中国中文信息学会
中国科学院软件研究所 主办

ISSN 1003-0077
CN 11-2325/N
CODEN ZXXHAU

JOURNAL OF CHINESE INFORMATION PROCESSING

中国中文信息学会会刊

中文信息学报

ZHONGWEN XINXI XUEBAO

第39卷 第7期
Vol.39 No.7

2025.7

中文信息学报(月刊)
第 39 卷,第 7 期　2025 年 7 月

目　　次

(本期责任编辑：来雨轩　杨沐昀)

期刊基本参数：CN11-2325/N * 1986 * b * 16 * 172 * zh * P * 50.00 * 800 * 15 * 2025-07

JOURNAL OF CHINESE INFORMATION PROCESSING
Vol.39 No.7 July 2025

CONTENTS

(Executive Editor: LAI Yuxuan YANG Muyun)

文章编号：1003-0077(2025)07-0148-14

面向语言学习者的跨语言反馈评语生成方法

安纪元[1,2]，朱 琳[1,2]，杨尔弘[1,2]

(1. 北京语言大学 国家语言资源监测与研究平面媒体中心，北京 100083；
2. 北京语言大学 信息科学学院，北京 100083)

摘 要：反馈评语生成是近年来自然语言处理研究的一个热点任务，旨在为语言学习者的作文提供纠偏及解释性的评价，以帮助学习者理解并内化语言规则，从而提高写作水平。现有研究主要聚焦于单一语言的反馈评语生成，忽略了非母语学习者可能面临的理解障碍，以及评语中存在陌生语言知识等问题。该文提出了一种新的跨语言反馈评语生成(CLFCG)任务，其目的是为汉语母语者学习英语提供汉语的反馈评语。首先，通过构建首个英-汉跨语言反馈评语数据集，探索了大语言模型(如 GPT-4)和预训练语言模型(如 mBART、mT5)在该任务上的性能，并针对预训练语言模型，分析了修正编辑、线索词语和语法术语等附加信息对反馈评语生成效果的影响。其次，该文提出了一种基于大语言模型的评估方法，以更加准确地评估反馈评语生成效果。实验结果显示，基于微调的预训练语言模型能够更好地对齐人类教师的评语，但其生成的准确性略逊于采用少样本学习策略的 GPT-4 模型。最后，该文对实验结果进行了深入讨论和分析，以期为跨语言反馈评语生成任务提供更多思路和见解。

关键词：智能辅助语言学习；反馈评语生成；跨语言文本生成；预训练语言模型；大语言模型

中图分类号：TP391 **文献标识码**：A

Cross-Lingual Feedback Comment Generation for Language Learners

AN Jiyuan[1,2], ZHU Lin[1,2], YANG Erhong[1,2]

(1. National Language Resources Monitoring and Research Center for Print Media,
Beijing Language and Culture University, Beijing 100083, China;
2. School of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: Feedback comment generation has emerged as a practical demand in recent years, aiming to provide both corrective and explanatory evaluations for language learners' writings. This paper describes a Cross-Language Feedback Comment Generation (CLFCG) method, designed to provide Chinese feedback comments for Chinese native speakers learning English. First, we construct the first English - Chinese cross-lingual feedback comment dataset and examine the performance of large language models (e.g., GPT-4) and pre-trained language models (e.g., mBART, mT5) on this task. Second, we introduce a large language model-based evaluation method to assess the quality of the generated feedback. Experimental results indicate that fine-tuned pre-trained language models align more closely with human teachers' comments, although their accuracy is slightly lower than that of GPT-4 when it employs few-shot learning.

Keywords: intelligent computer-assisted language learning; feedback comment generation; cross-lingual text generation; pretrained language models; large language model

收稿日期：2024-06-03 **定稿日期**：2024-12-12

基金项目：国家语委科研项目(ZDA145-17)；教育部人文社会科学研究一般项目(23YJCZH264)；北京语言大学梧桐创新平台(中央高校基本科研业务费专项资金)项目(21PT04)；北京语言大学研究生创新基金(中央高校基本科研业务费专项资金)项目(24YCX069)

0　引言

智能辅助写作是智能计算机辅助语言学习(Intelligent Computer-Assisted Language Learning, ICALL)的一个重要研究领域，旨在帮助语言学习者纠正写作错误，提升语言运用能力。近年来，自动语法错误纠正任务(Grammatical Error Correction, GEC)受到了广泛的研究关注，其通过识别学习者文本中存在的错误，生成修正后的正确语句。该任务旨在减轻语言教师的负担，并使语言学习的过程更加即时和便捷。然而，语法错误纠正任务通常只能告诉学习者“改哪里”和“如何改”，而忽略了最为本质的问题——“为什么要修改”。这种缺乏解释的模式可能导致学习者机械性地纠正错误，却无法真正掌握错误背后的语言知识。

反馈评语生成任务(Feedback Comment Generation, FCG)正是为了弥补这一不足而设计的。与语法错误纠正任务不同，反馈评语不仅提供具体的错误纠正，还包含对句子结构、内容和语言运用的综合评价，兼具指导性与激励性。这样的解释性反馈能够更好地帮助学习者理解并内化语言规则，从而提高写作水平。

然而，现有的反馈评语生成研究大多基于单一语言，忽略了非母语学习者对反馈内容的接受度问题。如表 1 所示，母语为汉语的英语学习者更易理解汉语反馈评语，而非单纯的英语修正建议或英语反馈评语。

表 1　跨语言反馈评语生成任务

任务	语法错误纠正	单语言反馈评语生成	跨语言反馈评语生成
学习者输入	We reached to the station.		
输出示例	We reached the station. (to → to)	Since the ＜verb＞ ＜＜reach＞＞ is a ＜transitive verb＞, the ＜preposition＞ ＜＜to＞＞ is not required.	由于＜动词＞＜＜reach＞＞是＜及物动词＞，所以不需要介词＜＜to＞＞。
输出类型	修正编辑或正确语句	源语言反馈评语(英语)	目标语言反馈评语(汉语)
学习者反应	为什么要删除“to”?	“transitive verb”是什么意思?	我明白了!

为解决这一问题，本文提出了新颖的跨语言反馈评语生成任务(CLFCG)，并构建了首个针对汉语母语者学习英语的跨语言反馈评语数据集。在数据集构建过程中，本文充分利用现有的反馈评语生成资源，基于构建的语法术语翻译对照表，利用大语言模型的少样本学习能力，以无监督的方式进行评语翻译，并通过人工标注对结果进行校对。

为实现跨语言反馈评语生成，本研究分别探索了采用少样本学习策略的大语言模型和基于微调的预训练语言模型的性能，并尝试将修正编辑、线索词语及语法术语等反馈评语中的关键信息整合，作为生成模型输入的附加信息，从而起到对输出的引导和提示作用，提高预训练语言模型的生成效果。同时，本文提出了一种基于大语言模型的自动化评估方法，与传统自动评估指标和人工评估共同衡量反馈评语生成效果。实验结果表明，基于微调的预训练语言模型虽然准确率不及最先进的 GPT-4 模型，但大幅优于 GPT-3.5 模型，并且与人类评语的一致性更高，使用场景更加灵活。同时，引入不同的附加信息对评语生成效果影响显著，与直接微调的基线模型相比，加入附加信息后的模型能够生成质量更高、内容更具解释性的反馈评语。最后，本文深入讨论和分析了实验结果，以期为今后的研究提供经验和参考。

本文的主要贡献包括：

(1) 构建了首个针对汉语母语者学习英语的跨语言反馈评语数据集。

(2) 探索了大语言模型和预训练语言模型在跨语言反馈评语生成任务中的潜力。

(3) 分析了不同附加信息对预训练语言模型生成反馈评语性能的影响。

1　研究现状

智能辅助写作是计算机智能辅助语言学习领域的重要研究方向之一。然而，传统的语法错误纠正任务因缺乏可解释性而无法向语言学习者提供修改的原因，大大限制了其教育效果。因此，Nagata[1] 提出了反馈评语生成任务，旨在为语言学习者提供关于错误原因的解释性说明，从而帮助其提升写作水

