NIST MT 训练数据；测试集采用 WMT22 中英测试集，共 2,037 个句对；易错词检索库则由训练集中筛选出的高质量句对构建。

英德翻译(En→De)：训练集规模为 121 万句对，来源于 WMT 训练数据；测试集为 WMT22 英德测试集，共 1,984 个句对；易错词检索库同样从训练集中筛选得到。

阿英翻译(Ar→En)：训练集包含 99 万句对，来自 WikiMatrix；测试集从 WMT22 英德测试集选取前 2,000 个句对；易错词检索库由训练集高质量句对构成。

3.1.2 模型与实现细节

在基础模型方面，我们选用了 Llama2-7B 和 Llama3.1-8B，后者在多语言处理上具备更强的能力。此外，Llama 系列模型在开源大模型体系中具有较强的代表性，许多后续模型（如 Mistral、Qwen、Yi 等）均基于其架构改进。因此，在 Llama 系列模型上得到的规律更具可推广性与研究参考价值。

推理阶段基于 vLLM v0.3.0 框架，生成参数设定为 temperature=0、max_tokens=150、seed=42，批处理大小为 32。

微调阶段采用 LoRA 方法，参数设置为 rank=16、alpha=16、dropout=0.1，目标模块包括注意力层的 Q、K、V 投影以及前馈网络的 gate、up、down 投影层。优化器使用 AdamW，学习率为 $1e-4$ ，权重衰减 0.01。训练配置为批次大小 12，梯度累积步数 4，共训练 1 个 epoch。实验在 NVIDIA A100 40GB GPU 上完成。

检索系统方面，采用示例质量与示例语义相似度（相似度计算使用 BGE 模型）的混合方法，混合权重通过验证集确定为 $\alpha=0.8$ 、 $\beta=0.2$ 。

3.1.3 评估指标

我们使用三类互补的自动评估指标：

COMET：基于 XLM-R 的神经评估指标，采

用 wmt22-comet-da 模型，与人工评分具有高相关性。

COMET-ref-free：无需参考译文的质量评估，使用 wmt22-cometkiwi-da 模型。

BLEU：基于 n-gram 匹配的经典指标，使用 SacreBLEU 实现。

3.2 易错词识别结果

3.2.1 易错词统计

表 3 展示了不同语言对和模型下识别出的易错词数量：

表 3 易错词识别统计结果

Table 3 Error-prone word identification and statistics results

翻译方向	模型	易错词数量
Zh→En	Llama2-7B	5217
	Llama3-8B	6289
En→De	Llama2-7B	1522
	Llama3-8B	789
Ar→En	Llama2-7B	4791
	Llama3-8B	3247

结果显示，中英、阿英翻译方向的易错词数量显著多于英德翻译，这反映了语言结构差异对易错词翻译难度的影响。语言结构差异越大，源端易错词现象越显著。

3.2.2 易错词影响验证

表 4 展示了包含易错词的句子与整体测试集的翻译质量对比：

表 4 易错词对翻译质量的影响

Table 4 The impact of error-prone words on translation quality

翻译方向	模型	测试集整体 COMET	含易错词句 子 COMET
Zh→En	Llama2-7B	76.38	73.89
	Llama3-8B	79.21	78.48
En→De	Llama2-7B	78.33	75.27
	Llama3-8B	81.90	78.64
Ar→En	Llama2-7B	73.29	71.52
	Llama3-8B	84.84	84.02

