分词很重要：改进对印度语言的零样本命名实体识别

1st Priyaranjan Pattnayak
University of Washington
Seattle
ppattnay@uw.edu

2nd Hitesh Patel
New York University
New York
hitesh.patel945@gmail.com

3rd Amit Agarwal
Liverpool John Moores University
Liverpool
amit.pinaki@gmail.com

***Abstract*—分词是自然语言处理（NLP）的一个关键组成部分，尤其是对于资源稀缺的语言，子词分割会影响词汇结构和下游任务的准确性。尽管字节对编码（BPE）是多语言模型中的标准分词方法，但由于在处理形态复杂性方面的局限性，其在低资源的印度语言中用于命名实体识别（NER）的适用性仍未被充分探索。在这项工作中，我们系统地比较了使用 IndicBERT 在低资源的印度语言（如阿萨姆语、孟加拉语、马拉地语和奥迪亚语）以及极低资源的印度语言（如桑塔利语、曼尼普尔语和信德语）中用于 NER 任务的 BPE、SentencePiece 和字符级分词策略。我们评估了内在的语言特性——分词效率、超出词汇表（OOV）率和形态保留，以及外在的下游表现，包括微调和零样本跨语言迁移。我们的实验表明，SentencePiece 在低资源的印度语言 NER 任务中是一种比 BPE 更具持续性良好表现的方法，尤其是在零样本跨语言设置中，因为它能更好地保持实体的一致性。虽然 BPE 提供了最紧凑的分词形式，但由于在未见过的语言上进行测试时错误分类甚至无法识别实体标签，它无法实现泛化。相比之下，SentencePiece 构成了一种更好的语言结构保留模型，对极低资源和形态丰富的印度语言（如桑塔利语和曼尼普尔语）提供了更优越的实体识别，并在跨文字脚本（如用阿拉伯语书写的信德语）上实现了高度泛化。结果表明，SentencePiece 是多语言和低资源印度 NLP 应用中 NER 的更有效的分词策略。**

***Index Terms*—Tokenization, Named Entity Recognition (NER), IndicBERT, Indic Language, Low Resource Language**

I. 介绍

多语种预训练模型已经变革了自然语言处理任务，但资源匮乏的印度语言由于语料库有限、形态复杂和文字多样性而被低估。印度语言有庞大的使用者基础，包括阿萨姆语（2500 万）、孟加拉语（超过 2.7 亿）、马拉地语（8300 万）、和奥里亚语（3800 万）[1] 。然而，极度资源匮乏的语言如桑塔利语（760 万）、曼尼普尔语（180 万）、和信德语（2500 万）由于数字化资源有限和语言复杂性面临显著的自然语言处理挑战 [2] 。命名实体识别（NER）作为信息提取的一个非常关键的任务，严重依赖于适当的分词去正确地分隔实体 [3] 。与分类或翻译相比，命名实体识别模型也受错误分词、实体拆分或合并的影响，从而阻碍性能。虽然这很重要，但分词方法对印度语言中命名实体识别的影响仍未被探索。在多语种模型如 mBERT [4] 和 IndicBERT [5] 中，字节对编码（BPE）[6] 是常见方法，但其如何处理印度语言形态和零样本推理尚未被研究。较早的研究 [4], [7], [8] 曾考察多语种模型的分词 [9] ，但跨语言实体识别仍然是一个尚待研究的话题。

分词影响了 Transformer 模型中的词表示。BPE 和 WordPiece [4] 等子词方法使用基于频率的合并，但由于词汇不足和复杂的词形变化，它们在资源较少的语言中效果不佳 [10] 。SentencePiece [11] 提供了更灵活的分段方法，比基于频率的方法更能保持语言结构。字符级分词可以完全保留词形结构，但会增加序列长度 [12] 。

本文通过内在和外在分析评估了三种分词策略——BPE、SentencePiece 和字符级。在内在评估中，涉及分词效率、OOV 率以及跨印度语言的形态保留，因字符级分词效率低且序列较长而被排除。在外在评估中，对 IndicBERT [5] 在孟加拉语和印地语上进行微调，并零样本转移到阿萨姆语、奥里亚语、马拉地语、曼尼普尔语、信德语和桑塔利语。

