基于词典注入的藏汉机器翻译模型预训练方法

桑杰端珠[1,2]　　才让加[1,2]

[1] 青海师范大学，计算机学院，西宁，810000

[2] 青海师范大学，藏语智能信息处理及应用国家重点实验室，西宁，81000

sangjeedondrub@live.com　　zwxxzx@163.com

摘要

近年来，预训练方法在自然语言处理领域引起了广泛关注，但是在比如藏汉机器等低资源的任务设定下，由于双语监督信息无法直接参与预训练，限制了预训练模型在此类任务上的性能改进。考虑到双语词典是丰富且廉价的先验翻译知识来源，同时受到跨语言交流中人们往往会使用混合语言增加以沟通效率这一现象启发，本文提出一种基于词典注入的藏汉机器翻译模型的预训练方法，为预训练提供学习双语知识关联的广泛可能。经验证，该方法在藏汉和汉藏翻译方向测试集上的 BLEU 值比 BART 强基准分别高出 2.3 和 2.1，证实了本文所提出的方法在藏汉机器翻译任务上的有效性。

关键词： 藏汉 ； 机器翻译 ； 预训练 ； 词典注入

Dictionary Injection Based Pretraining Method for Tibetan-Chinese Machine Translation Model

Sangjie Duanzhu[1,2]　　Cairangjia[1,2]

[1] School of Computer Science, Qinghai Normal University, Xining, 810000

[2] The State Key Laboratory of Tibetan Information Processing and Application, Qinghai Normal University, Xining 810000

sangjeedondrub@live.com　　zwxxzx@163.com

Abstract

In recent years, pretrained models have attracted extensive attention in the field, however, due to bilingual supervision can not directly participate in the pretraining process, pretrained models are not contributive under low-resource settings such as Tibetan-Chinese machine translation. Given bilingual dictionaries are rich and low-cost source of prior translation knowledge and inspired by the phenomenon that people often use mixed lexicons for better communication in cross-lingual conversations, this paper proposes a technique to pretrain the Tibetan-Chinese machine translation model via dictionary injection, which provides a wide range of possibilities for bilingual knowledge interaction. Empirical results show the proposed method can produce improvements of 2.3 and 2.1 in BLEU scores on test set for Tibetan-Chinese and Chinese-Tibetan translation directions over strong BART baselines, indicating the effectiveness of the proposed method.

Keywords: Tibetan-Chinese , Machine Translation , Pretraining , Dictionary Injection

基金项项目：青海省重点研发与转化计划项目 (2022-GX-104) 、青海省中央引导地方科技发展资金项目 (2022ZY006)

通信作者：才让加 (zwxxzx@163.com)

1 介绍

目前神经机器翻译 (Neural Machine Translation, NMT) (Sutskever et al., 2014; Gehring et al., 2017; Ashish et al., 2017) 已经成为最主流机器翻译方法，在性能上全方位超越传统短语统计翻译模型 (Statistical Machine Translation, SMT) (Brown et al., 1990)，并成为工业界机器翻译服务系统的标准实现方法 (Wu et al., 2016), 甚至研究者声称在特定领域和语言对上 NMT 的性能可以接近甚至超越人类的翻译水平 (Hassan et al., 2018)。与 SMT 不同的是 NMT 以端到端风格的建模方式将翻译决策过程视为单个条件概率模型的参数估计过程，从而摒弃了 SMT 不同组件独立优化各自训练目标的建模范式。但是目前 NMT 卓越的性能表现是以具备大规模、高质量和多领域对齐数据为重要前提的，受制于市场规模较小、数据标注成本高昂等客观因素，现阶段藏汉机器翻译的质量距离中英等主流语言存在巨大的差距。

在对齐数据受限的条件下，对于多数语言而言，单语数据的来源相对较为广泛且容易收集，研究者自然地探索了各类在 NMT 框架内有效利用目标端和源端单语数据的方法。其中最简单和直接的是回译方法 (Sennrich et al., 2016), 该方法利用监督式方法训练一个初始的反向模型，将目标端的单语数据进行翻译，用于扩充训练正向模型的数据。回译方法不仅能改善低资源场景下的翻译性能，同时在富资源场景中也能缓解领域适应等问题 (Kumari et al., 2021)。回译方法要求初始回译模型本身有较高的性能，但是在现实中很多低资源语言的对齐数据无法保证初始回译模型的性能。

