专题 多语种智能信息处理

蒙古文日常问答语料数据集

特日格勒呼[1]，王斯日古楞[1*]，韩永顺[1]，爱丽雅[1]，娜何雅[1]

1. 内蒙古师范大学，呼和浩特市 010022

摘要：蒙古文自动问答研究发展缓慢，其中问答语料的稀缺是重要的原因之一。本研究通过对现有中文问答语料进行收集后通过规则筛选、汉蒙翻译、人工校正构建了5万对蒙古文问答语料。通过自动评价发现，该语料的问句和答复句具有较好的多样性，人工评价结果显示97%的语料符合日常问答逻辑。该语料范围主要是开放领域的日常对话，可应用在端到端的一问一答形式问答模型中，在蒙古文自动问答的研究中具有重要的使用价值。

关键词：蒙古文；问答语料；语料库构建；语料校正

www.csdata.org

ISSN 2096-2223

CN 11-6035/N

文献 CSTR：

32001.14.11-6035.csd.2021.0095.zh

文献 DOI：

10.11922/11-6035.csd.2021.0095.zh

数据 DOI：

10.11922/sciencedb.j00001.00347

文献分类：信息科学

收稿日期：2021-12-24

开放同评：2022-01-28

录用日期：2022-06-15

发表日期：2022-06-30

* 论文通信作者

王斯日古楞：siriguleng@inmu.edu.cn

数据库（集）基本信息简介

数据库（集）名称	蒙古文日常问答语料数据集
数据作者	特日格勒呼、王斯日古楞
数据通信作者	王斯日古楞（siriguleng@inmu.edu.cn）
数据时间范围	2019–2021年
地理区域	世界各地
数据量	4.47 MB
数据格式	*.xlsx
数据服务系统网址	http://www.doi.org/10.11922/sciencedb.j00001.00347
基金项目	内蒙古自治区科技计划项目(2021GG0139)；国家自然科学基金资助项目(61762072)。
数据库（集）组成	数据集共包括1个数据文件，表中有2列数据，分别是蒙古文问句和蒙古文答句，共计100000句。

引 言

问答系统是人工智能领域的重要研究方向，它作为人与机器交互的沟通桥梁，具有重大的研究意义和发展前景。在当代老龄化严重的社会背景下，智能问答系统可以陪伴老人，同时也能减轻年轻人的工作压力和困扰。问答系统主要分为任务型和非任务型，其中非任务型问答系统是面向开放领域，与用户进行闲聊对话，而任务型问答系统是为了完成用户提出的某个特定任务工作。

随着互联网数据的暴涨、深度学习技术的崛起以及硬件设备性能的提高，越来越多的智能交互设备融入到我们日常生活中。但是主流的产品或模型主要以中文、英文等高资源语言为主，而蒙古文问答系统发展缓慢。蒙古文信息处理研究中，机器翻译、语音识别、语音合成等方向的研究已经取得了较好的成果。但是，

蒙古文自动问答领域的研究处于起步阶段，内蒙古大学常泽晖[1]研究了面向开放领域的蒙古语语音交互系统，其中问答系统部分是在约 2 万条问答语料上使用序列到序列（Sequence to Sequence，Seq2Seq）框架实现的。谭铭言[2]利用构建的蒙古文知识图谱以及命名实体识别系统和关系抽取系统，搭建了面向旅游领域的蒙古文问答系统。王光义[3]构建了 32156 条纪检监察领域的蒙古文问答语料，并通过问句意图识别和问答匹配两个模块实现了蒙古文问答系统。

问答语料资源的稀缺是影响蒙古文自动问答技术发展的重要因素之一。因此，本文通过获取开源中文问答语料库并通过筛选、翻译、校正等方法构建了 5 万句对蒙古文问答语料，相比，其他蒙古文自动问答研究使用的语料具有更大的数据量和更贴切的内容。蒙古文问答语料库的建设可以有效促进蒙古文信息处理的研究，对促进民族之间的交流与合作具有十分重要的意义。

