China Scientific Data. http://www.csdata.org
Submitted: 2019-04-08

汉维可比语料数据集

冯韬[1,2]，李淼[1*]，曹宜超[1,2]，曾伟辉[1]

1. 中国科学院合肥智能机械研究所，合肥 230031

2. 中国科学技术大学，合肥 230026

* 论文通信作者：李淼（mli@iim.ac.cn）

摘要：语料库的构建是自然语言处理领域的重要工作。但是，双语平行语料库的规模和领域并不能满足实际的需求，尤其是在维吾尔语信息处理中表现得更加明显。因此，从互联网上挖掘汉维双语资源的工作，对于汉维双语资源的建设、促进民族之间的交流具有十分重要的作用。本文针对维吾尔语复杂多变以及汉维语言形态差异大等特点，研究并设计了汉维可比语料挖掘系统。本系统主要包括汉维网页正文抽取，汉维可比语料候选获取以及跨语言相似度计算等几个部分。目前已经有 5000 个汉维可比语料篇章，主要是新闻领域语料和政府公文等。该语料库对于少数民族语言分析与教学，汉维机器翻译等领域具有十分重要的作用。为了使用的便利，本数据集对汉语和维吾尔语进行了进一步的加工和规范化操作。

关键词：语料库建设；可比语料；汉维；数据挖掘

A Chinese-Uighur comparable corpus

Feng Tao[1,2], Li Miao[1], Cao Yichao[1,2], Zeng Weihui[1]

1. Institute of Intelligent Machines, Chinese Academy of Science, Hefei 230031, China

2. University of Science and Technology of China, Hefei 230026, China

*Email: mli@iim.ac.cn

Abstract: Corpus construction is a prerequisite for natural language processing. But the fact is that existing parallel corpora do not meet actual needs for their hardly unsatisfactory scale, which is especially true regarding Uighur information processing. Against this background, our work of constructing Chinese-Uighur corpus based on Internet resources plays an important role in preserving Chinese-Uighur bilingual resources and promoting ethnic exchanges. This studies designs a Chinese-Uighur comparable corpus mining system that fully considers the complexities of Uighur language and the great differences between Chinese and Uighur language forms. This process mainly includes web content extraction, acquisition of candidate comparable corpora and cross-language similarity calculation. Till now, we have collected more than 5000 comparable Chinese and Uigur texts, mainly from news and government documents. The corpus plays an important role in minority language analysis and teaching, and in Chinese-Uigur machine translation. For convenience, Chinese and Uighur language pairs have been further processed and normalized.

Keywords: corpus construction; comparable corpus; Chinese- Uighur; data mining

Article DOI: 10.11922/csdata.2019.0010.zh
Data DOI: 10.11922/sciencedb.748

数据库（集）基本信息简介

数据集名称	汉语-维吾尔语可比语料数据集
数据作者	冯韬，李淼，曹宜超，曾伟辉
数据通信作者	李淼（mli@iim.ac.cn）
数据时间范围	2016-2019
数据量	5000篇章
数据格式	*.txt
数据网址	http://202.127.200.3/sc/kbyl http://www.sciencedb.cn/dataSet/handle/748
基金项目	中国科学院信息化专项科学大数据工程（一期）多民族语言资源特色数据库课题（XXH13505-03-203）
数据集组成	该数据集由从互联网上挖掘的汉语和维吾尔语的可比语料构成，汉语和维吾尔语是篇章对应的。汉维可比语料主要是新闻领域的语料，包括新闻标题、时间、正文等。该数据集包含两个数据文件，它们分别为ch_corpus.zip和uy_corpus.zip，其中：每一个压缩包中包含4个文档文件，分别是document_1，document_2，document_3和document_4。每个文档文件包含两个文件夹uy和ch，其中uy表示维吾尔语，ch表示汉语，每一个文件夹中又包含多个txt文档，维吾尔语和汉语的txt文档是按照名称一一对应的。

Dataset Profile

Title	A Chinese-Uighur comparable corpus
Data corresponding author	Miao Li (mli@iim.ac.cn)
Data authors	Feng Tao, Li Miao, Cao Yichao, Ceng Weihui

