专题 2023 多语种智能信息处理数据集

TN-SUM：藏文文本摘要数据集

黄硕[1,2,3]，闫晓东[1,2,3*]，田金超[1,2,3]

1. 中央民族大学信息工程学院，北京 100081
2. 国家语言资源监测与研究少数民族语言中心，北京 100081
3. 国家安全研究院语言信息安全研究中心，北京 100081

摘要： 自动文本摘要是自然语言处理领域的一个重要研究方向，有助于解决信息过载、提高文本数据的可用性和可理解性的一门技术。藏语是中国少数民族语言之一，属于低资源语言，拥有自己独特的文字和语法结构。与中英文这些主要语言相比，藏文在自动文本摘要领域的研究仍然相对滞后，主要原因之一是因为缺乏大规模的可用数据集。为了弥补这一缺失，使用爬虫的方法从各大藏文新闻门户网站抓取了 2 万篇真实藏文新闻，每篇新闻使用标题作为摘要，创建了一个包含丰富、多样的藏文文本摘要数据集 TN-SUM，并寻找了 10 位藏语作为母语的学生对数据进行打分，以此对数据进行质量控制和评估，进而满足科研人员的需要，来推动藏文在自动文本摘要领域的发展。

关键词： 自动文本摘要；数据集；藏文新闻；标题

数据库（集）基本信息简介

数据库（集）名称	TN-SUM：藏文文本摘要数据集
数据通信作者	闫晓东（yanxd3244@sina.com）
数据作者	黄硕，闫晓东，田金超
数据量	2万篇藏文新闻
数据格式	*.csv
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01018
基金项目	国家自然科学基金（61972436）；中央民族大学研究生精品示范课程（GRSCP202316，2023QNYL22）；国家语委重点项目（ZDI145-61）。
数据库（集）组成	数据集是以csv格式存储。第一列是数据的ID；第二列是Title，藏文新闻的标题；第三列是Content，藏文新闻的正文内容。

引 言

自动文本摘要是自然语言处理领域中备受瞩目的研究方向之一，它为信息过载提供了一种强有力的解决方案，有助于从海量文本中提取关键信息，提高了文本数据的可用性和可理解性，实现自动文本摘要也是人工智能走向成熟的一个重要标志[1]。数据集是深度学习模型训练的基础，深度学习模型依赖于大量的数据来学习、优化参数以及提高模型的泛化能力，从而实现各种下游任务。随着大模型时代的到来，数据集的重要性愈发不可替代。高质量的数据集能有效推动自动

www.csdata.org

ISSN 2096-2223

CN 11-6035/N

文献 CSTR：

32001.14.11-6035.csd.2024.0004.zh

文献 DOI：

10.11922/11-6035.csd.2024.0004.zh

数据 DOI：

10.57760/sciencedb.j00001.01018

文献分类：

收稿日期：2024-01-19

开放同评：2024-05-14

录用日期：2024-12-24

发表日期：2024-12-30

* 论文通信作者

闫晓东：yanxd3244@sina.com

文本摘要的研究进展。然而，目前公开的大规模数据集的数量十分稀少，且人工构建成本过大。

在中文方面，清华新闻（THUCNews）数据集[2]是由清华大学自然语言处理实验室根据新浪新闻 RSS 订阅频道 2005 至 2011 年间的历史数据整理而成，包含 80 多万篇新闻文档，利用新闻正文-标题构成摘要数据集。NLPCC-2017 摘要数据集[3]是 2017 年 NLPCC 比赛 Task3 任务的数据集，包含 5 万篇文档。LCSTS 数据集[4]是哈尔滨工业大学整理，基于中国微博网站和新浪微博构建而成的一个大型中文短文本摘要数据集，该数据集由 200 多万篇真实的中文短文本组成，每篇文本作者都给出了简短的摘要。搜狗新闻（SogouCS）数据是搜狗实验室整理的 1 245 835 个样本，同样利用正文-标题构成摘要数据集。在英文方面，DUC2004 数据集[5]是只用于测试摘要文档的数据集，由 500 篇新闻文章组成，每篇文章都配有 4 篇人工摘要。Gigaword 语料库[6]包含 950 万篇左右文章，使用正文-标题构成摘要数据集。CNN/Daily Mail 数据集[7-8]由新闻文章及其对应的人工撰写摘要构成，大约包含 100 万条新闻报道数据。该数据集是一个单文本摘要语料库，在语料库中含有大量的摘要篇章，每个篇章中又包含若干个摘要句子，随后对其进行了简单的修改，形成了一个用于文本摘要技术的语料库。NYTAC 数据集[9]包含 180 万篇文章，其中超过 150 万篇文章由专业人员手工标注，常将其作为抽取式文本摘要工作的数据集。在藏文方面，Ti-SUM 数据集[10]是藏文多文本摘要数据集，由 1000 篇真实藏文新闻组成，每一条新闻都给出了简短的摘要。此外还针对每篇新闻构建了超过 3500 个文章关键词，用以辅助文本摘要任务。

