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
jiyuanan.blcu@gmail.com

摘要

反馈评语生成任务旨在为语言学习者的产出提供纠偏及解释性的评价，促进学习者写作能力的发展。现有研究主要聚焦于单语的反馈评语生成，如为英语学习者提供英文反馈评语，但这忽略了非母语学习者可能面临的理解障碍问题，尤其当评语中存在陌生的语言知识时。因此，本文提出跨语言反馈评语生成任务（CLFCG），目的是为语言学习者生成母语的反馈评语。本研究构建了首个英-中跨语言反馈评语生成数据集，该数据集包含英语学习者产出的语句与相应的中文反馈评语，并探索了基于流水线的预训练语言模型引导增强生成方法，将修正编辑、线索词语和语法术语等作为输入的附加信息，引导和提示生成模型。实验结果表明，附加引导信息的预训练语言模型流水线方法在自动评估（BLEU: 50.32）与人工评估（Precision: 62.84）上表现良好。本文对实验结果进行了深入分析，以期为跨语言反馈评语生成任务提供更多见解。

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

Cross-Lingual Feedback Comment Generation for Language Learners

Jiyuan An[1,2], Lin Zhu[1,2], Erhong Yang[1,2*]
[1]National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, Beijing
[2]School of Information Science, Beijing Language and Culture University, Beijing
jiyuanan.blcu@gmail.com

Abstract

The task of generating feedback comments is designed to furnish corrective and interpretative evaluations for the outputs of language learners, thereby facilitating the enhancement of their writing competencies. Predominantly, existing studies have concentrated on monolingual feedback generation, such as providing feedback in English for learners of English. However, this approach neglects the comprehension barriers that non-native speakers may encounter, particularly when the feedback incorporates unfamiliar linguistic elements. To address this issue, this paper introduces the task of Cross-Linguistic Feedback Comment Generation (CLFCG), which aims to produce feedback in the learner's native language. This research develops the inaugural English-Chinese cross-linguistic feedback comment generation dataset, comprising sentences

* 通讯作者
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）；中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）

crafted by English learners paired with corresponding feedback in Chinese. The study explores a methodology based on a pipeline of pre-trained language models, enhanced by guided generation techniques that incorporate modification edits, cue words, and grammatical terminology as supplementary input to steer and prompt the generative model. The experimental outcomes demonstrate that the enhanced pre-trained language model pipeline with additional guiding information yields robust performance, achieving a BLEU score of 50.32 and a Precision of 62.84 in automatic and human evaluations respectively. An in-depth analysis of these results is presented, offering substantive insights for the advancement of cross-linguistic feedback comment generation tasks.

Keywords: Intelligent Computer-Assisted Language Learning , Feedback Comment Generation , Cross-Lingual Text Generation , Pretrained Language Models

1 引言

智能辅助写作是计算机智能辅助语言学习（ICALL）中的一个重要研究领域，其主要目的在于帮助语言学习者纠正写作中的错误，以提高他们的语言运用能力。近年来，自动语法错误纠正任务（GEC）受到了广泛的研究关注。该任务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，旨在减轻语言教师的负担，使语言学习过程更加即时和便捷。然而，自动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”。这种方法可能导致语言学习者在一知半解中修改了本次的错误，但是在未来的表述中出现同样的错误。反馈评语生成任务正是为了解决这一问题而设计的。反馈评语是帮助作者（学习者）提高写作技巧的字符串，它不仅提供关于语法错误的具体反馈，还可以包括关于话语、结构和内容等其他方面的评价。这种反馈既可能是对当前写作的具体指导，也可能是对作者的鼓励。总之，为语言学习者的写作提供针对性的建议或解释性评语，对于提高他们的语言运用水平是极其有益的。

在本研究之前，由于平行训练和评估数据的可获得性，有关反馈评语生成任务的研究主要集中于单一语言环境。然而，这些研究往往忽略了一个重要的实际问题，即生成的评语可能包含对语言学习者来说陌生的词汇。换言之，这更适合为希望提高写作水平的英文母语者生成反馈评语。如图1 所示，如果以汉语为母语的英语学习者想要知道错误的原因，相比仅提供结果的英文语法错误纠正和英文反馈评语，中文反馈评语更易于被理解和接受。

因此，本文提出了一项新颖的跨语言反馈评语生成（TLFCG）的新任务，并构建了首个针对英文学习者语句采用中文反馈评语的跨语言反馈评语数据集。在数据集的构建过程中，本研究利用现有的反馈评语生成数据资源，在构建语法术语翻译映射表的基础上，利用大语言模型以无监督的方式进行翻译，并通过人工标注对结果进行修正。为了实现跨语言反馈评语生成，本研究尝试将修正编辑、线索词语及语法术语等反馈评语中的关键信息整合入流水线，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，提高模型生成效果。研究结果表明，引入不同的附加信息会对生成效果产生显著影响，与直接微调的基线模型相比，本文所提方法能够生成更高质量的反馈评语。此外，我们对实验结果进行了深入分析，以期为今后的相关研究提供参考。本文的主要贡献有：构建首个针对中文母语者学习英语外语的反馈评语语料库；探索预训练语言模型和大语言模型跨语言反馈评语生成的能力；分析添加不同附加信息对预训练语言模型生成反馈评语效果的影响。

