计算机科学
COMPUTER SCIENCE

CINOSUM:面向多民族低资源语言的抽取式摘要模型

翁彧, 罗皓予, 超木日力格, 刘轩, 董俊, 刘征

引用本文

翁彧, 罗皓予, 超木日力格, 刘轩, 董俊, 刘征. CINOSUM:面向多民族低资源语言的抽取式摘要模型[J]. 计算机科学, 2024, 51(7): 296-302.

WENG Yu, LUO Haoyu, Chaomurilige, LIU Xuan , DONG Jun, LIU Zheng. CINOSUM:An Extractive Summarization Model for Low-resource Multi-ethnic Language [J]. Computer Science, 2024, 51(7): 296-302.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100201

CINOSUM:面向多民族低资源语言的抽取式摘要模型

翁 彧[1] 罗皓予[1] 超木日力格[1] 刘 轩[1] 董 俊[1] 刘 征[1,2]
1 中央民族大学民族语言智能分析与安全治理教育部重点实验室 北京 100081
2 中央民族大学中国少数民族语言文学学院 北京 100081
(wengyu@muc.edu.cn)

摘 要 针对现有的模型无法处理多民族低资源语言自动摘要生成的问题,基于 CINO 提出了一种面向多民族低资源语言的抽取式摘要模型 CINOSUM。为扩大文本摘要的语言范围,首先构建了多种民族语言的摘要数据集 MESUM。为解决以往模型在低资源语言上效果不佳的问题,构建了一个框架,采用统一的句子抽取器,以进行不同民族语言的抽取式摘要生成。此外,提出采用多语言数据集的联合训练方法,旨在弥补知识获取上的不足,进而扩展在低资源语言上的应用,显著增强模型的适应性与灵活性。最终,在 MESUM 数据集上开展了广泛的实验研究,实验结果表明 CINOSUM 模型在包括藏语和维吾尔语在内的多民族低资源语言环境中表现卓越,并且在 ROUGE 评价体系下取得了显著的性能提升。

关键词:抽取式摘要;多语言预训练模型;低资源语言信息处理;知识迁移
中图分类号 TP391

CINOSUM:An Extractive Summarization Model for Low-resource Multi-ethnic Language

WENG Yu[1],LUO Haoyu[1],Chaomurilige[1],LIU Xuan[1],DONG Jun[1] and LIU Zheng[1,2]
1 Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance,Ministry of Education,Beijing 100081,China
2 School of Chinese Ethnic Minority Languages and Literatures,Minzu University of China,Beijing 100081,China

Abstract To address the issue of existing models being unable to handle abstractive summarization for low-resource multilingual languages,this paper proposes an extractive summarization model,CINOSUM,based on CINO(a Chinese minority pre-trained language model). We construct a multi-ethnic language summarization dataset,MESUM,to extend the linguistic scope of text summarization. To overcome the poor performance of previous models on low-resource languages,a unified sentence extraction framework is employed for extractive summarization across various ethnic languages. In addition,we introduce a joint training strategy for multilingual datasets that effectively expands applications in low-resource languages,thereby greatly improving the model's adaptability and flexibility. Ultimately,this paper conducts extensive experimental study on the MESUM dataset,and the results reveal that the CINOSUM model demonstrates superior performance in multilingual low-resource linguistic environments,including Tibetan and Uyghur languages,achieving significant improvements in the ROUGE evaluation metric.

Keywords Extractive summarization,Multilingual pre-trained model,Low-resource language processing,Knowledge transfer

1 引言

中国低资源民族语言的文本摘要抽取在自然语言处理领域中的多语言任务中具有重要地位。受限于训练语料的匮乏,目前大多数少数民族语言抽取式文本摘要任务都是基于传统的机器学习算法,如 TextRank[1],TF-IDF[2],或者直接使用文档的前两句作为摘要。随着 BERT[3]等一系列预训练模型的提出,将这些通过大规模数据集训练的预训练模型用于其他 NLP 下游任务的工作层出不穷,在文本摘要抽取任务中也发挥了十分重要的作用。然而,这些方法并未深入探索多语言预训练模型在少数民族语言抽取式文本摘要任务中的应用。

目前,研究人员提出了许多将预训练模型应用于文本摘要抽取任务的方法。BERTSUM[4]中首次提出了使用 BERTSUM 进行抽取式文本摘要的思路。Liu[4]通过在 BERT 上加入一些分类层来进行抽取式的文本摘要,其通过[SEP]和[CLS]将文本分句,让模型学习是否应该选取某个句子作为摘要。Zhang 等[5]提出利用现在非常流行的 ChatGPT 来进行摘要,探讨了情境学习和思维链推理对于提高其性能的有效性。此外,他们还提出了 DiffuSum[6],即一种

到稿日期:2023-11-30 返修日期:2024-03-14
基金项目:国家重点研发计划(2020YFB1406702-3);国家自然科学基金(61772575,62006257)
This work was supported by the National Key R&D Program of China(2020YFB1406702-3) and National Natural Science Foundation of China (61772575,62006257).
通信作者:刘征(liuzheng@muc.edu.cn)

通过扩散模型直接生成所需的摘要句子表示,并根据句子表示匹配提取句子的方法,这种新颖的抽取式摘要方法在英文上得到了目前最佳的效果。

然而,BERT 等其他基于高资源语言训练的预训练模型可能无法准确捕捉到低资源民族语言的特殊语言特征和结构,导致这些模型在低资源民族语言抽取式文本摘要任务中的性能下降。

中国低资源民族语言自动摘要生成的另外一大痛点在于数据严重不足。近年来,随着深度学习模型的广泛运用,RNN(Recurrent Neural Network)[7],Transformer[8] 等序列到序列的模型已经能很好地在英文或者其他高资源语言中完成摘要抽取任务。但这些模型需基于大规模的标注数据集进行训练之后,才能得到较好的效果。中国低资源民族语言自动摘要模型的训练缺乏大规模的标注数据集,因此,如何在小规模的数据集上取得较好的效果,一直都是低资源语言信息处理的重要问题。