Data volume	5000 documents
Data format	*.txt
Data service system	http://202.127.200.3/sc/kbyl http://www.sciencedb.cn/dataSet/handle/748
Sources of funding	Science Big Data Project (Phase I) of the Chinese Academy of Sciences Informatization Program; Multi-ethnic Language Resource Characteristic Database Project (XXH13505-03-203).
Dataset composition	The dataset is composed of comparable corpus of Chinese and Uigur, obtained from the Internet. Chinese and Uigur language pairs are textually corresponding. The dataset is mainly from news, including news headlines, time and text. The dataset contains two data files: ch_corpus.zip and uy_corpus.zip. Each package contains four documents, namely document_1, document_2, document_3 and document_4. Each document contains two folders: uy and ch, where uy represents Uyghur, ch represents Chinese, and each folder contains multiple text documents. Uighur and Chinese language pairs are organized correspondingly according to their names.

引 言

语料库是自然语言处理工作的基础资源，具有非常大的应用价值。根据语料库包含的语种数量，可以分为单语语料库、双语语料库以及多语语料库。其中，双语语料库是最常用也是最主要的语料库资源，根据语料库中语料资源的对应关系，其包含平行语料库和可比语料库两种形式。平行语料库中的双语数据严格互译，其按照不同的对齐粒度可以分为词级、句级、段级以及篇章级。平行语料由于其良好的互译性、双语资源严格对齐等特点，已经被广泛应用于自然语言处理的许多方面。但是，平行语料库的构建是一项非常艰巨的任务，需要借助语言学专家的知识，耗时费力，周期较长。而且，从互联网上获取平行语料也是比较困难的，因为互联网中严格互译的文档资源比较稀少，无法从网络中挖掘大规模的平行语料资源。因此，目前平行语料库中的双语资源数量并不能达到实际的应用需求，尤其是在类似于维吾尔语的少数民族语言方面，该问题更加明显。

Article DOI: 10.11922/csdata.2019.0010.zh
Data DOI: 10.11922/sciencedb.748

可比语料作为平行语料的补充，日益受到了人们的重视。可比语料是指内容具有一定的相似性但是并不是严格互译的双语资源。两篇可比语料文档的主题相似，描述的是同一个事件，但是独立的产生于各自的语言中，文本之间并不是互译的，这些特点使得可以利用机器学习算法从大规模的互联网文本中获取可比语料。首先利用网络爬虫技术从互联网上挖掘源语言文本，其次采用主题建模算法获取文本的主题，然后从互联网上挖掘类似主题的目标语言候选文本，最后利用跨语言相似度算法获取最终的目标文本，并将其放入到可比语料库中[1]。可比语料也可以应用于自然语言处理的其他任务中，如机器翻译，跨语言信息计算，语言模型等。因此，可比语料对于自然语言处理领域具有十分重要的意义。

我国是一个统一的多民族的国家，维吾尔语信息处理对于促进民族之间的交流与合作具有十分重要的意义，汉维可比语料库的建设可以有效的促进汉维机器翻译的研究。目前神经机器翻译已经取得了很好的进展，在多种语言对上的性能超过了传统的机器翻译方法。但是，神经机器翻译是“数据驱动”的方法，其性能严重的依赖于平行语料的规模、质量和领域覆盖面，只有大量的数据才能充分的发挥神经网络的性能。所以，汉维平行语料资源的匮乏严重制约了汉维机器翻译的发展，但是人工构建汉维平行语料库又是非常困难的。因此，在汉维平行语料资源不足的情况下，从互联网上挖掘高质量的汉维可比语料具有重要的意义，可以为汉维机器翻译的研究以及维吾尔语信息处理提供语料资源和技术支撑。

1 数据采集和处理方法

汉语和维吾尔语文本数据是利用网络爬虫技术从互联网上获取的，然后对其进行数据预处理、特征提取、相似度计算等步骤，最终决定是否将其放入到汉维可比语料库中。汉维可比语料挖掘系统框架结构如图 1 所示。

