引入反思机制的机器译文质量估计方法

万洁　李茂西*
1.江西师范大学，计算机信息工程学院，江西省，南昌市
2.江西师范大学，管理科学与工程研究中心，江西省，南昌市
{202341600070, mosesli}@jxnu.edu.cn

摘要

在缺乏人工参考译文对照的情况下，如何自动地评估机器译文的质量？现有一种机器译文质量估计方法利用异构翻译系统对源语言句子进行直接翻译，把生成的译文作为伪参考译文，将机器译文和伪参考译文进行对比来评估机器译文的质量。为了使生成的伪参考译文能够帮助机器译文质量估计方法准确地识别当前机器译文中存在的错误，本文提出引入反思机制的伪参考译文生成方法，并将其应用在机器译文质量估计任务中。生成伪参考译文的异构翻译系统是一个反思智能体，该反思智能体将待评估机器译文作为生成伪参考译文过程中的关键元素，它的推理步骤包括对机器译文进行回译、对源语言句子和回译进行智能反思、基于反思结果生成对机器译文的修正意见以及生成候选伪参考译文。在WMT'23句子级别机器译文质量估计任务基准数据集上的实验结果表明，所提方法显著提高了机器译文质量估计的效果。

关键词： 机器翻译；机器译文质量估计；大语言模型；智能体；反思

Quality Estimation of Machine Translation Based on Reflection

Jie Wan　Maoxi Li*
1.School of Computer Information Engineering, Jiangxi Normal University
Jiangxi, Nanchang, China
2.Management Science and Engineering, Jiangxi Normal University
Jiangxi, Nanchang, China
{202341600070, mosesli}@jxnu.edu.cn

Abstract

How can we automatically assess the quality of machine translation output in the absence of human reference translations? A common Quality Estimation of Machine Translation method employs heterogeneous translation systems to directly translate the source language sentence, using the resulting translation as a pseudo reference to compare with the machine translations for quality assessment. To enable pseudo reference translations to effectively assist Quality Estimation of Machine Translation methods in accurately identifying errors in the machine translations, we propose a pseudo reference generation method incorporating a reflection mechanism, applied to the Quality Estimation of Machine Translation task. The heterogeneous translation

*为通讯作者
基金项目：国家自然科学基金(62366020)

system is implemented as a reflection-based agent, which integrates the machine translations under evaluation as a critical component in the pseudo reference generation process. It's inference process includes four key steps: generating a back translation from the MT; performing intelligent reflection by comparing the source language sentence and the back translation; producing correction suggestions for the machine translations based on reflection outcomes; and generating candidate pseudo reference translations. Experimental results on the WMT'23 sentence-level Quality Estimation of Machine Translation benchmark dataset demonstrate that the proposed method significantly enhances Quality Estimation of Machine Translation performance.

Keywords: Machine translation , Quality estimation of machine translation , Large language model , AI Agent , Reflection

1 引言

机器译文质量估计(Quality Estimation)是指在没有人工参考译文的情况下，评估机器翻译系统输出的机器译文质量 (翟煜锦 et al., 2020)。研究机器译文质量估计方法对机器翻译系统的开发完善以及机器译文的后编辑等具有重要的实践意义 (Huang et al., 2022)。

一般的机器译文质量估计方法利用传统机器学习或深度学习技术从源语言句子和机器译文构成的二元组中提取反映翻译质量的特征，然后利用支持向量机 (Mariano Felice and Lucia Specia, 2012)或循环神经网络(Recurrent Neural Network, RNN)预测机器译文质量。其中典型代表包括Specia等人 (Specia et al., 2013)提出的QuEst模型，该模型利用外部语言资源 (赵阳 et al., 2019)或语言学分析工具从复杂度、流利度、忠实度和置信度共四大特征对源语言句子和机器译文进行特征提取，在模型构建时采用支持向量回归器、随机森林等算法评估机器译文质量；Kim等人 (Kim et al., 2017)提出的预测器-估计器模型(Predictor-Estimator)利用基于双向RNN带注意力机制 (Bahdanau et al., 2014)的预测器，从源语言句子和机器译文中提取翻译质量的特征，而后利用RNN模型预测机器译文质量；Fan等人 (Fan et al., 2019)提出的双语专家模型(Bilingual Expert)利用Transformer模型 (Vaswani et al., 2017)构建预测器进行特征提取，并通过基于双向长短期记忆网络构建的估计器来预测机器译文的质量分数。随着预训练语言模型的发展，学者们提出了基于预训练语言模型的机器译文质量估计方法。Ranasinghe等人 (Ranasinghe et al., 2020)提出了一种基于跨语言预训练模型的机器译文质量估计框架TransQuest，该框架中的MonoTransQuest模型将源语言句子和机器译文进行拼接，将拼接后的内容输入至跨语言预训练模型XLM-R (Conneau et al., 2020)中获取词向量，随后通过softmax层对机器译文进行质量评估。但是由于该方法没有对源语言句子和机器译文之间的语义对应关系进行直接建模，其估计结果存在一定的翻译忠实度偏见。针对这一问题，Huang等人 (Huang et al., 2023)提出了ConRegQE模型，该模型利用对比学习作为正则技术，通过创建有效的负样本并聚焦于双语语义对齐，从而缓解机器译文质量估计任务中的翻译忠实度偏见，同时提升模型性能。

