专题 多语种智能信息处理

www.csdata.org

ISSN 2096-2223
CN 11-6035/N

少数民族语言分词技术评测数据集 MLWS2021

赵小兵[1,2]，高璐[1,2]，高定国[3]，包乌格德勒[4]，米尔阿迪力江·麦麦提[5]，刘洋[5]，才智杰[6,7]，孙媛[1,2*]

1. 中央民族大学，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. 西藏大学，拉萨 850013
4. 呼和浩特民族学院，呼和浩特 015501
5. 清华大学，北京 100084
6. 青海师范大学，西宁 810016
7. 藏语智能信息处理及应用国家重点实验室，西宁 810016

摘要：依据蒙古文、藏文和维吾尔文词汇的构词规律和特点，制定适合计算机信息处理的蒙古文、藏文和维吾尔文分词评测标准，构建蒙古文、藏文和维吾尔文的分词标注语料，形成标准评测数据集（MLWS2021），为解决自动分词、词性标注、信息检索、语料库构建等研究课题提供依据。MLWS2021 共包含 2.5 万句藏文、6.5 万句蒙古文、6.5 万句维吾尔文。本评测数据集将面向社会，提供免费评测服务，逐步建成权威的少数民族语言分词技术评测平台，推动少数民族语言信息处理技术的发展。

关键词：少数民族语言；分词标注；评测数据集；分词标准规范

文献 CSTR：
32001.14.11-6035.csd.2021.0091.zh
文献 DOI：
10.11922/11-6035.csd.2021.0091.zh
数据 DOI：
10.11922/sciencedb.j00001.00340
文献分类：信息科学

收稿日期：2021-12-20
开放同评：2022-01-28
录用日期：2022-06-07
发表日期：2022-06-30

数据库（集）基本信息简介

数据库（集）名称	少数民族语言分词技术评测数据集 MLWS2021
数据作者	赵小兵，高璐，高定国，包乌格德勒，米尔阿迪力江·麦麦提，刘洋，才智杰，孙媛
数据通信作者	孙媛（tracy.yuan.sun@qq.com）
数据时间范围	2017–2021
数据量	9.8 MB
数据格式	*.txt
数据服务系统网址	http://www.doi.org/10.11922/sciencedb.j00001.00340
基金项目	国家语委科研重点项目（ZDI135-118）
数据库（集）组成	数据集共包括3个数据文件，其中：(1) Tibetan.zip 是藏文分词语料，数据量2.5万个句子，文件大小1.52 MB；(2) Mongolian.zip 是蒙古文分词语料，数据量6.5万个句子，文件大小3.16 MB；(3)Uyghur.zip 是维吾尔文分词语料，数据量6.5万个句子，文件大小5.12 MB。

* 论文通信作者
孙媛：tracy.yuan.sun@qq.com

引　言

少数民族语言信息处理技术起步较晚，目前还处于初中级阶段，要解决人与计算机交互、系统问答等顶层问题，首先要从能够独立表义的最小单位即词汇开始研究。由于各少数民族语言分词标准的不统一以及语料的不开放性，极大地限制了少数民族语言信息处理发展的进程。因此，迫切需要运用评测语料库的科学计算方法进行公开、公正的评测，建构适用于自动分词评测的规范标准，从而推动少数民族语言分词的规范与统一。

本数据集从计算机的角度出发，考虑蒙古文、藏文、维吾尔文分词的规范原则，依据蒙古文、藏文和维吾尔文词汇的构词规律和特点，制定适合计算机信息处理的蒙古文、藏文和维吾尔文分词评测标准，设计三个语种的分词评测分析软件，构建蒙古文、藏文和维吾尔文的分词语料，形成标准评测数据集，为解决自动分词、词性标注、信息检索、语料库构建等研究课题提供依据。

1　数据采集和处理方法

1.1　制定分词评测标准

根据目前少数民族语言分词和词性标注研究的现状来看，数据集标准普遍采用了“规范+参考语料”的方式给出。首先从计算机理解的角度考虑蒙古文、藏文、维吾尔文词类体系和分词的规范原则，并邀请少数民族语言信息处理领域的相关专家，分别制定适用于计算机信息处理的蒙古文、藏文、维吾尔文分词评测标准。