实验结果证实了易错词的存在确实显著影响翻译质量，为后续优化方法提供了理论依据。

3.3 主要实验结果

3.3.1 上下文学习优化结果

表 5 展示了不同检索策略的效果

表 5 上下文学习优化实验结果

Table 5 Context learning optimization experiment results

方法	中文-英语			阿拉伯语-英语			英语-德语		
	COMET	COMET-kiwi	BLEU	COMET	COMET-kiwi	BLEU	COMET	COMET-kiwi	BLEU
Llama2-7b									
直接翻译	76.38	73.51	19.34	73.29	66.91	15.61	78.33	73.73	21.27
随机示例	76.34	73.67	18.83	73.76	67.60	16.39	79.24	75.42	21.91
含易错词示例	76.74	73.87	19.33	74.09	67.98	16.63	79.39	75.61	22.02
BM25 检索(易错词库)	76.86	73.84	19.32	74.64	68.44	16.95	79.25	75.46	21.90
BM25 检索(全库)	76.68	73.71	19.18	74.19	68.02	16.65	79.31	75.60	22.04
BGE 语义检索	76.67	73.86	19.19	74.42	68.03	17.28	79.32	75.66	22.09
质量导向选择	76.97	74.06	19.73	74.41	68.16	16.83	79.34	75.49	22.08
混合策略(质量+BGE)	77.14	74.29	19.47	74.82	68.51	17.15	79.36	75.60	22.10
Llama3.1-8B									
直接翻译	79.21	76.31	22.77	84.84	79.35	32.50	81.90	78.35	26.69
随机示例	79.14	76.27	22.17	84.99	79.59	33.05	82.20	78.78	26.97
含易错词示例	79.53	76.53	23.33	84.83	79.46	32.85	82.17	78.79	26.94
BM25 检索(易错词库)	79.65	76.53	23.44	84.94	79.56	32.87	82.19	78.73	26.84
BM25 检索(全库)	79.46	76.43	23.28	84.93	79.53	32.97	82.08	78.59	27.12
BGE 语义检索	78.81	75.89	22.25	84.86	79.43	32.85	82.19	78.65	26.79
质量导向选择	79.57	76.64	23.47	84.92	79.57	33.06	82.25	78.90	27.05
混合策略(质量+BGE)	79.76	76.87	23.18	84.99	79.62	32.99	82.18	78.80	26.98

3.3.2 LoRA 微调结果

表 6 展示了不同训练数据策略的微调效果。所有方法使用相同数量的训练数据(约 15,000 条)

表 6 LoRA 微调实验结果

Table 6 LoRA fine-tuning experiment results

训练数据策略	中文-英语			阿拉伯语-英语			英语-德语		
	COMET	COMET-kiwi	BLEU	COMET	COMET-kiwi	BLEU	COMET	COMET-kiwi	BLEU
Llama2-7b									
Baseline(无微调)	76.38	73.51	19.34	73.29	66.91	15.61	78.33	73.73	21.27
随机采样	76.20	72.87	16.89	78.26	72.08	22.03	81.15	78.47	22.35
高 COMET 采样	77.81	75.43	18.20	78.10	71.89	21.89	81.72	79.53	23.21
ALMA 高质量数据	78.51	75.71	21.42	-	-	-	81.32	78.65	24.21
易错词加权采样	78.23	75.83	19.02	79.07	73.63	23.06	81.13	78.09	20.93
Llama3.1-8B									
Baseline(无微调)	79.21	76.31	22.77	84.84	79.35	32.50	81.90	78.35	26.69
随机采样	79.05	76.18	21.49	85.18	79.54	35.11	84.24	81.04	27.50
高 COMET 采样	80.07	77.81	21.98	85.59	80.41	35.41	84.64	81.88	28.14
ALMA 高质量数据	80.04	77.32	24.81	-	-	-	84.46	81.63	28.85
易错词加权采样	80.23	77.74	23.01	85.84	80.42	36.16	83.76	80.93	25.23

3.4 消融实验

尽管易错词本身并不一定被模型直接误译,但大量实证分析表明,它们往往出现在句法结构更复杂、语义歧义更强的句子中,因而可被视为复杂翻译难点的稳定指示器。基于这一认识,本文设计的消融实验并非用于验证对易错词进行词级修复的有效性,而是用于检验:易错词这类“复杂度指示信号”是否能够被随机选取的词汇线索所替代。

