# 藏文虚词 BPE 的藏汉机器翻译方法研究

严松思[1,2]，珠杰[1,2]*，汪超[1,2]，刘亚姗[1,2]，许泽洲[1,2]，徐泽辉[1,2]

（1. 西藏大学信息科学技术学院，西藏 拉萨 540000；
2. 省部共建西藏信息化协同创新中心，西藏 拉萨 540000）

**摘要：**本文针对藏文虚词的文法特点，设计了基于藏文虚词的 BPE 方法，首先通过全部藏文虚词 BPE、过滤兼类虚词 BPE、单音节虚词 BPE 和多音节虚词 BPE，得到四种对应语料，其次将其在 Transformer 模型和 mBART 模型上进行了实验，使用轮数集成和不同网络结构集成来提高最终模型的泛化能力。对比实验证明，藏文虚词 BPE 算法与模型集成策略可以提升藏汉机器翻译的翻译效果，最高可以达到 38.05 个 BLEU。

**关键词：**藏文虚词 BPE；机器翻译；模型集成

**中图分类号：**TP391　　**文献标志码：**A

## 1. 引言

### 1.1 研究背景

藏族是中国历史最悠久的民族之一，有着自己的语言和文字。自吐蕃王朝大臣吐弥桑布扎创制藏文起，藏文距今已有 1300 多年的历史。历史长河中，藏族人民创造了灿烂的民族文化，在文学、音乐、舞蹈、绘画、雕塑、建筑艺术等方面，留下了极为丰富的文化遗产。一方面，深入研究藏汉机器翻译有助于非藏文母语的其他学者了解和掌握藏族的历史和文化；另一方面，深入研究藏汉机器翻译能够促进民族之间的交往交流交融，有助于民族团结，同时还能够促进藏区的经济发展和对外交流。因此藏汉机器翻译的研究有非常重要的现实意义。

### 1.2 研究现状

藏汉机器翻译与其他高资源机器翻译，如英德，英汉一样，经历了基于规则的藏汉机器翻译、基于统计的藏汉机器翻译和基于神经网络的藏汉机器翻译。

在基于规则的藏汉机器翻译方面，2001 年陈玉忠等人[1]设计并实现了实用化汉藏机器翻译系统，该系统共设计了三千多个规则，并建立 20 余万词条的汉藏词典作为辅助工具提高了翻译效果，奠定了基于规则的汉藏机器翻译理论基础。

在基于统计的藏汉机器翻译方面，2011 年才让加等人[2]开展了面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究，为开发和研究汉藏统计机器翻译奠定了语料基础；而后 2012 年诺明花等人[3]提出了 CMWEPM 模型，该模型提高了汉藏多词单元等价对的召回率，从而能够间接地提高汉藏辅助翻译系统的翻译质量；2013 年董晓芳[4]进行了藏汉统计机器翻译短语抽取技术研究，该研究在 Och 短语抽取算法的基础上，对其进行改进，在一定程度上提高了翻译模型的质量。

在基于神经网络的藏汉机器翻译方面，2017 年李亚超等人[5]研究了藏汉神经网络机器翻译，该研究在藏汉语对上进行了基于注意力的神经网络机器翻译的实验，并采用迁移学习方法缓解藏汉平行语料数量不足问题；2020 年沙九等人[6]探究了不同切分粒度的藏汉双向神经机器翻译，该文章提出了藏汉双向机器翻译的具有音节、词语以及音词融合的多粒度训练方法；2020 年慈祯嘉措[7]研究了低资源语言条件下的藏汉（汉藏）机器翻译关键技术，作者通过单语语言模型的融合，迭代式回译策略的应用提高了藏汉（汉藏）机器的效果；2021 年刘赛虎[8]进行了基于注意力机制的藏汉双语机器翻译技术研究，该研究在不同切分粒度的实验中，探讨了以藏字汉字、藏词汉词、藏词汉词+BPE