近年来，受到计算机视觉研究的启发 (He et al., 2016), 在未标注的海量文本数据、高阶的分布式优化方案、强大的序列学习模型和高性能计算加速设备的共同加持下自监督式预训练 (Self-supervised Pretraining) 模型 (Devlin et al., 2019; Brown et al., 2020; Liu et al., 2019) 激起了自然语言处理 (Natural Language Processing, NLP) 领域内的研究热潮。预训练模型使得研究者可以不用从头训练昂贵和复杂的大规模模型，直接使用现有预训练模型在下游的目标任务上结合任务自身特点进行微调，就往往可以获得比监督式训练更好的性能表现。在诸多的预训练模型中具有代表性的包括掩码语言模型 (Masked Language Model, MLM) BERT (Devlin et al., 2019); 自回归语言模型 (Autoregressive Language Model, ALM) GPT (Radford et al., 2019); 置换语言模型 (Permuted Language Model, PLM) XLNet (Yang et al., 2019); 降噪自编码器模型 (Denoising Auto Encoder, DAE) BART (Lewis et al., 2020) 等。其中 BERT 和 XLNet 语言模型是 Transformer (Ashish et al., 2017) 的编码器，能对语言序列进行双向的表示学习，主要用于序列的语义理解。GPT 使用了 Transformer 的解码器，结合已生成的解码片段和当前时刻的输入，以自回归的方式逐词生成目标序列。而 BART 可以视为结合 BERT 和 GPT 泛化的预训练模型，与 BERT 和 GPT 不同的是，BART 采用序列到序列的建模方式，使用单个 Transformer 模型对编码器端完成各类加噪操作的输入序列在解码器端完成重构，通过降噪自编码为优化目标完成整个解码器和编码器的联合预训练，然后在下游的目标任务上通过标注数据进行微调，非常适合于机器翻译和知识问答等采用编码器-解码器构架的建模任务。BART 是针对单一语言 (英语) 的预训练，而随后提出的 mBART (Liu et al., 2020) 则是将 BART 的建模方式扩展到多语言场景下，完成多语言模型的预训练。同样是采用 BART 训练目标的 M2M-100 (Fan et al., 2021) 更是进一步扩大了所覆盖的语言种类，支持 100 个语言之间的多对多翻译。对于藏文这种低资源语言而言，多语言预训练是一个非常具有吸引力的设想，因为除了支持多语言翻译外，M2M-100 级别的大规模预训练模型本身能够有效支持通用语义知识的迁移。但是 mBART 和 M2M-100 的训练都没有包含藏文。本文旨在探索训练 BART 风格的藏汉翻译预训练模型的有效方法，为后续的藏语多语言翻译课题提供研究基础。

BART 在预训练过程中主要学习当前输入语言的表示和分布，缺乏双语对齐监督信号的直接参与，没有显式地学习语言对之间的映射关系。这种预训练方式不利于平行资源匮乏的藏汉语言对的预训练效果。考虑到双语词典是重要的先验知识来源，人类语言学者在学习一门新语言时，往往会

第二十一届中国计算语言学大会论文集，第374页–第383页，南昌，中国，2022年10月14日至16日。

借助双语词典探索所要学习的语言，通过词典建立新语言和其他已掌握的语言之间的关联；此外人类翻译人员也会使用双语词典推敲用词、查询专业词汇，以改善翻译工作的质量。另外，受到跨语言交流过程中使用混合语言往往能够增加沟通效率 (Matras, 2000) 这一现象的启发，本文提出了一种基于双语词典注入的藏汉预训练翻译模型的训练方法，即基于词典注入的藏汉机器翻译预训练模型 (Pretrained Translation Model with Dictionary Injection, PTMDI)。通过构建较大规模双语词典，然后利用词典对大规模的藏汉单语数据进行跨语言数据注入，以降噪自编码为训练目标完成藏汉机器翻译模型的预训练。词典的数据注入如表 1 所示。

