www.csdata.org

ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2021.0105.zh
文献 DOI：
10.11922/11-6035.csd.2021.0105.zh
数据 DOI：
10.11922/sciencedb.j00001.00356
文献分类：信息科学

收稿日期：2021-12-31
开放同评：2022-02-08
录用日期：2022-05-17
发表日期：2022-06-29

专题 多语种智能信息处理

机器翻译辅助的中蒙、维汉语音翻译数据集子集

李宁[1]，朱丽平[1,2*]，赵小兵[1,2]，木尼热•艾尔肯[1]

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： 目前，语音翻译的公开数据集稀少，中文与其他低资源语言的双向语音翻译数据集尤其匮乏，阻碍了相关语言端到端语音翻译研究的推进。本文参考国际语音翻译数据集研究思想，将公开的语音识别数据集（AISHELL、THUYG-20）通过机器翻译，转换成语音翻译数据集，进行数据处理后交由专家审核、校验，从而得到高质量语音翻译数据集。本数据集包括中蒙语音翻译数据集和维汉语音翻译数据集两部分，音频采样率是 16 kHz。中蒙语音翻译数据集包含样本 1919 条，大小为 238 MB。维汉语音翻译数据集包含样本 3692 条，大小为 652 MB。本数据集可用于端到端语音翻译的研究，为探索中文与少数民族语言的语音翻译提供数据支撑，也可结合语音识别数据集用于研究机器翻译。

关键词： 语音翻译；中蒙；维汉；低资源

数据库（集）基本信息简介

数据库（集）名称	中-蒙和维-汉语音翻译数据集
数据作者	朱丽平，李宁
数据通信作者	朱丽平（2007014@muc.edu.cn）
数据时间范围	2021年
地理区域	内蒙古自治区，新疆维吾尔自治区
数据量	890 MB（压缩前）
数据格式	*.wav, *.txt
数据服务系统网址	http://www.doi.org/10.11922/sciencedb.j00001.00356
基金项目	国家社科基金项目（17BGL199）
数据库（集）组成	数据集包含中蒙语音翻译数据集和维汉语音翻译数据集两部分。数据包括音频文件以及对应翻译文本，音频文件格式为wav格式，采样率是16 kHz，文本文件格式是txt文本。中蒙语音翻译数据集包含样本1919条，大小为238 MB。维汉语音翻译数据集包含样本3692条，大小为652 MB。

引 言

语音自古以来就是人际交流最基本的方式，在使用不同语言的人与人之间实现无障碍语音交流一直是世界各国人民的愿望。语音翻译，通过计算机技术实现语音到语音的翻译（S2ST）或语音到文本的翻译（AST），是实现跨语言人际交流的重要工具。

* 论文通信作者
朱丽平：2007014@muc.edu.cn

传统的语音翻译系统采用级联方式，语音到文本翻译由自动语音识别（ASR）模块和机器翻译（MT）模块两级级联实现，语音到语音翻译由 ASR、MT 和语音合成模块（TTS）三级级联实现，通过单独训练和调整每个模块提升整体性能。随着语音识别、机器翻译和语音合成技术的日趋成熟，级联方式语音翻译的整体性能较高，但也存在一些固有的问题，如只有语音没有文字的语言的语音翻译问题[1]，因系统级联而产生的误差传播问题[2]等。为了解决这些问题，端到端模型[3]成为近年来的研究热点。研究表明，当有足够多的数据可用时，端到端模型的性能优于级联方式，但在低数据情况下表现不佳[4]。与现有的语音识别、机器翻译和语音合成数据集相比，语音到语音翻译和语音到文本翻译均面临严重的数据稀缺问题，尤其是低资源小语种语音翻译数据集非常匮乏[5]。

针对语音翻译数据稀缺问题，数据集建设成为当前语音翻译的研究方向之一。在语音到文本翻译数据集建设方面，国内外研究者目前广泛采用的方法是在现有公开数据集基础上，利用机器翻译得到数据集。根据构建方式不同，这种方法又可分为两类，一类是利用 ASR 数据，将源文本翻译成目标语言文本，生成 AST 数据集；另一类是利用 MT 数据，将某一语言的文字进行语音合成，生成 AST 数据集[6]。

