

文章编号: 1003-0077(2026)05-0052-09

基于事实验证的五链式机器翻译幻觉检测智能体

魏嘉琴, 李茂西

(江西师范大学 人工智能学院, 江西 南昌 330022)

摘要: 机器翻译幻觉现象指机器译文中包含与源语言句子完全无关的翻译信息。当前基于模型内部状态的幻觉检测方法, 与基于外部工具的幻觉检测方法均存在显著不足, 前者依赖翻译模型内部推理过程产生的额外信息, 而后者使用的底层模型优化目标与幻觉检测任务存在偏差。根据机器翻译幻觉现象的特点, 该文提出了一种基于事实验证的五链式机器翻译幻觉检测智能体, 该智能体通过构建的五条链式流程, 预测机器译文中的事实与源语言中的事实是否一致, 从而进一步检测机器翻译幻觉现象。在公开数据集 HalOmi 句子级高资源自然幻觉检测子任务上的实验结果表明, 基于事实验证的五链式机器翻译幻觉检测智能体显著提高了机器翻译幻觉检测的性能, 且在大多数语言对上的表现超过了参与对比的最优方法。

关键词: 机器翻译幻觉; 智能体; 大语言模型; 链式提示; 提示优化

中图分类号: TP391

文献标识码: A

A Fact-verification-based Five-link Chain Machine Translation Hallucination Detection Agent

WEI Jiaqin, LI Maoxi

(School of Artificial Intelligence, Jiangxi Normal University, Nanchang, Jiangxi 330022, China)

Abstract: Machine translation hallucination refers to the phenomenon that the translated text contains information unrelated to the source language sentence. This paper proposes a fact-verification-based five-chain machine translation hallucination detection agent. This agent uses five sequential processes to determine whether the facts in the machine translation align with those in the source language, thus enabling the detection of hallucination phenomena in machine translation. Experimental results on the HalOmi sentence-level high-resource natural hallucination detection subtask show that the proposed method significantly improves the performance of hallucination detection and outperforms the best competing methods for most language pairs.

Keywords: machine translation hallucination; AI agent; large language models; chain-of-thought prompting; prompt optimization

0 引言

神经机器翻译中的幻觉(Hallucination)现象是指机器译文中包含了与源语言句子完全无关的翻译信息, 幻觉现象被视为低质量机器译文的一种外在表现, 同时也是神经机器翻译系统产生“病态”翻译文本的特殊体现, 它不仅影响翻译的准确性, 还显著削弱了用户对机器译文的信心^[1-2]。因此, 机器翻译幻觉检测方法的研究作为神经机器翻译的新兴研究方向具有重要意义。

前人关于机器译文的质量评估主要集中在对机

器译文整体进行打分(如机器译文质量估计与机器译文自动评价等)或对机器译文从词级进行细粒度错误分类^[3-5], 然而, 这些方法在检测机器翻译幻觉时存在显著不足。基于质量估计或自动评价指标的分数仅反映机器译文的整体流畅性或忠实度, 难以捕捉如幻觉等特定类型的错误。由于幻觉在机器译文中的出现频率相对较低, 其对整体评分的影响可能被稀释甚至忽略, 导致评分无法真实反映幻觉的严重性。尽管词级错误分类能够将错误细化为词汇、语法等类别, 但现有分类体系通常未将幻觉单独列为一种错误类型^[5], 因此, 幻觉无法在细粒度分析中被明确识别和量化。这种模糊性使得面向幻觉的针对性检测难以实现。

收稿日期: 2025-06-26 定稿日期: 2025-11-03

基金项目: 国家自然科学基金(62366020, 61662031)

现有的机器翻译幻觉检测方法包括基于模型内部状态的方法和基于外部工具输出的方法^[6]。基于模型内部状态的机器翻译幻觉检测方法是指：利用神经机器翻译模型在翻译过程中产生的信息预测机器译文是否存在幻觉，包括标准序列对数概率(Seq-Logprob)^[6]、ALTI^[7]和 Attn-OT^[8]等方法，它们通过检测基于注意力机制的神经机器翻译模型在翻译过程中的异常注意力分布，来识别翻译过程中产生的幻觉现象。

基于外部工具输出的幻觉检测方法是指：借助无监督模型、质量过滤器^[9-10]、质量估计系统^[11-12]、预训练语言模型或大语言模型等作为外部检测工具预测机器译文是否存在幻觉现象^[1]，包括使用预训练语言模型计算源语言句子与机器译文的余弦相似度^[13-14]、使用 NLI^[15]计算源语言句子与机器译文的双向蕴含得分以及使用机器译文自动评价系统 xCOMET^[16]作为质量过滤器等^[17]。

然而，基于模型内部状态的机器翻译幻觉检测方法需要借助翻译推理过程中产生的注意力分布等特征，导致其无法适配未采用注意力机制的神经机器翻译架构，仅适用于可获取内部信息的“白盒”模型。随着大语言模型，如 GPT^①系列为代表的“黑盒”方法逐渐成为主流翻译工具，此类方法因无法获取模型内部状态而失效。基于外部工具的方法虽具有通用性，但其底层模型的优化目标与幻觉检测任务存在偏差^[18]，且大多忽略了存在幻觉现象的源语言句子与机器译文之间事实不一致性这一重要特性。

