专题 多语种智能信息处理

藏文文本摘要数据集

闫晓东[1,2,*]，王羿钦[1,2]，黄硕[1,2]，杨金朋[1,2]，赵小兵[1,2]

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

摘要： 自动文本摘要是自然语言处理中的一个关键任务，高质量的数据集能有效推动摘要的研究。深度学习算法模型在中英开源数据集上都取得了显著的成绩，甚至超过了人类的表现。然而，公开的、高质量的大规模摘要数据集仍然非常稀少，且不容易人工构建。目前在藏文文本摘要任务中，由于公开数据集较少，藏文文本摘要任务还处于起步阶段。为了推动藏文信息化发展，本文人工构建了一个小型藏文多文本摘要数据集 Ti-SUM，由 1000 篇真实藏文新闻组成，每一篇新闻都给出了简短的摘要。此外我们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。

关键词： 文本摘要；数据集；藏文；低资源

数据库（集）基本信息简介

数据库（集）名称	面向文本摘要的藏文数据集 Ti-SUM
数据作者	闫晓东
数据通信作者	闫晓东（yanxiaodong@muc.edu.cn）
数据时间范围	2020–2021
数据量	1000 条摘要
数据格式	*.txt
数据服务系统网址	http://www.doi.org/10.11922/sciencedb.j00001.00352
基金项目	国家语委中心项目（ZDI135-98）
数据库（集）组成	本数据集共包括多个 txt 文件，每个 txt 文件中存放的是已经聚好类的新闻文本以及其对应的人工摘要和新闻关键词信息，txt 文件中 title 后面的内容是人工摘要，keywords 后面的内容是关键词信息。

引 言

文档摘要是一项被广泛研究的自然语言处理任务。随着人工神经网络模型的出现，摘要性能不断提高，对训练数据的要求也越来越高。在一个好的摘要系统应该理解全文，并重新组织信息，以生成连贯、信息丰富且显著简短的摘要，从而传达原文的重要信息[1-2]。大多数传统的生成式摘要方法将过程分为两个阶段[3]。首先，使用监督方法或语言知识从原始文本中提取关键文本元素。然后，通过使用语言规则或语言生成技术，对提取的不清楚的成分进行重写或解释，以生成原始文本的简明摘要。尽管人们对摘要进行了广泛的研究，但摘要的语言质量仍不

ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2021.0098.zh
文献 DOI：
10.11922/11-6035.csd.2021.0098.zh
数据 DOI：
10.11922/sciencedb.j00001.00352
文献分类：信息科学

收稿日期：2021-12-30
开放同评：2022-01-29
录用日期：2022-05-24
发表日期：2022-06-30

* 论文通信作者
闫晓东: yanxiaodong@muc.edu.cn

尽如人意。最近，深度学习方法显示出通过利用 GPU 从大规模数据学习表征[4-5]和生成语言[6-7]的潜在能力。通过深度学习方法生成的摘要更接近于人工书写的摘要。高质量数据集的可用性能有效推动摘要的研究进展。然而，目前公开的、高质量的大规模摘要数据集仍然非常稀少，且不容易人工构建。例如近 10 年流行的英文摘要数据集 DUC 包括来自纽约时报和美联社有线服务的 500 篇新闻文章，每篇文章的参考摘要都由 4 位不同的人书写得到，摘要上限为 75 词，属于小型语料库。CNN/Daily Mail 数据集[8]由新闻文章和人工撰写摘要构成 320 KB 大小的英文单文本摘要数据集，NYTarticles 数据集[9]已广泛用于摘要研究[10-13]，是一个由《纽约时报》策划的文章和图书馆科学家撰写的摘要组成的 100 KB 数据集。Gigaword 语料库[14]，包含 950 万篇左右文章，使用标题作为参考摘要。

流行的中文数据集主要有清华新闻（THUCNews）数据根据新浪新闻 RSS 订阅频道 2005–2011 年间的历史数据筛选过滤生成，利用正文-标题构成摘要数据集，总共包含 830 749 个样本。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。以及 lcsts 摘要数据是哈尔滨工业大学整理，基于新闻媒体在微博上发布的新闻摘要创建，每篇短文约 100 个字符，每篇摘要约 20 个字符[15]。

