ICS
CCS

DB65

新 疆 维 吾 尔 自 治 区 地 方 标 准

DB XX/T XXXX—XXXX

维汉机器翻译平行语料库标注规则及建设要求

Annotation Rules and Construction Requirements for Uyghur-Chinese Machine Translation Parallel Corpora

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

2023－XX－XX 发布　　2023－XX－XX 实施

新疆维吾尔自治区市场监督管理局　　发 布

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国科学院新疆理化技术研究所提出。

本文件由新疆维吾尔自治区工业和信息化厅归口。

本文件起草单位：中国科学院新疆理化技术研究所。

本文件主要起草人：XX。

本文件实施应用中的疑问，请咨询中国科学院新疆理化技术研究所。

对本文件的修改意见及建议，请反馈至中国科学院新疆理化技术研究所（乌鲁木齐市新市区科学二街181号）、新疆维吾尔自治区市场监督管理局（乌鲁木齐市新华南路167号）。

中国科学院新疆理化技术研究所 联系电话：0991-3837795；传真：0991-3838957；邮编：830046

新疆维吾尔自治区市场监督管理局 联系电话：0991-2818750；传真：0991-2311250；邮编：830004

维汉机器翻译平行语料库标注规则及建设要求

1 范围

本文件规定了维汉机器翻译平行语料库标注及建设过程中的术语和定义、平行语料元数据以及语料库建设流程。

本文件适用于维汉机器翻译系统中双向平行语料的采集、平行语料库系统的设计与建设，维汉机器翻译平行语料库建设机构或个人参照执行。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 40035-2021 双语平行语料加工服务基本要求

GB/T 15237.1—2000 术语工作 词汇 第1部分:理论与应用

GB/T 40036-2021 翻译服务 机器翻译结果的译后编辑 要求

GB/T 18793-2002 信息技术 可扩展置标语言(XML)1.0

GM/T 0125.1-2022 JSON Web 密码应用语法规范 第1部分：算法标识

3 术语和定义

下列术语和定义适用本文件。

3.1

语料 corpus

语言材料或资料。

[来源:GB/T 40035-2021,3.2]

3.2

语料库 corpora

集中起来供分析用的语料集合。

[来源:GB/T 15237.1—2000,3.6.9,有修改]

3.3

机器翻译 machine translation；MT

使用计算机应用程序将文本从一种自然语言自动翻译成另一种自然语言。

[来源:GB/T 40036-2021,3.1.1]

3.4

双语平行语料 bilingual parallel corpus

由两种语言构成，并在章节、段落、句子或其他级别平行对齐的语料。

[来源:GB/T 40035-2021,3.3]

3.5

元数据 metadata

关于数据的内容、质量、状况和其他特性的描述性数据。

[来源:GB/T 40035-2021,3.7]

3.6

平行词对 parallel word pairs
双语平行语料(3.3)中具有“翻译”关系的两个单词。

3.7

平行短语对 parallel phrase pairs
双语平行语料(3.3)中具有“翻译”关系的两个短语。

3.8

平行句对 parallel sentence pairs
双语平行语料(3.3)中具有“翻译”关系的两个句子。

3.9

标注 annotation
将双语语料进行篇章、段落、句子或其他级别的对齐，构成平行对照的形式。
[来源:GB/T 40035-2021,3.11]

3.10

源语言 source language
待翻译内容的语言。
[来源:GB/T 40036-2021,3.2.2]

3.11

目标语言 target language
由源语言（3.10）内容翻译而成的语言。
[来源:GB/T 40036-2021,3.2.4]

3.12

领域 domain
语料（3.1）按照内容划分所属的类别。

3.13

外来词 loan word
从其他语言音译或简单直译而来的词语。

3.14

可扩展置标语言 XML Extensible Markup Language
XML是标准通用置标语音（Standard generic markup language，SGML）的一个子集。XML描述了一类成为XML文件的数据对象，同时也部分地描述了处理这些数据对象的计算机程序的行为。
[来源:GB/T 18793-2002,引言]