除机器翻译幻觉外，大语言模型在问答与摘要生成等任务中也存在幻觉现象，它们常表现为：大语言模型有时会提供给用户流利但完全虚构、错误的回答内容^[19-20]。针对此类幻觉现象的检测，常借助外部数据库知识对生成内容进行检索验证，从而确保回答的真实性与逻辑性^[21-24]。但在机器翻译幻觉检测任务中，虚构信息体现为机器译文中存在的、与源语言句子完全无关的翻译。这意味着，即使源语言句子存在不符合逻辑或不真实的内容，机器译文仍需在不添加任何其他内容的情况下尽可能忠诚地传达其原意。因此，机器译文幻觉检测任务无法借助外部数据库或知识库直接验证机器译文的准确性，但也正因如此，验证机器译文与源语言句子表达事实是否一致将成为进行机器翻译幻觉检测的一种可能有效方式。

基于上述可能性，为了解决现有机器翻译幻觉检测方法中存在的不足，本文提出了一种基于事实验证的五链式机器翻译幻觉检测智能体(Machine Translation Hallucination Agent, MTHaluAgent)。MTH-

aluAgent 仅依赖源语言句子与机器译文即可完成机器翻译幻觉检测，兼容“白盒”与“黑盒”翻译系统。MTHaluAgent 通过五链式模块：幻觉初筛链、原子事实提取链、无效事实剔除链、事实蕴含推理链及补充判断链，直接针对机器翻译幻觉检测任务进行细粒度验证。幻觉初筛链首先对机器译文进行初步幻觉检测；鉴于机器翻译幻觉现象的关键特征在于机器译文中出现了源语言句子中并不存在的信息，事实蕴含推理链根据源语言句子与机器译文的事实关系，进一步检测机器译文中是否存在幻觉。实验中将本文所提方法与基线系统在公开数据集 HalOmi 句子级高资源自然幻觉检测子任务上进行对比。结果表明，本文所提方法在 8 个语言对上的平均性能优于进行对比的机器译文幻觉检测中的最优方法，并在 5 个语言对上的马修斯相关系数超过了进行对比的机器译文幻觉检测的最优结果。

1 相关工作

1.1 基于模型内部状态的机器翻译幻觉检测方法

当模型产生幻觉时，其置信度较低。基于此启发式假设，Guerreiro 等^[6]提出使用 Seq-Logprob 进行幻觉检测。Ferrando 等^[7]提出的 ALTI+方法计算源语言句子和机器译文的目标前缀对机器译文的贡献，并在扰动幻觉中可视化该贡献分布，由于幻觉在此方法中体现为源语言句子贡献的下降，因此 Dale 等^[17]将 ALTI+用于自然幻觉检测。Xu 等^[25]聚焦于神经机器翻译模型在推理过程中的表现，通过模型自省机制在翻译推理阶段识别其潜在的幻觉症状，从而预测翻译是否存在幻觉，他们设计并提出了一个逐层相关性传播^[26]的脱节幻觉检测器。但该检测器仅在构建的扰动幻觉数据集上进行训练，且被设计用来检测扰动幻觉下的脱节幻觉。此外，基于模型内部状态的幻觉检测方法只适用于“白盒”机器翻译模型，对于以 GPT 系列等的“黑盒”神经机器翻译模型，由于无法获得翻译过程中产生的信息，且内部方法受限于基于注意力机制构建的神经机器翻译模型，因此基于模型内部状态的幻觉检测方法将在这些情况下失效。

1.2 基于外部工具输出的机器翻译幻觉检测方法

Zhou 等^[27]在预训练语言模型 BART^[28]上合成的

① <https://chat.openai.com/>

中英扰动幻觉数据集上,在有人工参考译文的情况下对预训练语言模型 XLM-R^[29]进行微调,微调后的 XLM-R 能够在有人工参考译文或无人工参考译文的情况下预测中英词级幻觉现象。Dale 等^[30]使用机器译文质量估计系统 BLASER-QE 进行幻觉检测,并取得优于标准序列对数概率、ALTI+方法等内部方法以及其他外部方法的性能^[1]。Benkirane 等^[31]评测了通过计算大语言模型 SONAR^[32]、GPT text-embeddings3-large^①、Cohere Embed v3^②以及 Mistral mistral-embed^[33]作为词嵌入空间时的源语言句子与机器译文的余弦距离来检测幻觉现象,此外,受 Liu 等^[34]在摘要任务上使用的 G-eval 提示启发,以及数据集 HalOmi 的人工标注指南和幻觉严重程度级别定义,他们从中提取出了幻觉检测的 G-eval 提示,并评测了 GPT4o^①、Command-R^③、Command-R Plus^③、LLaMa-70B^[35]等共八种大语言模型在 G-eval 提示下进行幻觉检测时的性能^[31]。

但是, Benkirane 等^[31]提出的幻觉检测 G-eval 提示并未针对机器翻译幻觉检测任务进行适配性调整,提示词中忽略了源语言与机器译文之间事实一致性的深入逻辑推理过程。受 Cheng 等^[36]的黑盒提示优化方法启发,本文设计了一种基于错误响应样本反馈的提示优化方法,通过此方法得到的两样本提示(2-shot prompt)更加符合机器翻译幻觉检测任务偏好,展现出更高的适用性与有效性。此外,使用 BLASER-QE 和基于计算余弦距离等机器翻译幻觉检测方法,都为单一检测机制,这在一定程度上制约了其性能,且易将低质量翻译或其他非幻觉错误预测为幻觉现象。相比之下,本文提出的基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent,通过构建五链式流程,实现了多层次、多维度的检测策略,并且能够在多种高资源语言对之间进行幻觉检测。除却初步的幻觉检测环节, MTHaluAgent 的事实蕴含推理链运用事实验证机制,通过推理判断机器译文中的事实与源语言中的事实是否一致,从而进一步检测机器翻译幻觉现象,显著提高了检测的全面性和准确性。

