# CHARSPAN：利用词汇相似性为极低资源语言实现零样本机器翻译

**Kaushal Kumar Maurya**[1], [3]*和**Rahul Kejriwal**[2]**Maunendra SankarDesarkar**[1] 和**Anoop Kunchukuttan**[2]印度海得拉巴印度理工学院 NLIP 实验室[2]Microsoft，印度 [3]MBZUAI，阿联酋
cs18resch11003@iith.ac.in，maunendra@cse.iith.ac.in{rahul.kejriwal, anoop.kunchukuttan}@microsoft.com

arXiv:2305.05214v2 [cs.CL] 4 Feb 2024

## 摘要

我们通过利用与密切相关的高资源语言（HRL）的跨语言迁移，解决从极低资源语言（ELRL）到英语的机器翻译（MT）任务。为极低资源语言开发机器翻译系统颇具挑战，因为这些语言通常缺乏平行语料库和单语语料库，且大型多语语言模型中没有它们的表示。许多极低资源语言与某些高资源语言存在词汇相似性，这提供了新的建模机会。然而，现有的基于子词的神经机器翻译模型并未明确利用这种词汇相似性，因为它们只是隐式地对齐高资源语言与极低资源语言的潜在嵌入空间。为克服这一局限，我们提出一种新颖的 CHARSPAN 方法，将字符跨度噪声增强应用于高资源语言的训练数据。这是一种正则化技术，可增强模型对高资源语言与极低资源语言之间词汇差异的鲁棒性，从而促进有效的跨语言迁移。在零样本设置下，对于来自三个不同语系的密切相关高资源语言与极低资源语言语对，我们的方法显著优于强大的基线方法，成为极低资源语言机器翻译的最先进模型。

## 1 引言

多语言建模的最新进展，通过将知识从高资源语言迁移到低资源语言，扩大了自然语言处理技术对许多低资源语言的覆盖范围。因此，这些进展推动了多项自然语言处理任务取得显著进步，包括机器翻译、音译、自然语言理解和文本生成（Johnson 等人，2017年；Kunchukuttan 等人，2018年；Conneau 等人，2020年；Liu 等人，2020年），以支持低资源语言。然而，现有研究大多聚焦于排名靠前的几百种语言

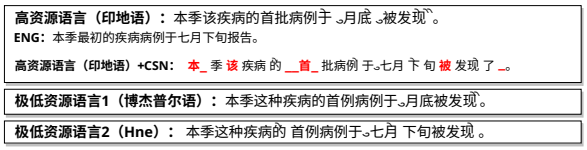


图 1：印地语（HIN；高资源语言）、博杰普尔语（BHO；极低资源语言）和恰蒂斯加尔希语（HNE；极低资源语言）的平行句子。此外，还展示了对应的带有字符跨度噪声的印地语噪声示例。博杰普尔语和恰蒂斯加尔希语与印地语密切相关。

在网络上有所体现（Joshi 等人，2020年）。这些语言的单语语料库和/或平行语料库的可用性，推动了这一进展的实现，其途径包括直接训练、少样本训练，或利用大型多语语言模型（多语大语言模型）进行学习。这使得学习共同的嵌入空间成为可能，从而促进跨语言迁移（Nguyen 和 Chiang，2017 年；Khemchandani 等人，2021年）。然而，仍有大量语言缺乏单语语料库或平行语料库，且未被纳入多语大语言模型。这些语言被称为极低资源语言（ELRLs）。本文旨在为极低资源语言构建机器翻译系统迈出一步。

幸运的是，许多极低资源语言在词汇方面与某些高资源语言具有相似性。词汇相似性是指不同语言共享形式（拼写和发音）及含义相似的词语。[1] 这包括同源词、横向借词和借词。我们探究能否通过明确地考虑词汇相似性，来实现或改善极低资源语言的跨语言迁移。具体而言，我们研究从一种极低资源语言到另一种语言（英语）的机器翻译，其中源语言端通过一种相关的高资源语言实现迁移。我们的关键洞见是，极低资源语言中与高资源语言词语拼写相似的同源词，可以被视为后者的拼写错误。例如，印地语（高资源语言）中的单词 लगता（*lagta*）在博杰普尔语（低资源语言）中的拼写为 लागवता（*laagata*）。如果我们使高资源语言模型能够适应拼写变体，就能改善向相关极低资源语言的跨语言迁移。为了实现拼写变体的处理，

*这项工作是在第一作者于 Microsoft 实习期间完成的。当时他是海得拉巴印度理工学院的研究生。

[1]https://en.wikipedia.org/wiki/ Lexical_相似度

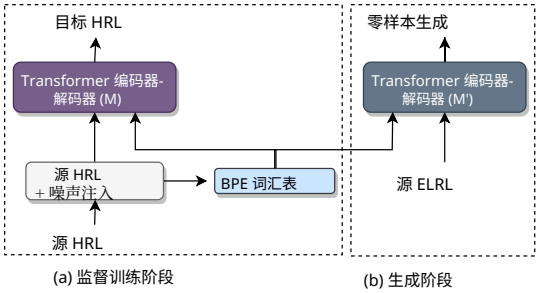


图 2：所提出的 CHARSPAN 模型概览

鲁棒性，我们提出一种新颖的字符跨度噪声增强（CHARSPAN），并将其应用于高资源语言的训练数据中。图 1展示了一个示例。这种方法充当正则化器，使模型能够更好地应对相关语言中词语表示的扰动，并提升模型对词汇相似语言的泛化能力。

我们的主要贡献如下：(1) 我们提出了一种新颖的模型 CHARSPAN：字符跨度噪声增强，该模型考虑表层词汇相似性，以改善亲缘关系密切的高资源语言与低资源语言之间的跨语言迁移。与基线神经机器翻译模型相比，该方法在所有纳入研究的极低资源语言上均实现了 12.5% 的 chrF 提升。我们的模型相较于各种数据增强基线方法也展现出更优的性能。(2) 我们证明，该方法能够推广至三个类型学多样的语系，涵盖 6 种高资源语言和 12 种极低资源语言。(3) 我们进行了详细的消融实验与分析，以深入了解并验证该方法的有效性。

## 2 相关工作

传统上，字符级噪声被用于提升机器翻译系统对拼写错误和自动语音识别错误的鲁棒性（Sperber 等人，2017年；Vaibhav 等人，2019；Karpukhin 等人，2019）。不过，这些方法主要研究的是它们对鲁棒性的影响，而非跨语言迁移。近来，有研究提出了词元/BPE 级通用噪声增强方法，例如词丢弃（Sennrich 等人，2016年（a））和 SwitchOut（Wang 等人，2018年），但这些方法的跨语言迁移能力有限。与我们的工作较为接近的是，Aepli 和 Sennrich (2022年) 和 Blaschke 等人(2023年) 表明，使用字符级噪声增强数据有助于跨语言迁移。他们使用自然语言理解任务对模型进行了评估。相比之下，我们的工作聚焦于机器翻译这一自然语言生成任务；在零样本设置下，它比自然语言理解任务更具挑战性。此外，我们还探究了 SpanNoise 增强方法。该方法考虑了更大的词汇差异（即高资源语言与极低资源语言之间的词汇相似性更低），从而能够实现更好的跨语言迁移。

在其他利用词汇相似性的研究中，Patil 等人（2022年）提出了 OverlapBPE，在学习 BPE 词汇表时考虑高资源语言与低资源语言之间的词汇重叠。Provilkov 等人（2020年）提出了 BPE-Dropout，在训练过程中即时提供非确定性分词。软解耦编码（SDE）Wang 等人（2019年）通过解耦词汇表示与语义表示，在不预先对数据进行分词的情况下利用词汇信息。SDE 需要少量单语数据进行建模。相比之下，CHARSPAN 模型不需要任何极低资源语言的训练资源，只需要高资源语言与极低资源语言之间具有文字系统相似性。

## 3 CHARSPAN模型

图2展示了所提出的 CHARSPAN 模型概览，该模型用于极低资源语言到英语的机器翻译任务。该模型包含两个阶段：使用带噪声的高资源语言进行监督训练，以及使用极低资源语言进行零样本生成。

模型训练与生成：在监督训练阶段，对高资源语言语言对（$\mathcal{D}_{\mathcal{H}}$）的源端训练数据添加字符跨度噪声（稍后介绍），以创建增强平行语料库（$\mathcal{D}'_{\mathcal{H}} = \eta(\mathcal{D}_{\mathcal{H}})$），其中 $\eta$ 是噪声函数。$\eta(\mathcal{D}_{\mathcal{H}})$可视为极低资源语言-英语翻译任务的代理平行语料。接下来，我们使用 $\mathcal{D}'_{\mathcal{H}}$学习子词词汇表（$\mathcal{V}$），也就是说，在学习词汇表之前先添加噪声。随后，在监督设置下，从头开始使用 $\mathcal{D}'_{\mathcal{H}}$ 和$\mathcal{V}$ 训练标准的编码器-解码器 Transformer 模型（$\mathcal{M}$；Vaswani 等人（2017）），从而获得训练好的模型 $\mathcal{M}'$。最后，在零样本生成阶段，对于给定的源极低资源语言$\mathcal{L}$，使用$\mathcal{M}'$ 和 $\mathcal{V}$ 在零样本设置下获得目标英语译文。

**字符跨度噪声函数：**噪声函数旨在使模型能够适应相关语言之间的拼写变体。这起到正则化器的作用，有助于改进跨语言表示和迁移。直观而言，现有的一元字符噪声或许可以应对高资源语言与极低资源语言之间有限的词汇变体。为了解决更大的

词汇差异，我们提出了 CHARSPAN，在其中增强了跨度噪声。形式上，对于给定句子，$x \in \mathcal{X}$来自 $\mathcal{D}_{\mathcal{H}}(\mathcal{X}, \mathcal{Y})$，其索引为 $I = 1, 2, \ldots, |x|$，我们会随机、均匀地选取这些索引的一个子集$I_s \subset I$，作为噪声增强的起始位置。随后，迭代采样 1–3 个字符的跨度，直到耗尽噪声增强预算（即字符总数的 9%–11%）。我们采用跨度删除和跨度替换操作，将跨度替换为一个随机的极低资源语言字符，两种加噪操作的概率相同[2]。这种 CHARSPAN 的设计受到 SpanBERT（Joshi 等人，2020年）的启发[3]。算法 1 给出了形式化算法。我们对全部三种操作（包括插入）进行了实验，并测试了不同的噪声比例及其他多种实验设置，详见附录表13。我们发现，所展示的噪声增强配置效果最佳。

## 4 实验设置

我们试图回答以下问题：（1）跨度噪声增强能否改善跨语言迁移，即提升机器翻译任务中相关极低资源语言的零样本性能？（2）模型的跨语言迁移为何有所改善？——从学习到的嵌入空间中寻找见解。（3）所提出的方法能否扩展到类型学多样的语系？

### 4.1 数据集与语言

我们在三个语系上评估了所提出模型的性能：印度-雅利安语支、罗曼语族和马来-波利尼西亚语族。我们纳入了六种高资源语言和十二种低资源语言（每个语系包含两种高资源语言和若干种极低资源语言）。如图所示，所有极低资源语言在词汇上都与对应的高资源语言相似，并使用相同的文字系统4（附录D）。高资源语言的平行训练数据选自公开数据集。模型的性能在 FLORES-200 devtest 集上进行评估（Costa-jussà 等人，2022年）。数据集统计信息见附录。