**基金项目：**国家自然基金项目(62066042)；教育部人文社会科学研究项目(21YJCZH059)；2021 年西藏自治区高校人文社会科学研究项目(SK2021-24)；西藏大学提升计划项目(ZDTSJH21-07)；西藏大学培育计划项(ZDCZJH21-10)；西藏大学珠峰学科建设计划项目(zf22002001)；西藏大学研究生“高水平人才培养计划”项目(2020-GSP-S176)

* **通信作者：**rocky_tibet@qq.com

三种切分粒度形态下的 Bi-LSTM 和 Transformer 模型的藏汉机器翻译效果；2021 年头旦才让等人[9]研究了基于改进字节对编码的汉藏机器翻译，该研究改进了字节对编码算法，提出了带字数阈值的藏文字节对编码算法,优化了基于注意力机制的汉藏神经机器翻译模型。

综上所述，一直有学者从事藏汉机器翻译的研究，并且在翻译效果上取得了长足的进步，特别是大规模平行语料的构建，深度学习技术的应用，翻译效果接近于人类的水平。在语料构建、方法改进上虽然有不少亮点，但鲜少有文献从藏文自身文法特点出发研究藏汉机器翻译方法。在藏文虚词方面，目前存在一些针对藏文虚词切分的分词研究，但尚未存在藏文虚词 BPE 的藏汉机器翻译方法研究。因此，本文结合传统的 BPE 算法，利用藏文虚词丰富、虚词语法结构特殊的特点，研究了藏文虚词 BPE 的 Transformer、mBART 模型集成策略的藏汉机器翻译方法。

## 2. 藏文虚词 BPE 算法

字节对编码（Byte Pair Encoding,BPE），被用来在固定大小的词表中实现可变长度的子词，是一种数据压缩算法。该方法首先将词分成单个字符，然后依次用另一个字符替换频率最高的一对字符，直到循环次数结束。以往的一些工作也证明了 BPE 分词对藏汉机器翻译的提升作用。

而在藏文文本中，虚词用途广且出现频率相当高，在句子结构中用法和意义十分复杂。藏文虚词丰富，并且不同传统藏语文法对藏语虚词描述不尽相同。按照传统的藏文文法“三十颂”（སུམ་རྩུ་པ།）描述，藏文虚词共有 85 个。在传统的藏文文法“三十颂”之外，很多虚词与其他的字或词组合成新词。该类新词有的具有实词意义，但有的仍为虚词含义，例如 ཇི་ཙམ་；ཚད་དུ་；འདི་ཕྱིར་。本文将这些仍具有虚词意义的词称作组合性虚词。根据文献[11][12]总结了 95 个组合性虚词，同时还收集了部分助动词、存在动词等，共计 197 个。将这些虚词当成词库，称为虚词词典，经过筛选虚词，并做 BPE 处理，称为全部虚词 BPE。

在这些藏文虚词中，有部分藏文虚词还具有实词的含义，我们将其称为虚词兼类。根据文献[13]所述，共有以下 25 个既是虚词又是实词，如表 1 所示，建立兼类虚词词典，对其进行了过滤兼类虚词的 BPE 操作。

表 1 兼类虚词表

Tab.1 Tibetan Compound function words

| 虚词 | 兼类含义 | 虚词 | 兼类含义 | 虚词 | 兼类含义 | 虚词 | 兼类含义 | 虚词 | 兼类含义 |
|---|---|---|---|---|---|---|---|---|---|
| ས་ | 谁 | ན་ | 年龄、生病 | གོ་ | 懂 | ལོ་ | 年 | ནས་ | 青稞 |
| དུ་ | 烟、多少 | ནས་ | 天 | ངོ་ | 脸 | སོ་ | 牙齿 | མི་ | 人、不 |
| ར་ | 山羊 | ལས་ | 路 | མོ་ | 女 | བོ་ | 吻 | བ་ | 奶牛 |
| ཅ་ | 队 | ཡང་ | 再、又、轻 | ཏོ་ | 一双、二 | ཞིང་ | 田地 | མ་ | 母亲、不 |
| ལ་ | 坡 | ལམ་ | 职业、从 | རོ་ | 尸体 | ཤིང་ | 木 | ཤིག་ | 虱子 |

