专题 多语种智能信息处理

蒙汉机器翻译校正数据集

申影利[1,4]，包乌格德勒[2]，赵小兵[3,4*]

1. 中央民族大学中国少数民族语言文学学院，北京 100081
2. 呼和浩特民族学院，呼和浩特 010051
3. 中央民族大学信息工程学院，北京 100081
4. 国家语言资源监测与研究少数民族语言中心，北京 100081

摘要：机器翻译数据集的精确度对翻译模型的性能起决定性作用。传统蒙古语由于字符编码的特殊性，拼写错误十分普遍，网络开放资源字符编码准确性不足20%，这给其文本智能处理造成重大障碍。本文以第十七届全国机器翻译大会（CCMT 2021）蒙汉双语公开评测数据集作为原始语料，进行蒙文文本自动校正，构建面向机器翻译的高质量蒙汉句对校正数据集。在CWMT2017测试集上的实验结果表明，经过蒙文文本校正后的蒙汉双语平行句对在蒙汉、汉蒙两个方向上均优于原始评测数据的翻译效果，验证了蒙文校正文本的使用对提升下游自然语言处理任务性能的有效性及实用性。

关键词：机器翻译；传统蒙古文；文本校正；数据集

数据库（集）基本信息简介

数据库（集）名称	蒙汉机器翻译校正数据集
数据作者	申影利、包乌格德勒、赵小兵
数据通信作者	赵小兵（nmzxb_cn@163.com）
数据时间范围	2021年
地理区域	中国
数据量	5万句对蒙汉双语平行语料
数据格式	*.txt
数据服务系统网址	http://www.doi.org/10.11922/sciencedb.j00001.00354
基金项目	国家语委重点项目（ZDI135-118）；中央民族大学研究生科研实践项目（BZKY2021062）。
数据库（集）组成	本数据集共包含两部分。1. 校正后蒙文句子级文本：mn_correct.txt； 2. 中文句子级文本：zh.txt。

引 言

传统蒙古文（又称回鹘式蒙古文）是一种黏着型拼音文字，包含“名义字符”和“变形显现字符”。名义字符是蒙古文字符的独立体存在形式，显现字符则是字符居于词首、词中、词尾时由于变形而产生的不同显示形态[1]。蒙古文Unicode字符编码“以音编码”，其文本存在“形同音异”的现象，因而造成以国际标准编

ISSN 2096-2223
CN 11-6035/N

文献CSTR：
32001.14. 11-6035.csd.2021.0102.zh
文献DOI：
10.11922/11-6035.csd.2021.0102.zh
数据DOI：
10.11922/sciencedb.j00001.00354
文献分类：信息科学

收稿日期：2021-12-31
开放同评：2022-02-07
录用日期：2022-06-07
发表日期：2022-06-27

* 论文通信作者
赵小兵：nmzxb_cn@163.com

码存储的传统蒙古文文本常常错误地录入形状相同，但读音不同的变形显现字符。从字形上看，该单词是完全相同的，但其内部编码却是不同的，这种文本拼写错误对蒙古文信息处理研究造成重大障碍[2]。

蒙古文的文本校对工作是蒙古文信息处理的基础性工作之一。早期的校正工作依赖于人工校对，准确性高，但耗时耗力，效率低下。很多学者针对传统蒙古文的自动校对问题提出了可行的方案。华沙宝[3]依据蒙古文正字法规则开发 MHAHP 校对系统，受限于词典规模，该系统对动词构形附加成分、格附加成分之外的错误校对效果欠佳。苏传捷[4]等人利用机器翻译模型来构建拼写校对模型，在小规模文本上纠错后正确词比例达到 97.55%。蔡祝元[5]通过建立音节与真词混淆集，实现了对蒙古文非词错误与真词错误的查错与纠错。