3.15

JSON Javascript Object Notation
Javascript对象标记，一种轻量级、基于文本的、语言独立的数据交换格式。
[来源:GM/T 0125.1-2022, 3.1]

4 平行语料元数据

4.1 基本信息

基本信息（basic information）定义了语料库编号、创建人、创建时间、领域、类型、描述、状态、规模、来源及版本等信息。

4.1.1 编号

中文名称：编号
英文名称：id
定　　义：双语平行语料在语料库中的唯一标识符
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例："202105060123"

4.1.2 创建人

中文名称：创建人
英文名称：creator
定　　义：双语平行语料创建机构或个人名称
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例："中国科学院新疆理化技术研究所"

4.1.3 创建时间

中文名称：创建时间
英文名称：create_time
定　　义：语料库的创建时间
数据类型：字符串
值　　域：用阿拉伯数字将年、月、日、时、分、秒标全，具体格式应为YYYY-MM-DD HH:mm:ss
是否必填：是
取值示例："2021-05-06 11:12:38"

4.1.4 领域

中文名称：领域
英文名称：domain
定　　义：语料内容所属的类别
数据类型：字符串
值　　域：无要求
是否必填：否
取值示例："体育"

4.1.5 描述

中文名称：描述
英文名称：description
定　　义：语料库内容的简要描述

数据类型：字符串
值　　域：无要求
是否必填：否
取值示例："2021年5月2日新闻联播"

4.1.6　类型

中文名称：类型
英文名称：type
定　　义：平行语料库中每条语料记录的类型
数据类型：字符串
值　　域：“词对”、“短语对”及“句对”
是否必填：否
取值示例："句对"

4.1.7　状态

中文名称：状态
英文名称：state
定　　义：语料库的对齐状态
数据类型：字符串
值　　域：“编辑”、“未对齐”及“对齐”
是否必填：是
取值示例："对齐"

4.1.8　规模

中文名称：规模
英文名称：size
定　　义：语料库中平行元素对的数量
数据类型：整数
值　　域：大于零的整数
是否必填：是
取值示例：100

4.1.9　来源

中文名称：来源
英文名称：origin
定　　义：语料库数据的来源信息
数据类型：字符串
值　　域：无要求
是否必填：否
取值示例："XX公司"

4.1.10　版本

中文名称：版本
英文名称：version
定　　义：语料库的版本信息

数据类型：字符串
值　　域：无要求
是否必填：是
取值示例："1.0"

4.2　源语言信息

源语言信息（source language information）定义了源语言语种代码及源语言相关的描述。

4.2.1　语种代码

中文名称：语种代码
英文名称：language_code
定　　义：源语言的语种代码
数据类型：字符串
值　　域：“zh”为中文、“ug”为现行维吾尔文、“ru-ug”为以斯拉夫文书写的维吾尔文、“en-ug”为以拉丁文书写的维吾尔文
是否必填：是
取值示例："zh"

4.2.2　描述

中文名称：描述
英文名称：description
定　　义：语种的文字描述
数据类型：字符串
值　　域：无要求
是否必填：否
取值示例："中文"

4.3　目标语言信息

目标语言信息（target language information）定义了目标语言语种代码及目标语言相关的描述。

4.3.1　语种代码

中文名称：语种代码
英文名称：language_code
定　　义：目标语言的语种代码
数据类型：字符串
值　　域：“zh”为中文、“ug”为现行维吾尔文、“ru-ug”为以斯拉夫文书写的维吾尔文、“en-ug”为以拉丁文书写的维吾尔文
是否必填：是
取值示例："ug"

4.3.2　描述

中文名称：描述

英文名称：description
定　　义：语种的文字描述
数据类型：字符串
值　　域：不做要求
是否必填：否
取值示例："现行维吾尔文"