我们的结果显示，SentencePiece 在跨语言的命名实体识别 (NER) 中表现优于 BPE。尽管 BPE 能够实现紧凑的分词，但在泛化能力上不如 SentencePiece，并且在零样本环境中往往过于激进地使用 “O” 标签。SentencePiece 保留了更多的语言信息，提高了对 Santali 和 Manipuri 等形态复杂语言的实体识别。此外，它在处理书写形式的变化上表现更好，尤其是在阿拉伯文书写的信德语中，BPE 未能发挥作用。内在分析表明，虽然 BPE 降低了未登录词的比率，但在形态保持方面表现不佳，而 SentencePiece 在灵活性和泛化之间取得了平衡。主要贡献：

- 低资源印度语言的 BPE、SentencePiece 和字符级标记化比较。
- 评估内在语言特性和外在下游性能。
- 从内在研究结果中排除基于外在评估的字符级标记化。
- SentencePiece 优越的零样本跨语言迁移能力的展示。
- 关于分词对多语言自然语言处理的影响的见解，以及对低资源语言的更广泛意义。

我们的研究强调了分词在提升欠代表语言的自然语言处理中的重要性。通过展示 SentencePiece 在跨语言命名实体识别中的优越性以及字符级分词的不切实际性，我们为未来的多语言自然语言处理研究和实际应用提供了指导 [13] ，有助于开发出更高效和包容的语言模型。

II. 相关工作

A. 低资源自然语言处理的分词策略

分词是自然语言处理的核心，影响词汇量、模型泛化 [14] 和跨语言迁移，因为分词决定了如何在基于 transformer 的模型中表示单词和子词，特别是对于资源稀缺的语言 [15] 。字级分词在处理罕见单词时表现不佳，导致资源稀缺语言中不可接受的高词汇外（OOV）率。像字节对编码 (BPE) [6] 和 WordPiece [4] 这样的子词技术通过将单词拆分为频繁出现的子词单元来解决这个问题。

BPE 在诸如 mBERT 和 XLM-R [7] 等多语言模型中被广泛使用，它根据字符序列的频率进行组合，而 WordPiece 利用基于概率的子词选择进一步优化了这一过程 [16] 。SentencePiece [11] 消除了对空格的依赖，使得字符感知分割成为可能。字符级分词虽然能够保留词形，但非常冗长且计算成本高 [12], [17], [18] 。

尽管有所进展，分词仍然为高资源语言进行优化。BPE 在处理丰富形态 [19] 和未见实体时表现不佳，影响了具有屈折变化、词汇碎片化和多样化字符 [20], [21] 的印度语言。本研究系统评估了用于印度 NER 的分词策略，以在微调和零样本设置中应对这些挑战。

像 XLM-R [7] 和 IndicBERT [22] 这样的多语言模型依赖于子词分词，以在词汇量和表示效率之间取得平衡。虽然 BPE 支持精简的词汇表，但在形态丰富的低资源语言中表现不佳 [23]–[25] 。

分词会影响跨语言性能 [12] 。BPE 在相似语言中表现出色，但在零样本环境中表现不佳，经常将命名实体分割为不可识别的子词。SentencePiece 的字符感知分割是一种有前途的替代方法，因为先前的研究表明它比基于频率的方法更能保留形态结构 [26] ，尽管其在零样本命名实体识别中的有效性仍未得到充分探索。我们的研究通过比较 BPE 和 SentencePiece 在印度语言零样本实体识别中的表现来填补这一空白。Wang 等人（2022）[27] 的研究进一步支持了这一点，显示了对分词感知的适应可以增强跨语言迁移。

B. 印度语中的命名实体识别

NER 依赖于准确的分词以保留实体边界。与使用广泛文本表示的分类不同，NER 模型依赖于精确的分割——对齐错误会导致分类错误。

此前的印度语言命名实体识别（NER）研究主要集中于数据集创建和模型架构 [24] ，常常忽视了分词的作用。尽管已经探索了低资源 NER 的迁移学习 [23] ，但分词对零样本泛化的影响尚未被检验。

SentencePiece 的自适应分词可能会超越 BPE，尤其是在像桑塔利语和信德语这样语言体系多样的语言中，因为它可以在不同的书写体系间保持语言结构。我们通过对已知语言进行微调以及对未知语言进行零样本转移来评估这一点。

C. 与以往工作的比较

多语言自然语言处理研究已经广泛地探讨了分词技术，但其在跨语言命名实体识别（NER）泛化中的作用仍未被检验。BPE [6] 被广泛使用，但在低资源语言中仍未被充分探索。SentencePiece [11] 提供了一种不依赖空格的方式，但其在 NER 中特定的有效性尚未经过大量测试。

尽管已知在零样本 NER 转移中存在局限性，但像 XLM-R [7] 和 IndicBERT [22] 这样的模型依赖于 BPE。之前的工作 [23], [24] 强调了印度语言的 NER 挑战，但没有研究分词对跨语言实体识别的影响。