在缺乏人工参考译文对照的情况下，另一类机器译文质量估计方法利用机器翻译系统或对话大语言模型对源语言句子进行重新翻译，把生成的翻译作为伪参考译文(Pseudo Reference Translation)，将机器译文质量估计任务转换为机器译文自动评价任务(Automatic Evaluation)，或从源语言句子、机器译文和伪参考译文构成的三元组中提取反映翻译质量的特征，进而对机器译文质量进行自动评估。其中Wang等人 (Wang et al., 2020)提出的PEAQE模型采用在线翻译系统生成基于源语言句子的伪参考译文，将伪参考译文与源语言句子进行拼接编码，用于提取质量特征，最后实现机器译文的质量预测；Cui等人 (Cui et al., 2021)和Geng等人 (Geng et al., 2023)通过生成器构建伪机器译文质量估计的训练数据，用于训练基于神经网络的机器译文质量估计模型。

然而，前人基于伪参考译文的机器译文质量估计方法在创建伪参考译文的过程中完全忽略了与待评估机器译文的联系，仅对源语言句子进行翻译，这导致其生成的伪参考译文不能准确地为识别当前待评估机器译文中存在的翻译错误和语法错误等提供指导和对照。针对这个问题，本文提出了引入反思机制的机器译文质量估计方法(Quality Estimation of Machine

Translation Based on Reflection, ReflectQE)。该方法利用当前大语言模型链式提示中的反思(Reflection)，将待评估机器译文作为大语言模型生成伪参考译文过程中的一环，从而构建与待评估机器译文密切相关的伪参考译文。随后，采用多视角表征融合的方法对机器译文的质量进行定量评估。本文所提方法在WMT'23句子级别机器译文质量估计评测任务的基准数据集上，与参与测评的系统和当前最先进的机器译文质量估计方法进行了对比，实验结果表明引入反思机制的机器译文质量估计方法显著提高了与人类评价的相关性。

2 知识背景

2.1 基于伪参考译文的机器译文质量估计

"对于同一源语言句子，若两个独立的翻译系统生成的译文A和B相似，那么译文A的质量通常较高。"，Soricut等人 (Radu Soricut and Sushant Narsale, 2012)在此假设的基础上提出了基于伪参考译文的机器译文质量估计方法。随后Shah等人 (Shah et al., 2013)通过对多个数据集进行句子级别的特征分析，进一步验证了伪参考译文在机器译文质量估计任务上的有效性。Scarton等人 (Carolina Scarton and Lucia Specia, 2014)通过引入字符串相似度度量，对基于伪参考译文的机器译文质量估计方法进行优化。为了进一步凸显伪参考译文在机器译文质量估计任务中的优势，Duma等人 (Melania Duma and Wolfgang Menzel, 2018)利用在线机器翻译系统生成的伪参考译文和回译(Back Translation)，对机器译文进行表征提取，并将其引入传统机器学习核函数中，从而提高机器译文质量估计方法的性能。

2.2 基于反思方法的机器翻译智能体

近年来，大语言模型(Large Language Models, LLMs)，如GPT (Brown et al., 2020)、Llama (Grattafiori et al., 2024)和DeepSeek (Liu et al., 2024; Guo et al., 2025)等，由于其卓越的生成能力，显著推动了自然语言处理领域的发展。而以LLMs为核心构建的智能体(AI Agent)，凭借其集成多步骤推理与任务导向流程的特性，进一步拓展了其应用潜力。

基于智能体的这一特性，许多学者开始探索如何将大语言模型的动态推理能力应用于机器翻译领域，以提升翻译的质量和准确性。吴恩达提出的翻译智能体(Translation Agent)，通过大语言模型自动分析源语言句子与机器译文之间的差异，并基于该差异改进对源语言句子的翻译，从而显著提升翻译的质量。翻译智能体的核心在于其独特的"反思工作流"，该反思流模仿了人类翻译专家的思考过程，将翻译任务分解为初始翻译、反思与改进以及优化输出三个步骤。但是由于该反思流是基于跨语言得到的反思，存在一定的语义偏差。Chen等人 (Chen et al., 2024)提出了双重反思(Dual Reflect)，该反思流在获得源语言句子初始翻译后，使用回译的方式将源语言句子和机器译文保持在同一语言空间，克服跨语言存在的语义偏差，并利用LLMs反思回译与源语言句子之间的差异以获得反馈信号，最终优化翻译结果。