1 数据采集和处理方法

蒙古文问答语料来源是中文公开数据集，通过对其进行规则筛选、汉蒙机器翻译、人工校正等步骤构建了蒙古文问答语料，其构建流程如图 1 所示。

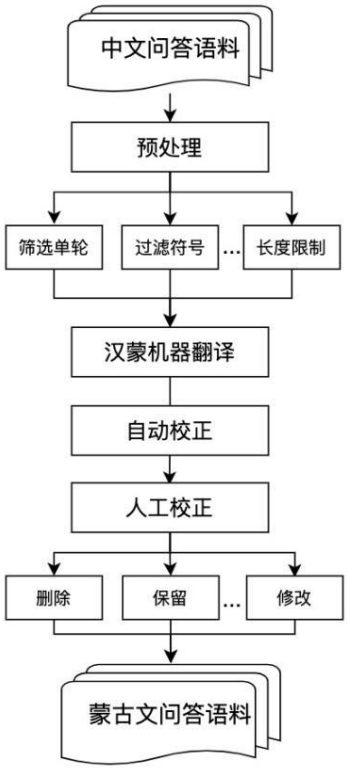

图 1　蒙古文问答语料构建流程

Figure 1　Flow chart of Mongolian question and answer corpus construction

1.1 数据采集方法

语料库的质量和规模对问答系统的研究发展有直接的影响，因此语料的正确选择和处理非常重要。关于蒙古文问答的研究较少，更没有公开可用的蒙古文问答语料库。

本研究首要任务是构建适当规模的蒙古文问答语料库。使用的问答语料来源是 2020 年清华大学公开的中文问答数据集 LCCC[4]中的 LCCC-base。该数据集的原始对话数据来自微博对话，这

一数据过滤流程包括一系列手工规则以及若干基于机器学习算法所构建的分类器，已经对脏字脏词、特殊字符、颜表情、语法不通的语句、上下文不相关的对话等噪声进行了初步过滤。与“小黄鸡”“青云”等公开的同类数据集相比，具有更好的内容质量和更大的数量。

1.2 数据预处理

源语料是由单轮和多轮问答交替组成的 json 文件，语料样例如表 1 所示。首先，从源语料中筛选了 100 万对单轮问答语料，为了获取更高质量的问答对，通过编写以“？”“吗”“么”“嘛”“了”等常用的疑问句结尾字以及多种自定义的规则设定为约束条件进行筛选获得了 10 万对问答语料。

表 1 中文原始语料样例

Table 1 Samples of original Chinese corpus

```
[
"我 饿 了 。",
"去 相 机 家 里 吃 … …",
"相 机 今 年 木 有 回 去 T . T"
],
[
"网 络 大 实 话 里 说 的 是 也 许 你 能 在 网 络 里 找 到 你 想 要 的 友 情 但 永 远 不 会 找 到 你 想 要 的 爱 情",
"你 过 来 我 们 什 么 关 系"
],
[
"老 铁 家 好 吃 贾 三 不 好 吃",
"我 不 挑 食"
],
[
"你 有 翘 臀 啊 ！ ！ ！ ！ 你 的 脸 还 不 够 小 啊 ？ ？ ？ ？ ？",
"死 鱼 皮 真 会 安 慰 人 那 不 是 翘 臀 是 肥 肉 不 ！ 是 赘 肉 ！"
],
```

通过分析发现，句子仍然包含连续重复多次的问号、感叹号、逗号和含有一些“\、'、～、「」”等不规则符号等，因此把句末和句中的问句、感叹号和逗号替换为单个符号，对不规则符号和句首的符号进行过滤，并且去除了长度超过 100 个字的句子。表 2 列举了几种代表性的语料清洗样例。

表 2 中文原始语料清洗样例

Table 2 Sorting-out samples of original Chinese corpus

处理前	处理后
是 呀 ' ' ' 能 吃 能 睡！！！	是 呀 ， 能 吃 能 睡！
很 社会 ， ， ， 我 都 不 喝酒	很 社会 ， 我 都 不 喝酒

处理前	处理后
情人节 快乐 啦 ～～～～	情人节 快乐 啦
我 怀疑 你 大脑 有 问题 2333333	我 怀疑 你 大脑 有 问题
正在 追 「 暖 爱 」	正在 追 暖 爱
【 旁友 能 把 你们 大陆 表情 包 发给 我 吗 】	旁友 能 把 你们 大陆 表情 包 发给 我 吗
不 … 兴 … 奋 … 啊 …	不兴 奋 啊
你 知道 为了 秀 ， 我 手肘 磕青 了 么 ····	你 知道 为了 秀 ， 我 手肘 磕青 了 么
我 小腿 真的 瘦 了 ～ ⊙ ▽ ⊙	我 小腿 真的 瘦 了
？ ？ ？ ？ 多久 生 的 二胎 ？ ？ ？	多久 生 的 二胎 ？