表 1: 词典注入样例

原始输入	当代中国，江山壮丽，人民豪迈，前程远大。新时代为我国文艺事业发展提供了前所未有的广阔舞台。
词典替换	དེང་རབས་ གྲུང་གོ་ ，江山壮丽，མི་དམངས་ 豪迈，前程远大。གསར་པ་ དུས་རབས་ 为 རང་རྒྱུད་ 文艺 ལས་དོན་འཛེལ་རྒྱས་ 提供了前所未有的广阔 གངས་སྤེལ་གས་ 。

被替换的汉文词都使可以被视为是一种对原始文本的加噪，另外由于藏汉两种语言在语序上的偏差，比如藏文的定语普遍后置，新时代的正确翻译是 དུས་རབས་ (时代) གསར་པ་ (新)，进行词典替换之词序变为 གསར་པ་ དུས་རབས་ ，但是这种词序颠倒可视为额外的加噪操作。基于掩码的降噪自编码的训练中，遮蔽连续的词比单独的词（比如 BERT）能使模型学习到更好的表示 (Joshi et al., 2020)。同样地, 在基于词典注入的预训练方法中，连续词条的替换（比如 དེང་རབས་ (当代) གྲུང་གོ་ (中国)），同样符合这种思想。与 BART 中的加噪方式（见图 1）不同的是，通过词典替换的加噪方案，在加噪的同时为模型提供一个学习双语知识关联的广泛可能，客观上要求模型在联合学习双语语义对齐信息和序列上下文进行以降噪为目标的解码，学习到跨语言的联合表示，为在平行数据受限的条件下进行有效进行微调提供了便利。此外词典注入的预训练方法能够借助词典学习到目标领域的翻译知识，对机器翻译的领域适应提供了一个可行且低廉的可选方案。

在规模分别为 6.9 M 和 5.2 M 句子规模的藏汉单语数据、500 K 句对的藏汉平行数据和 314 K 词条双语词典的数据设定下，本文中的 PTMDI 模型在藏汉和汉藏翻译方向的测试集上的 BLEU 值比 BART 这一强基准模型分别高出 2.3 和 2.1，充分证实了本文所提出的预训练方法在藏汉机器翻译任务上的有效性。

综上，本文的贡献为：

1. 考虑到双语词典能在预训练过程中提供有效的监督信号，同时受跨语言交流中使用混合的多语言词汇能提高沟通效率这一现象启发，提出一种利用藏汉双语词典和藏汉单语数据进行词典注入的机器翻译预训练方法，即 PTMDI;

2. 在通过与包括监督式 Transformer, 回译，BART 的性能对比实验，证实本文提出的 PTMDI 方法比各类基准模型在测试数据集上都有大幅的性能提升；

3. 由于使用了藏汉双语词典，本文的提出的 PTMDI 模型适用于翻译模型的领域适应问题，能够借助领域词典和单语数据学习平行数据中缺乏的翻译知识。

2 相关工作

近年来随着人工智能领域技术的迅猛发展和日益密切的跨语言交流需求，藏汉机器翻译技术取得了长足发展。和其它低资源机器翻译研究课题一样，藏汉机器翻译的研究集中在致力于在平行数据资源受限的条件下探索提高机器翻译性能的方法。其中包括优化藏汉翻译模型的词表大小和分布 (孙义栋 et al., 2022; 头旦才让 et al., 2020), 利用大规模单语数据进行迭代式回译 (慈祯嘉措 et al.,

第二十一届中国计算语言学大会论文集，第374页-第383页，南昌，中国，2022年10月14日至16日。

2020)，迁移学习 (李亚超 et al., 2017), 融合藏文多层次先验特征 (沙九 et al., 2020), 融合目标端语言模型的方法 (慈祯嘉措 et al., 2019) 等。此外还有一些与藏文预训练语言模型相关的研究工作，比如 中国少数民族预训练语言模型 CINO (Yang et al., 2022)。该模型使用了 XLM-R (Conneau et al., 2020) 风格的预训练方法，是至迄今为止规模最大的支持藏文的公开跨语言预训练语言模型。CINO 虽然只在文本分类任务上进行了测试和验证，由于该模型可以进行跨语言的表示，可以用于初始化藏汉机器翻译的解码器、编码器或者整个模型的参数。