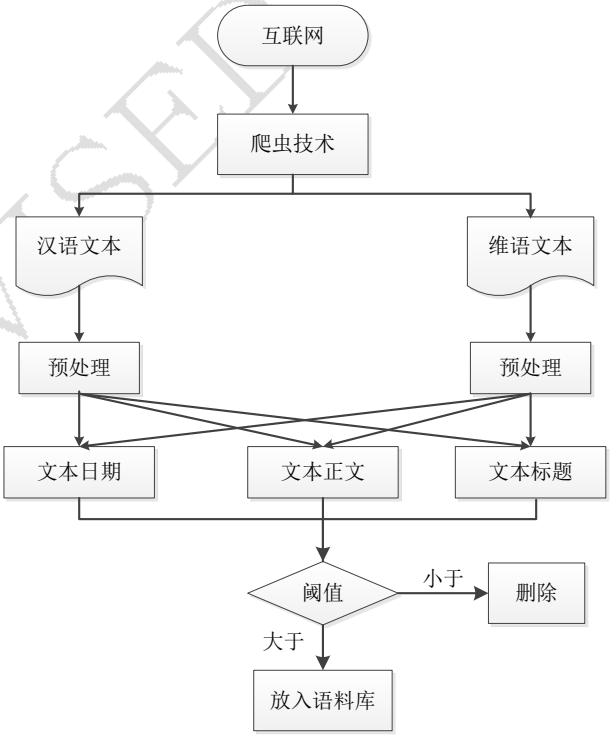

图 1　汉维可比语料系统示意图

该系统利用最大连续文本密度和的方法对汉语和维吾尔语的网页内容进行抽取。根据

Article DOI: 10.11922/csdata.2019.0010.zh
Data DOI: 10.11922/sciencedb.748

现有的网页正文抽取方法，本方法提出了一个融合结构和语言特征的统计模型，将网页文档转化为正、负交替的文本密度序列。为避免丢失短小正文行，采用高斯平滑技术，通过邻近行内容的连续性，增加短文本行的文本密度[2-3]。最后，结合最大间隔距离，利用动态规划的方法计算最大连续文本密度和来抽取网页正文内容，这样可以有效避免将网页评论等篇幅较长的噪声误判为正文内容的情况发生。

在获取汉语和维吾尔语网页文本之后，对其进行相似度计算[4]。在汉维可比语料挖掘系统中，采用融合多特征的汉维网页文本相似度计算方法。该方法首先抽取预处理后的网页文本的发布时间、标题和正文信息等特征，这里的预处理主要是先去噪，然后翻译维吾尔语标题和关键字，再使用中科院的 ICTCLSA（Institute of Computing Technology, Chinese Lexical Analysis System）系统进行分词、过滤停用词等处理[5-6]。然后根据上述特征计算双语文档发布日期的差异、正文长度关系、正文阿拉伯数字相似度、标题重合程度以及正文重合程度 5 种启发信息，并将它们作为特征来判断汉语文本和维吾尔语文本的相似程度。在该方法中利用正则表达式匹配文本的标题和发布日期并且抽取文本的正文内容，然后利用正则表达式提取正文中的阿拉伯数字。选择双语文档发布日期作为相似度计算的特征是因为不同语言文本对同一事件的描述一般是在事件发生后的一段时间内，两篇可比语料文档的发布日期应该是相近的[7-8]。

对于网页文本内容，选择正文长度关系、正文阿拉伯数字、标题重合度以及正文重合程度作为相似度计算的特征。选择正文长度关系是由于两篇可比语料文本对同一事件的描述应基本一致，内容长度比应该在某个值附近分布，可将长度关系转换为长度关系度；选择正文阿拉伯数字相似度是因为可比语料的不同语言文档是对同一事件的描述，那么出现在正文中的量词等阿拉伯数字应基本一致，可以利用欧式距离计算汉维文本中的阿拉伯数字的相似度；选择标题重合程度是因为新闻标题是对内容的概要，可比语料的源语言标题经翻译后应与目标语言标题基本一致，即有较多相同的词汇；选择正文重合程度是因为两篇可比语料文档的主题是一致的，源语言新闻正文经翻译后的文本应与目标语言新闻正文相似，即两篇新闻文档的主旨是相同的。为了提高模型的效率，减少其计算时间，本文取 300 个字符作为处理的阈值，即文本长度超过 300 个字符的数据不参与正文重合度的计算。最后通过神经网络训练得到各启发信息的权重并将 5 种启发信息进行加权融合，从而得到两篇汉维新闻文档的相似度得分。

