恢复节奏：使用 Transformer 模型对孟加拉语（一种低资源语言）的标点符号进行修复

Md Obyedullahil Mamun[1], Md Adyelullahil Mamun[2[0000−0002−6314−228X]], Arif Ahmad[3[0000−0002−2649−8981]], and Md. Imran Hossain Emu[1]

[1] Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh
{ obyedullahilmamun,mdihemu } @gmail.com
[2] BRAC University, Dhaka, Bangladesh
md.adyelullahil.mamun@g.bracu.ac.bd
[3] North East University Bangladesh (NEUB), Sylhet, Bangladesh
arif@neub.edu.bd

Abstract. 标点符号恢复增强了文本的可读性，并且对于自动语音识别（ASR）中的后处理任务至关重要，尤其是对于像孟加拉语这样的低资源语言。在这项研究中，我们探讨了基于 Transformer 的模型的应用，特别是 XLM-RoBERTa-large，用于自动恢复无标点孟加拉语文本中的标点符号。我们专注于预测四种标点符号——句号、逗号、问号和感叹号，适用于不同文本领域。为了应对带注释资源的稀缺性，我们构建了一个大型的、多样的训练语料库，并应用了数据增强技术。我们表现最佳的模型在增强因子为 $\alpha = 0.20$ 的情况下，在新闻测试集上达到了 97.1 % 的准确率，在参考集上达到了 91.2 %，在 ASR 集上达到了 90.2 % 的准确率。结果显示对参考和 ASR 转录有很强的泛化能力，证明了模型在真实世界噪音场景中的有效性。这项工作为孟加拉语标点符号恢复建立了强有力的基线，并贡献了公开可用的数据集和代码，以支持未来对低资源自然语言处理的研究。

Keywords: punctuation restoration · punctuation marks · deep learning · transformer models · natural language processing.

1 介绍

标点符号恢复是一个关键的后处理步骤，它增强了 ASR 生成的转录稿的可读性和可用性，支持一系列下游自然语言处理（NLP）任务，如翻译、摘要和情感分析 [10] [15] 。没有适当的标点符号，句子之间的语义边界会变得模糊，导致歧义和 NLP 管道的有效性降低。

早期的模型通过使用词汇特征和统计方法解决这个挑战，例如在大规模语料库上训练的条件随机场（CRF）[14] [23] 。深度学习的发展引入了更有效的技术，如长短期记忆网络（LSTM）、卷积神经网络（CNN），以及最近的基于转换器的模型 [4] [8] [24] [20] 。

尽管像 BERT [6] 和 RoBERTa [13] 这样的转换器模型在各种 NLP 任务中已经取得了成功，但它们在低资源语言（例如孟加拉语）中的标点恢复应用仍然有限。主要的挑战包括注释语料库的稀缺、缺乏标准化的基准测试，以及训练数

据（干净且结构化的文本）与真实世界用例（嘈杂的 ASR 输出）之间的领域不匹配。

为了解决这些问题，我们利用包括《Prothom Alo》和《The Daily Star》等主要报纸以及书籍抄本和在线平台在内的公开可用资源，构建了一个大型、多样化的孟加拉语数据集。此外，我们为不同类型文本的标点恢复奠定了强有力的基准，包括结构化新闻文章、一般参考文本和嘈杂的 ASR 转录文本。本研究的主要贡献如下：

1. 探索使用基于 transformer 的语言模型进行孟加拉语标点符号还原。
2. 提出一种使用数据增强技术提高模型性能的新策略。
3. 创建和评估孟加拉语的训练数据集，并为此任务提供基准结果。
4. 解决恢复四个关键标点符号的问题：句号（|）、逗号（,）、问号（?）和感叹号（!）。
5. 公开源码和数据集 [4] 以促进未来的研究。

标点恢复在引入基于 transformer 的架构后有了显著进展，其准确性和稳健性方面始终优于传统的 CNN 和 RNN 模型。例如，[20] 突出显示了 transformer 如何改善联合标点预测和语音转文本翻译，从而显著提升了转录质量。

由于自注意机制能够捕捉长距离依赖关系和上下文提示，Transformer 在序列标注任务中特别有效 [19] 。它们能够并行处理输入的能力进一步提高了训练效率 [11] 。对于标点恢复而言，这些特性至关重要，尤其是当标点提示距离它们在句子中的相关位置较远时 [2] 。

