

文言文翻译工具包

清华大学自然语言处理与社会人文计算实验室

2017 年 8 月

1 介绍

本工具包采用 moses 翻译工具¹进行翻译，使用 55 万文言-白话句对进行训练，达到了较好的翻译效果。并且，本工具包中包含已经训练完毕的模型，不需要用户提供数据或者是模型，通过简单的命令行就可以使用。

2 使用指南

2.1 解压缩

使用解压缩命令 `tar -zxvf guwen_translate.tar.gz` 进行解压。

2.2 设置路径

假设解压路径为 `/home/Tom/projects/guwen_translate/`，将 `/home/Tom/projects/guwen_translate/code/translate.py` 中第 7 行 `root='/path/to/guwen_translate'` 替换为 `root='/home/Tom/projects/guwen_translate/'`。

2.3 文件翻译

本工具包可以通过简单的命令行完成以文件为输入输出的翻译，在文件翻译时需要提供以下参数：

- i 或 --input-file 输入文件路径，默认为 `/path/to/guwen_translate/code/input.txt`，
- o 或 --output-file 输出文件路径，默认为 `/path/to/guwen_translate/code/output.txt`，
- r 或 --reference-file 参考翻译路径，可选。

其中，输入文件不需要进行分词，但需要每句文言文独占一行，并且请注意最后一句文言文的行末同样需要换行。输出文件是并未进行分词的白话文翻译结果。如果设置了参考翻译，则本工具包会自动计算 bleu 值并记录在 `/path/to/guwen_translate/code/evaluate.res` 中，请注意参考翻译需要进行分词。

以下是命令行指令的例子：

```
python /path/to/guwen_translate/code/translate.py -i /path/to/myinput.txt \
-o /path/to/myoutput.txt -r /path/to/myref.txt
```

2.4 单句翻译

本工具包支持通过简单的命令行进行单句翻译，在单句翻译时需要提供唯一的参数为：

- wenyan 需要翻译的文言文

在设置了需要翻译的文言文的情况下，文件翻译的三个参数即使设置也是无效的，工具包将直接在控制台中输出翻译后的白话文。命令行例子如下：

¹ <http://www.statmt.org/moses/>

```
python /path/to/guwen_translate/code/translate.py --wenyan 公所事者且十主,皆面谀  
以得亲贵
```