我们的工作通过研究标记化在跨语言命名实体识别中的作用，特别是在零样本设置中的应用，扩展了之前的研究 [28] 。我们使用内在和外在的评价方法，对多种标记化策略在印度语言中的表现进行评估：包括阿萨姆语、孟加拉语、马拉地语、信德语、桑塔利语、曼尼普尔语和印地语。

尽管标记化在基础层面上具有重要作用，但其对低资源印度语言中的命名实体识别（NER）和零样本迁移的影响尚未得到充分探索。我们的研究提供了：

我们的研究结果表明，SentencePiece 的灵活性增强了跨语言的泛化能力，而 BPE 在未见过的语言中处理实体破碎方面却较为困难。这项研究为优化多语言自然语言处理的分词，特别是对于欠代表的语言，提供了关键的见解。

在本研究中，我们系统地评估了低资源印度语言中的命名实体识别（NER）标记化策略。我们通过内在分析检验了三种标记化方法——字节对编码（BPE）、SentencePiece 和字符级别。基于内在评估的结果，我们淘汰了字符级别，接着仅对 BPE 和 SentencePiece 进行外在评估。具体而言，我们在印地语和孟加拉语的 NER 数据集上微调 IndicBERT，并评估其对阿萨姆语、马拉地语、奥里亚语、信德语、桑塔利语和曼尼普尔语的跨语言零样本泛化能力。

我们的方法论包括四个关键阶段：

- 使用 FLORES-200 数据集进行内部分析 [29] 。
- 在印地 XMATHX 与孟加拉语上对 IndicBERT 进行 NER 微调。
- 零样本迁移到六种未见的印度语言。
- 比较内在和外在性能的分词策略。

我们的数据集选择是基于需要在内在（无任务特定模型）和外在（通过微调 NER 模型）上评估分词。我们使用两个主要的数据来源：

- 用于内在分词评估的 FLORES-200++ 数据集。
- Naamapadam [30] NER 数据集用于 Assamese、Oriya & Marathi 的外部评估，以及一个用于 Sindhi、Santali & Manipuri 外部评估的手动注释数据集

1) 内在分词分析： 我们选择了 FLORES-200++ 数据集的一个子集，如表 I 所示，这是一个高质量的平行语料库，用于内在评估。它是少数覆盖低资源印度语言的公共资源之一，如阿萨姆语、桑塔利语、信德语和曼尼普尔语。其平行句子由人工翻译，支持单语和多语的分词分析。其规模（每种语言约 1K 句话）在计算效率和语言多样性之间达到平衡，实现稳健评估。我们分析了四种形态丰富语言中的分词行为：曼尼普尔语、桑塔利语、阿萨姆语和信德语。我们的评估比较了以下分词方法：

TABLE I
用于内在评估的 FLORES++ 子集概述

Language	Script	Sentences	Linguistic Properties
Assamese	Bengali	997	Inflectional
Sindhi	Arabic	997	Agglutinative
Manipuri	Bengali	997	Morphologically rich
Santali	Ol Chiki	997	Complex morphology

- 字节对编码 (BPE)：一种子词分割方法，它合并频繁字符序列，广泛用于变压器模型。
- SentencePiece：一种灵活的无监督分词模型，它去除了对空白的依赖，可以在字符或子词级别进行词的分割。
- 字符级：一种基于字符的分词策略，可以消除词汇表外（OOV）问题，但会导致序列长度显著增加。

我们使用三个内在指标来评估这些方法：

- 分词效率：测量每个单词产生的平均标记数。

- 词汇压缩率：衡量分词对比原始文本在减少词汇量方面的效率。
- 词汇表外词（OOV）率：计算被分割为多个子词的词的百分比。
- 形态保持：评估在分词后是否保留有意义的词结构。

针对外部评估的过滤分词策略：在内在分析之后，我们将字符级别从进一步研究中排除，因为它在 Fig 1 中表现出对印度语言次优的行为。字符级别显著增加了序列长度，使得基于 Transformer 的命名实体识别任务在计算上开销较大。它在词汇外率和形态保持方面表现更差，使其不适合于微调和零样本迁移。

基于这些发现，我们仅使用 BPE 和 SentencePiece 来微调 IndicBERT，并评估 NER 性能。

TABLE II
分词策略的比较（设计选择）

Tokenization	Efficiency	OOV Rate	Morphological Preservation	Fine-Tuned?
BPE	High	Moderate	Moderate	Yes
SentencePiece	Moderate	Low	High	Yes
Character-Level	Low	None (0 %)	Very High	No