1.3 汉蒙机器翻译与语料校正

将预处理后的中文问答语料经过本实验室现有的汉蒙机器翻译模型从中文翻译成蒙古文。由于中文问答语料内容存在一些噪声，以及翻译后的蒙古文译文中有语序错误和错别字等问题，最后，我们对蒙古文语料进行校正。

本文对汉蒙机器翻译过后的蒙古文问答语料内容采用了自动校正和人工校对相结合的方法。自动校正是针对蒙古文语料中存在的编码错误和名词格附加成分使用不当等拼写错误，使用自动校对工具进行修正。

人工校正是一项费时费力的工作，同时，我们开发了一款语料管理及修改的平台，该平台支持多人在线校正双语平行语料，并且可以自由地分配任务，也支持实时监督和统计任务进度，可以提高工作效率，平台展示如图 2 所示。

图 2　蒙古文问答语料校正平台

Figure 2　The correction platform of Mongolian question and answer corpus

校正平台将修改的内容展示成四列，中文问答句为修改蒙古文问答句提供参考。通过平台可以对语料进行一一校正，校正的主要工作内容有：

（1）抛弃中文问题和答案不匹配、质量较差、句子逻辑有误的句子，相反保留质量很好的蒙古文问答对，不需要其进行改动。

（2）对中文问答语料质量较好，但翻译后的蒙古文句子不通顺、不完整情况进行补充修正，构成符合蒙古文语法的句子。校正过程中遇到的部分典型例子如表 3 所示。

表 3　蒙古文问答语料校正样例

Table 3　Correction Samples of Mongolian question and answer corpus

中文问句	中文答复句	蒙古文问句	蒙古文答复句	蒙古文问句（修改）	蒙古文答复句（修改）
你在干什么	打球	ᠴᠢ ᠶᠠᠮᠠᠷ ᠨᠢᠭᠡ ᠪᠠᠢᠳᠠᠯ ᠪᠤᠢ	ᠪᠤᠴᠤᠭᠤᠷ ᠮᠢᠨᠢ ᠪᠠᠢᠨ᠎ᠠ	ᠴᠢ ᠶᠠᠮᠠᠷ ᠨᠢᠭᠡ ᠪᠠᠢᠳᠠᠯ ᠪᠤᠢ	ᠪᠤᠴᠤᠭᠤᠷ ᠮᠢᠨᠢ ᠪᠠᠢᠨ᠎ᠠ．
这饮料好喝吗	不知道我没喝	ᠡᠨᠡ ᠤᠭᠠᠭᠠᠯᠠ ᠮᠠᠭᠠ ᠤᠭᠤ	ᠪᠢ ᠪᠠᠢᠬᠤ ᠮᠡᠳᠡᠬᠦᠯ ᠦᠬᠡᠢ ᠴᠢ ᠮᠠᠭᠠ ᠮᠠᠨᠠ	ᠡᠨᠡ ᠤᠭᠠᠭᠠᠯᠠ ᠮᠠᠭᠠ ᠤᠭᠤ	ᠮᠠᠭᠠ ᠮᠠᠨᠠ ᠪᠢ ᠪᠠᠢᠬᠤ ᠮᠡᠳᠡᠬᠦᠯ ᠦᠬᠡᠢ
回新疆了？	记得找我玩	《 ᠮᠠᠨᠠᠢ ᠭᠠᠵᠠᠷ ᠰᠢᠨᠵᠢᠶᠡᠩ ᠪᠤᠴᠤᠵᠤ ᠪᠤᠢ ᠤᠤ? 》 ᠤᠭᠤ	ᠮᠠᠭᠠ ᠴᠢ ᠮᠠᠨᠠ ᠵᠤᠭᠰᠠᠭᠠᠵᠤ ᠪᠤ ᠮᠠᠭᠠᠳᠠᠯ	ᠰᠢᠨᠵᠢᠶᠡᠩ ᠭᠠᠵᠠᠷ ᠰᠢᠨᠵᠢᠶᠡᠩ ᠪᠤᠴᠤᠵᠤ ᠪᠤᠢ?	ᠮᠠᠭᠠ ᠴᠢ ᠮᠠᠨᠠ ᠵᠤᠭᠰᠠᠭᠠᠵᠤ ᠪᠤ ᠮᠠᠭᠠᠳᠠᠯ