虽然目前藏文已有一个公开的文本摘要数据集[10]，但是过于少量，难以用于深度学习模型的训练和调参过程。本研究构建了一个藏文文本摘要数据集，是由 2 万篇藏文新闻和标题构成（表 1），为研究人员和从业者提供了一个重要的资源，以推动藏文文本摘要领域的发展。

表 1　文本摘要数据集概况

Table 1　Overview of the text summarization datasets

数据集	大小	语言	内容
THUCNews	80 多万篇文档	中文	正文-标题
NLPCC-2017	5 万篇文档	中文	正文-标题
LCSTS	200 多万篇	中文	10666 篇人工标注
SogouCS	1 245 835 个样本	中文	正文-标题
DUC2004	500 篇文档	英文	每篇文章有 4 个摘要
Gigaword	950 万篇左右	英文	正文-标题
CNN/Daily Mail	100 万条	英文	人工标注
NYTAC	180 万篇	英文	150 万篇人工标注
Ti-SUM	1000 篇	藏文	人工标注
TN-SUM	2 万篇	藏文	正文-标题

1　数据采集和处理方法

由于缺乏大规模的藏文文本摘要语料库，且为了避免信息传播中的错误，不能直接使用其他语言的语料进行翻译，因此，采取以下步骤来构建藏文文本摘要数据集。

1.1　数据采集

首先从各大藏文网站进行语料的抓取。这些网站包括人民网藏文版、中国西藏网、香格里拉藏文网、中国藏族网通等多家藏文新闻门户网站，涵盖了不同的藏文新闻文本类型。使用 Python 的 requests 库发送 HTTP 请求，然后使用 BeautifulSoup 库解析 HTML 内容，提取所需的新闻信息，通过网络爬虫工具，获取了 58642 篇藏文原始文本数据，包括新闻标题和正文内容。

1.2　数据清洗

对于抓取到的原始数据，进行数据清理操作。这包括去除 HTML 标签、特殊字符、过滤掉冗余信息，以保留文本的纯净内容。在这一阶段，只保留了新闻标题和正文纯文本，并去除了其他与文本摘要任务无关的信息。

1.3　数据筛选

为确保数据集的质量，对清理后的文本进行筛选。根据设置的特定长度阈值，删除了过长或过短的新闻文本，以确保数据集的一致性和适用性。这有助于减少在后续处理中的噪声和不必要的复杂性。

1.4　摘要审核

爬取到的原始数据中包含很多标题抽象、不能很好地涵盖新闻大意的文章，因此又组织另外一组人员进行摘要质量的把控，对构建的初始摘要进行审核，剔除抽象标题的新闻文章，最终保留了 2 万篇新闻文本作为最终的 TN-SUM 数据集。