目前中文开源的文摘数据集大部分都是由“文章-标题”这样的伪摘要语料构成，流行的人工摘要数据集只有数百个，且都是针对短文本数据。就我们所知在藏文文本摘要领域目前还没有公开的数据集，由于缺乏相应的数据集，藏文文本摘要任务还处于起步阶段。为了进一步推动藏文文本摘要的发展，同时为了满足相关研究人员对高质量的藏语文本摘要数据集的需求，本文构建了一个藏文文本摘要数据集，其中包含 1000 篇新闻内容与人工摘要对和超过 3500 个文章关键词（表 1）。

表 1 文本摘要数据集概况

Table 1 Overview of text summarization datasets

数据集	大小	内容	语言
Gigaword	950 万篇文章	文章—标题摘要	英文
THUCNews	80 万个样本	文章—标题摘要	中文
SogouCS	12 万个样本	文章—标题摘要	中文
DUC	500 篇新闻文章	文章—人工摘要	英文
CNN/Daily Mail	320K	文章—人工摘要	英文
NYTarticles	100K	文章—人工摘要	英文
Ti-SUM	1000 条	文章—人工摘要	藏文

1 数据采集和处理方法

由于藏文文本摘要没有规范的语料，并且由于机器翻译的限制，并不能直接将其他语种语料直接翻译，以免造成信息的错误传播。所以首先从各大藏文网站进行语料爬取。对爬取下来的语料需要进行数据清理，过滤掉 HTML 标签以及其他冗余信息，只留下新闻标题以及新闻内容。首先对爬取的原始新闻进行挑选，删除篇幅过长或过短的新闻文本，并对文本内容进行清洗。获取到清洗好的数据集后，将参与构建人员分成两组，一组负责在清洗后的数据集上进行摘要的人工构建，另一组负责验证摘要的质量，对初始摘要进行审核，对低于标准的摘要进行删除或人工复构建操作。

1.1 数据采集方法

从香格里拉藏文网站、人民网藏文版等多家新闻媒体网站上爬取 20000 余篇新闻文本。

1.2 数据预处理

将爬取到的语料进行挑选，删除篇幅过长或过短的新闻文本。数据集清洗时，仅进行了简单的规则清洗，采用正则表达式对新闻文本进行数据清洗，对图像标记，表格等文本、非文本数据进行删除操作。