3 方法

NMT 给定源端句子 $x = \{x_1, \cdots, x_N\}$ 和目标端句子 $y = \{y_1, \cdots, y_M\}$，NMT 将句子级别的翻译概率建模问题转换为词级别的条件概率的积,

$$P(y|x;\theta) = \prod_{j=1}^{M} P(y_i|x, y_{<i};\theta) \tag{1}$$

其中 θ 为模型所要估计的参数，$y_{<i} = \{y_1, \cdots, y_i\}$ 为 i 时刻已生成的目标序列片段。在 Transformer 构架中序列转导 (Sequence Transduction) 建模任务由自注意力网络完成。NMT 模型通常采用交叉熵损失作为训练目标，通过最大化整个训练数据上预测序列和目标序列词级别的似然进行训练；为了提高译文的质量，往往会采用束搜索策略，在牺牲一定推理速度的条件下扩大模型的搜索空间。

机器翻译预训练模型 BERT 之类的掩码语言模型能够对序列的双向上下文表示进行建模，但是其训练是按照分类任务进行的，即将编码器的输出输入到 `Softmax` 层预测被掩码的词在整个词表上的概率分布。GPT 之类的自回归模型和传统的语言模型的训练方式一致，即通过当前已生成序列的信息预测下一个词。BART 将类似 BERT 具有双向表示能力的构架作为编码器学习加噪序列的表示，而将类似于 GPT 的自回归构架运用于解码器，用于逐词生成原始未加噪的序列。其训练的优化目标为在整个训练集 D 上加噪序列片段与原始序列片段的似然概率，即

$$\arg\max_{\theta} \mathcal{L}_{\theta} = \arg\max_{\theta} \sum_{x \in D} log(P(x|\mathcal{N}(x);\theta)) \tag{2}$$

其中 $\mathcal{N}(\cdot)$ 表示加噪函数，在 BART 在预训练过程中采用了多个加噪方法，包括：BART 在预训练程中采用了多个加噪的方法，包括 1) 词的遮蔽、2) 句子顺序扰动、3) 文档转换、4) 词删除、5) 序列片段替换等，这些加噪方法的示意请见图1。

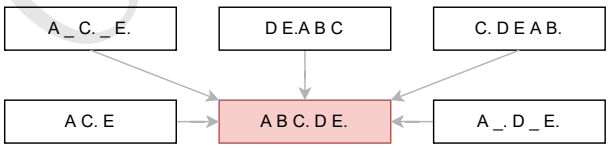

图 1: BART 的加噪方法示意

词典注入的藏汉机器翻译预模型训练方法 PTMDI 的预训练沿用了 BART 加噪并重构的建模方法, 但是与 BART 不同的是 PTMDI 中词典注入代替了各类加噪方案。词典的注入不仅能起到加噪的作用，同时也在客观上要求编码器学习跨语言的联合表示。有关双语词典的获取、筛选和注入请见 4.1 节。本文中在完成词典注入的单语数据上进行预训练之后，并在规模为 500 K 平行数据上进行微调。具体的预训练和微调的示意请见图 2 和图 3。考虑到收集的双语词典的词条大部分为名词，在进行词典注入时优先替换单语数据中的名词，同时保证被替换的词的数量不超过整个句子词长度的 15 %.