然而，尽管有这些优势，孟加拉语标点恢复仍然相对未被深入研究 [3] 。大多数孟加拉语研究依赖单语架构或有限的资源。缺乏大规模的、标注的语料库和标准化的评估指标为模型的训练和基准测试带来了重大挑战。

近期的研究提出使用数据增强来缓解数据稀缺问题。受一般策略 [21] 启发，诸如同义词替换、随机插入和反向翻译 [17] 等技术已经被应用于序列任务。然而，孟加拉语复杂的形态使得简单的增强方法出现问题，例如，修改后缀或引入不连贯的短语。因此，维护语法和形态的语言学知情方法更为有效 [18] 。此外，最初为文本分类任务设计的技术 [7] 已经成功地被调整和改进用于序列标注，显著提高了模型的鲁棒性。这些策略与多语言建模结合，为孟加拉语标点符号恢复开发出更具普适性的系统做出了贡献。

像 XLM-RoBERTa 这样的多语言 Transformer 模型在低资源环境中显示出潜力 [19] 。这些模型利用跨语言的共享嵌入，可以将知识从高资源语言转移到低资源语言。此外，诸如适配层和参数高效方法 [9] 等微调技术进一步增强了它们在资源受限场景中的适用性。

此外，文献强调了仔细的预处理分词、标准化、降噪以及使用精确度、召回率和 F1-score [2] [16] 进行严格评估的重要性。这些做法有助于开发能够在不同文本领域中进行泛化的稳健模型。

总之，文献支持在标点恢复方面，基于 Transformer 的架构、多语言建模和智能数据增强在像孟加拉语这样的低资源语言中的日益重要性。我们的工作基于这些策略，为未来的研究奠定了坚实的基础。为了推进孟加拉语的标点恢复，特别是考虑到其作为低资源语言的地位，我们开发了一个由多样化文本来源组成的综合数据集。主要语料库来源于一个公开可用的孟加拉语报纸文章数据集，该数据集提供了大部分的训练资料。具体来说，在总共约 210 万个标记中，大约 130 万个标记来自这个参考语料库。数据集中这一重要组成部分确保了当代书面孟加拉语的基础语言多样性。

为了拓宽标点符号现象的范围，特别是感叹号的出现，我们纳入了更多来自文学和表达领域的文本。这些来源，包括 Kishor Alo (www.kishoralo.com)、Kali O Kalam (www.kaliokalam.com) 和 E Banglalibrary (www.ebanglalibrary.com) 等叙述性和文学性网站，经常具有更具情感和动态的标点符号使用。

这些补充材料通过增加更多细微的标点模式实例丰富了数据集，从而为标点恢复创建了一个更均衡和具有挑战性的测试环境。

数据集结构和标注：如表 **??** 所示，数据集被划分为训练集、开发集和测试集。训练集包含约 217 万标记，开发集包括约 20.7 万标记，合并的测试集总计约 12.7 万标记。所有子集都进行了手动标点符号注释，确保了一个全面的训练资源，涵盖了多种语言形式和风格。

补充数据集：为了评估模型在口语场景下的表现，我们整合了两个额外的数据集。首先，手动转录来自约 65 分钟的孟加拉语短篇故事阅读的独白风格。这些文本主要来自感叹号使用更频繁的领域，并通过手动标注标点符号为清晰的、人工生成的转录提供基准。

其次，通过 Google Cloud 的语音 API 处理相同的 65 分钟音频，获得了 ASR 转录结果。由于 ASR 输出缺乏标点符号，对其进行了人工标注，以评估模型处理嘈杂、自动生成的转录文本的能力。这种方法模拟了现实场景，在这些场景中，模型必须恢复语音识别输出中的标点符号，以及其他嘈杂文本流中的标点符号。

我们的标注框架侧重于恢复五种标点符号类别：句号（| ）、逗号（,）、问号（?）、感叹号（!），以及 O（表示无标点符号）。在人工创建的数据集和 ASR 生成的数据集中，这些标点符号都被全面标注，以确保一致性，并在不同文本形式中实现稳健的模型评估。