2 基于事实验证的五链式机器翻译幻觉检测智能体

大语言模型作为自然语言处理的核心工具,具备理解、生成及深入分析文本的能力。为了更有效地管理与大型语言模型的交互过程,本文引入了能够以结

构化方式与大语言模型进行决策和协作的 LangChain 框架^①。为了通过事实验证,即通过从机器译文中提取事实并验证它们是否能够与源语言句子保持一致,从而预测机器译文是否存在幻觉现象,本文使用 LangChain 与大语言模型协同作用,构建了一个高效且灵活的轻量级机器译文幻觉检测智能体——MTHaluAgent。为进一步提升机器译文幻觉检测方法的准确性与鲁棒性,本文提出了一种基于错误响应样本反馈的提示优化方法,对 Kenza 等提出的 G-eval 幻觉检测提示进行精细化改进,系统地提高了提示词的语义适配性和人类需求的契合度。

2.1 MTHaluAgent: 基于事实验证的五链式机器翻译幻觉检测智能体

为充分挖掘源语言句子与机器译文中隐含的信息, MTHaluAgent 采用多层次的处理方式,通过不同模块逐步进行优化,由于这些模块均由 LangChain 构建,因此将这些模块称为“链”。具体来说, MTHaluAgent 由五条链组成,分别是:幻觉初筛链、原子事实提取链、无效事实剔除链、事实蕴含推理链以及补充判断链,其系统框架如图 1 所示。

幻觉初筛链旨在初步评估机器译文是否存在幻觉现象。该链采用基于错误响应样本反馈的提示优化方法(见 2.2 节)获得的 2-shot prompt 作为输入引导,将源语言句子与对应的机器译文组成二元组,并输入大语言模型中。为了挖掘大语言模型识别幻觉的潜在能力,本文将温度固定在一个较高的数值以获得更加丰富的判断理由,在此过程中,幻觉初筛链将经历四轮重复独立的幻觉检测,以确保评估的全面性与准确性,如式 (1) 所示。

$$\begin{aligned} (\text{Labels}, \text{Reasons}) &= \text{Chain}_{2\text{-shot prompt}}(\text{src}, \text{mt}) \\ \text{Labels} &= (l_1, l_2, l_3, l_4) \\ \text{Reasons} &= (\text{reason}_1, \text{reason}_2, \text{reason}_3, \text{reason}_4) \end{aligned} \quad (1)$$

其中, $\text{Chain}_{2\text{-shot prompt}}$ 为幻觉初筛链, src 为源语言句子, mt 为机器译文, Labels 为大语言模型四次评估结果,分别为 l_1, l_2, l_3, l_4 , 取值为“Hallucination”或“No Hallucination”; Reasons 表示大语言模型做出该判断的原因,四轮检测的判断原因分别为 $\text{reason}_1, \text{reason}_2, \text{reason}_3, \text{reason}_4$ 。若 $l_1=l_2=l_3=l_4$, 则根据评估结果直接

① <https://chat.openai.com/>

② <https://cohere.com/blog/introducing-embed-v3>

③ <https://cohere.com/blog/command-series-0824>

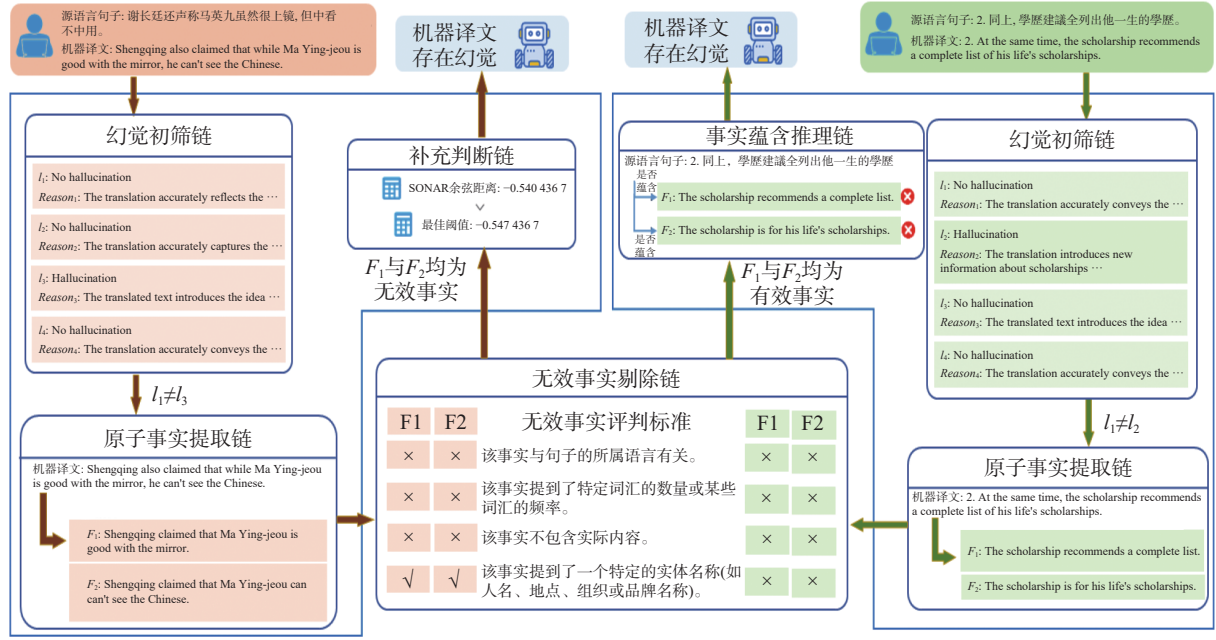


图1 MTHaluAgent 的系统框架图

判断该机器译文是否存在幻觉。

原子事实提取链则在 $l_i \neq l_j (i, j=1, 2, 3, 4)$ 的情况下, 从机器译文中提取若干候选原子事实 $\text{fact}_{1:k}$, 其中原子事实意为只包含一条不同命题的基本信息单元^[37], 如式 (2) 所示。

$$\text{fact}_{1:k} = \text{Chain}_{\text{extract facts}}(mt) = (\text{fact}_1, \text{fact}_2, \dots, \text{fact}_k) \quad (2)$$

