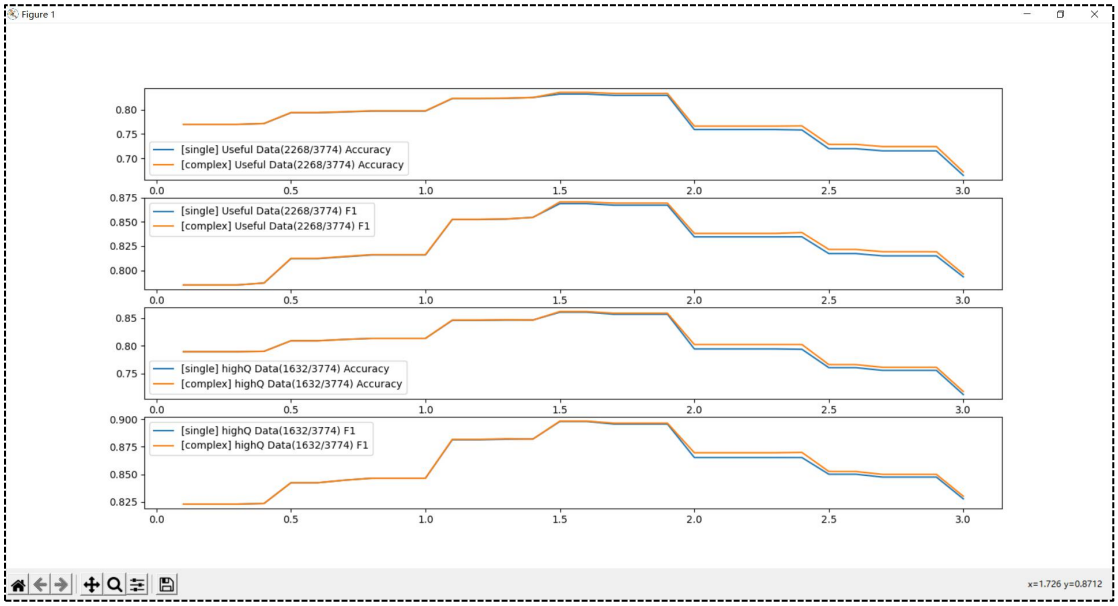
2021.07.20 ~ 2021.08.04 工作进展

1 扩充标准答案，校验数据

2 设计距离度量算法

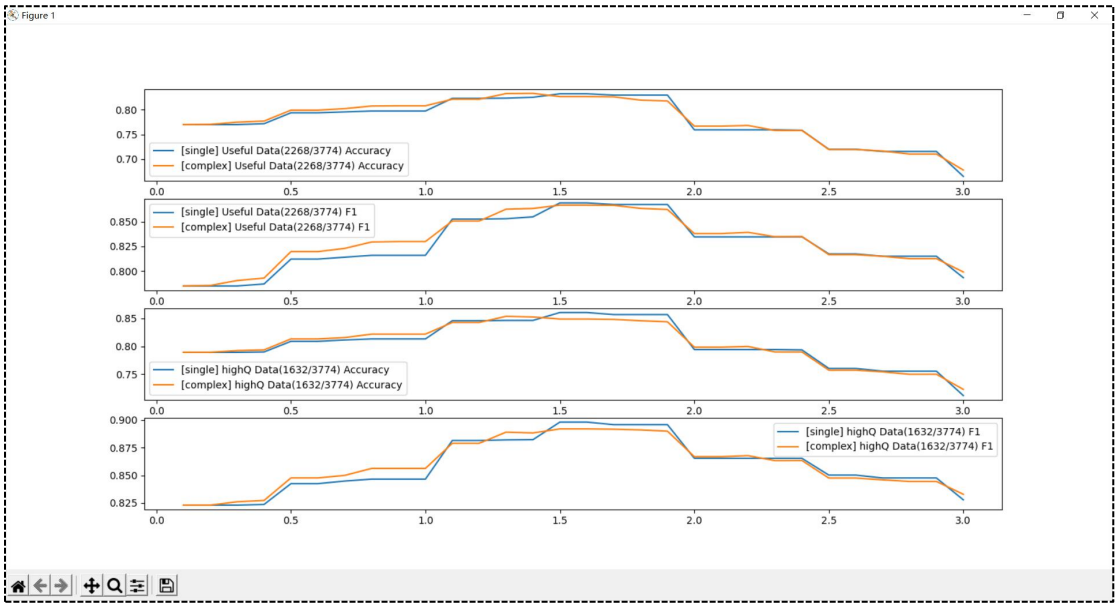
将单字拼音拆为声母和韵母，分别求编辑距离。尽管汉字拼音之间的距离难以人工加权，但由于声母、韵母数量不多（仅仅是组合 声母 * 韵母 的组合情况多），可以人工分别维护不同版本的距离权值矩阵，进行测试。