平。随后,Nagata等人[2]基于亚洲英语学习者国际语料库网络(ICNALE)[3]标注反馈评语,并公开了标注细节和部分数据。该数据集采用英语和日语标注反馈评语,涵盖了多种错误类型。其中,反馈评语中的语法术语和引文采用两个特殊符号进行标记。语法术语使用"<"和">"标记(如<不及物动词>)以便于学习者对照语法书中相应的语法项目。而引用符号"<<"和">>"用于标明其中的单词是从原始学习者语句中引用的(如<<agree>>),这使得反馈评语更加灵活和具体。在INLG 2022针对语言学习者反馈评语生成的共享任务(GenChal)中,Nagata等人[4]进一步定义了反馈评语生成任务,即输入是学习者语句和需要进行评语位置的索引,输出则是针对该位置的具体反馈评语。同时,Nagata等人[5]公开了一个基于上述数据集翻译而来的、专注于介词使用错误类型的英文单语言数据集。

在反馈评语生成方法的研究中,研究者对不同的模型架构和训练策略均有所探索。Hanawa等人[6]对基于检索、直接生成以及检索与编辑结合的三种方法进行了研究,发现尽管检索与编辑结合的方法可以对检索结果进行修改,但过度编辑导致生成效果不佳;基于检索的方法虽然稳定,却因其仅能检索已有评语库中的匹配项而缺乏灵活性;相较之下,直接生成方法取得了最佳的性能,其使用序列到序列模型直接生成评语,能够提供多样化的反馈,但生成评语的准确性和可用性仍亟待提高。

为了提升模型的泛化能力和性能,研究者还探索了多种数据增强技术,以适用于低资源的错误类型。Babakov等人[7]通过使用依存句法分析对学习者语句进行裁剪,并采用语言模型GPT-Neo对剪裁后的部分进行扩展以生成伪数据。Behzad等人[8]在已有数据集的基础上,通过在ICNALE语料库中标注未用于训练和验证的其他文章中的介词使用类型错误,从而扩大了数据规模。

为了在生成过程中充分利用和参考已有的数据,Ihori等人[9]提出了基于检索的生成方法,该方法包括三个主要模块:检索模块、屏蔽模块和生成模块。首先,检索模块从训练数据中检索与输入学习者语句最相似的实例。然后,屏蔽模块将屏蔽检索到的反馈评语中与输入语句关联度较低的词汇。最后,生成模块依据输入语句及经过屏蔽的反馈评语来生成最终的结果。检索和屏蔽模块基于BERT模型,而生成模块使用预先微调的T5模型。Jimichi等人[10]采用了预训练的T5模型作为生成器,并使用RoBERTa作为分类器来获取名词、介词等语法术语标签。这些预测出的语法术语标签被用作生成模型中的一个额外信息源。

语法错误纠正任务中对错误进行解释的研究也备受关注。Fei等人[11]认为错误的原因(线索词语)及其对应的错误类型是解释错误的两个关键因素。为了通过解释来增强GEC模型,该研究引入了一个配有线索词语和语法错误类型标注的大型数据集——EXPECT,并基于此数据集提出了结合语法分析和错误修正机制的两个基线模型。

近年来,大语言模型(LLMs)在上下文学习(In-context Learning)和少样本学习(Few-Shot Learning)中表现出了令人惊叹的性能。大语言模型能够直接在推理阶段利用输入中少量的任务示例学习任务模式,无须修改任何模型参数就能生成符合预期的输出。Brown[12]提出的GPT-3是这一领域的重要里程碑,其凭借规模化的参数训练和丰富的语料预训练,能够在没有专门微调的情况下,完成包括翻译、总结、问答和代码生成等多种任务。与传统需要大量标注数据的监督学习方法相比,上下文学习显著降低了数据标注成本,并增强了模型的通用性和迁移能力。

2 数据

本研究基于现有的反馈评语生成数据资源,构建了首个英-汉跨语言反馈评语生成数据集。本节将详细介绍数据来源及数据集构建的具体过程。

2.1 数据来源

现有的反馈评语生成数据集主要包括基于亚洲英语学习者国际语料库网络(ICNALE)标注的数据集,以及基于INLG 2022语言学习者反馈评语生成共享任务(FCG GenChal)的数据集。前者从ICNALE语料库中选取了来自中国、日本、韩国、泰国、印度尼西亚等国家的1 194篇文章,使用日语对其进行了反馈评语标注,其中400篇文章的作者来自中国,约占33.5%。尽管该数据集属于跨语言数据集,但该数据集并未得到实际的使用。而后者则是在前者的基础上筛选出有关介词使用错误的子集,并将其反馈评语翻译为英语。这也是目前唯一广泛应用于反馈评语生成的数据集,其规模如表2所示。

表 2　反馈评语数据集规模

数据集名称		学习者语句数量	总词语数量	反馈评语数量
ICNALE 数据集		17 938	300 289	10 463
共享任务数据集（FCG GenChal）	训练集	4 868	110 906	4 868
	开发集	170	3 142	170
	测试集	215	4 446	215

2.2　数据标注

2.2.1　数据清洗

本研究首先对原始数据集中存在的错误进行了修正，主要包括以下几个方面：

错误符号　语法术语符号、原句引用符号及引号的使用需要满足两两匹配关系。在这一步骤中，修正了包括非法语法术语（如＜＜verb＞＞ → ＜verb＞）、非法原句引用（如＜＜couple＞ → ＜＜couple＞＞）和非法引号（如 'of ' ' → "of"）等错误。

符号混淆　语法术语符号和原句引用符号分别应为“＜…＞”和“＜＜…＞＞”，然而在原始数据集中，部分符号存在混淆使用的情况。例如，语法术语“＜verb＞”被错误标记为“＜＜verb＞＞”。本研究通过人工逐条检查的方式，修正了此类混用错误。

非法索引　目标索引的起始位置和结束位置必须精确对应错误单词的起始位置和结束位置，且不能包含单词之间的空格。例如，对于句子“It is fun to me.”，索引 10:12 是正确的，而 9:12 的索引则是非法的。本研究通过将字符级别索引自动转化为单词级别索引以解决这一问题。

字符编码　其他一些较为少见的错误，如语法错误和非 ASCII 字符的使用（如全角符号 'A' 和半角符号 'A'），也在数据清洗过程中得到了修正。

2.2.2　评语翻译

对于生成由学习者母语撰写的跨语言反馈评语，一个最自然的思路是在现有单语言反馈评语生成结果的基础上增加一个翻译模块。因此，本研究首先尝试使用商业机器翻译模型进行从英语到汉语的评语翻译。然而，根据本文的测试结果，反馈评语中包含的表示语法术语和学习者语句引用等的特殊符号（如“＜”、“＞”、“＜＜”和“＞＞”）通常无法被传统机器翻译模型有效处理。具体来说，传统机器翻译模型在面对包含此类特殊符号的反馈评语时，往往无法准确地翻译其中的语法术语和学习者语句引用，从而进一步造成学习者的困惑。增加机器翻译模型的方法存在对语法术语翻译不准确（如将“＜intransitive verb＞”错误地翻译为“＜in 及物动词＞”）、无法理解评语中特殊符号（部分翻译或过度翻译）等问题，这也间接证明了开展跨语言反馈评语生成研究的必要性。基于此，本文弃用直接使用翻译模型的方法，转而采用人工校对＋基于语法术语翻译对照表的大语言模型辅助翻译的策略，以构建跨语言的反馈评语生成任务数据集。