对于外部评估，我们使用 Naamapadam 数据集来评估印地语、孟加拉语、阿萨姆语、马拉地语和奥里亚语，而对信德语、桑塔利语和曼尼普尔语，我们则使用每种语言 200 个句子的手动标注数据集。所有这三种语言的标注者间一致性 F1 得分均高于 72 %，表明标注质量良好。表 III 总结了数据集的详细信息。

TABLE III
命名实体识别数据集统计

Language	Train Size	Validation Size	Test Size
Assamese (Naamapadam)	10,266	52	51
Bengali (Naamapadam)	961,679	4,859	607
Hindi (Naamapadam)	985,787	13,460	867
Marathi (Naamapadam)	455,248	2,300	1,080
Oriya (Naamapadam)	196,793	993	994
Hand-Annotated			
Sindhi	–	–	200
Santali	–	–	200
Manipuri	–	–	200

对于 NER 微调，我们在 IndicBERT 中比较两种分词策略：

字节对编码（BPE）：IndicBERT 中的默认分词方法，BPE 通过合并常见字符序列构建一个紧凑的词汇表。然而，由于对罕见词的过度切分，它可能在零样本迁移中表现出较差的泛化能力。

SentencePiece：与 BPE 不同，SentencePiece 可以在字符级别操作，从而更好地处理未见过的词汇和形态复杂的结构。这种灵活性可能会提高命名实体识别和零样本性能。

我们使用两种分词策略微调 IndicBERT，这是一种在多种印度脚本上训练的变压器模型。

训练配置：

- 学习率: 2e-5
- 批量大小：16
- 轮数：3，权重衰减：0.01
- 优化器：AdamW
- 评估指标：F1-得分（宏观和实体级别）

单独的 IndicBERT 模型针对 BPE 和 SentencePiece 分词进行了微调，之后在各自的测试集和零样本语言上进行评估。

为了评估分词的影响，我们使用：

- 词级准确率：衡量正确分类词元的百分比。
- 实体级 F1 分数：用于评估命名实体识别的精确度、召回率和 F1 分数。
- 零样本迁移性能：比较模型应用于未见语言时的性能：阿萨姆语、奥里亚语、马拉地语、信德语、桑塔利语和曼尼普尔语。

D. 实现细节

所有模型都使用 Hugging Face Transformers 库实现。微调和评估是在具有 16GB 内存的 NVIDIA A10 GPU 上进行的。模型检查点和分词数据集被存储以便于重现。

III. 结果与讨论

A. 内在评估：*FLORES-200* 上的分词分析

内在评价使用 FLORES-200 对四种低资源的印度语言：曼尼卑尔语、桑塔利语、阿萨姆语和信德语，不同的分词策略——BPE、SentencePiece 和字符级别——进行评估。此次评估侧重于分词效率、词汇表外（OOV）率、词汇压缩和形态学保留，以确定最适合下游 NER 任务的方法。

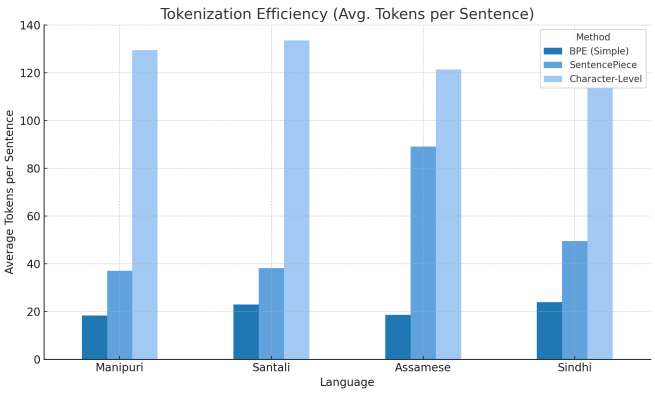

Fig. 1. 分词效率（平均每句标记数）

图 1 显示了作为每句话平均标记数的标记化效率。BPE 给出了最少的标记，提供计算效率，但可能在形态丰富的语言上分段不足。SentencePiece 生成更多标记，通过有意义的子词在效率和语言结构上达到平衡。字符级标记化导致最高的标记数，造成了分散的表示和增加的计算负担。