表中蓝色字体表示保持原文，红色字体表示对原文进行了修改。

第一行中，现在将来事态形动词“ᠨᠠ”“ᠪᠠ”，以该形动词结尾的词一般不能当作句子结尾。所以应当根据问句的事态和人称对句子进行修改，补充助动词构成完整正确的蒙古文句子。

第二行中，由于中文问答语料缺少停顿标点符号，导致翻译的蒙古文句子含义发生了变化。

第三行中，中文源句中的句子是祈使句或者感叹句，导致翻译后的蒙古文句子含有“《》”“ᠤᠤ”、“ᠤᠤ”等词的情况。

校正后的语料由问题和答案组成，属于开放领域的单轮日常问答语料。

2　数据样本描述

本文公开的语料包含通过人工校正后的蒙古文问答语料，由 5 万句对一一对应的问题和答复组成，词表大小为 20927 字，问答句平均长度为 6.94 个字。图 3 展示了 10 行蒙古文问答语料样例，第一列是蒙古文问句，第二列为所对应的回复句。图 4 根据问答句的长度分布进行了绘制。

1	ᠠᠰᠠᠭᠤᠭᠴᠢ：	ᠬᠠᠷᠢᠭᠤᠯᠭᠴᠢ：
2	ᠠᠭᠤᠤ ᠠᠴᠠ ᠬᠠᠪᠤᠰᠤᠷᠤᠯ ᠤᠤ ？	ᠠᠮᠠᠰᠢᠭᠤ ᠪᠠ ，ᠰᠠᠨᠠᠭᠠᠴᠠᠯ ᠠᠴᠠ ᠠᠭᠤᠯᠠᠮ ᠠᠭᠤᠯᠬᠤ ᠪᠣᠭᠣᠵᠤ ．．
3	ᠨᠢᠭᠡ ᠬᠣᠭᠣᠷ ᠤᠭᠠᠭᠠᠯᠠ ᠰᠠᠨᠠᠭᠠᠴᠠᠯ ᠠᠭᠤᠯᠬᠤ ᠪᠤᠢ ？	ᠤᠭ ᠤᠭ ᠬᠤᠢᠨ ᠬᠠᠶᠠ ᠨᠠᠮᠠᠭᠤᠷ ᠤᠭᠠᠭᠠᠯᠠ ，ᠠᠭᠤᠯᠬᠤ ᠤᠭᠠᠭᠠᠯᠠ ᠪᠤᠢ ᠠᠭᠤᠯᠬᠤ ᠬᠣᠭᠣᠷ ᠤᠭᠠᠭᠠᠯᠠ ，ᠠᠭᠤᠯᠬᠤ ᠬᠣᠭᠣᠷ ᠤᠭᠠᠭᠠᠯᠠ ᠪᠤᠢ ᠠᠭᠤᠯᠬᠤ ᠤᠭᠠᠭᠠᠯᠠ
