基于端到端预训练模型的藏文生成式文本摘要

黄硕 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 h17852656271@163.com

闫晓东* 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 yanxd3244@sina.com

欧阳新鹏& 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 2855973230@qq.com

杨金朋 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 15713862215@163.com

摘要

近年来，预训练语言模型受到了广泛的关注，这些模型极大地促进了自然语言处理在不同下游任务中的应用。文本摘要作为自然语言处理中的一个重要分支，可以有效的减少冗余信息，从而提高浏览文本速度。藏文作为低资源语言，缺乏用于大规模的训练语料，藏文生成式文本摘要研究还处于起步阶段，为了解决藏文生成式文本摘要的问题，本文首次提出将端到端的预训练语言模型CMPT（Chinese Minority Pre-Trained Language Model）用于藏文生成式文本摘要研究，CMPT模型通过对其他不同低资源语言文本进行去噪和对比学习，同时为了提高编码器的理解能力，在编码器的输出层增加一个单层掩码语言模型(MLM)解码器，进行Seq2Seq的生成和理解的联合预训练。通过进一步微调可以有效地提高在藏文文本摘要任务上的性能。为了验证模型的性能，我们在自己构建的5w条藏文文本摘要数据集和公开数据集Ti-SUM上进行实验，在两个数据集上的实验表明，我们提出的方法在藏文生成式文本摘要的评测指标上有显著提升。同时，该方法不仅可以应用于藏文文本摘要任务，也可以拓展到其他语言的文本摘要任务中，具有较好的推广价值。

关键词： 预训练语言模型；藏文；文本摘要；CMPT；Seq2Seq

Abstractive Summarization of Tibetan Based on end-to-end Pre-trained Model

Shuo Huang, Xiaodong Yan*, Xinpeng Ouyang&, Jinpeng Yang
Minzu University of China
School of Information Engineering
National Language Resources Monitoring and Research Center on Minority Languages
Language Information Security Research Center
Institute of National Security MUC
h17852656271@163.com, yanxd3244@sina.com, 2855973230@qq.com, 15713862215@163.com

Abstract

In recent years, pre-trained language models have received widespread attention, and these models have greatly facilitated the application of natural language processing in different downstream tasks. Text summarization, as an important branch of natural language processing, can effectively reduce redundant information and improve the speed of text browsing. As a low-resource language, Tibetan lacks large-scale training corpus, and research on Tibetan generative text summarization is still in its infancy. In order to solve the problem of Tibetan generative text summarization, this paper

国家语委项目，多语言网络谣言检测研究，ZDI145-61
基金项目：国家自然科学基金项目（61972436）
* 通讯作者：yanxd3244@sina.com
& 同等贡献

proposes for the first time to use the end-to-end pre-trained language model CMPT (Chinese Minority Pre-Trained Language Model) for Tibetan generative text summarization research. The CMPT model denoises and compares other different low-resource language texts. At the same time, in order to improve the comprehension ability of the encoder, the encoder A single-layer masked language model (MLM) decoder is added to the output layer for joint pre-training of Seq2Seq generation and understanding. The performance on the Tibetan text summarization task can be effectively improved by further fine-tuning. In order to verify the performance of the model, we conducted experiments on the 50,000-entry Tibetan text summary dataset constructed by ourselves and the public dataset Ti-SUM. There is a significant improvement in the evaluation metrics of summaries. At the same time, this method can not only be applied to Tibetan text summarization tasks, but also can be extended to text summarization tasks in other languages, which has good promotion value.

Keywords: Pre-trained language model , Tibetan , Text summarization , CMPT , Seq2Seq

1 引言

随着互联网技术的高速发展和信息化时代的到来，人们面临的信息量呈爆炸式增长，想要高效、快速、准确地获取有价值的信息成为一大难题，文本摘要技术的出现有效缓解了这个问题。作为自然语言处理的一个重要分支，文本摘要技术可以帮助人们从文本中快速、准确获得重要信息,从而提高了人们浏览文本的效率。

根据实现技术的不同，文本摘要一般可以分为两大类：抽取式摘要和生成式摘要。抽取式摘要是指从原始文本中选择最相关的，包含最多信息的，以及最能代表整篇文章的多个句子作为文章的摘要。生成式摘要则是通过理解文本的含义和语法规则，从原始文本中生成全新的概括性文本。由于生成式摘要需要理解和创造语言表达方式，其可靠性和准确性通常比抽取式摘要要低，因此大多数研究都集中在抽取式摘要上(Gambhir and Gupta, 2017)。近年来，随着深度学习技术的不断发展，生成式摘要逐渐成为研究热点。