BÉRARD A 以 LibriSpeech 公开数据集为基础，对该数据集进行法语对齐与谷歌翻译，生成语音翻译数据集[7]，该数据集已被 LIU Y 用于基于知识蒸馏的端到端语音翻译研究[8]。KANO T 通过英日机器翻译语料库，通过语音合成的方式生成语音数据，进行端到端的英语日语语音翻译研究[9]。PINO J 利用机器翻译模型，将英文文本翻译成法语和罗马尼亚语和利用语音合成技术将 WMT14 进行语音合成生成音频增强数据[6]。KANO T 使用 BTEC 英语日语平行语料库，并使用谷歌语音合成技术生成语音语料库研究远距离语言对的端到端语音翻译[1]。TU M 使用 IWSLT2019 提供的由并行数据和机器翻译生成的合成语料库研究端到端语音翻译[10]。PINO J 证明了两类语音到文本翻译数据集，并证明利用 ASR 生成 AST 数据集比利用 MT 生成 AST 数据集效果更好[6]。

由于目前国内语音翻译相关数据集几乎是空白，国际数据集多集中在英语方面，在汉语方面仅仅开展了英汉领域的研究，蒙古语、维吾尔语研究工作由于缺少相关数据集支撑而无法开展。本研究在现有公开数据集 AISHELL[11]、THUYG-20[12]基础上，利用机器翻译和人工校对相结合，构建了两种语音到文字翻译数据集：中文语音到蒙文文字数据集和维语语音到中文文字数据集，可用于端到端语音翻译模型的研究，开展汉语方面的语音翻译相关研究。本数据集内容涵盖智能家居、无人驾驶、工业生产、新闻等多方面，覆盖面广，可用于多种场景。数据集生成方法较国际公开方法，增加了人工校对步骤，更加科学可靠地保证了数据质量。

1 数据采集和处理方法

本数据集包含两部分，由中文语音蒙文文字语音翻译数据集和维语语音中文文字语音翻译数据集组成。中蒙语音翻译数据集包含 1919 条中文语音，以及中文语音翻译对应的蒙古文文字。维汉音数据集包含 3692 条维吾尔语语音，以及维吾尔语语音翻译对应的中文文字。

1.1 中蒙语音翻译数据集

中文语音蒙文文字语音翻译数据集的中文语音语料直接取自于 AISHELL 语音识别数据集[11]，对应的蒙文文本原始语料由 AISHELL 数据集中的中文文本经过预处理、机器翻译和后处理得到。用中文语音和原始蒙文文本训练语音翻译模型，从训练结果中筛选出准确（Bilingual Evaulation

Understudy，即 BLEU 值为 1）的中蒙语音翻译数据共计 25842 条，得到形成中蒙语音翻译数据集原型。再采用随机抽样的方式，从数据集中随机抽取 2000 条数据，经过专家审核、校对、删除和更新，得到最终的中蒙语音翻译数据集。数据处理方案如图 1 所示。

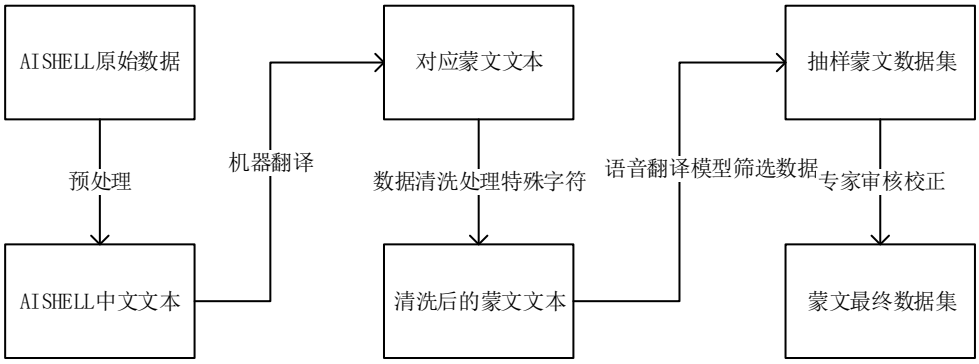

图 1　蒙文文本数据处理流程

Figure 1　Data processing flow of Mongolian text

具体处理步骤如下：

1）预处理：将 AISHELL 数据集中带空格的中文文本数据去空格。

2）机器翻译：把中文文本翻译成蒙文文本。

3）后处理：数据清洗，处理特殊字符，包括过滤蒙文语句中的特殊符号，比如书名号，双引号等，以及用计算机辅助方法对蒙古语中的不可见字符，如蒙古元音分隔符等进行批处理，消除不可见字符造成的蒙古文变形现象。