4	ᠪᠢ ᠰᠠᠶᠢᠨ ᠤᠴᠤᠷ ᠬᠣᠯᠪᠣᠭᠳᠠᠭᠤᠯᠬᠤ ？	ᠨᠢᠭᠡ ᠬᠤᠷᠢᠮ ᠲᠣᠭᠣᠭᠣᠷ ᠲᠤᠯ ᠪᠢᠴᠢᠭᠴᠢᠶᠡᠨ ．．
5	ᠬᠠᠭᠤᠭᠤᠯ ᠠᠭᠤᠯᠬᠤ ᠤᠭᠠᠭᠠᠯᠠ ᠤᠭᠤ ？	ᠬᠤᠭᠤᠯᠤᠭᠴᠢ ᠬᠡᠰᠡᠭ ᠪᠠᠭᠠᠰᠢ ᠤᠭᠤ ？
6	ᠪᠠᠶᠠᠷᠢᠯᠠᠯ ᠠᠮᠣᠭᠠᠨᠲᠠᠢ ᠤᠤ ？	ᠪᠠᠶᠠᠷᠢᠯᠠᠯ ᠠᠮᠣᠭᠠᠨᠲᠠᠢ ᠰᠠᠨᠠᠭᠠᠴᠠᠯ ᠭᠤ ᠲᠠᠨᠢᠮᠳᠤ ᠪᠤᠢ ᠪᠠᠭᠤᠯᠤᠨ ᠬᠡᠯᠡᠭ ．．
7	ᠬᠣᠨᠣᠭᠲᠤᠭᠤᠷ ᠰᠠᠶᠢᠨ ᠭᠡᠵᠦᠦᠰᠡᠨ ᠪᠠᠭᠠᠰᠢᠮᠠᠭ ᠪᠤᠢ ？	ᠬᠤᠭᠤᠷ ᠪᠠᠭᠤᠷ ᠬᠡᠭᠡᠭ ᠪᠠᠭᠠᠰᠢ ．
8	ᠮᠣᠩᠭᠣᠯ ᠬᠡᠯᠡᠨᠳᠡ ᠭᠠᠷᠴᠠᠭ ᠲᠠᠢ ᠬᠤᠪᠢᠴᠠᠭᠠᠷ ᠬᠡᠯᠡᠭᠴᠢᠯᠡᠬᠦ ᠪᠤᠯᠤᠨ ᠤᠤ ？	ᠠᠭᠤᠭᠤᠯᠤᠭ ᠬᠠᠮᠲᠤᠭᠤᠯᠤᠭ ，ᠬᠡᠯᠡᠭᠴᠢᠯᠡᠭ ᠲᠣᠭᠣᠭ ，ᠬᠤᠪᠢᠴᠠᠭᠠᠷ ᠤᠭ ᠪᠤ ᠭᠡᠵᠦᠦᠰᠡᠨ ᠪᠠᠭᠠᠰᠢᠮᠠᠭ ．．
9	ᠤᠭ ᠬᠠᠭᠠᠨ ᠪᠤ ᠵᠢᠶᠠᠨ ᠬᠣᠷᠢᠯᠠᠭᠤᠯᠬᠤ ？	ᠬᠣᠷᠢᠯᠠᠭᠤᠯᠬᠤ ᠬᠣᠷᠢᠯᠠᠭᠤᠯᠬᠤ ，ᠪᠢ ᠪᠠᠭᠤᠴᠢ ᠬᠣᠷᠢᠯᠠᠭᠤᠯᠬᠤ ．．
10	ᠤᠭ ᠲᠣ ᠮᠠᠰᠢ ᠬᠡᠴᠦ ᠬᠡᠭᠴᠢᠭᠡ ᠤᠭ ᠪᠠᠶᠢᠴᠠᠭᠠᠭᠤᠯᠬᠤ ？	ᠵᠠᠭᠤ ᠲᠠᠭ ᠪᠤ ᠪᠠᠶᠢᠴᠠᠭᠠᠭᠤᠯ ᠠᠭᠤᠯ ！
11	ᠰᠤᠪᠤᠭᠴᠠᠢᠪᠠᠭ ᠵᠤᠭᠰᠤᠨᠠ ᠤᠤ ？	ᠬᠣᠭᠣᠷ ᠪᠤ ᠵᠠᠭᠠ ，ᠰᠤᠪᠤᠭᠴᠠᠢ ᠬᠡᠪᠡᠷ ᠬᠡᠪᠡᠲᠦᠢ ．．