其中, $\text{Chain}_{\text{extract facts}}$ 为原子事实提取链, k 为候选原子事实的个数。

鉴于候选原子事实中可能包含与机器译文所属语言相关特征, 故将此类事实定义为无效事实, 无效事实定义如图 2 所示。无效事实定义的核心原则是过滤掉那些仅与机器译文所属语言特征相关内容的原子事实, 以确保后续验证的准确性。图 2 中“定义 2”中的“特定词汇”主要指的是那些自身不构成完整语义断言的词汇, 例如对于一个典型的机器译文“GGRREEAATT”和源语言句子“棒极了!!”, 根据“定义 2”, 本文将候选原子事实“G 出现了 2 次”作为无效事实。其根本原因在于: 此类事实是基于机器译文独特的语言表现形式得出的, 它无法作为一个跨语言的、稳定的语义单元进行真伪验证。剔除此类事实, 是为了确保本文的系统聚焦于语义内容的忠实度, 而非表面形式的机械对齐。第 4 条定义将“提及特定实体名称(如人名、地点、组织或品牌名称)”的事实归类为无效事实, 其根本原因在于跨语言场景下实体表征的不确定性。在单语言任务(如摘要生成或问答系统等的幻觉

检测任务)中, 实体名称始终处于同一语言上下文, 其作为关键语义单元具有稳定的检测价值。然而, 机器翻译幻觉检测涉及跨语言转换, 机器译文中的实体名称常为源语言实体的音译转化, 当分别从语言句子与机器译文中提取原子事实时, 音译差异将使同名实体的指称一致性丧失, 进而影响后续推理过程, 因此本文将将其明确定义为无效事实予以剔除。

定义1: 该事实与句子的所属语言有关。

定义2: 该事实提到了特定词汇的数量或某些词汇的频率。

定义3: 该事实不包含实际内容。

定义4: 该事实提到了一个特定的实体名称(如人名、地点、组织或品牌名称)。

图2 无效事实定义

无效事实剔除链将根据无效事实定义逐一评估候选事实是否有效, 并最终得到有效原子事实的集合 $\text{facts}'_{1:v}$, 其中 v 为有效事实的个数。

事实蕴含推理链将依次评估源语言句子是否蕴含若干有效事实, 如式 (3) 所示。

$$\begin{aligned} \text{present}_{1:v} &= \text{Chain}_{\text{infer facts}}(\text{facts}'_{1:v}) \\ &= (\text{present}_1, \text{present}_2, \dots, \text{present}_v) \end{aligned} \quad (3)$$

其中, $\text{Chain}_{\text{infer facts}}$ 为事实蕴含推理链, $\text{present}_{1:v}$ 表示源语言句子是否蕴含各有效事实。若源语言句子蕴含所有有效事实, 则判定机器译文不存在幻觉; 否则, 判定为幻觉。

补充判断链将在候选原子事实均为无效事实的情况下,依据样本的语言方向进行后续处理。若样本为“英语到其他语言”方向(即 English→High-Resource Languages, HRLs),则使用幻觉初筛链进行第五次判断,得到 $Labels=(l_1, l_2, l_3, l_4, l_5)$, 根据此时 Labels 的结果,若幻觉判定频次超过非幻觉判定,则判定该机器译文存在幻觉,否则为非幻觉。对于“其他语言到英语”方向(即 HRLs→English),补充判断链将计算源语言句子与机器译文在 SONAR 语义空间中的余弦距离,当余弦距离超过最佳阈值时,则判定机器译文中存在幻觉。

为了进一步说明基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent 的框架及工作流程,本文在中英翻译方向上抽取两个实例进行分析,图 1 的左半部分为第一个实例,右半部分为第二个实例。在第一个实例中,机器译文中的“he can't see the Chinese”为源语言句子中不存在的内容,因此此条机器译文存在幻觉。由于提取的原子事实中包含人名,因此两条候选事实均被无效事实剔除链判定为无效事实,当无有效事实时, MTHaluAgent 将调用补充判断链判断源语言句子与机器译文间的余弦距离与最佳阈值的大小关系,由于该实例的余弦距离大于最佳阈值,因此补充判断链预测该条机器译文存在幻觉。在第二个实例中,机器译文中的“the scholarship recommends a complete list of his life's scholarships”为源语言句子中不存在的内容,因此此条翻译存在幻觉。由于提取的原子事实均为有效事实,因此事实蕴含推理链将判断源语言句子是否蕴含每一条事实,由于两条事实与源语言句子均不存在蕴含关系,因此事实蕴含推理链成功预测该条机器译文存在幻觉。

2.2 基于错误响应样本反馈的提示优化方法

基于错误响应样本反馈的提示优化方法通过迭代机制在反馈回路中不断精化提示,以实现多轮调整和自适应更新。本文通过此方法对 Benkirane 等提出的幻觉检测 G-eval 提示进行优化,从而得到 2.1 节中幻觉初筛链所用提示。基于错误样本反馈的提示优化方法框架如图 3 所示,它由三阶段构成:反馈收集阶段、错误分析阶段与提示生成阶段。反馈收集阶段将初始提示得到的预测结果进行分类,并从中提取一例错误响应样本,随后通过多维度的错误分析提取原提示中的潜在问题,进而在提示生成阶段生成改进后的提示,并将其与错判样本进行融合,形成闭环式反馈收集与优化机制。这一方法形成的检测提示能够有效提高幻觉检测提示的可靠性。