上述机器翻译的方法仅从源语言句子出发，其生成的翻译可能无法有效地应用于机器译文质量估计任务中。为此我们提出改进方案：将反思机制引入机器译文质量估计任务中，同时使用机器译文作为生成伪参考译文过程的起始，克服前人工作中伪参考译文与待评估机器译文关联性不足的问题。一方面该方法可以提高伪参考译文质量，另一方面生成与待评估机器译文密切相关的伪参考译文能够指导和辅助机器译文质量估计。

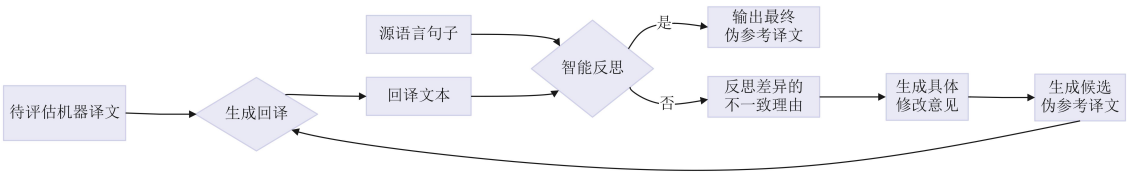

Figure 1: 反思智能体的迭代流程

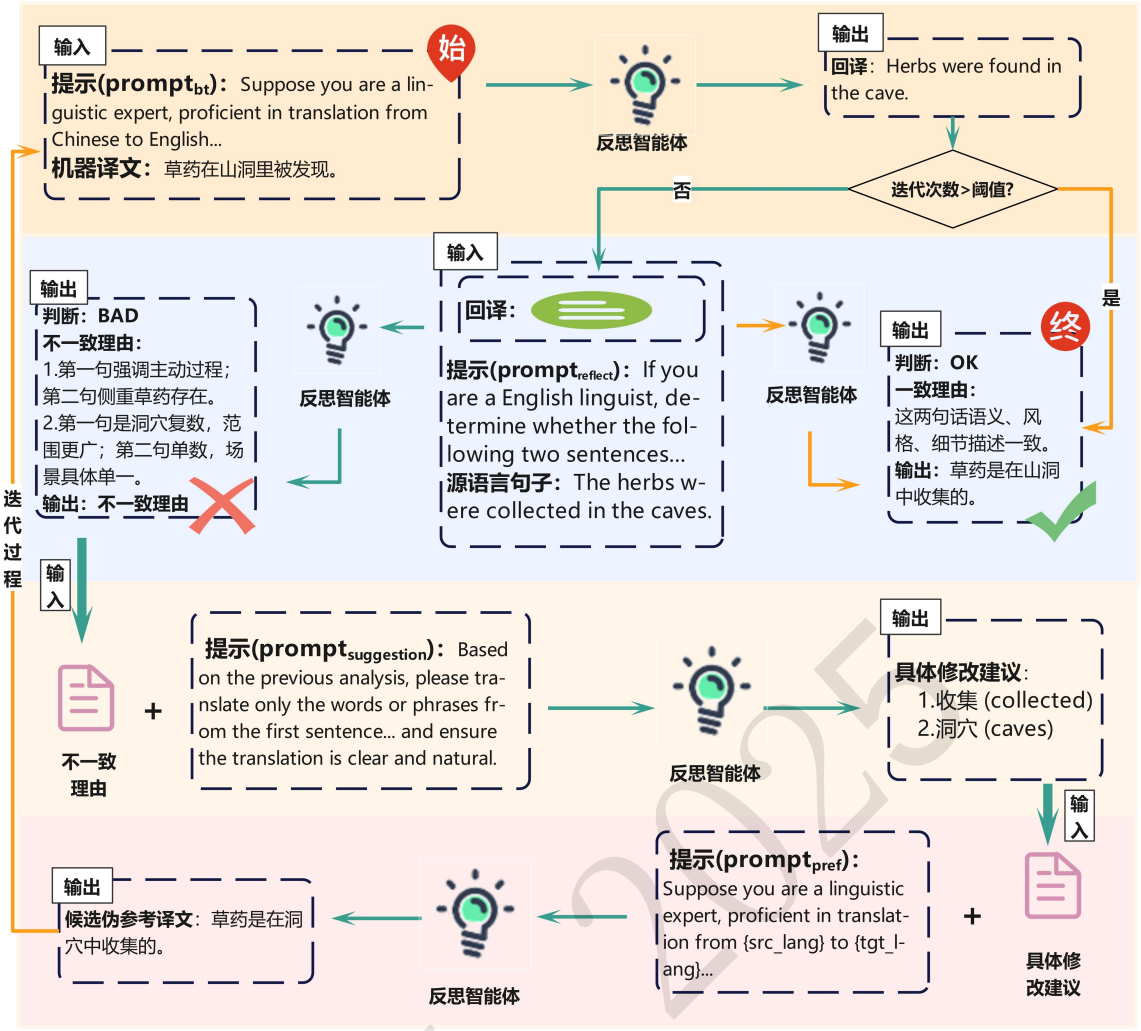

Figure 2: 反思智能体的工作流程

3 引入反思机制的机器译文质量估计方法

引入反思机制的机器译文质量估计方法主要分为两个部分：利用反思机制生成伪参考译文部分和多视角表征融合部分。在利用反思机制生成伪参考译文部分，为了得到高质量且与待评估机器译文高度契合的伪参考译文，我们构建了一个基于智能反思的机器翻译智能体：反思智能体，其总体流程框架如图1所示。该方法的核心在于智能反思，通过动态分析源语言句子和回译之间的语义、风格特征以及对齐细节的一致性，实现对机器译文的自适应优化。我们将生成伪参考译文的过程分解为以下四个步骤：由机器译文生成回译、对源语言句子和回译进行智能反思、基于反思结果对机器译文提出修正建议以及生成候选伪参考译文。上述过程将循环迭代直至生成的候选伪参考译文质量满足预定要求或循环次数达到预定上限，从而得到最终的伪参考译文。反思智能体能根据上下文反馈自适应地调整生成策略，从而逐步提升语义一致性。在多视角表征融合部分，我们将反思智能体得到的伪参考译文，与源语言句子和机器译文相结合，进行多视角表征融合，得到输出的质量分数。

3.1 利用反思智能体生成伪参考译文

反思智能体是一个循环迭代的推理流程，主要包括对机器译文进行回译、对源语言句子和回译进行智能反思、基于反思结果生成对机器译文的修正意见以及生成候选伪参考译文四个步骤，其详细流程框架可见图2。

在“对机器译文进行回译”步骤中，为了使生成的伪参考译文与待评估机器译文密切相关，从而为识别待评估机器译文中的错误提供有效对照，我们利用反思智能体，通过回译提示引导大语言模型生成机器译文的回译。如式(1)所示：

$$bt=Agent(prompt_{bt}, mt) \tag{1}$$

其中$Agent$为反思智能体，$prompt_{bt}$为回译提示，mt为机器译文，bt为生成的回译。该步骤与吴恩达等提出的翻译智能体中的从源语言句子出发进行翻译步骤有显著差异。

