**附件3：语音翻译评测任务数据文件格式说明**

　　本文档对评测中的相关数据文件及格式进行说明，文件包括评测组织方发放的数据文件以及参评单位需要提交的结果文件。

　　所有文件均要求为UTF-8编码，其中评测组织方发放的开发集（含参考答案）、测试集以及参评单位最终提交的翻译结果文件为XML文件（所使用的文档类型定义请参见本附件第二部分），其它文件均要求为纯文本文件。

**一、评测组织方发放的数据格式说明**

　　评测组织方发放的数据有三种，分别为：训练集、开发集和测试集。以下是数据格式说明。

**1. 训练集**

　　训练集为综合领域，包含音频文件、文本文件以及补充说明文件。

**1.1 音频文件**

|  |  |
| --- | --- |
| 数据格式 | WAV |
| 采样率 | 16kHZ |
| 样本位数 | 16 |
| 音频通道 | 单声道 |
| 文件名格式 | Baidu\_XXXX.wav，其中XXXX为演讲者id |
| 文件大小 | 取决于演讲内容的长短，一般为5~50分钟 |

**1.2 文本文件**

对音频内容进行语音人工转写、人工翻译，以及断句等加工后形成的文本文件。包含时间轴信息、转录文本、演讲者ID、翻译文本以及文件信息。整体文件采用文本格式存取，样例如下：

|  |
| --- |
| {“duration”: 11.850000000000009, “offset”: 61.06, “speaker\_id”: “1922”, “wav”: “baidu\_1922.wav”, “wav\_id”: “baidu\_1922”, “transcript”:“各位百度的同学，下午好。”, “translation”:“Good afternoon, every classmates from Baidu.”} |

字段说明

|  |  |
| --- | --- |
| duration | 音频文件长度，单位秒 |
| offset | 音频文件在当前演讲稿中的起始位置，单位秒 |
| speaker\_id | 演讲者ID |
| wav | 音频文件名 |
| wav\_id | 音频文件编号 |
| transcript | 转录文本 |
| translation | 翻译文本 |

**1.3 补充说明文件**

存储演讲者ID和演讲者名字的映射关系，数据格式为：

|  |  |
| --- | --- |
| 演讲者ID | 演讲者名字 |

以及语音来源渠道等版权申明信息。

**1.4 数据样例**

WAV：baidu\_0032.wav

offset: 258 (4分18秒)

Duration:11

speaker\_id: lichao

wav\_id: baidu\_0032

wav:baidu\_0032.wav

transcript: 呃97%的用户体验就是97%的字准的体验应该是什么样的呢，今天我们大家来体验一把，就是这是我们从线上呢捞下来很多那个线上真实用户的一些query.

translation: What kind of difference would 97%-accuracy-rate of recognition bring to users? Let's try it out today. We got a lot of queries from real users online.

**2. 开发集**

　　开发集包括源语言音频文件、转写后文本文件（经过人工校对）和参考译文文件。

1. **源语言音频文件**

每个音频文件时长1-2分钟。

**（2）转写后文本文件**

与其他翻译任务采用一致的xml格式。

源语言文件包含一个srcset元素，这个元素包含以下属性：

* setid: 开发集的id
* srclang: 源语言标识，值为：en（英语）、zh（汉语）、mn（蒙古语）、uy（维语）、ti（藏语）或jp（日语）
* trglang: 目标语言标识，值为：en、zh、mn、uy、ti或jp

srcset元素包含一个或多个DOC元素，每个DOC具有属性：

* docid: 对应音频文件名字

每个doc对应一个音频文件，并且包含多个seg。每个seg内容代表一串连续字符串。

每个seg有一个timestamp属性，在开发集中标记为0。在提交的结果文件中，参评单位此属性处标明系统开始产生译文的时间，单位秒，即当前演讲者演讲语音的时间戳。以此来计算系统时延。由于测试集只包含语音文件，参赛单位可以利用组织方提供的语音识别基准系统输出语音识别的中间结果，以此来获取当前翻译的时间戳。

　　图3给出了一个开发集源语言文件的示例。

|  |
| --- |
| <?xml version="1.0" encoding="UTF-8"?>  <srcset setid="zh\_en\_news\_trans" srclang="zh" trglang="en">  <DOC docid="id">  <p>  <seg id="1" timestamp="0">句子1</seg>  <seg id="2" timestamp="0">句子2</seg>  ……  </p>  ……  </DOC>  </srcset> |
|  |

