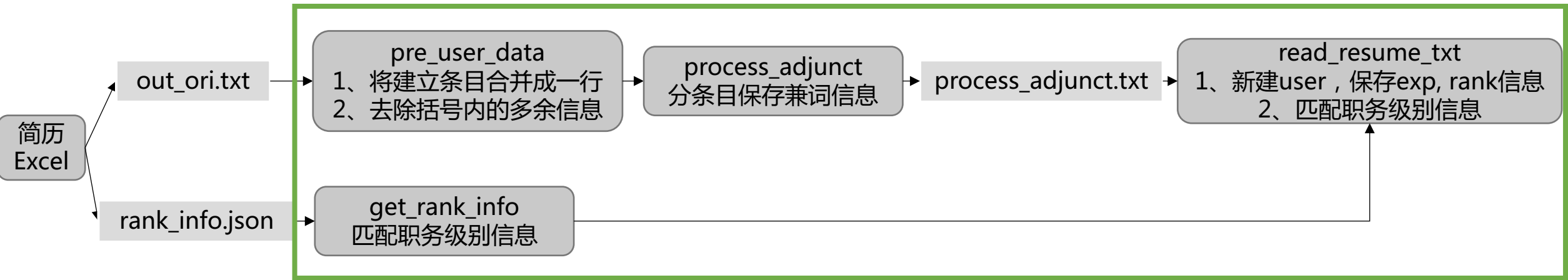
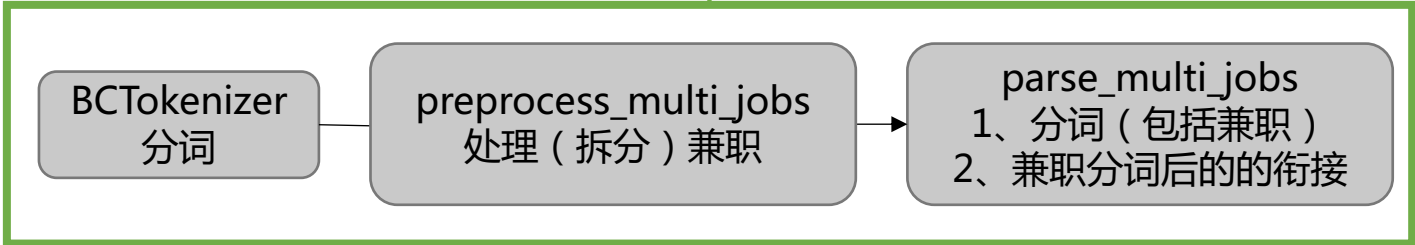


流程概览

数据前处理



Raw data写入catalog



分词 数据后处理

分词结果保存至catalog和数据库

新加入的模块

1、预处理阶段

➤读取、保存职务级别（rank）信息

rank_info.json、并将rank信息根据时间与experience信息匹配

如：

7272 | 12 | 2010.08—2010.11 | 2010.08—2010.11 深圳市福田区委区政府办公室综合科科长 | 深圳市L 福田区L 委区政府办公室S 综合科S 科长P | NULL | {'null': 3} |

7272 | 14 | 2012.04—2014.07 | 2012.04—2014.07 深圳市福田区委区政府办公室副主任 | 深圳市L 福田区L 委区政府办公室O 副主任P | NULL | {'副处级': 27}

新加入的模块

➤Logger

(a) 出现时间问题的log

```
2021-08-06 15:35:01,890 - Career_Platform.util.logger - ERROR - Invalid date format: 1988.17-1991.11 深圳市罗湖区黄贝街道办事处司法助理员
2021-08-06 15:35:01,890 - Career_Platform.util.logger - ERROR - 111.xls
2021-08-06 15:35:05,709 - Career_Platform.util.logger - ERROR - Date is after today: 1990.11-9506.03 深圳市荔园招待所所长
2021-08-06 15:35:05,709 - Career_Platform.util.logger - ERROR - 1.xls
```