针对上述问题,本文提出了面向多民族低资源语言的抽取式摘要模型 CINOSUM。本文的主要贡献如下:

1)针对低资源民族语言缺乏标注数据集的问题,构建了多语言多民族摘要数据集 MESUM(Multi-Ethnic Summarization),并在此数据集上训练了所提模型。

2)采用在大规模无标注的低资源民族语言上预训练好的模型 CINO(A Chinese Minority Pre-trained Language Model)[9] 作为少数民族语言文本编码和理解的主要模型,设计并实现了一系列的摘要抽取层,并将其无缝集成到预训练模型中。这一创新的设计使得 CINOSUM 能够在复杂的多民族语言环境中高效地进行抽取式摘要的生成。它不仅使得信息的获取更为精准,同时也为理解和处理低资源的民族语言信息提供了新的途径。

3)成功地实现了从高资源语言到低资源语言的知识补充和迁移,这一策略的运用不仅揭示了知识迁移和补充在自然语言处理中的重要性,更在实质上提高了模型的处理能力。这一策略的实施,很好地提升了模型的通用性和灵活性,使得模型在面临复杂多变的语言环境时,仍能维持良好的表现。

基于以上工作,本文在公开的数据集如藏语的 Ti_Sum[10]、中文的 LCSTS[11] 等数据集上进行评测,在 ROUGE-L-F1 值上得到了较好的效果。我们公开了模型的代码和使用方法[1)]。

本文第 2 章介绍自动摘要生成和低资源语言摘要的相关研究工作;第 3 章对本文提出的面向多民族低资源语言的抽取式摘要模型进行详细说明;第 4 章对本文的实验方法、MESUM 数据集的构建方法、实验结果和实验评价进行介绍,并对实验结果进行分析;最后为总结全文并展望未来。

2　相关工作

2.1　抽取式文本摘要

抽取式摘要是自动文摘的一种方法,其主要思想是从原文中直接选择一些句子或段落组成摘要。在中文抽取式摘要方面,Hou 等将文本中主题下重要的一些关键词语的信息与文本语义信息综合起来实现对摘要的引导生成[12]。该研究提出了一种新的中文抽取式文本自动摘要方法,综合考虑关键词信息和词序关系。Huang 等[13] 将外部语料库的信息以词向量的形式融入 TextRank 算法,提出了一种新的中文新闻抽取式自动摘要方法。这些研究主要针对的是中文或英文文本摘要,提供了针对这些语言自动摘要方法的研究基础,但对于低资源语言的文本摘要研究提供的帮助较少。

2.2　单种低资源语言的自动文本摘要

藏文自动文本摘要技术是处理大量藏文信息、提取关键信息,生成简洁、完整和连贯的文本摘要的重要方式。这种技术可以分为生成式和抽取式两种主要类型。藏文自动文本摘要的主要步骤包括:分词和预处理、关键信息抽取、内容组织和生成摘要。分词和预处理是将原始文本转化为计算机可以处理的格式的过程,主要包括去除无关信息、分词、词性标注等,在这方面,Chen 等提出的藏文的分词和词性标注[14] 为之后的藏文研究奠定了基础。

2.2.1　单种低资源语言生成文本摘要

生成式自动文本摘要通过深度理解原文本的内容,来生成反映该内容的新的文本。这种类型的摘要技术通常需要理解文本的语义,甚至需要一定的推理能力,以便生成新的、简洁的文本来传达原文的主要意思。Li[15] 提出的基于统一模型的藏文新闻摘要采用两层双向 GRU,通过基于注意力机制的 Seq2Seq 在藏文生成式摘要上得到了不错的结果。之后 Huang 等[16] 提出了基于端到端预训练模型的藏文生成式文本摘要,使用 CMPT(Chinese Minority PreTrained Language Model)预训练模型作为基础,在藏文生成式摘要方面获得了很大的提升。

2.2.2　单种低资源语言抽取式自动摘要

抽取式自动文本摘要则是从原始文本中直接抽取关键的句子或短语,然后将这些信息拼接或稍作修改,形成摘要。在藏文的抽取式自动文本摘要领域,Li 等提出了基于改进的 TextRank 的自动摘要方法[17],在原有的传统方法上取得了较大的提升。

这些方法聚焦于对单一语言的研究,它们在很大程度上依赖于特定语言的资源和语料。尽管大量的研究提供了探索低资源语言处理的思路,但目前尚缺乏有效的多语言摘要方法。

因此,本文提出了一种新颖的多语言自动文本摘要框架,旨在弥补当前研究中的这一空白。本文方法不仅适用于资源充足的语言,也能够有效地应用于低资源语言,如藏语、维吾尔语等。这种方式可以为低资源语言的自动摘要能力提升提供更强大的技术支持。