反馈收集阶段通过使用原提示,将源语言句子 src 与机器译文 mt 输入大语言模型,以获得幻觉评估结果,如式 (4) 所示。

$$\left[(X_i, Y_i, \bar{Y}_i'), (X_j, Y_j, \bar{Y}_j') \right] = \text{LLM}(\text{prompt}_{\text{halu}}, X), \quad (4)$$

$$Y_i, Y_j \in Y$$

其中, $\text{prompt}_{\text{halu}}$ 表示幻觉评估任务描述提示,初始为经由 G-Eval 推导出的幻觉评估任务描述提示。 X 表示由所有源语言句子与机器译文构成的全部样本,具体为 $X = \{x_r\}_r^n$, 其中, $x_r = (\text{src}_r, \text{mt}_r)$, n 为样本个数。 Y 表示样本的真实标签,具体为 $Y = \{y_r\}_r^n$, 样本 x_r 的真实标签为 y_r 。经由大语言模型通过提示得到的幻觉评估结果若与真实标签一致,则预测标签为 $\bar{Y}_i'$, 否则为 $\bar{Y}_j'$ 。 $(X_i, Y_i, \bar{Y}_i')$ 表示所有正确预测的样本集合, $(X_j, Y_j, \bar{Y}_j')$ 表示所有错误预测的样本集合。

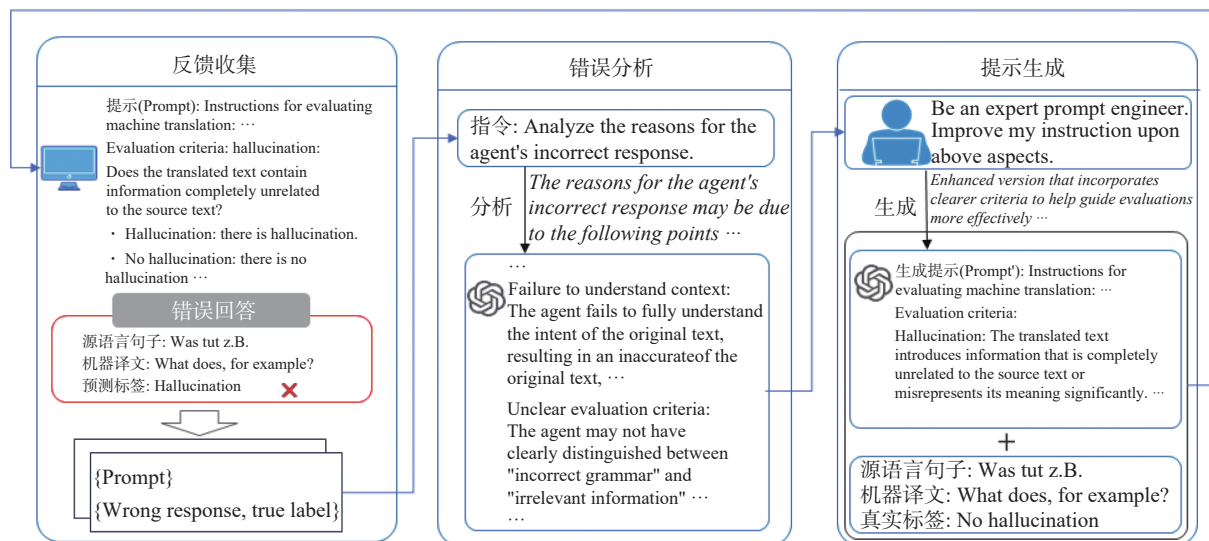


图 3 基于错误样本反馈的提示优化方法

错误分析阶段将原提示与预测错误样本、真实标签、预测标签组合为错误分析阶段所需的二元组。通过使用错误分析提示,将该二元组输入大语言模型,以获得原提示判断错误的分析结果,如式(5)所示。

$$\begin{aligned} \text{reasons} = \text{LLM}(\text{prompt}_{\text{halu}}, (x_m, y_m, \overline{y'_m})) \\ (x_m, y_m, \overline{y'_m}) \in (X_j, Y_j, \overline{Y'_j}) \end{aligned} \quad (5)$$

其中, reasons 为分析结果, $(x_m, y_m, \overline{y'_m})$ 为错判样本集合中的一个子样本。

提示生成阶段将原提示与分析结果组合为提示生成阶段所需的二元组,并将该二元组输入大语言模型,以得到更贴近人类意图的新提示,如式(6)所示。

$$\begin{aligned} \text{prompt}' = \text{LLM}(\text{prompt}_{\text{halu}}, \text{reasons}) \\ \text{prompt}_{\text{halu}} = [\text{prompt}', (x_m, y_m, \overline{y'_m})] \end{aligned} \quad (6)$$

其中, prompt' 为所得到的新提示,将新提示与错误分析阶段所选择的样本组合作为反馈收集阶段所需的二元组。经过两次迭代,得到最终的两样本提示(2-shot prompt)。

3 实验

3.1 实验设置

为了测试基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent 的效果,本文在公开数据集 HalOmi 句子级高资源自然幻觉检测子任务上进行了实验验证。性能测量方法遵循 Benkirane 等人的二元检测标准,使用马修斯相关系数(Matthews Correlation Coefficient, MCC)评价幻觉检测性能。该值越大,表明方法性能越好。

考虑到数据集中幻觉与非幻觉标签极度不平衡的情况,本文使用马修斯相关系数作为主要度量标准,它在面对不平衡数据时,能够给出更加全面的性能评估。马修斯相关系数的取值范围为 $[-1, 1]$ 。其中, -1 表示完全错误的预测, 0 表示预测与随机猜测相似, $+1$ 表示预测完全正确。