在“对源语言句子和回译进行智能反思”步骤中，前一步骤生成的回译使得源语言句子与回译置于同一语言空间，便于直接进行语义对比。相较于待评估机器译文与源语言句子的跨语言对比可能忽略的潜在语义偏差或翻译错误，源语言句子与回译的语义分析能够更清晰地揭示这些差异。反思智能体使用设计好的智能反思提示引导大语言模型对回译和源语言句子进行多维度的一致性判断，其中包括语义、风格以及对齐细节三个维度。反思智能体根据判断结果分两种情况处理：如果判断为不一致(如图2中“BAD”表示未传达一致语义)，则反思智能体需要输出源语言句子和回译之间产生差异的不一致理由，该理由需包含具体某方面的不一致、对源语言句子的反思分析、对回译的反思分析以及总结反思分析，为后续输出具体的修正意见提供精准指导。如式(2)所示：

$$judge, reason=Agent(prompt_{reflect}, src, bt) \tag{2}$$

其中src为源语言句子，$prompt_{reflect}$为智能反思提示，$judge$为一致性判断的结果，$reason$为判断后总结的不一致理由。如果判断结果为一致(如图2中“OK”表示传达一致语义)，则终止“反思”阶段，得到最终的伪参考译文。

在“基于反思结果生成对待评估机器译文的修正建议”步骤中，此时一致性判断结果为不一致，根据上一步所提供的不一致理由，反思智能体利用修正建议提示驱动大语言模型对源语言句子与回译之间的差异进行深入分析，该分析包括识别两者之间的不同和不合理之处，明确将差异细节落实到句子的具体位置，最后得到修正建议的结果，如式(3)所示：

$$suggestion=Agent(prompt_{suggestion}, reason) \tag{3}$$

其中$prompt_{suggestion}$为生成修正建议提示，$suggestion$为对待评估机器译文的具体修正建议。

在“生成候选伪参考译文”步骤中，结合机器译文修正得到的具体修改建议，反思智能体由源语言种类朝目标语言种类的方向，对源语言句子进行重新翻译，得到候选伪参考译文。如图2中，通过分析源语言句子和回译之间的不一致理由，提出了两处具体的修正建议：“收集($collected$)”和“洞穴($caves$)”。根据这些修正建议，反思智能体将“$found$”替换为“$collected$”，以更好地体现主动动作的语义；同时将“$cave$”改为复数形式“$caves$”，以匹配范围的表达。基于这些修正建议，反思智能体重新翻译源语言句子，生成了候选伪参考译文“草药是在洞穴中收集的。”，如式(4)所示：