本文以第十七届全国机器翻译大会（The 17th China Conference on Machine Translation , CCMT 2021，网址见 http://sc.cipsc.org.cn/mt/conference/2021/）蒙汉双语翻译项目公开评测数据集作为原始语料。根据分析，评测中提供的未经处理的蒙文语料存在诸多文本错误，这将严重影响机器翻译的性能。因此，本文开展蒙文自动校正工作，构建面向机器翻译任务的高质量蒙汉双语数据集。

1 数据采集和处理方法

1.1 原始语料数据收集

原始数据来自第十七届全国机器翻译大会机器翻译评测任务（CCMT 2021 MT Evaluation），CCMT 2021 蒙汉双语翻译任务的评测训练、开发语料数据的情况见表 1。

表 1 CCMT 2021 蒙汉双语翻译任务数据情况

Table 1 Data of CCMT 2021 Mongolian and Chinese bilingual translation task

数据资源		提供单位	领域	句对	总计
训练数据	IMU-CWMT2013	内蒙古大学	综合(政府文献，法律法规，日常对话，文学等)	104,790	262,458
	IMU-CWMT2015	内蒙古大学	综合	24,978	
	IIM-CWMT2015	中国科学院合肥智能机械研究所	新闻	1,682	
	IMU-corpus-CWMT2017	内蒙古大学	综合	100,001	
	ICT-MC-corpus-CWMT2017	中国科学院计算技术研究所	新闻	30,007	
	CWMT2018-TestSet-MC	内蒙古大学	综合	1,000	
开发数据	CWMT2019-TestSet-MC	内蒙古大学	综合	1,001	1001

1.2 数据处理

1.2.1 噪声数据清洗

在对蒙古文进行文本校正工作之前，我们发现原始评测集中蒙汉平行语料，存在源端、目标端

语言混杂的情况。例如，在 IMU-CWMT2015 文件夹中在源语言训练语料中存在大量的目标端语言句子，反之亦然，如图 1 所示。另外，训练数据中的重复句子会增加模型的负担，影响翻译效果，因此在对蒙汉双语句对中的蒙古文文本进行校正前，首先需要进行清洗、过滤蒙汉平行句对中的“噪声”数据。这样不但可以降低文本校正工作量，还能缓解低质量语料引起的翻译性能下降问题。针对以上情况，分别利用语种检测技术删除混杂语种、重复语句及空行，由实验最初设定的 262,458 句对训练语料得到经过清洗后的 248,438 句对，共删除 14,020 句对。

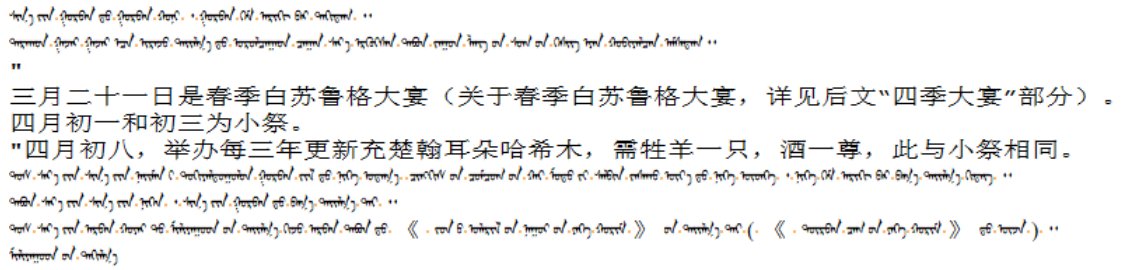

图 1　蒙汉语种混杂示例

Figure 1　Samples of mixed Mongolian and Chinese

1.2.2　蒙文文本校正

（一）数字、英文、中文符号、蒙古文非 Unicode 字符的转换处理

CCMT2021 提供的蒙古文语料为 Unicode 编码语料，因此，首先将蒙文语料中的数字、英文、符号及蒙古文非 Unicode 字符进行转换处理。

（二）文本校对

（1）通过正则表达式对部分字符进行修正

连续的变形控制符（\u180B，\u180C，\u180D）只保留第一个；对分写的附加成分进行统一处理；对\u182C（ᠬ）和\u182D（ᠭ）字符进行修正；对混用的阳性元音和阴性元音进行修正；对\u1836（ᠶ）字符进行修正。以上操作结束后把蒙古文语料转换为拉丁转写形式，对拉丁转写语料进行校对。