2 相关工作

智能辅助写作是计算机智能辅助语言学习领域的一个重要研究方向。由于传统的语法错误纠正任务缺乏可解释性，无法向语言学习者提供修改的原因，因此Nagata(2019)提出反馈评语生成任务，旨在为语言学习者提供关于错误原因的解释性说明，从而帮助语言学习者提升写作水平。Nagata(2020)发布了基于亚洲英语学习者国际语料库网络（ICNALE）(Ishikawa, 2013)标注反馈评语的细节，并公开了部分标注结果。该数据集采用英语（标注结果被弃用）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

图 1: 跨语言反馈评语生成任务

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“>>”用于表明其中的单词是从原始语句中引用的（如<<agree>>），这使得反馈意见更加灵活和具体。在INLG 2022针对语言学习者反馈评语生成的共享任务（GenChal）中，Nagata(Nagata et al., 2021)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的位置索引，输出是针对该位置的具体反馈评语。同时，Nagata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(Nagata et al., 2023)。

在反馈评语生成方法的研究中，研究者尝试了不同的模型架构和训练策略。Hanawa(2021)对基于检索、生成以及检索与编辑结合的三种方法进行了初步探索。研究发现，检索与编辑结合的方法虽然能够对检索结果进行修改，但其过度编辑的问题导致生成效果不佳；基于检索的方法虽稳定，却因仅限于检索已有评语库中的匹配项而缺乏灵活性；相较之下，简单生成的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，能够提供更多样化的反馈，但其生成评语准确性和可用性仍是一个挑战。

为了提高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。Babakov(2022) 通过使用依存句法分析对原始学习者语句进行了裁剪，并采用语言模型GPT-Neo对剪裁后的部分进程扩展以生成伪数据。而Behzad(2022) 则采取了一种不同的伪数据生成方式，该方法通过在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(2023) 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句子最相似的实例。然后，屏蔽模块将屏蔽检索到的反馈评语中与输入句子不太相符的词汇。最后，生成模块依据输入句子及已屏蔽的反馈评语来生成最终的反馈评语。检索和掩蔽模块基于BERT模型，而生成模块使用预先微调的T5模型。另一方面，Jimichi(2023)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。

此外，在语法纠错任务中，关于可解释性的研究也并未止步。Fei(2023) 认为对于语法错误纠正任务的可解释性而言，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。

3 数据

本研究利用现有反馈评语生成数据资源，构建了首个英-中跨语言反馈评语生成数据集。本章节将介绍数据来源和数据集构建的详细过程。

3.1 数据来源

现有反馈评语生成数据集，主要有基于亚洲英语学习者国际语料库网络（ICNALE）标注的数据集和INLG 2022 针对语言学习者反馈评语生成共享任务（FCG GenChal）的数据集（Nagata，2020）。前者是从ICNALE语料库中抽取了中国大陆、中国台湾、日本、韩国、泰国、印度尼西亚等国家和地区的1,194篇文章，使用日语对其标注了反馈评语，其中400篇来自中国的作者，约占33.5%。虽然这也是一个跨语言的数据集，但该数据集并未得到实际的使用。而后者是对前者中关于介词使用错误类型进行筛选并将评语语言翻译为英文后的结果，这也是目前唯一被广泛用于反馈评语生成任务的数据集。语料库规模如表1所示。

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	训练集	4,868	110,906	4,868
	开发集	170	3,142	170
	测试集	215	4,446	215

表 1: 反馈评语数据集规模

3.2 数据标注

3.2.1 数据清洗

本研究首先对原始数据中存在错误进行了人工修改，包括以下几个方面:

错误符号 语法术语符号、原句引用符号和引号的使用需要满足匹配关系。在这一步骤中，包括非法语法术语（例如<verb >>→ <verb>）、非法原句引用（例如<<couple>→ <<couple>>）和非法引号（例如，'of" → "of"）等错误已得到纠正。

符号混淆 语法术语符号和原句引用符号分别为“<...>”和“<<...>>”，但是在原始评语数据集中，二者存在一部分混合使用的情况，例如语法数据“<verb>”被标记为“<<verb>>”。本文通过人工逐条筛查的方式修正了这类错误。

非法索引 目标索引的开始和结束必须分别对应于错误单词的开始和结束，而不能包含单词之间的空格。例如，对于输入“It is fun to me .”，10:12 的索引是正确的，而9:12 的索引是非法的。本文通过将字符级别的索引自动转化为单词级别的索引解决了这一问题。