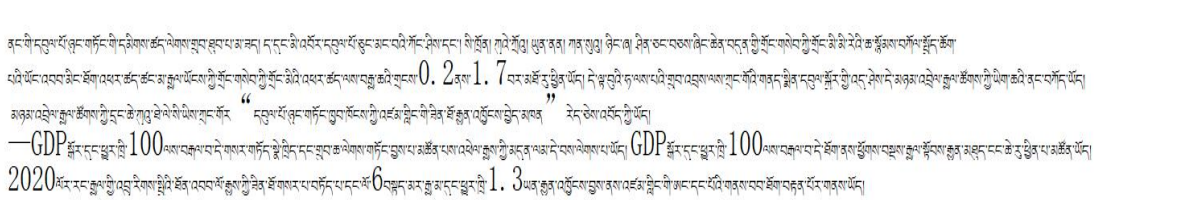
2 数据样本描述

主要包括原始文章、人工摘要和文章主题关键词，摘要展示如表 2 所示。

表 2 藏文文本摘要数据集样样例

Table 2 Sample of Tibetan text summarization datasets

新闻文章

……

……
（到 2020 年，我国国内生产总值（GDP）首次突破 100 万亿元，达到 1015986 亿元。按可比价格计算，比上年增长 2.3%。GDP 突破 100 万亿元，集共同奋斗、共享发展于一体，人人有所得。云南提出，傈僳族，独龙族怒族普米族人口全部脱贫，88 个贫困县全部脱贫。广西要求，荣水苗族自治县等 8 个贫困县摘帽，54 个贫困县摘帽。贵州省宣布，全省已有 9 个贫困县退出贫困县。中国经济上了一个新台阶。经初步核算，2020 年我国国内生产总值（GDP）首次突破 100 万亿元，达到 1015986 亿元。按可比价格计算，比上年增长 2.3%。--GDP 突破 100 万亿元，来之不易，非常罕见。这反映出中央的判断能力，决策能力行为能力。到 2020 年，百年未遇过的非典疫情突如其来，世界经济处于二战结束以来最严重的跌宕起伏。面对突如其来的冲击，我国一季度 GDP 同比下降 6.8%，实现了季度统计数据以来的首次负增长。“只有充分显示我国发展的巨大潜力和强大动力，才能实现今年经济社会发展的目标任务。”
……
不仅完成了家中的减贫目标，在贫困人口较多的广西、四川、贵州、云南、甘肃、宁夏、新疆 7 省农村居民人均可支配收入名义增速均比全国农村居民增速高 0.2 个百分点至 1.7 个百分点。正是这样了不起的成就，中国精准扶贫理念被写入联合国文件。联合国秘书长古特雷斯将中国称为“减贫领域的世界纪录保持者”。--GDP 突破 100 万亿元，反映创新引领和结构优化，发展前景更好。GDP 突破 100 万亿元，标志着综合国力持续增强。2020 年我国粮食

总产量创历史新高，连续 6 年保持在 1.3 万亿斤以上，稳居世界第一……）

人工摘要

ཀྲུང་གོའི་དཔལ་འབྱོར་སྐལ་རེམ་གསར་བ་ཞིག་ཏུ་སྐེབས་པ་ཡིན།

2020ལོར་རང་རྒྱལ་གྱི་རྒྱལ་ནང་ཐོན་སྐྱེད་ཀྱི་རིན་ཐང་བསྡོམས་འབོར(GDP) སྟོང་དང་ཕྱར་ཞུ་100ལས་ཇེངས་དང་ཡོ་བརྒལ་ནས་སྟོང་དང་ཕྱར་1015986ཟེན་ཡོད།

བསྟར་གཞི་ཡོད་པའི་རིན་གོང་ཤུར་ཚེས་ནུ། ལོ་གོང་མ་ལས་2.3%འཕར་ཡོད།

གོ་ལ་ཉེལ་ཡིའི་ནང་དཔལ་འབྱོར་དུང་འཕར་ཐུབ་པའི་དཔལ་འབྱོར་ཁྱེད་ཡོ་གཙོ་ཡོ་འབའ་ཞིག་དེར་འགྱུར་རྒྱ་དང་འཛོམ་སྤྱིང་གི་དཔལ་འབྱོར་ཁྲོད་ཀྱི་ཐོག་སྐལ་ཡང་2019ལོའི་16.3%

ནས་17%ཙམ་བར་འཕར་ནས་ལོ་རྒྱལ་ཀྱི་ཟེན་ཐོ་གསར་པ་བཏོད་ཡོད་པས། ཀྲུང་གོ་གསར་པའི་ལོ་རྒྱལ་ཐོག་སྤྱིར་བཏང་གཏན་ནས་མི་ན་དམངས་སྤྲོ་ཡིད་ཚོམ་པ་དང་།

འཛམ་སྤྱིང་སྐེ་ཡོ་ཀུན་ཀྱིས་དོ་སྣང་ཅེད་པ། ལོ་རྒྱལ་དེབ་ཐོག་འགོད་ཚོག་པའི་རྒྱགས་ལན་ཞིག་ཤུལ་ཡོད།

（中国经济上了一个新台阶。到 2020 年，我国国内生产总值（GDP）首次突破 100 万亿元，达到 1015986 亿元。按可比价格计算，比上年增长 2.3%。全球经济将成为唯一公平增长的经济主体，在世界经济中的份额也从 2019 年的 16.3%上升到 17%左右，创历史新高。新中国历史上极不平凡的一年，人民满意。举世瞩目。交出了载入史册的答卷。）