藏文是作为藏族人民的书面交际工具，是世界公认的成熟的文字之一，在藏语信息化处理的过程中，同样也涉及到自然语言处理的各种任务。因此藏文的文本摘要也是一个值得关注的问题。相对于中英文来说，藏文的文本摘要还面临着许多问题和困难，首先，藏语有丰富的语法和语义特征，如主谓宾语的排列、合成词和分词等，这增加了文本摘要的难度。对于机器学习算法来说，这些语言学特征需要进行更加细致和复杂的处理，以确保正确的理解和提取。其次，目前藏语文本摘要缺乏大规模的标注数据集，这使得使用监督学习方法进行文本摘要变得困难，同时缺乏标注数据也使得评估文本摘要算法的性能变得更加困难，最后对基于语义表示的方法存在一定的局限性，现有的模型无法捕捉文本中的全部语义信息，这会影响文本摘要的质量。

本文提出了使用预训练少数民族语言模型(CMPT)(Li et al., 2022)来完成藏文生成式摘要任务，首先，本文使用的CMPT模型采用中英和多种少数民族语言进行预训练，中国少数民族语言具有文化传播的相似性和邻接性特征，模型通过对不同低资源语言文本进行去噪和对比学习的预训练，提高了语言理解能力。为了提高编码器模型的理解能力，该模型参考CPT(Shao et al., 2021)的设置，在编码器输出层增加一个单层掩码语言模型(MLM)(Taylor, 1953)解码器，进行生成和理解的联合训练，在一定程度上克服了语义表示方法存在的局限性。与其他模型相比，该模型能够有效地生成藏文摘要。

本文的主要贡献如下：

1）首次将预训练语言模型用于藏文生成式文本摘要研究，取得了较好的效果，为后续的藏文生成式摘要研究提供了参考；

2）把标题作为文章摘要，并人工对数据集进行校对，构建了5万条藏文文本摘要数据集，解决藏语文本摘要缺乏大规模的标注数据集的问题。

3）训练多个模型完成藏文生成式文本摘要任务，进行结果对比分析。

2 相关工作

预训练语言模型在自然语言处理相关的下游任务上取得巨大进步，本文先梳理预训练语言模型在生成式摘要的研究进展，再介绍藏文生成式文本摘要的发展状况。

受预训练Transformer句子编码器的工作的启发，Zhang等人(Zhang et al., 2019b)提出了HIBERT进行文档编码，以及一种使用未标记数据对其进行预训练，然后将预训练的HIBERT 应用于摘要模型。同年由OpenAI开发的一种基于Transformer架构的预训练语言模型GPT-2(Radford et al., 2019)，其在大规模语料上进行了无监督的预训练，并可以通过微调适应不同的自然语言处理任务。研究人员通过对GPT-2进行微调，将其应用于生成式文本摘要任务取得优异效果。Zhang等人(Zhang et al., 2019a) 提出了一种新颖的基于预训练的编码器-解码器框架，在编码器端使用BERT 将输入序列编码为上下文表示。对于解码器端有两个阶段，在第一阶段，使用基于Transformer 的解码器来生成草稿输出序列。在第二阶段，屏蔽草稿序列的每个单词并将其提供给BERT，然后通过组合输入序列和BERT 生成的草稿表示，然后使用基于Transformer 的解码器来预测每个掩码位置的改进单词，并将该方法应用于文本摘要任务。Song等人(Song et al., 2020)期望通过改进通用单文档摘要的框架来实现生成不同文本重用比例的摘要，提出一个基于Transformer但仅包含解码器的模型来控制生成摘要的复制率，在训练和解码阶段采取了多种策略生成从完全抽取到高度生成度的不同摘要。Google Research开发的一种基于Transformer架构的序列到序列的预训练语言模型PEGASUS(Zhang et al., 2020)，以间歇句生成为预处理目标，为生成式文本摘要定制。Facebook AI在2020年提出了一个新的预训练序列到序列模型的去噪自动编码器BART(Lewis et al., 2019)，通过用任意噪声函数破坏文本来训练的，以及学习一个模型来重构原始文本，被许多研究者应用在文本摘要任务上取得了优异的效果。

由于缺乏大规模的训练语料，目前针对藏语的文本摘要研究多数停留在抽取式方法。安见才让提出基于句子抽取的文本摘要算法，将每个句子的权重分解为特征词权重和句子结构权重，根据权重挑选候选句子，然后进行平滑处理，抽取出一定质量的摘要(安见才让, 2010)。南奎娘若在此基础上又基于不同特征加权，然后根据权重进行度量来实验基于敏感信息的藏文摘要抽取(南奎娘若and 安见才让, 2016)。李维提出了两种藏文文本摘要方法，一种改进TextRank的藏文抽取式摘要生成方法。该方法将外部语料库的信息以词向量的形式融入到TextRank算法，通过TextRank与词向量的结合，把句子中每个词语映射到高维词库形成句向量，进行迭代为句子打分，并选取分值最高的句子重新排序作为文本的摘要(李维et al., 2020)；另一种是一种将抽取式摘要和生成式摘要相结合的藏文摘要生成统一模型，使用双向Bi-GRU神经网络从藏文新闻中提取句子。其次，将指针网络融入到基于注意力的seq2seq模型中生成摘要(Yan et al., 2020)，此方法为藏文生成式摘要任务提供了一个可以参考的基线。李亮通过预训练一个藏文的ALBERT模型完成藏文抽取式文本摘要任务(李亮, 2020)，主要思想是把藏文抽取式任务转化为句子分类任务，验证了预训练语言模型在藏文文本摘要任务上的有效性。