2.1 原始拼音 vs 声母、韵母分离



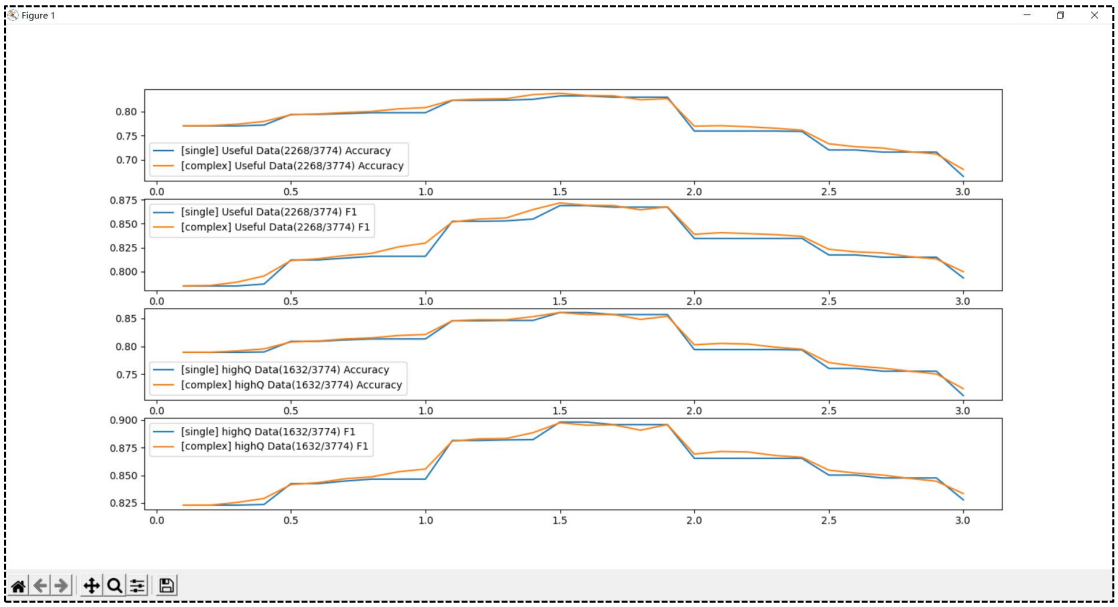

```
---------All Data Best Performance---------
[single] threshold: 1.50 , accuracy: 83.20%
[complex] threshold: 1.50 , accuracy: 83.55%


[single] threshold: 1.50 , F1: 86.89%
[complex] threshold: 1.50 , F1: 87.07%


---------HighQ Data Best Performance---------
[single] threshold: 1.50 , accuracy: 86.03%
[complex] threshold: 1.50 , accuracy: 86.15%


[single] threshold: 1.50 , F1: 89.79%
[complex] threshold: 1.50 , F1: 89.83%
```

