Yiran Fei
fyraimar
October 10, 2014

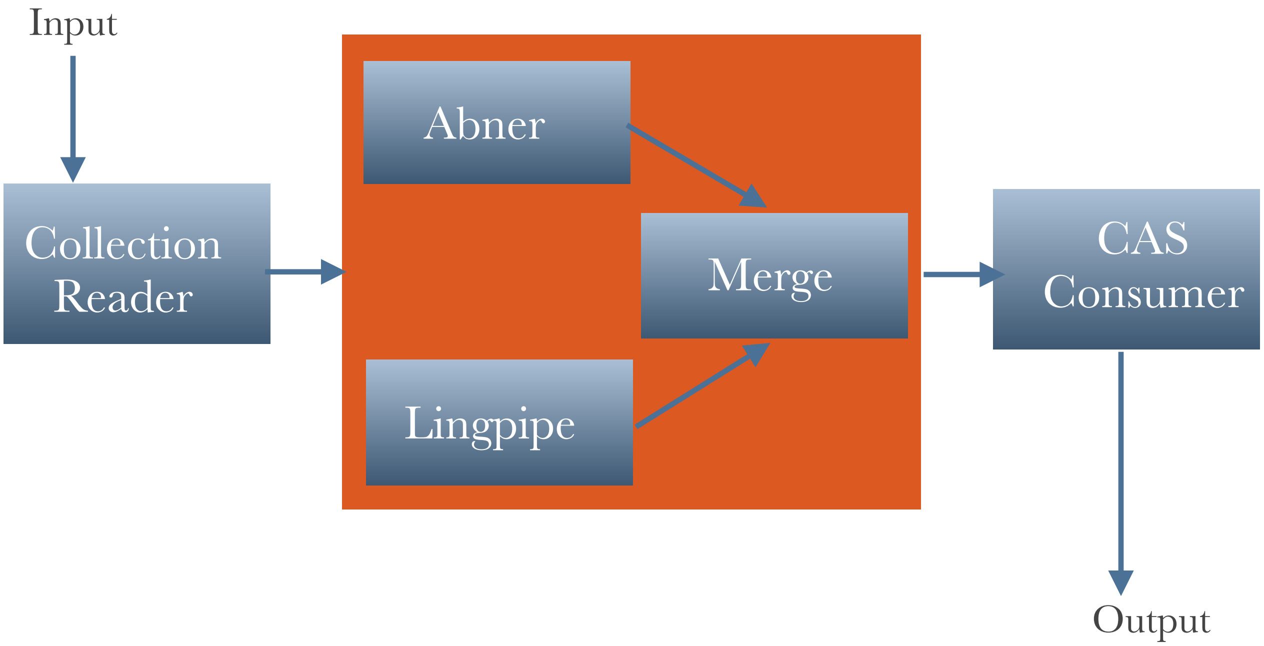
HW2-Report

11-693 Software Methods for Biotechnology

1. Requirement

Design and implement an aggregate analysis engine in UIMA framework to detect Gene name entity in given dataset.

2. Architecture Diagram



3. Design

3.1 Type System

In this system, there are 2 different AEs and a merge engine, so it's important to create a middle Type which can store the information of engines number and confidence value for a candidate token. The other par are quiet similar to the Type system in hw1.

3.2 Collection Reader

The collection reader retrieves information from the original dataset sentence by sentence, and it also separates the sentence ID and context. The ID is sent directly to the consumer and the context is used by the aggregate AE.