在此基础上,我们构建了一个伪易错词驱动的对照消融设置。具体而言,在不改变训练数据总规模、

以确保公平比较。另外,由于 ALMA 高质量数据集中阿英语向训练集数量不足,这里不进行比较。

采样流程及质量筛选标准的前提下,我们从源端词表中去掉易错词后随机选取一组词汇,使其在数量及频次分布上与易错词集合保持一致,并将其作为“伪易错词”参与训练数据的构建过程。随后,按照与易错词加权采样完全相同的策略,从包含这些随机词的句子中筛选高质量样本,用于对基线模型进行 LoRA 微调。通过这一设计,我们确保两种设置在数据规模、质量控制及训练配置上完全一致,唯一差异在于所使用的词汇是否具备“指示器”属性。

实验结果如表 7 所示,在中英与阿英翻译任务中,伪易错词采样在 COMET 指标上明显低于易错词

加权采样。该结果表明,仅依赖词频匹配的非易错词线索不足以有效引导模型学习复杂翻译模式,而基于易错词构建的训练数据能够更集中地覆盖高难度翻译情形,从而带来更显著的性能提升。上述消融实验进一步验证了本文的核心假设:易错词并非通过词级纠错直接提升翻译质量,而是作为复杂句法与语义现象的稳定指示器,在训练阶段引导模型更高效地对齐困难样本分布,这也解释了其相较于随机或仅基于质量筛选的数据构建方法所表现出的优势。

表 7 伪易错词采样消融实验

Table 7 Pseudo Error-Prone Word Sampling Ablation Experiment

翻译方向	模型	易错词采样 COMET	伪易错词采样 COMET
Zh→En	Llama2-7B	78.23	77.29
	Llama3-8B	80.23	79.43
Ar→En	Llama2-7B	79.07	78.13
	Llama3-8B	85.84	85.31

4 实验结果与分析

4.1 易错词识别有效性分析

实验结果表明,我们提出的源端易错词识别方法能够有效识别出影响翻译质量的关键词汇。从表 2 可以看出,语言结构差异对易错词的产生有显著影响。作为分析型语言与综合语的代表,中英、阿英方向识别出的易错词数量(分别为 5217-6289 个和 3247-4791 个)远多于同属印欧语系的英德翻译方向(789-1522 个)。这说明源语言与目标语言的结构差异越大,由词汇用法、句法结构引发的翻译难点就越多,易错词现象也越普遍。

表 3 的结果进一步验证了易错词对翻译质量的负面影响。在所有语言对和模型上,包含易错词的句子其 COMET 评分普遍低于整体测试集 1-3 分,这证明了我们后续针对易错词进行优化的必要性和有效性。

4.2 上下文学习优化效果分析

4.2.1 检索策略的普遍有效性

从表 4、5 和 6 的结果来看,精心设计的检索策略在不同语言对上均表现出优越性:

1. 易错词导向优于随机选择:在中英和阿英这两个结构差异较大的语言对上,“含易错词示例”的提示效果明显优于“随机示例”,证明了针对易错词进行提示的有效性。

2. 混合策略表现最佳:结合质量和语义的“混合策略”在绝大多数情况下都取得了最好的 COMET 分

数。例如,在阿英翻译上,Llama2-7B 的 COMET 分数提升了 1.53 分,中英翻译提升了 0.76 分。这说明,一个优秀的示例不仅要与待翻译句子语义相关,其自身的翻译质量也至关重要。

4.2.2 语言对特异性分析

不同语言对的实验结果揭示了方法的适用性边界:

1. 结构差异大的语言对(中英、阿英):在这两个方向上,不同检索策略之间的性能差异较大,优化空间也更大。例如,在阿英翻译中,Llama2-7B 的最佳策略(混合策略)比最差的非基线策略(随机示例)高出 1.06 分 COMET。这说明,对于这类语言对,必须采用更精细的检索策略才能获得显著提升。

2. 结构相似的语言对(英德):在英德翻译中,即使是简单的随机示例也能带来可观的性能提升(Llama2-7B: +0.91 COMET)。而基于易错词的优化方法则提升的并不明显。这表明,对于结构相似的语言,模型更容易从任何形式的上下文中学到翻译模式,且易错词较少,故不适用基于易错词的优化方法。