### 4.2 基线方法与评估指标

基于近期低资源机器翻译领域的研究，我们将本方法与以下强有力的基线方法：（a）采用字节对编码分词的原始神经机器翻译（字节对编码；Sennrich 等人(2016年（b））），（b）通用数据增强方法：(Sub)WordDropout 和 (Sub)WordSwitchout，（c）利用词汇相似性的方法：重叠 BPE、BPE-Dropout、SDE 和 unigram 字符加噪（Aepli 和 Sennrich，2022年）。基线方法和模型训练详情见附录。遵循近期关于极低资源语言机器翻译的研究（Costa-jussà 等人，2022年；Siddhant 等人，2022年），我们采用 chrF（Popovi´c，2015）作为主要评估指标。此外，我们还在附录Papineni 等人，2002）中报告 BLEU，以及两种神经指标，即 BLEURT（Sellam 等人，2020年）和 COMET（Rei 等人，2020年）得分，详见附录C。

## 5 结果与分析

所提出的 CHARSPAN 方法与基线模型在不同语系中的结果见表1。主要观察结果如下：

**噪声增强与基线方法的比较：**所有提出的噪声增强模型均优于原始神经机器翻译以及所有利用词汇相似性的基线模型（即 OBPE、BPE-Dropout 和 SDE）。这一趋势在所有语系和极低资源语言中都一致。此外，现有基于词汇相似性的基线方法并未显著提升原始神经机器翻译的翻译质量。可能有两个原因：（1）大多数极低资源语言要么没有单语数据（OBPE 和 SDE 均需要单语数据），要么数据量较少；（2）我们发现，在 OBPE 中，高资源语言与极低资源语言之间约有 90% 的词汇词元已经重叠，因此几乎没有空间再学习额外的重叠词元。这是意料之中的，因为这两组语言关系密切。所提出的 CHARSPAN 方法也优于诸如 (Sub)WordDropout 和 (Sub)WordSwitchout 等通用数据增强方法，体现了其有效性。

**一元噪声与字符跨度噪声：**我们首次探索了将一元字符噪声（Aepli 和 Sennrich，2022年）用于相关语言机器翻译。我们发现，一元字符噪声对这项任务有益。然而，我们提出的 CHARSPAN 相较于一元字符噪声取得了显著提升。我们认为，我们提出的数据增强更能有效地拉近语言表示之间的距离。

[2]我们探索了一些基于语言学原理的加噪方案，但这些方案并无益处。[3]SpanBERT 对子词词元进行去噪，而我们则在字符层面进行去噪。

**算法 1** CHARSPAN：字符跨度噪声增强算法

**输入：** **[输入项]** 来自 $H$-$En$ 平行语料库的高资源语言数据（$\mathcal{D}_{\mathcal{H}}(\mathcal{X},\ \mathcal{Y})$），噪声增强范围百分比 $[P1, P2]$、噪声增强候选项集合 $C$（见图 3），加噪时考虑的最大字符 $n$-gram 大小 $N$将被考虑用于加噪

**确保：** **[输出]**加噪后的高资源语言数据（$\mathcal{D}_{\mathcal{H}}^{'}$）

增强百分比（$I_p$）= 随机浮点数(P1, P2) # 找出 $P1$ 和 $P2$ 之间的随机浮点值
增强因子 ($\alpha$) =int($I_p/N$)
**遍历** $h$ 中的 $\mathcal{X}$ **执行**
    设 $sz$ 为 $h$中的字符数。
    设 $Indices = \{\lceil(N/2)\rceil,\ \cdots,\ sz - \lceil(N/2)\rceil\}$ # 从开头和结尾各留出 $\lceil(N/2)\rceil$ 个字符索引
    从 $Indices$中随机选取 $S = N * \alpha$ 个字符索引
    **遍历** $k$ 中的 $S$ **执行**
        跨度语法（$Sp_N$）= 从 $\{1, 2, \ldots, N\}$ 中以等概率均匀采样字符跨度大小
        操作（$O_p$）= 从 { 删除、替换 } 中以等概率均匀采样操作
        $C_d$ ={}
        **如果**（$O_p$）是替换**则**
            候选字符（$c$）= 从 $C$ 中以等概率均匀采样单个字符
            将候选字符追加 $c$ 到 $C_d$中
        **结束条件判断**
        **如果** $Sp_N == 1$，**则**
            使用 $C_d$ 在索引 $k$ 处执行操作（$O_p$）
        **else**
            使用 $C_d$ 在从 $k - int$（$(Sp_N - 1)/2$）到 $k + int$（$(Sp_N - 1)/2$）的索引范围内执行操作（$O_p$）
        **结束条件判断**
    **结束循环**
**结束**

| 模型 | 印度-雅利安语支 | | | | | | | | 罗曼语族 | | 马来-波利尼西亚语族 | | 平均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gom | Bho | Hne | San | Npi | Mai | Mag | Awa | Cat | Glg | Jav | Sun | |
| BPE* | 26.75 | 39.75 | 46.57 | 27.97 | 30.84 | 39.79 | 48.08 | 46.28 | 33.32 | 53.75 | 31.44 | 32.21 | 38.06 |
| 词丢弃 | 27.01 | 39.57 | 46.19 | 28.13 | 31.91 | 40.31 | 47.37 | 46.48 | 34.20 | 52.21 | 32.03 | 32.52 | 38.16 |
| 子词丢弃 | 27.91 | 40.11 | 46.26 | 29.46 | 32.56 | 40.99 | 47.91 | 47.43 | 35.09 | 52.28 | 33.38 | 33.47 | 38.90 |
| WordSwitchOut | 25.17 | 38.81 | 45.87 | 26.21 | 29.95 | 39.69 | 47.53 | 44.54 | 32.98 | 51.81 | 31.84 | 32.49 | 37.24 |
| SubwordSwitchOut | 26.08 | 38.84 | 45.84 | 28.19 | 30.81 | 40.19 | 47.28 | 45.93 | 33.26 | 53.71 | 31.24 | 32.06 | 37.78 |
| OBPE | 27.90 | 40.57 | 47.46 | 28.52 | 31.99 | 40.71 | 49.10 | 47.16 | 32.33 | 52.77 | 29.98 | 30.88 | 38.28 |
| SDE | 28.01 | 40.91 | 47.88 | 28.66 | 32.03 | 40.82 | 48.96 | 47.30 | 33.72 | 53.95 | 31.84 | 31.24 | 38.77 |
| BPE-Dropout* | 28.65 | 40.84 | 46.58 | 28.80 | 31.88 | 40.79 | 47.86 | 47.32 | 34.56 | 55.83 | 32.01 | 32.97 | 39.00 |
| 一元字符噪声** | 28.85 | 42.53 | 49.35 | 29.80 | 34.61 | 42.67 | 50.97 | 49.43 | 43.16 | 54.81 | 35.42 | 36.69 | 41.52 |
| BPE → SpanNoise*** (*ours*) | 28.66 | 41.94 | 49.48 | 30.49 | 35.66 | 44.75 | 50.55 | 49.21 | 43.11 | 54.89 | 36.12 | 37.11 | 40.16 |
| CHARSPAN (*ours*) | 29.71 | 43.75 | 51.69 | **31.40** | 36.52 | 45.84 | 51.90 | 50.55 | 43.51 | 55.46 | 36.24 | 37.31 | 42.82 |
| CHARSPAN + BPE-Dropout (*ours*) | **29.91** | **44.02** | **51.86** | 30.88 | **37.15** | **46.52** | **52.99** | **51.34** | **44.93** | **55.87** | **36.97** | **38.09** | **43.37** |

表1：极低资源语言 → 英语机器翻译的零样本 chrF 分数结果。我们使用配对自助法抽样（Post，2018年）进行了统计显著性检验，以比较 CHARSPAN 与多种基线方法：BPE、BPE-Dropout、一元字符噪声和 BPE → SpanNoise。CHARSPAN 相较于这些基线方法的改进具有统计显著性，显著性水平分别为 *（$p < 0.0001$）、**（$p < 0.001$）和***（$p < 0.05$）。在附录中展示的其他评估指标上也观察到了类似结果。

**何时引入噪声？** 为了了解噪声增强何时有效，我们在基线方法中学习完词汇表后再加入噪声（字节对编码 → SpanNoise）。与所有基线方法相比，这种做法提升了性能。由于在学习词汇表后加入噪声，这种方法能够应用于具有固定词汇表的大语言模型，因此具有可扩展性。不过，结果表明，像 CHARSPAN那样，在学习词汇表之前加入噪声，效果略好。此外，我们使用配对自助法抽样进行了统计显著性检验，比较字节对编码 → SpanNoise 与字节对编码、BPE-Dropout 以及一元字符噪声基线方法（Post，2018年）。我们观察到，字节对编码 → SpanNoise 模型优于基线 BPE 和 BPE-Dropout 方法（在 $p < 0.001$处具有统计显著性），这表明在分词后添加噪声同样非常有效。此外，我们注意到，在 $p < 0.05$ 条件下，BPE → SpanNoise 在 12 种语言中的 11 种上优于一元字符噪声。因此，在分词后引入字符跨度噪声能够带来相较于基线方法具有统计显著性的改进，这在使用预训练模型时可能很有优势。

**结合噪声与 BPE-dropout：** 我们发现，将 CHARSPAN 与 BPE-dropout 结合可获得最佳效果。

**较不相似语言上的表现：** 我们

| 语言 | BPE | 一元噪声 | 字符跨度噪声 | Sim |
|---|---|---|---|---|
| 古吉拉特语-天城文 | 34.36 | 36.17 | 38.09 | 0.42 |
| 旁遮普语-天城文 | 29.18 | 33.34 | 36.50 | 0.40 |
| 孟加拉语-天城文 | 25.35 | 28.42 | 30.28 | 0.34 |
| 泰卢固语-天城文 | 23.30 | 24.05 | 24.12 | 0.27 |
| 泰米尔语-天城文 | 13.81 | 13.69 | 14.40 | 0.15 |

表 2：加入词汇相似性较低的语言后的零样本 chrF 分数。HRL：印地语和马拉地语；相似度：词汇相似性

评估模型在与所考察语言词汇相似性较低且采用不同文字系统的语言上的表现。这些语言包括古吉拉特语（古吉拉特语）、旁遮普语（旁遮普语）、孟加拉语（孟加拉语）、泰卢固语（泰卢固语）和泰米尔语（泰米尔语）。我们首先按照 Kunchukuttan （2020年）将这些语言进行文字转换，使其转换为高资源语言。训练设置与印度-雅利安语族类似。表2 显示，与高资源语言词汇相似的极低资源语言表现出更大的性能提升，而词汇相似性较低的语言提升有限。这表明，模型在 CHARSPAN 中的有效性与语言间的词汇相似性密切相关。