语法术语翻译　在原始评语数据集中，相同的语法点往往存在多种表述方式（如“determiner”和“qualifiers”均表示限定词）。此类多种表述方式与反馈评语生成任务的定义相悖，既不利于语言学习者的学习与记忆，也不利于构建与语法知识点相对应的详细解释。因此，本文参考一线大学英语授课教师的建议，通过人工翻译和整理，将原有的 2 202 个日语语法术语和 348 个英语语法术语全部对应并归纳为 1 745 个汉语语法术语，减少了其中的重复表述，并使其更符合中国英语教学的实际需求。

机器翻译辅助　本文基于上述构建的语法术语翻译对照表，使用 GPT-4-Turbo 模型对英语反馈评语进行翻译，以解决传统机器翻译模型在处理反馈评语中特殊符号之间的内容经常出现错误的问题。为了确保大语言模型的翻译结果符合预期，本文在 Prompt 中加入了对翻译要求的详细描述，并提供了输入输出示例作为提示。此外，本研究根据构建的语法术语对照表，将语法术语翻译和原句引用进行了替换操作，通过 Prompt 的形式输入 GPT-4 模型，从而在翻译过程中保留语法术语和原句引用中的特殊符号。所用 Prompt 如表 3 所示。为了使模型拥有更好的翻译效果，本文在调用 API 时使用了以下 System Prompt：“You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture.\nKnowledge cutoff: 2023-12\nCurrent date: 2024-03-10”。

人工校对　为了进一步提高汉语反馈评语的质量，本研究招募了 10 名汉语母语者、所学专业为日语或日语水平达到 N2 以上的本科生和研究生，基于 GPT-4 模型的翻译结果，对汉语反馈评语进行校对。校对的主要任务包括：①检查翻译结果是否符合汉语的表达习惯；②检查翻译结果与原始反馈评语的含义是否一致；③检查翻译后的反馈评语是否

表3 使用GPT-4模型翻译反馈评语所用Prompt

日-汉评语翻译 Prompt	英-汉评语翻译 Prompt
请将下面的日语句子翻译成汉语，不要保留任何日语。在翻译中，请使用以下给定的词语替换： {term_str}：{reference_str}。直接输出翻译后的汉语句子，不要包含任何其他内容。 日语句子：{feedback} 汉语句子：	将下面的英语句子翻译成汉语。在翻译中，请使用以下给定的词语替换： {term_str}：{reference_str}。直接输出翻译后的汉语句子，不要包含任何其他内容。 英语句子：{feedback} 汉语句子：
term_str = {非母语语法术语} → {汉语语法术语} reference_str = {原句引用} → {原句引用}	

包含与原始评语相同的语法术语和原句引用等；④检查翻译后的反馈评语中包含的特殊符号是否符合两两匹配关系，以及其他格式是否正确。本研究为标注者组织了反馈评语生成任务翻译的培训会议，并让其对20条翻译结果进行了试标注。标注者正确完成试标注后，方可进行正式的标注任务。标注结束后，我们对标注结果进行了最终审核。为了更加准确地评估模型在实际应用中的效果，本文还邀请了两位英语教师对测试集进行了二次校对。

3 实验

基于上一节中构建的跨语言反馈评语数据集，本研究首先明确了任务定义，随后分别采用了附加引导提示信息的流水线预训练语言模型和基于少样本提示的大语言模型开展实验，最后详尽说明了本文使用的评估指标。

3.1 任务定义

跨语言反馈评语生成通常指根据给定的学习者语句，自动生成以学习者母语为载体的反馈评语。在本研究中，输入数据是一个由英语语句及其错误位置索引构成的集合，输出则是针对指定错误位置的汉语反馈评语。其中，错误位置索引是一个整数区间，用于明确标识学习者语句中需要解释说明的具体错误位置。生成的反馈评语旨在帮助作者（语言学习者）提升写作技能，内容通常包含对语法错误的点评，同时可能涉及话语、结构以及内容方面的建议。这些建议不限于改善学习者当前写作质量，还可能包含对学习者的鼓励或表扬，以增强其信心。

介词使用是非英语母语学习者常见的难点之一，其中包括介词区分、固定搭配、特殊用法以及习惯用语等方面的挑战。介词使用错误类型在现有数据集中拥有最大的数据量，在所有错误类型中占有最高的比例。因此，本研究以介词使用错误类型的跨语言反馈评语生成作为研究起点。已有研究表明，对于低资源的反馈评语类型，通过数据增强方法可显著提升模型性能[7-8]。因此，对于数据量较少的其他错误类型，可以通过数据增强的方式扩大数据规模，并采用与介词使用错误类型相同的方法生成反馈评语。

3.2 方法

3.2.1 预训练语言模型方法

反馈评语生成任务是一种序列到序列的生成任务，其目标是将语言学习者撰写的输入文本转换为解释语法规则的另一段文本。这意味着，在其他生成式任务中已被证明有效的方法（例如复制机制）可能同样适用于该任务[5]。

以下示例表明，在一条反馈评语中通常包含四个特殊的部分，即错误引用、修正编辑（Correction Edit，CE）、线索词语（Evidence Word，EV）以及语法术语（Grammatical Terms，GT）。

Input：The small steps will lead to a complete ban of smoking not only at restaurants，but also at any other indoor places. 90：92

Output：＜介词＞＜＜at＞＞*可以与*＜＜place＞＞*一起使用来指示某事发生的地点，但更常见的是使用*'in'*代替。*

反馈评语通常会引用输入文本中出现的单词或短语。错误引用“at”已经在输入中被标记，而未被标记线索词语“places”也同样出现在原始语句中。此外，反馈评语中使用符号“＜”和“＞”标注的语法术语，以及引号“ ”之间的修正编辑，也可能对提升反馈评语生成的效果有所助益。如果将这些信息附

加到生成模型的输入中，模型能够更方便地引用输入文本中的相关片段，从而有效降低生成任务的难度。

对于生成模型而言，同时识别原始语句中的错误引用，并预测修正编辑、线索词语和语法术语，然后将这些要素组织成完整语句，是一项极具挑战性的任务。因此，我们提出将任务拆分为多个步骤，采用流水线的方式处理。首先，前置模型会预测修正编辑、线索词语和语法术语等信息；接着，基于这些附加信息由生成模型输出最终的反馈评语。这种步骤化的分解有助于模型更高效地处理复杂信息，从而提升反馈评语生成的准确性。在跨语言反馈评语生成任务中，本研究基于前文构建的数据集，选用多语言预训练语言模型 mBART[13] 和 mT5[14] 作为反馈评语生成模型，并对其进行微调。任务流程如图 1 所示。

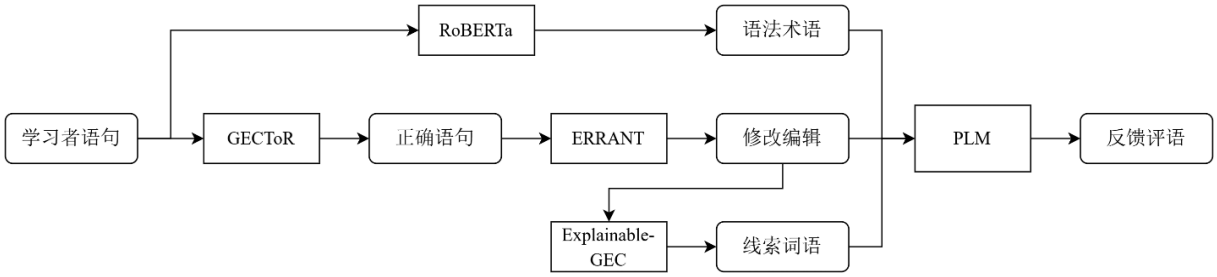

图 1　跨语言反馈评语生成任务流程

直接微调　在前文所构建数据集的基础上，本文对 mBART 和 mT5 预训练语言模型进行了直接微调，并将其作为跨语言反馈评语生成任务的基线模型。考虑到预训练语言模型无法有效处理输入中的索引位置，本文对原始输入中基于字符的错误位置索引进行了调整，改为基于单词的位置索引，并在需要进行评语的起始位置和结束位置分别添加符号“[”和“]”作为标记。

