# Penguin拼音标注软件

开题报告

# 为什么选拼音标注

- 本组员的目标
- 尚壬鹏：团队合作的经验
- 王栋、王长海：为实习找工作打基础
- 秦子童、邹永平、肖洒：积累提高linux下C语言编程
- 庞博、任望：熟悉开发流程、提高编程能力

# 为什么选拼音标注

- 综合考虑大家的需要，认为拼音标注最为符合大家的期望。
- 该任务难度较大，适合王栋、王长海的目标
- 任务中的合作任务可以提高我和庞博的团队合作能力
- 通过共同编程，可以提高秦子童和肖洒的linux下C语言编程能力

# 拼音标注问题的难点

- 切分方法
- 1） 正向最大匹配法（由左到右的方向）
- 2） 逆向最大匹配法（由右到左的方向）
- 3） 最少切分（使每一句中切出的词数最小）。

# 拼音标注问题的难点

- 用正向最大匹配法得到的结果是：
- 长春市/长春/节/致辞（分成4个词，其中“节”未匹配到，语义错误）
- 长春市/长春/药店（分成3个词，都匹配到，语义正确）

- 用逆向最大匹配法得到的结果是：
- 长春/市长/春节/致辞（分成4个词，都匹配到，语义正确）
- 长春/市长/春药店（分成3个词，都匹配到，语义错误）

# 拼音标注问题的难点

- 语料库结构未知，不知道是给语料库还是要自己做

# 注音软件怎么做

- 根据字库结构建模
- 最大正向匹配
- 最大反相匹配
- 结果筛选
- 命令行和界面两种使用方法
- 0.1版：字词统计，项目组上手
- 0.2版：句子切分，实现核心功能
- 0.3版：选音、保存，逻辑完善

谢谢大家!