ICS
CCS

DB65

区　　准

DB XX/T XXXX—XXXX

则及

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

2023－XX－XX 发　　2023－XX－XX

区　　发

则及

前

XX

0991-2818750 0991-2311250 830004

则及

1

2 件

; , ( )

GB/T 40035-2021
GB/T 15237.1 2000 1 :
GB/T 40036-2021
GB/T 18793-2002 (XML)1.0
GM/T 0125.1-2022 JSON Web 1

3 义

3.1

corpus

[:GB/T 40035-2021,3.2]

3.2

corpora

[:GB/T 15237.1 2000,3.6.9,]

3.3

machine translation；MT

[:GB/T 40036-2021,3.1.1]

3.4

双 bilingual parallel corpus

[:GB/T 40035-2021,3.3]

3.5

元 metadata

[:GB/T 40035-2021,3.7]

3.6

parallel word pairs
(3.3)

3.7

parallel phrase pairs
(3.3)

3.8

parallel sentence pairs
(3.3)

3.9

annotation

[:GB/T 40035-2021,3.11]

3.10

source language

[:GB/T 40036-2021,3.2.2]

3.11

target language
3.10
[:GB/T 40036-2021,3.2.4]

3.12

domain
3.1

3.13

loan word

3.14

XML Extensible Markup Language
XML Standard generic markup language， SGML XML
XML
[:GB/T 18793-2002,]

3.15

JSON Javascript Object Notation
Javascript
[:GM/T 0125.1-2022, 3.1]

4 元

4.1 信

basic information

4.1.1

id

"202105060123"

4.1.2 创 人

creator

" "

4.1.3 创

create_time

YYYY-MM-DD HH:mm:ss

"2021-05-06 11:12:38"

4.1.4

domain

" "

4.1.5

description

"2021 5 2 "

4.1.6

type

" "

4.1.7

state

" "

4.1.8

size

100

4.1.9

origin

"XX "

4.1.10

version

"1.0"

4.2 信

source language information

4.2.1 代

language_code

zh ug ru-ug en-ug

"zh"

4.2.2

description

" "

4.3 信

target language information

4.3.1 代

language_code

zh ug ru-ug en-ug

"ug"

4.3.2

description

" "

4.4

parallel corpus

4.4.1

index

0 1

10

4.4.2 匹

match

true false

false

4.4.3

source_text

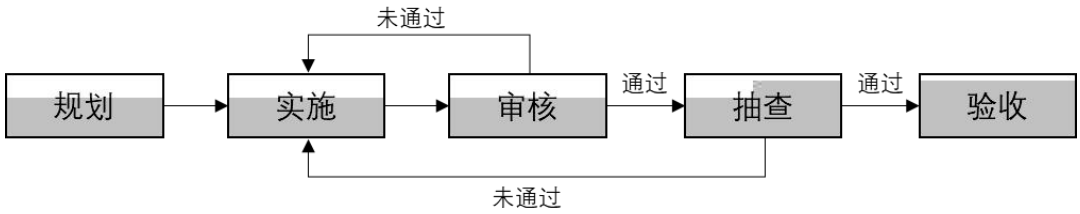
" 2021 5 10 "

4.4.4

target_text

"بۇگۈن 2021-يىلى 5-ئاينىڭ 10-كۈنى."

5



1

5.1 划

5.1.1

a)
b)
c)

5.1.2 划

5.1.3 专

5.1.4 则

5.1.4.1 一

5.1.4.2 内 一

a)
b)

5.1.4.3 一

5.1.4.4 —

a)

b)

c) - -

5.1.4.5 —

5.1.4.6

a)

b)

5.2

5.2.1

5.2.2 任务创

5.2.3 任务分发

5.2.4 任务

5.3

5.3.1 制 准

5.3.2

5.3.3 任务

5.3.4 与反

5.4

5.4.1

5.4.2 任务

5.4.3 与反

5.5

5.5.1 准

5.5.2 与交付

5.5.3

XML JSON UTF-8

1. XML 件 例

```
<?xml version="1.0" encoding="utf-8"?>
    <corpus>
        <basic_information>
            <id>102</id>
            <creator>                                    </creator>
            <create_time>2021-05-03 13:34:09</create_time>
            <domain></domain>
            <description>  -              </description>
            <type>      </type>
            <state>      </state>
            <size>10000</size>
            <origin>XX          </origin>
            <version>1.0</version>
        </basic_information>
        <source_language_information>
            <language_code>ug</language_code>
            <description>                  </description>
        </source_language_information>
        <target_language_infromation>
            <language_code>zh</language_code>
            <description>      </description>
        </target_language_infromation>
        <parallel_corpus>
            <index>0</index>
            <match>true</match>
            <source_text>ئاراخان</source_text>
            <target_text>            </target_text>
        </parallel_corpus>
        <parallel_corpus>
            <index>1</index>
            <match>true</match>
            <source_text>ئادەم</source_text>
            <target_text>          </target_text>
        </parallel_corpus>

    </corpus>
```

2. JSON 件 例

```
{
    "basic_information": {
        "id": "102",
        "creator": "                                    ",
        "create_time": "2021-05-03 13:34:09",
        "domain": "",
        "description": "  -              ",
```

```
        "type": "        ",
        "state": "        ",
        "size": 100,
        "origin": "XX            ",
        "version": "1.0"
    },
    "source_language_information": {
        "language_code": "uy",
        "description": "              "
    },
    "target_language_information": {
        "language_code": "zh",
        "description": "        "
    },
    "parallel_corpus": [
        {
            "index": 0,
            "match": true,
            "source_text": "ئاساخان",
            "target_text": "              "
        },
        {
            "index": 1,
            "match": true,
            "source_text": "ئادەم",
            "target_text": "            "
        },

    ]
}
```