在基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent 中,本文使用大语言模型 ChatGPT3.5-turbo-instruct 作为核心组件之一,处理自然语言理解与检测推理等任务,当大语言模型在幻觉检测链中进行推理时设置温度为 0.4 。使用 SONAR 语义空间计算源语言句子与机器译文的余弦距离,根据开发集中德英语言对无效事实样本的余弦距离,通过最大化准确率-召回率曲线中的 F_1 得分确定补充判断链中的最佳阈值。

为进一步评估本方法的实用性,本文对其计算经济成本进行了量化分析。以当前 ChatGPT-3.5-turbo-instruct 的 API 定价为例,若样本在幻觉初筛链阶段即判定完成,则单次检测的平均成本约为 ¥0.10~0.15。若该样本在幻觉初筛链中被判定为需进一步验证,则因需调用后续链路的模型,总成本将相应增加至约 ¥0.30~0.45。

实验将本文提出的方法与现有的多种幻觉检测技术进行了系统对比分析。具体而言,这些对比方法包括使用机器译文质量估计系统 Blaser-QE 作为外部检测工具,通过计算源语言句子和机器译文在大语言模型 SONAR、GPT text-embeddings3-large、Cohere Embed v3 等嵌入空间的余弦距离来进行幻觉检测。此外,本文还将所提方法与大语言模型 GPT4-turbo、GPT4o、Command-R、Command-R Plus、Llama-70B 等在使用 G-eval 提示进行幻觉检测时的性能进行了对比。其中 Llama-70B 在该提示下性能最优。

在验证集与测试集的选择上,本文遵从 Benkirane 等人的设置,选择德英(De-En)与英德(En-De)两个语言对作为验证集,进行全部提示的调整与生成、无效事实的定义以及温度等参数的确定。使用其余八个语言对:英语-阿拉伯语(En-Ar)、英语-俄语(En-Ru)、英语-西班牙语(En-Sp)、英语-中文(En-Zh)、阿拉伯语-英语(Ar-En)、俄语-英语(Ru-En)、西班牙语-英语(Sp-En)、中文-英语(Zh-En)作为测试集。

3.2 实验结果

表1展示了在数据集 HalOmi 句子级高资源自然幻觉检测子任务上,本文提出的基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent 和对比的其他幻觉检测方法的马修斯相关系数。

表1的实验结果表明,基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent 的平均性能最好,与基线中在 G-eval 提示下的最优模型 Llama-70B 相比,平均性能提高了 $0.03(0.63 \text{ vs } 0.66)$;在英语到其他语言方向(En-HLRs)上,平均性能提高了 $0.04(0.59 \text{ vs } 0.63)$;在其他语言到英语方向(HLRs-En)上,平均性能提高了 $0.02(0.67 \text{ vs } 0.69)$;在 En-Ar、En-Ru、En-Sp、Ar-En、Ru-En、Zh-En 六个翻译方向上分别提高了 43.59% 、 24.53% 、 7.58% 、 3.90% 、 15.69% 。与使用 Blaser-QE 作为基座模型的外部方法相比,基于事实验证的智能体 MTHaluAgent 的马修斯相关系数平均提高了 $0.19(0.47 \text{ vs } 0.66)$;在 En-HLRs 翻译方向上,平均性能提高了 $0.26(0.37 \text{ vs } 0.63)$;在 HLRs-En 翻译

表 1 不同幻觉检测方法在数据集 HalOmi 句子级高资源自然幻觉检测子任务上的性能

(单位: %)

模型		马修斯相关系数										
		En-Ar	En-Ru	En-Sp	En-Zh	Ar-En	Ru-En	Sp-En	Zh-En	Avg.	En-HRLs	HRLs-En
词嵌入空间	BLASER-QE	0.41	0.13	0.61	0.31	0.69	0.64	0.51	0.45	0.47	0.37	0.57
	SONAR	0.40	0.17	0.41	0.51	0.76	0.75	0.50	0.65	0.52	0.37	0.67
	GPT text-embeddings3-large	0.35	0.20	0.66	0.78	0.77	0.61	0.58	0.60	0.57	0.50	0.64
	Cohere Embed v3	0.40	0.20	0.66	0.71	0.69	0.67	0.51	0.55	0.55	0.49	0.61
	Mistral mistral-embed	0.29	0.54	0.70	0.63	0.70	0.67	0.47	0.60	0.54	0.54	0.61
G-eval提示下的 大语言模型	GPT4-turbo	0.27	0.27	0.61	0.53	0.47	0.34	0.55	0.39	0.43	0.42	0.44
	GPT4o	0.38	0.34	0.53	0.35	0.52	0.43	0.50	0.41	0.43	0.40	0.47
	Command-R	0.43	0.34	0.52	0.57	0.56	0.39	0.59	0.43	0.48	0.46	0.49
	Command-R Plus	0.14	-0.03	0.55	0.36	0.57	0.50	0.63	0.41	0.39	0.25	0.53
	Mistral 8×22b	0.25	0.59	0.55	0.72	0.62	0.53	0.55	0.45	0.53	0.53	0.54
	Claude Sonnet	0.39	0.52	0.59	0.56	0.58	0.43	0.64	0.46	0.52	0.51	0.53
	Claude Opus	0.39	0.26	0.59	0.63	0.59	0.48	0.58	0.39	0.49	0.47	0.51
	Llama3-70B	0.39	0.53	0.66	0.78	0.77	0.72	0.68	0.51	0.63	0.59	0.67
本文方法	MTHaluAgent	0.56	0.66	0.71	0.59	0.80	0.75	0.60	0.59	0.66	0.63	0.69