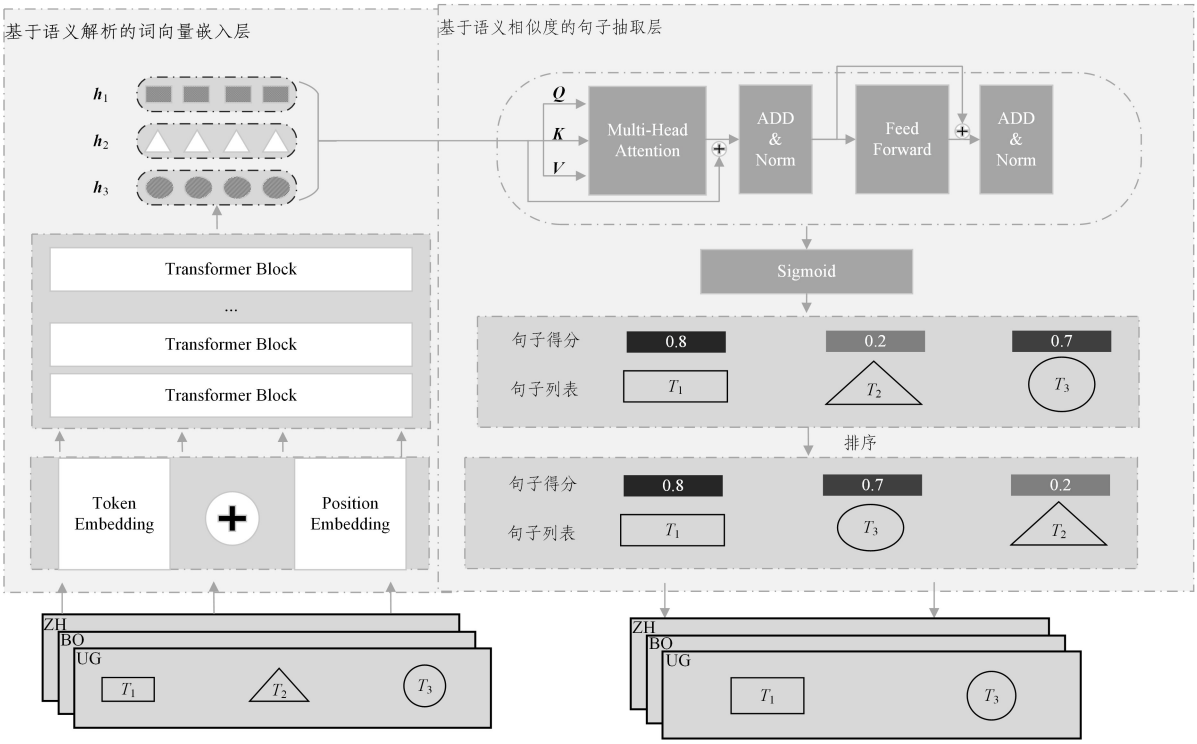
3　多民族低资源语言抽取式摘要模型

本章将详细介绍本文提出的模型,模型的总体结构如图 1 所示。具体地,其可以细分为两个主要组成部分:基于语义解析的词向量嵌入和基于语义相似度的句子抽取层。下文将分别对这两个部分进行详尽的阐述。

此外,本文尝试了两种微调策略:仅微调句子抽取层,

1) https://github.com/5i-wanna-be-the-666/CINOSUM

以及同时微调 CINO 和句子抽取层。实验发现，当数据集较小时，单独微调句子抽取层能快速取得较好的效果，但在多语言环境中的表现不佳。因此，在实验中，本研究主要采用了同时微调 CINO 和句子抽取层的策略。



注：在该架构中，采用 ISO 639-1 标准缩写表示不同的语言代码：ZH 代表汉语，BO 代表藏语，UG 代表维吾尔语。架构图以 3 个文本句子 T_1，T_2，T_3 为例来阐述摘要提取的流程。在经过语义解析以及词向量嵌入模块后，分别得到 T_1，T_2，T_3 对应的向量 $\boldsymbol{h}_1$，$\boldsymbol{h}_2$，$\boldsymbol{h}_3$。在基于语义相似度的句子选择模块中，使用 Transformer 模型作为示例，将 $\boldsymbol{h}_1$，$\boldsymbol{h}_2$，$\boldsymbol{h}_3$ 作为输入，并通过 Sigmoid 激活函数输出各句子的评分。通过评分排序机制，T_1 和 T_3 被挑选出来，构成最终的文本摘要。

图 1 CINOSUM 架构图

Fig. 1 Overall architecture of CINOSUM

3.1 基于语义解析的词向量嵌入层

少数民族语言预训练模型 CINO 在 2022 年首次被提出，它基于 XLM-R[18]进行训练，旨在提供一个覆盖众多少数民族语言的预训练模型。CINO 通过 BERT 提出的 MLM (Mask Language Model)任务，使用多种少数民族语言的文本进行预训练，设定目标函数为：

$$L_{\mathrm{MLM}}=-\sum_{\hat{x}\in m(x)}\log p(\hat{x}\mid x_{\backslash m(x)}) \tag{1}$$

其中，$m(x)$和$x_{\backslash m(x)}$分别表示来自 x 的屏蔽字和剩余字。

本文采用 CINO 预训练模型作为语义解析以及词向量嵌入的模块，文本信息被输入涵盖少数民族语言的 CINO 大模型中，实现了词汇的切分、句子的分割以及词嵌入的操作，以便对文本内容进行处理。

3.2 基于语义相似度的句子抽取层

在从 CINO 模型获取句子向量之后，我们在 CINO 的输出上构建了几个专为摘要提取设定的层级，旨在捕捉文档级别的特征，分别使用了线性分类器(Classifier)、循环神经网络，以及 Transformer 模型作为句子抽取层。

本文将 CINO 模型得到的词向量结果以及句子列表作为句子抽取层的输入，由句子抽取层对句子列表中的每个句子进行打分，使用选择该句子作为摘要的概率作为得分。

1)Classifier。简单分类器仅在 CINO 输出上添加线性层，最后通过 Sigmoid 函数来获取预测分数：

$$\hat{S}_i=\sigma(\boldsymbol{W}_o T_i+b_o) \tag{2}$$

其中 σ 是 Sigmoid 函数；$\hat{S}_i$是输出句子的分数；T_i是输入的句子，用于将输出值限定在 0～1。对每个神经元的输出进行归一化，得到一个概率值，即选择此句作为摘要的概率。Sigmoid 函数公式如式(3)所示：

$$\sigma=\frac{1}{1+\mathrm{e}^{-x}} \tag{3}$$

其中，e 是自然对数的底数，x 是输入值。

2)LSTM。循环神经网络是在文献[7]中首次提出的，它在序列到序列(Seq2Seq)的任务中已经取得了显著的成果。后续的研究[19]在此基础上继续深化，提出的长短期记忆网络(LSTM)弥补了传统 RNN 在处理长距离依赖问题上的缺陷，被成功地应用于长文本处理任务中。

