www.csdata.org

ISSN 2096-2223
CN 11-6035/N

OPEN ACCESS
CC BY

文献 CSTR：
32001.14.11-6035.csd.2024.0038.zh

文献 DOI：
10.11922/11-6035.csd.2024.0038.zh
数据 DOI：
10.57760/sciencedb.j00001.01084

文献分类：信息科学

收稿日期：2024-02-18
开放同评：2024-05-20
录用日期：2024-12-18
发表日期：2024-12-26

* 论文通信作者
刘征：liuzheng@muc.edu.cn

中-蒙-藏-维文多文档摘要数据集

翁彧[1,2]，邢天娇[1,2]，叶旭明[1,2]，刘征[1,2*]，超木日力格[1,2]，刘轩[1,2]

1. 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. 中央民族大学信息工程学院，北京 100081

摘要： 多文档摘要是自然语言处理中的一个关键任务，也是理解和处理大量文档中核心内容的关键技术。尽管目前英文和中文的多文档研究已取得显著进展，低资源语言相关研究却因数据不足而相对滞后。因此，本研究在构建中文多文档摘要数据集的基础上，采取机器翻译加专家校对的方式构建了蒙文、藏文和维文对齐语料，组成了中文、蒙文、藏文和维文版本的面向多文档摘要生成的数据集MMDS。每种语言包含1044个新闻簇（6234篇新闻文章），覆盖2018至2023年间的重大新闻事件。此外，通过人工评价确保了数据集的高质量，使其更加适用于多语言的深度学习模型训练和NLP研究。本数据集的发布，对于促进低资源语言信息处理的发展具有重要的价值。

关键词： 多文档摘要；数据集；蒙文；藏文；维文；低资源

数据库（集）基本信息简介

数据库（集）名称	面向多文档摘要的中-蒙-藏-维文数据集 MMDS
数据通信作者	刘征（liuzheng@muc.edu.cn）
数据作者	刘征、邢天娇、叶旭明、翁彧、超木日力格
数据时间范围	2018–2023年
数据量	307 MB
数据格式	*.xlsx
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01084
基金项目	国家重点研发计划（2020YFB140670002-03）；北京市科技计划（Z231100001723002）
数据库（集）组成	本数据集由中文、蒙文、藏文、维文多文档摘要数据构成，其中每种语言数据之间是高度对齐的，每种语言都包括1044簇新闻，共6234篇新闻文章，每篇新闻存储在txt文件中，由xlsx表格存储新闻事件名称以及对应新闻簇的摘要。不同语言的多文档摘要数据分别存储在各自压缩包中。

引 言

多文档摘要（MDS）是自然语言处理领域的一个关键任务，旨在从多个文档中提取信息，生成一个连贯、准确、信息丰富的摘要[1-2]。随着互联网信息的爆炸式增长，MDS 对于帮助用户快速理解大量文档中的核心内容变得尤为重要。

尽管多文档摘要任务取得了显著进展，但大部分研究依赖于资源丰富的语言数据，如英文和中文，而针对低资源语言的研究则相对较少。到目前为止，多文档摘要研究的数据集主要集中在英文，如 DUC[3]和 TAC[4]系列数据集。这些数据集为英文 MDS 研究提供了标准化的评估平台。CNN/Daily Mail 数据集[5]虽主要用于单文档摘要，但也被应用于多文档摘要研究，提供了大量新闻故事及其紧凑摘要。Demian[6]等人利用维基百科时事门户（WCEP）构建了大规模的多文档摘要数据集，该门户提供简明且中立的人工撰写的新闻事件摘要以及外部源文章的链接，并且通过在 Common Crawl 存档中查找相关文章来自动扩展这些源文章。然而现有的中文多文档摘要数据集数量和质量都有限。例如，清华新闻数据集（THUCNews）和搜狗新闻数据集（SogouCS）主要用于单文档摘要任务，且主要基于新闻标题生成，未能完全满足多文档摘要研究的需求。中国科学院计算技术研究所发布的 Large-scale Chinese Social Media Processing（LCSTS）[7]数据集虽然提供了大量的微博数据用于摘要生成，但其侧重于短文本，与传统的多文档摘要任务在文本长度和结构上存在差异。

本研究针对蒙文、藏文和维文这些低资源语言，通过创新的数据集构建方法，构建了一个多语言多文档摘要数据集。从中文网站爬取大量新闻报道作为源文档，经过严格的筛选和人工摘要编写过程，确保了数据集的高质量和实用性。此外，通过机器翻译和专家校对的方式，将数据集扩展到了蒙文、藏文和维文，满足了低资源语言机器学习模型研究的需求。

本文统计了数据集中文档簇数、文章数、文章包含的最大、最小字符数以及平均每簇所包含的文章数量，由于数据集中的藏语、蒙语、维吾尔语文章数均相同，以下数据描述以其中一种语言为例进行展示，如表 1 所示。

表 1　多文档摘要数据集概况

Table 1　Overview of the multi-document summary dataset

项目	数量
数据集所有文档簇数	1044 簇
文章数	6234 篇
文章最大字符数	31544 个
文章最小字符数	75 个
平均每簇包含文章数	5.97 篇

1　数据采集和处理方法

由于获取低资源语言新闻报道的多文档数据较为困难，不能形成大规模文档簇，所以首先从各大中文网站上爬取多文档新闻，并对获取的语料进行初步筛选和过滤，去除杂质新闻文档，根据新闻事件名称构成多文档簇。由参与人员进行人工摘要的撰写，同时进行原始新闻的第二遍筛选，充分确保原始新闻簇的质量。将构建完成的中文多文档摘要数据集进行机器翻译，形成蒙文、藏文、维文多文档数据集，由各语言专业人员分别对其进行人工数据校对，提高数据可用性。最后由所有参与人员进行数据审核，对于低于标准的数据条目进行删除。

1.1　数据采集方法

从人民网、新华网、中国网、腾讯新闻、新浪新闻等 20 多家新闻媒体网站上共爬取 5300 余篇

新闻文本，初步构成 7700 余簇原始新闻。

1.2 数据预处理

对收集到的新闻多文档簇进行了初步筛选和清洗。首先，使用 K-means 聚类的方法将簇内新闻进一步细分，剔除文档簇中内容不相关的新闻文本和内容高度重复的新闻文本。接着，根据文档长度对初步清洗后的文档簇进行排序，过滤掉字符数少于 70 的文档。这一操作会导致部分文档簇内包含的新闻数量过少，无法满足多文档摘要的训练需求，因此将包含文档数少于 3 的文档簇丢弃，最终保留了 1500 个结构良好的新闻簇。为确保数据的整洁性和标准化，采用正则表达式对新闻文本内容进行清洗，精确地定位和修改特定的文本结构，如特殊符号、格式错误和不规则排版等。