2　数据样本描述

本研究构建的藏文文本摘要数据集包含藏文新闻正文以及藏文新闻标题，摘要展示如表 2 所示。

表 2　藏文文本摘要数据集样例

Table 2　Sample of Tibetan text summaries

新闻正文	ས་དེའི་དུས་ཚོད་ཟླ་བཅུ་གཉིས་པའི་ཚེས་དགུ་ཉིན་གྱི་སྔ་དྲོར།རྒྱལ་ཁབ་ཀྱི་ཀྲུའུ་ཞི་ཞི་ཅིན་ཕིང་གིས་ལིས་ཡ་ཏེ་ནས་ཁ་ཐར་གྱི་ཨེ་མུས་ཨར་ཐ་མུས་མུའུ་ལ་མཇལ་འཕྲད་གནང་།ཞི་ཅིན་ཕིང་གིས་ཐོག་མར་ཁ་ཐར་གྱིས་འཛམ་གླིང་བུམ་པའི་རྐང་སྤྲུག་འགྲན་སྡུར་སྒྲིག་པ་ལེགས་འགྲུབ་འབྱུང་རྒྱུའི་སྨོན་འདུན་ཞུས།ཁོང་གིས་བསྟན་དོན།ཁ་ཐར་གྱིས་འཛམ་གླིང་བུམ་པའི་རྐང་སྤྲུག་འགྲན་སྡུར་གཙོ་སྒྲུབ་བྱས་པ་དེས་ཁ་ཐར་གྱི་ནུས་པ་མངོན་ཡོད།སང་ལོ་ནི་ཀྲུང་གོ་དང་ཁ་ཐར་བར་འབྲེལ་བ་བཙུགས་ནས་ལོ་སོ་ལྔ་ལོན་པའི་ལོ་ཡིན་པས།ཀྲུང་གོས་ཁ་ཐར་དང་མཉམ་དུ་དེ་གོ་སྐབས་སུ་བྱས་ཏེ་ཀྲུང་གོ་དང་ཁ་ཐར་གྱི་འཐབ་ཇུས་རོགས་རེས་འབྲེལ་བའི་ནང་དོན་ཕུན་སུམ་ཚོགས་སུ་བཏང་སྟེ།ཀྲུང་གོ་དང་ཁ་ཐར་གྱི་འཐབ་ཇུས་གྲོགས་པོའི་འབྲེལ་བ་ཆུ་ཚད་གསར་པར་གོང་འཕེལ་དུ་གཏོང་འདོད་ཡོད།ཞི་ཅིན་ཕིང་གིས་བསྟན་དོན།ཀྲུང་གོས་ཁ་ཐར་བར་གྱི་ཁྱབ་ཁོངས་སོ་སོའི་མཉམ་ལས་རྒྱ་བསྐྱེད་འདོད་ཡོད།ཀྲུང་གོས་ཁ་ཐར་དང་མཉམ་དུ་ཀྲུང་གོ་དང་ཨ་རབ་རྒྱལ་ཁབ་ཀྱི་འབྲེལ་བ་དང་།ཀྲུང་གོ་དང་མཚོ་ཁུལ་རྒྱལ་ཁབ་ཀྱི་འབྲེལ་བ་གོང་དུ་སྤེལ་འདོད་ཡོད།ཐ་མུས་མུའུ་ཡིས་བཤད་རྒྱུར།ཁ་ཐར་དང་ཀྲུང་གོའི་འབྲེལ་བ་ཧ་ཅང་བཟང་ལ།ཆབ་སྲིད་དང་དཔལ་འབྱོར།རིག་གནས།ལུས་རྩལ་སོགས་ཁྱབ་ཁོངས་ནས་གྲུབ་འབྲས་ལེགས་པོ་བླངས་ཡོད།ཁ་ཐར་གྱིས་ཀྲུང་གོ་གཅིག་གི་སྲིད་ཇུས་ལག་བསྟར་མཐའ་འཁྱོལ་བྱེད་པ་དང་།ཕྱི་ཕྱོགས་སྟོབས་ཤུགས་ཀྱིས་ཀྲུང་གོའི་ནང་སྲིད་ལ་ཐེ་གཏོགས་བྱེད་པར་ངོ་རྒོལ་བྱེད་པ་ཡིན་ཟེར།ཏིན་ཞུའེ་ཞང་དང་།ཁང་དབྱི་ཧྲི་ལེས་ཆིན་བཅས་མཇལ་འཕྲད་ལ་ཞུགས།
	当地时间 12 月 9 日上午，国家主席习近平在利雅得会见卡塔尔埃穆尔塔穆姆。马巴足球赛事的主办充分展示了卡塔尔的潜力，明年是中卡建交三十五周年，中国将与卡塔尔一道，以此为契机，充实中卡战略协作关系的内涵，共同打造中卡战略伙伴、合作习近平指出，中国愿