关键词

ཀྲུང་གོའི་དཔལ་འབྱོར། རྒྱལ་ནང་ཐོན་སྐྱེད།

（中国经济，国内生产）

3 数据质量控制和评估

文摘的撰写由中央民族大学藏语言文学专业学生负责，藏文是他们的母语，又具备本专业文学功底，完全能够胜任摘要撰写工作。基于以下摘要构建要求对文摘进行构建：舍弃与藏文新闻主题无关的内容；简略说明次要材料；摘要紧扣中心，突出新闻重点；顺序结构严谨，摘要层次分明。此外，为了进一步提高数据集的质量，采用交叉验证对构建的摘要进行选择。获取到初始摘要后，对摘要的质量进行验证。验证组分别从语句的流畅程度、语义的完整度以及新闻的覆盖度对初始摘要进行打分，剔除低质量摘要。打分规则如表 3 所示。去除或对平均分数低于 3.5 的摘要进行重写。最终，人工校对出 1000 个新闻和新闻文摘对。

表 3 人工摘要打分规则

Table 3 Grading standards

语句流畅程度	不流畅 1-2	流畅 2-4	流畅，且逻辑关系清晰 4-6
语义完整度	不完整 1-2	语义完整，指代不明 2-4	语义完整，容易理解 4-6
新闻内容覆盖	重点内容覆盖不完全 1-2	全面覆盖重点内容 2-4	按时空逻辑全面覆盖 4-6

4 数据价值

藏文是一种具有一千多年历史的拼音文字，是藏族人们交流思想的工具，是世界公认的成熟的文字之一。信息时代对藏文信息的处理提出了新的课题——用计算机来处理藏文信息。从 20 世纪 80 年代起，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，研制开发了许多藏文信息处理系统，推动了藏文信息处理技术的发展[16]。随着科学技术的快速发展，西藏的研究和建设也进入了快速增长期。同时，由于汉英语言文字信息处理研究技术的

不断迭代和更新，藏文信息处理技术也逐渐从文字信息处理[17]扩展到语言语音信息处理[18]。然而对藏语自然语言的处理还没有大规模的发展。为了藏语能够跟上信息时代社会发展的步伐，更好地满足西藏社会进步和发展的需要，促进西藏社会文明发展。藏语信息化发展已成为一项紧迫的任务。

文本摘要的目的是将原始文档压缩成几个能够涵盖文档主题的短句，通过该技术可以自动化地生成摘要，能有效缓解互联网高速发展带来的信息爆炸和信息冗余的问题。这样，无论是用户还是搜索引擎都能快速通过摘要捕获到原始文本中所包含的主要意思。藏文自动文摘的研究发展缓慢，目前还没有用于训练的大规模藏语文摘语料，且文摘的训练数据构建需要大量的时间和资源，因此最新提出来的一些神经模型只能应用在有限的领域。本研究人工构建的数据集有助于推动藏文文本摘要的发展，满足相关研究人员对高质量的藏语文本摘要数据集的需求。

致 谢

特别感谢香格里拉藏文网站、人民网藏文版，云藏网以及参与本数据集工作的藏语专业人员。

作者分工职责

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。

王羿钦（1998—），女，天津市人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据采集、论文撰写。

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、数据校对、论文撰写。

杨金朋（1997—），男，吉林省（市）通化市人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据采集。

赵小兵（1967—），女，内蒙古自治区呼和浩特市人，博士，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。

参考文献

[1] HOVY E, LIN CY. Automatic text summarization and the Summarist system. Proceedings of the Tipster Text Program, phase III October 1996–October 1998. 1999:197-214.

[2] TAS O, KIYANI F. A survey automatic text summarization[J]. Pressacademia, 2017, 5(1): 205–213. DOI:10.17261/pressacademia.2017.591.

[3] BING L, LI P, LIAO Y, et al. Abstractive multi-document summarization via phrase selection and merging. arXiv preprint arXiv:1506.01597. 2015 Jun 4.

[4] HU B T, LU Z D, LI H, et al. Convolutional neural network architectures for matching natural language sentences[J]. InAdvances in neural information processing systems 2014 (pp. 2042-2050). [5] ZHOU X Q, HU B T, CHEN Q C, et al. Recurrent convolutional neural network for answer selection in community question answering[J]. Neurocomputing, 2018, 274(Jan.24): 8–18. DOI:10.1016/j.neucom.2016.07.082.