修正编辑　由于约有 34% 的反馈评语中包含了对错误位置的具体修改建议，我们使用了 GECToR[15] 模型对语料中的所有语句进行了自动修正，以确保语句中非评语部分的语法正确性。随后，为了精确获取针对错误位置的修改编辑（CE），本文采用了 ERRANT[16] 工具，从原始语句和修正后的语句之间提取编辑操作，并仅保留发生在评语索引范围内的编辑操作进行保留。例如，上例中的修正语句为：“The small steps will lead to a complete ban of smoking not only at restaurants, but also in any other indoor places.”，保留的 ERRANT 结果是：“Orig：[11，12，'at']，Cor：[11，12，'in']，Type：'R:PREP'”，表示原始语句中的介词‘at’被替换为了‘in’。

线索词语　线索词语（EV）通常能提供有关语法错误产生的关键线索或指示。例如，在处理介词错误时，介词及其周边的名词短语往往是至关重要的线索。通过在生成模型的训练过程中加入这些线索词语的标注，可以显著提升模型在错误识别和修正方面的准确率。基于已经获取的修正编辑，本文直接应用了 Fei 等人[11] 公开的线索词语预测模型，以获得相应的线索词语。

语法术语　本文注意到几乎所有的反馈评语中均包含有语法术语（GT）。在生成反馈评语之前，如果能预测并将可能使用的语法术语附加到输入中，可能会显著提升模型的性能。在前一节中，本研究已经构建出一个完整的语法术语表。考虑到一条反馈评语中可能包含多个语法术语，因此可将其视为一个多标签分类任务。基于此，本研究借鉴了 Jimichi 等人[10] 提出的方法，在 RoBERTa[17] 模型的基础上附加一个线性层来预测语法术语。该模型在汉语数据集上训练，并且同样仅保留出现频率前十的语法术语作为输入的附加信息。

本研究将上述三种附加提示分别追加在以制表符“|”作为分割的原始学习者语句之后作为输入，并尝试将这些提示两两组合或全部组合作为输入，分别对 mBART 和 mT5 预训练语言模型进行了微调，旨在探究不同输入组合对模型性能的影响。

3.2.2　大语言模型方法

鉴于大语言模型在语言理解和指令遵循方面的强大能力，以及其在通用领域及多种自然语言处理任务中的优异表现，本文评估了大语言模型（基于 Few-Shot 策略的 GPT-3.5 和 GPT-4）在跨语言反馈评语生成任务上的效果，并探索了在此任务上提升其效果的方法。

本研究主要关注构建有效的 Prompt，以最大限度地利用大语言模型的上下文理解和指令遵循能力。为了使大语言模型更好地理解并适应跨语言反

馈评语生成任务，本文在详细描述任务的基础上，尝试了 Zero-Shot、One-Shot 和 Few-Shot 等不同的 Prompt 构造策略。通过不断优化 Prompt 内容，从众多版本中筛选出了表现最佳的 Prompt。在实验中，本研究要求大语言模型扮演中国英语教师的角色，以使其与中国英语教师的行为模式对齐。此外，本研究还在 Prompt 中添加了针对错误位置的添加、修改、删除三种类型的评语示例，以帮助模型更好地理解任务描述、需求和输出模式。

为了深入探索大语言模型在跨语言反馈评语生成上的表现，我们选择使用 OpenAI 的闭源商用大语言模型 GPT-3.5-Turbo 和 GPT-4-Turbo 进行实验，二者均使用撰写本文时的最新版本，即 GPT-3.5-Turbo-0125 和 GPT-4-Turbo-2024-04-09。本研究采用的 Prompt 为：

> 你是一位经验丰富的中国英语教师。你需要对于给定的一个学习者语句和其中指定的错误提供具体且简洁的汉语反馈评语，让学习者能够理解这个错误的本质原因。学习者语句中使用方括号[和]标记错误位置。
> 每一个反馈评语应该包含：为什么学习者语句中指定的结构是错误或不合适的，其本质的原因或者规则是什么？
> 请直接给出反馈评语。
> 如下是几个例子：
> 学习者语句：And how to propaganda and let people [agree this] rule?
> 反馈评语：由于＜动词＞＜＜agree＞＞是一个＜不及物动词＞，所以＜介词＞需要在＜宾语＞之前。在字典中查找＜动词＞＜＜agree＞＞以找到合适的＜介词＞。
> 学习者语句：We must consider other people especially [at] the public places.
> 反馈评语：＜介词＞＜＜at＞＞可以与＜＜place＞＞一起使用来指示某事发生的地点，但更常见的是使用 'in' 代替。
> 学习者语句：It was miserable and I thought that I did n't want to face [to] such a situation.
> 反馈评语：由于＜动词＞＜＜textless face＞＞是一个＜及物动词＞，因此＜宾语＞不需要＜介词＞。
> 学习者语句：{{学习者语句}}
> 反馈评语：

在上述指令中，本研究将需要模型生成反馈评语的学习者语句以示例相同的形式输入模型，有助于其仅生成反馈评语而不包含无关内容。为了使大语言模型在跨语言反馈评语生成任务上得到更好的效果，本研究同样添加了如2.2.2节中所述的 System Prompt。

3.3 评估

3.3.1 自动评估指标

表层相似度 与现有关于反馈评语生成的研究相同，本文采用 BLEU（Bilingual Evaluation Un-derstudy）[18]作为评估生成结果表层相似度的指标。BLEU 最初用于评估机器翻译的质量，它通过统计模型输出文本与一组参考文本之间的 *N*-gram 匹配次数，从而反映出二者之间的相似程度。本研究直接使用了 Nagata 等人[4]在共享任务中提供的评估代码，用以计算 BLEU 分数。

语义相似度 对于语义相似度的评估，本研究采用了 BERTScore[19]指标。BERTScore 通过使用预训练语言模型 BERT 来衡量生成文本与参考文本之间的语义关联。具体而言，BERTScore 通过计算生成的反馈评语中每个词语的 BERT 嵌入向量与参考反馈评语中相对应的最相似词语的余弦相似度，来评估整体的语义匹配程度。此方法不仅关注词语的表层匹配，而且更加重视语义的深层相似性，从而更全面地评估生成文本的整体质量。

3.3.2 人工评估指标

在人工评估方法中，本文采用了 Nagata 等人[4]提出的评估框架。具体来说，三位独立的评价者在盲审的情况下，基于指定的评估标准，使用 0～2 的评分系统对每条生成的反馈评语进行评估：完全正确(2)、部分正确(1)或不正确(0)。其中，完全正确(2)指反馈评语不仅包含了与参考内容相似的信息，而且未包括任何与错误无关的内容。即使评语中包含了参考内容未提及的相关信息，只要这些信息与错误直接相关，亦被视为完全正确。部分正确(1)意味着反馈评语基本准确，但需要简单编辑以提高准确度，例如如果反馈正确指出了句子中的错误，并且只需修改几个词便能使其内容更准确，则归于此类。而对于那些与参考内容完全无关，未能指出错误原因的反馈评语，则被评为不正确(0)。

所有评价者均为拥有语言学背景的汉语母语者，其英语水平达到六级或专业四级，并且了解跨语言反馈生成任务。此外，为了确保评估的一致性和可靠性，本研究进行了 Kappa 统计分析，其值为 0.817 2，显示评价者之间有较高的一致性。

3.3.3 大语言模型评估方法

由于传统的基于参考内容的表层相似度指标(如 BLEU)与人类判断的相关性较低，而人类评估的耗时较长且成本高昂。因此，本文针对需要一定创造性和多样性的跨语言反馈评语生成任务，尝试采用 GPT-3.5 Turbo 和 GPT-4 Turbo 等大语言模型作为评估工具。此方法不仅可以验证本文所提方法在更广泛语境中的有效性，还能够探索如何利用