字符编码 其他一些更为少见的错误，如语法错误和非ASCII 字符的使用（例如全角符号'Ａ'和半角符号'A'）也在数据清洗中被修改。

3.2.2 评语翻译

对于生成学习者母语的跨语言反馈评语，一个最自然的想法是在现有单语言反馈评语生成结果的基础上挂载一个翻译模块。因此，本研究首先尝试使用商业机器翻译模型进行英译中的评语翻译。然而，根据本文的测试结果，由于反馈评语中包含表示语法术语和学习者语句引用的特殊符号（即“<”、“>”、“<<”和“>>”），传统机器翻译模型通常无法对其进行处理，传统机器翻译模型通常无法处理含有表示语法术语和学习者语句引用的特殊符号（如“<”、“>”、“<<”和“>>”）的反馈评语，因而导致反馈评语中包含的语法术语无法被正确地翻译的，从而进一步造成学习者的困惑。外挂机器翻译模型的方法存在对语法术语翻译不准确（例如将<intransitive verb>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，这也间接证明了对跨语言反馈评语生成进行研究的必要性。基于此，本文放弃直接使用翻译模型的方法构建跨语言的反馈评语生成任务数据集，转而采用人工校对基于语法术语对照的大语言模型辅助翻译方法。

语法术语翻译 在原始数据中，相同的语法点存在多种表述方式（例如“determiner”和“qualifiers”均表示限定词），这是与反馈评语生成任务的定义相悖的，既不利于语言学习

者的学习和记忆，也不利于构建语法知识点对应的详细解释。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个日文语法术语和348个英文语法术语全部对应至1,745个中文语法术语，减少其中相同描述，并使其更加符合中国的英语教学实际需要。

机器翻译辅助 本文基于上述构建的语法术语翻译对照，使用GPT-4-Turbo模型对英文反馈评语进行翻译，以解决使用传统机器翻译模型时对于特殊符号之间内容处理有误的问题。为了使其生成的翻译结果符合预期，本文在Prompt中加入了对翻译要求的描述和一些输入输出示例作为提示。除此之外，本研究基于上述构建语法术语对应集，将语法术语翻译和原句引用作为替换操作，以Prompt的形式输入GPT模型，从而在翻译过程中保留对于语法术语和原句引用的特殊符号，所用Prompt如表2所示。同时，为了使模型拥有更好的效果，本文在调用API时使用了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"。

日-中评语翻译Prompt	英-中评语翻译Prompt
请将下面的日语句子翻译成中文，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}。直接输出翻译后的中文句子，不要包含任何其他内容。 日语句子：{feedback} 中文句子：	将下面的英文句子翻译成中文。在翻译中，请使用以下给定的词语替换： {term_str}；{reference_str}。直接输出翻译后的中文句子，不要包含任何其他内容。 英文句子：{feedback} 中文句子：
term str = {非母语语法术语} ->{中文语法术语} reference str = {原句引用} ->{原句引用}	

表 2: 使用GPT-4模型翻译反馈评语所用Prompt

人工校对 为进一步提升中文反馈评语的质量，本研究招募了10名以中文为母语所学专业为日语或专业水平N2以上的本科生和研究生，在GPT-4模型翻译结果的基础上，对中文反馈评语进行校对。其主要任务是：(1)检查翻译结果是否符合中文的表达习惯；(2)检查翻译结果与原始反馈评语是否表达相似的含义；(3)检查翻译后的反馈评语是否包含与原始反馈评语相同的语法术语、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满足两两匹配关系以及其它格式是否正确。我们组织标注者参加反馈评语生成任务翻译的培训会议，并进行了对20条翻译结果的校对试标注，标注者完成试标注后可以进入正式的标注任务，我们对标注结果进行最终审查。为了评估模型在实际使用场景下的效果，我们邀请了两位英语教师来校对测试集。

4 实验

基于前一章节构建的跨语言反馈评语数据集，本研究首先对任务定义进行了明确，然后分别采用附加引导的流水线预训练语言模型方法和少量样本提示的大语言模型方法进行了实验，并对生成结果的评估指标进行了详细介绍。

4.1 任务定义

跨语言反馈评语生成通常指在对于给定的学习者语句自动生成学习者母语中的反馈评语。在本研究中，输入数据是一个由英文语句及其错误位置索引组成的集合，输出则是针对指定错误位置的反馈评语。其中错误位置索引是一个整数区间，用于指出学习者语句中需要进行解释性说明的具体位置。所生成的反馈评语旨在帮助作者（语言学习者）提升写作技能，通常包含对语法错误的点评，同时也可能涉及话语、结构及内容方面的建议。这些建议不仅能改善当前的写作，也可能包含对作者的鼓励或赞美。

由于介词使用对于非母语学习者存在一定难度，其中含有介词区分、固定搭配、特殊用法和习惯用语等难点。并且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。而Babakov（2022）和Behzad（2022）等已经证明，对于低资源的反馈评语类型，可以通过数据增强的方法显著提升模型的性能。因此，对于数据量较少的错误类型，可以通过数据增强的方式扩展数据规模，而后采用与介词错误相同的方法生成反馈评语。

4.2 方法

4.2.1 预训练语言模型方法

反馈评语生成任务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文本的一种序列到序列生成任务，这意味着在其他生成式任务中被证明有效的方法（如复制机制）可能同样有益于这一任务（Nagata，2023）。

通过以下示例，我们发现一条反馈评语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）和语法术语（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。