藏文分词评测标准的制定借鉴了《信息处理用现代汉语分词规范》（GB/T 13715-1992）[5]、《信息处理用藏文分词规范》（GB/T 36452-2018）[3]、《信息处理用藏语词类标记集》（GB/T 36337-2018），对每一词类制定详细的切分细则。

蒙古文按照特定的规范，把词表示为词干和构形词缀的形式。蒙古文分词评测标准的制定主要依据《信息处理用现代汉语分词规范》（GB/T 13715-1992）[5]和《信息处理用蒙古文词语标记》（GB/T 26235-2010）[5]确定大类词类，并对每个词类制定详细的切分规则。

维吾尔文的分词是词干提取的过程，其制定主要依据《信息处理用现代维吾尔语词类标注标记规范集》。规范集对词形变化丰富的名词、动词、形容词进行了详细的规则介绍，并举例说明。

1.2　语料媒体来源

蒙、藏、维 3 个语种的语料均来源于由新闻、经济、法律、娱乐等各领域组成的综合语料，因此语料爬取的媒体来源广泛，表 1 展示了部分新闻媒体渠道。

表 1　部分新闻媒体来源

Table 1　Part of news media sources

语种	来源	网址
蒙古文	中国蒙古语新闻网	http://www.mgyxw.cn
	人民网蒙文版	http://mongol.people.com.cn
	呼伦贝尔蒙古语新闻网	http://hb.mgyxw.cn/mdls/am/amview.aspx?pid=0&alias=hulunbuir&iid=580901&mid=10662&wv=U

语种	来源	网址
	鄂尔多斯蒙古语新闻网	http://www.ordosmgy.cn
	中国蒙古语广播网	http://www.mongolcnr.cn
	央视网蒙文	http://mongol.cctv.com/index.html
	锡林郭勒盟蒙文政务门户网站	http://mgl.xlgl.gov.cn/U_index.html
	阿拉善蒙古语网	http://www.alsmgy.com
藏文	青海藏语网络广播电视台	http://www.qhtb.cn
	人民网藏文版	http://tibet.people.com.cn
	中国藏族网通	https://ti.tibet3.com
	中国共产党新闻网藏文版	http://tibet.cpc.people.com.cn
	中国西藏网	http://tb.tibet.cn
	中国西藏新闻网	http://tb.xzxw.com
维吾尔文	天山网	http://www.uy.ts.cn
	新疆昆仑网	http://uyghur.xjkunlun.gov.cn
	新疆语言文字网	http://www.xjyw.gov.cn
	新疆维吾尔自治区政府网	http://uygur.xinjiang.gov.cn

1.3 数据采集及预处理

1.3.1 数据爬取

每个语种团队的技术小组负责数据爬取及预处理工作。通过构建并行分布式爬虫框架，按照之前整理的蒙、藏、维各领域媒体渠道，采用合适的机制对网页数据进行爬取并保存在本地。爬取过程中，我们对网页的具体内容并不处理，以提升爬取的速度和效率。爬取结束后使用相应的预处理解析模块，提取需要保存的内容。

1.3.2 数据预处理

该模块将蒙古文、藏文、维吾尔文的网页数据转换成统一编码。此外，由于少数民族语言字符输入较为复杂，如输入蒙古文时需要考虑分写词缀、分写元音、特殊字符等，部分人员在使用时为了提高输入效率，会尽可能地减少使用这些繁琐的控制符，导致输入文本的后端编码出错，因此在进行文本处理之前势必要进行字符的校对。在数据集构建过程中，按照不同语种文本预处理的需要，应用和开发了相关的加工软件对语料进行预处理。开发的软件包括垃圾信息滤除软件、编码转换软件、语料校对软件等。

1.3.3 自动初步分词

少数民族语言分词技术已经有一定的研究基础，也积累了相应的分词工具，取得了领域内较好的分词结果。为了加快语料的处理，提高分词的准确性，利用清华大学、中央民族大学、西藏大学等依托单位现有的分词工具对各个语种的语料进行了初步分词。由于没有任何分词工具能够达到百分之百的准确性，初步分词的结果需要进一步的人工校对。