4）语音翻译模型筛选数据：采用编码器解码器结构的端到端语音翻译模型，将文本正确，BLEU 值为 1 的蒙文翻译文本筛选出来。

5）抽样校验：利用随机抽样，从抽样数据集中抽出部分数据，由专家审核，挑选出存在偏差的数据，交由后续专家人工校对，纠正文中的错词、错字及语义不清的文本，形成最终数据集。

1.2　维汉语音翻译数据集

维汉数据集中的维语语音语料取自于清华大学和新疆大学发布的 THUYG-20 语音识别数据集[12]，对应的中文文本原始语料由 THUYG-20 数据集中拉丁化的维文文本数据经过预处理、机器翻译、后处理、专家校验、最终整合得到，如图 2 所示。

具体处理步骤如下：

1）预处理：将 THUYG-20 数据集，利用 THUYG-20 官方提供的工具包解码拉丁化，得到维吾尔文字。

2）机器翻译：把维吾尔语文本翻译成中文文本。

3）后处理：数据清洗，处理特殊字符，包括过滤维文语句中的特殊符号，比如书名号，双引号等，以及一些机器翻译无法识别的语句。

4）专家校验：通过随机抽样，从数据集中抽出部分数据，由专家审核、校对。

5）整合处理：将专家校对后的数据整理、去除标记，形成最终数据集。

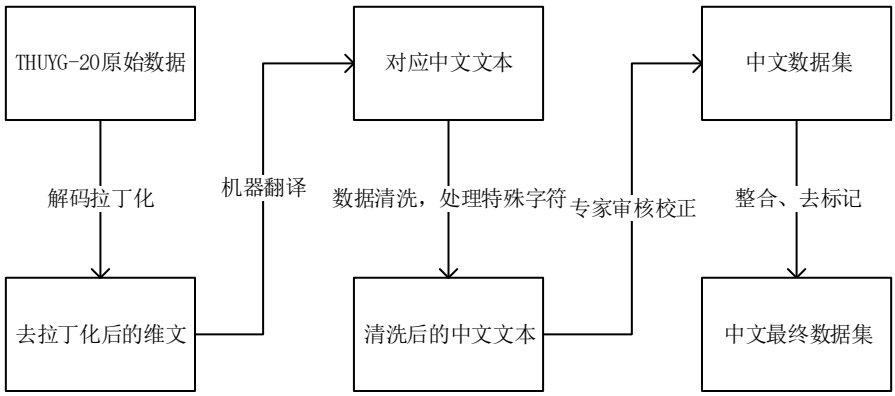

图 2　维文文本数据处理流程

Figure 2　Data processing flow of Uyghur text

2　数据样本描述

本数据集包含中蒙语音翻译数据集和维汉语音翻译数据集两部分。数据包括音频文件以及对应翻译文本，音频文件格式为 wav 格式，采样率是 16 kHz，文本文件格式是 txt 文本。中蒙语音翻译数据集包含样本 1919 条，大小为 238 MB。维汉语音翻译数据集包含样本 3692 条，大小为 652MB。