（2）通过词典和规则的方法对文本进行校正

采用基于词典和规则的方式对蒙古文进行自动校正，使用国家语言资源监测与研究少数民族语言中心（https://nmlr.muc.edu.cn/）构建整理的 20 万蒙古文的单词词典和构形附加成分词典。校正流程如图 2 所示。

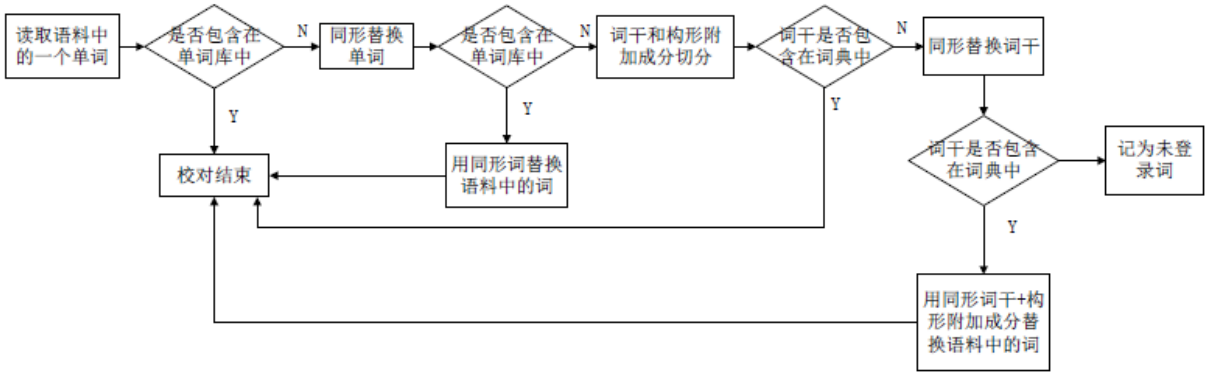

图 2　蒙古文文本校对流程

Figure 2　Procedure of Mongolian text proofreading

蒙文文本校正示例如表 2 所示。表 2 通过举例说明 CCMT 2021 蒙汉评测数据中原始蒙文文本的错误形式以及经过蒙文文本校正后的正确蒙文形式。从字形上看，错误蒙文文本、校正蒙文文本基本相同，但通过将二者进行相应的拉丁转写，就可以发现其内部编码的不同之处。在表 2 的例子中，我们将错误蒙文文本中的格错误部分进行标红，该类型是指蒙古文单词在连写附加成分时由于阴阳性或者其他构词方面的语法原因导致的错误；紫色及蓝色标记单词分别表示单音字、多音字错误。