反馈评语通常引用输入文本中出现的单词和短语，错误引用“at”已经在输入中被标记，而未被标记线索词语“places”也会出现在原始语句中。同时，在反馈评语中使用符号“<”和“>”标记的语法术语，以及引号“”之间的修正编辑，对提升反馈评语生成的效果也可能是有益的。如果这些信息被附加在生成模型输入中，即可便于该模型引用源文本中的这些片段，从而降低生成任务的难度。

对于生成模型而言，同时识别原句中的错误引用，并预测修正编辑、线索词语和语法术语，然后将它们组织成一个完整的句子，是一项极具挑战性的任务。因此，我们考虑将任务细分为多个步骤，采用流水线的方式处理。首先预测修正编辑、线索词语和语法术语等信息，然后基于这些附加信息生成最终的反馈评语。这样的步骤化分解有助于模型更好地处理复杂信息，从而提高反馈评语的准确性。对于跨语言反馈评语生成任务，我们基于前文构建的数据集，选择采用多语言预训练语言模型mBART(Tang et al., 2020)和mT5(Xue et al., 2021)作为反馈评语生成模型进行微调。我们的任务流程如图2所示。

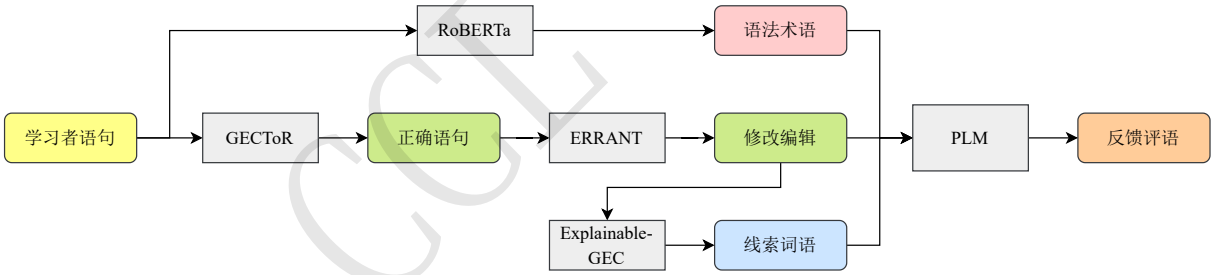

图 2: 跨语言反馈评语生成任务流程

直接微调 基于前文构建的数据集，我们直接微调了mBART和mT5预训练语言模型并作为跨语言反馈评语生成任务的基线。由于预训练语言模型无法有效处理输入中的索引位置，因此我们将原始输入中基于字符的错误位置索引修改为基于单词的位置索引，并在需要评语的位置起始和结束分别添加符号“[”和“]”作为标记。

修正编辑 由于约34%的反馈评语中包含了对于错误位置的修改建议，因此我们使用GECToR(Omelianchuk et al., 2020)模型对语料中的所有语句进行了自动修正，以得到语法正确的语句。然后，为了进一步得到错误位置的修改编辑，我们使用了ERRANT(Bryant et al., 2017)工具，提取原始语句和修正语句之间的编辑操作，对于编辑操作在待评语索引范围内的编辑进行保存。针对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结果为：“Orig: [11, 12, 'at'], Cor: [11, 12, 'in'], Type: 'R:PREP'”，即将介词‘at’替换为‘in’。

线索词语 线索词语通常可以提供有关语法错误发生原因的线索或指示。例如，在处理介词错误时，介词本身及其周围的名词短语常常是关键线索。通过在生成模型训练中加入这些线索词语的标注，可以显著提高模型的错误识别和修正准确率。我们基于已经获得的修正编辑，直接使用了Fei(2023)公开的线索词语生成模型，得到线索词语。

语法术语 我们注意到几乎所有反馈评语中均包含有语法术语（GT），如果在生成反馈评语之前，将预测出可能使用的语法术语附加到输入中，可能会显著提高模型的性能。在上一章节中，我们已经构建出了完整的语法术语表。由于每条反馈评语中可能使用了多个语法术语，因此可将这视为一个多标签分类任务。因此，我们借鉴了Jimichi(2023) 提出的基于RoBERTa(Liu et al., 2019)附加一个线性层的语法术语预测模型，并在中文数据集上进行训练，同样仅保留出现频率前十的语法术语作为附加信息。

我们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，并尝试将上述三者分别两两组合或全部组合作为输入，分别微调了mBART和mT5预训练语言模型。

4.2.2 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，以及其在通用领域和众多自然语言处理任务中的优异表现，本文评估了大语言模型（使用GPT-3.5和GPT-4，采用few-shot方法）在跨语言反馈评语生成任务上的效果，并初步探索了提升其在此任务上表现的方法。

本研究重点关注于构建有效的Prompt，以充分利用大语言模型的指令遵循能力。为了使大语言模型更好地理解和适应跨语言反馈评语生成任务，我们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，我们从数十个版本中筛选出了效果最佳的Prompt。我们要求大语言模型扮演中国英语教师的角色，使其与对英语外语教师的行为对齐。此外，我们分别添加了对于错误位置进行添加、修改、删除等三种类型的示例，以使大语言模型更好的理解跨语言反馈评语任务描述、需求和输出模式。详细的实验设置请见附录A。