如图 3，每个数据集包括 wav 文件夹和 doc 文件夹两个文件夹，其中 doc 文件夹中存放的是翻译文本，wav 文件夹中存放音频文件，如下图 4 所示。

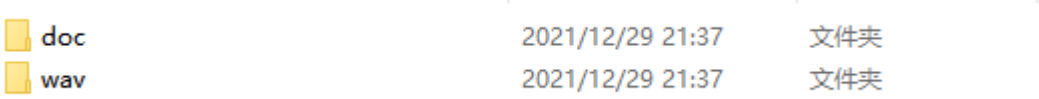

图 3　数据集包含文件夹

Figure 3　Dataset Contains Folders

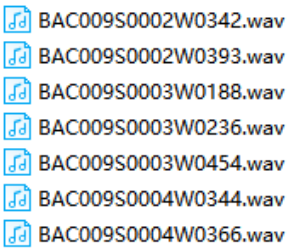

图 4　中蒙语音翻译数据集音频文件

Figure 4　Audio files of Chinese-Mongolian speech translation dataset

图 5 是中蒙语音翻译数据集中的蒙文文本，第一列是音频文件名，对应 wav 文件夹中的音频文件，中间采用水平制表符“\t”分隔，第二列是音频对应的蒙文文本。音频文件名中的第 7–11 个字符，比如 BAC009S0113W0155 中的 S0113 代表是由用户 idS0113 所录制，中间用户 id 不同，代表音频录制人不同。

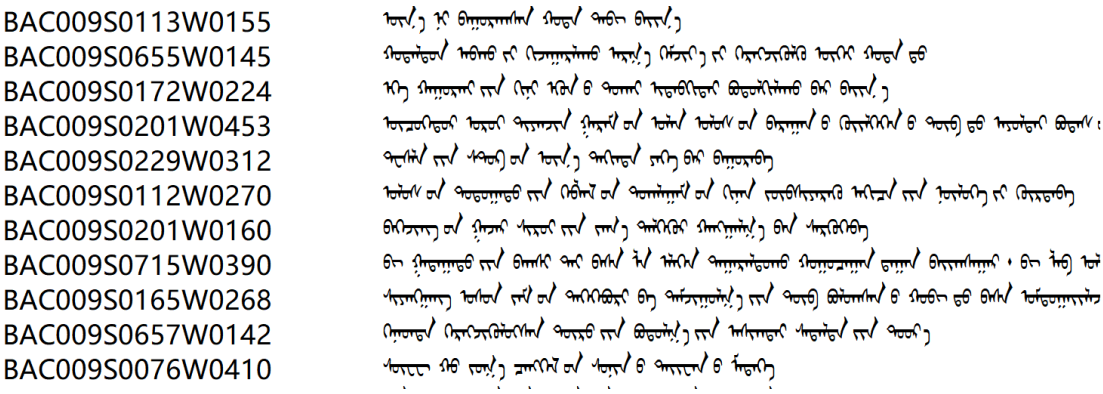

图 5　中蒙语音翻译数据集蒙文文本

Figure 5　Mongolian text of Chinese-Mongolian speech translation dataset

3　数据质量控制和评估

本数据通过机器翻译将源语言文本翻译成目标语言文本，从而得到了语音翻译数据集，但机器翻译的结果存在一定偏差，故后续邀请蒙语、维语语言专家进行打分评价，人工校验数据集，将数据质量高的数据整理成为最终的语音翻译数据。

如图 6 是蒙语专家对中蒙机器翻译数据审核的结果，蒙语专家将根据偶数行的中文数据审核、判断蒙文数据是否存在差错，以及存在怎样的差错。

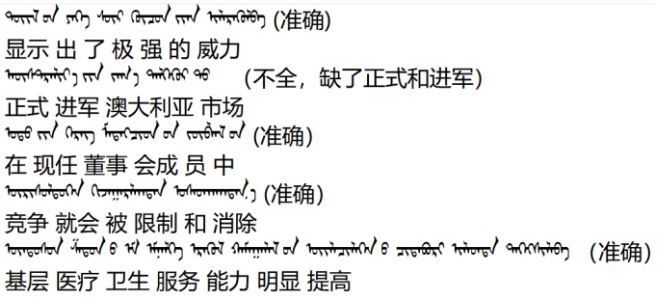

图 6　蒙汉文本审核图

Figure 6　Review of Mongolian-Chinese text

图 7、图 8 给出了蒙文专家校正的文本数据对比图，图中左列均为音频文件名，右列是音频文件所对应的中文文本和蒙文文本，图 7 为专家校验之前的机器翻译原文，图 8 是专家校正之后的结果。