$$pref=Agent(prompt_{pref}, src, suggestion) \tag{4}$$

其中$prompt_{pref}$为生成候选伪参考译文提示，$pref$为候选伪参考译文。

为了获得高质量的伪参考译文，反思智能体将生成的候选伪参考译文继续输入至步骤一进行迭代，直至候选伪参考译文的回译与源语言句子语义一致或迭代次数超过设置的阈值。

我们在中英翻译的方向上抽取了一个实例说明利用反思智能体生成伪参考译文的工作流程，如图2所示。在“对机器译文进行回译”步骤中由机器译文“草药在山洞里被发现。”生成回译“*Herbs were found in the cave.*”，经由“对源语言句子和回译进行智能反思”步骤中得到对源语言句子和回译“BAD”的判断，并得到不一致理由，其中包含侧重点不同和单词使用不准确。再到第三步“基于反思结果生成伪参考译文的修正意见”步骤中将理由落实至回译的“收集($collected$)”和“洞穴($caves$)”两处具体位置，并将其输入至“生成候选伪参考译文”步骤中得到候选伪参考译文。随后进入迭代过程，将候选伪参考译文重新回译、智能反思，此时的判断为“OK”，则输出该候选伪参考译文作为最终的伪参考译文。

第二十四届中国计算语言学大会论文集，第237页-第246页，济南，中国，2025年8月11日至14日。

3.2 多视角表征融合

为了充分考虑全局语义信息，我们结合源语言句子和机器译文之间的表征、机器译文和伪参考译文之间的差异表征以及源语言句子、机器译文、伪参考译文之间的交互表征进行多视角表征融合。具体而言，我们以句对的形式构造三种序列($src+mt$、$mt+pref$、$src+mt+pref$)，并将其分别输入预训练语言模型提取统一质量特征向量,各特征向量再通过前馈神经网络层进行预测质量得分，生成的三个质量分数$score_{src+mt}$、$score_{mt+pref}$、$score_{src+mt+pref}$通过线性融合得到多视角表征融合的质量分数。

$$QEscore = \alpha_1 \times score_{src+mt} + \alpha_2 \times score_{mt+pref} + (1-\alpha_1-\alpha_2) \times score_{src+mt+pref} \quad (5)$$

其中α_1，α_2为线性插值系数，介于0与1之间，用于平衡三种视角的表征信息所预测的质量评估分数。

为了训练模型的参数，我们在训练集上使用均方差损失进行优化模型目标:

$$loss = \frac{1}{N}\sum_{i=1}^{N}\left(score^{(i)} - h^{(i)}\right)^2 \quad (6)$$

其中$score^{(i)}$为多视角表征融合部分中各视角对待评估机器译文的评分，$h^{(i)}$为人类评价的结果，N为训练集包含的样本数量。

最后，我们结合Sato等人 (Sato et al., 2024)提出的TMU-HIT[1]系统，它采用手动设计的提示驱动经过微调后的GPT-4o-mini模型进行多次百分制的评分，通过对多个评分进行加权处理后得到质量分数，进一步提高了我们系统的性能。

4 实验

在实验部分，我们旨在回答以下研究问题:

1)所提方法在质量估计任务中的不同语言对上表现如何?

2)智能反思机制相较于其他方法的优越性体现在哪里?

3)不同大语言模型所生成的伪参考译文对所提方法的影响如何?

4)保留分数的分布范围对方法的性能有怎样的影响?

为了回答问题1，实验在WMT'23的机器译文质量估计任务中的五个语言对上测试了所提方法的性能，具体实验可见4.2节；为探究问题2，实验对仅使用大语言模型得到的伪参考译文、其他反思相关系统得到的伪参考译文以及本文引入反思机制得到的伪参考译文，三者的实验性能进行对比，具体实验可见4.3节的第一部分；对于问题3，实验分别尝试大语言模型GPT-4o-mini和DeepSeek-V3生成不同的伪参考译文，对比两者的性能结果，具体实验可见4.3节的第二部分；最后，为分析问题4，实验比较了原始分数与缩放后分数对相关系数的影响，具体实验可见4.3节的第三部分。

4.1 实验设置

为了验证本文提出的引入反思机制的机器译文质量估计方法效果，我们在WMT'23句子级别质量估计评测任务中的数据集上进行了实验验证。系统性能测量方法遵循WMT评测官方的评分标准，使用斯皮尔曼相关系数作为计算机器译文质量估计结果与人类评价结果的主要测评指标，其中斯皮尔曼系数越大，相关性越好。在WMT'23句子级别质量估计评测任务中包括五个翻译方向的翻译质量评估，即英文-马拉地语(En-Mr)、英语-印地语(En-Hi)、英语-泰米尔语(En-Ta)、英语-泰卢固语(En-Te)以及英语-古吉拉特语(En-Gu)。