表 2　CCMT 2021 蒙文文本错误及校正示例

Table 2　Samples of CCMT 2021 Mongolian text errors and correction

	中文源语言	扩大谈话范围，听取不同声音。
校正前	错误蒙文文本	ᠨᠢᠭᠡ ᠳᠤ ᠶᠠᠷᠢᠯᠴᠠᠬᠤ ᠬᠦᠷᠢᠶᠡ ᠡᠪᠴᠢᠶᠡ ᠪᠠᠨ ᠦᠷᠭᠡᠳᠵᠡᠶᠤ ᠠᠲᠠᠯᠢ ᠪᠤᠰᠤ ᠰᠠᠨᠠᠯ ᠰᠠᠭᠠᠨᠠᠭᠤᠯᠭ ᠠ ᠶᠢ ᠬᠤᠯᠢᠶᠠᠨ ᠠᠪᠴᠤ ᠪᠠᠶᠢᠨ ᠠ ᠃
	错误拉丁转写	**nige-dv , yarilcahv huriy_a hebciy_a ban urgedgejv , atali bvsv sanal sanagvlg_a-yi huliyan abcv bayin_a .**
校正后	校正蒙文文本	ᠨᠢᠭᠡ ᠳᠦ ᠶᠠᠷᠢᠯᠴᠠᠬᠤ ᠬᠦᠷᠢᠶᠡ ᠡᠪᠴᠢᠶᠡᠨ ᠦᠷᠭᠡᠳᠵᠡᠶᠦ ᠠᠲᠠᠯᠢ ᠪᠤᠰᠤ ᠰᠠᠨᠠᠯ ᠰᠠᠭᠠᠨᠠᠭᠤᠯᠭ ᠠ ᠶᠢ ᠬᠤᠯᠢᠶᠠᠨ ᠠᠪᠴᠤ ᠪᠠᠢᠨ ᠠ ᠃
	校正拉丁转写	**nige-d'u` , yarilcahv huriy_e hebciy_e-ben orgedgeju , adali bvsv sanal sanagvlg_a-y'i huliyen abcv bain_a .**

2　数据样本描述

本数据集为蒙汉机器翻译双语平行句对，共包含两部分：5 万句校正后蒙文文本，文件名称为：mn_correct.txt；5 万句中文文本，文件名称为：zh.txt。如下图 3 所示。

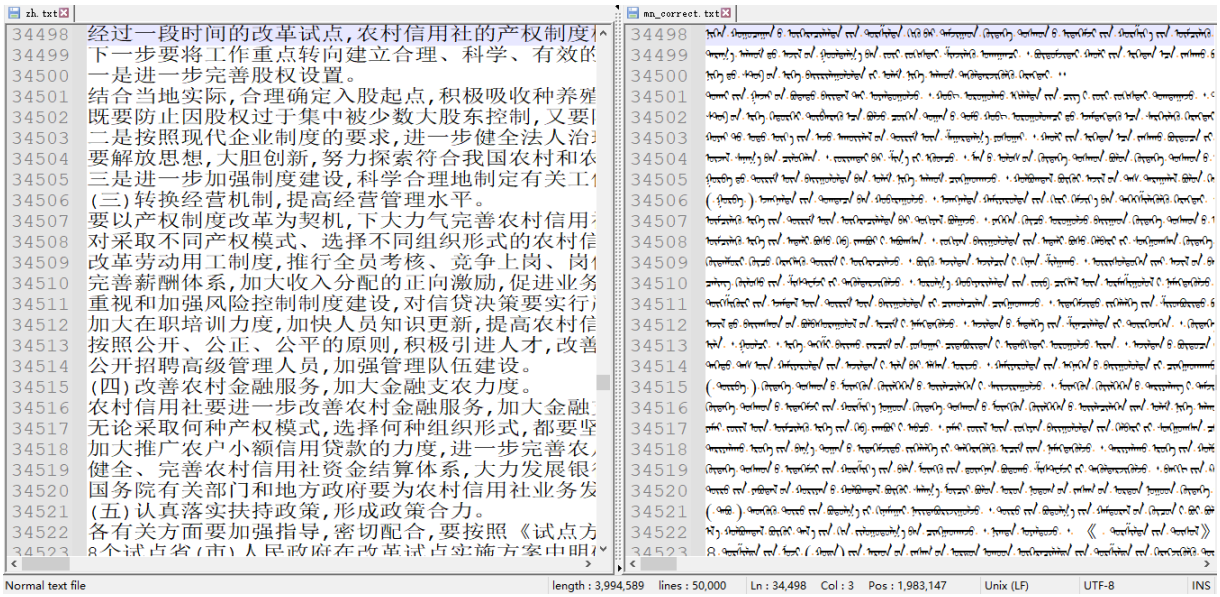

图 3　校正后蒙汉双语句对语料

Figure 3　Corrected Mongolian and Chinese bilingual sentence pairs

3　数据质量控制和评估

为验证上述蒙文文本校正工作是否对下游机器翻译质量有提升作用，我们使用全部经过蒙文校

正的CCMT2021蒙汉评测集及原始蒙汉评测集，在当前主流的神经机器翻译框架Transformer[6]上进行对比实验，使用BLEU[7]作为评测指标。由于CCMT2021主办方未提供蒙汉双语测试数据，我们选取CWMT2017提供的蒙汉双语测试集共1001句对。实验结果如表3所示，其中2021_dev、2017_test分别表示CCMT2021验证集和CWMT2017测试集。