4.4　平行语料

平行语料（parallel corpus）定义了语料库中平行句对（平行词对、平行短语对）的索引、匹配、源文及译文等数据。

4.4.1　索引

中文名称：索引
英文名称：index
定　　义：平行句对（平行词对、平行短语对）在语料库中的序号
数据类型：整数
值　　域：0至语料库规模减1
是否必填：是
取值示例：10

4.4.2　匹配

中文名称：匹配
英文名称：match
定　　义：表示当前平行句对（平行词对、平行短语对）是否匹配
数据类型：布尔
值　　域：true表示匹配，false表示不匹配
是否必填：是
取值示例：false

4.4.3　源文

中文名称：源文
英文名称：source_text
定　　义：当前平行句对（平行词对、平行短语对）中的源语言句子（单词、短语）
数据类型：字符串
值　　域：无要求
是否必填：是
取值示例："今天2021年5月10号。"

4.4.4　译文

中文名称：译文
英文名称：target_text
定　　义：当前平行句对（平行词对、平行短语对）中源语言句子（单词、短语）对应的目标语言句子（单词、短语）
数据类型：字符串

值　　域：无要求
是否必填：是
取值示例："بۈگۈن 2021-يىلى 5-ئاينىڭ 10-كۈنى."

5　建设流程

维汉平行语料库建设及标注工作流程如下：

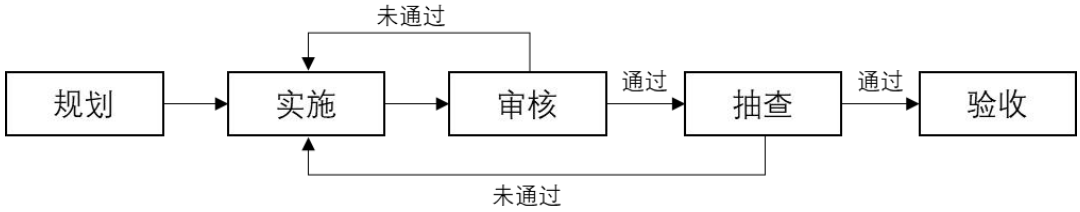

图1 语料库建设及标注流程

5.1　规划

5.1.1　明确要求

开始标注双语平行语料之前，负责人应明确如下要求：
a)　明确标注内容和标注规则；
b)　明确标注任务完成的时间节点；
c)　明确数据验收规则。

5.1.2　明确计划

负责人应根据标注需求制定标注计划，包括进度计划、人员计划、资金计划、质量控制计划、验收计划等。

5.1.3　专项培训

按照标注计划和标注规则，对标注人员进行有针对性的培训。

5.1.4　标注规则

5.1.4.1　语气一致

平行句对必须同属一种语气，如：中文句子属于陈述句，维吾尔文句子也要属于陈述句；中文句子属于疑问句，维吾尔文句子也要属于疑问句。

5.1.4.2　内容一致

a)　平行句对所描述的事物、时间、地点须一致；
b)　平行句对中所用的人称须一致。

5.1.4.3　时态一致

平行句对的时态须一致，如：中文句子采用将来时，维吾尔文句子也须采用将来时。

5.1.4.4 标点符号一致

a) 单引号、双引号、圆括号、方括号、六角括号、单书名号、双书名号必须成对出现，缺失部分需要补齐；
b) 句末的句号、感叹号、问号以及省略号等符号在源语言和目标语言句子中保持一致；
c) 中文里的破折号（——）、一字线（—）、短横线（-）在维吾尔文里用短横线（-）表示。