大语言模型的强大功能来进一步提升跨语言反馈评语的生成质量。

本文分别采用了无监督评估和有监督评估两种方式。在无监督评估中，本文参考 Liu 等人[20]和 Wang 等人[21]提出的大语言模型评估框架，分别探索了在 Prompt 中指定评估维度和未指定评估维度情况下大语言模型的评估效果；对于有监督评估，本研究将上一节中所述的人工评估标准融入 Prompt 中，以指导大语言模型进行评分。

4　分析与讨论

本节采用第 4.3 节中介绍的评估指标，对反馈评语生成模型的性能进行了评估，并对结果进行了深入分析与讨论。

4.1　预训练语言模型方法结果与分析

本节分析了采用预训练语言模型方法的结果，基于 BLEU 分数和 BERTScore（包括精确度、召回率以及 F_1 分数）的评估结果如表 4 所示。

通过分析 mBART 模型的结果，本研究观察到当模型输入中组合附加修正编辑和线索词语（Edit＋EV）时，BLEU 分数达到最高值（50.317%），而 BERTScore 的精准率、召回率和 F_1 分数分别为 86.982%、86.308%和 86.589%，均高于其他单一策略或组合策略的分数。这一结果表明，修正编辑和线索词语相拼接组合的加入显著提升了模型生成反馈的质量和准确性，这可能得益于附加的语法信息为模型提供了关键的上下文支持。

表 4　预训练语言模型实验结果　（单位：%）

模型	mBART				mT5			
	BLEU	BERTScore			BLEU	BERTScore		
		P	R	F_1		P	R	F_1
Bare	47.739	84.497	83.463	83.906	40.521	83.668	82.787	83.158
Edit	43.472	85.571	84.154	84.784	42.126	84.946	84.473	84.625
EV	45.141	84.135	83.381	83.700	41.180	83.515	82.664	83.011
GT	48.951	86.973	86.144	86.490	**47.953**	**85.662**	84.693	**85.100**
GT＋Edit	49.581	86.942	85.979	86.391	46.211	84.972	**84.725**	84.781
GT＋EV	47.546	86.793	85.153	85.896	45.901	85.116	84.329	84.644
Edit＋EV	**50.317**	**86.982**	**86.308**	**86.589**	42.693	84.198	83.413	83.736
GT＋Edit＋EV	48.503	86.803	85.961	86.313	44.538	85.069	83.961	84.442

对于 mT5 模型而言，采用单独附加语法术语（GT）的输入策略表现出最优的 BLEU 分数（47.953%），并在 BERTScore 的精准率和 F_1 分数上获得了相对较高的评分（分别为 85.662%和 85.100%）。这一现象说明语法术语信息对于 mT5 生成反馈评语有极大的辅助作用。

当采用多元输入组合策略（GP＋Edit＋EV）时，本文注意到，虽然这一组合在 BLEU 分数上并未达到最佳的效果，但在 mBART 和 mT5 模型上的 BERTScore F_1 分数均有所提升，分别达到 86.313%和 84.442%。这表明了一个多元化的输入组合可能为模型平衡精确度与召回率提供了额外的帮助，尽管这种平衡并未在 BLEU 分数上反映出显著的提升。

在两种模型的对比中，mBART 在大多数评价指标上均优于 mT5，这可能表明 mBART 在处理英语到汉语的跨语言反馈生成任务上具有更为适宜的架构和更强的处理能力。

考虑到人工评估的时间和成本，我们仅对在 mBART 和 mT5 两个模型上取得最佳效果的情况进行了人工评估，其准确率分别为 62.84%和 59.27%。

4.1.1　流水线中的前置模型效果

修正编辑和线索词语预测模型效果　由于预测线索词语需要依赖修正编辑结果，本节将二者合并进行探讨。目前，最先进语法错误纠正模型的准确率约为 67.5%。此外，由于在反馈评语数据集中，有部分实例是针对语言习惯的评语，即其原始的语句

在语法上并非存在错误，但是存在更符合语言习惯的表述方式，这类实例目前难以从现有GEC模型中获取更多的修正信息。总体来看，GECToR能够识别和修正的语法错误实例尚不及全部数据条目的50%，如表5所示。

表5 GECToR修正错误数量及比例

数据集	评语数量	修改数量	修改比例/%
训练集	4 868	2 383	48.95
开发集	170	80	47.06
测试集	215	90	45.58

虽然在全部数据中仅有一部分数据含有附加的修正编辑和线索词语信息，但是本研究发现仅有部分数据带有的信息对提升模型效果仍然发挥了重要作用。如表6所示，对于mBART模型，尽管添加修正编辑和线索词语信息导致了在全部测试数据上模型效果的下降，但是针对带有附加信息的部分实例，效果的下降大大缓解；对于mT5模型，带有附加信息的实例比整个测试集展现出了更好的效果。因此，可以推测的是，随着GEC模型效果的提升，这一方法有望显著增强跨语言反馈评语生成的效果。

表6 直接微调和以修正编辑或线索词语作为附加信息的结果分析 （单位：%）

模型	mBART		mT5	
	全部数据	部分附加信息数据	全部数据	部分附加信息数据
Bare	47.739 1	48.522 5	40.521 0	41.046 0
Edit	43.472 3（−4.266 8↓）	45.213 0（−3.309 6↓）	42.126 0（1.605 0↑）	46.677 8（5.631 8↑）
EV	45.140 6（−2.598 5↓）	48.115 5（−0.407 0↓）	41.180 2(0.659 3↑）	43.266 2（2.180 3↑）

语法术语预测模型效果 本研究基于附加额外线性层的RoBERTa模型，通过改进损失函数的计算方式和索引位置的计算策略，提高了模型对Top-10高频语法术语预测的准确性。语法术语预测的最终效果如表7所示，表中展示了该模型在开发集上的多标签预测性能。

表7 汉语语法术语预测模型结果评估 （单位：%）

EMR	P	R	F_1
18.82	85.68	72.93	78.55

如表8所示，当使用全部而非Top-10的语法术语作为生成模型输入中的附加信息时，生成模型的BLEU分数相较于当前的结果有较大幅度的提升，因此，在未来的研究中，进一步提高语法术语预测的准确性或增加可预测语法术语的范围（如扩展至前20个高频语法术语），都将有助于进一步提升反馈评语生成模型的性能。

表8 使用全部语法术语推理结果 （单位：%）

模型	mBART	mT5
GT-ALL	55.049 6	51.975 8

4.1.2 mBART和mT5表现的差异分析

本研究观察到，在跨语言反馈评语生成任务中，尽管mT5模型的参数量较大，但其表现却不如参数量较小的mBART模型。这可能由以下几个原因导致：首先，mBART是专为机器翻译任务设计并进行预训练的模型，其训练目标和数据处理方式与跨语言反馈生成任务的需求更为吻合。这种专门化的预训练架构可能赋予mBART在相关任务上取得更优表现的能力。相比之下，虽然mT5是一个用途广泛的多语言模型，但其预训练任务和目标可能与跨语言反馈生成任务的具体需求不完全匹配。此外，mBART在微调和推理阶段需要明确指定输入和输出语言，而mT5仅将任务视为通用的文本到文本生成，这种差异可能增加了mT5在此任务上的复杂性。再者，由于mBART与mT5在模型结构上的不同，mBART可能因其结构或内部机制的优化，在处理特定类型的语言生成任务时，能更有效地捕捉语言间的转换和生成规律。

另外一个有趣的现象是，在模型的输入中附加完全相同的提示信息，对mBART和mT5模型的性能产生了截然不同的影响。对于mBART模型而言，添加修正编辑和线索词语后，其BLEU值仅有微小幅度提升，甚至不及对初始模型进行直接微调所取得的效果。相反，mT5模型在加入这一附加信息后，其BLEU值显著高于直接微调的原始模型。