图3 开发集源语言文件示例

**（3） 参考译文文件**

　　参考译文文件包含一个refset元素，DOC元素包含docid和site两个属性，其中docid表示文档名称，site用以区分不同的参考译文。本次评测中，汉英、英汉和日汉开发集数据的每个句子将有1个不同的参考译文，蒙汉、藏汉和维汉开发集数据的每个句子将给出4个不同的参考译文。

　　图4给出了一个开发集参考译文的示例。

|  |
| --- |
| <?xml version="1.0" encoding="UTF-8"?>  <refset setid="zh\_en\_news\_trans" srclang="zh" trglang="en">  <DOC docid="news" sysid="ref" site="1">  <p>  <seg id="1">参考译文11 </seg>  ……  </p>  ……  </DOC>  <DOC docid="news" sysid="ref" site="2">  <p>  <seg id="1">参考译文21 </seg>  ……  </p>  ……  </DOC>  <DOC docid="news" sysid="ref" site="3">  <p>  <seg id="1">参考译文31</seg>  ……  </p>  ……  </DOC>  <DOC docid="news" sysid="ref" site="4">  <p>  <seg id="1">参考译文41 </seg>  ……  </p>  ……  </DOC>  </refset> |
|  |

图4 开发集参考文件示例

**3. 测试集**

　　在评测阶段组织方将只发放测试集音频文件，其格式与开发集源语言文件格式相同。参赛单位可以利用组织方提供的语音识别系统进行语音识别。

**二、参评单位需要提交的数据格式说明**

　　参评单位仅需要提供最终的结果文件及系统描述信息，其格式说明如下。

**1. 文件命名**

　　所有需要提交的文件的命名方式请遵循下表要求：  
（其中：项目代号以ce为例，参评单位代号以ict为例，评测语料年份以2018为例）

|  |  |  |
| --- | --- | --- |
| 文件 | 文件名模式 | 文件名举例 |
| 最终翻译结果 | 项目代号-评测语料年份-参评单位代号-主/对比系统-参评系统代号.xml | ce-2018-ict-primary-a.xml  ce-2018-ict-contrast-b.xml |

注：文件名模式中主/对比系统-参评系统代号必须以英文字母顺序排序，顺序中包含主及对比系统，例如：主系统为a，则对比系统从b开始。

**2. 最终翻译结果**

　　最终翻译结果文件为解码器输出文件经过后处理的结果。

　　最终翻译结果文件格式与测试集源语言文件格式基本相同。

　　最终翻译结果文件包含一个tstset元素，tstset元素包含一个system元素，system元素包含site（说明单位名称）和sysid（说明系统标识）两个属性。

　　其中，system元素应给出参评系统的描述信息，即对以下内容给出说明：

* 软硬件环境：包括操作系统及其版本、CPU数量、CPU类型及其频率、系统内存大小等等；
* 运行时间：参评系统从接受输入到产生全部输出所花费的时间；
* 技术概要：简要说明参评系统所采用的主要技术和重要参数，如果采用了系统融合技术，这里要进行说明；
* 训练数据说明：说明参评系统所使用的训练数据和开发数据，并注明是受限训练还是非受限训练，对于英汉、汉英项目，还应注明使用的数据为WMT数据，CCMT数据，还是两者皆有；
* 外部技术说明：说明除了参评单位自己的技术外，还采用了哪些外部技术，包括各种开源代码、自由软件、共享软件或商业软件。

　　tstset元素中的DOC元素及其内部的p元素、seg元素应与测试集源语言文件中相应的元素及其属性保持一一对应关系。

　　图5示例说明了最终翻译结果文件的格式。

|  |
| --- |
| <?xml version="1.0" encoding="UTF-8"?>  <system site="单位名称" sysid="系统标识">  参评系统的描述信息  ............  </system>  <DOC docid="文档名称" sysid="系统标识">  <p>  <seg id="1">提交译文1</seg>  <seg id="2">提交译文2</seg>  ……  </p>  ……  </DOC>  </tstset> |
|  |