表 IV 比较了在处理 OOV 和词汇压缩方面的分词策略。由于激进的子词合并，BPE 显示 0 % OOV，提供了空间效率但对未见的语言形式适应性有限。SentencePiece 的 OOV 率在 4.34 % –7.81 % 之间，提供了更灵活的分割和更好的泛化。字符级分词由于过度的碎片化，产生了最高的 OOV（40.70 % –50.32 %）。如图 2 所示，词汇压缩反映了这些趋势：BPE 保持紧凑的 1.0 比例，SentencePiece 提供适度的压缩（4.33–7.80），而字符级分词显著扩大词汇（高达 73.35），使其在计算上效率低下。这些结果证明了在外部评估中排除字符级方法的合理性，转而主要关注

TABLE IV
未登录词率（%）和词汇压缩比率

Method	OOV Rate (%)			
	Manipuri	Santali	Assamese	Sindhi
BPE (Simple)	0.0	0.0	0.0	0.0
SentencePiece	-341.59	-333.65	-609.60	-680.12
Character-Level	-7235.11	-6710.92	-5091.77	-4042.75
Vocabulary Compression Ratio				
BPE (Simple)	1.0	1.0	1.0	1.0
SentencePiece	4.41	4.33	7.09	7.80
Character-Level	73.35	68.10	51.91	41.42

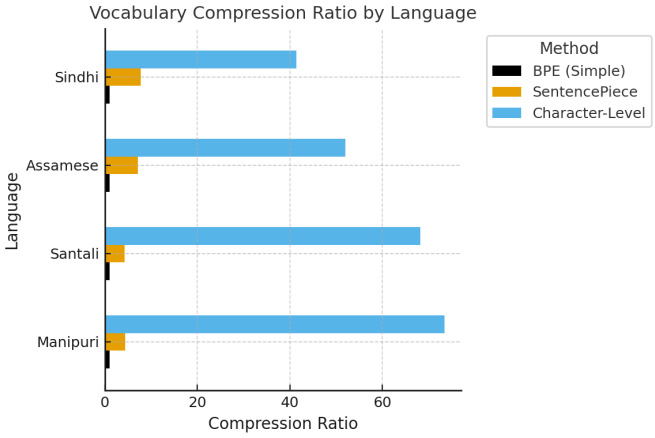

Fig. 2. 词汇压缩率

BPE 的紧凑性和 SentencePiece 在低资源 NER 任务中的适应性。

TABLE V
形态保留分析

Method	Findings
BPE (Simple)	Can't preserve morphemes in agglutinative languages
SentencePiece	Can preserve morphemes in low-resource languages
Character-Level	Over-fragments words, causing semantic loss

表 V 展示了每种分词方法如何保持语言的形态学结构。BPE 在处理黏着语时表现较差，无法保持结构。具有灵活子词单位的 SentencePiece 在跨多种文字脚本时更好地保持了形态学结构 [31] 。字符级分词能够捕捉结构但会过度分解词语，削弱语义一致性。

这些内在结果强烈支持在下游命名实体识别中使用 SentencePiece 而不是 BPE。尽管 BPE 的效率很高，但其对形态复杂性的处理较差，限制了零样本转移。SentencePiece 提供了效率与语言保真度之间最佳的权衡。字符级方法虽然全面，但计算成本过高，在实际的自然语言处理应用中不切实际。

B. 用于外部评估的过滤分词策略

内在评价的结果为选择用于实体识别模型外在评价的分词策略提供了指导。由于字符级分词的效率低下，它被淘汰。因此，只有 BPE 和 SentencePiece 被选为进行微调和零样本跨语言评价。

C. 外部评估：命名实体识别微调和零样本跨语言迁移

IndicBERT 在印地语和孟加拉语 NER 数据集上使用 BPE 和 SentencePiece 分词器进行了微调。然后模型在未见过的语言：阿萨姆语、马拉地语、奥里亚语、信德语、桑塔利语和曼尼普尔语上进行了测试。

1) 在印地语 & 孟加拉语上进行微调并在印地语 & 孟加拉语上进行测试 ： 为了评估分词策略对命名实体识别（NER）性能的影响，我们在印地语和孟加拉语数据集（Naamapadam）上使用 BPE 和 SentencePiece 分词对 IndicBERT 进行了微调，并在各自的测试集上评估了模型。总体结果如表 VI 所示。

TABLE VI
INDICBERT 在相同语言上微调和测试的 NER 性能

Language	Tokenizer	Precision	Recall	F1-Score
Hindi	BPE	0.9572	0.9571	0.9569
Hindi	SentencePiece	0.9544	0.9543	0.9543
Bengali	BPE	0.9485	0.9499	0.9488
Bengali	SentencePiece	0.9493	0.9501	0.9495