**跨语言迁移的影响：**在本分析中，我们研究句子的编码表示，以深入了解字符跨度噪声增强如何提升性能。我们收集了采用字节对编码、一元字符噪声（UCN）和*CharSpan* 模型的高资源语言和低资源语言在所有平行测试样例上的编码器池化后的最后一层表示。随后，我们计算了整个测试集的平均余弦相似度分数，结果见表3。值得注意的是，*CharSpan*模型表现出较高的相似度，表明其嵌入空间对齐良好，有助于增强跨语言迁移。

| 模型 | Bho | Hne | San | Npi | Mai | Mag | Awa |
|---|---|---|---|---|---|---|---|
| BPE | 0.761 | 0.793 | 0.701 | 0.744 | 0.762 | 0.809 | 0.792 |
| UCN | 0.853 | 0.888 | 0.765 | 0.821 | 0.849 | 0.897 | 0.883 |
| CHARSPAN | **0.871** | **0.909** | **0.789** | **0.858** | **0.868** | **0.913** | **0.901** |

表 3：印度-雅利安语族源高资源语言与源极低资源语言表示之间的平均余弦相似度。其他语族的结果见附录F。

**选择合适的高资源语言的重要性：**表4 分析了用于训练的词汇多样性不同的高资源语言所产生的影响。结果表明，在对词汇相似的高资源语言进行噪声注入时，CHARSPAN 模型的性能有所提升。而当高资源语言之间的词汇相似度较低时，模型性能则会下降。这些发现表明使用词汇相似的高资源语言的重要性 Ls.

| 模型 | Hne | Mag | Mai | Npi | San |
|---|---|---|---|---|---|
| 使用词汇相似的高资源语言进行训练：印地语、马拉地语、*Pan*、古吉拉特语、孟加拉语 | | | | | |
| BPE | 43.04 | 45.08 | 39.51 | 31.92 | 29.29 |
| 字符跨度噪声 | 45.89 | 45.82 | 41.67 | 34.40 | 30.34 |
| 使用词汇相似度较低的高资源语言进行训练：印地语、*Tel*、*Tam*、*Mal*、*Ora* | | | | | |
| 字节对编码 | 41.87 | 42.27 | 36.95 | 30.50 | 26.95 |
| 字符跨度噪声 | 39.93 | 40.34 | 37.98 | 29.20 | 25.84 |

表4：使用词汇多样化的高资源语言进行零样本 chrF 分数分析实验。受计算资源限制，我们为每种高资源语言使用了 100 万条平行语料。

**少量极低资源语言平行语料的影响：**在此，我们将少量极低资源语言平行语料与高资源语言训练数据结合起来，用于训练字节对编码和 CHARSPAN 模型。结果见附录14 E中的表格。额外数据提升了两个模型的性能，而 CHARSPAN 仍然优于字节对编码模型。

**错误分析：** 在附录 G中，我们进行了两项错误分析： 音译错误 和语法良构性。在图 7中可以观察到，对于许多输入词，单字模型常常进行音译，而非翻译。然而，所提出的模型没有出现此类错误，且音译错误的影响很小。这一观察结果在不同测试数据中均成立。这可能是因为 CHARSPAN 扩展了跨度，从而实现更强的正则化，并促进更具上下文性的零样本跨语言迁移。在表16中，我们比较了句子合规性，结果表明，与 CharSpan 相比，单字模型的零样本生成在语法上并不合规。

## 6 结论

本研究提出了一种简单而有效的新型字符跨度噪声增强模型 CHARSPAN，旨在促进从高资源语言向密切相关的极低资源语言进行更好的跨语言迁移。该方法适用于来自三个类型学多样语系的密切相关高资源语言—濒危低资源语言配对。所提出的模型始终优于所有基线方法。据我们所知，我们是首个将噪声增强应用于自然语言生成任务的团队。未来，我们将把 CHARSPAN 扩展到其他自然语言处理任务，将其与预训练模型结合，并研究噪声增强在英语到濒危低资源语言的机器翻译任务中的应用。

## 局限性

目前的工作仅涉及从极低资源语言翻译成英语时的跨语言迁移。噪声增强是否有助于从英语翻译成极低资源语言，仍有待研究。我们假设相关语言也使用相同的文字系统，或者使用彼此易于映射／转写的文字系统。对于以差异很大的文字系统书写的相关语言之间的迁移，这种方法可能并不有效。例如，印地语使用天城文书写，而信德语使用波斯-阿拉伯文字。

## 伦理声明

我们将低资源语言表述为高资源语言的一种拼写错误版本。我们想澄清的是，我们并非认为低资源语言是高资源相关语言的拼写错误版本。这并非一项语言学主张；正如比较语言学所表明的那样，这类情形大多可能是相关语言共同演化的结果。将相关语言视为彼此关联，只是一种技术工具，用来利用这样一个事实：相关语言共同演化的最终结果，是它们彼此“看起来像”对方的拼写变体。因此，应用于神经机器翻译的鲁棒性方法也可以调整后用于这种情形。

本研究未涉及任何新的数据收集，也未聘用任何标注人员进行数据收集。我们使用公开可用的数据集进行本研究所报告的实验。其中一些数据集来源于网络爬取；我们并未明确尝试识别这些数据集中的任何偏差，而是直接使用原始数据。

## 参考文献


Noëmi Aepli 和 Rico Sennrich。2022年。通过注入字符级噪声，改进密切相关语言之间的零样本跨语言迁移。载于 计算语言学协会发现：*ACL 2022*，第 4074–4083 页，爱尔兰都柏林。计算语言学协会。

Mikel Artetxe、GorkaLabaka、Eneko Agirre 和 Kyunghyun Cho。2018年。无监督神经机器翻译。载于第六届国际学习表征会议论文集。

Mikel Artetxe、Sebastian Ruder、Dani Yogatama、Gorka Labaka 和 Eneko Agirre。2020年。呼吁更多无监督跨语言学习的严谨性。在第*58*届计算语言学协会年会论文集中，第7375–7388页，线上。计算语言学协会。

Verena Blaschke、Hinrich Schütze 和 Barbara Plank。2023年。操纵分词有助于跨语言迁移吗？一项关于非标准化语言词性标注的研究。载于 第十届相似语言、变体与方言自然语言处理研讨会（*VarDial 2023*），第 40–54 页，克罗地亚杜布罗夫尼克。计算语言学协会。

JunyoungChung、Kyunghyun Cho 和 YoshuaBengio。2016。用于神经机器翻译的无需显式分段的字符级解码器。载于第*54*届计算语言学协会年会论文集（第*1*卷：长篇论文），第1693–1703页，德国柏林。计算语言学协会。

Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。2020年。大规模无监督跨语言表征学习。载于第*58*届计算语言学协会年会论文集，第8440–8451页，线上。计算语言学协会。

Marta RCosta-jussà、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Maillard等。2022年。不让任何语言掉队：扩展以人为中心的机器翻译。*arXiv*预印本 *arXiv:2207.04672*。

Marta R. Costa-jussà、Carlos Escolano 和 José A. R. Fonollosa。2017年。基于字节的神经机器翻译。载于自然语言处理中的子词和字符级模型首届研讨会论文集，第 154–158 页，丹麦哥本哈根。计算语言学协会。

Angela Fan、Shruti Bhosale、Holger Schwenk、Zhiyi Ma、AhmedEl-Kishky、Siddharth Goyal、Mandeep Baines、Onur Celebi、Guillaume Wenzek、Vishrav Chaudhary等。2021年。超越以英语为中心的多语机器翻译。机器学习研究期刊，22(107):1–48。

Philip Gage。1994年。一种新的数据压缩算法。*C*语言用户杂志，12(2):23–38。

Google-2018。2022年。开源 BERT 中的 Wordpiece 算法。见 *https://github.com/google-research/bert/blob/master/ tokenization.py#L335-L358*。检索日期：*2023*年*11*月*1*日。

Rohit Gupta、Laurent Besacier、Marc Dymetman 和 Matthias Gallé。2019年。采用 Transformer 的基于字符的神经机器翻译。*CoRR,abs/1911.04997*。

Georg Heigold、Stalin Varanasi、Günter Neumann 和 Josef van Genabith。2018年。基于字符的词嵌入在标注和机器翻译中面对词语扰乱或随机噪声时有多稳健？载于第十三届美洲机器翻译协会会议（第*1*卷：研究专题），第68–80页，波士顿，马萨诸塞州。美洲机器翻译协会。

Melvin Johnson、Mike Schuster、Quoc V. Le、Maxim Krikun、Yonghui Wu、Zhifeng Chen、Nikhil Thorat、Fernanda Viégas、Martin Wattenberg、Greg Corrado、Macduff Hughes 和 Jeffrey Dean。2017年。Google 的多语言神经机器翻译系统：实现零样本翻译。计算语言学协会会刊，5:339–351。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S Weld、Luke Zettlemoyer 和 Omer Levy。2020年。SpanBERT：通过表示和预测跨度改进预训练。计算语言学协会会刊，8：64–77。

Pratik Joshi、Sebastin Santy、Amar Budhiraja、Kalika Bali 和 Monojit Choudhury。2020年。NLP 世界中语言多样性与包容性的现状与前景。载于第*58*届计算语言学协会年会论文集，第 6282–6293 页，线上。计算语言学协会。

Vladimir Karpukhin、Omer Levy、Jacob Eisenstein 和 Marjan Ghazvininejad。2019年。通过合成噪声训练可提升机器翻译对自然噪声的鲁棒性。载于第五届噪声用户生成文本研讨会论文集（*W-NUT 2019*），第42–47页，中国香港。计算语言学协会。

Yash Khemchandani、Sarvesh Mehtani、Vaidehi Patil、Abhijeet Awasthi、Partha Talukdar 和 Sunita Sarawagi。2021年。利用语言相关性进行低网络资源语言模型适配：一项印度语言研究。载于第*59*届计算语言学协会年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第1312–1323页，线上。计算语言学协会。

Philipp Koehn 和 Rebecca Knowles。2017年。神经机器翻译面临的六大挑战。载于首届神经机器翻译研讨会论文集，第28–39页，温哥华。计算语言学协会。

Taku Kudo 和 John Richardson。2018年。SentencePiece：一种用于神经文本处理的简单、语言无关的子词分词器和去分词器。载于*2018*年自然语言处理实证方法会议论文集：系统演示，第 66–71 页，比利时布鲁塞尔。计算语言学协会。

Anoop Kunchukuttan。2020年。IndicNLP 库。https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf。

Anoop Kunchukuttan、MiteshKhapra、Gurneet Singh 和 Pushpak Bhattacharyya。2018年。利用正字法相似性进行多语言神经音译。计算语言学协会会刊，6：303–316。

Guillaume Lample、Alexis Conneau、Ludovic Denoyer 和 Marc'Aurelio Ranzato。2018年。仅使用单语语料库进行无监督机器翻译。第六届国际学习表征会议论文集。

Jason Lee、Kyunghyun Cho 和 Thomas Hofmann。2017年。无需显式分段的完全字符级神经机器翻译。计算语言学协会会刊，5：365–378。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于第*58*届计算语言学协会年会论文集，第7871–7880页，线上。计算语言学协会。

Jindˇrich Libovický 和 Alexander Fraser。2020年。通过微调子词系统，实现规模适中的字符级 Transformer 神经机器翻译。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 2572–2579 页，线上。计算语言学协会。

