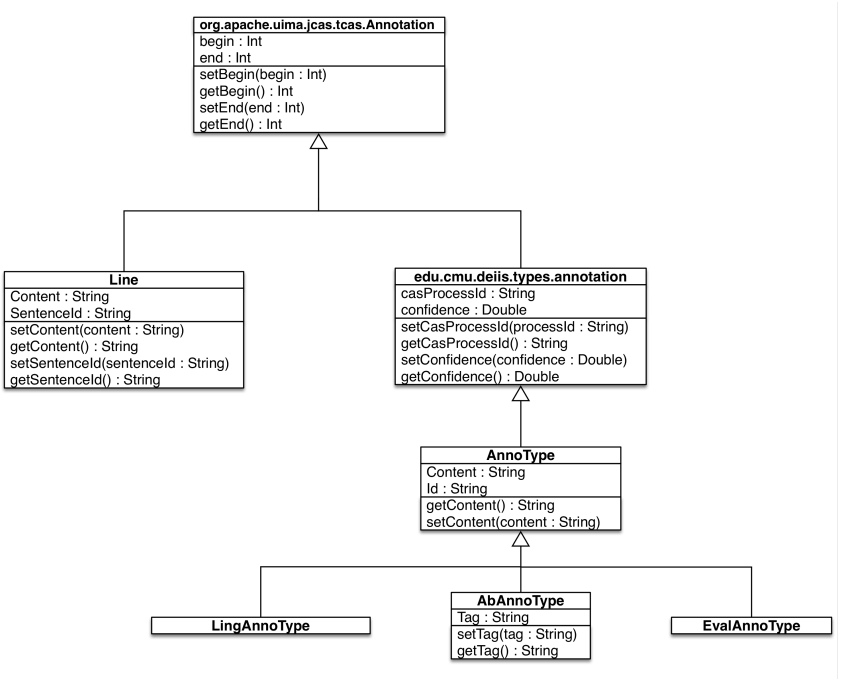
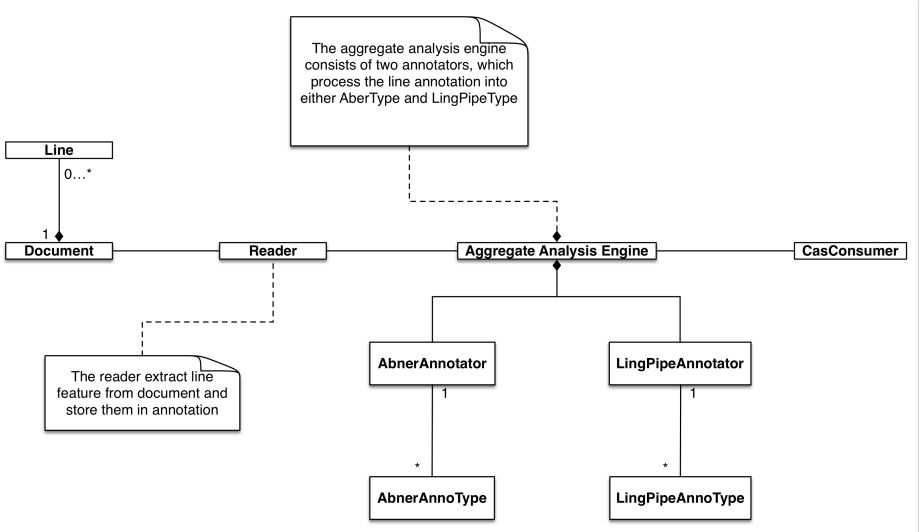
Xinyun (Victor) Zhao

HW2 Final Report

Type System Class Diagram



Collection Processing Engine Design Diagram



Type System

The pipeline has a pipeline of two phases. The first one extracts a gene feature from the type of *Line* (created by collection reader) and add to one type call *LingType*, which is processed by *LingAnno*. On the other hand, the annotation of *AbAnnoType*, produced by *AbAnno* could also be created based on

the features of *LineType*. Finally, the annotations merge into a single type called *EvalType*, which synthesis the all of the feature that CasConsumer needs. Since *AbAnnoType* is the only type that has unique *Tag* feature comparing to other types of annotation, a base type class that inherited from *edu.cmu.deiis.annotation* has been created with additional Id feature and Content feature, which reduce the redundancy of the type system.

Collection reader

Comparing to the design in HW1, the reader will read one line of document into system at one time rather than read the whole document together, which takes the advantage of the CPE pipeline.

Aggregate Annotators

The aggregate analysis engine consists of three annotators.

The first one is based on LingPipe, which recognize the name entity based on the metric of confidence by Confidence Named Entity Chunking.

The second annotator is based on Abner. This annotator will give each entity a special tag based on its category of biological terms.

The last annotator will merge the results produced by the previous two analysis engines. In terms of the threshold of confidence, all of the entities which have a confidence below that threshold will be excluded from the results given by LingPipe-based annotator. However, the corresponding results from Abner-based annotator are taken into consideration by the last annotator, which may increase the F-measure result.

CAS Consumer

CASConsumer will write all of the gene that the system analyzed into one single file according to the given format. In addition, the precision, recall and F-measure can also be calculated according to the given sampleout file.

Running result

```
Precision: 0.614186064144
Recall: 0.886996988776
F1 Score: 0.725802477432
```

Miscellaneous

Comparing to the LingPipe package, Abner performs worse either in the speed of process or the accuracy. Before I designed this parallel pipeline, where LingAnno and AbAnno process the annotation produced from the reader individually, the original pipeline was sequential. The AbAnno can receive the annotation from LingAnno and refine it. Nevertheless, the speed of this old pipeline is 10 time slower than the parallel version. Therefore, the previous design has been discarded.

Reference:

Alias-i, L. (2008). 4.1. 0. URL http://alias-i. com/lingpipe.

ABNER: A Biomedical Named Entity Recognizer. (n.d.). Retrieved October 9, 2014. URL http://pages.cs.wisc.edu/~bsettles/abner/