3 模型架构

3.1 模型描述

对于低资源语言，多语言预训练可以比单一语言预训练表现的更好，但是多语言模型对方言和少数民族语言建模是个很困难的问题，尽管如此，哈工大讯飞联合实验室为中国少数民族语言开发了第一个预训练语言模型CINO，该模型提供了藏语、蒙语（回鹘体）、维吾尔语、哈萨克语（阿拉伯体）、朝鲜语、壮语、粤语等少数民族语言与方言的理解能力(Yang et al., 2022)，但是在下游生成任务上表现不尽人意。受CPT工作的启发，将理解和生成任务结合到CMPT模型中，CMPT是一个基于Transformer(Vaswani et al., 2017)，在BART的基础上，加入DeepNorm预训练的超深层生成模型,支持多种语言。它有256个隐藏状态、8个注意力头、128个编码器层和128个解码器层。如图1所示，为了更好的适应理解和生成任务，对Transformer结构做了以下四部分修改:

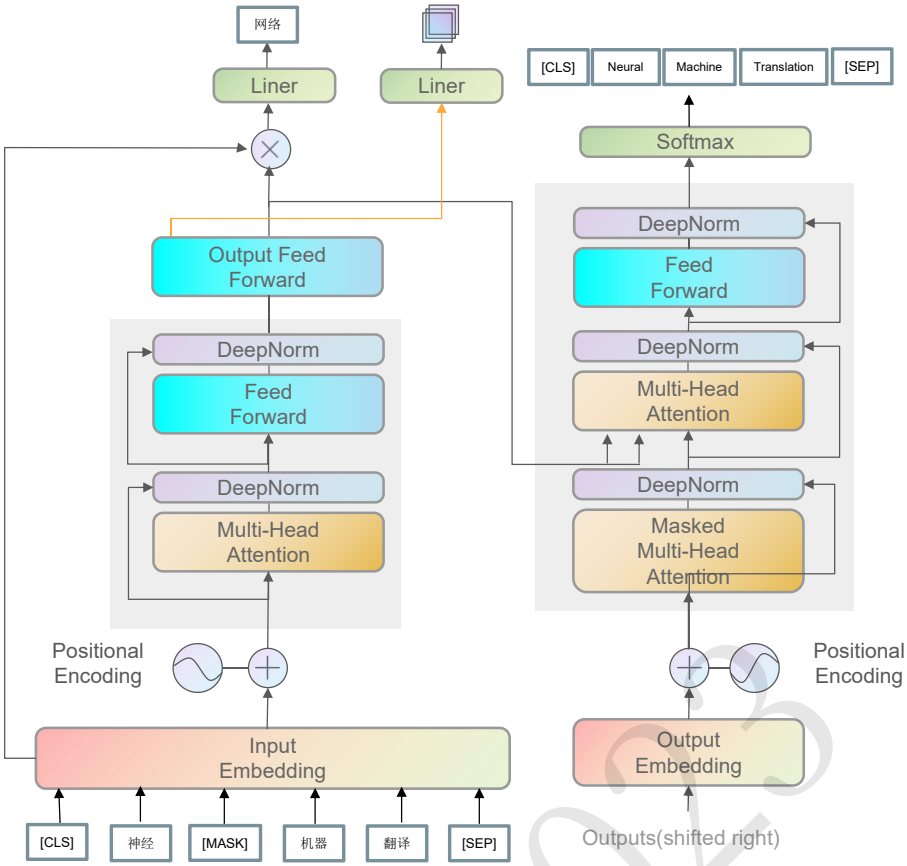

图1.CMPT(Chinese Minority Pre-Trained)语言模型结构图

1）双向编码器:使用双向自注意力编码器，它可以利用语义表示和文本含义。

2）掩码解码器：对双向编码器的输出采用单个线性层，其中输入嵌入(input embedding)乘以输出，被称为MLM头来支持MLM预训练任务的训练。

3）自回归解码器:采用交叉注意实现自回归解码。

4）相似解码器：将编码器的CLS 向量输入到单层相似度解码器中以提取语义向量。

3.2 模型预训练

本文主要研究藏文生成式摘要方法，所以对模型预训练主要介绍和生成任务相关的方法。CMPT为了更好地利用低资源语料库来学习语言知识，采取了四个具有两阶段策略的预训练任务。