2 数据样本描述

本数据集的一个样本共包含两个文件：第一个是 txt 格式的汉语语料文本，第二个是 txt 格式的维吾尔语语料文本，汉语文本和维吾尔语文本是一一对应的，图 2、图 3 分别表示汉语语言文本和其相对应的维吾尔语语言文本。

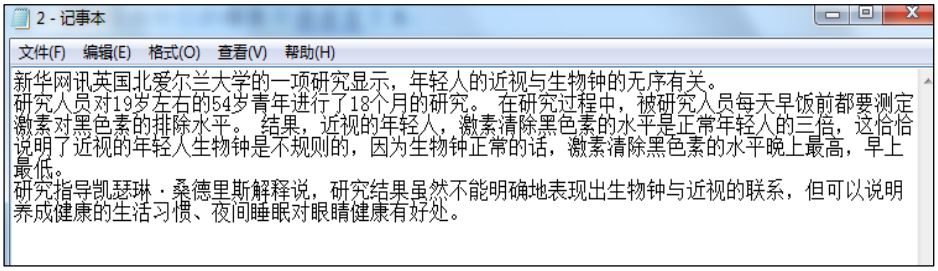

图 2　汉语语言文本

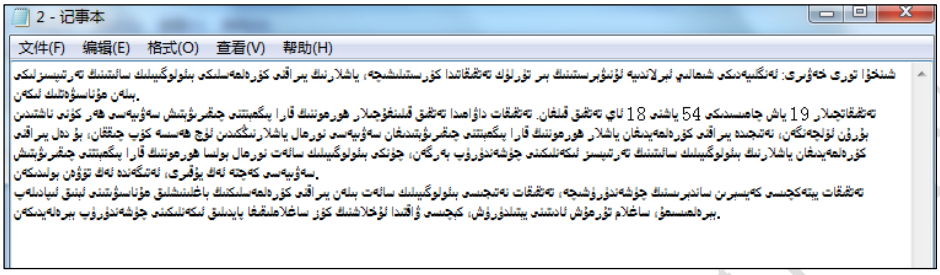

图 3　维吾尔语语言文本

整个数据集由 5000 个样本数据构成，即数据集包含 5000 个汉语语言文本和 5000 个维吾尔语语言文本。图 4 和图 5 分别表示汉语文本的数据结构和维吾尔语文本的数据结构。汉语的文件名是 ch，维吾尔语的文件名是 uy，每一个文件夹中包含多个文本数据，它们是一一对应的关系。如图 4 中的 1_cn.txt 与图 5 中的 1_uy.txt 是一组可比语料对。

图 4　汉语文本的数据结构

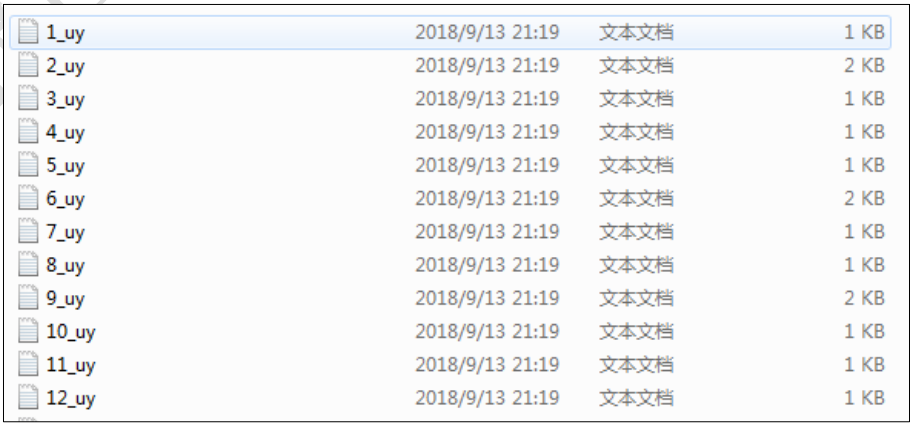

图 5　维吾尔语语言文本数据结构

Article DOI: 10.11922/csdata.2019.0010.zh
Data DOI: 10.11922/sciencedb.748

3 数据质量和评估

为了保证可比语料数据的质量，将汉维可比语料加入到数据库后，审核人员会对这些数据进行进一步筛选和审查。并且为了更好的服务审核人员，我们开发了远程 Web 网页系统供审核人员使用，在网页中显示汉维可比语料供专家审查。因此，维吾尔语语言专家们可以通过远程登录网页的方式对汉维可比语料进行审核，对于审核结果不达标的可比语料，将它们从汉维可比语料库中删除。