此外，由于藏文音节是藏语中重要的语言单位，像汉语一样，音节的界限很明显，但词没有明确的界限。按上文所述，藏文虚词包含单音节和多音节，在文本中切分并识别出单音节和多音节。多音节包括前文提到的组合性虚词 98 个、“三十颂”中提到的 3 个多音节虚词、以及助动词、存在动词 14 个，共计 115 个。将此类虚词合并构建了多音节虚词词典。剩余的 83 个虚词构成单音节词典。

藏文虚词 BPE 算法的整体流程如图 1 所示。首先读入待处理的藏文文本对其进行传统 BPE 操作，并利用 197 个虚词建立的虚词词典库，对句子进行虚词识别和标记处理，增加“@@”作为虚词的 BPE 标志，即图 1 中的全部虚词 BPE，得到第II类语料。其次根据上文总结的兼类虚词词典，做还原操作，过滤兼类虚词的 BPE 操作，得到第III类语料。然后，判断其是否为单音节虚词，若是过滤多音节虚词 BPE，保留单音节虚词 BPE，得到第IV类语料；否则过滤单音节虚词 BPE，保留多音节虚词 BPE，得到第V类语料。语料对应关系如表 3 所示。

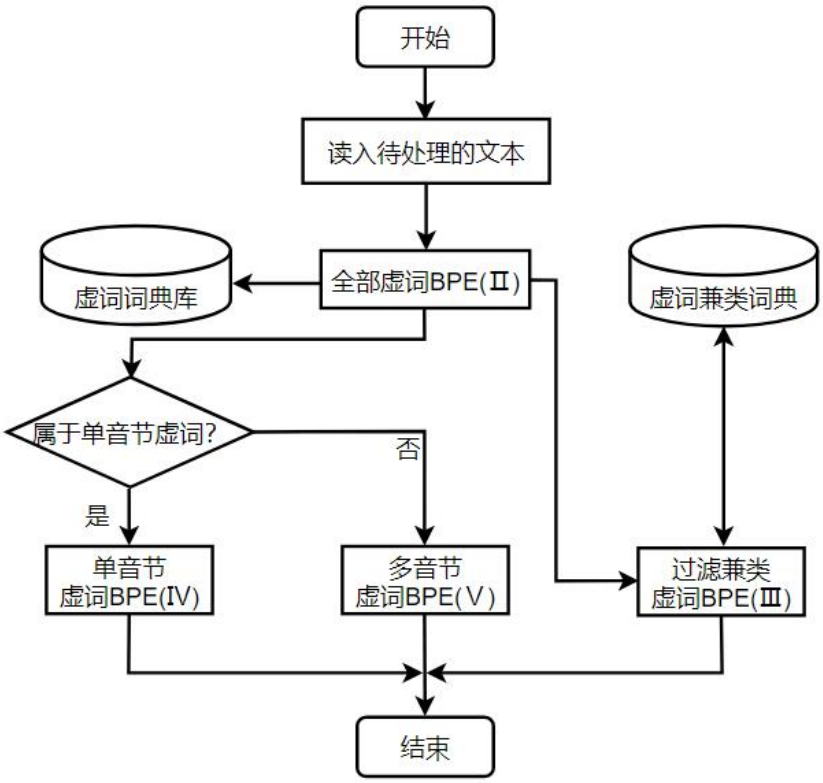


图 1 藏文文本中虚词 BPE 过程

Fig. 1 The process of function words BPE in Tibetan texts

把“བག་མ་ དང་ ཐོག་ ཏུ་ ཤེན་ཏུ་ མཇོས་པ་ ར་ སྣང་ །”按图 1 过程进行藏文虚词 BPE 后的切分结果如表 2 所示。

表 2 切分结果展示

Tab.2 Example of slitting results

| | |
|---|---|
| 汉语原始句子 | 相片 上 新娘 照 得 很 好 。 |
| 藏语原始句子 | བག་མ་ དང་ ཐོག་ ཏུ་ ཤེན་ཏུ་ མཇོས་པ་ ར་ སྣང་ ། |
| 传统 BPE 处理 | བག་མ་ དང་@@ ཐོག་ ཏུ་ ཤེན་ཏུ་ མཇོས་པ་ ར་ སྣང་ ། |
| 传统 BPE+全部藏文虚词 BPE 处理 | བག་མ་ དང་@@ ཐོག་ ཏུ་@@ ཤེན་ཏུ་@@ མཇོས་པ་ ར་@@ སྣང་ ། |
| 传统 BPE+过滤兼类虚词后 BPE 处理 | བག་མ་ དང་@@ ཐོག་ ཏུ་@@ ཤེན་ཏུ་@@ མཇོས་པ་ ར་ སྣང་ ། |