在引入反思机制的机器译文质量估计方法中，反思智能体使用大语言模型"GPT-4o-mini"作为核心模型，最大迭代次数阈值设置为2。在多视角表征融合部分中，使用预训练语言模型"mdeberta-v3-base"对三种视角进行表征向量提取，子词向量维度为768。在模型微调时使用AdamW优化器，初始学习率设置为2e-5，训练批次大小为32。在多视角的分数线性融合公式(5)中，α_1和α_2分别设置为0.5，0.3。

实验将本文提出的方法与WMT'23句子级别机器译文质量估计任务中比较流行的方法：Cometkiwi (Rei et al., 2023)、CrossQE (Li et al., 2023)、IOL Research (Zeyu Yan ,

[1]https://colab.research.google.com/drive/1p8VMnAkRfuVpbvM$_revV2ZaN76sSxmiE?usp=sharing/$

模型	斯皮尔曼相关系数					
	En-Mr	En-Hi	En-Ta	En-Te	En-Gu	Avg.
CrossQE	0.692	**0.644**	**0.775**	**0.394**	0.691	**0.639**
CometKiwi	**0.704**	0.598	0.739	0.388	**0.714**	0.628
IOL Research	0.505	0.600	0.740	0.376	0.695	0.583
EnsembleTQ	0.596	0.551	0.674	0.349	0.649	0.563
MMT	0.650	0.494	0.547	0.337	0.540	0.514
TMU-HIT	0.676	0.614	0.726	0.354	0.674	0.609
TransQuest	0.673	0.551	0.660	0.329	0.628	0.568
CometKiwi-da	0.605	0.385	0.577	0.276	0.559	0.481
ReflectQE	0.694	**0.656**	**0.797**	**0.409**	0.700	**0.651**

Table 1: 不同机器译文质量估计方法在WMT'23句子级别机器译文质量估计数据集上的性能

2023)、EnsembleTQ (Sindhujan et al., 2023)和MMT (Wu et al., 2023)，同时与基于预训练语言模型的经典方法TransqQuest，以及在WMT'22句子级别机器译文质量估计任务上较为出色的模型CometKiwi-da (Rei et al., 2022)，在WMT'23句子级别机器译文质量估计任务数据集上进行实验并比较性能差异。

4.2 实验结果

所提方法的性能在WMT'23句子级别机器译文质量估计任务数据集的五个语言对上得到验证。实验结果如表1所示，引入反思机制的机器译文质量估计方法ReflectQE在平均相关系数表现最佳，达到0.651，相较参加WMT'23句子级别机器译文质量估计任务评测的最优模型CrossQE的平均相关系数提高了0.012(0.639 VS 0.651)。这一增益在三个语言对上尤为显著：En-Hi、En-Ta、En-Te，ReflectQE的相关系数分别比相应语言对的最优模型CrossQE提高了1.2 %、2.2 %、1.5 %，显示出其在多语言对场景下的优越性。与基于预训练语言模型的经典方法TransQuest相比，ReflectQE的平均性能提高了0.083(0.568 VS 0.651)。相较在WMT'22句子级别机器译文质量估计任务上表现出色的模型CometKiwi-da，性能增幅达到0.170(0.481 VS 0.651)。这些结果表明，本文提出的引入反思机制的机器译文质量估计方法有效地增强了机器译文质量估计的精度，显著提高了评估的全面性和鲁棒性，验证了该方法在机器译文质量估计任务中的应用潜力。

4.3 消融实验

(1)为了研究本文提出的反思智能体相较其他方法的优越性，我们设计了两个维度的对比实验：一方面是反思机制的有无，另一方面是反思细节处理的差异。

在第一个维度中，实验尝试对比仅利用大语言模型直接生成的伪参考译文与引入反思智能体推理流程生成的伪参考译文的性能表现。为了确保实验的公平性，大语言模型的选用与反思智能体使用的GPT-4o-mini保持一致。

实验结果可见表2，与直接利用大语言模型生成伪参考译文的质量估计方法DirectQE相比，ReflectQE在每个语言对上都有不同程度的提高，平均性能提高了0.027(0.624 VS 0.651)，这验证了反思机制的有效性。

在第二个维度中，实验关注点在反思细节的处理差异，实验使用吴恩达提出的翻译智能体生成的伪参考译文作为对比基准，该方法通过模型跨语言自动分析源语言句子和机器译文，并基于分析结果改进自身输出，从而得到最终的生成伪参考译文，而我们提出的反思智能体不仅分析了机器译文质量，同时深入分析了同一语言空间下的源语言句子和回译的语义一致性，能够发现在跨语言对比时容易忽略的语义偏差，生成更加细节、贴合待评估机器译文的改进建议。