4.3 评价

4.3.1 自动评估指标

表层相似度 与前人关于反馈评语生成的研究相同，本文使用BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002)作为评估结果表层相似度的指标。BLEU 最初被用于评估机器翻译的质量，其通过统计模型输出与一组参考之间的N-gram 匹配的次数来可以反映二者之间的接近程度。本研究直接使用Ryo Nagata（2021）在共享任务种提供评估代码工具来计算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 (Zhang et al., 2020)。BERTScore 利用预训练的语言模型BERT 来计算生成反馈与参考反馈之间的语义相关性。具体而言，BERTScore 通过计算反馈文本中每个词语的BERT 嵌入向量与参考文本中最相似词语的余弦相似度，来评估整体的语义匹配度。这种方法不仅关注词语的表层匹配，而且更加重视语义的深层次相似性，能够更全面地评估生成文本的质量。

4.3.2 人工评估指标

在人工评估方法中，我们采用了Nagata(Nagata et al., 2021)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，使用0至2的评分系统基于特定评估标准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。即便反馈中包含了参考内容未提及的相关信息，只要这些信息直接关联到错误，也被认定为完全正确。**部分正确（1）**的评分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的错误，且只需修改几个词便能让内容更准确，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，则被评为**不正确（0）**。

所有评价者均为拥有语言学背景的中文母语者，其英语都达到了六级或专业四级水平，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，我们还进行了Kappa 统计分析（0.8172），保证了评价者之间较高的一致性。

4.3.3 大语言模型评估方法

由于传统的基于参考的表层相似度指标（BLEU）与人类判断的相关性相对较低，而人类评估的耗时较长且成本较高。因此，对于需要一定创造力和多样性的跨语言反馈评语生成任务，本文尝试使用GPT 3.5-Turbo和GPT 4-Turbo等大型语言模型作为评估的指标，这不仅可以验证我们的方法在更广泛的语境中的有效性，也可以探索如何利用LLM的强大能力来进一步提升跨语言反馈生成的质量。

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，我们参考Liu(Liu et al., 2023)和Wang(Wang et al., 2023)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下大语言模型的评估效果；对于有监督评估，我们将上小节所述的人工评估标准作为Prompt的一部分指导大语言模型进行评分。经过验证发现，使用大语言模型评估与人类评估的一致性较低，因此我们仅在附录C中报告和分析了这部分结果。

5 分析

本章使用如4.3节所述的评价指标对模型反馈评语生成效果进行了评估。由于大语言模型方法在与人类教师评语的语言一致性、评语的简洁性、解释的准确性等方面的问题（详见附录B），在本节中我们仅对预训练语言模型方法进行了分析，本文所述预训练语言模型方法的BLEU分数和BERTScore（包括精确度，召回率以及F1分数）评估结果如表3所示。

	mBART				mT5			
	BLEU	BERTScore			BLEU	BERTScore		
		P	R	F1		P	R	F1
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
GTs	48.951	86.973	86.144	86.490	**47.953**	**85.662**	84.693	**85.100**
GP+Edit	49.581	86.942	85.979	86.391	46.211	84.972	**84.725**	84.781
GP+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
GP+Edit+EV	48.503	86.803	85.961	86.313	44.538	85.069	83.961	84.442

表 3: 预训练语言模型实验结果

通过对mBART模型结果的分析，我们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（50.317），而BERTScore的精准率、召回率和F1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的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，这可能是因为附加的语法信息为模型提供了关键的上下文支持。

对于mT5模型而言，采用单独附加语法术语（GTs）的输入策略表现出最优的BLEU分数（47.953），并在BERTScore的精准率和F1分数上获得了相对较高的评分（分别为85.662和85.100）。这一现象说明了语法术语信息对于mT5生成反馈评语有极大的辅助作用。

当采用组合策略（GP+Edit+EV）时，我们注意到尽管这一组合在BLEU分数上均不是效果最佳的情况，但在mBART和mT5模型上的BERTScore F1分数有所提升，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在处理英文到中文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。

鉴于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价，其准确率分别为62.84和59.27。

第二十三届中国计算语言学大会论文集，第134页-第149页，太原，中国，2024年7月25日至28日。
卷1：主会论文

5.1 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，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对于语言习惯的评语，即其原始的写法在语法上并不完全错误但是存在更符合习惯的说法，这部分实例也是目前的GEC模型无法提供更多修正信息的。总的来说，GECToR能够标记出语法错误修正的实例不及全部数据条目的50%，如表4所示。

	反馈评语数	错误修改数	错误修改比例
训练集	4,868	2,383	48.9523%
开发集	170	80	47.0588%
测试集	215	90	45.5814%