Jindˇrich Libovický、Helmut Schmid 和 Alexander Fraser。2022年。为什么人们不使用字符级机器翻译？载于计算语言学协会发现：*ACL 2022*，第 2470–2485 页，都柏林，爱尔兰。计算语言学协会。

Yinhan Liu、Jiatao Gu、Naman Goyal、Xian Li、Sergey Edunov、Marjan Ghazvininejad、Mike Lewis 和 Luke Zettlemoyer。2020年。用于神经机器翻译的多语言去噪预训练。计算语言学协会会刊，8:726–742。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Kelly Marchisio、Kevin Duh 和 Philipp Koehn。2020年。无监督机器翻译何时有效？载于第五届机器翻译会议论文集

翻译，第571–583页，线上。计算语言学协会。

I. Dan Melamed. 1995年。用于诱导 n 最佳翻译词典的自动评估与统一过滤级联。载于第三届超大规模语料库研讨会。

Benjamin Minixhofer、Fabian Paischer 和 Navid Rekabsaz。2022年。WECHSEL：为单语语言模型的跨语言迁移有效初始化子词嵌入。载于计算语言学协会北美分会 *2022*年会议论文集：人类语言技术，第 3992–4006 页，美国西雅图。计算语言学协会。

Toan Q. Nguyen 和 David Chiang。2017年。面向神经机器翻译的低资源相关语言间迁移学习。载于第八届自然语言处理国际联合会议论文集（第 *2* 卷：短文），第 296–301 页，中国台湾台北。亚洲自然语言处理联合会。

Myle Ott、Sergey Edunov、Alexei Baevski、Angela Fan、Sam Gross、Nathan Ng、David Grangier 和 Michael Auli。2019年。fairseq：一款快速、可扩展的序列建模工具包。载于*NAACL-HLT 2019* 论文集：演示。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002 年。BLEU：一种用于机器翻译自动评估的方法。载于计算语言学协会第 *40* 届年会论文集，第 311–318 页，美国宾夕法尼亚州费城。计算语言学协会。

Vaidehi Patil、Partha Talukdar 和 Sunita Sarawagi。2022年。基于重叠的词汇生成方法能够改善相关语言之间的跨语言迁移。载于《第*60*届计算语言学协会年会论文集》（第*1*卷：长篇论文），第219–233页，爱尔兰都柏林。计算语言学协会。

Maja Popovi´c. 2015. chrF：用于自动机器翻译评估的字符n元语法 F 分数。载于第十届统计机器翻译研讨会论文集，第 392–395 页，葡萄牙里斯本。计算语言学协会。

Matt Post。2018年。呼吁清晰地报告 BLEU 分数。载于第三届机器翻译会议论文集：研究论文，第186–191页，比利时布鲁塞尔。计算语言学协会。

Ivan Provilkov、Dmitrii Emelianenko 和 Elena Voita。2020年。BPE-Dropout：简单且有效的子词正则化。载于第*58*届计算语言学协会年会论文集，第1882–1892页，线上。计算语言学协会。

Gowtham Ramesh、Sumanth Doddapaneni、Aravinth Bheemaraj、Mayank Jobanputra、Raghavan AK、Ajitesh Sharma、Sujit Sahoo、Harshita Diddee、Mahalakshmi J、Divyanshu Kakwani、Navneet Kumar、Aswin Pradeep、Srihari Nagaraj、Kumar Deepak、Vivek Raghavan、Anoop Kunchukuttan、Pratyush Kumar 和 Mitesh Shantadevi Khapra。2022年。Samanantar：规模最大的公开可用平行语料库集合，涵盖11种印度语言。计算语言学协会会刊，10：145–162。

Reinhard Rapp。2021年。使用 Marian神经机器翻译对加泰罗尼亚语、葡萄牙语和西班牙语进行相似语言翻译。载于第六届机器翻译会议论文集，第292–298页，线上。

Ricardo Rei、Craig Stewart、Ana C Farinha 和 Alon Lavie。2020年。COMET：用于机器翻译评估的神经框架。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第2685–2702页，线上。计算语言学协会。

Thibault Sellam、Dipanjan Das 和 Ankur Parikh。2020年。BLEURT：学习用于文本生成的稳健指标。载于计算语言学协会第*58*届年会论文集，第7881–7892页，线上。计算语言学协会。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年（a）。用于 WMT 16 的爱丁堡神经机器翻译系统。载于《第一届机器翻译会议论文集：第 *2* 卷，共享任务论文》，第 371–376 页，德国柏林。计算语言学协会。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年（b）。使用子词单元进行稀有词的神经机器翻译。载于《第*54*届计算语言学协会年会论文集（第*1*卷：长篇论文）》，第 1715–1725 页，德国柏林。计算语言学协会。

Uri Shaham 和 Omer Levy。2021年。无需嵌入的神经机器翻译。载于《*2021*年北美计算语言学协会分会会议论文集：人类语言技术》，第 181–186 页，线上。计算语言学协会。

Aditya Siddhant、Ankur Bapna、Orhan Firat、Yuan Cao、Mia Xu Chen、Isaac Caswell 和 Xavier Garcia。2022年。迈向多语机器翻译领域的下一个 1000 种语言：探索监督学习与自监督学习之间的协同作用。*arXiv* 预印本 *arXiv:2201.03110*。

Matthias Sperber、Jan Niehues 和 Alex Waibel。2017年。面向噪声输入序列的稳健神经机器翻译。载于第*14*届口语翻译国际会议论文集，第90–96页。

Chau Tran、Shruti Bhosale、James Cross、Philipp Koehn、Sergey Edunov 和 Angela Fan。2021年。Facebook AI 提交的 WMT21 新闻翻译任务论文。载于第六届机器翻译会议论文集，第205–215页，线上。计算语言学协会。

Vaibhav Vaibhav、Sumeet Singh、Craig Stewart 和 Graham Neubig。2019年。利用合成噪声提升机器翻译的鲁棒性。载于*2019*年计算语言学协会北美分会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），第1916–1920页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力机制就是你所需的一切。神经信息处理系统进展，第30卷。

Xinyi Wang、Hieu Pham、Philip Arthur 和 Graham Neubig。2019年。采用软解耦编码的多语神经机器翻译。载于第七届国际学习表征会议，*ICLR 2019*，新奥尔良，路易斯安那州，美国，*2019*年*5*月*6*日至*9*日。OpenReview.net。

Xinyi Wang、Hieu Pham、Zihang Dai 和 Graham Neubig。2018年。SwitchOut：一种用于神经机器翻译的高效数据增强算法。收录于*2018*年自然语言处理实证方法会议论文集，第856–861 页，比利时布鲁塞尔。计算语言学协会。

Barret Zoph、Deniz Yuret、Jonathan May 和 Kevin Knight。2016。低资源神经机器翻译的迁移学习。收录于*2016*年自然语言处理实证方法会议论文集，第 1568–1575 页，得克萨斯州奥斯汀。计算语言学协会。

## A 基线方法

我们将所提出模型的性能与以下强基线方法进行比较：

**•原始神经机器翻译（字节对编码；Sennrich 等人(2016年（b））：**使用标准 BPE 算法训练神经机器翻译模型。

**• 词丢弃（Sennrich 等人，2016年（a））：**在此基线方法中，随机选取源句或目标句中的词，并将其嵌入设为 0。由于噪声增强是在源句中进行的，我们选取了源句中 10% 的词。

**•子词丢弃：** 这是词丢弃基线的一种变体，其中丢弃的是 BPE 子词单元，而不是单词。

**•WordSwitchOut（Wang 等人，2018年）：**该基线采用了一种数据增强技术，即将源句和目标句中的随机单词替换为分别从其词汇表中随机选取的单词。我们采用了官方发布的实现，并将单词替换率设为 10%。

**•SubwordSwitchOut：** 这是 WordSwitchOut 基线的一种变体，我们使用字节对编码词元而不是单词。

**• 重叠 BPE（OBPE；Patil 等人（2022年））：**该方法通过修改 BPE 算法，鼓励高资源语言和低资源语言的词汇表中包含更多共享词元。该模型需要极低资源语言的单语数据集。根据数据可用情况，我们为极低资源语言使用了一个小型单语数据集。早期研究仅将 OBPE 用于自然语言理解任务，而我们是首个探索其在机器翻译中应用的研究。

**•软解耦编码（SDE；（Wang 等人，2019年））：**在 SDE 方法中，作者设计了一种框架，将词嵌入有效地解耦为两个相互作用的组成部分：表示单词的拼写，以及捕捉单词的潜在语义。这种建模技术已证明能够有效提升低资源语言的性能。在本研究中，我们采用了官方发布的 SDE 实现。

**•BPE-Dropout（Provilkov 等人，2020年）：**它利用 BPE 算法学习词汇表，并在训练过程中（即时）为输入文本采样不同的切分。

**•一元字符噪声（UCN；Aepli 和 Sennrich（2022年））：**受 UCN 模型启发，我们添加字符级噪声（包括全部三种操作），而不是字符跨度；其余设置与 CHARSPAN 类似。

**•字节对编码 → 字符跨度噪声：** 在这项消融实验中，我们首先使用干净的高资源语言学习词汇表。随后，将字符跨度噪声加入训练数据中。这将

展示了使用含噪数据集学习字节对编码词汇表的重要性。

- **字符跨度噪声 + BPE-Dropout：** 在该模型中，我们使用字符跨度噪声增强的高资源语言训练数据集来训练 BPE-Dropout 模型。

## B 模型训练细节

我们使用 FairSeq 库（Ott 等人，2019年）训练了所提出的 CharSpan 以及其他基线模型。训练和实现细节见表 6。我们根据验证损失选择最佳检查点。印度-雅利安语族语言的训练时间约为 8 小时；罗曼语族语言约为 7 小时；马来-波利尼西亚语族语言则不到 1 小时。每种语言的推理均在 5 分钟内完成。受计算资源限制，模型性能报告基于单次运行。在生成过程中，批量大小设为 64，束宽设为 5，其余参数均采用 FairSeq 提供的默认值。对于印度语言的数据预处理和文字系统转换，我们使用 Indic NLP 库[6]。

## C 使用 BLEU、BLEURT 和 COMET 指标进行性能评估

BLEU[7], BLEURT 和 COMET 分数分别列于表 7、8 和 9。我们观察到与正文中报告的 chrF[8]相同的趋势。

## D 语言相似度直方图

如图所示。4，图中以热力图形式展示了所选语系和语言的相似性分析。分析表明，极低资源语言（ELRLs）与高资源语言（HRLs）关系密切。语言之间的词汇相似性采用字符级最长公共子序列率（LCSR）指标进行测量（Melamed，1995）。较不相似语言的类似热力图也展示于图中5。这些语言用于多项分析。

## E 额外少量极低资源语言平行语料的影响

在这里，我们将少量极低资源语言的平行语料与高资源语言的训练数据相结合，用于训练 BPE 和 CHARSPAN 模型。结果如表 14 所示。额外数据的加入提升了两种模型的性能，而 CHARSPAN 仍然优于 BPE 模型。