值得注意的是，虽然感叹号在训练和开发集中的标记中所占比例不到 1 %，但在某些测试集中，尤其是 Test (Ref.) 和 Test (ASR) 子集中，由于包含文学和叙述性来源，这些感叹号的比例超过了 1 %，因为在这些来源中感叹更为普遍。相比之下，基于新闻的测试集保持了较低的感叹号使用频率，反映了新闻文本较为正式和客观的风格。

在模型训练之前，所有文本数据均经过预处理以确保质量和一致性。此处理流程包括噪声去除、格式规范化，并通过 `subword-nmt` 库实现的字节对编码（BPE）等技术进行子词标记化。鉴于孟加拉语的复杂性和资源匮乏性质，子词标记化在处理这种语言的广泛词汇和书写变体时起到了重要作用。

自定义 Python 脚本然后将所有标点符号转换为适合监督学习任务的相应标签。结合前述的分词和预处理步骤，这些措施提供了一个干净、标准化的输入，支持稳健且可重现的模型开发。

2 方法论

图 1 展示了我们基于 XLM-RoBERTa 的架构，因其在低资源语言处理中的有效性而被选中。XLM-RoBERTa 通过多语言掩码语言模型（MLM）目标进行预训练，基于 XLM 的结构，但训练在一个显著更为广泛和多样化的数据集上。这个数据集包括超过两个 TB 的过滤后的 Common Crawl 数据，涵盖一百种语言 [22] 。XLM-RoBERTa 的多语言能力和在多样化数据集上的预训练 [5] 使其特别适合捕捉孟加拉文文本和标点的复杂性。该模型进一步在孟加拉语数据集上进行了微调，以提升其在标点恢复上的性能。

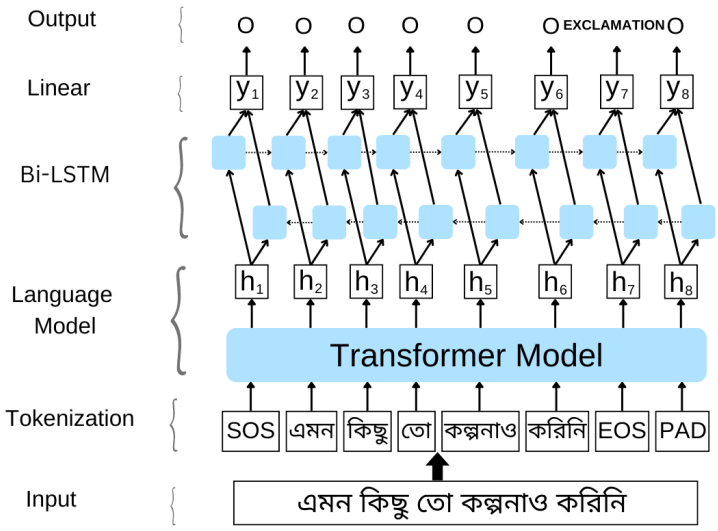

Fig. 1. 孟加拉语标点符号恢复的模型架构

在模型训练过程中，我们选择了适当的超参数，包括学习率、批处理大小和训练周期的数量。使用 transformer 模型进行标点符号恢复的方法与 [1] 中描述的类似。

在我们的方法中，文本中的每个标记由预训练语言模型获得的 d 维嵌入向量表示。该嵌入被传递到一个具有 h 个隐藏单元的双向长短期记忆（BiLSTM）层。BiLSTM 架构使模型能够在进行预测时利用过去和未来的上下文，这对于准确恢复标点符号至关重要。LSTM 层的前向和后向传递输出在每个时间步被连接。这些连接的输出然后由一个具有五个输出神经元的全连接层处理。每个神经元对应于四种标点符号（句号、逗号、问号、感叹号）中的一种和一个额外的 O 标记，代表非标点符号。

我们提出了一种数据增强方法，灵感来源于 [21]，并针对我们的数据集进行了定制。该方法解决了自动语音识别（ASR）中的常见错误，如替换、插入和删除。值得注意的是，与逗号、句号和问号等其他标点符号相比，我们的数据集中感叹号的实例较少。为了平衡这一点，我们应用了增强技术以引入更多的感叹号，尽管分布未能达到完美平衡。