表 4: GECToR修正错误数量及比例

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，但是我们发现这些仅有部分数据带有的信息仍然起到了重要作用。如表5所示，对于mBART模型，尽管添加修正编辑和线索词语导致了在全部测试数据上效果的下降，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，这一方法势必带来更好的跨语言反馈评语生成效果。

	mBART		mT5	
	全部数据	标记附加信息部分	全部数据	标记附加信息部分
Bare	47.7391	48.5225	40.5210	41.0460
Edit	43.4723 (-4.2668↓)	45.2130 (-3.3096↓)	42.1260 (1.6050↑)	46.6778 (5.6318↑)
EV	45.1406 (-2.5985↓)	48.1155 (-0.4070↓)	41.1802 (0.6593↑)	43.2662 (2.1803↑)

表 5: 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

语法术语预测模型效果 本研究基于附加额外一层线性层的RoBERTa模型，通过修改损失函数的计算方式和索引位置的计算策略，提高了其预测出现频率Top-10语法术语的准确性。最终，语法术语预测的效果如表6所示，该表显示了所提交的模型在开发数据集上的多标签性能。

如表7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，生成模型的BLEU值相较于目前的结果有较大幅度的提升，因此，未来对于预测语法术语的多分类任务，提高其准确性或是增加可预测语法术语的数量（如前20个高频率的语法术语）都将对提升反馈评语生成模型的表现提供帮助。

EMR	Precision	Recall	F1 Score
18.82	85.68	72.93	78.55

表 6: 中文语法术语预测模型结果评估

	mBART	mT5
GTs-ALL	55.0496	51.9758

表 7: 使用全部语法术语推理结果

5.2 mBART和mT5表现的差异分析

我们观察到，在跨语言反馈评语生成任务中，尽管mT5的模型参数量较大，但其表现却不如参数量较小的mBART。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，其训练目标与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使得mBART在相关任务上具有更优的表现。相比之下，虽然mT5是一个用途广泛的多语言模型，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，这可能增加了任务的复杂性。再者，由于mBART与mT5在模型结构上的差异，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，将完全相同=的提示信息附加到模型输入中对mBART和mT5两个模型产生了截然不同的影响。对mBART模型而言，添加修正编辑和线索词语后，其BLEU值仅有很小幅度提升，甚至不及直接对初始模型进行微调的表现效果。相反，mT5模型在加入这些信息后，其BLEU值显著高于直接微调的原始模型。

我们推测，这种差异主要源于模型架构的不同。尽管mT5和mBART均基于Transformer架构的Seq2Seq模型，它们处理输入和输出的方式却有所区别。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，因此更适合处理结构化的输入，这使得模型可以将这些输入视为明确的任务指令。相反，mBART以自编码方式预训练，专注于文本恢复，如重构乱序或掩盖的文本，这使其可能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。附加信息丰富了输入文本的上下文，使mT5能够更有效地利用这些信息，从而适应多变的输入格式。相对而言，mBART可能更依赖输入的一致性和结构，它在处理更直接和传统的文本生成任务中可能表现更佳。然而，在生成反馈评语的任务中，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，从而使得额外信息反而干扰了对原始文本的处理，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了更丰富的上下文逻辑信息，使其表现更好。

5.3 更多附加信息导致模型效果下降

根据我们的假设，当模型的输入包含更多附加的提示信息时，其生成效果理应有所提升，这也是很符合直觉的。但是，我们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添加其中单一附加信息更有效，即添加更多附加信息反而导致了模型效果的下降。例如，表3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及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。

本研究认为导致这一现象的是由于信息过载导致的注意力分散，即将语法术语、修正答案和线索词语同时附加到输入可能导致信息量过大。对于基于Transformer架构的模型来说，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不直接相关，模型可能难以从中提取有效的特征，从而影响学习效果。虽然单独使用三者对模型生成反馈评语都是有用的信息，但它们可能在语义上存在重叠或者相互干扰，这会使模型在处理输入时产生混淆，难以区分哪部分信息更重要。如果语法术语、修正答案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。

5.4 案例分析

表8展示了基于流水线的预训练语言模型在生成反馈评语时，使用不同附加信息，导致生成效果差异的一个典型的示例。

该表前两行分别为学习者语句和人类教师提供的反馈评语，之后五行分别为不同方式下mBART模型的附加信息及生成的反馈评语等。表中红色字体为评语中存在错误的内容，而绿色字体为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，解释他们花时间的目的。但是，学习者错误的使用了介词“to”与“learning”进行搭配。不含任何附加信息的端到端方法生成的解释完全错误；包含语法术语或修正编辑的方法正确的解释了修改方式但错误原因分析错误；而附加线索词语的方式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。这一结果与添加各个附加信息的初衷相符，证明了方法的有效性。

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。	✓

表 8: 案例分析

6 结语

在本研究中，我们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（TLFCG）。首先，我们基于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-中跨语言反馈评语数据集。为应对这一任务，我们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。通过广泛的实验验证，我们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，我们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多个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错误类型生成反馈评语的效果差异。

致谢

感谢各位评审专家对本文的细致审阅和宝贵建议。他们的建议丰富了本文的内容，提高了本文的质量，更加深了我们的研究视角。衷心感谢他们对本研究的贡献。

参考文献

Nikolay Babakov, Maria Lysyuk, Alexander Shvets, Lilya Kazakova, and Alexander Panchenko. 2022. Error syntax aware augmentation of feedback comment generation dataset, December.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2022. Sentence-level Feedback Generation for English Language Learners: Does Data Augmentation Help?, December.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July. Association for Computational Linguistics.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing Grammatical Error Correction Systems with Explanations. May.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring Methods for Generating Feedback Comments for Writing Learning. In *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing, pages 9719–9730, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Mana Ihori, Hiroshi Sato, Tomohiro Tanaka, and Ryo Masumura. 2023. Retrieval, Masking, and Generation: Feedback Comment Generation using Masked Comment Examples. In *International Conference on Natural Language Generation*.