## F 跨语言迁移的影响

我们进行了以下研究，以了解噪声为何有帮助。跨语言迁移的效果取决于 HRL 和 ELRL 的表示对齐程度。我们的假设是，使用字符级噪声进行正则化，可以使 *HRL* 和 *ELRL* 的表示彼此更加接近，从而改善跨语言迁移。为衡量这些情况，我们计算了 3 种不同模型（基线 BPE、一元字符噪声、CHARSPAN）中平行 HRL 和 ELRL 句子的编码器表示的余弦相似度。编码器表示通过对编码器顶层的词元表示进行均值池化计算得出。表格 -15展示了结果（我们报告了测试集上的平均结果）。我们可以清楚地看到，经过噪声增强的模型，其编码器表示的相似度显著提高。此外，CHARSPAN 的表现优于一元字符噪声，反映出翻译质量有所提升。

## G 错误分析

### G.1 基线生成结果经过了转写

图 7展示了几个基线模型生成错误的示例。在这里，我们关注音译错误。可以观察到，在基线模型的目标生成中，许多源词会被直接音译；然而，所提出的 CHARSPAN 模型成功缓解了这些错误。

### G.2 语法良构性

人们经常会观察到，生成结果在语法上并不规范，而性能评估指标很容易忽略这些特征

[6]https://github.com/anoopkunchukuttan/ 印度尼西亚语_自然语言处理_库[7]使用 SacreBLEU 计算的 BLEU 签名：参考译文数量：*1*|大小写：混合|有效性：否|分词：*13a*|平滑：指数|版本：*2.3.1*[8]使用 SacreBLEU 计算的 chrF 签名：参考译文数量：*1*|大小写：混合|有效性：是|字符数：*6*|词数：*0*|空格：否|版本：*2.3.1*

| 语系 | Code | 语言 | 文字系统 | 语系 | 分支 | Res. | 训练 | Dev | Test | 数据来源 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Hin | 印地语 | 天城文 | 印欧语系 | 印度-雅利安语支 | High | 10M | 1000 | 2390 | Ramesh 等人(2022年) |
| | Mar | 马拉地语 | 天城文 | 印欧语系 | 印度-雅利安语支 | High | 3.6M | 1000 | 2390 | Ramesh 等人（2022年） |
| | Bho | 博杰普尔语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| | Gom | 孔卡尼语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 2000 | ILCI[4] |
| | Hne | 恰蒂斯加尔希语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| | San | 梵语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| | Npi | 尼泊尔语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| | Mai | 迈蒂利语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| | Mag | 马加希语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| | Awa | 阿瓦德语 | 天城文 | 印欧语系 | 印度-雅利安语支 | Low | - | - | 1012 | FLORES-200 |
| 2 | Spa | 西班牙语 | 拉丁语 | 印欧语系 | 罗曼语族 | High | 6.6M | 670 | 1131 | Rapp （2021年） |
| | Pot | 葡萄牙语 | 拉丁语 | 印欧语系 | 罗曼语族 | High | 4.8M | 681 | 1103 | Rapp （2021年） |
| | Cat | 加泰罗尼亚语 | 拉丁语 | 印欧语系 | 罗曼语族 | Low | - | - | 1012 | FLORES-200 |
| | Glg | 加利西亚语 | 拉丁语 | 印欧语系 | 罗曼语族 | Low | - | - | 1012 | FLORES-200 |
| 3 | Ind | 印度尼西亚语 | 拉丁语 | 南岛语系 | 马来-波利尼西亚语族 | High | 0.5M | 2500 | 3000 | OPUS[5] |
| | Zsm | 马来语 | 拉丁语 | 南岛语系 | 马来-波利尼西亚语族 | High | 0.3M | 1500 | 2000 | OPUS |
| | Jav | 爪哇语 | 拉丁语 | 南岛语系 | 马来-波利尼西亚语族 | Low | - | - | 1012 | FLORES-200 |
| | Sun | 巽他语 | 拉丁语 | 南岛语系 | 马来-波利尼西亚语族 | High | - | - | 1012 | FLORES-200 |
| 其他 | Pan | 旁遮普语 | 古尔穆基文 | 印欧语系 | 印度-雅利安语支 | Low | 1M* | 1000* | 1012 | FLORES-200 |
| | Guj | 古吉拉特语 | 古吉拉特语 | 印欧语系 | 印度-雅利安语支 | Low | 1M* | 1000* | 1012 | FLORES-200 |
| | Ben | 孟加拉语 | 孟加拉语 | 印欧语系 | 印度-雅利安语支 | High | 1M* | 1000* | 1012 | FLORES-200 |
| | Tam | 泰米尔语 | 泰米尔语 | 印欧语系 | 印度-雅利安语支 | Low | 1M* | 1000* | 1012 | FLORES-200 |
| | Tel | 泰卢固语 | 达罗毗荼语系 | 印欧语系 | 印度-雅利安语支 | Low | 1M* | 1000* | 1012 | FLORES-200 |
| | Mal | 马拉雅拉姆语 | 马拉雅拉姆语 | 印欧语系 | 印度-雅利安语支 | Low | 1M* | 1000* | 1012 | FLORES-200 |
| | Ora | 奥里亚语 | 奥里亚语 | 印欧语系 | 印度-雅利安语支 | Low | 1M* | 1000* | 1012 | FLORES-200 |

表5：数据集详情与统计数据。* 引自 Ramesh 等人（2022年 ）

| 语系 | 文字系统 | 候选字母表 |
|---|---|---|
| 印度-雅利安语支 | 天城文 | 'ँ', 'ः', 'ऩ', 'ऽ', 'ा', 'ॊ', 'ञ', 'ऐ', 'अ', 'ऑ', 'र', 'फ', 'ग', 'ह', 'इ', 'न', 'ॅ', 'स', 'ए', 'ओ', 'ल', 'थ', 'ई', 'ऊ', 'ो', 'ा', 'ठ', 'ॄ', 'म', 'ी', 'छ', 'ॉ', 'ि', 'क', 'ण', 'भ', 'ट', 'ॅ', 'ॏ', 'ळ', 'ऋ', 'ष', 'ड', 'ु', 'ठ', 'ऌ', 'श', 'ब', 'ल', 'ी', 'ऒ', 'त', 'झ', 'ख', 'ज', 'थ', 'उ', 'ॢ', 'ॅ', 'ॾ', 'ः', 'ד', 'ऐ', 'ॠ', 'ो', 'ओ', 'ा', 'द', 'ह', 'ो', 'घ', 'च', 'ढ', 'ॣ', 'ञ', 'य', 'औ', 'व', 'आ', 'ऍ' |
| 斜体和马来语-波利尼西亚语 | 拉丁语 | A, a, B, b, C, c, D, d, E, e, F, f, G, g, H, h, I, i, J, j, K, k, L, l, M, m, N, n, O, o, P, p, Q, q, R, r, S, s, T, t, U, u, V, v, W, w, X, x, Y, y, Z, z, ñ, ó, ã, à, ç, í, é, ñ |

图3：用于噪声增强的候选字母表。对于印度-雅利安语支语系，使用天城文字母表；而罗曼语族和马来-波利尼西亚语系则使用拉丁字母表。

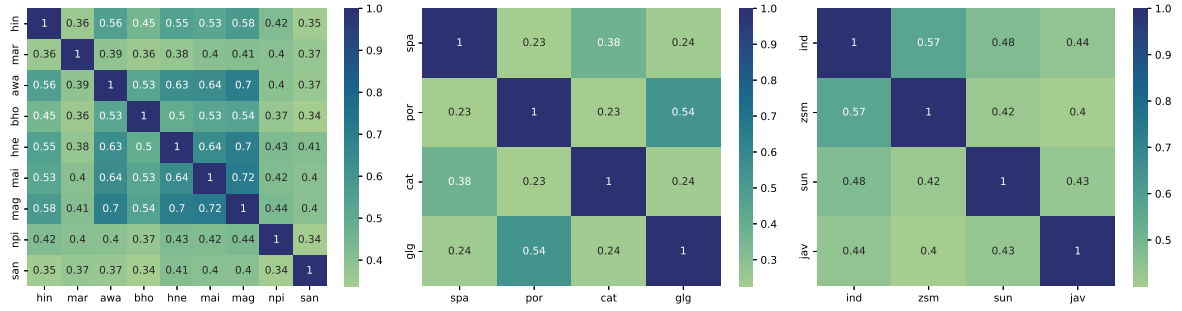


图4：三个语系的词汇相似性（最长公共子序列率）热图。印度-雅利安语被认为使用天城文，而其他两个语系使用拉丁字母。

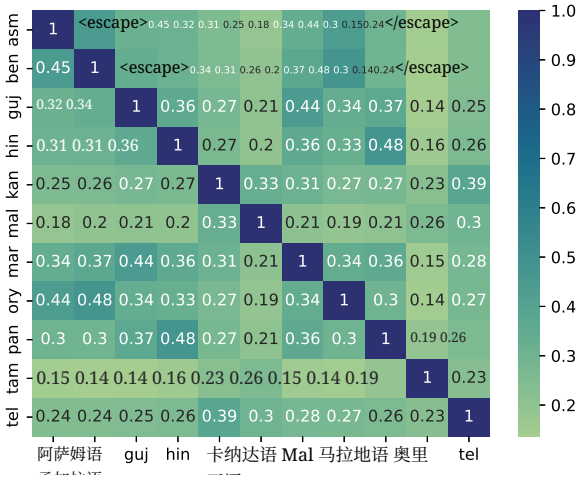


图 5：分析部分中使用的其他语言的词汇相似度热图。此处展示了阿萨姆语（asm）、孟加拉语（ben）、古吉拉特语（guj）、旁遮普语（pan）、印地语（him）、马拉地语（mar）、奥里亚语（ory）、马拉雅拉姆语（mal）、卡纳达语（kan）、泰米尔语（tam）和泰卢固语（tel）的相似度得分。