在获取汉维可比语料的过程中，我们使用了正则匹配算法对维吾尔语和汉语语料文本进行去噪处理。针对网页文本杂乱无序、不规范等特点，我们把网页中的一些冗余标签，如“<script>”、“<!---->”等替换成空白符，并删除网页文本数据中的一些无用的字符，如“/n”“/r”等。此外，我们还对挖掘到的语料文本数据进行了相应的处理，主要是删除网页文本中的一些杂乱字符，如将获取到的语料文本数据中的“ ”替换成空格符，将“&ldquo”替换成上引号，将“&rdquo”替换成下引号等操作。

4 数据价值

本数据集共分享了 5000 篇章的汉语和维吾尔语的可比语料，对于汉维机器翻译和维吾尔语信息处理具有重要的意义。本数据集可以用于少数民族语言教学和语法语义分析研究，也可以用于训练维吾尔语语言模型和词嵌入等实际任务中，具有广泛的科研价值和较高的社会应用价值。

可比语料库是具有相近含义但不是严格互译的两种语言文本的集合，因此，对于研究两种语言的语法特点和跨语言相似度计算具有十分重要的意义。可比语料库作为自然语言处理领域的重要资源，日益受到了人们的重视，已经被广泛应用于计算语言学的许多方面。

数据作者分工职责

冯韬（1993—）男，江苏徐州人，硕士研究生，研究方向为自然语言处理、机器翻译。主要承担工作：数据收集与整理。

李淼（1955—）女，安徽合肥人，研究院，研究方向为人工智能，自然语言处理。主要承担工作：总体方案设计与组织实施。

曹宜超（1994—）男，山东枣庄人，硕士研究生，研究方向为自然语言处理、机器翻译。主要承担工作：软件系统的构建与调试。

曾伟辉（1982—）女，陕西宝鸡人，博士研究生，研究方向为人工智能，计算机视觉。主要承担工作：数据质量评估。

参考文献

[1] 马颖华, 王永成, 苏贵洋, 等. —种基于字同现频率的汉语文本主题抽取方法[J]. 计算机研究与发展, 2003, 40(6): 874-878.

[2] 安增文, 王超, 徐杰锋. 基于机器学习的网页正文提取方法[J]. 微型机与应用, 2010 (12):

Article DOI: 10.11922/csdata.2019.0010.zh
Data DOI: 10.11922/sciencedb.748

4-6.

[3] 肖根胜. 改进 TFIDF 和谱分割的关键词自动抽取方法研究[D]. 武汉: 华中师范大学, 2012.

[4] 郭华庚, 赵英. 跨语言信息检索研究与应用[J]. 现代情报, 2008, 28(9):142-145.

[5] 杨宇娜. 基于统计的中文词义消歧技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2006.

[6] 梁建飞, 吐尔根·依布拉音, 田生伟, 等. 汉维主题网页自动获取技术的研究[J]. 计算机应用与软件, 2012, 29(01): 42-45.

[7] 热西旦·塔依, 吐尔根·依布拉音. 汉文-维吾尔文双语语料库中段落对齐技术研究[J]. 新疆大学学报(自然科学版), 2010, 27(01): 102-105.

[8] 任高举, 吐尔根·伊布拉音, 艾山·吾买尔. 统计机器翻译中汉维短语对抽取的研究[J]. 新疆大学学报(自然科学版), 2010, 27(03): 349-352.

论文引用格式

冯韬, 李淼, 曹宜超, 曾伟辉. 汉维可比语料数据集[J/OL]. 中国科学数据, 2019. (2019-04-23). DOI: 10.11922/csdata.2019.0010.zh.

数据引用格式

冯韬, 李淼, 曹宜超, 曾伟辉. 汉维可比语料数据集[DB/OL]. Science Data Bank, 2019. (2019-04-08). DOI: 10.11922/sciencedb.748.