2.2 原始拼音 vs 声母、韵母分离（相似拼音按 0.5，1.5 等比例加权）




```
---------All Data Best Performance---------
[single] threshold: 1.50 , accuracy: 83.20%
[complex] threshold: 1.40 , accuracy: 83.29%


[single] threshold: 1.50 , F1: 86.89%
[complex] threshold: 1.50 , F1: 86.67%


---------HighQ Data Best Performance---------
[single] threshold: 1.50 , accuracy: 86.03%
[complex] threshold: 1.30 , accuracy: 85.36%


[single] threshold: 1.50 , F1: 89.79%
[complex] threshold: 1.50 , F1: 89.18%
```

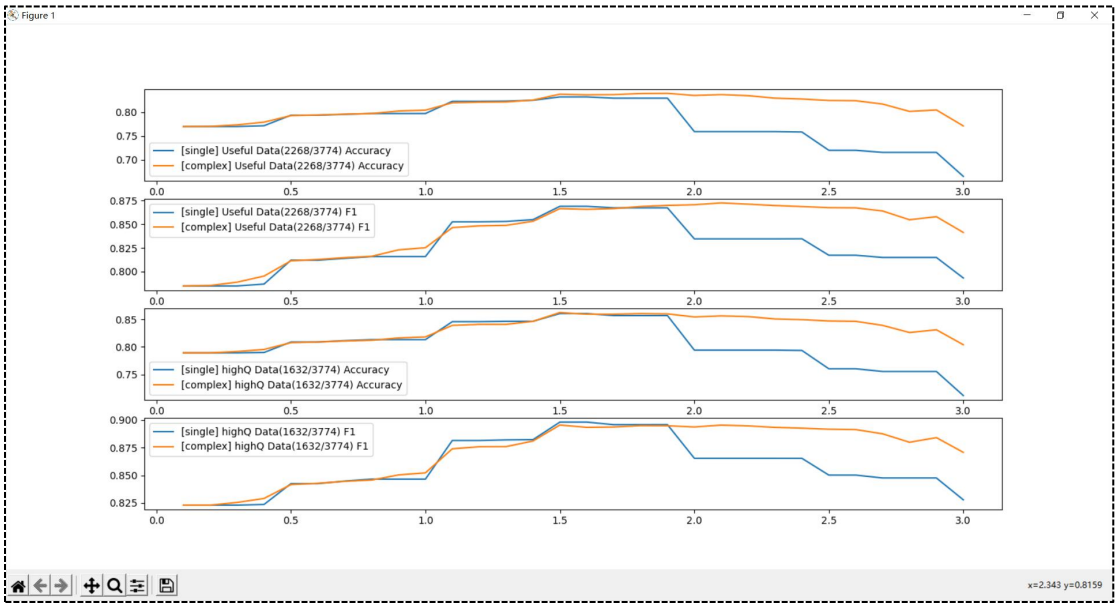
2.3 原始拼音 vs 声母、韵母分离（相似拼音按各种比例加权）




```
---------All Data Best Performance---------
[single] threshold: 1.50 , accuracy: 83.20%
[complex] threshold: 1.50 , accuracy: 83.73%


[single] threshold: 1.50 , F1: 86.89%
[complex] threshold: 1.50 , F1: 87.17%


---------HighQ Data Best Performance---------
[single] threshold: 1.50 , accuracy: 86.03%
[complex] threshold: 1.50 , accuracy: 86.03%


[single] threshold: 1.50 , F1: 89.79%
[complex] threshold: 1.50 , F1: 89.73%
```

3 查阅拼音相似度研究相关论文

当声母和韵母距离都大于 0 时，添加惩罚机制，即总距离*1.5




```
---------All Data Best Performance---------
[single] threshold: 1.50 , accuracy: 83.20%
[complex] threshold: 1.90 , accuracy: 83.95%


[single] threshold: 1.50 , F1: 86.89%
[complex] threshold: 2.10 , F1: 87.25%


---------HighQ Data Best Performance---------
[single] threshold: 1.50 , accuracy: 86.03%
[complex] threshold: 1.50 , accuracy: 86.21%


[single] threshold: 1.50 , F1: 89.79%
[complex] threshold: 1.50 , F1: 89.53%
```