对图 **??** 中各类别度量指标的更深入分析表明，SentencePiece 在多个实体类别中的召回率始终较高。在印地语中，B-ORG 的召回率从 80.99 % (BPE) 提高到 80.86 % (SentencePiece)，而 B-LOC 从 83.03 % (BPE) 提高到 84.50 % (SentencePiece)。类似地，在孟加拉语中，I-PER 的召回率从 89.73 % (BPE) 提高到 92.92 % (SentencePiece)，这强化了 SentencePiece 的灵活子词分割增强了模型在不同语言结构中泛化实体识别能力的观念。这与我们的内在分析一致，后者展示了 SentencePiece 比 BPE 更好地保留了形态结构，使其能够更有效地适应未见过的子词。

另一方面，BPE 始终保持稍高的准确率，特别是在资源丰富的环境中，分词的歧义较少。在印地语中，B-PER 准确率为 87.33 %（BPE），相比于 SentencePiece 的 85.92 %。在孟加拉语中，B-LOC 准确率仍然较高，使用 BPE 时为 84.61 %，而 SentencePiece 则为 84.72 %。这些结果表明，虽然 BPE 确保了较保守的分段，SentencePiece 在捕获子词结构方面的灵活性有助于在细粒度实体分段中提高召回率。

总体而言，这两种方法在语言内微调中表现良好，F1 评分差异较小。然而，它们的精准率与召回率之间的权衡揭示了各自的优势——BPE 适合以精准率为重点的任务，尤其是有频繁出现的标记，而 SentencePiece 更适合需要灵活分割、以召回率为驱动的场景。这些区别在零样本跨语言转移中更加显著，强调了在不同的印度语言中，平衡分割精度与广泛适应能力的重要性。

2) 零样本命名实体识别在未知语言中的评估 ： 为了评估不同分词策略的泛化能力，我们在密切相关但未见过的语言上测试了孟加拉语和印地语微调模型。具体来说，我们在阿萨姆语、奥里亚语、圣塔利语和曼尼普尔语上测试了孟加拉语微调模型，而在马拉地语和信德语上测试了印地语微调模型。句子片段和 BPE 分词器的结果汇总在表 VII 中。

结果显示这两种分词策略之间存在显著差异：

- SentencePiece 展现出强大的泛化能力：以 88.38 %（阿萨姆语），81.08 %（奥里亚语），81.09 %（马拉地语）和 51.98 %（曼尼普尔语）的 F1 分数，模型

TABLE VII
零样本 NER 评估：使用 BPE 和 SentencePiece 分词法在未见语言上进行孟加拉语和印地语微调模型

Target Language	Source Model	F1-score	Accuracy
Assamese (as)	Bengali SP	88.38 %	87.55 %
	Bengali BPE	0.00 %	91.22 %
Oriya (or)	Bengali SP	81.08 %	86.94 %
	Bengali BPE	0.00 %	87.02 %
Santhali (sat)	Bengali SP	46.12 %	78.54 %
	Bengali BPE	12.67 %	69.12 %
Manipuri (mni)	Bengali SP	51.98 %	80.23 %
	Bengali BPE	9.34 %	65.78 %
Marathi (mr)	Hindi SP	81.09 %	83.02 %
	Hindi BPE	67.79 %	75.49 %
Sindhi (sd)	Hindi SP	33.28 %	45.04 %
	Hindi BPE	20.69 %	25.04 %

保留了实体识别能力，特别是在语言相似度高的语言中。然而，在资源较少的语言如桑塔利语中，性能有所下降，其中 SentencePiece 的 F1 分数为 46.12 %，而 BPE 的 F1 分数仅为 12.67 %。

- BPE 在某些情况下完全失败：模型在阿萨姆语和奥里亚语中仅预测 “O”（非实体）类，导致 F1 分数为 0.00 % 。在马拉提语中，BPE 保持了一些性能（67.79 % ），但显著不如 SentencePiece。在桑塔利语和曼尼普尔语中，BPE 难以泛化，仅获得 12.67 % 和 9.34 % 的 F1 分数。
- 对于像信德语这样极度低资源的语言，两个分词器的性能都表现不佳：由于训练数据有限，两个模型在信德语的实体识别上都存在困难，尽管 SentencePiece 比 BPE 取得了稍高的 F1 分数（33.28 % 对 20.69 %）。这表明虽然 SentencePiece 提供了更好的子词分割，但极度低资源的环境需要额外的适应技术。