3.3 Aggregate Analysis Engine

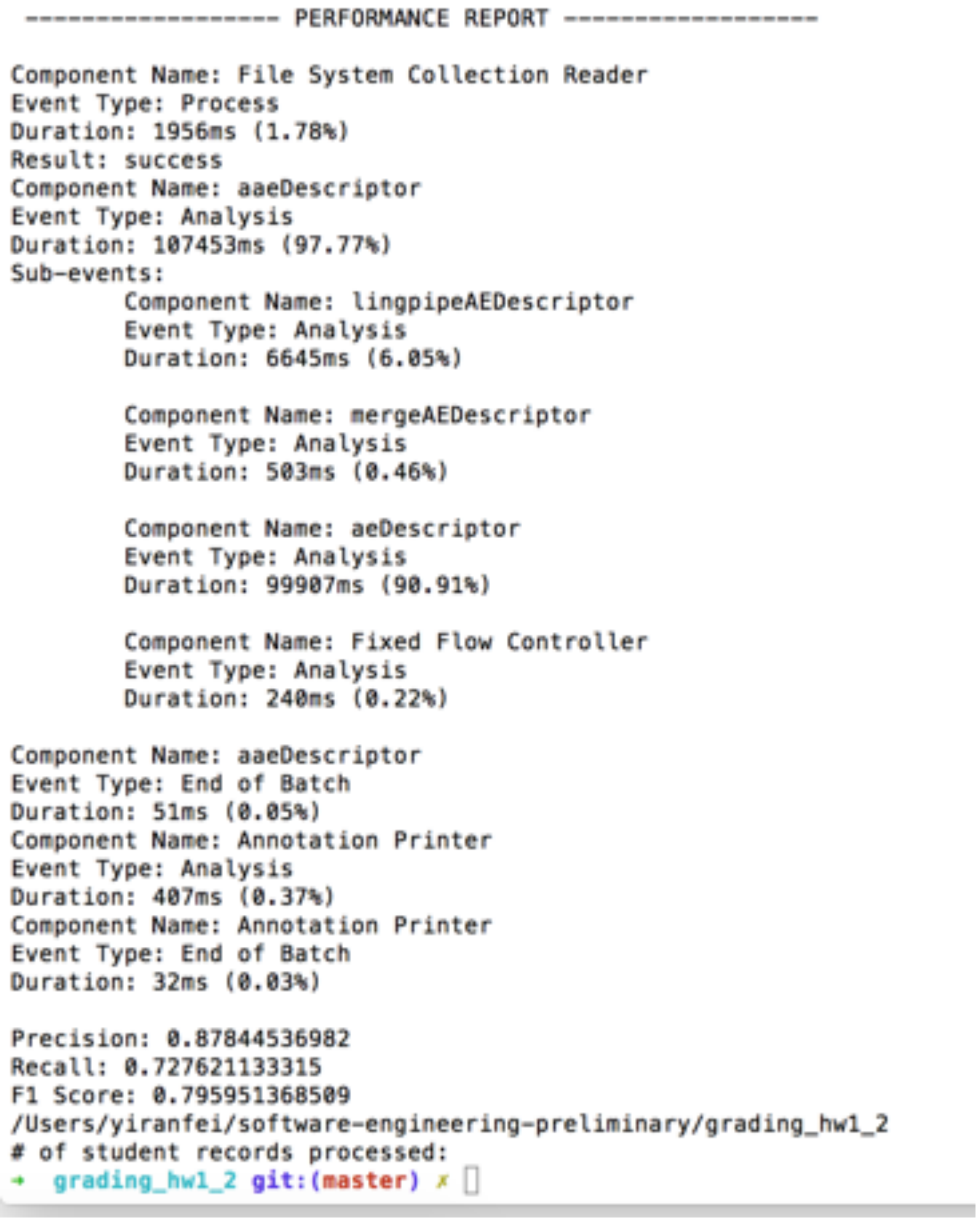
The AAE is designed to evaluate a token using many different tools. In this project, 2 AEs work in a pipeline and output a score for each token. The merge AE evaluate each candidates according to their values and marks some high-score entities as Gene name.

3.4 CAS Consumer

Same with the hw1, the CAS Consumer is just a file printer. It retrieves the gene names from the AAE and the sentence ID from the reader, erases the whitespace, calculates the positions and writes them into an output file.

4. Evaluation

I used to run the evaluation with my own program, but now there's a grading script that can test the path problem and do the evaluation. The following are a snapshot of the result of the grading script:



```
 ------------------ PERFORMANCE REPORT ------------------

Component Name: File System Collection Reader
Event Type: Process
Duration: 1956ms (1.78%)
Result: success
Component Name: aaeDescriptor
Event Type: Analysis
Duration: 107453ms (97.77%)
Sub-events:
        Component Name: lingpipeAEDescriptor
        Event Type: Analysis
        Duration: 6645ms (6.05%)

        Component Name: mergeAEDescriptor
        Event Type: Analysis
        Duration: 503ms (0.46%)

        Component Name: aeDescriptor
        Event Type: Analysis
        Duration: 99907ms (90.91%)

        Component Name: Fixed Flow Controller
        Event Type: Analysis
        Duration: 240ms (0.22%)

Component Name: aaeDescriptor
Event Type: End of Batch
Duration: 51ms (0.05%)
Component Name: Annotation Printer
Event Type: Analysis
Duration: 407ms (0.37%)
Component Name: Annotation Printer
Event Type: End of Batch
Duration: 32ms (0.03%)

Precision: 0.87844536982
Recall: 0.727621133315
F1 Score: 0.795951368509
/Users/yiranfei/software-engineering-preliminary/grading_hw1_2
# of student records processed:
→  grading_hw1_2 git:(master) ✗ □
```