1) 掩码语言模型(MLM)任务。以15%的概率随机屏蔽输入文本。并且要求Mask 解码器分别预测掩码标记，以便模型可以学习更深层次的语义信息。输入嵌入（input embedding）与编码器的输出一起用于此MLM 任务。

2) 去噪自动编码(DAE)任务。使用噪声函数来随机破坏输入文本，然后使用掩码填充相应的位置。动机是自回归解码器可以学习重建原始噪声输入。 3) 文本翻译(TT)任务。在第二阶段，将DAE任务更改为监督训练，将多语言翻译对输入到预训练的语言模型中，而MLM任务保持其原始设置。

4) 跨语言对比学习(CCL)任务。在第二阶段，添加相似度解码器来比较和学习相互翻译对的CLS输出，从而缩短具有相同语义的文本之间的向量空间距离。

在预训练阶段，首先使用了Xavier Norm(Glorot and Bengio, 2010)来初始化模型参数，其中E是编码器的层数，D是解码器的层数。

$$\alpha^{Encoder} = 0.81(E^4 \cdot D)^{\frac{1}{16}} \tag{1}$$

$$\alpha^{Decoder} = (3D)^{\frac{1}{4}} \tag{2}$$

$$\beta^{Encoder} = 0.87(E^4 \cdot D)^{-\frac{1}{16}} \tag{3}$$

$$\beta^{Decoder} = (12D)^{-\frac{1}{4}} \tag{4}$$

参考DeepNet(Wang et al., 2022)设置，模型为标准参数归一化设置α和β值:

$$std_{Encoder} = \beta^{Encoder} \times \sqrt{\frac{2}{fan\_in + fan\_out}} \tag{5}$$

$$std_{Decoder} = \beta^{Decoder} \times \sqrt{\frac{2}{fan\_in + fan\_out}} \tag{6}$$

$$W_{Encoder} \sim N(0, std_{Encoder}) \tag{7}$$

$$W_{Decoder} \sim N(0, std_{Decoder}) \tag{8}$$

其中fan_in是输入网络连接的数量，fan_out是该层输出网络连接的数量。

模型为每一层在LayerNorm中添加了残差结构。

$$Layer^{Output}_{Encoder} = LyerNorm(x \times \alpha^{Encoder} + f(x)) \tag{9}$$

$$Layer^{Output}_{Decoder} = LyerNorm(x \times \alpha^{Decoder} + f(x)) \tag{10}$$

首先使用编码器将句子编码为特征矩阵H，$H \in R^{x \times d \times t}$，然后将其输入到三个不同的解码器层。生成任务主要是应用自回归解码器，模型采用交叉注意力机制将特征矩阵H融合到自回归编码器中，注意力函数可以描述为查询(Q) 和一组键值对(K-V)映射到输出，其中Q、K、V和输出都是向量。输出可以通过值的加权和而计算得出。这些Q、K、V 之间的计算如下所示:

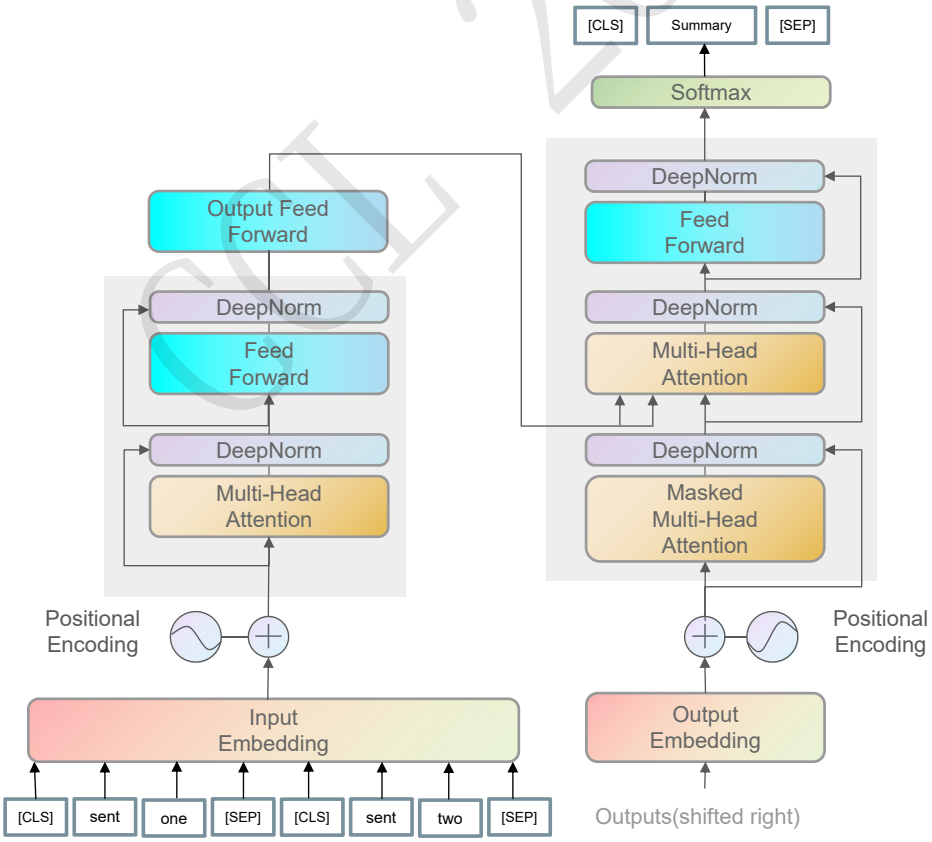

图2.用于生成式文本摘要任务的CMPT模型图结构图

$$\dot{H}_D^t = MultiHead\_Self\ Att(H_D^t\ ) \tag{11}$$

$$\ddot{H}_D^t = MultiHeadAt(\dot{H}_D^t, H, H) \tag{12}$$

$$H_D^{t+1} = LyerNorm(H_D^t \times \alpha^{Decoder} + \ddot{H}_D^t)) \tag{13}$$

其中t表示当前时间，并且整个计算被实现为进一步自回归的递归过程。CMPT模型在超过10G的汉英维藏蒙语料中进行受限预训练，最终的模型大小是390M，具有较强的理解与生成性能。

3.3 摘要任务微调

CMPT模型在翻译任务上的表现，说明预训练的CMPT模型有较好的语义理解和生成性能，我们把该模型用在藏文文本生成式摘要任务上，如图2所示，并在下游摘要任务微调阶段，使用CMPT模型的权重进行初始化，选用自回归解码器作为CMPT模型的解码器，把数据集按照8:1:1的比例划分为训练集，验证集以及测试集，为了更好的适应藏文，我们依旧选用CINO的词表，使用交叉熵作为损失函数进行实验，相比于其他方法，我们的实验结果有显著的提升。

4 实验

4.1 数据集

本文使用Python爬虫工具从香格里拉藏文网、新华网、人民网藏文版等多家新闻媒体网站上爬取58642篇藏语新闻文本作为文本摘要的原始语料，我们采用新闻标题作为参考摘要，剔除过长或过短的新闻原文以及过短的新闻标题，对于新闻标题，我们又进行了人工校验，剔除了标题较为抽象的新闻篇幅，对于藏文原始文本进行了剔除HTML标签以及标点符号过滤等数据清洗操作，最终保留了51221条语料作为实验的数据集。

4.2 超参数设置

表1展示了实验阶段的超参数设置情况。

Parameter	Value
batch_size	8
epochs	10
learning_rate	1e-04
warmup_steps	500
weight_decay	0.001
max_input_length	1024
max_targe_length	128
vocab_size	135259

表1.超参数设置

4.3 评测方法

文本摘要的评价方法分为两种：人工评价方法和自动评价方法。人工评价就是由专家对生成的摘要进行评价，但是评价成本高不利于大规模语料评测，另外人工评价带有主观性容易受外界因素干扰。自动评价是比较模型生成的摘要和参考摘要的相似度。目前，Lin等人参考机器翻译自动评测方法Bleu(Papineni et al., 2002)，提出了ROUGE(Recall-Oriented Understudy for Gisting Evaluation)评测方法(Lin, 2004)，其基本思想是通过将由一系列算法或技术自动生成的摘要或翻译与一组通常由人工生成的理想摘要或翻译进行比对，通过对两者之间的重叠单元（n元语法，单词序列和单词对）进行计数，从而得出分值，以衡量自动生成的摘要或翻译与

参考文本之间的相似性，来评价算法有效性。ROUGE系列评价指标包括ROUGE-N、ROUGE-L、ROUGE-S、ROUGE-W。最常见的评价指标是ROUGE-N,它基于n-gram共现统计,n的范围是从1到4。计算如公式(14)所示：

$$ROUGE-N=\frac{\sum_{S\in\{Refsummaries\}}\sum_{n-grams\in S}Count_{match}(n-gram)}{\sum_{S\in\{Refsummaries\}}\sum_{n-grams\in S}Count(n-gram)} \tag{14}$$

其中$Refsummaries$表示引用摘要，$Count(n-gram)$表示引用摘要中的个数，$Count_{match}(n-gram)$表示生成的摘要和引用摘要中的公用个数。

ROUGE-L是基于最长公共子串的统计，ROUGE-W所做的工作就是给连续的匹配给到更多的权重，让连续匹配的比非连续匹配的有更高的分数。ROUGE-S是ROUGE-N的一种扩展，N-gram是连续的，Skip-bigram是允许跳过中间的某些词，同时结合了ROUGE-L的计算方式。不同的方法对不同类型的总结评价有不同的影响。