| | | |
|---|---|---|
| 博杰普尔语 → 英语 | 来源： 年份 2017年 的 首次 出现 西米诺夫，QVC 购物<br>在电视购物频道 上出现。 | 参考： 2017年末，Siminoff 出现在电视购物频道 QVC 上。<br>生成： 2017年年底，Siminauff 出现在 QVC 购物电视频道上。 |
| 孔卡尼语 → 英语 | 来源： 现在 这 建筑 一座 博物馆 中<br>改建了。 | 参考： 现在这座监狱已改建成博物馆。<br>生成： 现在，这座监狱已变成博物馆。 |
| 摩揭陀语 → 英语 | 来源： 罗蒂 萨阿 吉塔纳 乔特 霍泰, 恰ে巴基 乌马纳<br>के संबंध मे तारा उतना ही कम सकरयि होगा। | 参考： 罗斯贝数越小，恒星在磁反转方面的活动越不活跃。<br>通用： 罗斯贝数越小，恒星在磁演化方面的活动性就越低。 |
| 恰蒂斯加尔希语 → 英语 | 来源： 罗宾 乌塔 帕 里 以 最高 分 、打出 11 分<br>四 边和 2 个六 分，打出仅 41 个球，得分 70 分 | 参考： 罗宾·乌塔帕以仅 41 个球打出全场最高分 70 分，其中包括 11 个四分球和 2 个六分球。<br>生成： 罗宾·乌塔帕仅用 41 个球就取得最高分 70 分，其中包括 11 个边界球和 2 个六分球。 |
| 迈蒂利语 → 英语 | 来源： 电视报道 显示.白 色 烟雾 从 工厂 中。冒 出来了<br>正在显示。 | 参考： 电视报道显示，白色烟雾正从工厂冒出。<br>生成： 电视报道显示，烟雾从工厂冒了出来。 |
| 阿瓦德语 → 英语 | 来源： 该 系统 首 先 由西蒙 分 别 配 置<br>曾在多部节目中工作。 | 参考： 在《辛普森一家》之前，Simon 曾担任多个节目的不同职位。<br>生成： 在《辛普森一家》之前，Simon 曾担任多个节目的不同职位。 |
| 尼泊尔语 → 英语 | 来源： हिब्रु परिवारको अधिकाश जीवन ख。ला हावामा<br>बाे। | 参考： 希伯来家庭的大部分生活都是公开的。<br>创世记： 希伯来家庭的大部分生活都发生在公开场合。 |
| 梵语 → 英语 | 来源： ষু একমে ম 大金字塔<br>在吉萨 इति आचि थत्तम् अ。 | 参考： 吉萨的大金字塔是七大奇迹中唯一一座至今仍屹立不倒的建筑。<br>生成： 吉萨的大金字塔是七大奇迹之一，至今仍屹立不倒。 |
| 加泰罗尼亚语 → 英语 | 来源： 最初，服饰深受<br>东方拜占庭文化的影响。 | 参考： 最初，这种服饰深受东方拜占庭文化的影响。<br>生成： 位于吉萨的大金字塔是东方至今仍屹立的七大奇迹之一。 |
| 加利西亚语 → 英语 | 来源： 与此同时，德国海军，<br>主要依靠U型潜艇，设法.....<br>阻止这类流量。 | 参考： 与此同时，德国海军主要使用潜艇，试图阻止这条航线上的运输。<br>生成： 与此同时，德国海运业主要使用潜艇，试图阻止这一运输活动。 |
| 爪哇语 → 英语 | 来源： 虚拟团队成员通常是联络点<br>用于其直接物理群体。 | 参考： 虚拟团队成员通常是其所在实体小组的联络人。<br>生成： 虚拟团队成员是其所在实体小组成员的直接联络人。 |
| 巽他语 → 英语 | 来源： 中东的美国人正面临<br>与欧洲人或<br>阿拉伯<br>语 | 参考： 身处中东的美国公民可能会面临与欧洲人或阿拉伯人不同的情况。<br>世代： 中东的美国人面临一种情况，或从欧洲公民或阿拉伯人那里受益。 |

图 6：使用 CHARSPAN 模型为极低资源语言生成的零样本示例。

| 示例句子类型 | | 源语/目标语/生成 |
|---|---|---|
| 博杰普尔语转 ENG | 源输入 | 他回答道："我们有一只4个月大的猫咪，它之前感染过中耳疾病，但现在它已经从这种疾病中康复了。 |
| | 参考目标：他补充说："我们现在有4个月大的非糖尿病小鼠，它们以前患有糖尿病。" | |
| | BPE | "我们有一只4个月大的Mous Ba Jawan Pahil，患有糖尿病，但现在已经摆脱了这种疾病，"他补充道。" |
| | UCN | "我们有一个4个月大的女儿，她起初患有糖尿病，但现在我们已经摆脱了这种疾病，"他补充道。 |
| | CharSpan | 我们有4个月大的非糖尿病小鼠，但现在已经摆脱了糖尿病，"他补充道。" |
| HNE 至 ENG | 源输入 | 我们同意 USOC 的声明，即解散可能并不符合我们的运动员和俱乐部的利益，也不符合他们所从事的体育项目的利益。切实的 变革 是 前进 的方式 ， 或许 能够 带来 更好的结果。 |
| | 参考目标 | 我们同意 USOC 的声明：我们的运动员和俱乐部，以及他们所从事的体育项目，或许可以得到更好的保障通过在我们的组织内部推动有意义的变革，而非取消认证。 |
| | BPE | 我们同意 USOC 的观点，解散不如让我们的运动员和俱乐部的利益以及他们的体育项目 Syed Hamro Bhitra 有意义的变革，连同 Ah Ramro Service Day Sakinch。 |
| | UCN | Hami同意 USOC 的看法，即解散 Bhanda Baru Hamra运动员俱乐部 Bahruko 感兴趣蒂尼赫鲁科游戏说赛义德·哈姆罗与我们携手，共同迎接阿兹希·拉姆罗服务日萨金奇。 |
| | CharSpan | 我们同意 USOC 的看法，即解散符合运动员和俱乐部及其体育项目的利益并推动友好且有意义的转型，在组织中庆祝这一转型，而不是取消认证。 |

图7：不同基线模型的生成错误（音译）。所提出的 CHARSPAN 模型成功缓解了这些错误。颜色表示生成过程中对应的音译。

| | |
|---|---|
| 架构 | 编码器-解码器（Transformer） |
| # 编码器层 | 6 |
| # 解码器层 | 6 |
| # 参数 | 46,956,544 共享 |
| 学习率（lr） | $5e^{-4}$ |
| 优化器 | adam |
| 丢弃率 | 0.2 |
| 输入大小 | 210 个词元（两侧） |
| 训练轮数 | 15 |
| 每批次词元数 | 32768 |
| 梯度裁剪范数 | 1.0 |
| 学习率调度器 | 逆平方根 |
| GPU 数量 | 8 |
| GPU 类型 | Nvidia V100 |
| 生成批次大小 | 64 |
| 束宽 | 5 |

表 6：模型实现与训练详情

例如 chrF 和 BLEU。通过这项错误分析，我们旨在考察不同基线模型生成结果的语法良构性。为评估语法良构性，我们使用 L'AMBRE 工具[9]。结果见表16。为简明起见，我们仅展示印度-雅利安语族的结果。所有极低资源语言中，*CharSpan* 的语法构成均优于字节对编码和一元字符噪声模型。

这些错误分析进一步证明，CHARSPAN 模型的性能提升是真实的。

[9]https://github.com/adithya7/lambre

## H 文献综述

在本节中，我们详细介绍了与所提出的工作相关的三条文献综述脉络。相关内容总结于主论文的第 2 节。

### H.1 低资源语言的机器翻译

由于缺乏低资源语言的大规模双语文本数据集，现有的大部分研究都集中在多语言机器翻译上。这种方法能够实现跨语言迁移（Nguyen and Chiang, 2017; Zoph et al., 2016），并使相关语言能够相互学习（Fan et al., 2021; Costa-jussà et al., 2022; Siddhant et al., 2022）。尽管这一研究方向已受到广泛关注，但与高资源语言相比，低资源语言的性能提升仍然有限（Tran et al., 2021），这仍是一个有待深入研究的领域。另一方面，研究人员也尝试直接利用单语数据集构建机器翻译模型（Artetxe et al., 2018; Lample et al., 2018; Lewis et al., 2020）。这些无监督方法展现出了一定的潜力，但仍然需要大量单语数据，而这些数据理想情况下应与高资源语言的数据领域相匹配（Marchisio et al., 2020）。然而，对于许多低资源语言而言，单语数据集并不可用（Artetxe et al., 2020）。与此不同，我们提出了一种无需任何双语文本或单语数据集的模型，并且该模型能够扩展至任意数量的低资源语言或方言。

| 模型 | 印度-雅利安语支 | | | | | | | | 罗曼语族 | | 马来-波利尼西亚语族 | | 平均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gom | Bho | Hne | San | Npi | Mai | Mag | Awa | Cat | Glg | Jav | Sun | |
| BPE | 4.36 | 10.62 | 15.76 | 3.43 | 4.36 | 9.36 | 16.7 | 15.6 | 5.23 | 22.99 | 5.74 | 6.02 | 10.01 |
| 词丢弃 | 4.62 | 11.21 | 15.71 | 4.11 | 5.47 | 9.96 | 16.76 | 16.31 | 6.19 | 22.26 | 5.90 | 6.02 | 10.37 |
| 子词丢弃 | 4.57 | 9.99 | 14.47 | 3.93 | 5.25 | 9.08 | 15.53 | 16.03 | 5.85 | 20.72 | 4.78 | 4.93 | 09.59 |
| WordSwitchOut | 4.03 | 10.75 | 15.86 | 3.56 | 4.92 | 9.91 | 16.85 | 15.54 | 5.27 | 21.97 | 5.95 | 6.35 | 10.08 |
| SubwordSwitchOut | 4.13 | 10.56 | 15.93 | 3.76 | 4.49 | 9.69 | 16.61 | 16.69 | 5.19 | 23.82 | 6.02 | 6.01 | 10.24 |
| OBPE | 4.65 | 10.62 | 16.31 | 3.63 | 4.95 | 9.18 | 16.88 | 15.69 | 5.03 | 22.91 | 5.33 | 5.81 | 10.08 |
| SDE | 4.77 | 10.69 | 16.21 | 3.66 | 5.42 | 9.86 | 16.80 | 16.03 | 5.47 | 23.51 | 5.88 | 6.39 | 10.39 |
| BPE-Dropout | 5.24 | 11.33 | 15.64 | 3.71 | 4.94 | 10.00 | 16.62 | 16.63 | 5.94 | 24.07 | 5.79 | 6.65 | 10.54 |
| 一元字符噪声 | 5.21 | 12.62 | 18.29 | 3.81 | 6.55 | 11.29 | 19.47 | 18.95 | 11.82 | 24.09 | 7.35 | 6.87 | 12.19 |
| 字节对编码 → SpanNoise (*ours*) | 5.39 | 13.06 | 19.00 | 4.48 | 7.01 | 13.17 | 20.30 | 19.69 | 11.91 | 24.27 | 7.51 | 7.30 | 12.75 |
| CharSpan (*ours*) | 5.77 | 13.01 | 19.52 | 4.63 | 7.13 | 13.43 | 20.81 | 20.36 | 12.21 | 24.72 | 7.52 | 7.32 | 13.03 |
| CHARSPAN + BPE-Dropout（*ours*） | **5.81** | **13.81** | **21.03** | **4.64** | **8.10** | **14.33** | **22.11** | **21.25** | **12.64** | **25.35** | **7.52** | **7.31** | **13.65** |