	扩大同卡塔尔各领域合作，中国愿同卡塔尔共同发展中阿关系、中国同海湾国家关系。中国关系很好，在政治、经济、文化、体育等领域卓有成效。“卡塔尔坚决奉行一个中国政策，反对外部势力干涉中国内政。”
新闻标题	ཞི་ཇིན་ཕིང་གིས་ཀ་ཏ་ར་ཀྱི་ཨེ་རྒྱས་ཨར་ཏ་རྒྱས་སྒྲུབ་ལ་མངོལ་འཇུག་གནང་།
	习近平会见卡塔尔埃穆尔塔穆姆

3 数据质量控制和评估

3.1 数据质量控制

为了确保新闻标题概括新闻内容的准确性，构建了一个打分系统，如图 1 所示。10 位藏语作为母语的学生对数据进行打分，为了减少主观因素的影响，将打分人员分成了 5 组，分别对同一篇文章进行打分，最后得分取其平均值，只保留平均分不少于 6 分的文章和标题作为最后的摘要数据。

图 1　数据评估打分系统

Figure 1　Data evaluation scoring system

3.2 数据质量评估

3.2.1 数据集评估实验

为了评测 TN-SUM 数据集的质量，本研究使用预训练少数民族语言模型（CMPT 模型）在 TN-SUM 数据集和公开的 Ti-SUM 数据集上来完成藏文生成式摘要任务。CMPT 模型是一个基于 Transformer，在 BART（Bidirectional and Auto-Regressive Transformers）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，CMPT 的最终模型大小为 390MB，属于预训练语言模型的基本版本。将本研究构建的 2 万多条生成式文本摘要数据集按 8:1:1 划分为训练集、验证集和测试集，生成式文本摘

要是通过模型生成新的、精炼的文本，以概括或提炼源文本的主要信息，以原始文档的标题作为参考摘要以此为目标来训练模型，使用交叉熵作为损失函数进行实验。实验的超参数设置情况如图 2 所示。

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

图 2　实验的超参数设置

Figure 2　Hyperparameter setting for the experiment

3.2.2　实验结果及分析

本研究使用 CMPT 模型在 TN-SUM 数据集上完成藏文生成式文本摘要任务，并把 CMPT 模型在公开的 Ti-SUM 数据集上的表现作为基线进行实验结果对比，对比结果如表 3 所示：

表 3　CMPT 模型不同数据集上的实验结果（单位：%）

Table 3　Experimental results on different datasets of CMPT models (unit: %)

数据集	ROUGE-1	ROUGE-2	ROUGE-L
TN-SUM 数据集	49.16	33.43	48.66
Ti-SUM 数据集	39.53	26.42	38.02

通过实验结果，分析得知，CMPT 模型在 TN-SUM 数据集上的得分比在 Ti-SUM 数据集上的得分分别提升了 24.36 个百分点、26.53 个百分点、27.99 个百分点，说明 TN-SUM 数据集能较好地满足文本摘要任务的模型训练。