方向上, 平均性能提高了 0.22(0.57 vs 0.69); 在 En-Ar、En-Ru、En-Sp、En-Zh、Ar-En、Ru-En、Sp-En、Zh-En 八个翻译方向上分别提高了 36.59%、407.69%、16.39%、90.32%、15.94%、17.19%、17.65%、31.11%。这些结果表明本文提出的基于事实验证的智能体 MTHaluAgent 显著地提高了机器翻译幻觉检测的性能。

3.3 消融实验

(1)为了说明采用机器译文与源语言句子的事实一致性验证对本文所提方法性能的影响, 本文对比在

处于如下情况时机器翻译幻觉检测的马修斯相关系数。

情况 1: Chain_{2-shot prompt} 为仅用幻觉初筛链在温度为 0 且仅进行一次幻觉检测, 其中温度固定为 0 是为了确保生成回答的稳定性。

情况 2: + Chain_{extract facts}&Chain_{infer facts}, 在情况 1 的基础上增加使用原子事实提取链, 在不去除无效事实的情况下直接对原子事实进行蕴含推理过程。

情况 3: MTHaluAgent, 即使用完整的基于事实验证的五链式机器翻译幻觉检测。

三种情况的结果如表 2 所示。

表 2 MTHaluAgent 中不同链组合后进行机器翻译幻觉检测的马修斯相关系数

(单位: %)

方法	马修斯相关系数								
	En-Ar	En-Ru	En-Sp	En-Zh	Ar-En	Ru-En	Sp-En	Zh-En	Avg.
Chain _{2-shot prompt}	0.41	0.66	0.55	0.50	0.62	0.66	0.44	0.54	0.55
+ Chain _{extract facts} & Chain _{infer facts}	0.47	0.56	0.61	0.52	0.60	0.63	0.62	0.49	0.57
MTHaluAgent	0.56	0.66	0.71	0.59	0.80	0.75	0.60	0.59	0.66

可以看到在 En-Ru、Ar-En、Ru-En 及 Zh-En 语言对上, 在添加事实提取链后马修斯相关系数反而下降, 这是由于候选原子事实中存在无效事实, 在没有去除无效事实时直接用候选原子事实对源语言句子进行蕴含推断导致的误判。随后添加无效事实剔除链, 可以看到马修斯相关系数出现明显回升, 并且最终结果超过情况 1 与情况 2, 这表明了引入有效事实提取与机器译文和源语言句子事实一致性验证能有

效提高机器译文幻觉检测的性能。

(2)为了验证基于错误响应样本反馈的提示优化方法对基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent 的影响, 本文将 ChatGPT3.5-turbo-instruct 的温度固定为 0, 只使用 2-shot prompt 的幻觉初筛链进行幻觉检测, 并将其结果与使用 G-eval 提示的 GPT4-turbo 与 GPT4o 的结果进行对比, 结果如表 3 所示。对比使用 G-eval 提示的 GPT4-turbo 与 GPT4o 的幻觉检测

结果,只使用基于错误响应样本反馈的提示优化方法得到的 2-shot prompt 的幻觉初筛链在仅使用 ChatGPT3.5-turbo-instruct 的情况下提升了至少 0.10(0.44 vs 0.47 vs 0.57)的平均性能,证明了该提示优化方法的有效性。

表 3 基于错误响应样本反馈的提示优化方法与 G-eval 提示下机器翻译幻觉检测的马修斯相关系数 (单位: %)

方法	马修斯相关系数										
	En-Ar	En-Ru	En-Sp	En-Zh	Ar-En	Ru-En	Sp-En	Zh-En	Avg.	En-HRLs	HRLs-En
GPT4-turbo _{G-eval}	0.27	0.27	0.61	0.53	0.47	0.34	0.55	0.39	0.43	0.42	0.44
GPT4o _{G-eval}	0.38	0.34	0.53	0.35	0.52	0.43	0.50	0.41	0.43	0.40	0.47
Chain _{2-shot prompt}	0.41	0.66	0.55	0.50	0.62	0.66	0.44	0.54	0.55	0.53	0.57

4 结论与展望

本文根据机器翻译幻觉检测任务的特点,结合最新的智能体技术,提出了基于事实验证的五链式机器翻译幻觉检测智能体 MTHaluAgent。实验结果表明该方法显著提高了机器翻译幻觉检测的效果。

在未来的工作中,我们将尝试使用更为健壮的提示或新的推理过程,继续在低资源语言上^[38]进行幻觉检测,进一步提高机器翻译幻觉检测的性能。

参考文献

[1] DALE D, VOITA E, LAM J, et al. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation[C]//Proceedings of the EMNLP, 2023: 638-653.

[2] LEE K, FIRAT O, AGARWAL A, et al. Hallucinations in neural machine translation[C]//Proceedings of ICLR, 2019.

[3] 罗兰, 何贤敏, 李茂西. 句子级别机器译文质量估计研究综述[J]. 情报工程, 2022, 8(2): 034-050.

[4] 李良友, 贡正仙, 周国栋. 机器翻译自动评价综述[J]. 中文信息学报, 2014, 28(3): 81-91.

[5] LOMELL A, USZKOREIT H, BURCHARDT A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics[J]. *Advances in Tradumàtica*, 2014(12): 455-463.

[6] GUERREIRO N M, VOITA E, MARTINS A F T. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation[C]//Proceedings of the EACL, 2023: 1059-1075.

[7] FERRANDO MONSONÍS J, GALLEGO OLSINA G I, ALASTRUEY LASHERAS B, et al. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer[C]//Proceedings of the ACL, EMNLP, 2022: 8756-8769.