S. Ishikawa. 2013. The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian Learners of English. March.

Kunitaka Jimichi, Kotaro Funakoshi, and Manabu Okumura. 2023. Feedback comment generation using predicted grammatical terms. In *International Conference on Natural Language Generation*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. GPTEval: NLG Evaluation using GPT-4 with Better Human Alignment. March.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating Corpora for Research in Feedback Comment Generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France, May. European Language Resources Association.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared Task on Feedback Comment Generation for Language Learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK, August. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, and Masato Mita. 2023. A Report on FCG GenChal 2022: Shared Task on Feedback Comment Generation for Language Learners. In *International Conference on Natural Language Generation*.

Ryo Nagata. 2019. Toward a Task of Feedback Comment Generation for Writing Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China, November. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, August.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer, March.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT, February.

附录A.大语言模型反馈评语生成

为了探索大语言模型在跨语言反馈评语生成任务上的表现，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，二者均采用作者撰写本文时的最新版本，即gpt-3.5-turbo-0125和gpt-4-turbo-2024-04-09。

大语言模型反馈评语生成Prompt：

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中文反馈评语，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><<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，我们将需要模型生成反馈评语的学习者语句以示例同样的形式输入模型，有助于其仅生成反馈评语而不包含其他内容。为了使大语言模型得到更好的效果，我们同样在System Prompt中添加了ChatGPT的指令：

You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture.\n Knowledge cutoff: 2023-12\n Current date: 2024-05-12

附录B.大语言模型方法结果与分析

由于以下原因我们决定仅使用大语言模型辅助人工进行评语翻译，而最终使用预训练语言模型进行反馈评语生成：

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要准确，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学习效益。我们的研究显示，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。

在实验中，我们人工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，预训练模型的输出在语言与人类教师的评语更为接近，这有助于增强反馈的接受度和效果。相反，大语言模型虽然能够生成语法结构正确的句子，但其风格和用词往往缺乏针对性和指导性细节，有时可能显得过于机械或与教学语境不够贴合。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，但由于对其的微调需要大量的计算资源和训练数据。因此，导致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，更加难以对齐人类教师反馈评语的内容、语言和组织形式。

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了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加长的内容。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于人工标注结果更为接近，结果如表9所示。

解释的准确性：在生成具体针对学习者错误的反馈评语时，准确性至关重要。我们评估了在few-shot情况下，大语言模型GPT-3.5、GPT-4和使用修正编辑+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。在这里我们额外添加了宽松的人工评价指

模型/方法	平均评语长度（字符）
人工标注	59.72
mBart-Large（Edit+EV）	56.26
GPT-3.5 （few-shot）	115.09
GPT-4 （few-shot）	119.45

表 9: 不同方法生成反馈评语的平均长度

标，即生成的反馈评语中只要包含有错误相关的信息，我们即认为它是正确的。我们的实验表明，GPT-3.5表现较差，而GPT-4展现出显著的鲜果提升，特别是在严格的人工评估中超过了GPT-3.5约32.56%。同时，使用Edit+EV方法的mBART模型也表现出色，特别是在宽松评估中接近80的高分。这一结果突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准确性的重要性。

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↑)

表 10: 大语言模型于预训练语言模型效果的人工评估

从上表可以看出，GPT-4在few-shot设置下对反馈生成任务的适应性和效果显著优于GPT-3.5，而mBART-Large则利用其特有的附加修正编辑和线索词语的策略，表现出稳定而高效的反馈生成能力。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GPT-3.5，较小幅度低于包含约1800B参数规模的GPT-4，我们认为这足够令人兴奋。这意味着相较于通用的大语言模型，预训练模型在本研究的特定任务上进行了深入优化，同样能够很好地适应语言学习反馈的需求。虽然我们的结果尚不及最新的GPT-4-Turbo模型，但其仍然提供了有价值的见解和创新，我们希望以此为未来的研究奠定基础。

其他原因：

1.资源的可持续性：大型模型如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，特别是资源有限的环境。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。

2.可定制性：预训练模型更容易根据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，更好地适应教育场景。