3.2.3　实验结果示例

生成摘要结果的示例如表 4 所示：

表 4　生成摘要示例

Table 4　Example of a summary generated in the dataset

新闻原文	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་འཐུས་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་ཙུན་གྱི་ཆེས་མཐོ་མི་དམངས་ཞིབ་དཔྱོད་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན།རྒྱལ་སྲིད་ཨུ་ཡོན་ནས་རྒྱལ་སྲིད་སྤྱི་ཁྱབ་ཁང་གི་དྲུང་ཡིག་ཆེན་མོ་ཞའོ་ཅེ་ཡི་རྒྱལ་སྲིད་སྤྱི་ཁྱབ་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་ཚོན་སྐབས་བཅུ་བའི་དེའི་ཐོག་ཚོགས་ཁང་ཆེན་གྱི་ནང་ཁྲིད་ཚང་འཛོམས་ཐོག་ཚོགས་ཁང་གཉེར་བ་འཚོགས་པ།
	十四届全国人大一次会议第二次全体会议举行
生成摘要	རྒྱལ་ཚོགས་ཚེན་སྐབས་བཅུ་བའི་དེའི་ཐོག་ཚོགས་ཁང་ཆེ་ནང་ཁྲིད་ཚང་འཛོམས་ཐོག་ཚོགས་ཁང་གཉེར་བ་འཚོགས་པ།
	十四届国会一次会议举行第二次全体会议

4 数据价值

藏文是作为藏族人民的书面交际工具，是世界公认的成熟文字之一，历史之悠久，在国内仅次于中文。我国一直致力于藏文信息技术标准化的研究工作。在信息时代，党和国家领导人的高度重视下，北京、上海、西藏、甘肃、青海等地的一些院校及科研机构纷纷开始了藏文信息处理的研究，推动了藏文信息处理技术的发展，同时取得了较好的成绩[11]。藏文处理技术虽然有一定的研究成果，并且在政策上国家予以了更多的支持，但是在最终的藏文信息处理的过程中仍然存在相应的问题。这些问题制约和影响着藏文信息处理技术的发展[12]。

数据集在深度学习中扮演着关键的角色，它们为模型训练、算法研究、应用开发和社区合作提供了基础。高质量和多样化的数据集是深度学习领域取得成功的基础之一，对于解决各种现实世界的问题具有重要价值，但是藏语一直缺乏开源的大规模数据集，阻碍着藏文自然语言处理前进的步伐。构建的包含高质量藏文文本摘要的数据集，促进了自动文本摘要研究在藏文语境中的发展。使用本数据集可以训练基于深度学习的自动文本摘要模型，从而使模型能够提取并概括原文的关键信息，生成表达原文主要内容的摘要。

致　谢

感谢人民网藏文版、中国西藏网、香格里拉藏文网、中国藏族网通等多家藏文新闻门户网站提供了大量可靠的数据。

数据作者分工职责

黄硕（1998—），男，山东省菏泽市人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据集的预处理和整合、数据采集、论文撰写。

闫晓东（1973—），女，内蒙古自治区赤峰市人，博士，副教授，研究方向为自然语言处理。主要承担工作：数据集质量控制与综合管理。

田金超（2000—），男，安徽省阜阳市人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据校对平台搭建、数据校对。

参考文献

[1] CCF 中文信息技术专委会.文本自动生成研究进展与趋势[C]. CCF2014-2015 中国计算机科学技术发展报告会，北京，中国, 2015. [CCF Chinese Information Technology Committee. Text automatically generates research progress and trends[C].CCF2014-2015 Proceedings of China Computer Science and Technology Development Conference, Beijing, China.]

[2] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. DOI: 10.1109/TASLP.2021.3124365.

[3] QIU X P, GONG J J, HUANG X J. Overview of the NLPCC 2017 shared task: Chinese news headline categorization[C]. National CCF Conference on Natural Language Processing and Chinese Computing. Dalian, China, 2018. DOI: 10.1007/978-3-319-73618-1_85.

[4] HU B T, CHEN Q C, ZHU F Z. LCSTS: a large scale Chinese short text summarization dataset[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015. DOI: 10.18653/v1/d15-1229.

[5] SCHUMANN R, MOU L L, LU Y, et al. Discrete optimization for unsupervised sentence summarization with word-level extraction[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020. DOI: 10.18653/v1/2020.acl-main.452.

[6] NAPOLES C, GORMLEY M, VAN DURME B. Annotated gigaword[J]. Association for Computational Linguistics, 2012, Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction(AKBC-WEKEX): 95–100.

[7] SEE A, LIU P J, MANNING C D. Get to the point: summarization with pointer-generator networks[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017. DOI: 10.18653/v1/p17-1099.