1.3.4 人工校对

人工校对工作繁重，需要大量有经验的母语人士参与。蒙古文、藏文、维吾尔文分别由中央民族大学、西藏大学、清华大学负责。每个语种团队分为技术小组、标注小组、校验小组，其中标注小组和校验小组由母语人士构成，负责语料的人工校对工作。标注小组和校验小组通力协作，不断推动着数据集的构建过程。在标注小组与校验小组标注结果达成一致时，该条数据才会成功入库；若有不一致现象产生，则移交相关负责人研判。在所有数据条目构建完成后，按照整体 10%的比例，引入第三方机构进行数据集抽检。

为保证标注的一致性及规范性，在标注之前，由相关团队对母语人士进行语言水平测试，筛选有经验的母语人士入组；对母语人士进行标准规范的相关培训，确保标注人员按照同一标准分词；校验小组与标注小组背对背，互不干扰，当二者标注不一致时，提交相关负责人研判，确保标注的准确率。

数据集整体构建流程如图 1 所示。

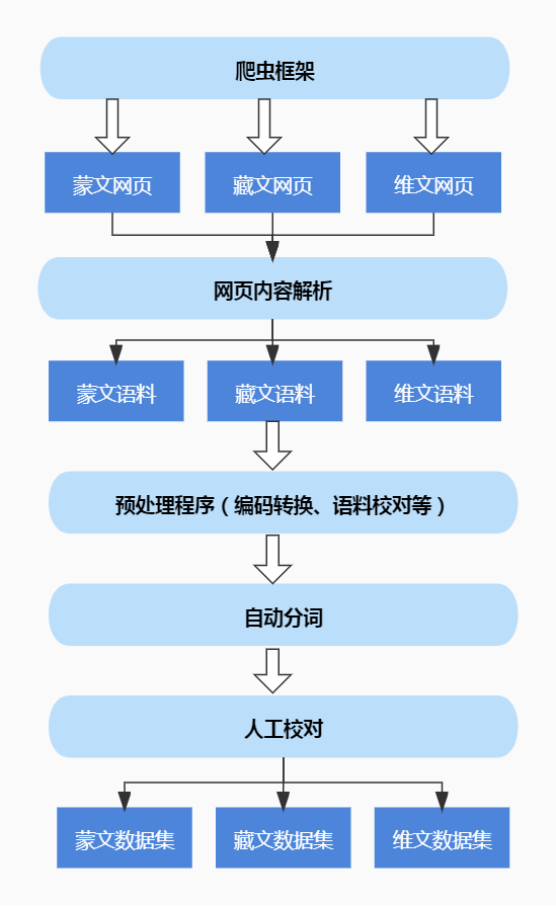

图 1 MLWS2021 构建流程

Figure 1 Construction procedures of MLWS2021

2 数据样本描述

MLWS2021 包含蒙、藏、维 3 个语种，评测对象是蒙古文、藏文、维吾尔文三个语种的自动分词核心技术。数据集在 MLWS2017 的基础上，由之前单一的新闻领域扩充到新闻、经济、法律、娱

乐等综合领域；数据规模也由之前的 3 万句，扩大到目前的 15.5 万句。MLWS2021 数据集中，蒙古文由中央民族大学提供，共计 6.5 万句；藏文由西藏大学提供，共计 2.5 万句；维吾尔文由清华大学提供，共计 6.5 万句。评测数据集概况如表 2 所示，标注样例见表 3。

表 2　MLWS2021 概况

Table 2　Overview of MLWS2021

序号	代号	数据集名称	语种	数目	大小	语料领域
1	MO	Mongolian	蒙古文	6.5 万	3.16 MB	综合领域
2	TI	Tibetan	藏文	2.5 万	1.52 MB	综合领域
3	UY	Uyghur	维吾尔文	6.5 万	5.12 MB	综合领域

表 3　标注样例

Table 3　Annotation samples

序号	语种	标注样例
1	Tibetan	[illegible]
2	Mongolian	[illegible]
3	Uyghur	[illegible]

3　数据质量控制和评估

为保证评测数据集质量的可靠性、稳定性，评测工作委员会启动数据集质量评估工作，成立数据集抽检小组，并进行抽检排期。对于蒙、藏、维 3 个语种的标注语料，按照 10%的比例抽取。其中，藏文以步长为 10，均匀抽取 10%，共抽取样本 2500 句；蒙古文和维吾尔文分别将原语料打散，随机抽取 10%，分别抽取 6500 句。将抽取的数据样本（15500 句）委托第三方机构进行人工校对。