第二十一届中国计算语言学大会论文集，第374页-第383页，南昌，中国，2022年10月14日至16日。

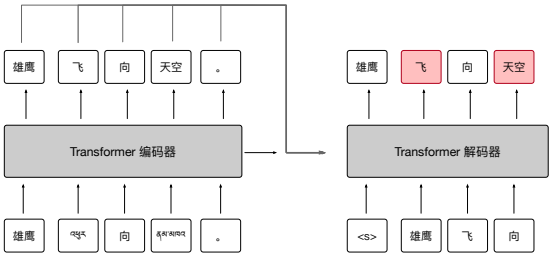

图 2: 预训练过程

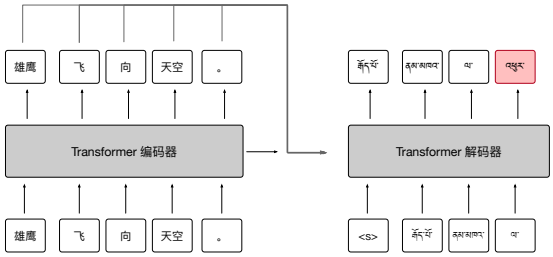

图 3: 微调过程

因为编码器需要学习藏汉两种语言的表示，需要模型有更大的学习容量，所以本文中使用了相较于解码器更深的网络构架。此外编码器的表示和理解性能相对而言比解码器的自回归生成和掩码自编码性能，对翻译终表现有更加重要的影响 (Kasai et al., 2020), 因而在多语言机器翻译任务研究者有使用较深的编码器较浅的解码器的应用实践 (Kong et al., 2021)，在翻译性能不退化的前提下，提高翻译速度。

PTMDI 训练方法能通过注入词典的方式进行翻译模型的预训练，因为词典的对齐特性使得模型在预训练阶段就开始跨语言的信息交互，学习跨语言信息的关联。此外，这种词典注入方式使得离散的词典特征能够很好地整合到端到端序列序列学习的连续过程中，是一种在机器翻译模型中有效融合先验知识的方法。考虑到相较于特定领域内的对齐数据，领域词典和领域单语数据比较容易获取和收集，所以 PTMDI 也是一种能以较为低廉的代价进行机器翻译领域适应的方法，尤其是适用于藏汉语言对这样的低资源机器翻译任务。

4 实验

4.1 数据设定

词典 为了藏汉双语词典涵盖较为广泛的领域，尤其是学习到受限的藏汉对齐文本之外的翻译知识，本文使用藏汉、汉藏、藏英、英藏四个方向的双语词典资源和利用统计词对齐工具 `FastAlign`[1] (Dyer et al., 2013) 在藏汉平行数据中获取的藏汉对齐词表。其中所有词典数据中只提取有单个释义的词条，另外对于藏英、英藏词典先将英文通过 Google 在线翻译系统翻译为汉文，然后再进行筛选处理; 对于统计对齐词表设定筛选的词对齐概率阈值为 `0.3` ，若有多个超过该阈值的对齐词表项则随机选择。词典词源的统计信息请见表 2, 藏汉和汉藏词典的领域包括日常用词、法律、生物、化学、医疗、数学、计算机等，藏英和英藏词典则主要是日常用词。对如表 2 所示的总计 `384654` 个筛选的词条进行正则化和去重处理之后，最终获得 `314500` 个独立词条。

表 2: 词典资源统计表

词典	词典数量	总词条数目	筛选的词条数目	筛选比例 (%)
藏汉词典	7	451200	153020	33.9
汉藏词典	5	341000	120000	35.2
藏英词典	3	130200	29949	23.0
英藏词典	2	177876	36685	20.6
统计对齐词表	-	95699	45000	47.0
总计		1195975	384654	32.2

[1] https://github.com/clab/fast_align

双语数据 与英文等具有显式的词分隔符不同，比如藏文和汉文如果直接使用纯粹基于频率统计的子词分词方法，将可能会生成大量在语言学上无实际意义的子词结构，这一现象对藏文这种拼音文字尤其明显。在低资源的机器翻译任务设定中，这些冗余的子词使得机器翻译模型需要学习额外的构词规律，在客观上加大了模型的学习负担。除了低资源机器翻译任务之外，涉及汉文、日文、朝鲜文等语言的富资源机器翻译任务中一般也是采用先分词再学习子词的数据预处理流程 (Alexis and Guillaume, 2019)。本文中数据的预处理也是采用了这种策略，汉文分词使用了 `jieba`[2]分词工具进行分词，藏文藏文分词采用了 (桑杰端珠 and 才让加, 2018) 提出的藏文分词方法。对文本进行分词处理之后使用了 `SentencePiece`[3] (Kudo and Richardson, 2018) 子词学习。为了过滤平行数据中的噪音样本，本文通过 `fasttext`[4] (Joulin et al., 2016) 中的语言标识模型去除藏文句子中的汉文和汉文句子中的藏文，同时也删除了数据样本中的非 Unicode 字符。本文限制了对齐句对的最大长度为 120 个词，同时剔除了藏汉词长度比大于 4 的句对。通过取重方法保证训练集、验证集和测试集没有交集。最终的藏汉平行数据规模见表 3 。