4.4 实验结果和分析

本节使用CMPT预训练模型在下游藏文生成式文本摘要任务上微调，分别在我们构建的数据集和公开的Ti-SUM数据集(闫晓东et al., 2022)上进行了实验。因为目前除CMPT模型以外，还没有其他的预训练语言模型可以做藏语的生成式任务，所以我们选择了同样使用标题作为摘要且数据量为5W条的基于统一模型的藏文新闻摘要方法作为基线进行了实验结果对比，对比结果如表2所示：

	ROUGE-1	**ROUGE-2**	**ROUGE-L**
统一模型	19.81	13.27	16.90
CMPT模型（本文数据集）	49.16	33.43	48.66
CMPT模型（Ti-SUM数据集）	39.53	26.42	38.02

表2.不同模型和数据集的实验结果

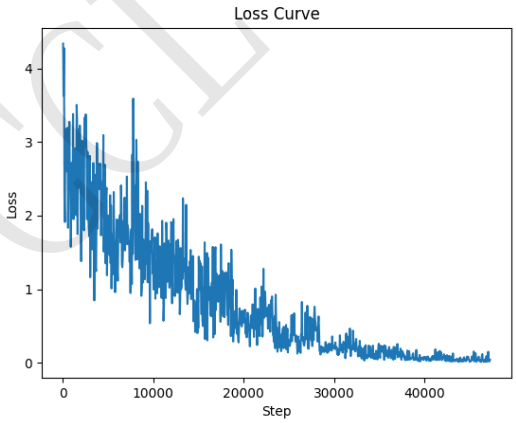

图3.CMPT模型在微调训练过程中的Loss曲线

通过实验结果分析，我们得出如下结论：

1）使用CMPT模型的方法比统一模型方法取得了更好的性能，ROUGE评分分别提高了29.35、20.16、31.76，我们认为CMPT能够取得如此优异的表现主要取决于模型对不同低资源语言文本进行去噪和对比学习的预训练，并在编码器输出层增加一个单层掩码语言模型解码器，进行生成和理解的联合训练，使得CMPT模型具有强大的语义理解和文本生成能力。

2）在Ti-SUM数据集CMPT模型的评测结果略有下降，相较于本文自己构建的数据集ROUGE评分分别下降了24.21%、27.48%、25.53%。对于这样的结果，我们分析认为可能

是两方面原因导致的，首先，Ti-SUM数据集的特点是摘要通常为两到三句，我们参数在设置的时候生成最大长度为128，这样的设置会影响到ROUGE评分的结果。其次，Ti-SUM数据集的数据量只有1000条，微调数据量太小可能会导致模型过拟合Ti-SUM数据集，而忽略了预训练过程中所学到的更广泛的知识，我们将会在4.5节具体展开分析不同数据量对模型的影响情况。

3）我们发现，在使用本文自己构建的数据集训练过程中，CMPT模型的Loss曲线前期震荡幅度较大，我们分析可能有以下原因导致，首先是学习率调整的问题，微调大型模型时前期学习率大，模型参数更新幅度大，使得loss的值波动比较大。其次当从预训练模型切换到微调阶段时，输入数据的分布通常会发生变化，这可能包括数据集的不同领域、任务的不同类别等，模型需要适应新的数据分布，因此可能会导致前期loss的不稳定性。最后模型参数的初始值也会对训练过程和loss 曲线的动荡性产生影响。

4.5 分析数据量对模型的影响

为了分析不同数量的数据对模型的影响，我们把原始的数据集随机抽取成6份（1000/5000/10000/20000/30000/40000），然后进行实验，为了避免随机抽样的影响，我们用不同的样本重复了每个实验5次，并报告了它们的平均结果，如图4-5所示。 根据这组数据，