表 3　蒙汉双向翻译模型测试结果

Table 3　Test results of Mongolian-Chinese bidirectional translation model

	任务	模型	BLEU/%	
			2021_dev	2017_test
校正前	蒙→汉	Transformer	42.0	35.8
	汉→蒙		19.4	21.6
校正后	蒙→汉	Transformer	**59.7**(+17.7)	**36.7**(+0.9)
	汉→蒙		**41.1**(+21.7)	**28.0**(+6.4)

从表3中的实验结果可以看出：经过蒙文校正后的语料在蒙汉双向翻译任务中都获得了最优性能。在蒙语→汉语翻译任务中，与校正前的蒙汉双语数据在2021_dev验证集和2017_test测试集上的BLEU值相比，分别提升了17.7和0.9个百分点。另一方面，汉语→蒙语翻译BLEU提升均优于蒙语→汉语翻译任务，校正后分别提升了21.7%、6.4%。这是因为蒙语相比于汉语构词形态更加复杂，当翻译为蒙语时，解码端很难避免语法错误，所以高质量蒙汉双语数据训练的模型对汉语→蒙语方向翻译效果的提升优于蒙语→汉语翻译方向。实验结果发现，使用蒙文文字校正后的蒙汉语料在双向翻译任务上均能够显著提升翻译效果。

4　数据使用价值

数据稀疏是低资源语言神经机器翻译面临的主要问题，针对蒙古文信息处理研究，蒙古文高质量语料的获取一直是亟待解决的难题。本文在蒙汉机器翻译评测数据集的基础上，进行蒙古文文本校正工作，实验验证发现，经过文本校正后的蒙汉双语数据集，在下游机器翻译任务中的翻译质量有明显提升。本数据集除机器翻译任务外，还可用于文本校正、命名实体识别、信息检索等蒙古文自然语言处理工作。

致　谢

感谢全国机器翻译大会主办机构提供的宝贵原始数据资源，感谢对本数据集进行蒙文校正工作的蒙语研究专家。

数据作者分工职责

申影利（1994—），女，安徽亳州人，在读博士研究生，研究方向为自然语言处理、机器翻译。主要承担工作：数据筛选、处理、加工，数据集生成，论文的撰写。

包乌格德勒（1979—），男，内蒙古兴安盟人，博士，副教授，研究方向为计算语言学、蒙古文信

息处理。主要承担工作：数据集设计和整理，数据校准。

赵小兵（1967—），女，内蒙古呼和浩特人，博士，博士生导师，研究方向为自然语言处理、舆情分析等。主要承担工作：研究思路设计与论文撰写指导。

参考文献

[1] 金良，林民. 三种蒙古文编码之间的差异性研究[J]. 内蒙古师范大学学报(自然科学汉文版), 2013,42(2): 225-227.[JIN L, LIN M. On the differences among three essential Mongolian codings standard[J]. Journal of Inner Mongolia Normal University(Natural Science Edition), 2013, 42(2): 225-227.]

[2] 廉冰. 基于有限状态自动机的蒙古文同形词校对方法的研究[D]. 呼和浩特:内蒙古大学, 2014. [LIAN B. Research on proofreading algorithm of Mongolian homograph based on finite state automata[D]. Hohhot: Inner Mongolia University, 2014.]