我们采用了以下三种增强技术：

1. Substitution： 随机替换标记为特殊的未知标记以模拟替换错误。
2. Deletion： 随机删除标记以模拟 ASR 删除错误。
3. Insertion： 在随机位置插入未知标记以模拟插入错误。

我们假设替换、插入和删除错误的频率会有所不同，并以不同的方式影响模型的性能。为了管理这种变化，我们使用了三个可调参数：

1. Token Change Probability： 更改一个符号的总体概率。
2. Substitution Probability： 将一个符号替换为未知符号的概率。

3. Deletion Probability ： 从序列中移除一个符号的概率。

。

插入概率通过计算替换和删除概率后剩余的部分来计算。这种方法允许在数据集中有控制地引入错误，模拟现实中的 ASR 缺陷，帮助提高我们模型的鲁棒性。我们的标点恢复实验采用了从 HuggingFace Transformers 库中获取的预训练基于变压器的模型。输入文本经过使用 Byte-Pair Encoding (BPE) 的模型特定分词器进行预处理，能够高效地表示孟加拉语中的子词和罕见词标记。每个输入序列被截断或填充到最大长度为 256 个标记，并由特殊的开始和结束序列标记框住。在注意力机制中，填充标记被掩盖，以确保模型关注有意义的输入。

训练过程中使用的迷你批量大小为 8，并在每个周期前对数据进行打乱，以减轻过拟合并增强模型的泛化能力。大规模 Transformer 模型的学习率设置为 5e-6，基础模型的学习率设置为 1e-5，因为初步的探索性运行表明，这些学习率下的收敛稳定。Adam 优化器 [12] 被应用于共计十个周期，并将用于序列建模任务的 LSTM 隐藏维度设置为与词元嵌入大小相匹配，以保持架构一致性。除非另有说明，否则超参数保持 Devlin 等人 [6] 提出的默认设置。模型选择是通过在每个周期后评估开发集上的性能来进行的，性能最佳的模型将用于最终测试。

为了增强训练数据的鲁棒性和多样性，我们采用了受先前已建立方法启发的数据增强策略。增强强度由参数 α 控制，该参数决定一个句子中的多少比例的标记可以用于增强。我们实验了三个值：$\alpha = 0.10$ 、$\alpha = 0.15$ 和 $\alpha = 0.20$ 。对于每个设置，替换和删除操作被以固定的 0.4 的比率应用。

这些受控扰动在训练数据中引入了词汇和结构变化，有效地扩展了训练分布。这使得模型能够更好地泛化到多样化和未见过的输入。

2.1 评估数据集

最终模型在三个测试数据集上进行了评估，每个数据集反映了不同的文本领域和特征：

1. 新闻：一个精选的孟加拉新闻文章集，代表了结构化的正式文本。
2. 参考 (Ref)：一个广泛选择的参考文本集合，涵盖了更加通用、多体裁的书面孟加拉文。
3. 自动语音识别 (ASR)：由自动语音识别系统生成的转录文本，提供了模型在嘈杂和不太受控的输入条件下恢复标点符号能力的洞见。

3 结果与讨论

我们评估了模型在新闻、参考 (Ref.) 和 ASR 三个不同测试数据集上恢复四种标点符号的能力——逗号、句号、问号和感叹号。性能通过每种标点符号类别的精确度 (P)、召回率 (R) 和 F1-分数 (F1) 以及总体准确性进行衡量。表 1 提供了这些结果的详细总结。

该模型在新闻数据集上的表现最为突出，这可能反映了新闻文本的结构化和正式性。相反，在包含更多样化语言使用以及风格和领域复杂性更大的参考和 ASR 数据集上的表现有所下降。