| 传统 BPE+单音节藏文虚词 BPE 处理 | བག་མ་ དང་@@ ཐོག་ ཏུ་@@ ཤེན་ཏུ་ མཇོས་པ་ ར་@@ སྣང་ ། |
| 传统 BPE+多音节藏文虚词 BPE 处理 | བག་མ་ དང་@@ ཐོག་ ཏུ་ ཤེན་ཏུ་@@ མཇོས་པ་ ར་ སྣང་ ། |

# 3. 模型介绍

## 3.1 Transformer 模型

Transformer 模型架构[14]由 6 个编码器和 6 个解码器组成，其总体架构如图 2 所示，其中左半部分是编码器，右半部分是解码器。

每个编码器都包含两个子层。第一个子层就是多头注意力层（multi-head attention layer），第二个子层是全连接层。每个子层都增加残差连接（residual connection）和归一化（normalisation）。

自注意力机制的公式如式（1）所示，其输入由维度为 d 的查询（Q）和键（K）以及维度为 d 的值（V）组成，所有键计算查询的点积，并应用 softmax 函数获得值的权重。

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \quad (1)$$

其中$\sqrt{\mathrm{d_k}}$为缩放因子。$\mathrm{d_k}$为 Q、K、V 的维度。

多头注意力机制是通过 h 个不同的线性变换对 Q，K，V 进行投影，最后将不同的自注意力机制结果拼接起来得到结果向量。多头注意力的具体计算公式如（2）、（3）所示。

$$\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_n)\mathrm{W}^{\mathrm{O}} \quad (2)$$

$$\mathrm{head}_i = \mathrm{Attention}\left(\mathrm{QW}_i^{\mathrm{Q}},\mathrm{KW}_i^{\mathrm{K}},\mathrm{VW}_i^{\mathrm{V}}\right) \quad (3)$$

其中，$\mathrm{W}^{\mathrm{O}}$是不同自注意力机制拼接之后生成最终上下文的线性映射参数。$\mathrm{W}_i^{\mathrm{Q}}$，$\mathrm{W}_i^{\mathrm{K}}$，$\mathrm{W}_i^{\mathrm{V}}$为第 i 层权重。