表 3: 平行数据和单语数据规模

数据类型	平行数据 (句对)	单语数据 (藏/汉) (句对)
训练集	500 K	6.8 M / 5.2 M
验证集	5 K	63 K / 51 K
测试集	5 K	62 K / 51 K
总计	510 K	6.9 M / 5.2 M

单语数据 由于用于微调的平行数据主要是新闻领域的，为了更加有效的模型训练，本文在收集藏语和汉语的单语数据时也使用了新闻领域的数据。单语数据的主要来源是各类藏文新闻网站和这些网站对应汉文网站的对应栏目，以完成数据更好的领域适配。单语数据的预处理方式和平行数据的预处理方式是一致的，也是先分词，再学习子词。在进行正则去噪、去重等预处理之后，最终保留的藏文和汉文单语数据的规模为反而别为 6.9 M 和 5.2 M。

4.2 模型设定

本文中所有模型的训练和测试都是基于 Fairseq [5] (Ott et al., 2019) 框架实现的，使用了 4 张 Nvidia Quadro P1000 GPU。基准模型中纯监督式模型和回译模型使用了 6 层的 Transformer 编码器和解码器；藏文和汉文的词表大小分别为 8K 和 9K。PTMDI 模型使用了 10 层的 Transformer 编码器和 6 层的 Transformer 解码器，编码器共享了藏语和汉语的词表，解码器使用了独立的对应目标语言的词表。所有模型解码器和编码器的嵌入维度为 512，编码器和解码器的前馈网络的维度为 2048，使用了 Adam 优化器进行参数优化，初始学习率设置为 0.001，学习率衰减函数选用了平方根倒数，批处理大小为 4096 个词, 所有的模型都训练了 60 轮次。

4.3 实验结果

表 4 列出了纯监督式 Transformer 模型、回译模型、BART 和 PTMDI 模型在测试集上的最终 BLUE 的测定值。从表中可以看出，本文中的 PTMDI 模型比 BART 这一强基准模型在藏汉和汉藏翻译任务上 BLEU 值分别高出 2.3 和 2.1，用实证方法证实了 PTMDI 在藏汉机器翻译任务上的

[2] https://github.com/fxsjy/jieba

[3] https://github.com/google/sentencepiece

[4] https://github.com/facebookresearch/fastText

[5] https://github.com/pytorch/fairseq/

有效性。此外从图 4 中模型在验证集上的 BLEU 变化和图 5 中训练过程中的损失变化，可以得知 PTMDI 模型有更好的收敛特性，证实了模型在预训练阶段就通过词典学习双语映射关系确实能够帮助微调过程中模型的学习能力。

表 4: 各个模型在测试集上 BLEU 值

模型	藏 → 汉	汉 → 藏
Transformer	27.1	26.8
回译	28.3	27.2
BART	29.8	29.3
PTMDI	**32.1**	**31.4**

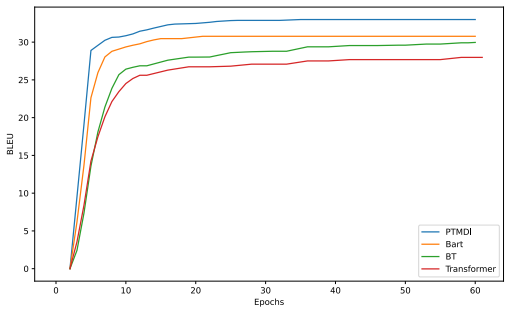

图 4: 各个模型验证集上的 BLEU 变化 (藏 → 汉)