在所有数据集中，一个一致的挑战是感叹号的检测。这一困难不仅在包含更多表现性和口语化内容的参考集和 ASR 集中显而易见，而且在新闻数据集中也

Table 1. XLM-RoBERTa-large 在孟加拉标点恢复任务中的表现：基于不同数据集，包含和不包含增强

Model	Test	Comma			Period			Question			Exclamation			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
XLM-RoBERTa-large	News	83.7	77.6	80.5	86.5	92.2	89.3	71.2	83.9	77.1	67.9	33.6	45.0	84.8	84.9	84.9
	Ref.	55.9	47.4	51.3	74.2	79.4	76.7	53.6	74.1	62.2	51.2	32.7	40.0	66.4	66.7	66.5
	ASR	54.9	49.7	52.1	69.5	75.0	72.1	47.7	71.3	57.2	52.2	37.7	43.8	62.4	65.0	63.6
XLM-RoBERTa-large + Aug. ($\alpha = 0.10$)	News	82.7	77.2	79.9	86.5	91.6	89.0	77.7	81.1	79.3	64.4	36.1	46.3	84.6	84.5	84.5
	Ref.	55.5	47.1	51.0	75.3	79.1	77.2	61.4	64.5	62.9	47.3	38.1	42.2	67.6	66.0	66.8
	ASR	53.8	49.9	51.8	69.9	73.5	71.6	53.4	60.6	56.8	42.1	39.0	40.5	62.4	63.3	62.8
XLM-RoBERTa-large + Aug. ($\alpha = 0.15$)	News	83.7	76.4	79.9	86.1	91.1	88.5	69.3	84.7	76.2	70.1	23.1	34.8	84.5	83.7	84.1
	Ref.	56.6	46.8	51.2	74.3	77.5	75.9	52.7	69.5	60.0	58.5	22.0	32.0	67.0	64.2	65.6
	ASR	55.3	49.0	51.9	70.2	73.0	71.6	45.8	70.7	55.6	59.4	27.1	37.2	62.9	62.8	62.9
XLM-RoBERTa-large + Aug. ($\alpha = 0.20$)	News	83.3	75.8	79.4	86.5	91.0	88.7	75.8	81.1	78.4	64.2	27.9	38.9	84.8	83.4	84.1
	Ref.	58.1	47.5	52.3	74.0	78.0	75.9	55.8	68.2	61.4	58.6	29.0	38.8	67.6	65.1	66.4
	ASR	57.1	48.3	52.3	69.2	73.5	71.3	49.7	65.1	56.4	55.1	35.7	43.3	63.5	63.0	63.3

存在。主要原因似乎是训练数据中感叹号相对较低的出现频率。由于可用的例子较少，模型难以学习准确预测这种标点类型的稳健模式。

4 消融研究

为了评估数据增强对模型性能的影响，我们进行了一个消融研究，对比了基本模型（XLM-RoBERTa-large，无增强）和使用不同增强强度训练的变体。增强参数 α 指定了每个句子中有资格进行修改的标记比例。我们评估了三个增强水平：$\alpha = 0.10$ 、$\alpha = 0.15$ 和 $\alpha = 0.20$ 。在所有情况下，替换和删除操作均以 0.4 的固定比率应用。

表 1 展示了在三个测试领域中——新闻 (News)、参考 (Ref.) 和自动语音识别 (ASR)——评估的四种标点符号类别（逗号、句号、问号和感叹号）的精确度 (P)、召回率 (R)、和 F1 分数 (F1)。这种比较使我们能够独立出增广相对于基础模型的效果。此消融研究的主要观察结果如下：

1. News Test Set: 基础模型在清晰的新闻文本上表现出色，F1 分数达到了 84.9 %。使用 $\alpha = 0.10$ 、$\alpha = 0.15$ 和 $\alpha = 0.20$ 扩展后的模型也达到了相当的性能（分别为 F1 = 84.5 %、84.1 % 和 84.1 %），这表明数据扩增在结构化领域中没有降低性能。
2. Reference Test Set: 在更为多样化的 Ref. 数据集上，数据扩增带来了明显的改进。总体 F1 分数从 66.5 %（无扩增）提高到 66.8 %（$\alpha = 0.10$ ）和 66.4 %（$\alpha = 0.20$ ），其中 $\alpha = 0.15$ （65.6 %）有轻微变化，表明增强了对多样语言结构的泛化能力。
3. ASR Test Set: 增强后的模型在嘈杂的语音识别（ASR）转录上保持或略有提高了性能。基础模型实现了 63.6 % 的 F1 分数，而包含 $\alpha = 0.10$ 、0.15 和 0.20 的模型分别取得了 62.8 %、62.9 % 和 63.3 %。这些结果表明在艰难条件下性能稳定。

4. Low-Frequency Punctuation: 数据扩增特别有助于处理稀有标点。例如，新闻集中的感叹号 F1 分数从 45.0 % （无扩增）提高到 46.3 % （$\alpha = 0.10$ ）。同样，对于 Ref. 和 ASR 集合，数据扩增提高或稳定了问号和感叹号的 F1 分数。