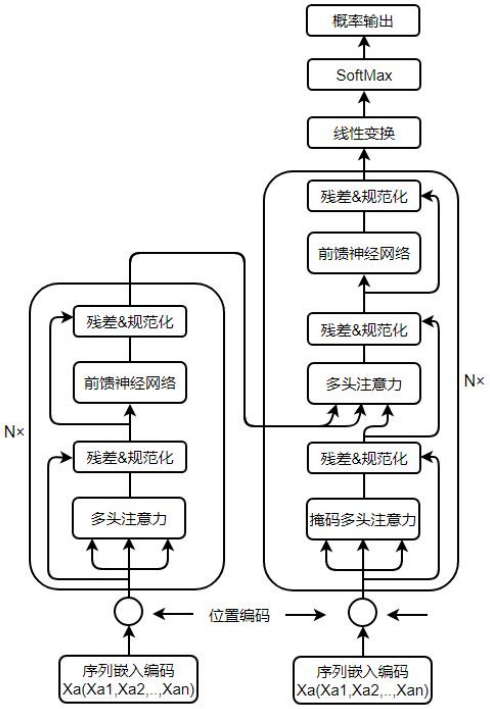


图 2 Transformer 模型架构

Fig. 2 The Transformer model architecture

第二个子层即全连接层，对每个位置变量进行相同的操作，包括两个线性变换和一个 ReLU 激活函数。

解码器结构与编码器类似。但比编码器多一层掩码注意力机制。

## 3.2 mBART 模型

提到 mBART 模型就不得不提到 BART 模型。BART 是由文献[17]提出的一种新的预训练范式，包括两个阶段：首先原文本使用某种 noise function 进行破坏，然后使用序列到序列模型还原原始的输入文本。

BART 模型在 BERT 模型和 GPT 模型[15][16]的基础上进行改进。在 BERT 模型中，随机令牌被替换为掩码，并且文档被双向编码。由于其缺失的令牌是独立预测的，因此 BERT 不能轻易地用于生成。而在 GPT 模型中，其令牌是自动回归预测的，这意味着 GPT 可以用于生成。然而，单词只能适应左向的上下文，所以它不能学习双向交互。

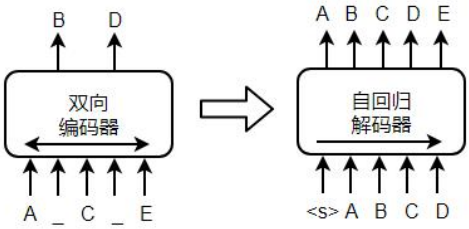


图 3 BART 的训练方式

Fig. 3 Training mode of BART

BART 模型，如图 3 所示，结合了上述两种模型的优点，它对编码器的输入不需要与解码器的输出对齐，允许任意噪声转换。在这里，文档由于用掩码符号替换文本跨度而损坏。将损坏的文档（左）用双向模型进行编码，在得到被破坏文本的编码后，使用一个类似 GPT 的结构，采用自回归的方式还原出被破坏之前的文本。BART 使用标准的 Transformer 架构，不过做了一些改变：同 GPT 一样，将 ReLU 激活函数改为 GeLU，并且参数初始化服从正态分布 N(0,0.02)，BART 解码器的各层对编码器最终隐藏层执行额外的交叉注意。

mBART[17]在 BART 模型的基础上，遵循 BART 序列到序列的预训练方案。mBART-base 模型使用标准的 Transformer 架构，包括 6 层编码器和 6 层解码器。同时在编码器和解码器的基础上包括了一个额外的层归一化层。mBART 模型的训练方式如图 4 所示。