1.3 低资源语言多文档摘要数据集的构建

目前，公开的低资源语言多文档摘要数据集较为稀缺。为填补这一空白，本研究在构建中文多文档摘要数据集的基础上，通过机器翻译生成了与之对齐的蒙文、藏文、维文语料，从而为低资源语言多文档摘要数据集的构建提供了基础数据。为确保新闻领域中专有名词翻译的准确性，本研究提前收集了相关专有名词，将其加入“术语词典”以提升翻译效果。随后，进行了人工修正和校对。将参与构建人员分成三组：一组负责新闻原文数据的内容校对与修改，另一组负责多文档摘要的内容校对与修改，第三组则对数据集的整体质量进行严格审核。参与本数据集的人工校对人员均为中央民族大学少数民族语言文学院的研究生，其中大多数都是低资源语言的母语者，进一步确保了数据集构建的质量和准确性。数据集具体构建过程如图 1 所示。

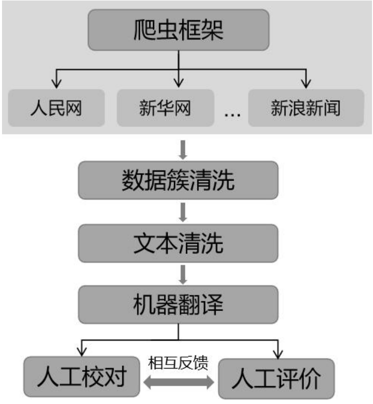

图 1 数据采集和处理过程

Figure 1 Flowchart of data acquisition and processing

2 数据样本描述

数据集中主要包括原始新闻簇、人工摘要和新闻事件名称，以藏文多文档摘要为例，数据展示如表 2 所示。其中，为突出各新闻文章中对摘要生成具有贡献的句子，分别使用黄色、蓝色和绿色进行高亮标注。

表 2　藏文多文档摘要数据集样例

Table 2　Sample of Tibetan multi-document summary datasets

新闻文章 1

ཨ་རིས་རྒྱལ་ཁབ་ཀྱི་སྟོབས་ཤུགས་ལ་བརྟེན་ནས་ཕྱོགས་ཡོངས་ནས་བཀག་སྡོམ་དང་སྐྱེ་བཙིར་བཏང་བའི་འོག ཧྭ་ཝེ་འགྱེལ་མེད་པ་མ་ཟད། ད་དུང་ཧུར་བརྩོན་རྒྱལ་སྐྱོང་གིས་རང་ངོས་ཀྱི་ཞིབ་འཇུག་གསར་སྤེལ་ལ་བརྟེན་ནས་འགྲན་རྩོད་ནུས་པ་དེ་བས་ཆེ་བའི་ཐོན་རྫས་གསར་པ"ཧྭ་ཝེ་Mate60"བཏོན་པ་དེ་ནི"རྡོ་གཤགས་གནམ་འདར"ཅེས་བརྗོད་ཆོག　ཧྭ་ཝེ་ཡིས་དོན་དངོས་ཀྱིས་ཨ་རིའི་གདོང་ལ་གཅར་རྡུང་དྲག་པོ་བྱས་པ་དེས་ཨ་རིས་ཐ་ཤལ་གྱི་བྱེད་ཐབས་སྤྱད་ནས་མཐུག་མཐར་ངལ་བ་འབྲས་མེད་ཡིན་པ་གསལ་བཤད་བྱས། ཧྭ་ཝེ་ནི་ཀུང་སིའི་ཁེ་ལས་ཞིག་ཡིན་མོད། འོན་ཀྱང་ཨ་རིས་ནམ་ཡང་གསོད་མི་ཐུབ་པའི"ཚན་རྩལ་སྟོབས་ཤུགས"ཞིག་ཏུ་གྱུར་ཡོད། ཨ་རིའི་སྲིད་ཚོང་བ་རྣམས་འདིའི་ཆེད་དུ་འདར་དུ་བཅུག

......

དེར་བརྟེན་རེན་གྲེང་ཧྥེ་ནི་སྤྱིར་བཏང་གི་ཁེ་ལས་པ་ཞིག་མིན་པར་འཐབ་ཇུས་ཀྱི་མིག་རྒྱང་ཡོད་པ་དང་། མི་རིགས་ཀྱི་ལས་འགན་ཐུག་ཏུ་འཁུར་ཐོད་པ། ཐུན་མོང་མ་ཡིན་པའི་འགོ་ཁྲིད་ཀྱི་འཛིན་ཐང་ལྡན་པ་བཅས་ཀྱི"མི་དམངས་ཀྱི་ཁེ་ལས་པ"ཞིག་ཡིན་པར་གདེང་འཇོག་དེ་འདྲ་བྱས་ཆོག་ཅིང་། ཁོང་དང་བསྡུར་ན་ཁེ་ལས་པ་ཆེས་མང་ཆེ་བ་ནི་ལེགས་འགྲུབ་བྱུང་བའི་ཚོང་པ་ཁོ་ནར་བརྩི་དགོས(དཔེར་ན་ལན་ཞང་གློག་ཀླད་ཀྱི་ལིའུ་ཁྲོན་གྱི་ལྟ་བུ) གལ་ཏེ་ཀྲུང་གོར་ད་དུང་ཧྭ་ཝེ་ཀུང་སི་རེན་གྲེན་ཧྥེ་ལྟ་བུའི་ཡུལ་བྱུང་ཁེ་ལས་པ་མང་པོ་འབྱུང་ཐུབ་ན། ཨ་རིས་ད་ནས་བཟུང་ཀྲུང་གོའི་འཕེལ་རྒྱས་ལ་ཚོད་འཛིན་བྱེད་མི་ཐུབ།

......

（在美国动用国家力量全面封锁扼杀之下,华为不仅没有倒下,而且还能励精图治靠自身研发推出更具竞争力的新产品“华为 Mate60”,堪称“石破天惊”！华为用事实狠狠抽打了美国的脸,说明美国用尽卑鄙手段最终也是枉然。

......