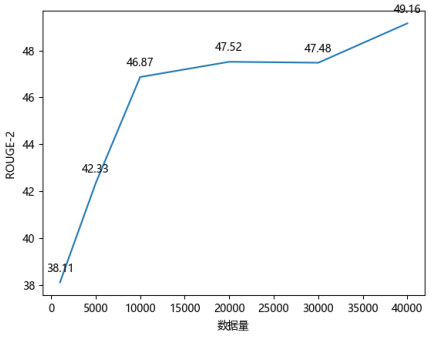

图4.不同数据量的ROUGE-1得分

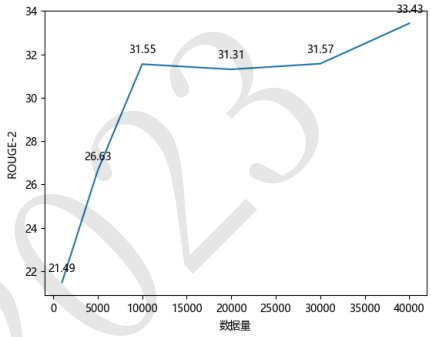

图5.不同数据量的ROUGE-2得分

可以发现随着数据量的增加，ROUGE得分有所提高，这可能是因为模型在更大的数据集上进行训练，获得更多的语言特征，从而提高了文本摘要的准确性。但是在1000-10000数据量时，ROUGE值大幅度增加，这可能是因为更大的数据集提供了更多的数据多样性，使得模型能够更好地捕捉到藏文生成式文本摘要任务中存在的一些规律和特征，从而提高了模型的性能，在这之后，随着训练数据量的增加，ROUGE得分逐渐趋于平缓，这说明此时模型的训练数据的规模已经不再是限制模型的重要因素，想要继续提高模型的性能，可以从其他因素继续研究。

4.6 分析数据长度对模型的影响

此外，本文进一步探究了文本长度对模型生成摘要效果的影响，我们挑选了文本长度在500-800、1000-1300、1500-1800和2000-2500的数据各2000条，重新对模型进行训练，同样，为了避免随机抽样的影响，我们用不同的样本重复了每个实验3次,实验结果取平均，实验结果如表3所示。 实验结果表明，文本的长度也是影响模型生成效果的一个重要因素。

	ROUGE-1	ROUGE-2	ROUGE-L
500-800	41.98	25.47	40.68
1000-1300	34.73	16.23	32.65
1500-1800	31.19	16.63	31.12
2000-2500	32.20	16.59	30.79

表3.不同数据长度的的实验结果

从表中可以看出，当文本长度小于1024时，ROUGE-1值稳定在42左右，而当文本长度大于1024时，ROUGE得分开始降低，模型生成摘要的效果明显变差。我们认为文本长度影响模型生成效果的主要原因是模型最大能编码的序列长度为1024，在对长文本进行生成时，模型无法编码完整的文本数据，会有一定的重要信息丢失，模型不能捕获到重要信息，导致生成效果不佳。在文本长度大于1024时，随着文本长度的继续增加，ROUGE评分波动幅度不大，这可能是因为我们用来构建数据集的文本主要为新闻文本，然而新闻文本的重要信息主要集中在文章的开头部分，为了验证我们的猜想，首先我们选取了5000条藏文新闻文本数据，并对文章分句编号处理，然后使用贪婪的方法把文章的每个句子和标题做ROUGE计算，我们把ROUGE得分作为评价文中句子重要程度的指标，并把ROUGE得分排序，返回句子的编号，结果如图6所示。接下来我们会继续增加数据集文本的类别，提高数据集的质量，推动藏文生成式文本摘要的研究。

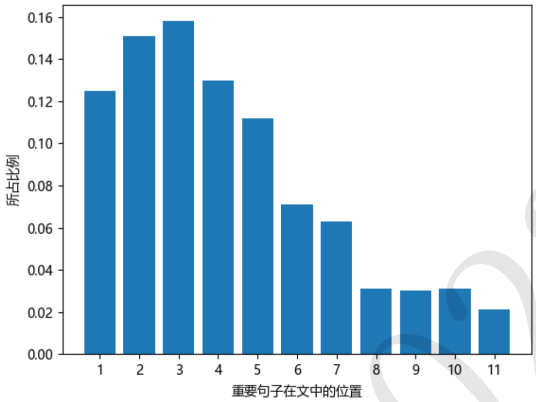

图6.重要句子在文中的位置所占比例

原文：
ཞིན་ཧྭ་གསར་འགྱུར་ཁང་གི་གློག་འཕྲིན་ལྟར།ཚེས7ཉིན་གྱི་ཕྱི་དྲོའི་ཆུ་ཚོད3པར།རྒྱལ་ཡོངས་དམངས་ཆེན་སྐབས་བཅུ་བཞི་པའི་གྲོས་ཚོགས་ཐེངས་དང་པོས་མི་དམངས་ཚོགས་ཁང་ཆེན་མོ་ནས་ཚང་འཛོམས་གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་ཤིང་།རྒྱལ་ཡོངས་དམངས་ཆེན་རྒྱུན་ཨུ་ལྷན་ཚོགས་ཀྱི་ཨུ་ཡོན་ཀྲང་ལི་ཀྲན་ཧྲུའུ་ཡིས་ཕུལ་བའི་རྒྱལ་ཡོངས་མི་དམངས་འདུས་མི་ཚོགས་ཆེན་རྒྱུན་ལས་ཨུ་ཡོན་ལྷན་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན་པ་དང་།ཆེས་མཐོ་མི་དམངས་ཁྲིམས་ཁང་གི་ཡོན་ཀྲང་ཀྲོའུ་ཆང་གི་ཆེས་མཐོ་མི་དམངས་ཁྲིམས་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན་ལ།ཆེས་མཐོ་མི་དམངས་ཞིབ་དཔྱོད་ཁང་གི་ཞིབ་དཔྱོད་དཔོན་ཀྲང་ཇུན་གྱི་ཆེས་མཐོ་མི་དམངས་ཞིབ་དཔྱོད་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན།རྒྱལ་སྲིད་ཨུ་ཡོན་ནས་རྒྱལ་སྲིད་སྲི་ཁྲབ་ཁང་གི་དྲུང་ཡིག་ཆེན་མོ་ཞའོ་ཅེ་ཡི་རྒྱལ་སྲིད་སྲི་ཁྲབ་ཁང་གི་སྒྲིག་གཞི་བཅོས་བསྒྱུར་ཇུས་གཞི་སྐོར་གྱི་གསལ་བཤད་ལ་མཉན།
据新华社电。7日下午3时。***十四届全国人大一次会议在人民大会堂举行第二次全体会议***。听取了全国人大常委会委员长栗战书关于全国人大常委会工作的报告。听取最高人民法院院长周强关于最高人民法院工作的汇报。听取最高人民检察院检察长张军关于最高人民检察院工作的报告。听取国务委员、国务院秘书长肖捷关于国务院机构改革方案的说明。