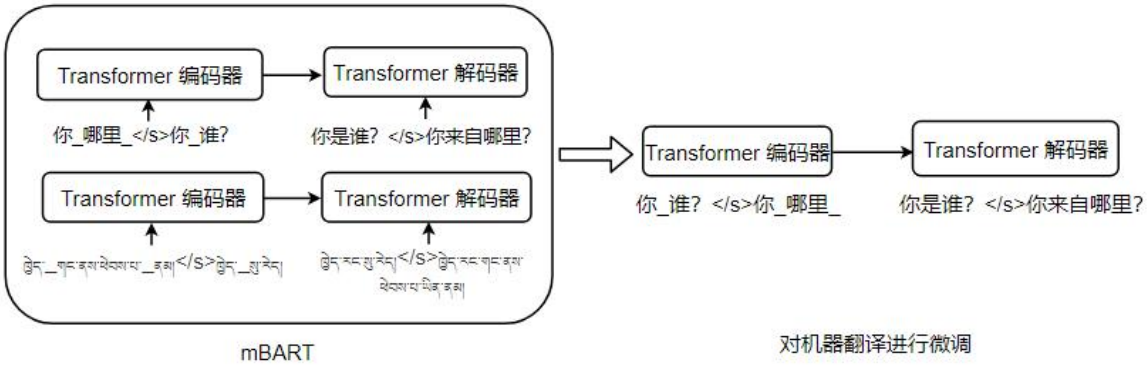


图 4 mBART 的训练方式

Fig. 4 Training mode of mBART

该模型同样采用降噪自编码器[19]，旨在采用部分损坏的输入，而后恢复原始的未失真输入。例如使用 MASK 破坏原始的序列，然后尝试模型恢复原始序列。该模型的噪声函数：在 g 中使用了两种类型的噪声，删除文本的跨度，并用 mask token 代替。如图 4 所示，第一，按照泊松分布（λ=3.5）随机抽取 token，然后对每个实例中 35%的词进行 mask；第二，对一个原始输入的不同句子进行调换顺序。

## 3.3 模型集成

模型集成是融合多个训练好的模型，基于某种方式实现测试数据的多模型融合，这样来使最终的结果能够“取长补短”，融合各个模型的学习能力，提高最终模型的泛化能力。本文采用同一模型不同训练轮数以及不同网络结构的模型集成方法。

同一模型不同训练轮数：若网络超参数设置得当，深度模型随着网络训练的进行会逐步趋于收敛，但不同训练轮数的结果仍有不同，无法确定到底哪一轮训练得到的模型最适用于测试数据。针对上述问题，一种简单的解决方式是将最后几轮训练模型结果做集成，一方面可降低随机误差，另一方面也避免了训练轮数过多带来的过拟合风险。该操作被称为“轮数集成”。

不同网络结构：不同网络结构也是一种有效的产生不同网络模型结果的方式。操作时可在不同的网络架构上训练模型，最后将不同架构网络得到的结果做以集成。

本文使用 Transformer 和 mBART 两种不同的网络结构，以及两种模型的集成来验证模型集成的效果。

# 4. 实验

## 4.1 数据集

本文实验训练数据集采用了 CCMT2021 提供的训练集和开发集，并选择 CWMT2017 的验证集和测试集分别做验证和测试。在语料预处理上对双语文本都做了分词+BPE 处理，汉文分词选择 jieba 工具包，藏文分词选择中科院提供的藏语分词软件，藏文虚词 BPE 处理过程参见第二节内容。

## 4.2 参数设置

本文采用标准的 Transformer、mBART 模型，同时对这两个模型分别进行了轮数集成，以及两种模型结构集成来验证虚词 BPE 对藏汉翻译效果的提升。实验中利用 pyTorch 框架的 Transformer、mBART 模型，参数均采用 adam 优化器，--adam-betas'(0.9,0.98)'，初始学习率均为 5e-4（防止过拟合，学习率采用逆开根号下降的方法，即$\mathrm{lr_t}=\frac{\mathrm{lr_0}}{\sqrt{\mathrm{t}}}$），采用交叉熵损失函数，以及标签平滑参数为 0.1，max tokens 为 2048；对于所有的隐藏层，都有 0.1 的随机失活率（Dropout）。在 mBART 模型中，在编码器和解码器的基础上额外添加了一个层归一化层。

## 4.3 评价指标

以 BLEU[20]分数来作为模型性能的评测指标，使用 mosesdecoder 工具中的 multi-bleu.perl 进行计算。

## 4.4 实验结果

从语料层面分析，实验分为传统 BPE 算法处理语料、传统 BPE+全部藏文虚词 BPE 处理语料、传统 BPE+过滤兼类虚词 BPE 处理语料、传统 BPE+单音节藏文虚词 BPE 处理语料和传统 BPE+多音节藏文虚词 BPE 处理语料五个部分，罗马数字与相应语料对应关系如表 3 所示。