(b) 分词拼接问题的log

```
2021-08-24 14:28:50,083 - Career_Platform.util.logger - INFO - L can not find L: 盐田港集团工作集团工会主管会计<end>|盐田港投资控股有限公司财务经理<end>
2021-08-24 14:28:50,083 - Career_Platform.util.logger - INFO - ['B0', 'M0', 'M0', 'M0', 'M0', 'M0', 'M0', 'M0', 'E0', 'BS', 'ES', 'BP', 'MP', 'MP', 'EP']|['BL', 'E
```

在7000人的数据上分词不能记录准确率，所以将分词不能拼接的数量作为间接的分词效果以及拼接效果的度量

优化ner分词模型以及拼接逻辑后大数据集上无法拼接语句约440条

新加入的模块

➤增量处理数据

根据不同的命令行选择读取和保存数据库的方式

`only_init_db` 建立数据库

`offline_process` 读入文本、清理格式、分词、打标签

`offline_process i` 清空并写入数据库

`offline_process a` 继续写入数据库

`offline_process u start_id` 更新数据库中从start_id开始的信息

新加入的模块

2、分词模型

➤分词标签的完善

原先ner标签包含SO，SL在内的18种字符标签，会导致分出单字的组织与地名等情形。在训练数据中进行标注修正，删去SO，SL使标签降低到16种，并配合转移矩阵确保不会出现任何单字分词成为实体的情形

现有字符标签16种：

["BO", "MO", "EO", "BP", "MP", "EP", "BS", "MS", "ES", "BL", "ML", "EL", "BU", "MU", "EU", "SU"]

新加入的模块

2、分词模型

➤分词标签的完善，加入四种U标签信息，在处理阶段进行删除

➤分词标注的完善以及后续校正

字符属性准确率提升至96%

整句分词（不计算字符属性）准确率达到94%

新加入的模块

➤转移矩阵的优化和完善

考虑了几乎所有转移对情况，并进行限制。其中与U相关转移由于识别率较低，删除了转移限制。

prohibitions = {

 "<start>": ["MO", "EO", "MP", "EP", "BS", "MS", "ES", "ML", "EL"],

 "EL": ["MO", "EO", "MP", "EP", "MS", "ES", "ML", "EL"],

 "EO": ["BO", "MO", "EO", "MP", "EP", "MS", "ES", "ML", "EL"],

 "ES": ["BO", "MO", "EO", "MP", "EP", "MS", "ES", "ML", "EL"],

 "EP": ["BO", "MO", "EO", "BP", "MP", "EP", "BS", "MS", "ES", "BL", "ML", "EL"],

}

禁止所有实体的结尾接上其它实体的中间或末尾，或者S实体后接O实体等不符合规则的实体连接，同时禁止句子开头出现单字实体情形。

新加入的模块

➤转移矩阵的优化和完善

```
restrictions = {  
    "BL": ["ML", "EL"],  
    "ML": ["ML", "EL"],  
    "BO": ["MO", "EO"],  
    "MO": ["MO", "EO"],  
    "BS": ["MS", "ES"],  
    "MS": ["MS", "ES"],  
    "BP": ["MP", "EP"],  
    "MP": ["MP", "EP"],  
}
```

- 禁止所有实体的中间部分接上其它实体的中间或末尾

在Debug过程中处理的问题

➤经历里含有双括号

如：深圳市益力矿泉水股份有限公司（深圳市益力实业有限公司）秘书、人事部干部（2001.10—2004.01在清华大学计算机技术工程领域专业学习，取得工程硕士）

➤对于时间缺失的补充

➤补充地名

如：市党委书记==>深圳市党委书记

➤兼词衔接的优化

如：LOSP|OLSP 共青团O深圳市L委员会S 等形式在子句中出现不进行拼接

在Debug过程中时处理的问题

➤整合类、项目文件

重点进行整合的类有：Catalog、BCTokenizer

目前脚本文件运行流程：demo/read_excel.py

demo/only_init_db.py

demo/offline_process.py i

network/octree.py

demo/query_social_networkAPI.py

工作与研究计划

➤NER方向

尝试词嵌入与字嵌入结合的方法

使用Chinese-Bert等中文预训练模型

尝试lattice/Flat等新分词模型

研究中文的嵌套分词等方法

➤Social network方向

Community detection

职业推荐