任正非已经不是一般的企业家了,而是有战略眼光、勇于肩负民族使命、具备卓越不凡领导才能的“人民企业家”,与他相对比,绝大多数企业家最多只能算是成功的商人了(比如联想电脑的柳传志)。中国若能再出现更多华为公司任正非这样的优秀企业家,那么美国就再也无法打压限制中国的崛起了　......）

新闻文章 2

Mate60 གློ་བུར་དུ་བྱུང་བ། ཧྭ་ཝེ་ཡིས་5Gཁ་པར་ཐོན་སྐྱེད་བྱེད་ཐུབ།འཐབ་ཇུས་རིམ་པའི་དོན་སྙིང་ལྡན་ཡོད།

གཅིག　ཧྭ་ཝེ་ཡིས་5Gཁ་པར་ཐོན་སྐྱེད་བྱེད་པའི་དཀའ་གནད་གང་ན་ཡོད།　ཐོག་མར་ལས་ཚབ་དང་ཅིང་ཕྲེལ་བཟོ་སྐྲུན་གྱི་དཀའ་ཚད་ཡིན། མི་མང་པོ་ཞིག་གིས་མིག་སྔར་ཧྭ་ཝེ་ཡིས་5Gཁ་པར་བཟོ་མི་ཐུབ་པའི་རྒྱུ་རྐྱེན་ནི་འཕྲུལ་ཚད་ཐོག་པའི་འཕྲུལ་ཆས་ཐོན་སྐྱེད་བྱེད་མི་ཐུབ་པ་དང་།ཁྱབ་ཁྱོན་ཕྲེང་འདི་ཏག་ཏག་དེ་ལས་ལྡོག་སྟེ་ཧྭ་ཝེ་རྨང་གཞིའི་ས་ཚིགས་ཀྱིས་དངོས་པོ་བསྐྱར་བ་རྒྱུན་ལྡན་ཡིན་པ་དང་འཁྲུལ་ཚད་ཐོག་པའི་འཕྲུལ་ཆས་ཀྱིས་ཐག་གཅོད་བྱེད་ཐུབ་པར་འདོད།　དཀའ་གནད་ངོ་མ་ནི་5Gཚ་བ་རྒྱས་པ་ཅུང་ཆེ་བ་དང་།ཐག་གཅོད་ཆས་དང་རྨང་གཞིའི་གླང་བྱ་ཅུང་དམའ་བའི་ནུས་ཚད་ཟད་གྲོན་དགོས་པ་དང་།ཉུང་མཐར་ཡང་7nmད་གཟོད་ཅུང་ལེགས་པའི་དོ་མཉམ་ནུས་པ་དང་ཚ་བ་རྒྱས་ཐུབ།

......

མི་ཁ་ཤས་ཀྱིས་ཨ་མེ་རི་ཀ་ཅན་འདོར་བ་ཞེས་པའི་གོ་དོན་ཅི་ཡིན་པ་མི་ཤེས་པར་ཨ་མེ་རི་ཀའི་དངོས་པོ་གང་ཡང་སྤྱད་མི་ཆོག་སྙམ་གྱི་ཡོད་ཀྱང་དོན་དངོས་ཐོག་དེ་འདྲ་མ་རེད།　ཨ་མེ་རི་ཀ་ཅན་འདོར་བའི་ནང་དོན་ནི་　བསྟུད་མར་ཐོན་སྐྱེད་བྱས་ནས་ཨ་རིའི་ཚོད་འཛིན་ཅན་དུ་འགྱུར་བར་སྔོན་འགོག་བྱས་པ་ཡིན།　གལ་ཏེ་ཁྱོད་ཀྱིས་ཨ་རིའི་དཔྱད་ཆས་ཉོས་ཡོད་ན།　དེ་ན་བཀོལ་དགོས་ན་བཀོལ།རྒྱུ་མཚན་ནི་ཨ་རིས་ཚོད་འཛིན་བྱེད་མི་ཐུབ།　ཁྱོད་ཀྱིས་wintelགློག་ཀླད་བཀོལ་བ་དང་།　ཨ་རི་བས་ཀྱང་ཁྱོད་ལ་ཚོད་འཛིན་བྱེད་དཀའ།རྒྱུ་མཚན་ནི་དངོས་པོ་འདི་སྲང་ལམ་གང་པོ་ཡིན་པས་ཚོད་འཛིན་བྱེད་མི་ཐུབ།　འཇར་པན་དང་ཁོ་རེ་ཡའི་སྒྲིག་ཆས་ཁ་ཤས་ཨ་རི་བས་གནོན་ཤུགས་སྤྲད་པས་མཁོ་སྤྲོད་བྱེད་ཐབས་བྲལ་བ་དང་།སྲིད་ཕྱོད་ཆེན་པོ་ནི་ཨ་མེ་རི་ཀ་ཅན་འདོར་དགོས་པའི་དངོས་པོར་གཏོགས་པ་དང་།རྒྱུ་མཚན་ནི་ཚང་མར་གྲབས་ཆས་མེད་པས་ཡུན་རིང་འགོར་མི་ཐུབ།

......

（Mate60 横空出世:华为能生产 5G 手机,意义是战略级的

华为要生产出 5G 手机难在哪里?首先就是代工,芯片制造的难度。和很多人以为当前华为造不出 5G 手机是射频器件无法生产消息相反的是,华为基站发货正常,射频器件是可以解决的。难的是因为 5G 发热较大,处理器和基带要求需要较低的功耗,至少要 7nm 才可以较好的平衡性能和发热。

......

一些人不懂去美化是什么意思,以为是不允许用到任何美国东西,其实不是这样的。去美化意思是连续生产去美国控制化。假如你已经买了美国仪表,那该用就用,因为美国控制不了。包括你就是用 wintel 电脑,美国人也难以控制你,因为这东西满大街都是,控制不了。而有些日韩设备,因为被美国人施压,无法供应,也属于大概能的去美化要去的,因为备件都没有,长久不了　.......）