考虑到大部分少数民族语言文本都较长，因此本文选择 LSTM 模型作为第二种句子抽取层。在时间步骤 t，LSTM 层的输入是上一层特征提取网络 CINO 的输出。这个输出是一个特征向量，包含了第 i 个句子在时间步骤 t 的重要信息。经过 LSTM 层处理后，这些特征向量会编码更多的时序关联信息，本文将输出计算表示如下：

$$H_t, C_t=LSTM(V, H_{t-1}, C_{t-1}) \tag{4}$$

$$Score(T_i)=\sigma(\boldsymbol{W}_s\cdot H_t+b_s) \tag{5}$$

其中，$LSTM()$代表了 $LSTM$ 的整个过程，包括遗忘门、

输入门、输出门的计算，以及隐藏状态和记忆细胞状态的更新。$\boldsymbol{V}$ 是输入向量，H_{t-1} 和 C_{t-1} 分别是上一时刻的隐藏状态和记忆细胞状态，H_t 和 C_t 是当前时刻的隐藏状态和记忆细胞状态。最终输出层也是一个 Sigmoid 层，具体公式如式(5)所示，其中，W_s 和 b_s 是权重参数和偏置项，将 LSTM 的输出 H_t 输入激活函数 Sigmoid 中，进行线性变换，得到每个句子 T_i 的评分 $Score(T_i)$。

3)Inter-sentence Transformer。在最近的深度学习研究中，Transformer 在各个领域的表现都极其突出，本文应用的 Inter-sentence Transformer 是一种基于 Transformer 模型架构的深度学习模型，它被特别设计用于理解和处理两个或更多句子之间的关系。这类模型适用于需要理解多个句子间复杂关系的任务，例如问答系统、多轮对话系统、文本摘要等。

本文采用 Inter-sentence Transformer 作为第三种句子抽取层，相比简单的 Classifier 和 LSTM，它能更好地理解文本句子的含义，并且选取合适的文本作为摘要。输出计算表示如下：

$$h^m = DN(h^{m-1} + MHAtt(h^{m-1})) \tag{6}$$

其中，DN 指差异化归一化(Differentiable Normalization)，$MHAtt$ 是多头注意力机制(Multi-head Attention)，$h^0 = PosEmb(\boldsymbol{T})$，$\boldsymbol{T}$ 是由 CINO 输出的句子向量，$PosEmb$ 是将位置嵌入(指示每个句子的位置)添加到 $\boldsymbol{T}$ 的函数，上标 m 表示堆叠层的深度。

在本文的模型架构中，终端输出层被配置为一个 Sigmoid 分类器，它是以概率的形式输出预测结果的典型选择。本研究采取了两层的 Transformer 结构作为句子抽取层进行实验验证。选择该结构是因为其具有基于 Transformer 的并行处理能力和自动捕捉序列中长距离依赖关系的优势。

4　实验

4.1　数据集与评价指标

因为算力不足等原因，本文采用了一定程度的约束措施，删除了长度不符合实验要求的摘要及文本数据，包括长度小于 10 的摘要、长度小于 10 的原文、长度大于 512 的原文句子。在处理爬取的数据时，我们设定了相应的策略以确保所获取的摘要信息准确地对应到标题，防止了文本路径的获取或者网页内错误标签的产生，以此提高数据的质量和实验的准确性。

本文使用了多个语言的数据集，具体信息如下：

1)中文 LCSTS 数据集[11]：在清洗数据集之后，本文使用原本数据集的训练集、测试集、验证集进行融合，再按照 8∶1∶1 的比例划分为训练集、验证集、测试集，清洗过后的训练集包含 162111 条摘要文本对。

2)藏文数据集：从多个藏文网站上爬取的新闻和文本对。本文将藏文数据集按照 8∶1∶1 的比例划分为训练集、验证集和测试集，得到训练集 37985 条数据，测试集 3800 条数据。

3)维吾尔语数据集：从维文网站上爬取的新闻文本对。同样，经过清洗之后，获得了 1 824 条训练数据对和 200 条测试数据对。

4)MESUM 数据集：将中、藏、维 3 种语言的训练集进行随机顺序的混合得到的数据集。

此外，本文使用公开数据集 Ti_Sum[10] 中的 592 条数据作为藏文的测试数据，进一步验证本文模型的泛化能力。

本研究主要采用 ROUGE-L-F1 评价指标来衡量生成摘要与标准摘要之间的符合程度，从而定量评估摘要生成模型的性能。

4.2　基线与实验细节

在对比实验中采用了多种经典的抽取式摘要生成方法作为比较对象，同时针对不同语言独立训练了单语摘要生成模型。对于不同的句子抽取层训练了多语的摘要模型并进行了比较分析，并引入了未经过预训练的 CINO 模型作为多语摘要生成的基准模型。

本文涉及的摘要方法如下：

1)Lead-N：在抽取式文本摘要方法中，Lead-N 是一种非常简单且广泛使用的策略。其中，N 代表摘要中包含的句子数。例如 Lead-2 就是简单地选取文章的前两句作为摘要。

2)TextRank：一种基于图的排序算法，用于从文本中提取关键词和句子，常用于自动摘要和关键词提取。TextRank[1] 算法会根据句子之间的相似性构建一个图模型，图中的节点代表句子，边代表句子间的相似性。然后 TextRank 算法会通过迭代传播节点的权重，直到达到稳定状态，最后选取权重最高(即最重要)的句子作为摘要。本文的 TextRank 基于 python 库 Gensim 中的 Word2Vec 模块，基于各个语言数据集的文本分别训练得到。

3)BERTSUM：使用最简单的 Classifier 作为句子抽取层，使用 Transformers 库中的“bert-base-uncased”作为语义解析的词向量嵌入层。使用 MESUM 的训练集进行训练，然后在每一种语言的测试集上都进行测试。

4)单语摘要模型：使用最简单的 Classifier 作为句子抽取层，分别只在对应的单语言上训练得到的模型，在对应的测试集上进行评测。下文将其简称为 CCS(CINOSUM-Classifier-Single)。

5)多语的摘要模型：分别使用 3 种不同的句子抽取层，并使用混合了多种语言的数据集进行训练，然后在每一种语言的测试集上都进行测试。比如使用 Classifier 作为句子抽取层的多语模型，下文将其简称为 CCC(CINOSUM-Classifier-Combine)。