4.2.3 模型规模的影响

Llama3.1-8B 作为更强大的基础模型,其本身翻译质量更高,因此通过上下文学习优化的绝对提升空间相对较小。例如,在阿英翻译中,Llama2-7B 的 COMET 最大提升为 1.53 分,而 Llama3.1-8B 仅为 0.15 分。这说明,更强大的模型已经内化了更多的语言知识,部分缓解了易错词带来的问题,但针对性的优化仍然能够带来稳定的性能增益。

4.3 LoRA 微调效果分析

4.3.1 数据质量与采样策略的关键性

LoRA 微调的成功与否高度依赖于训练数据的质量和相关性:

1. 高质量数据是基础:在所有语言对上,使用“ALMA 高质量数据”或“高 COMET 采样”数据进行微调,都取得了显著的性能提升,而“随机采样”数据甚至可能会导致性能下降(中英 Llama2: -2.45 BLEU)。这证实了数据质量是微调成功的先决条件。

2. 易错词加权采样的有效性:我们的“易错词加权采样”策略在结构差异大的中英和阿英翻译上表现尤为出色。在阿英翻译中,它在两个模型上都是最佳策略,使 Llama2-7B 的 COMET 分数提升了惊人的 5.78 分。在中英翻译中,其效果也与人工筛选的 ALMA 数据相当。这强有力地证明了,针对性地让模型学习如何处理已知难点,是一种极其高效的微调方法。

4.3.2 语言对特异性分析

1.中英、阿英翻译：易错词采样数据在这两个方向上取得了最佳或接近最佳的效果，说明对于结构差异大的语言对，基于易错词的针对性训练是一种高效且自动化的构建高质量训练集的方法。

2.英德翻译：在该方向上，“易错词加权采样”数据效果一般。这与上下文学习优化方法表现出同样的趋势，再次证明在结构相似的语言对上，不适用基于易错词的优化方法。

4.4 综合讨论

我们的系列实验不仅验证了方法的有效性，也与前文的易错词机理探究形成了闭环。机理探究发现，易错词是复杂翻译挑战的“指示器”，而我们的优化实验则证明，无论是通过上下文学习向模型“提示”这些难点的正确范例，还是通过微调让模型系统性地“学习”如何处理这些难点，都能显著提升翻译质量。

5 结束语

本研究提出了一个基于源端易错词的系统性大语言模型机器翻译优化框架，并通过在多个语言对上的详尽实验验证了其有效性。

首先，我们建立了从易错词识别、机理探究到靶向优化的完整流程。通过统计分析识别出影响翻译质量的关键易错词，并结合词语对齐和错误分类的深度学习分析，揭示了易错词作为复杂翻译挑战“指示器”的核心机理。我们发现，不同词性的易错词系统性地预示着特定类型的翻译错误，这一发现为后续的优化策略提供了理论依据。

其次，我们设计了基于检索增强的上下文学习策略。实验证明，结合翻译质量和语义相似度的混合检索策略能够最有效地选择高质量示例，在结构差异大的中英和阿英翻译任务上，分别带来了最高 1.53 分和 0.76 分的 COMET 提升。

同时，我们构建了针对易错词的高效 LoRA 微调方法。通过对包含易错词的高质量数据进行加权采样，我们的方法在结构差异大的语言对上取得了与人工筛选数据相当甚至更优的性能，在阿英翻译任务上实现了最高 5.78 分的 COMET 提升。

此外，跨语言对实验进一步揭示了本方法的适用规律。实验结果显示，在结构差异较大的语言对（如中英、阿英）中，易错词不仅数量更多，其所对应的翻译错误类型也更集中，因此基于易错词的上下文学习与微调策略能够显著提升模型性能。相比之下，在结构相似的语言对（如英德）中，易错词现象相对较少，模型本身更容易从通用上下文中学习翻译模式，