[8] GUERREIRO N M, COLOMBO P, PIAN TANIDA P, et al. Optimal transport for unsupervised hallucination detection in

neural machine translation[C]//Proceedings of the ACL, 2023: 13766-13784.

[9] 陈康, 刘尧舜, 李茂西, 等. 基于多语义空间的机器译文质量估计[J]. 中文信息学报, 2025, 39(5): 72-81.

[10] 翟煜锦, 李培芸, 项青宇, 等. 基于 QE 的机器翻译重排序方法研究[J]. *江西师范大学学报(自然科学版)*, 2020, 44(1): 46-50.

[11] 刘媛, 李茂西, 罗琪, 等. 基于神经网络的机器译文自动评价综述[J]. 中文信息学报, 2023, 37(9): 1-14.

[12] 支思威, 李茂西, 吴水秀, 等. 深层差异特征增强的机器翻译自动评价[J]. *中文信息学报*, 2024, 38(10): 46-53.

[13] FENG F, YANG Y, CER D, et al. Language-agnostic BERT sentence embedding[C]//Proceedings of the ACL, 2022: 878-891.

[14] HEFFERNAN K, ÇELEBI O, SCHWENK H. Bitext mining using distilled sentence representations for low-resource languages[C]//Proceedings of the ACL, EMNLP, 2022: 2101-2112.

[15] CONNEAU A, RINOTT R, LAMPLE G, et al. XNLI: Evaluating cross-lingual sentence representations[C]//Proceedings of the EMNLP, 2018: 2475-2485.

[16] GUERREIRO N M, REI R, VAN STIGT D, et al. Xcomet: Transparent machine translation evaluation through fine-grained error detection[J]. *Advances in TACL*, 2024(12): 979-995.

[17] DALE D, VOITA E, BARRAULT L, et al. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better[C]//Proceedings of the ACL, 2023: 36-50.

[18] HUANG C, GHADDAR A, KOBAYASHI I, et al. OTTAWA: Optimal transport adaptive word aligner for hallucination and omission translation errors detection[C]//Proceedings of the ACL, 2024: 6322-6334.

[19] HUANG L, YU W, MA W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[J]. *Advances in ACM Transactions on Information Systems*, 2025, 43(2): 1-55.

[20] XU Z, JAIN S, KANKANHALLI M. Hallucination is inevitable: An innate limitation of large language models[J]. *arXiv preprint arXiv*: 2401.11817, 2024.

- [21] GUO Z, SCHLICHTKRULL M, VLACHOS A. A survey on automated fact-checking[J]. *Advances in TACL*, 2022(10): 178-206.
- [22] HANSELOWSKI A, STAB C, SCHULZ C, et al. A richly annotated corpus for different tasks in automated fact-checking[C]//*Proceedings of the CoNLL*, 2019: 493-503.
- [23] ATANASOVA P, SIMONSEN J G, LIOMA C, et al. Generating fact checking explanations[C]//*Proceedings of the ACL*, 2020: 7352-7364.
- [24] HUO S, ARABZADEH N, CLARKE C L A. Retrieving supporting evidence for LLMs generated answers[J]. *arXiv preprint arXiv: 2306.13781*, 2023.
- [25] XU W, AGRAWAL S, BRIAKOU E, et al. Understanding and detecting hallucinations in neural machine translation via model introspection[J]. *Advances in TACL*, 2023(11): 546-564.
- [26] BINDER A, MONTAVON G, LAPUSCHKIN S, et al. Layer-wise relevance propagation for neural networks with local renormalization layers[C]//*Proceedings of the ICANN*, 2016: 63-71.
- [27] ZHOU C, NEUBIG G, GU J, et al. Detecting hallucinated content in conditional neural sequence generation[C]//*Proceedings of the IJCNLP*, 2021: 1393-1404.
- [28] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. *arXiv preprint arXiv: 1910.13461*, 2019.
- [29] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]//*Proceedings of the ACL*, 2020: 8440-8451.
- [30] DALE D, VOITA E, BARRAULT L, et al. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better[C]//*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023: 36-50, Toronto, Canada. Association for Computational Linguistics.
- [31] BENKIRANE K, GONGAS L, PELLER S, et al. Machine translation hallucination detection for low and high resource languages using large language models[C]//*Proceedings of the EMNLP*, 2024: 9647-9665.
- [32] DUQUENNE P A, SCHWENK H, SAGOT B. SONAR: Sentence-level multimodal and language-agnostic representations[J]. *arXiv preprint arXiv: 2308.11466*, 2023.
- [33] CHOI C, KIM J, LEE S, et al. Linq-embed-mistral technical report[J]. *arXiv preprint arXiv: 2412.03223*, 2024.
- [34] LIU N F, LIN K, HEWITT J, et al. Lost in the middle: How language models use long contexts[J]. *Advances in TACL*, 2024, 12: 157-173.
- [35] QIN M. The uniqueness of LLaMA3-70B series with per-channel quantization[J]. *arXiv preprint arXiv: 2408.15301*, 2024.
- [36] CHENG J, LIU X, ZHENG K, et al. Black-box prompt optimization: aligning large language models without model training[C]//*Proceedings of the ACL*, 2024: 3201-3219.
- [37] LIU Y, FABBRI A, LIU P, et al. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation[C]//*Proceedings of the ACL*, 2023: 4140-4170.
- [38] 赵阳, 周龙, 王迁, 等. 民汉稀缺资源神经机器翻译技术研究 [J]. *江西师范大学学报 (自然科学版)*, 2019, 43(6): 630-637.



魏嘉琴(2001—), 硕士, 主要研究领域为自然语言处理和机器翻译。

E-mail: weijiaqin0709@jxnu.edu.cn



李茂西(1977—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理和机器翻译。

E-mail: mosesli@jxnu.edu.cn