随着大语言模型的发展和微调技术研究，大模型在跨语言反馈评语生成任务上的表现会更加优异。我们也将持续关注相关的研究成果，并将其作为我们未来的研究工作。

附录C.大语言模型评估方法

C.1 Prompt内容

大语言模型有监督评估Prompt：

> 你是一位具有语言学背景的评价者。你需要对反馈评语生成模型针对语言学习者语句给出的反馈评语进行评价。
> 反馈评语生成任务是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中文反馈评语，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?
> 评价使用0至2的评分系统基于特定评估标准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。即便反馈中包含了参考内容未提及的相关信息，只要这些信息直接关联到错误，也被认定为完全正确。部分正确（1）的评分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的错误，且只需修改几个词便能让内容更准确，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，则被评为不正确（0）。
请直接给出评分，而不包含其他内容。
学习者语句：“{{学习者语句}}”
参考反馈评语：“{{参考反馈评语}}”
待评反馈评语：“{{待评反馈评语}}”
评分（0-2）：

大语言模型无监督评估且不提供评估维度Prompt：

你是一位具有语言学背景的评价者。你的任务是根据指标对反馈评语生成模型针对语言学习者语句给出的反馈评语进行评价。
反馈评语生成任务是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中文反馈评语，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。
请确保您仔细阅读并理解这些说明。请在审阅时保持本文档处于打开状态，并根据需要进行参考。
评估步骤：
1.仔细阅读学习者语句，找出错误位置错误的原因。
2.阅读反馈评语并将其与学习者语句中的错误进行对照。检查反馈评语是否解释了学习者语句中的错误和原因，以及是否以清晰且合乎逻辑的方式呈现。
3.根据评估维度，按照0到2的范围打分，其中0为最低，2为最高。
学习者语句：“{{学习者语句}}”
反馈评语：“{{待评反馈评语}}”
评估表（仅给出综合分数而不需要任何说明）：
- 综合分数（0-2）：

大语言模型无监督评估且提供评估维度Prompt：

你是一位具有语言学背景的评价者。你的任务是根据指标对反馈评语生成模型针对语言学习者语句给出的反馈评语进行评价。
反馈评语生成任务是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中文反馈评语，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。
请确保您仔细阅读并理解这些说明。请在审阅时保持本文档处于打开状态，并根据需要进行参考。
评估维度：
1.精确性（0-2），即评价反馈评语是否正确指出了学习者语句中的错误，并且正确解释了为什么这个结构是错误的或不合适的，前提是解释中的错误原因是真实存在的。评分标准：2-完全正确，清楚地指出错误并提供了正确的规则或原因；1-部分正确，可能识别了错误，但解释不够详细或部分错误；0-错误识别或解释错误，误导学习者。
2.明确性（0-2），即评价反馈评语是否表达清晰，无歧义，学习者能够轻易理解。评分标准：2-非常清晰，使用简洁的语言明确表达错误原因；1-表达较为清晰，但可能存在少量模糊或复杂的表述；0-表达含糊，难以理解，可能导致学习者混淆。
3.相关性（0-2），即评价反馈是否专注于学习者语句中标记的错误，没有偏离主题。评分标准：2-完全相关，专注于指定错误，没有无关内容；1-大体相关，但包含一些不必要或边缘的信息；0-反馈中包含大量无关信息，与标记的错误无关。
4.教育价值（0-2），即评价反馈是否提供了帮助学习者改正错误和避免未来错误的具体建

议或例子。评分标准：2-不仅指出错误，还提供了如何改正的建议或规则的深入解释；1-提供了基本的改正建议，但缺乏深入解释；0-没有提供改正建议或解释，无助于学习者的长期学习。
评估步骤:
1.仔细阅读学习者语句，找出错误位置错误的原因。
2.阅读反馈评语并将其与学习者语句中的错误进行对照。检查反馈评语是否解释了学习者语句中的错误和原因，以及是否以清晰且合乎逻辑的方式呈现。
3.根据评估维度，按照0到2的范围打分，计算其平均值并取整（综合分数），其中0为最低，2为最高。
学习者语句："{{学习者语句}}"
反馈评语："{{待评反馈评语}}"
评估表（仅给出综合分数而不需要任何说明）：
- 综合分数（0-2）：

C.2 结果与分析

使用大语言模型（GPT 3.5和GPT 4）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的结果如下表所示。

		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	
生成模型	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

表 11: 大语言模型评估结果

在无监督评估过程中，当Prompt没有指明评估维度时，GPT-3.5模型和GPT-4模型都对大语言模型生成的反馈评语给出了较高的分数。这是由于在没有评分参考和示例时，大语言模型并没有清晰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。但当Prompt包含评估维度时，GPT-4模型降低了对大预言模型结果的分数，并提高了对预训练语言模型的打分，这说明评分维度的加入使得大语言模型更加明确了这一任务及其评估方式。

在有监督的评估过程中，通过以人工标注的反馈评语作为参考，大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。一个有趣的现象是大语言模型都更加倾向于自己生成反馈评语，这说明大语言模型在作为模型效果的评估者时可能会带有偏见。

总体而言，应该使用能力更强的模型对能力较弱的模型进行评估，正如老师评价学生一样，所以在此我们倾向于采信GPT-4作为评估者的结果，尽管其会更偏向自己生成的答案。GPT-4与人工评估的Kappa一致性分别为4.52、15.43和34.30，表明大语言模型和人类评估者之间的一致性是公平的。这也证明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。