图 9 是维文专家校正的文本数据对比图，每一行从左至右依次为文本所属音频编号，机器翻译的中文文本，翻译检验标记。其中，0 代表翻译不准确，其后为翻译的问题，如漏翻或翻错，以及改正后的中文翻译结果；1 代表翻译正确。

通过专家审核、校验，改善机器翻译产生的偏差，进一步提高数据质量，使得数据更加真实、可靠。

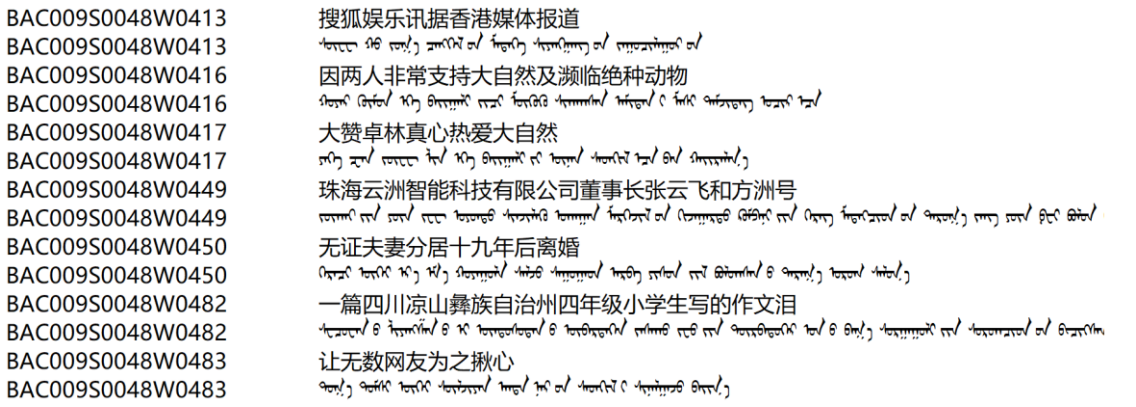

图 7　蒙汉文本原图

Figure 7　Original Mongolian-Chinese Text

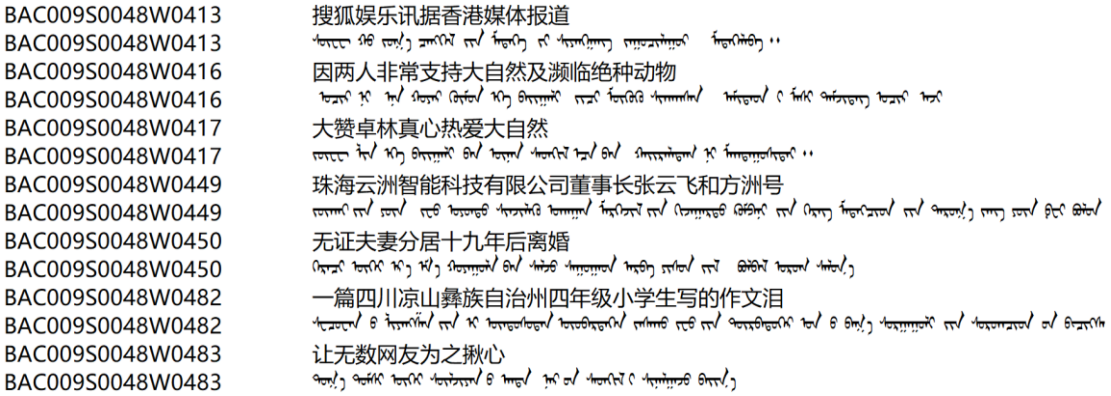

图 8　蒙汉文本校正图

Figure 8　Correction of Mongolian-Chinese Text

F001_012 那位孤寡老人总喜欢坐在茶馆里闲聊 0 漏翻 应为那位孤寡老人总喜欢坐在茶馆里和别人闲聊
F001_013 他看到周围都是陌生人顿时感到很孤独 1
F001_014 指责使人沮丧鼓舞使人振奋 1
F001_015 我们写作时要多观察生活 0 漏翻 应为我们为写作积累素材时要多观察生活
F001_016 学校的各项规章制度必须切实贯彻执行 1
F001_017 全国各地掀起了大规模的增产节约运动 1
F001_018 在孩子从幼年到成年的过渡期父母应该多关心 1
F001_019 老年人喜欢回忆过去年轻人喜欢回忆未来 0 翻错 应为老年人喜欢回忆过去年轻人喜欢畅想未来