本文推测，这种差异主要源于模型架构的不同。尽管mT5和mBART均是基于Transformer架构的Seq2Seq模型，但它们在处理输入和输出的方式

上有所区别。mT5 设计为将所有自然语言处理任务视为文本到文本的转换，利用前缀提示来明确任务类型，因此在处理结构化输入时更具优势，这使得其能够将输入的附加信息视为明确的任务指令并有效利用。相反，mBART 以自编码方式进行预训练，专注于文本恢复，如重构乱序或屏蔽的文本，这使其更适合对输入进行理解，而对输入的不同表示形式敏感度较低。

此外，mT5 模型的设计初衷就是处理多样化文本输入，并产生相应的输出。附加信息的加入为输入文本提供了更丰富的上下文，使得 mT5 能够更有效地利用这些信息，从而适应多变的输入格式。相比之下，mBART 可能更依赖输入的一致性和结构的统一性，因此在处理更直接和传统的文本生成任务时可能表现更佳。然而，在反馈评语生成任务中，不是所有输入都包含修正编辑和关键词，这种结构上的不一致可能影响了 mBART 对输入的理解，从而导致附加信息对其处理原始文本造成干扰，最终导致性能下降。

对于适用于文本到文本转化的 mT5 模型而言，输入中附加的语法术语有很大概率在输出中被直接使用，因此其取得了最佳的性能。而对于在理解输入方面能力更强的 mBART 模型，输入中附加的修正编辑和线索词语可能为其提供更丰富的上下文逻辑信息，使其表现更好。

4.1.3　更多附加信息导致模型效果下降

根据本文的研究假设，当模型的输入包含更多附加提示信息时，其生成效果应该有所提升，这一假设也是符合人的直觉的。然而，研究发现在输入中添加多种附加信息可能会导致模型效果不及仅添加其中单一附加信息时的表现，即添加更多附加信息反而导致了模型效果的下降。例如，表 4 中 mBERT 模型附加全部语法术语、修正编辑和线索词语的情况不如只附加修正编辑和线索词语的情况，这一现象在 mT5 模型中表现得更加明显，附加多个信息的策略均不及仅单独使用语法术语作为附加信息的效果。

本研究认为，这一现象的原因是信息过载导致的注意力分散。将语法术语、修正编辑和线索词语同时附加到输入中可能导致信息量过大。对于基于 Transformer 架构的模型而言，输入的每个组成部分都会通过注意力机制相互影响。如果输入信息过于复杂或与任务不直接相关，模型可能难以从中提取到有效的特征，从而影响学习效果。虽然单独提供三者对模型生成反馈评语都是有益的，但它们可能在语义上存在重叠或者相互干扰，这会导致模型在处理输入时产生混淆，难以区分哪部分信息更为重要。如果语法术语、修正编辑和线索词语被同时附加，模型的注意力可能会在这三类信息之间分散，使得关键信息的权重降低。

4.1.4　案例研究

表 9 展示了基于流水线的预训练语言模型在生成反馈评语时如何通过使用不同的附加信息来影响生成效果的一个典型例子。

表 9　预训练语言模型方法的案例分析

They spend time which should be for learning [to work], thus leading to decline of grades.			
使用结构＜介词＋动名词＞或＜动名词＞代替＜to 不定式＞来表达“在某事上浪费时间”。在字典中查找＜动词＞ ＜＜spend＞＞的正确用法。			
Bare	/	由于＜动词＞ ＜＜learn＞＞是一个＜及物动词＞，因此＜宾语＞前面不需要有＜介词＞。在字典中查找＜动词＞＜＜learn＞＞的用法。	×
GT	＜to 不定式＞、＜动名词＞	使用＜to 不定式＞的＜动名词＞形式来表达“为了某事努力工作”。	√
Edit	to work → working	使用结构＜介词＋动名词＞或＜动名词＞代替＜to 不定式＞来表达学习的东西。在字典中查找＜动词＞＜＜ work＞＞的用法。	√
EV	spend、for	使用结构＜动名词＞代替＜to 不定式＞来表达“在某事上浪费时间”。在字典中查找＜动词＞＜＜spend＞＞的正确用法。	√
Edit＋EV	to work → working | spend、for	使用结构＜介词＋动名词＞或＜动名词＞代替＜to 不定式＞来表达“在某事上浪费时间”。在字典中查找＜动词＞＜＜spend＞＞的正确用法。	√

该表的前两行分别为学习者语句和人类教师提供的反馈评语，接下来的五行分别为 mBART 模型在不同附加信息配置下生成的反馈评语。表中波浪线标识了评语中存在错误的内容，而下划线突出了正确且其他评语未涉及的内容。

在学习者语句中，“which should be for learning”

是修饰宾语“time”的定语从句，而“work”应该作为目的状语，以解释他们花时间的目的。然而，学习者错误地使用了介词“to”与“learning”进行搭配。不含任何附加信息方法生成的解释完全错误；采用语法术语或修正编辑的方法正确地解释了修改方式，但错误原因的分析有误；而附加了线索词语的方法首次正确识别出关联错误的词语为“spend”；取得最佳表现的修正编辑＋线索词语的方式进一步完善了修改方式，即提示出可以改为“working”或“on working”。这些结果与添加各种附加信息的初衷相符，证明了方法的有效性。

4.2 大语言模型方法结果与分析

本节评估了大语言模型（GPT-3.5 和 GPT-4）在少样本学习策略下生成跨语言反馈评语的表现，并将其与采用修正编辑＋线索词语（Edit＋EV）流水线的预训练语言模型 mBART 在本任务上的表现进行对比分析。最后，通过计算与人工评估的相似性，说明使用大语言模型进行评估的可靠性。

此外，本节中引入了宽松的人工评估指标：模型生成的反馈评语中只要包含有错误相关的信息，即认为它是正确的（不考虑生成内容中是否存在错误和无关信息），实验结果如表 10 所示。实验表明，GPT-3.5 表现较差，而 GPT-4 展现出显著的性能提升，尤其在严格的人工评估中超过了 GPT-3.5 约 32.56%。使用 Edit＋EV 策略的 mBART 模型同样表现优异，特别是在宽松评估中得分接近 80%。这一结果凸显了使用先进的预训练模型和专门的优化策略对于提高反馈评语生成准确性的重要意义。

表 10 大语言模型于预训练语言模型效果的人工评估

（单位：%）

模型/方法	人工评估（严格）	人工评估（宽松）
GPT-3.5 (Few-Shot)	37.21 (25.42↓)	56.98 (23.00↓)
mBART-Large (Edit＋EV)	62.84	79.98
GPT-4 (Few-Shot)	69.77 (7.14↑)	89.53 (9.55↑)

如表 10 所示，GPT-4 在 Few-Shot 策略下对跨语言反馈生成任务的适应性和效果显著优于 GPT-3.5，而 mBART-Large 则利用其附加修正编辑和线索词语的策略，展现出稳定而高效的反馈评语生成能力。在跨语言反馈评语生成任务中，仅拥有 0.68B 参数的预训练语言模型 mBART 在经过微调后，效果较大幅度超过拥有 175B 参数量的 GPT-3.5，较小幅度低于包含约 1 800B 参数规模的 GPT-4，本文认为这足够令人兴奋。这意味着相较于通用的大语言模型，拥有更小参数规模的预训练语言模型可以通过针对特定任务的深入优化，同样能够展现出较好的效果。虽然基于预训练语言模型方法的效果尚不及最新最先进的 GPT-4-Turbo 模型，但其仍然提供了有价值的见解和创新，并为未来的研究奠定基础。

4.2.1 与人类教师评语的语言一致性

面向语言学习者的反馈评语生成不仅需要准确性，还要具备教育意义和可理解性，以确保学习者能从中获得最大的教育效益。本研究显示，经过任务数据微调的预训练语言模型能够更好地模仿人类教师使用的自然和教育性语言风格。