本文使用 Pytorch 和 Transformers 库中的 CINO：“hfl/CINO-large-v2”版本来实现模型。

本文使用的是 Adam 优化器，学习率 lr 随着时间调整，依据 BERTSUM 的参数设置，学习率的计算式如式(7)所示。所有模型都在 3 个4090 GPU 上进行 50000 个 step 的训练，最先进的 Transformer 多语摘要模型大概耗时 8h。训练的前 10000 步进行预热(warmup)。

$$lr = 0.002 \cdot \min(step^{-0.5}, step * warmup^{-1.5}) \tag{7}$$

本文的 3 种句子抽取层的具体参数设置如下：

1)Classifier：使用 CINO 最后一层输出的隐藏层维度2048 作为输入维度，输出维度为 1，最后进行 Sigmoid 函数操作。

2)LSTM：使用 CINO 最后一层输出的隐藏层维度2048 作为输入维度，经过一层双向的 LSTM，将输出维度为2048

的输出经过残差连接，映射为维度为 1 的输出，最后进行 Sigmoid 函数操作。

3）Inter-sentence Transformer：设置多头数量为 4，FFN 的维度为 512，由两层的 Transformer block 组成。

4.3 实验结果及分析

本文利用了多个数据集进行评测和训练，因此我们将从两个方面来阐述实验结果。首先，本文将对比传统方法、单语模型以及最简单的线性层多语言模型的训练结果。其次，本文将对比不同句子抽取层的多语言模型的训练效果。这两方面的对比和分析将有助于我们更全面地理解模型的性能和特点。

表 1 详细列出了第一部分实验结果，使用 ROUGE-L-F1 作为性能衡量标准。Lead-2 表示将文章前两句作为摘要，TextRank 在分词后训练也选择两句作为摘要。其他自动摘要方法同样倾向于选择两句。本文采用这样的策略是出于对保留原文信息完整性和有效性的考虑，因为数据集中摘要的长度都较短，所以通过精选两个句子能更好地保证摘要的准确性与完整性。

表 1 不同训练集的 CINOSUM 模型与传统算法的 ROUGE-L-F1 值比较

Table 1 ROUGE-L F1 scores comparison of CINOSUM models trained on diverse datasets with conventional algorithms

methods	MESUM			Ti_Sum	AVG
	ZH	BO	UG		
Lead-2	35.12	25.63	24.30	21.81	26.715
TextRank	35.61	26.12	**24.31**	22.33	27.093
BERTSUM	33.17	11.24	19.35	5.16	17.231
CCS	39.21	46.75	21.52	31.10	34.645
CCC	**40.36**	**48.75**	22.54	**39.09**	**37.685**

从表 1 可以得到结论：

1）除了维文之外，在其他所有数据集上 CCS 都取得了优于传统方法的效果。

2）维文的摘要效果对比中依然是传统方法 TextRank 效果最佳，原因可能是维语的数据集相较于中文和藏文的较小，只有 1 824 条，二中文和藏语都有 10 000 条以上，说明在单语言上微调此模型可能需要较大的数据量，在低资源语言上单语的微调方法可能并不适用。

3）基于高资源语言训练的 BERT 模型在 MESUM 训练集和 Ti_Sum 测试集上表现差，其一是因为其预训练的目标语言中并没有中文、藏文和维文，所以不能很好地抓住这 3 种语言的语义特征；其二是其词表中没有这 3 种语言对应的 token，无法有效识别和分词。

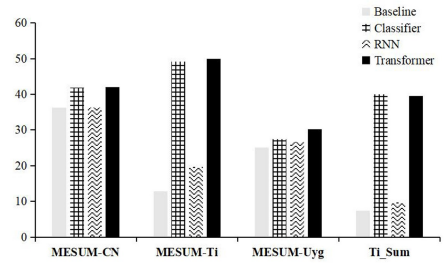
这些结果验证了 CINOSUM 在多语言摘要生成中的有效性。采用 ROUGE-L F1 评测标准，中文、藏文、维文的表现分别提升了 1.15，2.00 和 1.02。在未经训练的公开藏文摘要数据集 Ti_Sum 上，单语模型提升了 7.99，多语模型提升了 16.76，证明了 CINOSUM 的强大泛化能力。

第二部分的实验结果如表 2、图 2 所示，表 2 中的所有模型都是在 MESUM 训练集上进行训练的，表中第一列代表句子抽取层种类，也就是说它们都是多语言的摘要模型。如图 2 所示，本文对各种多语言摘要模型在多个数据集上的性能进行了定量评估。具体而言，采用了 ROUGE-1 和 ROUGE-2 这两种度量标准来衡量模型的性能。

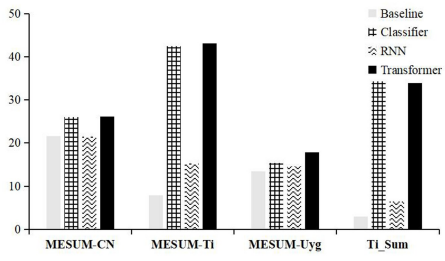
表 2 CINOSUM 不同句子抽取层基于 MESUM 数据集训练后的 ROUGE-L-F1 值

Table 2 ROUGE-L F1 score of CINOSUM model with various sentence extraction layers trained on MESUM dataset

methods	MESUM			Ti_Sum	AVG
	ZH	BO	UG		
Baseline	34.88	12.48	20.33	6.42	18.528
Classifier	40.36	48.75	22.54	**39.09**	37.685
RNN	34.80	19.41	21.78	9.35	21.335
Transformer	**40.56**	**49.44**	**24.52**	37.65	**38.043**



(a) ROUGE-1-F 值



(b) ROUGE-2-F 值

图 2 多种分类体系下多语言的文本摘要评估情况

Fig. 2 Evaluation of multilingual text summarization across various classification systems

其中，Baseline 是使用随机初始化的与 CINO："hfl/CINO-large-v2"版本模型完全相同的模型结构和参数作为语义解析以及词向量嵌入模块，使用 Classifier 作为句子抽取层进行句子抽取的模型。