新闻文章 3

ཚང་མས་ཤེས་གསལ་ལྟར། ཉེ་ཆར་ཧྭ་ཝེ་Mate60 བརྙན་པ་དེ་ནི་དངོས་གནས་འཛམ་གླིང་ཡོངས་ཀྱི་ཚན་རྩལ་ཁྱབ་ཁོངས་སུ"ཉིང་རྡུལ་འབར་མདེལ"ཞིག་གཡུགས་པ་དང་འདྲ། "ཉིང་རྡུལ་འབར་མདེལ"དེར་དོན་སྙིང་སྒྲིར་བཏང་མེན་པ་མ་ཟད་ཤུགས་རྐྱེན་ཡང་ཧ་ཅང་ཆེན་པོ་ཡོད།གང་ལ་ལགས་ཤེ་ན།དེ་ཉིང་སྙེབ་ཀྱི་བཀག་རྒྱ་དང་།5G·ཉིང་སྙེབ་ལག་རྩལ།གནོན་ཤུགས་དང་འགོག་རྒོལ་སོགས་དང་འབྲེལ་བ་ཡོད་པས"ཕྱེ་མ་ལེབ"ཀྱིས་གཤོག་པ་རྡེབ་ཙམ་བྱས་ཟིན་ཞེས་བཤད་ཆོག དེའི་འཕྲོ་ནས"རླུང་ནག་འཚུབ་མ"གང་ནས་གཡུག་པ་མི་ཤེས། མིག་སྔའི་གནས་ཚུལ་ལ་བལྟས་ན།ཨ་རི་ཡང་ཧྭ་ཝེ་Mate60བྱུང་བའི་རྐྱེན་གྱིས། ཐེལ་འཚུབ་དང་དཀའ་ཁག་གཉིས་ཀྱི་གནས་སུ་ལྷུང་། ཐོན་ལ་ཐེལ་འཚུབ་ཀྱི་ཕྱོགས་འདི་བཤད་ན། ཨ་རིའི་ཚོང་དོན་ལྷུའུ་ཀྲང་ལེ་མེན་ཏོ་ནི་ཐུང་དཔེ་མཚོན་ཅན་ཞིག་ཡིན་པ་ཐག་རེད། མོ་ཀྲུང་གོ་ན་ཡོད། དེ་ནས་ཧྭ་ཝེ་Mate60བཙོང་བཞིན་ཡོད། མོ་ལ་ཁ་ངོ་མེད་པར་བཟོས་སོང་།

......

འོན་ཀྱང་དེ་དང་མཉམ་དུ།ཨ་རི་ཡང་དཀའ་ཁག་གཉིས་ཀྱི་ཁྲོད་དུ་ལྷུང་སོང་།དེ་ནི་བཀག་རྒྱ་མུ་མཐུད་དུ་བྱེད་དགོས་སམ། དེ་སྔོན་བཀག་རྒྱ་དེ་འདྲ་ནན་པོ་ཡོད་པ་དང་ཀྲུང་གོས་རྒྱུན་ཆད་མེད་པར་འབད་བརྩོན་བྱེད་པ་དང་དེ་རྗེས་དུ་དྲང་བཀག་རྒྱ་ལས་བརྒལ་བ་དང་ཐ་ན་མི་མང་པོ་ཞིག་གིས་ཨ་རིས་བཀག་སྡོམ་བྱས་པའི་རྐྱེན་གྱིས་ཀྲུང་གོའི་ཉིང་སྙེབ་ཐོན་ལས་རྒྱུན་ཆད་མེད་པར་ཚོད་བརྒལ་ཡོང་རྒྱུར་ཐོག་སྐྱལ་བྱེད་ཀྱི་ཡོད་པས།ཨ་རིའི་བཀག་རྒྱ་དེ་འགྱུར་སྐུལ་རྫས་སུ་གྱུར་ཡོད་ཅེས་བརྗོད་ཀྱི་ཡོད། གལ་ཏེ་བཀག་རྒྱ་མ་བཞག་ན་ཨ་མེ་རི་ཀ་རང་ཉིད་ཀྱིས་ངོ་ཚ་གཏང་དང་པ་མ་ཡིན་ནམ།དེའི་ཐོག་ནས་ཨ་མེ་རི་ཀ་འཛམ་གླིང་ཡོངས་ཀྱི་ཉིང་སྙེབ་ཐོན་ལས་ཐོག་ནས་འགོ་ཁྲིད་ཀྱི་གོ་གནས་སུ་གནས་མེད་པ་གསལ་བཤད་བྱུང་བ་རེད།བཀག་རྒྱ་དེ་འདྲ་ནན་པོ་ཞིག་མཐར་ལེགས་འགྲུབ་མ་བྱུང་བ་དེ་ནི་གང་ཞིག་བཀག་སྡོམ་བྱས་པ་རེད།

......

（众所周知,最近华为 Mate60 的横空出世,真的是向全球科技领域,相当于扔了一颗“核弹”。这颗“核弹”意义非凡,且影响很大,因为它涉及到芯片禁令、5G、芯片技术、打压与反抗等等,可以说“蝴蝶”已经扇动了翅膀,接下来不知道哪里会刮起“龙卷风”了。从目前来看,美国也因为华为 Mate60 的出现,而陷入了慌乱和两难之中。先说慌乱这一块,美国商务部长雷蒙多肯定是比较典型的,她人在中国,然后华为 Mate60 打发售,搞的她很没面子。

......

但与此同时,美国接下来也陷入了两难之中,那就是禁令要不要继续? 之前禁令这么严格,中国在不断努力,然后还突破了封锁,很多人甚至称,正是因为美国的封锁,倒逼着中国芯片产业不断突破,美国的禁令,成为了催化剂。而如果不禁,那岂不是美国自己打脸,说明美国根本就没有在全球芯片产业上,处于领导地位,这么严格的禁令,最后没输了,这是禁了个寂寞? ......）

