Career Platform

Offline Chinese Segmentation

Ke Li

- Why does recall rate decrease?
- Changes in the Offline Process
    a. Enlarge training set
    b. Multiplier from 2 to 15
    c. Split tree before compress
    d. Conditional compress (exclude suffix list)
    e. remove single word
- Results and comparison
- Further Work

Hybrid Model Performance Comparison (Before)

预分词	precision	recall	F-1 score
Jieba	30.93%	44.65%	35.93%
PKUseg	37%	61.2%	46.16%

OCtree into user dictionary

再分词	precision	recall	F-1 score
Jieba→ Jieba	43.76%	41.69%	42.7%
PKUseg → Jieba	44.86%	43.21%	44%

Why does recall rate decrease

Recall rate $\frac{TP}{TP+FN} = \frac{\text{实际分词命中的个数}}{\text{正确分词的个数}}$

IMPORTANT：

Note that the resume entry tend to be segmented into longer parts after the compressed OCtree was added in.

1. Proportion
2. Compress problems (main reason)

EXAMPLES:

Ground True

A= [“深圳市”, “罗湖区”, “广播电影电视集团有限公司”, “委员会”, “主席”]

Pre segment

B1 = [“深圳市”, “罗湖区”, “广播”, “电影”, “电视”, “集团”, “有限公司”, “委员会”, “主席”] 4

After Octree:

B2 = [“深圳市罗湖区”, “广播电影电视集团有限公司’’, “委员会’’, “主席”]

3

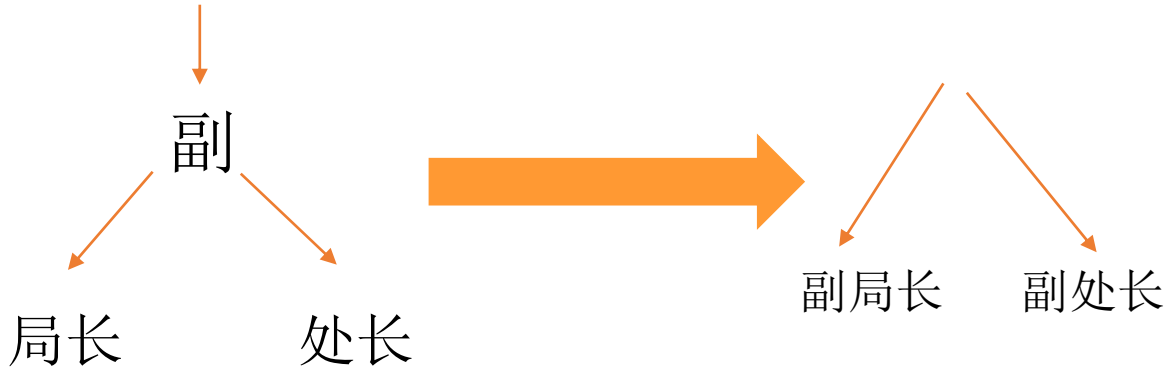
Changes in the Offline Process

- Conditional compress (exclude suffix list)

```
exclude_suffix = ['局','处','科','会','厅','委','司','室','省','市','县','区','所','队','部','系','大学','专业','学校','学院']
```

This way, “深圳市” “罗湖区” and “复旦大学” “新闻学专业” will not be compressed into one node.

- Split tree for single word



- remove single word, Increase train set and multiplier

These modifications can well solved the problem we talked in last meeting.

Results and Comparison (whole offline process)

预分词	precision	recall	F-1 score
Jieba	30.93%	44.65%	35.93%
PKUseg	37%	61.2%	46.16%

OCtree into user dictionary

	precision	recall	F-1 score
Jieba	43.76%	41.69%	42.7%
PKUseg	44.86%	43.21%	44%

Modifications talked in last slide

	precision	recall	F-1 score
Jieba	64.91%	68.56%	66.69%
PKUseg	68.41%	75.48%	71.77%

All Metrics increase a lot !!!

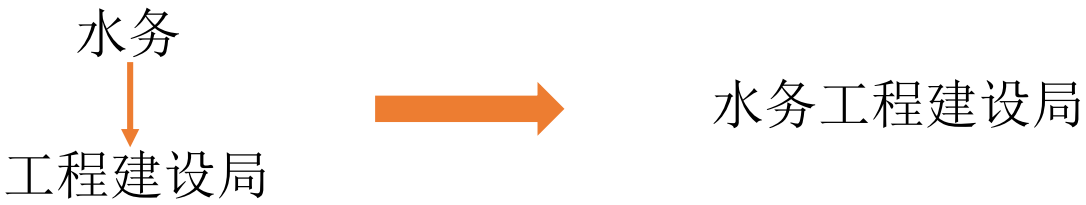
Real examples

['华南', '师范大学教育系教育', '专业', '学生']
['深圳市水务', '工程建设', '管理中心', '副', '主任']

- ['深圳市', '龙华新区', '观澜党工委', '副书记']
- ['深圳市', '龙华新区', '观澜党工委', '办事处', '副主任']
- ['深圳市', '罗湖区', '环境保护和水务局', '副局长']
- ['深圳市', '宝安区', '人民检察院', '渎职侵权检察科', '科长']
- ['深圳市', '深粮控股股份有限公司', '投资部', '部长']
- ['华融国际信托公司', '信托业务部门', '总经理', '助理']
- ['深圳市', '罗湖区', '教育局', '成教与社会办学科', '科长']

Further work Problems in not compressing leaf nodes

We compress the node and its child when the node has only one child.



we would not compress the last two nodes because we pre-assume that only the leaf node represents a position. (why?)

Therefore, we still separate “总经理” and “助理” in the final segmentation results. In the future, we consider trying part-of-speech tagging to solve it.