**附件4：翻译质量评估任务数据文件格式说明**

本文档对该评测任务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。

1. **词汇级翻译质量估计项目数据文件格式说明**

**1、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“汉-英词汇级翻译质量评估”项目为例说明这些文件的格式。

* 1. **训练集**

训练集由四个文件组成，包括由逐行对应的源语言文件train.source、译文文件train.target、对译文进行人工译后编辑的文件train.pe，每行为一个句子；以及待评估译文的Tags文件train.tags。文本数据均已分词，小写化处理。Tags文件每行为一个OK/BAD序列，包含译文词和词间的标注；序列奇数位置为词间标注，偶数位置为译文词标注，按序排列。

另外，Tags标注是使用TERCOM (http://www.cs.umd.edu/~snover/tercom)工具计算得到的，采用如下设置：分词（tokenized），大小写不敏感（case insensitive），精准匹配 （exactly matching only），使用“-d 0 ” 选项禁用shift（disabling shifts by using the ‘-d 0’ option）。

图1、图2、图3、图4示例说明了“汉-英翻译词汇级别质量评估”项目中源语言文件、译文文件、译后编辑文件、Tags文件的格式。

|  |
| --- |
| train.source |
| 文化 产业 的 高质量 发展 , 离不开 空间 载体 的 支撑 。  2018 年 , 全市 文化 产业 收入 保持 两位数 增长 …… |
| 图1 训练集的源语言文件 |

|  |
| --- |
| train.target |
| high quality develop of cultural industry , support without space carrier . in 2018 , the city 's cultural industry income maintained double-digit growth . …… |

图2 训练集的译文文件

|  |
| --- |
| train.pe |
| support of space carriers is indispensable for the high-quality development of the cultural industry .  in 2018 , the city 's cultural industry income maintained double-digit growth . …… |

图3 训练集的译后编辑文件

|  |
| --- |
| train.tags |
| BAD BAD OK BAD OK BAD OK BAD OK BAD OK BAD OK BAD OK BAD OK BAD OK BAD OK BAD OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK OK …… |

图4 训练集的Tags文件

* 1. **开发集**

其文件个数及格式与训练集相同，包括dev.source、dev.target、dev.pe、dev.tags四个文件。

* 1. **测试集**

在评测阶段组织方将只发放test.source和test.target两个文件，其格式与训练集相同。

**2、参评单位需要提交的数据格式说明**

参评单位可以选择提交任意子任务的预测结果（译文词或者译文词词间的OK/BAD 二分类标注），也可以提交两个子任务的全部预测结果。参评单位仅需要提供最终的翻译质量评估结果文件及系统描述信息文件，其格式说明如下。

**2.1 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-qe为例，参评单位代号以ict为例）

|  |  |  |
| --- | --- | --- |
| 文件 | 文件名模式 | 文件名举例 |
| 翻译质量评估  译文词标注结果 | 项目代号-参评单位代号-test.word | ce-qe-ict-test.word |
| 翻译质量评估  译文词词间标注结果 | 项目代号-参评单位代号-test.gap | ce-qe-ict-test.gap |
| 系统描述信息 | 项目代号-参评单位代号-word-sys.txt | ce-qe-ict-word-sys.txt |

**2.2 最终提交结果文件**

词汇级翻译质量评估文件以以下格式提交：

<METHOD NAME> <SUBTASK> <SEGMENT NUMBER> <WORD/GAP INDEX> <WORD> <BINARY TAG>

每个字段应该由一个制表符分隔。其中，

* <METHOD NAME> 表示质量评估方法。
* <SUBTASK> 表示词汇级别质量评估子任务类型；假如是译文词标注，为“word”；假如是词间标注，为“gap”。
* <SEGMENT NUMBER> 表示测试集译文的句子行号，从0开始。
* <WORD/GAP INDEX> 表示译文词或者译文词词间的标注在译文中的序号，从0开始。
* <WORD> 表示评估内容；假如是译文词标注，为该译文词；假如是词间标注，为“NULL”。
* <BINARY TAG> 表示译文词或者词间标注结果，为“OK”或者“BAD”

举例：

译文词标注文件：

word\_base word 0 0 high BAD

word\_base word 0 1 quality BAD

word\_base word 0 2 develop BAD

word\_base word 0 3 of BAD

……

译文词词间标注文件：

gap\_base gap 0 0 NULL BAD

gap\_base gap 0 1 NULL OK

gap\_base gap 0 2 NULL OK

gap\_base gap 0 3 NULL OK

……

在参评系统的描述信息文件中，需要对以下内容给出说明：

* 软硬件环境：包括操作系统及其版本、所用开发环境和开发软件等等；
* 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；
* 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，是否使用额外的数据资源，也需要进行说明。

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。

**二、句子级翻译质量估计项目数据文件格式说明**

**1、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。此外，以“汉-英翻译质量评估”项目为例说明这些文件的格式。

**1.1 训练集**

　　训练集由四个文件组成，包括由逐行对应的源语言文件train.source、译文文件train.target、对译文进行人工译后编辑的文件train.pe，每行为一个句子；以及待评估译文的HTER值文件train.hter，每行为一个区间[0,1]的数值。

　　其中，HTER值是使用TERCOM (<http://www.cs.umd.edu/~snover/tercom>)工具计算得到的，采用默认设置，即大小写不敏感、精确匹配，且值区间为[0,1]。汉英翻译结果以单词为单位进行计算，英汉翻译结果以字符为单位进行评估。英文句子进行tokenize的标准请参照翻译任务格式说明的相应描述。

　　图1、图2、图3、图4示例说明了“汉-英翻译质量评估”项目中源语言文件、译文文件、译后编辑文件、HTER文件的格式。

|  |
| --- |
| train.source |
| 一是树立健康的消费理念。 一种信息推送的方法及装置  …… |
|  |

图1 训练集的源语言文件

|  |
| --- |
| train.target |
| One is sets up the healthy consumption concept.  A method and a device for information push  …… |

图2 训练集的译文文件

|  |
| --- |
| train.pe |
| First one is to set up the healthy consumption concept.  A method and device for information pushing  …… |

图3 训练集的译后编辑文件

|  |
| --- |
| train.hter |
| 0.3636  0.2857  …… |

图4 训练集的HTER文件

**1.2 开发集**

　　其文件个数及格式与训练集相同，包括dev.source、dev.target、dev.pe、dev.hter四个文件。

**1.3 测试集**

　　在评测阶段组织方将只发放test.source和test.target两个文件，其格式与训练集相同。

**2、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翻译质量评估结果文件及系统描述信息文件，其格式说明如下。

**2.1 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-qe为例，参评单位代号以ict为例）

|  |  |  |
| --- | --- | --- |
| 文件 | 文件名模式 | 文件名举例 |
| 翻译质量评估结果 | 项目代号-参评单位代号-test.hter | ce-qe-ict-test.hter |
| 系统描述信息 | 项目代号-参评单位代号-sys.txt | ce-qe-ict-sys.txt |

**2.2 最终提交结果文件**

　　翻译质量评估结果文件为测试集中每一行待评估译文句子对应的预测的HTER值。

　　翻译质量评估结果文件格式与训练集的train.hter以及开发集的dev.hter相同，即每一行为一个区间[0,1]的数值。

　　在参评系统的描述信息文件中，需要对以下内容给出说明：

* 软硬件环境：包括操作系统及其版本、所用开发环境和开发软件等等；
* 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；
* 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，是否使用额外的数据资源，也需要进行说明。

　　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。