

# 基于词典和规则的汉语自动分词

孟磊 MF1833048

程序运行方式，目录下的 **seg** 为程序如果，参数为汉语句子，即 **./seg 句子**

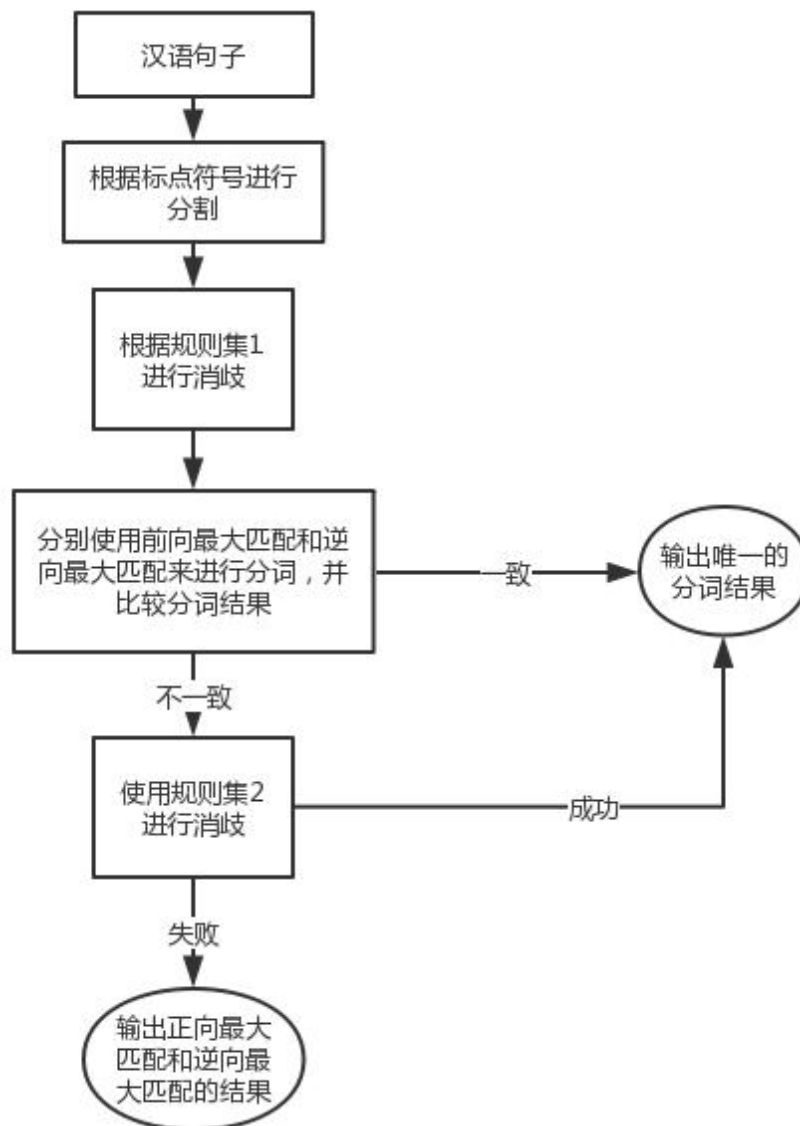
运行环境为 **python3.7**

## 1. 任务描述

实现程序，基于词典和规则对输入的汉语句子进行分词，并输出分词结果。

## 2. 技术路线

程序的流程为：

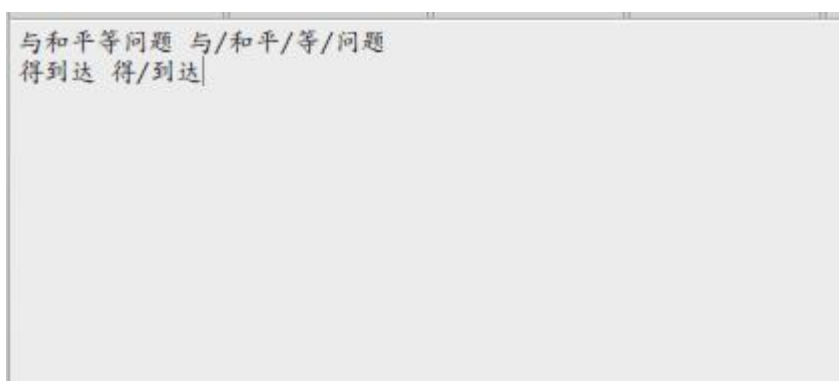


其中，规则集 1 中的规则是用于处理无法通过比较正向最大匹配（fmm）和逆向最大匹配（rmm）的结果来发现的歧义。如：一阵风 一/阵/风，把手移开 把/手/移开。

规则集 2 中的规则用于处理可以通过比较 fmm 和 rmm 的结果来发现的歧义。如：与和平等问题 与/和平/等/问题，得到达 得/到达。

所以对于要进行分词的句子，需要先通过规则集 1 来进行消歧。

规则集文件的结构为：歧义字段 分词方案（如下图所示）。



```
与和平等问题 与/和平/等/问题
得到达 得/到达
```

消歧的步骤为：依次将规则集中的歧义字段在句子中进行搜索，若找到，则将句子根据该歧义字段分成两段，然后分别对两段句子递归进行上述以及下面的步骤。若句子中不存在歧义字段，就可以进行 fmm 和 rmm 分词，此时两者的分词结果基本将会相同。

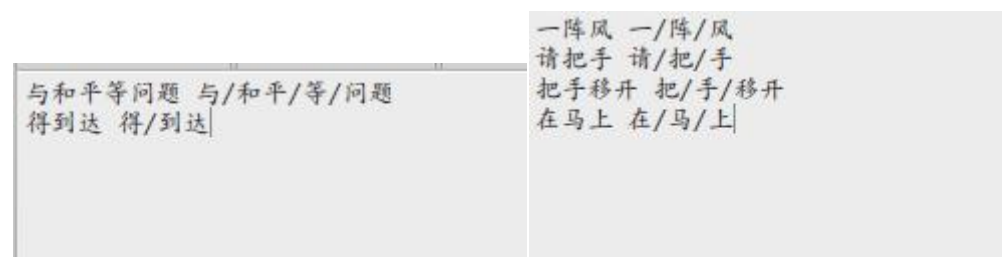
### 3. 用到的数据

（1）词典数据：

来源：<https://pan.baidu.com/s/1i37gKLZ> 目录下的：dic.txt

（2）规则数据

自己写了几条规则，分别是目录下的：guize.txt，guize1.txt



```
与和平等问题 与/和平/等/问题
得到达 得/到达

一阵风 一/阵/风
请把手 请/把/手
把手移开 把/手/移开
在马上 在/马/上
```

#### 4. 遇到的问题以及解决方案

(1) 有些歧义无法通过比较 fmm 和 rmm 的结果来发现,所以加了一组规则,在进行 fmm 和 rmm 之前使用,用于这些歧义的发现。

#### 5. 性能评价

添加规则集后,随着规则的增加,基本可以很准确的分词。但由于匹配规则的方法暂时为线性扫描比较,所以随着规则的增加,时间复杂度会线性增加。对于可发现的歧义,我认为可以试着给规则建立合适的索引,根据发生歧义的词和字来进行关键字查找。