表 7：极低资源语言 → 英语机器翻译的零样本 BLEU 分数结果

| 模型 | 印度-雅利安语支 | | | | | | | | 罗曼语族 | | 马来-波利尼西亚语族 | | 平均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gom | Bho | Hne | San | Npi | Mai | Mag | Awa | Cat | Glg | Jav | Sun | |
| BPE | 0.461 | 0.494 | 0.522 | 0.414 | 0.461 | 0.494 | 0.537 | 0.549 | 0.357 | 0.495 | 0.403 | 0.401 | 0.474 |
| 词丢弃 | 0.467 | 0.502 | 0.527 | 0.419 | 0.465 | 0.497 | 0.542 | 0.565 | 0.344 | 0.496 | 0.392 | 0.391 | 0.475 |
| 子词丢弃 | 0.454 | 0.493 | 0.513 | 0.393 | 0.459 | 0.481 | 0.526 | 0.554 | 0.319 | 0.468 | 0.382 | 0.383 | 0.460 |
| WordSwitchOut | 0.456 | 0.501 | 0.528 | 0.395 | 0.445 | 0.497 | 0.552 | 0.551 | 0.309 | 0.477 | 0.381 | 0.381 | 0.464 |
| SubwordSwitchOut | 0.459 | 0.494 | 0.519 | 0.415 | 0.455 | 0.496 | 0.535 | 0.555 | 0.365 | 0.496 | 0.383 | 0.385 | 0.467 |
| OBPE | 0.466 | 0.496 | 0.518 | 0.419 | 0.459 | 0.491 | 0.537 | 0.551 | 0.431 | 0.428 | 0.396 | 0.381 | 0.464 |
| SDE | 0.486 | 0.499 | 0.515 | 0.511 | 0.496 | 0.542 | 0.543 | 0.553 | 0.440 | 0.481 | 0.406 | 0.405 | 0.489 |
| BPE-Dropout | 0.474 | 0.494 | 0.501 | 0.413 | 0.461 | 0.481 | 0.522 | 0.555 | 0.443 | 0.443 | 0.407 | 0.412 | 0.467 |
| 一元字符噪声 | 0.471 | 0.523 | 0.547 | 0.403 | 0.456 | 0.486 | 0.571 | 0.592 | 0.495 | 0.501 | 0.403 | 0.405 | 0.487 |
| 字节对编码 → SpanNoise (*ours*) | 0.469 | 0.528 | 0.553 | 0.400 | 0.459 | 0.491 | 0.579 | 0.595 | 0.499 | 0.511 | 0.405 | 0.413 | 0.491 |
| CharSpan (*ours*) | 0.471 | 0.541 | 0.571 | 0.403 | 0.471 | 0.534 | 0.593 | 0.616 | 0.502 | 0.555 | **0.419** | 0.422 | 0.508 |
| CHARSPAN + BPE-Dropout (*ours*) | **0.478** | **0.548** | **0.582** | **0.421** | **0.478** | **0.535** | **0.604** | **0.623** | **0.505** | **0.567** | **0.419** | **0.429** | **0.515** |

表 8：极低资源语言 → 英语的零样本 BLEURT（使用*BLEURT-20* 检查点计算）得分结果 h

| 模型 | 印度-雅利安语支 | | | | | | | | 罗曼语族 | | 马来-波利尼西亚语族 | | 平均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gom | Bho | Hne | San | Npi | Mai | Mag | Awa | Cat | Glg | Jav | Sun | |
| BPE | 0.536 | 0.632 | 0.671 | 0.511 | 0.525 | 0.593 | 0.694 | 0.716 | 0.494 | 0.714 | 0.444 | 0.441 | 0.580 |
| 词丢弃 | 0.551 | 0.648 | 0.678 | 0.521 | 0.557 | 0.618 | 0.695 | 0.728 | 0.565 | 0.715 | 0.451 | 0.443 | 0.597 |
| 子词丢弃 | 0.541 | 0.638 | 0.659 | 0.528 | 0.548 | 0.607 | 0.684 | 0.717 | 0.524 | 0.686 | 0.437 | 0.428 | 0.583 |
| WordSwitchOut | 0.544 | 0.647 | 0.681 | 0.522 | 0.563 | 0.621 | 0.706 | 0.719 | 0.529 | 0.702 | 0.453 | 0.452 | 0.594 |
| SubwordSwitchOut | 0.542 | 0.641 | 0.668 | 0.521 | 0.528 | 0.601 | 0.694 | 0.721 | 0.567 | 0.718 | 0.452 | 0.451 | 0.592 |
| OBPE | 0.541 | 0.629 | 0.667 | 0.504 | 0.527 | 0.589 | 0.691 | 0.715 | 0.492 | 0.721 | 0.363 | 0.611 | 0.587 |
| SDE | 0.549 | 0.636 | 0.666 | 0.513 | 0.529 | 0.591 | 0.697 | 0.735 | 0.513 | 0.731 | 0.357 | 0.618 | 0.594 |
| BPE-Dropout | 0.549 | 0.638 | 0.644 | 0.506 | 0.531 | 0.589 | 0.677 | 0.721 | 0.504 | 0.747 | 0.373 | 0.626 | 0.592 |
| 一元字符噪声 | 0.562 | 0.679 | 0.701 | 0.536 | 0.573 | 0.634 | 0.728 | 0.754 | 0.554 | 0.741 | 0.408 | 0.621 | 0.624 |
| BPE → SpanNoise（*ours*） | 0.557 | 0.676 | 0.706 | 0.542 | 0.581 | 0.651 | 0.724 | 0.755 | 0.561 | 0.751 | 0.403 | 0.622 | 0.627 |
| CharSpan (*ours*) | 0.571 | 0.695 | 0.723 | **0.556** | 0.611 | 0.685 | 0.747 | 0.772 | 0.568 | **0.759** | **0.417** | 0.627 | 0.644 |
| CHARSPAN + BPE-Dropout（*ours*） | **0.579** | **0.705** | **0.733** | 0.551 | **0.616** | **0.687** | **0.757** | **0.778** | **0.572** | 0.756 | 0.414 | **0.631** | **0.648** |

Table 9：极低资源语言到英语的零样本 COMET（使用*Unbabel/wmt22-comet-da* 模型计算）分数结果 → Englis h

| XX →EN | 印度-雅利安语支 | | | | 罗曼语族 | | | | 马来-波利尼西亚语族 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 模型 | BLEU | | chrF | | BLEU | | chrF | | BLEU | | chrF | |
| | Hin | Mar | Hin | Mar | Spa | Pot | Spa | Pot | Ind | Zsm | Ind | Zsm |
| BPE | 37.44 | 26.31 | 64.04 | 54.47 | 41.44 | 35.38 | 68.71 | 63.27 | 29.61 | 21.76 | 58.31 | 49.14 |
| 词丢弃 | 36.54 | 26.31 | 63.27 | 53.96 | 39.32 | 32.73 | 66.89 | 60.86 | 27.59 | 20.42 | 56.72 | 48.22 |
| 子词丢弃 | 36.64 | 26.22 | 63.46 | 54.57 | 39.84 | 33.04 | 67.56 | 61.58 | 26.73 | 18.80 | 57.02 | 48.82 |
| WordSwitchOut | 34.12 | 23.84 | 60.98 | 51.84 | 35.27 | 30.63 | 63.25 | 58.38 | 27.04 | 19.60 | 55.69 | 46.93 |
| SubwordSwitchOut | 37.11 | 26.03 | 63.78 | 54.06 | 42.26 | 35.68 | 68.65 | 62.97 | 27.12 | 19.76 | 55.72 | 47.34 |
| OBPE | 37.32 | 26.90 | 64.05 | 55.03 | 41.81 | 36.44 | 68.17 | 63.45 | 28.14 | 21.83 | 57.11 | 49.21 |
| SDE | 37.22 | 26.19 | 63.98 | 55.44 | 41.41 | 35.51 | 68.61 | 62.89 | 29.11 | 21.52 | 58.25 | 48.98 |
| BPE-Dropout | 37.22 | 26.93 | 64.11 | 55.31 | 41.88 | 36.72 | 68.06 | 63.79 | 30.39 | 22.54 | 59.33 | 50.17 |
| 一元字符噪声 | 37.05 | 26.95 | 63.81 | 54.83 | 39.83 | 32.91 | 67.62 | 61.24 | 28.79 | 22.01 | 57.65 | 49.91 |
| BPE → SpanNoise（*ours*） | 36.66 | 26.93 | 63.80 | 54.84 | 39.92 | 32.22 | 66.83 | 61.06 | 27.84 | 22.16 | 57.15 | 50.19 |
| CharSpan (*ours*) | 36.68 | 26.70 | 63.87 | 54.59 | 40.04 | 32.36 | 66.95 | 61.03 | 27.84 | 21.87 | 56.75 | 49.58 |
| CHARSPAN + BPE-Dropout（*ours*） | 37.62 | 27.10 | 64.15 | 55.03 | 41.21 | 33.64 | 66.90 | 61.39 | 28.91 | 22.26 | 57.99 | 50.59 |

表1：BLEU 和 chrF 分数：三个语系的高资源语言表现。可以观察到，即使加入噪声增强，所提出的模型在高资源语言上的性能也仅略有下降。

| XX →EN | 印度-雅利安语支 | | | | 罗曼语族 | | | | 马来-波利尼西亚语族 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **模型** | **BLEURT** | | **COMET** | | **BLEURT** | | **COMET** | | **BLEURT** | | **COMET** | |
| | **Hin** | **Mar** | **Hin** | **Mar** | **Spa** | **Pot** | **Spa** | **Pot** | **Ind** | **Zsm** | **Ind** | **Zsm** |
| BPE | 0.775 | 0.726 | 0.891 | 0.857 | 0.769 | 0.720 | 0.871 | 0.830 | 0.687 | 0.561 | 0.821 | 0.701 |
| 词丢弃 | 0.774 | 0.725 | 0.891 | 0.854 | 0.755 | 0.701 | 0.86 | 0.814 | 0.681 | 0.555 | 0.815 | 0.693 |
| 子词丢弃 | 0.773 | 0.725 | 0.889 | 0.854 | 0.757 | 0.691 | 0.861 | 0.806 | 0.672 | 0.548 | 0.803 | 0.683 |
| WordSwitchOut | 0.756 | 0.706 | 0.879 | 0.842 | 0.707 | 0.651 | 0.826 | 0.775 | 0.665 | 0.547 | 0.804 | 0.688 |
| SubwordSwitchOut | 0.776 | 0.724 | 0.892 | 0.855 | 0.771 | 0.721 | 0.872 | 0.833 | 0.663 | 0.548 | 0.801 | 0.687 |
| OBPE | 0.777 | 0.731 | 0.893 | 0.861 | 0.766 | 0.727 | 0.863 | 0.821 | 0.672 | 0.551 | 0.811 | 0.697 |
| SDE | 0.772 | 0.721 | 0.889 | 0.856 | 0.765 | 0.721 | 0.866 | 0.832 | 0.679 | 0.558 | 0.818 | 0.699 |
| BPE-Dropout | 0.773 | 0.727 | 0.891 | 0.858 | 0.772 | 0.7281 | 0.881 | 0.839 | 0.706 | 0.586 | 0.838 | 0.729 |
| 一元字符噪声 | 0.775 | 0.731 | 0.892 | 0.857 | 0.756 | 0.683 | 0.861 | 0.798 | 0.681 | 0.574 | 0.815 | 0.716 |
| 字节对编码 →SpanNoise（*ours*） | 0.773 | 0.728 | 0.891 | 0.857 | 0.755 | 0.685 | 0.861 | 0.801 | 0.685 | 0.581 | 0.821 | 0.724 |
| CHARSPAN (*ours*) | 0.775 | 0.726 | 0.892 | 0.856 | 0.755 | 0.681 | 0.861 | 0.799 | 0.671 | 0.569 | 0.829 | 0.714 |
| CHARSPAN + BPE-Dropout (*ours*) | 0.775 | 0.726 | 0.892 | 0.856 | 0.768 | 0.683 | 0.877 | 0.801 | 0.685 | 0.582 | 0.823 | 0.726 |