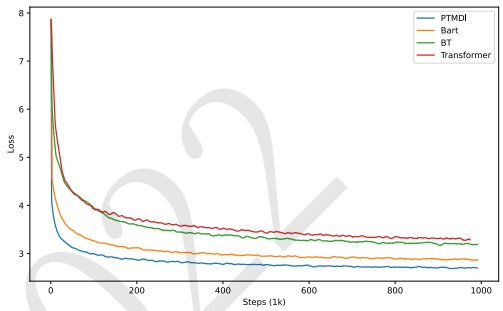

图 5: 各个模型的训练损失变化 (藏 → 汉)

从表 5 可以看出，在测试集样例中的专业词汇食用菌和羊肚菌在 PTMDI 模型中被较为准确地被译出，且译文更加流畅。除了验证模型在双语数据的领域有良好性能之外，本文还对其他跨领域场景下的性能进行了测试，如表 6 所示的是测试所有模型在计算机科学领域表现的一个样式示意，从该译文样例中可以发现比如汇编，编译器等双语平行数据中不存在的词条也被准确翻译出来。说明 PTMDI 确实在预训练过程中挖掘了先验的双语词典内的翻译知识。

表 5: 测试集中的译文样例

原文	ཞེང་ཡས་ཚན་རིག་ཁང་གི་ཟས་སྐྱོང་འབྲུ་ཐེབ་འཇུག་ཁང་གི་ཞེབ་འཇུག་པ་གཞོན་པས་མིའི་ཐབས་ཀྱི་གཞི་ཚོན་ཅན་གྱི་ལུག་ས་འབྲུ་ཐའེ་བས་འཇུག་སྐེད་པར་ཞེབ་འཇུག་རེད།
参考译文	农业科学院食用菌研究所副研究员研究人工规模化种植羊肚菌。
模型	译文
Transformer	农业科学院进食菌类研究所年轻研究员研究人为种植羊肚子细菌。
BT	农业科学院的吃饭细菌研究所的副研究员正在研究人为规模的羊肚子细菌的种植。
BART	来自农业科学院的可食用细菌研究所的助理研究员研究了规模种植羊菌的技术手段。
PTMDI	来自农业科学院的食用菌研究所的助理研究员研究了人工有规模种植羊肚菌。

表 6: 跨领域的译文样例

输入	把高级程序语言翻译成汇编语言或机器语言的工作称为编译，完成这项翻译工作的软件系统称为编译程序或编译器。
模型	译文
Transformer	མཐོ་རིམ་བྱ་རིམ་སྐད་ཡིག་སྐད་བརྒལ་འཕྲུལ་ཆས་སྐད་ཆ་ལ་བསྒྱུར་བའི་བྱ་བ་དེ་ལ་རྩོམ་སྒྲིག་ཟེར། ལོ་རྩྭའི་ལས་ཀ་འདི་འགྲུབ་པའི་ཚེད་དུ་མཉེན་ཆས་མ་ལག་འདི་ལ་རྩོམ་སྒྲིག་བྱ་རིམ་ཡང་ན་རྩོམ་སྒྲིག་ཟེར།
BT	མཐོ་རིམ་བྱ་རིམ་གྱི་སྐད་ཆ་དེ་རྩོམ་བསྒིགས་སྐད་ཆ་ཡང་ན་འཕྲུལ་ཆས་སྐད་ཆ་ལ་བསྒྱུར་བའི་བྱ་བ་དེ་མིང་ལ་རྩོམ་བསྒྱུར་ཟེར་ལ། ཡིག་སྒྱུར་ལས་དོན་འདི་འགྲུབ་པའི་ཚེད་དུ་མཉེན་ཆས་མ་ལག་ལ་རྩོམ་བསྒྱུར་བྱ་རིམ་ཟེར་པའོ།།
BART	མཐོ་རིམ་བྱ་རིམ་གྱི་སྐད་ཆ་དེ་གུང་སྒྲིག་པ་སྐད་ཆའམ་འཕྲུལ་ཆས་སྐད་ཆ་ཟེར། ལོ་རྩྭའི་བྱེད་པའི་ལས་ཀ་དེ་ལ་གུང་སྒྲིག་ཏུ་འགོད་པ་དང་། ལོ་རྩྭའི་ལས་ཀ་འདི་འགྲུབ་ཚེད་མཉེན་ཆས་མ་ལག་འདི་ལ་གུང་སྒྲིག་བྱ་རིམ་མམ་གུང་སྒྲིག་ལག་ཆམིང་བཏུགས་པ་རེད།
PTMDI	མཐོ་རིམ་བྱ་རིམ་གྱི་སྐད་བརྡ་དེ་སྡོགས་བསྒྲིགས་སྐད་བརྡའམ་འཕྲུལ་ཆས་སྐད་བརྡ་ལ་བསྒྱུར་བའི་བྱ་བ་དེར་རྩོམ་སྒྲིག་ཏུ་འགོད་པ་དང་། ལོ་རྩྭའི་ལས་དོན་འདི་ལེགས་འགྲུབ་ཐུབ་པའི་མཉེན་ཆས་མ་ལག་འདི་ལ་རྩོམ་སྒྲིག་བྱ་རིམ་མམ་རྩོམ་སྒྲིག་ཆས་ཞེས་འགོད་པ།