从表 2 的数据中可以清晰地观察到，具有强大少数民族语言理解能力的预训练模型 CINO 在模型性能提升方面的影响非常显著。Baseline 模型和 Classifier 模型的主要区别在于底层的 CINO 模型是否经过预训练，两者在各种语言处理效果上存在明显差异，特别是在处理藏文数据集的能力上，差距甚至可达 30%。

同时，表 2 的数据还显示：1）句子抽取层为 Transformer 的模型在所有测试中均取得了最佳效果。实验证明，更大的参数量可以显著提升拟合多语言任务的效果。相比使用 Classifier 作为句子抽取层的模型，采用 Transformer 作为句子抽取层的模型在除 Ti_Sum 外的其他数据集 MESUM-ZH，MESUM-BO，MESUM-UG 上，性能提升分别为 0.2，0.69，1.98。2）在多语言环境下，使用 MESUM 训练集对模型进行训练，这一方法在数据稀疏的语言环境中表现出色。具体来说，CINOSUM 在维吾尔语上的性能超过了传统的 TextRank 算法，提升了 0.21 分。这验证了本研究提出的知识

补充及迁移策略的可行性，即通过高资源语言向低资源语言进行知识迁移和补充，可以升高模型在多语种任务中的性能。表 3—表 6 列出了 CINOSUM 在 MESUM 的 3 个数据子集和 Ti_Sum 公开数据集上的摘要示例，涵盖 3 种语言。为了方便阅读，我们为所有少数民族语言的摘要提供了中文翻译，并将少数民族语言中的分隔符替换为逗号。

表 3　Ti_Sum 摘要例子

Table 3　Examples of Ti_Sum abstracts

原文	མི་དམངས་དྲ་བའི་འཕྲིན། གྲོང་མི་མང་པོ་ཞིག་གིས་པོ་བྲང་པོ་ཏ་ལའི་ཐང་ཆེན་དང་ཀློང་རྒྱལ་ཀླུ་ཁང་གྱི་གླིང་གི་གང་སར་སྐྱང་ཞུ་དམར་པོ་བཀལ་ནས་དཔྱིད་གསར་དུས་ཆེན་དང་བོད་ཀྱི་ལོ་གསར་བསུ་བའི་རྣམ་པ་མཐོང་ཐུབ་ཀྱི་ཡོད། ཀློང་རྒྱལ་ཀླུ་ཁང་གྱི་གླིང་དུ་མཐོན་པོའི་དང་བཀལ་བའི་སྐྱང་ཞུ་དམར་པོ་རླུང་ཐུབ་གཡོལ་བཞིན་ཡོད། དེ་དུས་དགུན་དུས་ཀྱང་ལྷང་ལྷང་གི་རྣམ་པ་དེ་དག་སྐྱང་ཞུ་དམར་པོས་མཛེས་རྒྱན་སྤྲས་པའི་རྐྱེན་གྱིས་དྲོད་ལྷང་ལྷང་གི་ཚོར་སྣང་ཞིག་སྐྱེར། གྲོང་མི་མང་པོས་ལག་འཁྱེར་ཁ་པར་བླངས་ནས་དྲན་རྟེན་གྱི་པར་རྒྱག་བཞིན་འདུག་ཚོགས་པའི་སྦྱོང་བརྡར་མཇུག་སྒྲིལ་ཟིན་པའི་པྲིན་སྒྲོན་གྱིས་སྐྱང་ཞུ་དམར་པོའི་འོག་ཏུ་རྩེད་མོ་རྩེ་བཞིན་པའི་ཚ་མོ་ལ་པར་རྒྱག་གི་འདུག, “གང་ས་གང་དུ་སྐྱང་ཞུ་དམར་པོ་བཀལ་འདུག་པ་དེས་སྤྱི་གླིང་ཉིད་མཛེས་པོ་སྤྲས་ཡོད་པས་དུས་ཆེན་གྱི་རྣམ་པ་རབ་ཏུ་རྒྱས་འདུག,” ཅེས་པྲིན་སྒྲོན་གྱིས་དེ་ལྟར་བཤད། པོ་བྲང་པོ་ཏ་ལའི་ཐང་ཆེན་དོ་དམ་ཁྲུའུ་ཡི་ཁྲུའུ་ཀྲང་གཞོན་པ་ཝུའུ་པིན་གྱིས་གསར་འགོད་པར་བཤད་རྒྱུར, ... 人民网讯许多市民可以看到布达拉宫广场和龙王庙公园内到处挂着红色灯笼，迎接春节和藏历新年的景象。龙王庙公园高挂的红灯笼在避风。此刻，那些冬日冷飕飕的景象，因为红灯笼的点缀，而感到暖洋洋的。很多市民拿起手机拍照留念，已经结束早操的边珍，正在红灯笼下给玩耍的孙女拍照。“边珍说。布达拉宫广场管理处副处长吴斌告诉记者……
生成摘要	句子 1： གྲོང་མི་མང་པོ་ཞིག་གིས་པོ་བྲང་པོ་ཏ་ལའི་ཐང་ཆེན་དང་ཀློང་རྒྱལ་ཀླུ་ཁང་གྱི་གླིང་གི་གང་སར་སྐྱང་ཞུ་དམར་པོ་བཀལ་ནས་དཔྱིད་གསར་དུས་ཆེན་དང་བོད་ཀྱི་ལོ་གསར་བསུ་བའི་རྣམ་པ་མཐོང་ཐུབ་ཀྱི་ཡོད 许多市民可以看到布达拉宫广场和龙王庙公园内到处挂着红色灯笼，迎接春节和藏历新年的情景 句子 2： ཐང་པོ་ཏ་ལའི་ཐང་ཆེན་དང་ཀློང་རྒྱལ་ཀླུ་ཁང་གི་ཟུར་ཕྱོགས་སོ་སོར་བཀྲ་ཤིས་པའི་སྐྱང་ཞུ་དམར་པོ་བཀལ་ཡོད་པ་མ་ཟད 在庆和龙王潭的各个角落悬挂着喜庆的红灯笼
标准摘要	གྲོང་མི་མང་པོ་ཞིག་གིས་པོ་བྲང་པོ་ཏ་ལའི་ཐང་ཆེན་དང་ཀློང་རྒྱལ་ཀླུ་ཁང་གྱི་གླིང་གི་གང་སར་སྐྱང་ཞུ་དམར་པོ་བཀལ་ནས་དཔྱིད་གསར་དུས་ཆེན་དང་བོད་ཀྱི་ལོ་གསར་བསུ་བའི་རྣམ་པ་མཐོང་ཐུབ་ཀྱི་འདུག་པ 许多市民可以看到布达拉宫广场和龙王庙公园内到处挂着红色灯笼，迎接春节和藏历新年的景象