[6] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. arXiv: 1409.0473. (2014-09-01). https://www.researchgate.net/profile/Y_Bengio/publication/265252627_Neural_Machine_Translation_by_Jointly_Learning_to_Align_and_Translate/.

[7] SUTSKEVER I , VINYALS O , LE Q V . Sequence to Sequence Learning with Neural Networks[C]// NIPS. MIT Press, 2014. [8] ADELE D. Advances in neural information processing systems 18[J]. Journal of Mathematical Psychology, 2007, 51(5): 339.

[9] MOZZHERINA E . An Approach to Improving the Classification of the New York Times Annotated Corpus[J]. communications in computer & information science, 2013.

[10] SEE A, LIU PJ, MANNING CD. 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). 2017.

[11] GEHRMANN S, DENG Y, RUSH AM. . Bottom-Up Abstractive Summarization[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[12] LEWIS M, LIU Y, GOYAL N, , et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[J]. 2019.

[13] ZHANG J Q, ZHAO Y, SALEH M, et al. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization[C]// 2019..

[14] NAPOLES C, GORMLEY MR, VAN DURME B. Annotated gigaword. InProceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX) 2012 Jun (pp. 95-100).

[15] HU B, CHEN Q, ZHU F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset[J]. Computer Science, 2015: 2667-2671.

[16] 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, GUANBAI. 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.]

[17] 陈夕林. 融合多特征的藏文句子相似度计算方法[D]. 青海师范大学,2021.DOI: 10.27778/d.cnki.gqhzy.2021.000185. [CHEN X L. A Mutil-feature Fusion Method for Tibetan Sentence Similarity Calculation[D]. Qinghai Normal University. 2021.DOI:10.27778/d.cnki.gqhzy.2021.000185.]

[18] 算太本. 基于深度学习的安多藏语语音识别技术研究[D]. 青海师范大学,2021.DOI:10.27778/d.cnki.gqhzy.2021.000270. [SUAN T B，Research on Amdo Tibetan Speech Recognition Technology Based on Deep Learning[D]. Qinghai Normal University. 2021.DOI:10.27778/d.cnki.gqhzy.2021.000270.]

论文引用格式

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.

数据引用格式

闫晓东. 面向文本摘要的藏文数据集[DS/OL]. Science Data Bank, 2022. (2022-06-27). DOI: 11.sciencedb.j00001.00352.

A dataset of Tibetan text summarization

YAN Xiaodong[1,2*], WANG Yiqin[1,2], HUANG Shuo[1,2], YANG Jinpeng[1,2], ZHAO Xiaobing[1,2]

1. School of information and Engineering, Minzu University of China, Beijing 100081, P.R. China
2. National language resource monitoring & Research Center, Minority Languages Branch, Beijing 100081, P.R. China

*Email: yanxiaodong@muc.edu.cn

Abstract: Automatic text summarization is a key task in natural language processing. High-quality datasets can effectively promote the research progress of summarization. Recent research is closer to generate abstractive summarizations by using the deep learning methods. However, there is a lack of high-quality and large-scale summarization datasets available to the public. Besides, it is difficult to construct this kind of dataset manually. The Tibetan text summarization task is still in its infancy due to the lack of public datasets. In order to promote the development of Tibetan informatization. we artificially constructed a small dataset of Tibetan text summarization in this paper, which is composed of 1,000 real Tibetan news articles, each with a short summary. In addition, we have also constructed more than 3,500 article keywords for each news article as a supplement to text summarization tasks.

Keywords: text summarization, dataset, Tibetan, low resource

Dataset Profile

Title	A dataset of Tibetan text summarization
Data corresponding author	YAN Xiaodong (yanxiaodong@muc.edu.cn)
Data author	YAN Xiaodong
Time range	2020-2021
Data volume	1,000 summarizes
Data format	*.txt
Data service system	<http://www.doi.org/10.11922/sciencedb.j00001.00352>
Source of funding	National Language Commission Center Foundation (ZDI135-98)
Dataset composition	This dataset includes a total of multiple “txt” files, each with the news texts that are clustered together with their corresponding artificial summaries and keywords information. The content after “title” in the “txt” file is artificial summarization, and the content after “keywords” is the keywords information.