本研究通过人工标注比较了预训练模型生成的评语与真实教师评语的语言风格和表达方式。结果表明，预训练模型生成的评语在语言风格上与人类教师的评语更为接近，这有助于提升反馈评语的被接受程度和实际效果。相比之下，大语言模型虽然能够生成语法结构正确的语句，但其风格和用词往往缺乏针对性和教育性细节，有时可能显得过于机械或与教学语境不够契合。虽然大语言模型在通用领域的能力远高于预训练语言模型，但其微调过程需要大量计算资源和训练数据。因此，在目前的情况下难以对其进行微调，以使其适配到跨语言反馈评语生成任务中，更不易于与人类教师的反馈评语在内容、语言及形式上进行对齐。

4.2.2 评语的简洁性

在实际语言教学场景中，反馈评语的长度直接影响到学习者的阅读兴趣和理解程度。一些研究已经表明，大语言模型在表达相同的意思时倾向于生成更长的内容。尽管我们已经在 Prompt 中要求大语言模型生成尽可能简洁的评语，并添加了一些示例，但是其生成的评语长度仍然达到了人工标注和预训练语言模型的近两倍，而这会大大影响学习者的学习体验和效果。预训练语言模型生成的评语长度则与人工标注结果更为接近，结果如表 11 所示。

表 11 不同方法生成反馈评语的平均长度

模型/方法	平均评语长度（字符）
人类教师	59.72
mBART-Large (Edit＋EV)	56.26
GPT-3.5 (Few-Shot)	115.09
GPT-4 (Few-Shot)	119.45

除此之外，预训练语言模型在资源的可持续性

和可定制性方面具有显著优势。大语言模型如GPT-3.5 和 GPT-4 在运行时需要大量的计算资源，这可能不适合所有使用场景，特别是在资源有限的环境中。相比之下，预训练语言模型通常具有更低的资源需求，更适合长期可持续发展。并且，预训练模型也更易于针对特定的教育需求进行调整和定制。例如，可以根据学习者的具体错误类型或学习阶段调整模型参数，更好地适应不同学习阶段和不同教育场景。

4.2.3　大语言模型评估方法的有效性

如第 3.3.3 节所述，本节使用大语言模型（GPT 3.5-Turbo 和 GPT 4-Turbo）分别对预训练语言模型和大语言模型生成的反馈评语进行评估，结果如表 12 所示。

表 12　大语言模型评估结果　　（单位：%）

生成模型	评估模型					
	GPT-3.5			GPT-4		
	无监督		有监督	无监督		有监督
	无维度	含维度		无维度	含维度	
GPT-3.5(Few-Shot)	94.76	60.47	69.70	74.42	59.30	46.18
mBart-Large(Edit+EV)	77.91	58.14	65.59	43.02	66.28	70.24
GPT-4(Few-Shot)	98.83	84.88	60.12	97.67	88.37	85.88

在无监督评估过程中，当 Prompt 没有包含评估维度时，GPT-3.5 模型和 GPT-4 模型均对大语言模型生成的反馈评语给出了较高的评分，这是由于在缺乏具体评分参考和示例的情况下，大语言模型难以准确地理解反馈评语生成任务，因此更加倾向于大语言模型自身生成的答案。然而，当 Prompt 包含评估维度时，GPT-4 模型降低了对大语言模型生成结果的评分，并提高了对预训练语言模型生成结果的评分，这说明评分维度的加入有助于大语言模型更加明确任务定义及其评估方式。

在有监督的评估过程中，通过使用人工标注的反馈评语作为参考，大语言模型可以更加了解任务的模式，以更加客观和符合任务要求的方式进行评估。值得注意的是，大语言模型都更加倾向于自己生成反馈评语，这表明其在作为评估模型效果的工具时可能存在偏见。

总体而言，应该使用能力更强的模型评估能力较弱的模型，类似于老师对学生的评价。所以，本研究倾向于采用 GPT-4 作为评估者的结果，尽管会更偏向自身生成的答案。在三种设置中，GPT-4 与人工评估的 Kappa 一致性分别为 4.52、15.43 和 34.30，表明其与人类评估者之间有较高的一致性。这进一步证实，通过增加评估指标和参考示例，可以提高大模型对评估任务的理解和与人类评估的一致性。

4.2.4　案例研究

本研究采用了 Few-Shot 策略来生成反馈评语，而未对大语言模型进行微调。由于大语言模型的上下文学习能力有限，可能导致其生成的反馈评语在语言风格和结构组织上与人类评语存在显著差异。表 13 中对比了人类教师、经过微调的 mBART（Edit+EV）、GPT-3.5 和 GPT-4 生成反馈评语的示例，表中最后一列展示了 GPT-4 模型在有监督情况下的评分。

表 13　大语言模型方法的案例分析

Recently, some people appeal the government to make a new law to ban [people smoking] in restaurants.		
人类	当＜动词＞＜＜ban＞＞的＜宾语＞是人时，禁止的行为用＜from＋动名词＞表示	
mBART（Edit+EV）	当＜动词＞＜＜ban＞＞的＜宾语＞是人时，禁止的行为用＜from＋动名词＞表示	2
GPT-3.5（Few-Shot）	在这个句子中，动词“appeal”后面接的宾语应该是直接跟在其后的，应使用“to ban smoking by people”来表达这个意思。在这种情况下，需要把动词“ban”及其后面的宾语“smoking”放在一起作为一个整体	0

续表

GPT-4 (Few-Shot)	“people smoking”这个短语在这里需要使用名词化形式，应为“the smoking of people”或者更简洁地说，可以使用“smoking”。在英语中，当我们希望表述一个动作的泛化行为时，通常使用名词形式。在此句中，使用“smoking”替换“people smoking”，可以让语句更加地道和清晰	1

学习者语句中的短语“ban people smoking in restaurants”并没有明显违反语法规则，但这种表述可能导致一些歧义。这一表述可能被解释为专门禁止“在餐厅吸烟的人们”，而不是常见的表达“禁止在餐厅吸烟”。而“ban”一般的用法是“ban sb. from sth.”或“ban from doing sth.”。

在这一示例中，微调后的 mBART 模型与大语言模型采用不同的修改方式。mBART 生成的评语中提供了英语语法规则，即使用“from + 动名词”来表达禁止的行为，使得语句更加符合英语的习惯用法和教学中的固定搭配。而 GPT-3.5 模型没有对输入中指定的位置做评语，并且提出的修改方式“to ban smoking by people”显得笨拙和生硬，造成了表达的冗长。GPT-4 模型建议将“people smoking”简化为“smoking”不仅使句子更流畅，也避免了不必要的重复和过度的具体化。但是，在专注于介词使用的教学场景下，使用微调预训练语言模型的方法显然是更适合的。此外，通过上例可以观察到，大语言模型生成的评语的长度远大于经过微调的预训练语言模型生成的评语的长度。

5 结语

本研究提出了一种新颖且具有挑战性的跨语言反馈评语生成任务（CLFCG）。首先，本文基于 GPT-4 模型，利用人工标注的语法术语，对现有资源进行了翻译，并通过人工校对提升翻译质量，从而构建了首个英-汉跨语言反馈评语数据集。为应对这一任务，本文使用了两种多语言预训练模型，并探讨了不同附加信息对模型效果的影响。同时，分析了最先进的大语言模型在这一任务上的应用效果。通过广泛的实验验证，本文提出的方法能够有效地处理跨语言反馈评语生成任务。通过对结果的深入分析，本文期望为自然语言处理社区贡献更多洞见。未来，我们计划探索输入中更多提示信息的拼接与排序对模型的影响，并进一步分析本文所提方法对不同类型的错误生成反馈评语的效果差异。

参考文献

[1] NAGATA R. Toward a task of feedback comment generation for writing learning[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 3206-3215.

[2] NAGATA R, INUI K, ISHIKAWA S. Creating corpora for research in feedback comment generation [C]//Proceedings of the 12th Language Resources and Evaluation Conference, 2020: 340-345.

[3] ISHIKAWA S. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English[J]. Learner Corpus Studies in Asia and the World, 2013(1): 91-118.

[4] NAGATA R, HAGIWARA M, HANAWA K, et al. Shared task on feedback comment generation for language learners [C]//Proceedings of the 14th International Conference on Natural Language Generation, 2021: 320-324.