多文档摘要

ཟླ་8ཚེས་29ཉིན་གྱི་ཉིན་དགུང་ཧྭ་ཝེ་ཡིས་ལག་ཐོགས་ཁ་པར་Mate60proའཚོང་འགོ་ཚུགས།སྐབས་དེ་དུས་ཨ་རིའི་ཚོང་དོན་ལྷུའུ་ཀྲང་ལེ་མོང་ཏུའོ་ཡིས་ཟླ་8ཚེས་27ཉིན་ནས་ཟླ་8ཚེས་30ཉིན་བར་དུས་ཡུན་ཉིན་4ཡི་ཀྲུང་གོར་འཚམས་འདྲིར་ཕེབས་པའི་བྱེད་སྒོ་སྤེལ་བ་རེད།ཧྭ་ཝེ་Mate60ཁྲོམ་རར་བཀྲམ་པ་དེས་ལེ་མོང་ཏུའོ་ལ་སྔ་ལྟས་ཅི་ཡང་མེད་པའི་ལེགས་སྐྱེས་ཤིག་ཕུལ། ཧྭ་ཝེ་Mate60ཁྲོམ་རར་བཀྲམ་པ་ལ་སྔ་ལྟས་ཅི་ཡང་མེད་ཟེར་བ་ནི་གནས་ལུགས་ལྡན་པ་ཞིག་རེད།རྒྱུ་མཚན་ནི་དེས་མི་ཚང་མ་ཅི་བྱ་གཏོལ་མེད་དུ་གཏོང་བ་དང་། དོན་དེ་ནི་དད་མོས་པ་དང་ལམ་འགྲོ་བ་གང་ཡིན་རུང་བསམ་ཡུལ་ལས་འདས་པ་ཞིག་རེད།དེའི་སྔོན་ཕྱིར་ལ་དོན་དེའི་"ཆ་འཕྲིན་ཙུང་ཙམ་ཡང་མ་མངོན།དེས་མ་ཚད་ད་ཐེངས་ཁྱབ་བསྒྲགས་བྱས་པའི་60བཀྲག་མདངས་འཚེར་ས་ཆེ་ཤོས་ནི་དེའི་སྒྲིག་གཙོད་ཆས་ལ་"ཧྭ་ཝེ་རང་ཉིད་ཀྱིས་ཞིབ་བཟོ་བྱས་པ་དང་ཀྲུང་ཤིན་རྒྱལ་སྤྱིའི་ལས་ཚབ་ཐོན་སྐྱེད་བྱས་པའི་7nmཡི་ཐོན་རྫས་སྤྱད་ཅིང་དེར་5Gདྲ་རྒྱ་སྐྱོང་ཆོག་པ་དེ་རེད། ཧྭ་ཝེ་ཡིས་ལེ་མེན་ཏོ་ཀྲུང་གོར་འཚམས་འདྲིར་ཡོང་བའི་སྐབས་སུ་དོན་འདི་ཁྱབ་བསྒྲགས་བྱེད་དོན་ནི་ལེ་མེན་ཏོ་ལ་ཧྭ་ཝེ་དར་བའི་རྣམ་པ་ཞིག་དངོས་སུ་མཐོང་དུ་འཇུག་འདོད་པ་ཡིན་པ་གདོན་མི་ཟ།འདི་ནི་ད་གཟོད་"མི་གསོད་སེམས་འདུལ"ཟེར་བ་རེད།

（8 月 29 日中午，华为开售 Mate60pro 手机。恰在此时，美国商务部长雷蒙多于 8 月 27 日到 8 月 30 日展开了为期 4 天的访华行程。华为 Mate60 的上市，给雷蒙多送上了一份毫无征兆的“礼物”。华为 Mate60 上市，说它毫无征兆是有道理的，因为它打了所有人一个措手不及，不管是粉丝还是路人都没能料到，此前一点没有透露出风声。而且，此次发布的 60，最大的亮点是，其处理器使用了华为自行研制、中芯国际代工生产的 7nm 产品，支持 5G 网络。之所以华为选择在雷蒙多访华之际公布这一消息，无疑是想让雷蒙多亲眼看到华为崛起的一幕，这才是所谓的“杀人诛心”。）

事件名称

ཧྭ་ཝེ·Mate60·སྣ་བྱར་དུ་བརྙན་པ།

（华为 Mate60 横空出世）

以蒙文多文档摘要为例，数据展示如表 3 所示。由于蒙文文档段落截取与上文所述藏文文档截取部分相同，此部分对于中文翻译不再赘述，可参考表 1 中文部分。

表 3　蒙文多文档摘要数据集样例

Table 3　Sample of Mongolian multi-document summary datasets

新闻文章 1

[illegible]

[illegible] 《[illegible] Mate60》 [illegible] 《[illegible]》 [illegible] !

[illegible]...

[illegible]

[illegible] 《[illegible]》 [illegible]

[illegible] ([illegible])

......

新闻文章 2

Mate60 [illegible] : [illegible] 5G [illegible]

[illegible] 5G [illegible] ? [illegible]

[illegible] 5G [illegible] 5G [illegible] 7nm [illegible]

......

[illegible] 5G [illegible]

[illegible] wintel [illegible]

......

新闻文章 3

[illegible] Mate60 [illegible] 《[illegible]》 [illegible]

([illegible])

[illegible] 《[illegible]》 [illegible] 5G [illegible] 《[illegible]》 [illegible] 《[illegible]》 [illegible]

[illegible] Mate60 [illegible]

[illegible] Mate60

[illegible]

……

[illegible]《[illegible]〈[illegible]〉[illegible]》

[illegible]

……

多文档摘要

8 [illegible] 29 [illegible] Mate60pro [illegible] 8 [illegible] 27 [illegible] 8 [illegible] 30 [illegible] Mate60 [illegible]《[illegible]》[illegible] Mate60 [illegible] Mate 60 [illegible] 7nm [illegible] 5G [illegible]《[illegible]》[illegible]

（8 月 29 日中午，华为开售 Mate60pro 手机。恰在此时，美国商务部长雷蒙多于 8 月 27 日到 8 月 30 日展开了为期 4 天的访华行程。华为 Mate60 的上市，给雷蒙多送上了一份毫无征兆的“礼物”。华为 Mate60 上市，说它毫无征兆是有道理的，因为它打了所有人一个措手不及，不管是粉丝还是路人都没能料到，此前一点没有透露出风声。而且，此次发布的 60，最大的亮点是，其处理器使用了华为自行研制、中芯国际代工生产的 7nm 产品，支持 5G 网络。之所以华为选择在雷蒙多访华之际公布这一消息，无疑是想让雷蒙多亲眼看到华为崛起的一幕，这才是所谓的“杀人诛心”。）

事件名称

[illegible] Mate60 [illegible]

（华为 Mate60 横空出世）

以维文多文档摘要为例，数据展示如表 4 所示。（由于维文文档段落截取与上文所述藏文文档截取部分相同，此部分对于中文翻译不再赘述，可参考表 1 中文部分）

表 4　维文多文档摘要数据集样例

Table 4　Sample of Uygur multi-document summary datasets

新闻文章 1

ئامېرىكىنىڭ دۆلەت كۈچىنى ئىشقا سېلىپ ئومۇمىيۈزلۈك كونترول قىلىشى ۋە بېسىمى ئارقىسىدا، خۇاۋېي گۇمران بولمايلا قالماستىن، بەلكى تېخىمۇ غەيرەتكە كېلىپ، ئۆزىنىڭ تەتقىق قىلىپ ئېچىش كۈچىگە تايىنىپ تېخىمۇ رىقابەت كۈچىگە ئىگە يېڭى مەھسۇلاتى خۇاۋېي《Mate 60》نى ئوتتۇرىغا چىقىرىپ ھەممىنى ھەيران قالدۇردى.