4.1 测试结果的比较分析

XLM-RoBERTa-large + 增强（ $\alpha = 0.20$ ）模型在新闻测试集上表现出高准确率 (97.1 %)，但在参考 (91.2 %) 和 ASR (90.2 %) 测试集上的表现有所下降。这些结果汇总在表 2 中。这种差异表明该模型在处理常规和经过良好编辑的文本时表现良好，但在处理结构较差、口语化或源自语音的内容时面临挑战。

Table 2. 模型在测试集上的准确率

Test	Accuracy (%)
News	97.1
Ref.	91.2
ASR	90.2

为深入了解模型的误分类，我们检查了从每个测试数据集中得出的混淆矩阵（图 2 ）。这些矩阵揭示了模型在预测缺乏标点符号（表示为 欧 ）方面在所有领域中始终表现出色，这与其较高的整体准确性一致。

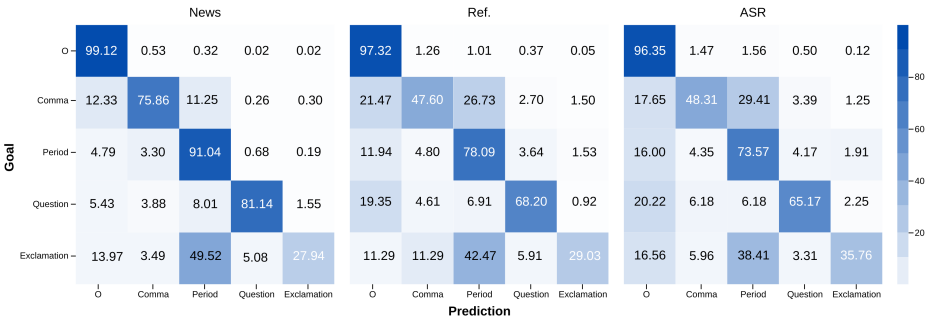

Fig. 2. 混淆矩阵显示了测试集中文种类的分类准确性。

尽管整体表现令人满意，该模型在区分某些标点符号方面表现出明显的困难，特别是在参考和 ASR 测试集中的逗号、句号和问号。这些挑战反映在其各自的混淆矩阵中增加的非对角线活动，表明错分类的比率较高。

例如，在参考集合中，只有 47.60 % 的逗号被正确预测，常被错误分类为句号 26.73 % 和无标点类别 21.47 % 。同样，ASR 集合显示逗号的正确分类率

为 48.31 % ，与句号 29.41 % 和无标点类别 17.65 % 有显著混淆。句号和问号也常被混淆。在 ASR 数据中，只有 65.17 % 的问号被正确识别，误分类为无标点 20.22 % 和逗号 6.18 % 的情况较多。感叹号尤其成问题，只达到 35.76 % 的准确率，与句号 38.41 % 和无标点实例 16.56 % 有很高的混淆度。

这种较高的混淆可以归因于数据集的固有特性。语音识别（ASR）转录通常包含不流利现象、不一致的句子边界以及口语中典型的韵律歧义，这些都使标点预测更加复杂。参考数据集来源于异构来源，表现出多样的句法和风格惯例，进一步增加了歧义性。相比之下，新闻数据集显示出更干净的模式，在混淆矩阵中有强烈的对角线对齐，例如对于句号的 91.04 % 准确性和对于问号的 81.14 % ，突显出模型在正式、结构化文本中的改进表现。

为了更好地优化真实世界语音识别数据的标点恢复，未来的工作可以探索使用带有标点和不流利标记的语音衍生语料库进行有针对性的微调。这将使模型能够学习与口语话语相关的上下文线索。在如对话记录、播客或者手动清理的语音识别输出等语音风格文本上进行领域自适应预训练或微调，可以进一步提高稳健性。此外，课程学习策略，即模型逐渐暴露于越来越多噪音的数据，可能会增强其对真实世界条件的泛化能力。最后，在多模态框架中整合与音频对齐的韵律特征（例如暂停持续时间、音高变化）也是很有前景的，尽管这超出了本工作的范围。我们的工作通过利用基于 Transformer 的架构（特别是 XLM-RoBERTa Large），为本质上资源匮乏的语言——孟加拉语——的标点恢复提出了一种有效的方法。鉴于标注数据集的稀缺性和这一领域先前工作的有限性，我们的研究旨在解决孟加拉语自然语言处理中的一个关键空白。为此，我们引入了有针对性的数据增强技术，旨在提高模型的鲁棒性，特别是在处理嘈杂的语音识别输出时。