[8] HERMANN K M, KOČISKÝ T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C]. Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. Montreal, Canada. ACM, 2015: 1693–1701. DOI: 10.5555/2969239.2969428.

[9] SANDHAUS E.The New York times annotated corpus overview[EB/OL]. (2021-09-02) [2024-12-27].https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf.

[10] 闫晓东，王羿钦，黄硕，等. 藏文多文本摘要数据集[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-27). DOI: 10.11922/11-6035.csd.2021.0098.zh. [YAN X D, WANG Y Q, HUANG S, et al. A dataset of Tibetan text summarization[J]. China Scientific Data (Chinese and English online version), 2022, 7(2). (2022-06-27). DOI: 10.11922/11-6035.csd.2021.0098.zh.]

[11] 高定国，关白. 回顾藏文信息处理技术的发展[J]. 西藏大学学报(社会科学版), 2009, 24(3): 18–27. DOI: 10.16249/j.cnki.1005-5738.2009.03.022. [GAO D G, GUAN B. Retrospect on the development of Tibetan information processing technology[J]. Journal of Tibet University, 2009, 24(3): 18–27. DOI: 10.16249/j.cnki.1005-5738.2009.03.022.]

[12] 何明华. 当代藏文信息处理的现状与展望[J]. 科技资讯, 2014, 12(23): 249. DOI: 10.3969/j.issn.1672-3791.2014.23.193. [HE M H. Present situation and prospect of contemporary Tibetan information processing[J]. Science & Technology Information, 2014, 12(23): 249. DOI: 10.3969/j.issn.1672-3791.2014.23.193.]

论文引用格式

黄硕, 闫晓东, 田金超. TN-SUM：藏文文本摘要数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-30). DOI: 10.11922/11-6035.csd.2024.0004.zh.

数据引用格式

黄硕, 闫晓东, 田金超. TN-SUM：藏文文本摘要数据集[DS/OL]. V2. Science Data Bank, 2024. (2024-12-30). DOI: 10.57760/sciencedb.j00001.01018.

A dataset of Tibetan text summaries—TN-SUM

HUANG Shuo[1,2,3], YAN Xiaodong[1,2,3*], TIAN Jinchao[1,2,3]

1. School of Information Engineering, Minzu University of China, Beijing 100081, P.R. China

2. National Language Resources Monitoring and Research Center for Minority Languages, Beijing 100081, P.R. China

3. Language Information Security Research Center Institute of National Security MUC, Beijing 100081, P.R. China

*Email: yanxd3244@sina.com

Abstract: Automatic text summarization is an important research direction in the field of natural language processing, a technology that helps address information overload and improve the usability and comprehensibility of textual data. As one of the minority languages in China, Tibetan is a low-resource language, with its own unique script and grammatical structure. Compared with major languages like Chinese and English, research in the field of automatic text summarization for Tibetan still remains relatively underdeveloped, primarily due to the lack of large-scale available datasets. To fill this gap, this study employed a web crawler to collect 20,000 real Tibetan news articles from major Tibetan news portals, using the title of each article as the summary, resulting in a rich and diverse dataset of Tibetan text summaries— TN-SUM. Furthermore, 10 native Tibetan-speaking students were invited to score the data to ensure quality control and evaluation, so as to further meet the needs of researchers and advance the development of Tibetan in the field of automatic text summarization.

Keywords: automatic text summarization; dataset; Tibetan news; title

Dataset Profile

Title	A dataset of Tibetan text summaries—TN-SUM
Data corresponding author	YAN Xiaodong (yanxd3244@sina.com)
Data authors	HUANG Shuo, YAN Xiaodong, TIAN Jinchao
Data volume	20,000 news items in Tibetan
Data format	*.csv
Data service system	<https://doi.org/10.57760/sciencedb.j00001.01018>
Sources of funding	National Nature Science Foundation (61972436); Minzu University of China Foundation (GRSCP202316, 2023QNYL22); Key Research Project of the National Language Commission (ZDI145-61).
Dataset composition	The datasets is stored in CSV format., with he first column representing the data ID; the second column the Title (the news title in Tibetan), the third column the Content (the body content of the news in Tibetan).