5 总结

本文受到双语交流中混和语言能有效增进交流这一现象启发，利用多个领域的藏汉双语词典和百万句子级别的藏汉单语数据，以 BART 风格降噪自编码为训练目标，通过在单语数据中有效注入词典，进行藏汉跨语言语言模型的预训练，并在已有藏汉平行数据上进行微调。经过广泛的实验验证，本文中的方法比 BART 强基准模型在测试集上的 BLUE 值在藏汉和汉藏方向上分别提高 2.3 和 2.1。结合利用更大规模的单语数据，更加准确有效的词典注入方式，混合 BART 和词典注入的训练方法，应该可以更进一步提高藏汉翻译的性能，我们将该设想在未来的工作中进行研究和探索。此外，本文方法能为后续一到多、多到一、多到多等藏文多语言翻译课题提供可靠的研究基础。

参考文献

Conneau Alexis and Lample Guillaume. 2019. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, N. Gomez Aidan, Kaiser Lukasz, and Polosukhin Illia. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato,

Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Hany Hassan, Anthony Aue, Chang Chen, and Chowdhary. 2018. Achieving human parity on automatic chinese to english new. *ArXiv preprint*, abs/1803.05567.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Armand Joulin, Edouard Grave, and Piotr an Bojanowski. 2016. Fasttext.zip: Compressing text classification models. *ArXiv preprint*, abs/1612.03651.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and A. Noah Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation.

Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online, April. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2021. Domain adaptation for NMT via filtered iterative back-translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 263–271, Kyiv, Ukraine, April. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, and Jingfei an Du. 2019. Roberta: a robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yaron Matras. 2000. Mixed languages: a functional–communicative approach. *Bilingualism: Language and Cognition*, 3(2):79–99, Aug.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. In *Proceedings of NIPS*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model.

头旦才让, 仁青东主, 尼玛扎西, 于永斌, and 邓权芯. 2020. 基于改进字节对编码的汉藏机器翻译研究. 电子科技大学学报, 50(02):249–255+293.

孙义栋, 拥措, and 杨丹. 2022. 基于 volt 的藏汉双向机器翻译. 计算机与现代化, (05):28–32+39.

慈祯嘉措, 桑杰端珠, 孙茂松, 色差甲, and 周毛先. 2019. 融合单语语言模型的藏汉机器翻译方法研究. 中文信息学报, 33(12):61–66.

慈祯嘉措, 桑杰端珠, 孙茂松, 周毛先, and 色差甲. 2020. 基于迭代式回译策略的藏汉机器翻译方法研究. 中文信息学报, 34(11):67–73+83.

李亚超, 熊德意, 张民, 江静, 马宁, and 殷建民. 2017. 藏汉神经网络机器翻译研究. 中文信息学报, 31(06):103–109.

桑杰端珠 and 才让加. 2018. 神经网络藏文分词方法研究. 青海科技, 25:15–21.

沙九, 冯冲, 张天夫, 郭宇航, and 刘芳. 2020. 多策略切分粒度的藏汉双向神经机器翻译研究. 厦门大学学报 (自然科学版), 59(02):213–219.