实验结果可见表2，与使用翻译智能体生成伪参考译文的机器译文质量估计方法TranslationQE相比，ReflectQE在所有语言对上都有不同程度的提高，平均性能提高了0.019(0.632 VS 0.651)，这进一步证明了本文提出的反思智能体的有效性。

(2)为了探究不同大语言模型所生成的伪参考译文对所提方法的影响，我们分别尝试使

基础质量估计方法	斯皮尔曼相关系数					
	En-Mr	En-Hi	En-Ta	En-Te	En-Gu	Avg.
DirectQE	0.681	0.624	0.753	0.377	0.684	0.624
TranslationQE	0.691	0.652	0.757	0.379	0.682	0.632
ReflectQE	**0.694**	**0.656**	**0.797**	**0.409**	**0.700**	**0.651**

Table 2: 通过不同方法生成的伪参考译文的机器译文质量估计方法与人类评价的相关性

用GPT-4o-mini模型和DeepSeek-V3模型生成伪参考译文。为确保实验的公平性，除了大语言模型的选择有所不同，其余实验流程、实验设置以及提示均保持一致。

表3展示了通过不同大语言模型生成伪参考译文的机器译文质量估计方法与人类评价的相关性，结果表明使用GPT-4o-mini模型生成伪参考译文的方法平均相关性高于使用DeepSeek-V3模型生成伪参考译文的方法(0.646 VS 0.651)。对不同大语言模型生成的伪参考译文自身进行分析，可以发现使用GPT-4o-mini模型生成的伪参考译文的整体质量要高于使用DeepSeek-V3模型生成的伪参考译文，机器译文质量估计方法可以从高质量的伪参考译文中提取更丰富的信息，从而提高机器译文质量估计方法的性能。

基础质量估计方法	大语言模型	斯皮尔曼相关系数					
		En-Mr	En-Hi	En-Ta	En-Te	En-Gu	Avg.
ReflectQE	DeepSeek-V3	0.691	0.655	0.793	0.397	0.693	0.646
	GPT-4o-mini	**0.694**	**0.656**	**0.797**	**0.409**	**0.700**	**0.651**

Table 3: 通过不同大语言模型生成的伪参考译文的机器译文质量估计方法与人类评价的相关性

(3)为了探究保留原始分数分布范围的计算策略是否更优这个问题，我们尝试将辅助系统TMU-HIT得到的百分制分数向多视角表征融合部分的分数对齐，将其缩放至[-3，1]的区间内，其中[-3，1]的范围是依据在WMT'23句子级别机器译文质量估计任务基准数据集中的直接评估分数(Direct Assessment)概括得到的。考虑到分数范围对最终相关系数的影响，我们统一量纲，旨在观察这种缩放是否会破坏系统间的权重平衡。将百分制系统进行线性缩放的公式如下:

$$Score_{scaled} = Score^{i}_{GPT} \times 0.04 - 3 \tag{7}$$

其中$Score^{i}_{GPT}$为原始分数，$Score_{scaled}$为放缩后的分数。

表4展示了缩放后分数计算的平均相关系数与原始分数计算的平均相关系数的实验结果。结果显示分数缩放后，平均相关系数出现了显著下降，变化幅度为0.023(0.628 VS 0.651)。对于各语言对，除En-Te语言对出现轻微上涨外，其余四个语言对出现了不同程度的下降，其中下降最明显的En-Hi变化幅度为0.053(0.603 VS 0.656)。结果表明原始分数分布中的百分制分数具有更广的动态范围，能更敏感地反映翻译质量的细微差异，而强制缩放会压缩其信息量，破坏模块间的权重平衡，导致性能下降。

基础质量估计方法	斯皮尔曼相关系数					
	En-Mr	En-Hi	En-Ta	En-Te	En-Gu	Avg.
ScoreScaled	0.682	0.603	0.782	**0.412**	0.693	0.628
ReflectQE	**0.694**	**0.656**	**0.797**	0.409	**0.700**	**0.651**

Table 4: 通过不同分数分布计算引入反思机制的机器译文质量估计方法与人类评价的相关性

5 结论与展望

本文提出了引入反思机制的机器译文质量估计方法，该方法构建了以大语言模型为核心的反思智能体，具备深层次的语义一致性分析能力，能够生成高质量且与机器译文紧密契合的伪参考译文，并将伪参考译文、源语言句子和机器译文三者相结合，进行多视角表征融合。实验

结果表明该方法不仅增强了伪参考译文的可解释性和使用性，还提升了机器译文质量估计任务的评估精度。

在未来工作中，我们计划探索更加科学合理的提示推理流程，以生成更高质量的伪参考译文，并采用更为灵活智能的实验工具，旨在从多个维度对质量估计进行更为全面和深入的评估。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2014. *Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:* 1409.0473.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal and Arvind Neelakantan, et al. 2020. *Language models are few-shot learners. Advances in NIPS*, 33: 1877-1901.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang and Tiejun Zhao, et al. 2024. *DUAL-REFLECT: Enhancing Large Language Models for Reflective Translation through Dual Learning Feedback Mechanisms. Proceedings of the ACL:* 693-704.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán and Edouard Grave, et al. 2020. *Alibaba-Translate China's Submission for WMT 2022 Quality Estimation Shared Task. Proceedings of the ACL:* 8440–8451.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang and Jiajun Chen. 2021. *Directqe: Direct pretraining for machine translation quality estimation. Proceedings of the AAAI*, 35(14): 12719-12727.