[3] 华沙宝. 现代蒙古文自动校对系统: MHAHP[J]. 内蒙古大学学报(人文社会科学版), 1997, 29(4): 49-53. DOI:10.13484/j.cnki.ndxbzsb.1997.04.007. [HUA S B. Modern Mongolian automatic proofreading system[J]. Journal of Inner Mongolia University(Philosophy and Social Sciences), 1997,29(4): 49-53. DOI:10.13484/j.cnki.ndxbzsb.1997.04.007.]

[4] 苏传捷, 侯宏旭, 杨萍, 等. 基于统计翻译框架的蒙古文自动拼写校对方法[J]. 中文信息学报, 2013, 27(6): 175–179. DOI:10.3969/j.issn.1003-0077.2013.06.026. [SU C J, HOU H X, YANG P, et al. A spelling correction method for traditional Mongolian based on statistical translation framework[J]. Journal of Chinese Information Processing, 2013, 27(6): 175–179. DOI:10.3969/j.issn.1003-0077.2013.06.026.]

[5] 蔡祝元. 基于蒙古文音节分析的文本校对方法研究[D]. 呼和浩特: 内蒙古大学, 2019. [CAI Y Z. Research on text proofreading method based on the analysis of the Mongolian syllable[D]. Hohhot: Inner Mongolia University, 2019.]

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]/Proceedings of the Thirty-first Conference on Neural Information Processing Systems. USA, California. Cambridge, MA, USA: Massachusetts Institute of Technology Press, 2017: 5998-6008.

[7] POST M. A call for clarity in reporting BLEU scores[C]/Proceedings of the Third Conference on Machine Translation: Research Papers. Belgium, Brussels. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 186–191. DOI:10.18653/v1/w18-6319.

论文引用格式

申影利, 包乌格德勒, 赵小兵. 蒙汉机器翻译校正数据集[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-07). DOI: 10.11922/11-6035.csd.2021.0102.zh.

数据引用格式

申影利, 包乌格德勒, 赵小兵. 蒙汉机器翻译校正数据集[DS/OL]. Science Data Bank, 2022. (2022-02-07). DOI: 10.11922/sciencedb.j00001.00354.

A dataset of Mongolian-Chinese machine translation correction

SHEN Yingli[1,4], BAO Wugedele[2], ZHAO Xiaobing[3,4*]

1. School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, P. R. China

2. Hohhot Minzu College, Hohhot 010051, P. R. China

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

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

*Email: nmzxb_cn@163.com

Abstract: The accuracy of machine translation datasets plays a decisive role in the performance of translation models. Due to the particularity of character encoding in traditional Mongolian, spelling errors are very common, and the accuracy of character encoding of open resources on the Internet is less than 20%, which poses a major obstacle to intelligent text processing. In this paper, we used the Mongolian-Chinese bilingual public evaluation dataset of the 17th China Conference on Machine Translation (CCMT 2021) as the original corpus to complete automatic Mongolian correction, and constructed a high-quality Mongolian-Chinese sentence pair correction dataset for machine translation. The experimental results on the CWMT2017 test set show that the Mongolian-Chinese bilingual parallel sentence pair after the Mongolian text correction is better than the translation effect of the original evaluation data in both Mongolian->Chinese and Chinese->Mongolian directions, which verifies the effectiveness and practicability of the Mongolian corrected text for improving the performance of downstream natural language processing tasks.

Keywords: machine translation; traditional Mongolian; text correction; dataset

Dataset Profile

Title	A dataset of Mongolian-Chinese machine translation correction
Data corresponding author	ZHAO Xiaobing (nmzxb_cn@163.com)
Data authors	SHEN Yingli, BAO Wugedele, ZHAO Xiaobing
Time range	2021
Geographical scope	China
Data volume	4.71 MB
Data format	*.txt
Data service system	<http://www.doi.org/10.11922/sciencedb.j00001.00354>
Sources of funding	National Language Commission Key Project (ZDI135-118); Graduate Research and Practice Projects of Minzu University of China (BZKY2021062).
Dataset composition	This dataset consists of two parts. One is thecorrected Mongolian sentence-level text: mn_correct.txt. The other is Chinese sentence-level text: zh.txt.