表 3 罗马数字与相应语料对应关系

Tab.3 Correspondence between Roman numerals and the corresponding corpus

| | |
|---|---|
| I | 传统 BPE 算法处理语料 |
| II | 传统 BPE+全部藏文虚词 BPE 处理语料 |
| III | 传统 BPE+过滤兼类虚词 BPE 处理语料 |
| IV | 传统 BPE+单音节藏文虚词 BPE 处理语料 |
| V | 传统 BPE+多音节藏文虚词 BPE 处理语料 |

表 4 和表 5 分别代表在验证集和测试集上的实验结果。

表 4 验证集上实验结果

Tab.4 Experimental results on the validation set

| 模型 | | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| Transformer | Valid | 26.97 | 30.19 | 29.84 | 25.79 | 30.03 |
| mBART | | 31.48 | 31.82 | 31.46 | 30.56 | 34.80 |
| 模型集成+轮数集成 | | **36.65** | **35.91** | **36.73** | **35.40** | **38.05** |

表 5 测试集上实验结果

Tab.5 Experimental results on the test set

| 模型 | | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| Transformer | Test | 28.13 | 28.19 | 29.47 | 26.04 | 30.53 |
| mBART | | 30.12 | 31.26 | 33.52 | 30.00 | 31.92 |
| 模型集成+轮数集成 | | **32.51** | **33.18** | **33.43** | **32.10** | **33.07** |

从对语料进行不同操作得到的实验结果来看，如表 4、表 5 所示。与传统 BPE 相比，全部藏文虚词 BPE 方法对两种模型而言没有太多提升，甚至在部分实验中出现了 BLEU 值下降情况，说明对全部虚词简单的筛选并不能提升翻译效果。而III相比于I的翻译效果有了改进，说明兼类虚词问题对藏汉翻译的影响较大。而IV相比于I、II来说都出现了 BLEU 的下降情况，这是由于兼类虚词问题集中于单音节虚词中。V相比于I的提升较为显著，这是由于组合性虚词在藏文中的意义较为重要，将其拆分开来的话反而会破坏句子原有的含义。上述实验证明了藏文虚词 BPE 算法在藏汉机器翻译上的可行性。

从模型层面分析，实验分为 Transformer、mBART 和模型集成+轮数集成三个部分。如表 4、5

所示。mBART 模型的翻译效果整体上要优于 Transformer 模型。在进行模型集成后，翻译效果较 Transformer 模型和 mBART 模型都有提升，其翻译效果达到最佳。在验证集中，BLEU 值均达到 35 以上；在测试集中，BLEU 值均达到 32.1 以上。这也充分说明了模型集成对藏汉机器翻译的效果改进。

# 5. 结论

本文从藏文自身文法特点出发，针对藏文虚词进行了全部虚词 BPE、过滤兼类虚词的 BPE、单音节虚词 BPE 和多音节虚词 BPE，得到对应语料。在 Transformer 模型和 mBART 模型的基础上，通过模型集成和轮数集成，探究了藏文虚词 BPE 对藏汉机器翻译的改进，同时探究了模型集成策略对翻译效果的影响。通过对比不同模型及它们的集成实验证明，通过藏文虚词 BPE，可以有效提升藏汉机器翻译的效果，效果最高可以达到 38.05 个 BLEU，较传统 BPE 提升 1.4 个 BLEU。

本文对藏文文法特点的挖掘还不够深入，以音节做切分的方式欠妥，未来需考虑词频，藏文虚词中紧缩词问题等影响因素。在未来的工作中，将更加深入研究藏文文法特点，并将其应用于下游的翻译任务中。

# 参考文献：


[1] 德盖才郎(陈玉忠)等. 实用化汉藏机器翻译系统的设计与实现智能计算机接口与应用进展[M]. 电子工业出版社, 2001,404-411.