图 9　维汉文本校正图

Figure 9　Uyghur Chinese Text Correction Map

4　数据价值

现在语音翻译数据稀少，国际英语相关的数据比较多，但国内研究较少，中蒙数据和维汉数据填补了中文相关语音翻译的稀缺数据。本文提供的语音翻译数据可以直接用于语音翻译的相关研究。本数据是由 AISHELL、THUYG20 数据集处理加工而来，便于使用 AISHELL、THUYG20 数据集的

科研工作人员快速开始训练，同时还便于将 AISHELL、THUYG20 的模型迁移到本数据集上。科研人员也可根据本数据集与 AISHELL、THUYG20 数据集音频命名规则一致，便于修改预处理流程，快速开展相应实验，用于机器翻译的相关研究。

致　谢

感谢中央民族大学中国少数民族语言研究院高娃教授，中国社会科学院民族学与人类学研究所哈斯其木格研究员，中国政法大学戚肖克博士对蒙文机器翻译质量评估给出的宝贵建议，感谢呼和浩特民族学院包乌歌德勒博士，九原区蒙古族学校娜日娜老师，中央民族大学赵美丽、都乐根、媛媛对蒙文数据的审校。

数据作者分工职责

李宁（1996—），男，山东省泰安市人，硕士研究生，研究方向为语音翻译。主要承担工作：数据集的预处理和整合、论文撰写。

朱丽平（1970—），女，湖南省株洲市人，博士，教授，研究方向为语音翻译。主要承担工作：总体质量管控，机器翻译结果审校组织、协调与管理，论文指导与修改。

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：数据质量控制与综合管理。

木尼热·艾尔肯（1999—），女，新疆省叶城县人，本科，研究方向为自然语言处理。主要承担工作：维语数据质量控制。

参考文献

[1] KANO T, SAKTI S, NAKAMURA S. End-to-end speech translation with transcoding by multi-task learning for distant language pairs[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1342–1355. DOI:10.1109/TASLP.2020.2986886.

[2] 杨政. 基于 Seq2Seq 模型的俄汉军事语音翻译关键问题研究[D]. 战略支援部队信息工程大学, 2019. [YANG Z. Research on Key Problems of Russian Chinese Military Speech Translation Based on Seq2Seq Model[D].Strategic Support Force Information Engineering University,2019.]

[3] JIA Y, JOHNSON M, MACHEREY W, et al. Leveraging weakly supervised data to improve end-to-end speech-to-text translation[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK. IEEE, 2019: 7180–7184. DOI:10.1109/ICASSP.2019.8683343.

[4] SPERBER M, NEUBIG G, NIEHUES J, et al. Attention-passing models for robust and data-efficient end-to-end speech translation[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 313–325. DOI:10.1162/tacl_a_00270.

[5] 中国中文信息学会. 中文信息处理发展报告(2021)[EB/OL]. (2021-12) [2021–12]. http://www.cipsc.org.cn/download.php?file=cips2021.pdf. [Chinese Information Society. Chinese

Information Processing Development Report(2021) [EB/OL]. (2021-12) [2021–12]. http://www.cipsc.org.cn/download.php?file=cips2021.pdf.]

[6] PINO J, PUZON L, GU J, et al. Harnessing Indirect Training Data for End-to-End Automatic Speech Translation: Tricks of the Trade[EB/OL]. Computer Science, 2019. (2019-09-14). https://www.semanticscholar.org/paper/6399cfc9e04cdc9f38bb50ab2288fc9180a08bea.

[7] BÉRARD A, BESACIER L, KOCABIYIKOGLU A C, et al. End-to-end automatic speech translation of audiobooks[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. IEEE, 2018: 6224–6228. DOI:10.1109/ICASSP.2018.8461690.