经第三方机构评估，反馈结果为：藏文正确率为 98.27%，蒙古文正确率 99.12%，维吾尔文正确率 86.39%。由评估结果可知，数据集质量稳定，可满足少数民族语言分词技术评测的要求。同时我们将第三方机构人工校对后的结果收集起来，对原数据集的对应错误进行替换。

目前 MLWS 数据集已经连续服务两届少数民族语言分词技术评测，在构建过程中形成了一套完整的迭代流程。未来，数据集维护小组会不定时对数据集按比例进行抽取、校验、反馈、优化，同时借助多语种信息处理专委会，成立评测工作组，利用 MLWS2021 数据集开展相关的评测工作。我们相信，在不断的公开评测及多轮迭代下，数据集会不断完善，推动少数民族语言信息技术的发展。

4 数据价值

目前评测集已成功服务两届少数民族语言分词技术评测，版本也由最初的 MLWS2017 迭代为 MLWS2021，质量和稳定性得到进一步巩固。表 4 展示了第二届少数民族语言分词技术评测中，藏文分词技术在 MLWS2021 数据集上的部分参评结果。结果采用准确率（Precision）、召回率（Recall）、F1 值作为评测指标，按照 F1 值进行高低排序并排名。未来该评测数据集将面向社会，提供免费评测服务，逐步构建权威的少数民族语言分词技术评测平台，推动少数民族语言信息处理技术的发展。

表 4　藏文评测部分结果

Table 4　Results of Tibetan evaluation

名次	Method	Precision(%)	Recall(%)	F1(%)
1	TI_4_TWS_GF_restoreN	95.1387	94.5030	94.81980
2	TI_13_GF	94.7225	94.7140	94.7182
3	TI_5_3	94.9343	94.4902	94.7117
4	TI_5_2	94.9276	94.4857	94.7061
5	TI_5_4	94.9276	94.4857	94.7061
6	TI_5_1	94.8831	94.4279	94.6550

致　谢

感谢中央民族大学硕士生金波搭建第二届少数民族语言分词技术评测平台！感谢中央民族大学博士生特尼格尔、依斯马依力·艾肯木、周毛克等在数据集校验过程中的辛苦付出！

数据作者分工职责

赵小兵（1967—），女，北京市人，博士，教授，研究方向为计算语言学。主要承担工作：评测数据集构建流程的整体把控，人员协调与安排。

高璐（1989—），女，河北省邯郸市人，博士生，讲师，研究方向为计算语言学。主要承担工作：数据集质量监控。

高定国（1972—），男，四川省若尔盖县人，硕士，教授，研究方向为藏文信息处理。主要承担工作：藏文数据集的收集、整理、分词标注、校对工作。

包乌格德勒（1979—），男，内蒙古兴安盟人，博士，副教授，研究方向为自然语言处理、人工智能。主要承担工作：蒙古文评测语料的收集、整理、分词标注、校对工作。

米尔阿迪力江·麦麦提（1989—），男，北京市人，博士，研究方向是：自然语言处理、机器翻译、多语言信息处理。主要承担工作：维吾尔文数据搜集、组织标注团队、清洗数据、分配标注任务。

刘洋（1979—），男，北京市人，博士，教授，研究方向为：自然语言处理、深度学习、机器学习、机器翻译。主要承担工作：维吾尔文数据集整体质量把控。

才智杰（1970—），男，青海乐都人，博士，教授，研究方向为藏语自然语言处理。主要承担工作：藏文训练集和测试集的标注质量校对。
孙媛（1979—），女，北京市人，博士，副教授，研究方向为自然语言处理。主要承担工作：数据集整体流程把控。