[2] 才让加. 面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究[J]. 中文信息学报, 2011(6)：157-161.

[3] 诺明花, 刘汇丹, 吴健, 等. 基于关联度的汉藏多词单元等价对抽取方法[J]. 中文信息学报, 2012(3)：98-103.

[4] 董晓芳. 藏汉统计机器翻译短语抽取技术研究[D]. 西北民族大学, 2013.

[5] 李亚超, 熊德意, 张民, 等. 藏汉神经网络机器翻译研究[J]. 中文信息学报, 2017,31 (06):103-109.

[6] 沙九, 冯冲, 张天夫, 等. 多策略切分粒度的藏汉双向神经机器翻译研究[J]. 厦门大学学报(自然科学版), 2020,59(02):213-219.

[7] 慈祯嘉措. 贫语言资源条件下的藏汉(汉藏)机器翻译关键技术研究[D]. 青海师范大学, 2020.DOI:10.27778/d.cnki.gqhzy. 2020.000541.

[8] 刘赛虎. 基于注意力机制的藏汉双语机器翻译技术研究[D]. 西藏大学, 2021.DOI:10.27735/d.cnki.gxzdx. 2021.000019.

[9] 头旦才让, 仁青东主, 尼玛扎西,等. 基于改进字节对编码的汉藏机器翻译研究[J]. 电子科技大学学报, 2021,50(02):249-255+293.

[10] 格桑居冕. 实用藏文文法[M]. 成都:四川民族出版社, 1987.

[11] 珠杰. 藏文文本自动处理方法研究[M]. 成都:西南交通大学出版社, 2018.

[12] 高定国, 扎西加, 赵栋材. 计算机识别藏语虚词的方法研究[J]. 中文信息学报, 2014,28(01),113-117.

[13] 拉巴顿珠, 欧珠, 赵栋材. 藏文自动分词系统中虚词识别算法研究[J]. 计算机应用与软件, 2017,34(09):299-301+333.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In Advances in Neural Information 7 Processing Systems, pages 5998-6008.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2018. Bert: Pre-training of deep bidirectional Transformers for language understanding. arXiv preprint arXiv:1810.04805.

[16] Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. arXiv preprint arXiv:1901.07291.

[17] Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." ACL 2020: 7871-7880

[18] Liu, Yinhan, et al. "Multilingual denoising pre-training for neural machine translation." TACL.2020

[19] Alexis Conneau, German Kruszewski, Guillaume Lam ple, et al. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In ACL.

[20] Papineni, K., Roukos, S., Ward, T., et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318 (2002).

# A Study on the Tibetan-Chinese Machine Translation Method for Tibetan Function Words BPE

YAN Songsi[1,2], ZHU Jie[1,2*], WANG Chao[1,2], LIU Yashan[1,2], XU Zezhou[1,2], XU Zehui[1,2]

(1. School of Information Science and Technology, Tibet University, Lhasa 540000, China; 2. Provincial and Ministerial Collaborative Innovation Centre for Informatization in Tibet, Lhasa 540000, China)

**Abstract:** In this paper, a BPE method based on Tibetan function words is designed to address the grammatical characteristics of Tibetan function words. Firstly, the four corresponding corpora are obtained through Tibetan function words BPE, filtered cum class function words BPE, monosyllabic function words BPE and polysyllabic function words BPE. And secondly, they are experimented on Transformer model and mBART model, using epoch ensemble and ensemble of different network structures to improve the generalization ability of the final model. The comparative experiments demonstrate that the Tibetan function words BPE algorithm and model integration strategy can improve the translation of Tibetan-Chinese machine translation up to 38.05 BLEU.

**Keywords:** Tibetan function words BPE; machine translation; model integration