表11：BLEURT 和 COMET 得分：三个语系中高资源语言的表现

| **实验设置** | **印度-雅利安语支** | | | | | | | **平均值** |
|---|---|---|---|---|---|---|---|---|
| | **Bho** | **Hne** | **San** | **Npi** | **Mai** | **Mag** | **Awa** | |
| **chrF 分数** | | | | | | | | |
| CHARSPAN 搭配印地语、马拉地语、Pan、古吉拉特语、孟加拉语 | 38.81 | 45.39 | 30.34 | 34.4 | 41.67 | 45.82 | 43.78 | 40.03 |
| CHARSPAN 搭配印地语、马拉地语、Pan、古吉拉特语 | 37.68 | 43.49 | 28.44 | 32.22 | 39.43 | 44.34 | 42.33 | 38.27 |
| CHARSPAN 搭配印地语, Mar, Pan | 33.32 | 38.81 | 25.71 | 29.21 | 54.82 | 39.17 | 26.47 | 35.35 |
| CHARSPAN 搭配印地语, Mar | 29.70 | 33.13 | 23.83 | 26.12 | 31.88 | 33.83 | 33.13 | 30.23 |
| CHARSPAN 搭配印地语 | 20.96 | 21.92 | 15.90 | 17.97 | 20.85 | 22.85 | 21.75 | 20.31 |
| **BLEU 分数** | | | | | | | | |
| CHARSPAN 与印地语、马拉地语、Pan、古吉拉特语、孟加拉语 | 10.46 | 15.97 | 4.87 | 7.02 | 11.83 | 16.32 | 14.65 | 11.58 |
| CHARSPAN 与印地语、马拉地语、Pan、古吉拉特语 | 9.55 | 14.32 | 3.92 | 5.99 | 9.85 | 14.71 | 13.47 | 10.25 |
| CHARSPAN 与印地语 , Mar, Pan | 7.41 | 10.21 | 2.91 | 4.63 | 7.88 | 11.01 | 9.89 | 7.70 |
| 使用印地语的 CHARSPAN , Mar | 5.30 | 7.06 | 2.40 | 3.20 | 5.00 | 7.28 | 6.96 | 5.31 |
| CHARSPAN 与印地语 | 2.03 | 2.27 | 0.6 | 0.97 | 1.77 | 2.23 | 2.39 | 1.75 |

表12：字符跨度噪声增强模型的零样本多语言性能。我们针对多语言设置考虑了多种高资源语言组合。受计算资源限制，每种语言采用了100万条平行训练数据。所有语言均选自FLORES-200测试集。

| **实验设置** | **BLEU（XX → EN）** | | | **chrF（XX → EN）** | | |
|---|---|---|---|---|---|---|
| | **Gom** | **Bho** | **Hne** | **Gom** | **Bho** | **Hne** |
| 字符噪声（仅用元音替换9%-11% +） | 4.77 | 11.21 | 15.17 | 28.08 | 40.36 | 46.13 |
| 字符噪声（仅用辅音替换9%-11%+ ） | 4.79 | 11.25 | 15.3 | 26.95 | 40.51 | 46.17 |
| 字符噪声（9%-11% +替换为声音相似的字符） | 4.55 | 10.7 | 15.78 | 27.86 | 40.45 | 46.98 |
| 字符噪声（9%-11% +包含数字和标点符号） | 5.13 | 12.07 | 17.66 | 27.66 | 41.43 | 48.68 |
| 字符噪声（9%-11% +仅插入） | 5.04 | 12.3 | 17.81 | 27.50 | 41.87 | 48.74 |
| 字符噪声（9%-11% +仅替换） | 5.58 | 12.8 | 18.75 | 28.85 | 42.43 | 49.68 |
| 字符噪声（9%-11%+ 仅删除） | 4.22 | 11.92 | 18.39 | 28.65 | 42.02 | 49.36 |
| 字符噪声（4%-6% + 三种操作 + 等概率） | 5.44 | 11.66 | 18.01 | 28.62 | 40.95 | 48.63 |
| 字符噪声（14%-16% +三种操作均为 + 等概率） | 5.17 | 11.4 | 17.01 | 27.93 | 40.32 | 47.61 |
| 字符噪声（9%-11% +三种操作均为 + 等概率） | 5.21 | 12.62 | 18.29 | 28.85 | 42.53 | 49.35 |
| 字符跨度噪声（9%-11% + 1-3 元 + 替换：N 个随机字符 -> 跨度） | 3.80 | 8.80 | 13.11 | 25.38 | 28.22 | 43.39 |
| 字符跨度噪声（9%-11% + 1-3 元 + 插入：1 个随机字符 -> 跨度） | **5.84** | 13.29 | 20.49 | 29.29 | 43.51 | 51.33 |
| 字符跨度噪声（9%-11% + 1-3 元 + 插入：N 个随机字符 -> 跨度） | 4.81 | 12.21 | 17.36 | 26.98 | 41.26 | 47.91 |
| 字符跨度噪声（9%-11% + 1-3 元 + 三种操作均为 + 等概率） | 4.01 | 10.41 | 16.33 | 27.99 | 36.66 | 46.13 |
| 字符跨度噪声（9%-11% + 1-2 grams + 替换和删除 + 等概率） | 5.42 | 12.08 | 18.02 | 29.17 | 42.21 | 49.17 |
| 字符跨度噪声（9%-11% + 1-4 元组 + 替换和删除 + 等概率） | 5.79 | 11.85 | 18.02 | **29.71** | 42.41 | 49.74 |
| 字符跨度噪声（9%-11% + 1-5 元组 + 替换和删除 + 等概率） | 5.56 | 11.36 | 17.06 | 24.13 | 26.35 | 29.55 |
| 字符跨度噪声（9%-11%+ 1-3 克拉姆 + 替换和删除 +不等概率） | 5.48 | 12.12 | 18.16 | 29.01 | 41.74 | 49.37 |
| **提议：**字符跨度噪声（9%-11% + 1-3 元语 + 替换和删除 + 等概率) | **5.81** | **13.81** | **21.03** | **29.71** | **43.75** | **51.69** |

表13：消融研究与不同实验设置。其他极低资源语言和语系也观察到了类似趋势。共进行了约200项实验。

| 设置 | Gom | Bho | Hne | San | Npi | Mai |
|---|---|---|---|---|---|---|
| BPE | 26.75 | 39.75 | 46.57 | 27.97 | 30.84 | 39.79 |
| 字节对编码+极低资源语言 par | 26.54 | 42.66 | 52.52 | 31.88 | 38.09 | 43.22 |
| CHARSPAN | 29.71 | 43.75 | 51.69 | 31.40 | 36.52 | 45.84 |
| CHARSPAN+极低资源语言 par | 29.65 | 45.39 | 53.38 | 33.92 | 39.66 | 47.18 |

表14：增加1000个极低资源语言-英语平行句（ELRL$_{par}$）后的翻译质量（chrF）。

| 模型 | 印度-雅利安语支 | | | | | | | 罗曼语族 | | 马来-波利尼西亚语族 | | 平均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bho | Hne | San | Npi | Mai | Mag | Awa | Cat | Glg | Jav | Sun | |
| BPE | 0.761 | 0.793 | 0.701 | 0.744 | 0.762 | 0.809 | 0.792 | 0.721 | 0.813 | 0.731 | 0.736 | 0.760 |
| UCN | 0.853 | 0.888 | 0.765 | 0.821 | 0.849 | 0.897 | 0.883 | 0.803 | 0.879 | 0.813 | 0.811 | 0.842 |
| CHARSPAN | **0.871** | **0.909** | **0.789** | **0.858** | **0.868** | **0.913** | **0.901** | **0.831** | **0.903** | **0.846** | **0.856** | **0.867** |

表15：源高资源语言与源低资源语言表示之间的平均余弦相似度。UNC：一元字符噪声

| 模型 | 印度-雅利安语支 | | | | | | |
|---|---|---|---|---|---|---|---|
| | Bho | Hne | San | Npi | Mai | Mag | Awa |
| BPE | 0.9782 | 0.9813 | 0.9444 | 0.9624 | 0.9647 | 0.9784 | 0.9812 |
| UCN | 0.9754 | 0.9616 | 0.9504 | 0.9592 | 0.947 | 0.9708 | 0.9753 |
| CHARSPAN | **0.9856** | **0.9865** | **0.9658** | **0.9735** | **0.9802** | **0.9842** | **0.9836** |

表16：使用 L'AMBRE 的不同模型的语法良构性

### H.2 面向机器翻译的词汇表适配

早期对基于字符的机器翻译的探索展现了其潜力（Chung 等人，2016；Lee 等人，2017），包括覆盖率和鲁棒性方面的优势（Provilkov 等人，2020；Libovický 和 Fraser，2020）。然而，近期的建模研究指出，与子词模型相比，这类方法在训练/推理时间和性能方面存在一些挑战（Gupta 等人，2019；Libovický 和 Fraser，2020）。具体而言，Shaham 和 Levy（2021）表明，字符级机器翻译和字节级机器翻译（Costa-jussà 等人，2017）的性能不如字节对编码（BPE；Sennrich 等人，2016年（b））模型，从而限制了它们的实际应用（Libovický 等人，2022年）。已有研究报告了 BPE 算法（Gage，1994年）在神经机器翻译（Sennrich 等人，2016年（b））和其他多个自然语言处理任务（Liu 等人，2019年）中的有效性。SentencePiece（Kudo 和 Richardson，2018年）和 Wordpiece（Google-2018，2022年）等其他算法与 BPE 类似。我们借鉴了现有研究，并提出了一个基于 BPE 的模型。

鉴于字节对编码模型的潜力，研究者已提出多种词汇修改/生成/适配方法（Provilkov 等人，2020年；Khemchandani 等人，2021年；Patil 等人，2022年；Minixhofer 等人，2022年）。尤其是Provilkov 等人（2020年）的工作利用 BPE 算法生成词汇表，并在训练期间对不同的切分进行采样。Patil 等人（2022年）提出了一种 BPE 扩展方法，称为重叠字节对编码（OBPE），该方法在创建词汇表时同时考虑高资源语言和低资源语言的词元。他们仅在自然语言理解任务中展示了该方法的有效性。相比之下，在本研究中，我们针对机器翻译任务，在含噪高资源语言数据上采用标准 BPE 模型。

### H.3 表层/词汇层面的机器翻译噪声

若干先前研究（Sperber 等人，2017年；Koehn 和 Knowles，2017年；Karpukhin 等人，2019年；Vaibhav 等人，2019年）研究了在机器翻译中采用噪声增强策略的做法，包括在不同文本粒度上进行替换、删除、插入、翻转和交换。这些策略旨在使模型在面对自然产生的噪声（如拼写错误）时更加稳定、提升其鲁棒性。此外，这些噪声处理方案还被用于在对抗性设置下生成非规范文本（Heigold 等人，2018年）。与我们的研究较为接近的是，Aepli 和 Sennrich （2022年）提出了一种基于字符的噪声模型，以便在零样本设置下将监督信号从高资源语言迁移到低资源语言。他们使用预训练模型，在两个自然语言理解任务上评估了所提出的模型。与此不同，我们从头开始训练模型来完成机器翻译任务，这与自然语言理解任务截然不同，且更具挑战性。此外，我们

探索了跨度去噪方法，该方法优于基于字符去噪的模型，并成为极低资源语言和方言中理想的机器翻译模型。