表 4　MESUM 的中文摘要例子

Table 4　Examples of ZH abstracts of MESUM

原文	此前，阿里巴巴以 11 亿美元收购了高德地图，而百度也已在地图领域发力。昨日，地图服务商四维图新发布公告称，腾讯 11.73 亿元投资四维，将占股 11.28%。有分析认为，腾讯入股四维图新意味着三巨头直接的 O2O 之战愈演愈烈
生成摘要	句子 1： 昨日，地图服务商四维图新发布公告称，腾讯 11.73 亿元投资四维，将占股 11.28% 句子 2： 有分析认为，腾讯入股四维图新意味着三巨头直接的 O2O 之战愈演愈烈
标准摘要	腾讯 11.73 亿元入股地图服务商四维图新占股 11.28%！

表 5　MESUM 的维吾尔语摘要示例

Table 5　Examples of UG abstracts of MESUM

原文	روشەن مەزگىلىدە دەسلەپكى راكىنىڭ تۇخۇمدان :تېلېگراممىسى بەرگەن بېيجىڭدىن كۈنى-26 ئاينىڭ-5 قانىلىنىڭ ئۇيغۇرچە تورى خەلق ئەھمىيەت تېخىمۇ شۇڭا ،ئاتىلىدۇ دەپ «قاتىل سۈكۈتتىكى» ،بولۇپ تۆۋەن نىسبىتى ياشاش يىل بەش ،قوزغىلىدۇ قايتا ئاسانلا ،بولمايدۇ ئالامىتى كېسەللىك راكىغا تۇخۇمدان يېڭىدىن يىلى ھەر بويىچە دۇنيا ،كۆرسىتىلىشىچە مەلۇماتتا سانلىق قىلغان ئېلان تەشكىلاتى سەھىيە دۇنيا .كېلىدۇ توغرا بېرىشكە بايقالغاندا بىمار 70% بولغان گىرىپتار راكىغا تۇخۇمدان .ئۆلگەن سەۋەبلىك بۇ ئادەم مىڭ 210 يىلى ھەر ،يېتىپ مىڭغا 310 سانى بولغانلارنىڭ گىرىپتار .....يەتمەيدۇ يىلغا بەش نىسبىتى قېلىش ھايات بىمارنىڭ 70% ،قوزغىلىدۇ قايتا ئىچىدە يىل ئۈچ بىماردا 70% ،بولىدۇ قالغان بېرىپ باسقۇچقا ئاخىرقى 新华社北京 5 月 26 日电(记者王宾)卵巢癌早期无明显症状，易复发，5 年生存率低，被称为“沉默的杀手”，更需引起重视。世界卫生组织公布的数据显示，全球每年新发卵巢癌人数达 31 万，每年有 21 万人因此死亡。70%的卵巢癌患者被发现时已是晚期，70%的患者 3 年内复发，70%的生存率不足 5 年……
生成摘要	句子 1： سانى بولغانلارنىڭ گىرىپتار راكىغا تۇخۇمدان يېڭىدىن يىلى ھەر بويىچە دۇنيا ،كۆرسىتىلىشىچە مەلۇماتتا سانلىق قىلغان ئېلان تەشكىلاتى سەھىيە دۇنيا .ئۆلگەن سەۋەبلىك بۇ ئادەم مىڭ 210 يىلى ھەر ،يېتىپ مىڭغا 310 世界卫生组织公布的数据显示，全球每年新发卵巢癌人数达 31 万，每年有 21 万人因此死亡 句子 2： 70% ،قوزغىلىدۇ قايتا ئىچىدە يىل ئۈچ بىماردا 70% ،بولىدۇ قالغان بېرىپ باسقۇچقا ئاخىرقى بايقالغاندا بىمار 70% بولغان گىرىپتار راكىغا تۇخۇمدان .يەتمەيدۇ يىلغا بەش نىسبىتى قېلىش ھايات بىمارنىڭ 70%的卵巢癌患者被发现时已是晚期，70%的患者 3 年内复发，70%的生存率不足 5 年
标准摘要	بولىدۇ قالغان بېرىپ باسقۇچقا ئاخىرقى بىمار 70% ،بايقالغاندا كېسىلى راك بۇنداق 当这样的癌症被发现时，70%的患者已经是晚期了

表 6　MESUM 的藏语摘要例子

Table 6　Examples of BO abstracts of MESUM

原文	སྤྱི་ཟླ་8པའི་ཚེས་3ཉིན་གྱི་ཕྱི་དྲོའི་དུས་ཚོད་16དང་སྐར་མ་30ཡས་མས་སུ, ཡུན་ནན་ཞིང་ཆེན་གྱི་ཀྲའོ་ཐུང་ས་ཁུལ་ལུའུ་ཏན་རྫོང་དུ་ཚད་རིམ་6.5ཅན་གྱི་ས་འགུལ་བྱུང་སྟེ"མང་ཚོགས་ཀྱི་ཚེ་སྲོག་དང་རྒྱུ་ནོར་ལ་གནོད་སྐྱོན་ཚབས་ཆེན་ཐེབས, བསྒྲགས་པའི་གནས་ཚུལ་ལྟར་ནསྤྱི་ཟླ་8་པའི་ཚེས་4ཉིན་གྱི་སྔ་དྲོའི་ཆུ་ཚོད་7པའི་བར་དུ,, ཀྲའོ་ཐུང་ས་ཁུལ་ལུའུ་ཏན་རྫོང་གི་ས་འགུལ་ཁྲོད་དུ་མི་379རྐྱེན་ལམ་དུ་ཤོར་འདུགས་འགྱུལ་བྱུང་རྫེས, ... 8 月 3 日下午 16 时 30 分左右，云南省昭通地区鲁甸县发生 6.5 级地震，给群众生命财产造成严重影响昭通地区鲁甸县地震造成 379 人遇难……