خۇاۋېي گەرچە بىر جۇڭگو كارخانىسى بولسىمۇ، لېكىن ئامېرىكىدىكى ھاياتىي كۈچكە ئىگە «پەن-تېخنىكا كۈچلۈك دۆلىتى»گە ئايلىنىپ، ئامېرىكا سىياسەتۋازلىرىنى قاتتىق چۆچۈتتى. بۇ خۇاۋېي شىركىتىنىڭ جاسارىتى، ئىرادىسى، غەيرىتى ۋە غايەت زور كۈچىنى تولۇق نامايەن قىلغان. ئۇ مىللىي يۇقىرى پەن-تېخنىكا كارخانىلىرىنىڭ ئۇلۇغ بايرىقى بولۇشقا مۇناسىپ بولۇپلا قالماي، پۈتكۈل جۇڭگو پۇقراسىنىڭ مىللىي ئىرادىسىنى ھەسسىلەپ ئۇرغۇتقان.
شۇڭا، بۇنىڭغا مۇنداق باھا بېرىشكە بولىدۇ:

……

新闻文章 2

<Mate60>نىڭ دۇنياغا كېلىشى، خۇاۋېي <5G> يان تېلېفوننىڭ ئىشلەپچىقىرىلىشى ئىستراتېگىيەلىك ئەھمىيەتكە ئىگە . بىرىنچىدىن، خۇاۋېي شىركىتىنىڭ G 5 لىق يان تېلېفون ئىشلەپچىقىرىشىدىكى قىيىنچىلىق زادى قەيەردە؟ ئالدى بىلەن ۋاكالىتەن ئىشلەپ چىقىرىش، ئۆزەك ياساشنىڭ قىيىنلىق دەرىجىسى مەسىلىسى.

نۇرغۇن كىشىلەرنىڭ ھازىر خۇاۋېيدا G5 لىق يان تېلېفون ياسىغىلى، رادىئو چاستوتىلىق زاپچاس ئىشلەپچىقارغىلى بولمايدۇ دەپ قارىشىنىڭ ئەكسىچە، خۇاۋېي ئاساسىي پونكىتىدا مال تارقىتىش نورمال، رادىئو چاستوتىلىق زاپچاس ئارقىلىق ھەل قىلغىلى بولىدۇ.

قىيىن يېرى شۇكى، G5 نىڭ قىززىشى بىر قەدەر چوڭ بولۇپ، بىر تەرەپ قىلغۇچ ۋە ئاساسىي بەلباغنىڭ بىر قەدەر تۆۋەن قۇۋۋەت سەرپىياتى تەلەپ قىلىنىدىغان بولۇپ، ئاز دېگەندە 7nm بولغاندا ئاندىن بىر قەدەر ياخشى تەڭپۇڭلۇق ئىقتىدارى ۋە قىزىش ئىقتىدارىغا ئىگە بولىدۇ.

......

خۇاۋېي G5 يان تېلېفوندىن ئىبارەت بۇ بەلگە خاراكتېرلىك مەھسۇلات بارلىققا كەلسە، ئامېرىكىلىقلار چوقۇم ئاچچىقتىن بوغۇلۇپ قالىدۇ. ئەمما بۇ پۈتۈن دۇنيا خەلقىگە ئامېرىكا بولمىغان تەقدىردىمۇ، جۇڭگونىڭمۇ ئىلغار ئۆزەك ياپراقچىسىنى ياساپ چىقالايدىغانلىقىغا، يېرىم ئۆتكۈزگۈچ كەسپىنى ياخشى ئىشلىيەلەيدىغانلىقىدىن ئىبارەت ئىشەنچ ئاتا قىلدى.

......

新闻文章 3

ھەممىگە مەلۇمكى، يېقىندا خۇاۋېي Mate 60 نىڭ دۇنياغا كېلىشى ھەققەتەن دۇنيا پەن-تېخنىكا ساھەسىگە بىر "يادرو بومبىسى" ئاتقانغا باراۋەر.بىر ئىش بولدى. (سۈرەت ماتېرىيالىدىن پايدىلىنىشقىلا بولىدۇ)

بۇ "يادرو بومبىسى" نىڭ ئەھمىيىتى ئالاھىدە، تەسىرى ناھايىتى چوڭ بولغان. چۈنكى ئۇ ئۆزەك چەكلەش بۇيرۇقى، G5، ئۆزەك تېخنىكىسى، سوقۇش ۋە قارشىلىق كۆرسىتىش قاتارلىقلارغا چېتىلىدىكەن. بۇ ھالەتنى كېيىنەك قانتىنى قېقىپ بولدى، ئەمدى قەيەردە قارا قۇيۇن چىقىدىغانلىقىنى بىلگىلى بولمايدۇ دېيىشكە بولىدۇ.

......

ئىلگىرى چەكلەش بۇيرۇقى مۇشۇنداق قاتتىق ئىدى، جۇڭگو ئۈزلۈكسىز تىرىشچانلىق كۆرسىتىپ، كېيىن يەنە قامالنى بۆسۈپ ئۆتكەن. نۇرغۇن كىشىلەر ھەتتا ئامېرىكىنىڭ قامال قىلىشى جۇڭگونىڭ ئۆزەك كەسپىنى ئۈزلۈكسىز بۆسۈپ ئۆتۈشكە مەجبۇر قىلغاچقا، ئامېرىكىنىڭ چەكلەش بۇيرۇقى ھەرىكەتلەندۈرگۈچ كۈچكە ئايلاندى، دېيىشتى.

......

多文档摘要

8-ئاينىڭ 29-كۈنى چۈشتە، خۇاۋېينىڭ mate60pro تىپىدىكى تېلېفونى رەسمىي تۈردە سېتىلىشقا باشلىغان. دەل مۇشۇ ۋاقىتتا، ئامېرىكا سودا مىنىستىرلىقىنىڭ مىنىستىرى لېيمېگدۇ 8-ئاينىڭ 27-كۈنىدىن 29-كۈنىگىچە 4كۈنلۈك جوڭگونى زىيارەت قىلىش سەپىرىنى باشلىدى.