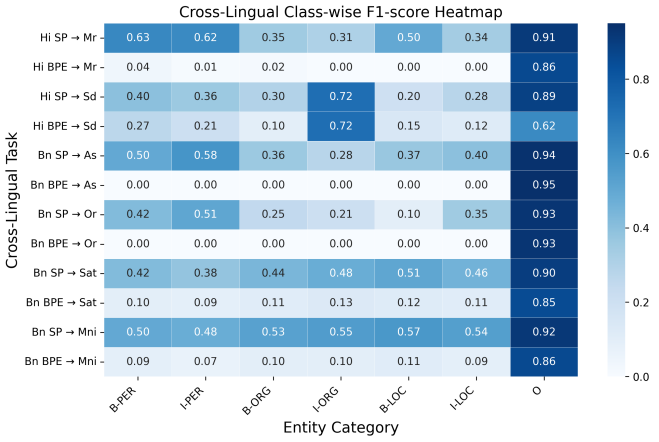

Fig. 3. SentencePiece 和 BPE 之间的零样本跨语言表现比较。

这些观察进一步加强了以下观点：SentencePiece 的子词分割有助于保留跨语言的实体结构，而 BPE 的激进合并策略在零样本场景中导致严重退化。这些发现提供了证据，表明子词标记化策略显著影响了命名实体识别（NER）模型的跨语言适应性。SentencePiece 始终能够在相关语言（如阿萨姆语、奥里亚语、马拉地语和曼尼普尔语）中保留实体区分，而 BPE 则难以有效推广。对于诸如信德语和桑塔利语这样资源极其稀缺的语言，可能需要额外的迁移学习技术来提高模型的稳健性。

我们的论文强调了在低资源的印度语言中进行零样本跨语言命名实体识别时，分词的重要性。尽管仍存在差距，但 SentencePiece 在泛化能力上优于 BPE。将字符级和子词级表示结合的混合方法可能会进一步增强实体的分割和转移。扩展到如博多语、多格拉语、孔卡尼语和克什米尔语等语言，可能会加深对不同语言环境下分词的理解。

IV. 结论

这项研究通过内在和外在实验，系统地研究了分词方法在资源匮乏的印度语言命名实体识别中的作用，对比了字节对编码（BPE）、字符级别和 SentencePiece 方法。内在研究表明，SentencePiece 更好地保留了形态结构，降低了超出词汇表的比例，并允许比 BPE 更动态的分段。在微调和零样本跨语言设置中，SentencePiece 的表现始终优于 BPE，尤其是在诸如阿萨姆语和奥里亚语以及马拉地语和信德语等密切相关的语言中。结果表明，BPE 中严格的子词合并会导致零样本迁移中的实体分段错误，而 SentencePiece 能够更好地保留实体边界。

我们的研究结果提供了实证证据，表明标记化策略显著影响低资源 NLP 中的跨语言泛化，表明 SentencePiece 是零样本 NER 任务的更有效选择。尽管 BPE 仍被广泛使用，但其在保持实体结构方面的局限性表明，需要更具适应性和语言学信息的标记化方法。未来的工作应集中在开发对标记化敏感的多语言预训练，完善混合标记化方法，扩展评估到其他低资源语言，以进一步优化在不同语言环境中的 NER 性能。

References

[1] I. Office of the Registrar General & Census Commissioner, “Census of india 2011: Data on language and mother tongue,” https://censusindia.gov.in/2011census/C-16/DDW-C16-STMT-MDDS-0000.XLSX, accessed: 2025-03-09.

[2] C. of India, “Eighth schedule to the constitution of india,” https://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India, accessed: 2025-03-09.

[3] P. Pattnayak, A. Agarwal, B. Kumar, Y. Bangera, S. Panda, T. Kumar, and H. L. Patel, “Review of reference generation methods in large language models,” *Journal ID*, vol. 9339, p. 1263.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, p. 4171–4186, 2019.

[5] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, and P. Kumar, “Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages,” 2023. [Online]. Available: https://arxiv.org/abs/2212.05409

[6] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 1715–1725, 2016.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 8440–8451, 2020.

[8] A. Agarwal, H. Patel, P. Pattnayak, S. Panda, B. Kumar, and T. Kumar, “Enhancing document ai data generation through graph-based synthetic layouts,” *arXiv preprint arXiv:2412.03590*, 2024.

[9] A. Agarwal, S. Panda, and K. Pachauri, "Fs-dag: Few shot domain adapting graph networks for visually rich document understanding," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*. Abu Dhabi, UAE: Association for Computational Linguistics, January 2025, pp. 100–114. [Online]. Available: https://aclanthology.org/2025.coling-industry.9/

[10] N. Yin, M. Wan, L. Shen, H. L. Patel, B. Li, B. Gu, and H. Xiong, "Continuous spiking graph neural networks," *arXiv preprint arXiv:2404.01897*, 2024.