（续表）

生成摘要	句子 1： ཡུན་ནན་ཞིང་ཆེན་གྱི་ཀྱའོ་ཐུང་ས་ཁུལ་ལྟའུ་ཏན་རྩོང་དུ་ཚད་རིམ་6.5ཙན་གྱི་ས་འགྱལ་བྱུང་སྟེ་ མང་ཚོགས་ཀྱི་ཚེ་སྲོག་དང་རྒྱུ་ནོར་ལ་གནོད་སྐྱེན་ཚབས་ཆེན་ཐེབས 云南省昭通地区鲁甸县发生 6.5 级地震，给群众生命财产造成严重损失 句子 2： ཀྱའོ་ཐུང་ས་ཁུལ་ལྟའུ་ཏན་རྩོང་གི་ས་འགྱལ་ཁོང་དུ་མི་379ཀྱེན་ལམ་དུ་ཤོར་འདུགས་འགྱལ་བྱུང་རྟེས 昭通地区鲁甸县地震中有 379 人遇难
标准摘要	ཡུན་ནན་ཞིང་ཆེན་གྱི་ཀྱའོ་ཐུང་ས་ཁུལ་ལྟའུ་ཏན་རྩོང་དུ་ཚད་རིམ་6.5ཙན་གྱི་ས་འགྱལ་བྱུང་། 云南省昭通地区鲁甸县发生 6.5 级地震

结束语 针对现有模型无法处理多民族低资源语言自动摘要生成的问题，本文提出了一种面向多民族低资源语言的抽取式摘要模型 CINOSUM，并构建了多民族语言摘要数据集 MESUM。在模型框架设计中，本文引入了统一分类器，使模型能够适应并处理不同民族语言的抽取式摘要生成任务，从而弥补以往模型在低资源语言上的效能不足。为了解决知识获取不足的问题，我们采用了多语言数据集的联合训练策略，这显著优化了模型的学习效率和深度，同时提高了模型的适用性和灵活性。

实验结果显示，CINOSUM 在 MESUM 数据集和其他公开数据集上表现优异，特别是在处理藏语和维吾尔语等多民族低资源语言的摘要抽取任务时，取得了 ROUGE 指标的显著提升。未来的研究将延续本研究的脉络，重点解决摘要连贯性较差的问题。我们将继续依托 CINO 模型，致力于在中国低资源民族语言领域实现更高效、精准的生成式摘要任务。

参 考 文 献

[1] MIHALCEA R, TARAU P. TextRank: Bringing order into text [C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 404-411.

[2] AIZAWA A. An information-theoretic perspective of tf-idf measures[J]. Information Processing & Management, 2003, 39(1): 45-65.

[3] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[4] LIU Y. Fine-tune BERT for extractive summarization [J]. arXiv: 1903.10318, 2019.

[5] ZHANG H, LIU X, ZHANG J. Extractive summarization via chatgpt for faithful summary generation[J]. arXiv: 2304.04193, 2023.

[6] ZHANG H, LIU X, ZHANG J. Diffusum: Generation enhanced extractive summarization with diffusion[J]. arXiv: 2305.01735, 2023.

[7] MEDSKER L R, JAIN L C. Recurrent neural networks[J]. Design and Applications, 2001, 5(64/65/66/67): 2.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[9] YANG Z, XU Z, CUI Y, et al. Cino: A chinese minority pre-trained language model[J]. arXiv: 2202.13558, 2022.

[10] YAN X D, WANG Y Q, HUANG S, et al. Tibetan Text Summarization Dataset. China Scientific Data: Online English and Chinese Edition, 2022, 7(2): 39-45.

[11] HU B, CHEN Q, ZHU F. LCSTS: A large scale chinese short text summarization dataset[J]. arXiv: 1506.05865, 2015.

[12] HOU L W, HU P, CAO W L. Research on Chinese Generative Automatic Summarization with Topic Keyword Information Fusion[J]. Acta Automatica Sinica, 2019, 45(3): 530-539.

[13] HUANG B, LIU C C. Chinese Automatic Text Summarization Based on Weighted TextRank[J]. Application Research of Computers, 2020, 37(2): 407-410.

[14] CHEN Y Z, LI B L, YU S W. Design and Implementation of Tibetan Word Segmentation System[J]. Journal of Chinese Information Processing, 2003, 17(3): 16-21.

[15] LI W. Research on Tibetan News Summary Generation Based on a Unified Model [D]. Beijing: Minzu University of China, 2021.

[16] HUANG S, YAN X, OUYANG X, et al. Abstractive Summarization of Tibetan Based on end-to-end Pre-trained Model [C]// Proceedings of the 22nd Chinese National Conference on Computational Linguistics. 2023: 113-123.

[17] LI W, YAN X D, XIE X Q. Tibetan Extractive Summary Generation Based on Improved TextRank[J]. Journal of Chinese Information Processing, 2020, 34(9): 36-43.

[18] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised Cross-lingual Representation Learning at Scale[J]. arXiv: 1911.02116, 2019.

[19] ZAYTAR M A, AMRANI C E. Sequence to Sequence Weather Forecasting with Long Short-Term Memory Recurrent Neural Networks [J]. International Journal of Computer Applications, 2016, 143(11): 7-11.

WENG Yu, born in 1980, Ph.D, professor, Ph.D supervisor. His main research interests include machine learning and cloud computing.

LIU Zheng, born in 1990, Ph. D. His main research interests include NLP, data mining, and AI.

（责任编辑：何杨）