خۇاۋېي تىلېفۇنىنىڭ بازارغا چىقىشى، ئامېرىكا سودا مىنىستىرلىقىنىڭ مىنىستىرى لېيمېگدۇغا ئالاھىدە بىر تەييارلانغان «سوغات» بولدى. خۇاۋېي mate 60pro نىڭ بۇنداق دېيىلىشىدە ھەرخىل ئاساسلار ھازىرلانغان بولۇپ، ئاساسلىقى ئۇ بارلىق كىشلەرنى تەمتىرىتىپ قويغانىدى.

مەيلى ئۇزاقتىن بېرى خۇاۋېينى قوللاپ كېلىۋاتقان مەستانىلەر بولسۇن ياكى تارتىپ خۇاۋېينىڭ يېڭى مەھسۇلاتنى بۇنچە تېز بازارغا سېلىشىنى ئويلاپ باقمىغان بولۇپ، خۇۋېي ماركىسىدىكى تېلغون رەسمىي تۈردە بازارغا چىقىشتىن بۇرۇن ھېچقانداق ئۇچۇرنى سىرتقا ئاشكارلىمىغانىدى.

ئۇنىڭ ئۈستىگە، ئۇنىڭدىن بۇرۇن بازارغا سېلىنغان خۇاۋېي 60تىپىدىكى تېلېفۇننىڭ ئەڭ چوڭ ئالاھىدىلىكى بولسا تېلفۇننىڭ ئۆزىكى خەلقئارا ۋاكالىتەن ئىشلىگەن 7 نانومېتىرلىق مەھسۇلات بولۇپ G5 تورنى قوللايدىكەن. يۇقارقىلاردىن كۆرۈۋېلىشقا بولىدۇكى، خۇاۋېينىڭ ئامېرىكا سودا مىنىستىرى لېيمېگدۇ جوڭگوغا كەلگەندە يېڭى مەھسۇلاتنى ئېلان قىلىشى، لېيمېگدۇنى جوڭگو خۇاۋېينىڭ باش كۆتۈرگەنلىكىنى ئۆز كۆزى بىلەن كۆرۈش پۇرسىتىگە ئىگە قىلىش خۇۋېينىڭ بازارغا سېلىنىشىنىڭ ھەققىي ئەھمىيىتى ئىدى.

（8 月 29 日中午，华为开售 Mate60pro 手机。恰在此时，美国商务部长雷蒙多于 8 月 27 日到 8 月 30 日展开了为期 4 天的访华行程。华为 Mate60 的上市，给雷蒙多送上了一份毫无征兆的“礼物”。华为 Mate60 上市，说它毫无征兆是有道理的，因为它打了所有人一个措手不及，不管是粉丝还是路人都没能料到，此前一点没有透露出风声。而且，

此次发布的 60，最大的亮点是，其处理器使用了华为自行研制、中芯国际代工生产的 7nm 产品，支持 5G 网络。之所以华为选择在雷蒙多访华之际公布这一消息，无疑是想让雷蒙多亲眼看到华为崛起的一幕，这才是所谓的“杀人诛心”。）

事件名称

خۇاۋېي دۇنياغا كەلدى

（华为 Mate60 横空出世）

3 数据质量控制和评估

为了确保数据的高质量和可靠性，本研究采取了一系列数据质量控制和评估方法。首先，从多个新闻网站爬取数据，经过初步的自动筛选和人工审核，去除质量低下或与主题不相关的新闻报道，确保所选新闻簇的准确性和相关性。接着，对于机器翻译生成的蒙文、藏文和维文版本，邀请了语言专家进行人工校对，以提高翻译质量和文本的自然流畅度。

此外，为了进一步验证数据集的可用性，聘请了 9 名研究生对本数据集中的多文档摘要质量进行评估，聘请人员其中包括：3 名蒙语母语人、3 名藏语母语人以及 3 名维吾尔语母语人。针对数据集中的每种语言，共计进行三轮评估。每论评估均随机选择 30 簇新闻文章，使其按照以下指标对多文档摘要质量进行判断打分，最终评价结果为三轮评估的平均分。（1）覆盖度：将摘要与实际文章进行比对，检查主要观点是否被充分覆盖。（2）翻译质量：将蒙语、藏语、维吾尔语文章与中文文章进行比对，检查翻译结果是否准确。（3）语法质量：考察数据集中的文本摘要在语义、语法上是否正确。（4）主题感知度：分析数据集中的摘要否遵循合适的写作风格，以反映其所属主题。9 名评估人员用从 0（最差）到 1（最好）给出三个指标的相应分数。结果显示，评估者普遍同意本数据集具有较高的覆盖度，这反映了摘要和原始文章之间的事实一致性。此外，摘要的语法正确度和主题契合度也达到了要求。具体评价结果如表 5 所示。

表 5 人工评价结果

Table 5 Mannual evaluation results

语言	覆盖度	翻译质量	语法质量	主题感知度
中文	0.926	—	0.954	0.886
藏语	0.862	0.846	0.908	0.892
蒙语	0.883	0.852	0.912	0.834
维语	0.845	0.730	0.874	0.891

4 数据价值

在当前的自然语言处理（NLP）领域，多文档摘要已成为一个重要的研究方向，尤其是在处理新闻报道和在线内容方面。本研究构建了一个多语言多文档摘要数据集，旨在为跨语言信息处理[8]、多文档摘要算法研究[9]以及机器翻译系统评估[10]提供数据资源。本研究中所构建的多文档摘要数据集，通过精心挑选和处理来自多个新闻网站的同一事件报道，提供了一个独特而全面的资源。这一方法不仅能捕捉到事件的各个方面，还能体现出不同来源在报道同一事件时的多样性和视角差异。

这种多元化的数据收集方式对于发展更复杂、更全面的文本摘要算法至关重要，因为它可以帮助模型学习如何从不同的信息源中提取和整合关键信息。

此外，本数据集的构成也特别关注了多语言和跨文化交流的需求。通过将中文原始新闻翻译成藏文、蒙文和维文，采取机器翻译和人工校对结合的方法，为这些低资源语言提供了较难大批量获取的数据资源。低资源语言在自然语言处理领域经常面临数据不足的问题，这严重限制了相关语言技术的发展。通过提供这些语言的高质量摘要，可以显著提高机器学习模型在这些语言上的表现，进而促进语言多样性和文化交流与发展，同时也为低资源语言的 NLP 应用开辟了新的可能性。