5.1.4.5 数字一致

平行句对中出现的数字所表示的数值必须一致。

5.1.4.6 外来词的用法

a) 维吾尔语借用的汉语外来词以音译形式书写；
b) 维吾尔语借用的外语缩略外来词以外文形式书写。

5.2 实施

5.2.1 数据获取

收集和准备需要标注的原始数据，确保数据的完整性、准确性和可用性，并对数据进行清洗、去噪等预处理操作。

5.2.2 任务创建

负责人利用语料库管理工具创建标注任务。

5.2.3 任务分发

负责人将待标注任务分派给标注人员。

5.2.4 任务实施

数据标注人员使用相应数据标注工具完成指派的数据标注任务。

5.3 审核

5.3.1 制定审核标准

确定明确的审核标准和指标，确保标注结果与预期结果一致。

5.3.2 数据标注验证

数据标注审查人员对标注好的数据进行审核，按标注要求比对标注结果，以确保准确性和一致性。

5.3.3 任务数据回收

对标注不合格数据进行回收，并重新分派进行标注。

5.3.4 问题记录与反馈

在审核过程中记录发现的问题和错误，并与标注团队进行问题反馈与沟通，确保标注团队理解存在的问题和改进要求。

5.4 抽查

5.4.1 随机抽样验证

从标注好的数据中随机选择一部分样本进行检查，并按标注要求比对标注结果，以确保准确性和一致性。

5.4.2 任务数据回收

对标注不合格数据进行回收，并重新分派进行标注。

5.4.3 问题记录与反馈

在审核过程中记录发现的问题和错误，并与标注团队进行问题反馈与沟通，确保标注团队理解存在的问题和改进要求。

5.5 验收

5.5.1 确定验收标准

需求方需要明确数据标注的验收标准和指标，例如数据的格式、质量、准确性等要求，以及与标注任务相关的特定要求。

5.5.2 数据整理与交付

对标注数据进行整理，确保数据的结构化和可访问性，同时提供详细的交付文件和文档，例如数据标注说明、数据格式说明、标签定义等。

5.5.3 验收报告和文档

撰写验收报告，记录验收过程、问题、改进措施和结果，报告应包括标注数据的准确性、一致性、完整性等评估结果。

附录

语料库文件应采用XML或JSON文件格式，由UTF-8编码方式存储。

1. XML 文件格式实例

```
<?xml version="1.0" encoding="utf-8"?>
    <corpus>
        <basic_information>
            <id>102</id>
            <creator>中国科学院新疆理化技术研究所</creator>
            <create_time>2021-05-03 13:34:09</create_time>
            <domain></domain>
            <description>维-汉人名词典</description>
            <type>词对</type>
            <state>对齐</state>
            <size>10000</size>
            <origin>XX出版社</origin>
            <version>1.0</version>
        </basic_information>
        <source_language_information>
            <language_code>ug</language_code>
            <description>现行维吾尔语</description>
        </source_language_information>
        <target_language_infromation>
            <language_code>zh</language_code>
            <description>中文</description>
        </target_language_infromation>
        <parallel_corpus>
            <index>0</index>
            <match>true</match>
            <source_text>ئانارخان</source_text>
            <target_text>阿娜尔汗</target_text>
        </parallel_corpus>
        <parallel_corpus>
            <index>1</index>
            <match>true</match>
            <source_text>ئالىم</source_text>
            <target_text>阿里木</target_text>
        </parallel_corpus>
        ⋮
    </corpus>
```

2. JSON 文件格式实例

```
{
    "basic_information": {
        "id": "102",
        "creator": "中国科学院新疆理化技术研究所",
        "create_time": "2021-05-03 13:34:09",
        "domain": "",
        "description": "维-汉人名词典",
```

```
        "type": "词对",
        "state": "编辑",
        "size": 100,
        "origin": "XX出版社",
        "version": "1.0"
    },
    "source_language_information": {
        "language_code": "uy",
        "description": "维吾尔文"
    },
    "target_language_information": {
        "language_code": "zh",
        "description": "中文"
    },
    "parallel_corpus": [
        {
            "index": 0,
            "match": true,
            "source_text": "ئانارخان",
            "target_text": "阿娜尔汗"
        },
        {
            "index": 1,
            "match": true,
            "source_text": "ئالىم",
            "target_text": "阿里木"
        },
        ⋮
    ]
}
```