基于文学作品的情景对话

肖达

怎样构建大规模个性化对话数据集？

- 可能的数据源
  - 经典文学作品
  - 影视剧本
  - 网文（言情类为主）
- 为什么选择网文？
  - 数据量大
  - 对话占比多，且对推动情节起主要作用
  - 对话基于特定场景/情节，且反映人物个性

前期工作

- 网文数据爬取
- 角色识别和抽取
- 性别识别
- 指代消解
- 对话提取
- 说话人识别
- 内心独白和独白人识别

待完善

- 角色别名识别
- 没有名字的人的识别
- 性别识别算法改进
- 指代消解算法改进
- 多人对话的提取和说话人识别

计划

- 构建基于文学作品的情景对话数据集
  - 有情节、有超长上下文、有动作、表情、内心os等非语言元素
- 提高transformer语言模型处理超长上下文的能力
- 训练基于作品人设和情节的对话互动bot

- 产品化的初步设想
  - 类似红袖男友的APP，人和bot的角色扮演游戏（多轮对话）

- 文学作品人物的性格分析和聚类
- 跑团