图 3　蒙古文问答语料样例展示

Figure 3　Sample display of Mongolian question and answer corpus

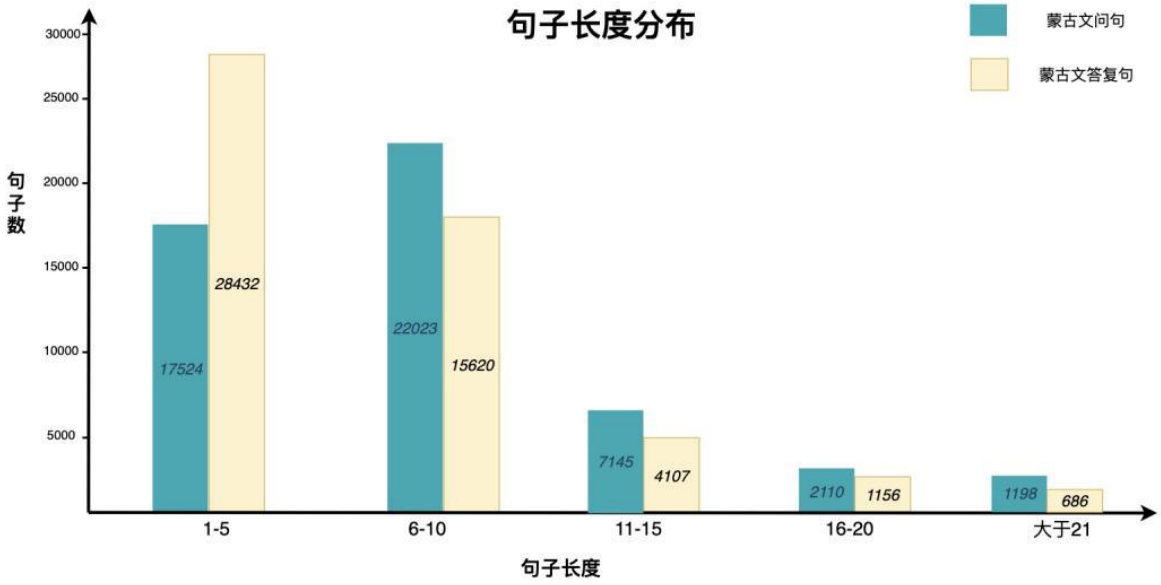

图 4　句子长度分布图

Figure 4　Sentence length distribution

从图 4 中可以看蒙古文问句长度主要分布在 6–10 字，而大量答复句长度在 2–5 字之间。

通过统计分析问答语料中的词频，并且去除符号、格附加成分、连词后对主要出现的词使用 WordArt（https://wordart.com/）平台进行了词云绘制，如图 5 所示。

图 5　问答语料中词云展示图

Figure 5　Word cloud display in question and answer corpus

该词云根据问答语料中的词频高低绘制而成，词频越高显示得越大。从图 5 中可以看出，疑问代词出现的概率较高，例如“ ”译为“为什么”、“ ”译为“怎么”、“ ”译为“什么”等。还有一些生活中交流的常用名词，例如“ ”译为“学校”、“ ”译为“朋友”、“ ”译为“饭”

等。说明符合日常对话逻辑。

3 数据质量控制和评估

为了验证问答语料的质量，我们使用了人工评价和自动评价两种方式。

首先通过 Distinct-N[5]对构建的 5 万句对语料进行了评价，Distinct-N 主要衡量问答系统中句子的多样性，避免出现一些“我不知道”等万能回复。Distinct-1、Distinct-2 分别由不同的一元词和二元词数量与生成单词总数相除得到，蒙古文问答语料多样性评测结果如表 4 所示，指标越高表示句子越好。

表 4　蒙古文问答句多样性评测

Table 4　Diversity evaluation of Mongolian question and answer corpus

蒙古文语料	Distinct-1	Distinct-2
问句	0.976	0.8573
回复句	0.977	0.7054

自动评价只能从客观的层面对语料进行评估，当数据量较大的情况下比较合适，可以考虑全局信息，但是无法从语义层面进行理解。因此，本文采用了三分制的人工评分方法，从语料库中随机抽样 500 个问答对，并邀请 5 位具有语料校正经历的人员对这些问答对进行打分，主要针对问答和答案的内容贴切度、句子流畅性、以及是否存在蒙古文语法错误等。打分标准如表 5 所示。

表 5　蒙古文问答语料打分标准

Table 5　Grading standards for Mongolian question and answer corpus

分数	回答标准
1	问题与回答内容不匹配，具有语法错误或错别字
2	问题与回答符合逻辑，但是提供的价值不高
3	问题与答案相关性很高、句子流畅

表 6 展示了蒙古文问答语料质量评价结果。

表 6　人工评价结果

Table 6　Result of manual evaluation

分数	得分
1	3%
2	20.6%
3	76.4%

评价结果显示，问题与回答内容不匹配，含有语法错误或错别字的问答对只占 3%；由于中文语料质量的限制，20.6%的回答提供的价值不高，但并没有逻辑错误；而剩余 76.4%的问答对句子流畅问题与答案相关性较高。评价结果证明了问答语料的质量以及有效性。

4 数据价值

目前，国内未见公开可用的蒙古文问答语料，本数据集的公开是蒙古文自动问答领域中的一次重要尝试，可以为蒙古文问答系统的发展提供重要的数据支撑，还可以用于训练生成式蒙古文问答模型、微调预训练模型和迁移学习等具体任务，从而获得更好的效果。本数据集具有广泛的科研价值和较高的社会应用价值。