在多文档摘要的应用方面，本数据集的实际用途不限于提高摘要算法的效率。它还可以用于训练机器学习模型，以更好地理解和处理复杂的信息，特别是在新闻和媒体内容的上下文中。此外，这种多文档摘要的方法能够帮助模型更好地理解文本中的主题、观点和关键信息，这对于提高新闻推荐系统、信息检索和内容聚合的质量至关重要。

致　谢

特别感谢参与本数据集工作的蒙语专业人员、藏语专业人员以及维吾尔语专业人员。

作者分工职责

翁彧（1980—），男，辽宁省大连市人，博士，教授，研究方向为民族语言智能分析。主要承担工作：数据集质量控制与综合管理。

邢天娇（2000—），女，辽宁省铁岭市人，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、数据预处理、论文撰写。

叶旭明（2000—），男，湖南省衡阳市人，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、论文撰写。

刘征（1990—），男，黑龙江省哈尔滨市人，博士研究生，研究方向为民族语言智能分析。主要承担工作：数据质量控制。

超木日力格（1988—），女，内蒙古自治区通辽市人，博士，讲师，研究方向为民族语言智能分析。主要承担工作：数据集质量评分。

刘轩（1992—），男，江西省南昌市人，博士，讲师，研究方向为最优化。主要承担工作：数据质量控制。

参考文献

[1] MA C B, ZHANG W E, GUO M Y, et al. Multi-document summarization via deep learning techniques: a survey[J]. ACM Computing Surveys, 2022, 55(5): 102. DOI: 10.1145/3529754.

[1] MA C B, ZHANG W E, GUO M Y, et al. Multi-document summarization via deep learning techniques: a survey[J]. ACM Computing Surveys, 2022, 55(5): 102. DOI: 10.1145/3529754.

[2] SEKINE S, NOBATA C. A survey for multi-document summarization[C]//Proceedings of the HLT-NAACL 03 on Text summarization workshop - Volume 5. ACM, 2003: 65 – 72. DOI: 10.3115/1119467.1119476.

[3] OVER P, YEN J. An introduction to DUC-2004[J]. National Institute of Standards and Technology, 2004.
[4] JI H, GRISHMAN R, DANG H T, et al. Overview of the TAC 2010 knowledge base population track[C]//Third text analysis conference (TAC 2010), 2010.
[5] ADELE D. Advances in neural information processing systems 18[J]. JOURNAL OF MATHEMATICAL PSYCHOLOGY, 2007, 51(5): 339.
[6] GHALANDARI D G, HOKAMP C, PHAM N T, et al. A large-scale multi-document summarization dataset from the Wikipedia current events portal[J]. arXiv preprint arXiv:2005.10070, 2020.
[7] HU B T, CHEN Q C, ZHU F Z. Lcsts: A large scale Chinese short text summarization dataset[J]. arXiv preprint arXiv:1506.05865, 2015.
[8] WANG J A, MENG F D, ZHENG D, et al. A survey on cross-lingual summarization[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 1304 – 1323. DOI: 10.1162/tacl_a_00520.
[9] 解晓庆. 朝鲜语多文档摘要生成的研究[D]. 北京：中央民族大学, 2021. [XIE X Q. A study of multi-document abstract generation in Korean[D]. Beijing: Minzu University of China, 2021.]
[10] 刘心怡. 面向低资源语言的无监督机器翻译双语语料挖掘研究[D]. 郑州：郑州轻工业大学, 2023. [LIU X Y. Research on Bilingual Corpus Mining of Unsupervised Machine Translation for Low-Resource Languages[D]. Zhengzhou: Zhengzhou University of Light Industry, 2023.]

论文引用格式

翁彧，邢天娇，叶旭明，等. 中-蒙-藏-维文多文档摘要数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-26). DOI: 10.11922/11-6035.csd.2024.0038.zh.

数据引用格式

刘征，邢天娇，叶旭明，等. 面向多文档摘要的中-蒙-藏-维文数据集[DS/OL]. V2. Science Data Bank, 2024. (2024-11-13). DOI: 10.57760/sciencedb.j00001.01084.

A dataset of Chinese-Mongolian-Tibetan-Uyghur multi-document summaries

WENG Yu[1,2], XING Tianjiao[1,2], YE Xuming[1,2], LIU Zheng[1,2*], CHAOMU Rilige[1,2], LIU Xuan[1,2]

1. Information Engineering and the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Beijing 100081, P.R. China
2. College of Information Engineering, Minzu University of China, Beijing 100081, P.R. China
*Email: liuzheng@muc.edu.cn

Abstract: Multi-document summarization is a key task in natural language processing and a crucial technology for understanding and processing the core content within a large volume of documents. Despite

significant progress in multi-document research for English and Chinese, research on low-resource languages has been slower due to a lack of data. This paper addresses this gap by first constructing a Chinese multi-document summary dataset and then using machine translation combined, with expert verification, to create aligned corpora in Mongolian, Tibetan, and Uyghur. Based on this, this study produced a multilingual multi-document summary dataset (MMDS) in Chinese, Mongolian, Tibetan, and Uyghur, each containing 1,044 news clusters (6,234 news articles) covering major events from 2018 to 2023. Additionally, the quality of the dataset is ensured through human evaluation, making it well-suited for training deep learning models and conducting NLP research across multiple languages. The release of this dataset is of significant value for promoting the development of information processing in low-resource languages.

Keywords: multi-document summary; dataset; Mongolian; Tibetan; Uygur; low resources

Dataset Profile

Title	A dataset of Chinese-Mongolian-Tibetan-Uyghur for multi-document summaries
Data corresponding author	LIU Zheng (liuzheng@muc.edu.cn)
Data authors	LIU Zheng, XING Tianjiao, YE Xuming, WENG Yu, CHAOMU Rilige
Time range	2018 – 2023
Data volume	307 MB
Data format	*.xlsx
Data service system	<https://doi.org/10.57760/sciencedb.j00001.01084>
Sources of funding	National Key R&D Plan of China (2020YFB140670002-03); Beijing Municipal Science & Technology Commission (Z231100001723002).
Dataset composition	This dataset comprises multi-document summaries in Chinese, Mongolian, Tibetan, and Uyghur, with data across different languages highly aligned. Each language includes 1,044 clusters of news articles, totaling 6,234 news articles. Each article is stored in a TXT file, with the name of the news event and the corresponding cluster summary saved in an XLSX spreadsheet. The multi-document summary data for each language are stored in their respective compressed files.