图5 最终翻译结果文件示例

**3. 文件细节**

　　关于最终提交的结果文件，请参评单位注意以下细节：

* 本次评测的输入输出文件格式采用严格的XML格式。XML格式与NIST评测的文件格式的最主要的区别在于，XML格式的文件中，标签以外的文本如果出现以下五个字符，必须采用相应的转义序列来表示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字符 | & | > | ＜ | " | ' |
| 转义序列 | &amp; | &lt; | &gt; | &quot; | &apos; |

* 汉语译文中，外国人名中间的中圆点统一采用UTF-8编码中十六进制为“ E2 80 A2”的样式，如“托德·西蒙斯” ；
* 英文token采用的标准：对语料进行后处理时，按照NIST的mteval-v11b.pl中normalizeText函数的tokenization方式。如图6所示，我们对其中的主要语句增加了汉语注释，供大家参考。

|  |
| --- |
| # language-dependent part (assuming Western languages): $norm*text = " $norm*text "; #把原文的最开头和最末尾各加上一个空格（最后再删去） $norm*text =~ tr/[A-Z]/[a-z]/ unless $preserve*case; #除非用户指定保留大小写，否则一般把英文大写字母转化成小写 $norm*text =~ s/([{-~[-` -&(-+:-\@\/])/ $1 /g; # tokenize punctuation #把下述符号两边各加上一个空格（对应ASCII字符集中的十六进制值标注在后）： #{ | } ~ (0x7b-0x7e) #[ \ ] ^ - ` (0x5b-0x60) #(空格）! " # $ % & (0x20-0x26) #( ) \* + (0x28-0x2b) # : ; < = > ? @ (0x3a-0x40) # / (0x2f) $norm*text =~ s/([^0-9])([.,])/$1 $2 /g; # tokenize period and comma unless preceded by a digit #当非数字0-9后面紧跟着句号"."或者逗号","时,在句号或者逗号两边各加一个空格（即数字后面紧跟句号或逗号时，不在句号或逗号两边加空格） $norm*text =~ s/([.,])([^0-9])/ $1 $2/g; # tokenize period and comma unless followed by a digit #当句号"."或者逗号","后面没有紧跟数字0-9时，在句号或者逗号两边各加一个空格 $norm*text =~ s/([0-9])(-)/$1 $2 /g; # tokenize dash when preceded by a digit #当连字符号"-"前面紧跟数字0-9时，在"-"两边各加一个空格 $norm*text =~ s/\s+/ /g; # one space only between words #把多个空格替换成一个空格 $norm*text =~ s/^\s+//; # no leading space #把文章打头的空格去掉 $norm\_text =~ s/\s+$//; # no trailing space #把文章末尾的空格去掉 |
|  |

**三、CCMT 2019 XML文件文档结构描述说明**

|  |
| --- |
| <?xml version="1.0" encoding="UTF-8"?> <!ELEMENT srcset (DOC+)> <!ATTLIST srcset setid CDATA #REQUIRED> <!ATTLIST srcset srclang (en | zh | mn | uy | ti | jp) #REQUIRED> <!ATTLIST srcset trglang (en | zh | mn | uy | ti | jp) #REQUIRED> <!ELEMENT refset (DOC+)> <!ATTLIST refset setid CDATA #REQUIRED> <!ATTLIST refset srclang (en | zh | mn | uy | ti | jp) #REQUIRED> <!ATTLIST refset trglang (en | zh | mn | uy | ti | jp) #REQUIRED> <!ELEMENT tstset (DOC+)> <!ELEMENT tstset (system+)> <!ATTLIST tstset setid CDATA #REQUIRED> <!ATTLIST tstset srclang (en | zh | mn | uy | ti | jp) #REQUIRED> <!ATTLIST tstset trglang (en | zh | mn | uy | ti | jp) #REQUIRED> <!ELEMENT system (#PCDATA) > <!ATTLIST system site CDATA #REQUIRED > <!ATTLIST system sysid CDATA #REQUIRED > <!ELEMENT DOC ( p\* )> <!ATTLIST DOC docid CDATA #REQUIRED> <!ATTLIST DOC site CDATA #IMPLIED> <!ELEMENT p(seg\*)> <!ELEMENT seg (#PCDATA)> <!ATTLIST seg id CDATA #REQUIRED> |
|  |