Melania Duma and Wolfgang Menzel. 2018. *The benefit of pseudo-reference translations in quality estimation of mt output. Proceedings of the WMT:* 776-781.

Kai Fan,Jiayi Wang, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen and Luo Si. 2019. *"Bilingual Expert" can find translation errors. Proceedings of the AAAI*, 33(01): 6367-6374.

Mariano Felice and Lucia Specia. 2012. *Linguistic features for quality estimation. Proceedings of the WMT:* 96-103.

Xiang Geng, Yu Zhang, Jiahuan Li, Hao Yang, Shimin Tao, Yimeng Chen and Ning Xie, et al. 2023. *Denoising pre-training for machine translation quality estimation with curriculum learning. Proceedings of the AAAI*, 37(11): 12827-12835.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle and Aiesha Letman, et al. 2024. *The llama 3 herd of models. arXiv preprint arXiv:* 2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu and Qihao Zhu, et al. 2025. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:* 2501.12948.

Hui Huang, Hui Di, Chunyou Li, Hanming Wu, Kazushige Ouchi, Yufeng Chen and Jian Liu, et al. 2022. *BJTU-Toshiba's Submission to WMT22 Quality Estimation Shared Task. Proceedings of the WMT:* 621-626.

Hui Huang, Shuangzhi Wu, Kehai Chen, Hui Di, Muyun Yang and Tiejun Zhao. 2023. *Improving translation quality estimation with bias mitigation. Proceedings of the ACL:* 2175-2190.

Hyun Kim,Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee and Seung-Honn Na. 2017. *Predictor-estimator: neural quality estimation based on target word prediction for machine translation. ACM TALLIP*, 17(1): 1-22.

Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang and Hao Yang. 2023. *Hw-tsc 2023 submission for the quality estimation shared task. Proceedings of the WMT:* 835-840.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu and Chenggang Zhao, et al. 2024. *Deepseek-v3 technical report. arXiv preprint arXiv:* 2412.19437.

Tharindu Ranasinghe, Constantin Orasan and Ruslan Mitkov. 2020. *TransQuest: Translation Quality Estimation with Cross-lingual Transformers. Proceedings of the COLING:* 5070-5081.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti and José G. C. de Souza, et al. 2022. *CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. Proceedings of the WMT:* 634-645.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur and José G. C. de Souza, et al. 2023. *Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. Proceedings of the WMT:* 841-848.

Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2024. *TMU-HIT's Submission for the WMT24 Quality Estimation Shared Task: Is GPT-4 a Good Evaluator for Machine Translation? Proceedings of the WMT:* 529-534.

Carolina Scarton and Lucia Specia. 2014. *Document-level translation quality estimation: exploring discourse and pseudo-references. Proceedings of the EAMT:* 101-108.

Kashif Shah, Trevor Conn, and Lucia Specia. 2013. *An investigation on the effectiveness of features for translation quality estimation. Proceedings of the MT Summit XIV:* Papers.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Tharindu Ranasinghe. 2023. *SurreyAI 2023 Submission for the Quality Estimation Shared Task. Proceedings of the WMT:* 849-855.

Radu Soricut and Sushant Narsale. 2012. *Combining quality prediction and system selection for improved automatic translation output. Proceedings of the WMT:* 163-170.

Lucia Specia, Kashif Shah, Jose G C de Souza and Trevor Cohn. 2013. *QuEst-A translation quality estimation framework. Proceedings of the ACL:* 79-84.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez and Łukasz Kaiser, et al. 2017. *Attention is all you need. Advances in NIPS:* 30.

Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei and Ying Qin, et al. 2020. *HW-TSC's participation at WMT 2020 quality estimation shared task. Proceedings of the WMT:* 1056-1061.

Yulong Wu, Viktor Schlegel, Daniel Beck, and Riza Batista-Navarro. 2023. *Mmt's submission for the wmt 2023 quality estimation shared task. Proceedings of the WMT:* 856-862.

Zeyu Yan. 2023. *Iol research's submission for wmt 2023 quality estimation shared task. Proceedings of the WMT:* 863-871.

翟煜锦, 李培芸, and 项青宇, et al. 2020. 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*)*, 44(1): 46-50.

赵阳, 周龙, and 王迁, et al. 2019. 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), 43(6): 630-637.