我们的实验结果表明，所提出的模型即使在语音转写的困难条件下，也能出色地恢复标点符号。通过提供公开可用的数据集和代码库，我们希望促进未来的研究并推动 NLP 社区内的合作。通过这一贡献，我们预期获得的见解不仅能增强孟加拉语标点符号恢复能力，还能作为解决其他低资源语言中类似挑战的蓝图。

References

1. Alam, T., Khan, A., Alam, F.: Punctuation restoration using transformer models for high-and low-resource languages. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). pp. 132–142. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.wnut-1.18, https://www.aclweb.org/anthology/2020.wnut-1.18
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate, http://arxiv.org/abs/1409.0473
3. Bijoy, M.H., Faria, M.F.A., E Sobhani, M., Ferdoush, T., Shatabda, S.: Advancing bangla punctuation restoration by a monolingual transformer-based method and a large-scale corpus. In: Proceedings of the First Workshop on Bangla Language Processing (BLP-2023). pp. 18–25. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.banglalp-1.3, https://aclanthology.org/2023.banglalp-1.3
4. Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector

5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747, https://www.aclweb.org/anthology/2020.acl-main.747
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, http://arxiv.org/abs/1810.04805
7. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 567–573. https://doi.org/10.18653/v1/P17-2090, http://arxiv.org/abs/1705.00440
8. Gale, W., Parthasarathy, S.: Experiments in character-level neural network models for punctuation. In: Interspeech 2017. pp. 2794–2798. ISCA. https://doi.org/10.21437/Interspeech.2017-1710, https://www.isca-archive.org/interspeech_2017/gale17_interspeech.html
9. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification
10. Jones, D.A., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D.A., Zissman, M.: Measuring the readability of automatic speech-to-text transcripts. In: 8th European Conference on Speech Communication and Technology (Eurospeech 2003). pp. 1585–1588. ISCA. https://doi.org/10.21437/Eurospeech.2003-463, https://www.isca-archive.org/eurospeech_2003/jones03_eurospeech.html
11. Kim, Y.J., Hassan, H.: FastFormers: Highly efficient transformer models for natural language understanding. In: Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing. pp. 149–158. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sustainlp-1.20, https://www.aclweb.org/anthology/2020.sustainlp-1.20
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization, http://arxiv.org/abs/1412.6980
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach, http://arxiv.org/abs/1907.11692
14. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields
15. Matusov, E., Hillard, D., Magimai-Doss, M., Hakkani-Tür, D., Ostendorf, M., Ney, H.: Improving speech translation with automatic boundary prediction. In: Interspeech 2007. pp. 2449–2452. ISCA. https://doi.org/10.21437/Interspeech.2007-644, https://www.isca-archive.org/interspeech_2007/matusov07_interspeech.html
16. Powers, D.M.W.: Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation . https://doi.org/10.48550/ARXIV.2010.16061, https://arxiv.org/abs/2010.16061, publisher: arXiv Version Number: 1
17. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. https://doi.org/10.48550/arXiv.1511.06709, http://arxiv.org/abs/1511.06709
18. Tariquzzaman, M., Anam, A.N., Haque, N., Kabir, M., Mahmud, H., Hasan, M.K.: BDA: Bangla text data augmentation framework. https://doi.org/10.48550/arXiv.2412.08753, http://arxiv.org/abs/2412.08753
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need, http://arxiv.org/abs/1706.03762

20. Wang, F., Chen, W., Yang, Z., Xu, B.: Self-attention based network for punctuation restoration. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2803–2808. IEEE. https://doi.org/10.1109/ICPR.2018.8545470, https://ieeexplore.ieee.org/document/8545470/
21. Wei, J., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks, http://arxiv.org/abs/1901.11196
22. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: CCNet: Extracting high quality monolingual datasets from web crawl data, http://arxiv.org/abs/1911.00359
23. Zhang, D., Wu, S., Yang, N., Li, M.: Punctuation prediction with transition-based parsing
24. elasko, P., Szymaski, P., Mizgajski, J., Szymczak, A., Carmiel, Y., Dehak, N.: Punctuation prediction model for conversational speech, http://arxiv.org/abs/1807.00543