参考文献

[1] MLWS2021 组委会. 第二届少数民族语言分词技术评测 MLWS2021 [EB/OL]. (2021-06-16) [2021-06-16]. https://nmlr.muc.edu.cn/info/1116/1651.htm. [Organizing Committee. The 2th ethnic Minority Language Word Segmentation technology evaluation (MLWS2021) [EB/OL]. (2021-06-16) [2021-06-16]. https://nmlr.muc.edu.cn/info/1116/1651.htm.]
[2] 西藏大学. 教育部、国家语委民族语言文字规范标准建设与信息化项目“大型藏文基础语料库建设”（MZ115-039）成果简介[R]. 2013-12. [Tibet University. Brief introduction to the achievements of the "Construction of large Tibetan basic corpus" (MZ115-039) of the national language standard construction and informatization project of the Ministry of Education and the State Language Commission.]
[3] 国家市场监督管理总局, 中国国家标准化管理委员会. 信息处理用藏文分词规范: GB/T 36452—2018[S]. 北京: 中国标准出版社, 2019. [General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. Specification on Tibetan segmentation for information processing: GB/T 36452—2018[S]. Beijing: Standards Press of China, 2019.]
[4] 国家质量监督检验检疫总局, 中国国家标准化管理委员会. 信息技术 信息处理用蒙古文词语标记: GB/T 26235—2010[S]. 北京: 中国标准出版社, 2011. [General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. Information technology—Mongolian word and expression marks for information processing: GB/T 26235—2010[S]. Beijing: Standards Press of China, 2011.]
[5] 国家标准化管理委员会. 信息处理用现代汉语分词规范: GB/T 13715—1992[S]. 北京: 中国标准出版社, 2004. [Standardization Administration of the People's Republic of China. Contemporary Chinese language word segmentation specification for information processing: GB/T 13715—1992[S]. Beijing: Standards Press of China, 2004.]

论文引用格式

赵小兵, 高璐, 高定国, 等. 少数民族语言分词技术评测数据集 MLWS2021[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-25). DOI: 10.11922/11-6035.csd.2021.0091.zh.

数据引用格式

赵小兵, 高璐, 高定国, 等. 少数民族语言分词技术评测数据集 MLWS2021[DS/OL]. Science Data Bank, 2022. (2022-01-28). DOI: 10.11922/sciencedb.j00001.00340.

A dataset of word segmentation technology evaluation for minority languages (MLWS2021)

ZHAO Xiaobing[1,2], GAO Lu[1,2], GAO Dingguo[3], BAO Wugedele[4], Mieradilijiang Maimaiti[5], LIU Yang[5], CAI Zhijie[6,7], SUN Yuan[1,2*]

1. Minzu University of China, Beijing 100081, P.R. China

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

3. Tibet University, Lhasa 850013, P.R. China

4. Hohhot Minzu College, Hohhot 015501, P.R. China

5. Tsinghua University, Beijing 100084, P.R. China

6. Qinghai Normal University, Xining 810016, P.R. China

7. State Key Laboratory of Tibetan intelligent information processing and Application, Xining 810016, P.R. China

*Email: tracy.yuan.sun@qq.com (Sun Yuan)

Abstract: According to the morphological rules and characteristics of Mongolian, Tibetan and Uyghur vocabulary, we formulated the evaluation standards for Mongolian, Tibetan and Uyghur word segmentation suitable for computer information processing, established the word segmentation and tagging corpus of Mongolian, Tibetan and Uyghur, and formed a standard evaluation corpus (MLWS2021), so as to solve the problems of automatic word segmentation, part of speech tagging, information retrieval. MLWS2021 contains 25,000 Tibetan sentences, 65,000 Mongolian sentences and 65,000 Uighur sentences. The evaluation dataset will provide free evaluation services available to the public, gradually build an authoritative minority language evaluation platform, and promote the development of minority language information processing technology.

Keywords: minority language; word segmentation; evaluation dataset; standard specification for word segmentation

Dataset Profile

Title	A dataset of word segmentation technology evaluation for minority languages (MLWS2021)
Data corresponding author	SUN Yuan (tracy.yuan.sun@qq.com)
Data authors	ZHAO Xiaobing, GAO Lu, GAO Dingguo, BAO Wugedele, Mieradilijiang Maimaiti, LIU Yang, CAI Zhijie, SUN Yuan
Time range	2017-2021
Data volume	9.8 MB
Data format	*.txt
Data service system	<http://www.doi.org/10.11922/sciencedb.j00001.00340>

Source of funding	Key Research Project of the National Language Commission (ZDI135-118)
Dataset composition	The dataset is composed of three data files: (1) Tibetan.zip is Tibetan word segmentation data, with a data volume of 25,000 sentences and a file size of 1.52 MB; (2) Mongolian.zip is Mongolian word segmentation data, with a data volume of 65,000 sentences and a file size of 3.16 MB; (3) Uyghur.zip is Uighur word segmentation data, with a data volume of 65,000 sentences and a file size of 5.12 MB.