[11] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 66–71, 2018.

[12] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 66, p. 673–717, 2021.

[13] P. Pattnayak, H. L. Patel, B. Kumar, A. Agarwal, I. Banerjee, S. Panda, and T. Kumar, "Survey of large multimodal model datasets, application categories and taxonomy," *arXiv preprint arXiv:2412.17759*, 2024.

[14] G. Ndomba, M. Mswahili, and Y. Jeong, "Tokenizers for african languages," *IEEE Access*, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10815724/

[15] A. Agarwal, S. Panda, A. Charles, B. Kumar, H. Patel, P. Pattnayak, T. H. Rafi, T. Kumar, and D.-K. Chae, "Mvtamperbench: Evaluating robustness of vision-language models," *arXiv preprint arXiv:2412.19794*, 2024.

[16] N. Hussain, A. Qasim, G. Mehak, and O. Kolesnikova, "Hybrid machine learning and deep learning approaches for insult detection in roman urdu text," *AI*, 2025. [Online]. Available: https://www.mdpi.com/2673-2688/6/2/33

[17] M. Bayram, A. Fincan, A. Gümüş, and S. Karakaş, "Tokenization standards for linguistic integrity: Turkish as a benchmark," *arXiv preprint arXiv:2502.07057*, 2025. [Online]. Available: https://arxiv.org/pdf/2502.07057

[18] O. Olaleye, H. L. Patel, and T. Sheng, "Pseudo-labelling based bootstrapping for semi supervised learning," Feb. 2025, uS Patent App. 18/237,234.

[19] M. Shahid, M. Iqbal, and M. Umair, "Leveraging cumeta for enhanced document classification in cursive languages with transformer stacking," *Multimedia Tools and Applications*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s11042-025-20681-w

[20] N. Kambhatla, "Augmented input representations in sequence generation models for decipherment and translation," *SFU Summit*, 2024. [Online]. Available: https://summit.sfu.ca/_flysystem/fedora/2025-02/etd23279.pdf

[21] H. L. Patel, A. Agarwal, B. Kumar, K. Gupta, and P. Pattnayak, "Llm for barcodes: Generating diverse synthetic data for identity documents," *arXiv preprint arXiv:2411.14962*, 2024.

[22] D. Kakwani, A. Varma, A. Kunchukuttan, M. M. Khapra, P. Kumar, and K. Shashi, "Indicnlp corpus: Monolingual corpora and word embeddings for indic languages," *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, p. 1173–1182, 2020.

[23] M. A. Hedderich, D. Klakow, G. Glavaš, O. Rohanian, J. Risch, and A. Bharadwaj, "A survey on recent approaches for natural language processing in low-resource scenarios," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, p. 2545–2568, 2021.

[24] A. Kumar, P. Mehta, and P. Bhattacharyya, "Named entity recognition for indian languages," *Proceedings of the 2022 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. 376–387, 2022.

[25] V. Dewangan, G. Suri, and R. Sonavane, "When every token counts: Optimal segmentation for low-resource language models," in *LoResLM Workshop*, 2025. [Online]. Available: https://aclanthology.org/2025.loreslm-1.24/

[26] A. Abdullah, S. Abdulla, and D. Toufiq, "Ner-roberta: Fine-tuning roberta for named entity recognition (ner) within low-resource languages," *arXiv preprint arXiv:2412.15252*, 2024. [Online]. Available: https://arxiv.org/abs/2412.15252

[27] Y. Wang, X. Jin, Y. Sun *et al.*, "Adaptive subword tokenization for low-resource nlp: Balancing efficiency and generalization," in *ACL 2022*, 2022.

[28] G. Suri, V. Dewangan, and R. Sonavane, "When every token counts: Optimal segmentation for low-resource language models," *arXiv preprint arXiv:2412.06926*, 2024. [Online]. Available: https://arxiv.org/pdf/2412.06926

[29] M. R. Costa-Juss'a, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[30] A. Mhaske, H. Kedia, S. Doddapaneni, M. M. Khapra, P. Kumar, R. M. V, and A. Kunchukuttan, "Naamapadam: A large-scale named entity annotated data for indic languages," 2023. [Online]. Available: https://arxiv.org/abs/2212.10168

[31] S. Kumar and R. Vavekanand, "Multiclass text classifications of sindhi newspaper articles," *Preprints*, 2025. [Online]. Available: https://www.preprints.org/frontend/manuscript/d40099f1eed56b67c6f65d138e209557/download_pub