[5] NAGATA R, HAGIWARA M, HANAWA K, et al. A report on FCGGenChal 2022: Shared task on feedback comment generation for language learners [C]//Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges, 2023: 45-52.

[6] HANAWA K, NAGATA R, INUI K. Exploring methods for generating feedback comments for writing learning [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 9719-9730.

[7] BABAKOV N, LYSYUK M, SHVETS A, et al. Error syntax aware augmentation of feedback comment generation dataset [J]. arXiv preprint arXiv: 2212.14293, 2022.

[8] BEHZAD S, ZELDES A, SCHNEIDER N. Sentence-level feedback generation for english language learners: Does data augmentation help? [J]. arXiv preprint arXiv:2212.08999, 2022.

[9] IHORI M, SATO H, TANAKA T, et al. Retrieval, masking, and generation: Feedback comment generation using masked comment examples[C]//Proceedings of the 16th International Natural Language Generation

Conference：Generation Challenges，2023：60-67.

[10] JIMICHI K，FUNAKOSHI K，OKUMURA M. Feedback comment generation using predicted grammatical terms [C]//Proceedings of the 16th International Natural Language Generation Conference：Generation Challenges，2023：79-83.

[11] FEI Y，CUI L，YANG S，et al. Enhancing grammatical error correction systems with explanations[J].arXiv preprint arXiv:2305.15676，2023.

[12] BROWN T B. Language models are few-shot learners [J].arXiv preprint arXiv:2005.14165，2020.

[13] TANG Y，TRAN C，LI X，et al. Multilingual translation with extensible multilingual pretraining and finetuning[J].arXiv preprint arXiv:2008.00401，2020.

[14] XUE L. mt5：A massively multilingual pre-trained text-to-text transformer [J]. arXiv preprint arXiv：2010.11934，2020.

[15] OMELIANCHUK K，ATRASEVYCH V，CHERNODUB A，et al. GECToR：Grammatical error correction：Tag，not rewrite [J]. arXiv preprint arXiv：2005.12592，2020.

[16] BRYANT C J，FELICE M，BRISCOE E. Automatic annotation and evaluation of error types for grammatical error correction[C]//Proceedings of the Association for Computational Linguistics，2017.

[17] LIU Y. RoBERTa：A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv：1907.11692，2019，364.

[18] 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，2002：311-318.

[19] ZHANG T，KISHORE V，WU F，et al.BERTScore：Evaluating text generation with bert [J]. arXiv preprint arXiv:1904.09675，2019.

[20] LIU Y，ITER D，XU Y，et al. G-eval：Nlg evaluation using GPT-4 with better human alignment[J]. arXiv preprint arXiv:2303.16634，2023.

[21] WANG J，LIANG Y，MENG F，et al. Is chatGPT a good nlg evaluator?：A preliminary study[J]. arXiv preprint arXiv:2303.04048，2023.

安纪元(2001—)，硕士研究生，主要研究领域为自然语言处理和智能语言学习。
E-mail：jiyuanan.blcu@gmail.com

杨尔弘(1965—)，通信作者，博士，教授，主要研究领域为语言信息处理、语言监测和语言资源建设。
E-mail：yerhong@blcu.edu.cn

朱琳(2000—)，硕士研究生，主要研究领域为自然语言处理和智能语言学习。
E-mail：nectarzl@163.com

关于开展2025年“中国中文信息学会博士学位论文激励计划”推荐工作的通知

为推动中国中文信息处理领域的科技进步，鼓励创新性研究，促进青年人才成长，现开展2025年“中国中文信息学会博士学位论文激励计划”推荐工作，具体事项通知如下：

一、申报资格

1. 论文作者在攻读博士学位期间，在中文信息处理技术及其相关领域的基础理论或应用研究中取得重要成果，或在关键技术或应用技术创新等方面成果显著。

2. 论文作者在申报受理日期当年和前一年(2024年、2025年)获得博士学位。

3. 论文作者须为中国中文信息学会个人会员。

二、推荐办法

CIPS接受各单位或三名学会理事/会士/院士联名推荐的候选博士论文。

1. 每个单位推荐候选博士学位论文不超过3篇。

2. 每位学会理事/会士/院士推荐候选博士论文不超过2篇。

3. 已经参评过的论文，不得再次参评。

4. 已获得其他一级学会推荐的博士学位论文(含提名)，不得再次参评。

三、评审程序

根据参评博士论文的研究方向选聘专家组成评审委员会，负责组织工作。评审过程分为格式审查、初审、终审和颁奖四个阶段。

受理材料：2025年7月11日—8月11日，受理材料，资格审查。

初审阶段：2025年8月12日—9月12日，专家初审，初审结果进行公示，为期3天。

终审阶段：2025年9月15日—9月30日，专家终审，终审结果进行公示，为期3天。

颁奖大会：时间待定

四、申报材料

1. 印刷版论文3份，电子版1份；

2. 推荐表(附件)纸质版3份(附相关证明材料)，推荐表电子版1份；

3. 攻读博士论文期间发表的论文(集)印刷版3份，电子版1份；

4. 其他有关证明材料。

涉密推荐材料，请按国家有关法律、法规进行审查，并提交保密审查证明。

五、联系方式

1. 电子版材料请发送至评选邮箱：pingjiang@iscas.ac.cn

(请在邮件主题标注“姓名，2025CIPS论文激励计划”字样)

2. 印刷版材料请邮寄至：

北京市海淀区中关村南四街4号7号楼201房间

中国中文信息学会 100190　　电话：010-62661047

联系人：肖老师(请在信封表面标注“姓名，2025CIPS论文激励计划”字样)

(中国中文信息学会)

本刊为下列检索期刊及数据库刊源

＊中国科技论文统计源期刊（中国科技核心期刊）
＊中国核心期刊（遴选）数据库收录期刊
＊全国中文核心期刊
＊中国学术期刊综合评价数据库收录期刊
＊中国期刊全文数据库全文收录期刊
＊中国科技期刊精品数据库收录期刊
＊中国科学引文数据库收录期刊
＊中文科技期刊数据库收录期刊
＊CEPS中文电子期刊服务数据库全文收录期刊
＊《中国学术期刊文摘》（中文版与英文版）收录期刊

中 文 信 息 学 报
（月刊，1986年创刊）
第39卷 第7期 2025年7月

Journal of Chinese Information Processing
(monthly)
(Started in 1986)
Vol.39 No.7 Jul. 2025

主管单位	中国科学技术协会	Managed by	China Association for Science and Technology
主办单位	中国中文信息学会 中国科学院软件研究所	Sponsored by	Chinese Information Processing Society of China Institute of Software, Chinese Academy of Sciences
主　　编	孙茂松	Editor in Chief	Sun Maosong
编　　辑	《中文信息学报》编辑部 北京市海淀区中关村南四街4号 邮编 100190 电话 010-62562916 E-mail:jcip@iscas.ac.cn http://jcip.cipsc.org.cn	Edited by	Editorial Board of Journal of Chinese Information Processing 4# South Fourth Street, Zhongguancun, Haidian District, Beijing 100190, China Tel: 010-62562916 E-mail: jcip@iscas.ac.cn http://jcip.cipsc.org.cn
出　　版	《中文信息学报》编辑部	Published by	Editorial Board of Journal of Chinese Information Processing
印　　刷	北京科信印刷有限公司	Printed by	Beijing Kexin Printing Co., Ltd.
国内发行	《中文信息学报》编辑部 全国非邮发报刊联合征订服务部	Distributed by Domestic	Editorial Board of Journal of Chinese Information Processing National United Distributing Department of Chinese Periodical and Serials
国外发行	中国图书进出口总公司	Foreign	China National Publications Import and Export Corporation

中国标准连续出版物号 ISSN 1003-0077 CN 11-2325/N
国外发行代号：BM0077T
定　　价：50.00元