[8] LIU Y C, XIONG H, ZHANG J J, et al. End-to-end speech translation with knowledge distillation[C]//Interspeech 2019. ISCA: ISCA, 2019: 1904. DOI:10.21437/interspeech.2019-2582.

[9] KANO T, SAKTI S, NAKAMURA S. Structured-based curriculum learning for end-to-end English-Japanese speech translation[C]//Interspeech 2017. ISCA: ISCA, 2017. DOI:10.21437/interspeech.2017-944.

[10] TU M, ZHANG F, LIU W. End-to-end speech translation with self-contained vocabulary manipulation[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain. IEEE, 2020: 7929–7933. DOI:10.1109/ICASSP40776.2020.9053431.

[11] BU H, DU J Y, NA X Y, et al. AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, Korea (South). IEEE, 2017: 1–5. DOI:10.1109/ICSDA.2017.8384449.

[12] 艾斯卡尔·肉孜, 殷实, 张之勇, 等. THUYG-20: 免费的维吾尔语语音数据库[J]. 清华大学学报(自然科学版), 2017, 57(2): 182–187. DOI:10.16511/j.cnki.qhdxxb.2017.22.012.[Aisikaer Rouzi, YIN S, ZHANG Z Y, et al. THUYG-20: a free Uyghur speech database[J]. Journal of Tsinghua University (Science and Technology), 2017, 57(2): 182–187. DOI:10.16511/j.cnki.qhdxxb.2017.22.012.]

论文引用格式

李宁, 朱丽平, 赵小兵, 等. 机器翻译辅助的中蒙、维汉语音翻译数据集子集[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-27). DOI: 10.11922/11-6035.csd.2021.0105.zh.

数据引用格式

朱丽平, 李宁. 中-蒙和维-汉语音翻译数据集[DS/OL]. Science Data Bank, 2022. (2022-06-27). DOI: 10.11922/sciencedb.j00001.00356.

A subset of Chinese-Mongolian and Uyghur-Chinese speech translation dataset aided by machine translation

LI Ning[1], ZHU LiPing[1,2*], ZHAO XiaoBing[1,2], MUNIRA[1]

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

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

*Email: 2007014@muc.edu.cn

Abstract: At present, there are few public datasets for speech translation, especially those between Chinese and other low-resource languages. The development of end-to-end speech translation is limited by resources. In light of the research idea of international speech translation datasets, in this paper, we used the public speech recognition datasets (AISHELL and THUYG-20) to convert them into speech translation datasets through machine translation. After data processing, they were reviewed and verified by experts, so as to obtain high-quality speech translation datasets. The dataset includes Chinese-Mongolian speech translation dataset and Uygur-Chinese speech translation dataset, and the audio sampling rate is 16 kHz. The Chinese-Mongolian speech translation subset contains 1,919 items with a size of 238 MB. The Uygur-Chinese speech translation subset contains 3,692 samples with a size of 652 MB. This dataset can be used for the research on end-to-end speech translation, and provide data support for exploring the speech translation between Chinese and minority languages. As the dataset has been reviewed and verified by experts, it can also be combined with speech recognition dataset to study machine translation.

Keywords: speech translation; Chinese-Mongolian; Uyghur-Chinese; low resource

Dataset Profile

Title	A subset of Chinese-Mongolian and Uyghur-Chinese speech translation dataset aided by machine translation
Data corresponding author	ZHU Liping (2007014@muc.edu.cn)
Data authors	ZHU LiPing, LI Ning
Time range	2021
Geographical scope	Inner Mongolia Autonomous Region, Xinjiang Uygur Autonomous Region
Data volume	890 MB
Data format	*.wav, *.txt
Data service system	<http://www.doi.org/10.11922/sciencedb.j00001.00356>
Source of funding	National Social Science Foundation of China (17BGL199)
Dataset composition	The dataset is composed of two parts: Chinese-Mongolian speech translation subset and Uygur-Chinese speech translation subset, including audio files and corresponding translated texts. The audio file format is “wav”; the sampling rate is 16 kHz, and the text file format is “tex” text. The Chinese-Mongolian speech translation subset contains 1,919 samples with a size of 238 MB. The Uygur-Chinese speech translation subset contains 3,692 samples with a size of 652 MB.