同时，希望同行能够分享更多蒙古文问答数据集，促进蒙古文自动问答研究的开放与发展。

5 数据使用方法和建议

本数据集以 xlsx 文件为存储格式，使用者可以根据自身需求将文件改为 txt 或者所需要的格式进行使用。任何组织和个人可以以非商业目的使用本数据集。

数据作者分工职责

特日格勒呼（1997—），男，内蒙古赤峰人，研究生在读，研究方向为自然语言信息处理、问答系统。主要担任工作：数据采集与管理，平台搭建与文章撰写。

王斯日古楞（1970—），女，内蒙古呼和浩特人，博士，教授，研究方向为自然语言信息处理、机器翻译。主要担任工作：提供研究思路、指导论文框架、修改文章内容。

韩永顺（1997—），男，内蒙古呼伦贝尔人，研究生在读，研究方向为自然语言信息处理。主要担任工作：数据采集与校正处理。

爱丽雅（1998—），女，内蒙古呼伦贝尔人，研究生在读，研究方向为自然语言信息处理。主要担任工作：数据采集与校正处理。

娜何雅（1998—），女，内蒙古通辽人，研究生在读，研究方向为自然语言信息处理。主要担任工作：数据采集与校正处理。

参考文献

[1] 常泽晖. 面向智能机器人的蒙古语语音交互系统的研发[D]. 呼和浩特市: 内蒙古大学, 2019. [CHANG Z H. Research and development of Mongolian speech interaction System for intelligent robot [D]. Hohhot: Inner Mongolia University, 2019.]

[2] 谭铭言. 面向旅游领域的蒙古文自动问答系统研究[D]. 呼和浩特市: 内蒙古大学, 2020. [TAN M Y. Research on Mongolian Automatic Question Answering System for Tourism [D]. Hohhot: Inner Mongolia University, 2020.]

[3] 王广义. 面向纪检监察领域的蒙古文自动问答系统研究[D]. 呼和浩特市: 内蒙古大学,2021. [WANG G Y. Research on Mongolian Automatic Question answering System for Discipline Inspection and Supervision [D]. Hohhot: Inner Mongolia University,2021.]

[4] WANG Y, KE P, ZHENG Y, et al. A Large-Scale Chinese Short-Text Conversation Dataset[J].international conference natural language processing,2020: 91-103.

[5] LI J W, GALLEY M, BROCKETT C, et al. A diversity-promoting objective function for neural conversation models[J]. Computer Science, 2016: 110-119.

论文引用格式

特日格勒呼，王斯日古楞，韩永顺，等. 蒙古文日常问答语料数据集[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-23). DOI: 10.11922/11-6035.csd.2021.0095.zh.

数据引用格式

特日格勒呼，王斯日古楞. 蒙古文日常问答语料数据集[DS/OL]. Science Data Bank, 2022. (2022-01-28). DOI: 10.11922/sciencedb.j00001.00347.

A dataset of Mongolian daily question and answer corpus

Terigelehu[1], WANG Siriguleng [1*], HAN Yongshun [1], Ailiya[1], Naheya[1]

1. Inner Mongolia Normal University, Hohhot 010022, P.R. China

*Email: siriguleng@inmu.edu.cn

Abstract: One of the important reasons of the slow pace of the Mongolian question and answer research lies in the scarcity of question and answer corpus. In this paper, we constructed a dataset containing 50,000 pairs of Mongolian question and answer corpus through rule selection, Chinese-Mongolian translation and manual correction after collecting the existing Chinese question answering corpus. The automatic evaluation shows that the corpus has a good diversity of question and answer sentences, and the manual evaluation results show that 97% of the corpus conforms to the daily question and answer logic. The entries in the corpus are mainly from daily conversations in various field. The corpus can used in the end-to-end question and answer model. It is of great values in the practice of Mongolian automatic question and answer research.

Keywords: Mongolian; question and answer corpus; corpus construction; corpus correction

Dataset Profile

Title	A dataset of Mongolian daily question and answer corpus
Data corresponding author	Siriguleng Wang (siriguleng@inmu.edu.cn)
Data author(s)	Terigelehu, WANG Siriguleng
Time range	2019 – 2021
Geographical scope	All over the world
Data volume	4.47 MB
Data format	*.xlsx
Data service system	<http://www.doi.org/10.11922/sciencedb.j00001.00347>
Sources of funding	Project of Inner Mongolia Autonomous Region Science and Technology Plan (No.2021GG0139). The National Natural Science Foundation of China under Grant (No.61762072).
Dataset composition	The dataset consists of one data file with two columns of data, namely Mongolian question and Mongolian answer, totaling 100,000 sentences.