参考摘要：
རྒྱལ་ཡོངས་དམངས་ཆེན་སྐབས་བཅུ་བཞི་པའི་གྲོས་ཚོགས་ཐེངས་དང་པོའི་ཚང་འཛོམས་གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་པ། །གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་པ།
十四届全国人大一次会议第二次全体会议举行第二次会议的召开

生成摘要：
རྒྱལ་ཚོགས་ཆེན་སྐབས་བཅུ་བཞི་པའི་གྲོས་ཚོགས་ཐེངས་དང་པོས་ཚང་འཛོམས་གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་པ།
十四届***国会***一次会议举行第二次全体会议

图7.生成摘要实例

4.7 摘要生成实例分析

从图7生成的摘要结果分析，我们的方法可以理解文本语义信息，能够捕捉到文本重要信息的所在位置，生成可读性和准确性较强的摘要，这主要归功于模型进行了生成和理解的联合预训练，又通过足够多的语料进行了监督训练，使得模型更好地学习数据的特征，从而提高模型

第二十二届中国计算语言学大会论文集，第113页-第123页，哈尔滨，中国，2023年8月3日至5日。

的泛化能力。另外生成的摘要里存在原文没有出现的词，这可能是因为模型在生成摘要时出现了一些语法或语义的歧义，导致生成的摘要略有瑕疵，接下来我们会针对这个问题对模型的解码端继续改进。

5 总结

文本摘要是自然语言处理的重要分支，在藏语信息化进程中，也需要跟上深度学习发展的步伐。本文提出使用预训练CMPT模型解决藏文生成式文本摘要问题，通过对不同低资源语言文本进行去噪和对比学习的预训练，并在生成式文本摘要任务上进行微调，通过在不同的数据集上的效果表明，生成的摘要具有较强的原文相关性和可读性。但是仍有许多问题亟待解决，首先，使用标题做参考摘要存在部分标题不足以总结全文的问题。其次，ROGUE评测主要基于匹配单词、短语等方式计算，与语法和语义的理解存在局限性。接下来，我们会继续完善藏文文本摘要的数据集，以及改进该模型并将其拓展到其他低资源语言的文本摘要任务中，同时，我们还要预训练一个支持蒙、藏、维三种少数民族语言的T5模型，致力推进中国少数民族语言文本摘要的发展。

参考文献

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022. A multi-tasking and multi-stage chinese minority pre-trained language model. In *Machine Translation: 18th China Conference, CCMT 2022, Lhasa, China, August 6–10, 2022, Revised Selected Papers*, pages 93–105. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8902–8909.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers.

Xiaodong Yan, Xiaoqing Xie, Yu Zou, and Wei Li. 2020. 基于统一模型的藏文新闻摘要(abstractive summarization of tibetan news based on hybrid model). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 479–490.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. *arXiv preprint arXiv:2202.13558*.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019a. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

南奎娘若and 安见才让. 2016. 基于敏感信息的藏文文本摘要提取的研究. 网络安全技术与应用, (4):58–59, 1.

安见才让. 2010. 藏文搜索引擎系统中网页自动摘要的研究. 微处理机, 31(5):77–80, 1.

李亮. 2020. 基于albert 的藏文预训练模型及其应用.

李维, 闫晓东, and 解晓庆. 2020. 基于改进textrank 的藏文抽取式摘要生成. 中文信息学报, 34(9):36–43.

闫晓东, 王羿钦, 黄硕, 杨金朋, and 赵小兵. 2022. 藏文文本摘要数据集. 中国科学数据*(*中英文网络版*)*.