使得基于易错词的优化策略增益有限。这一结果表明，易错词方法在本质上更适用于结构差异显著、跨语系特征突出的翻译任务，也进一步验证了易错词作为复杂翻译挑战指示器的跨语言稳定性。

未来的工作可以从以下几个方向展开：一是将易错词的机理分析与优化策略更紧密地结合，例如，为不同词性的易错词设计自适应的、差异化的检索或微调策略；二是将该框架扩展到更多语言对和领域，验证其普适性；三是探索易错词的动态更新机制，以适应语言模型的持续演进和新语言现象的出现。

参考文献

- [1] XU H, SHARAF A, CHEN Y, et al. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation[J]. arXiv preprint arXiv:2401.08417, 2024.
- [2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems. 2020, 33:1877–1901.
- [3] DONG Q, LI L, DAI D, et al. A survey on in-context learning[J]. arXiv preprint arXiv:2301.00234, 2022.
- [4] LIU J, SHEN D, ZHANG Y, et al. What makes good in-context examples for GPT-3?[J]. arXiv preprint arXiv:2101.06804, 2021.
- [5] LU Y, BARTOLO M, MOORE A, et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity[J]. arXiv preprint arXiv:2104.08786, 2021.
- [6] ZHAO Z, WALLACE E, FENG S, et al. Calibrate before use: Improving few-shot performance of language models[C]//International Conference on Machine Learning. PMLR, 2021:12697–12706.
- [7] AGRAWAL S, ZHOU C, LEWIS M, et al. In-context examples selection for machine translation[J]. arXiv preprint arXiv:2212.02437, 2022.
- [8] AYCOCK S, BAWDEN R. Topic-guided example selection for domain adaptation in LLM-based machine translation[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. 2024:175–195.
- [9] XU H, et al. A paradigm shift in machine translation: Boosting translation performance of large language models[J]. arXiv preprint arXiv:2309.11674, 2023.
- [10] ACHIAM J, et al. GPT-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
- [11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017.
- [12] ZHANG B, HADDOW B, BIRCH A. Prompting large language model for machine translation: A case study [C]//International Conference on Machine Learning. PMLR, 2023.
- [13] PETRONI F, ROCKTÄSCHEL T, LEWIS P, et al. Lan

- guage models as knowledge bases?[J]. arXiv preprint arXiv:1909.01066, 2019.
- [14] RUBIN O, HERZIG J, BERANT J. Learning to retrieve prompts for in-context learning[J]. arXiv preprint arXiv:2112.08633, 2021.
- [15] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[C]//International Conference on Learning Representations. 2022.
- [16] GURURANGAN S, MARASOVIC A, SWAYAMDIPTA S, et al. Don't stop pretraining: Adapt language models to domains and tasks[J]. arXiv preprint arXiv:2004.10964, 2020.
- [17] DODGE J, SAP M, MARASOVIC A, et al. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus[J]. arXiv preprint arXiv:2104.08758, 2021.
- [18] GHAZVININEJAD M, GONEN H, ZETTLEMOYER L. Dictionary-based phrase-level prompting of large language models for machine translation[J]. arXiv preprint arXiv:2302.07856, 2023.
- [19] BOGOYcheV N, CHEN P. Terminology-aware translation with constrained decoding and large language model prompting[J]. arXiv preprint arXiv:2310.05824, 2023.
- [20] 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. 2023:902–911.
- [21] FREITAG M, REI R, LO C, et al. Experts, errors, and context: A large-scale study of human evaluation for machine translation[J]. Transactions of the Association for Computational Linguistics, 2021, 9:1460–1474.

1. 联系人: 李扬韬 ; 2. 通讯地址 (邮政编码): 苏州大学 215006; 3. 电子信箱、电话 (务必提供且长期有效, 最好是手机号): 2524278328@qq.com 18950115169/15359236718。

李扬韬 CCF 学生会员 会员号: A16234G 其余作者非会员。