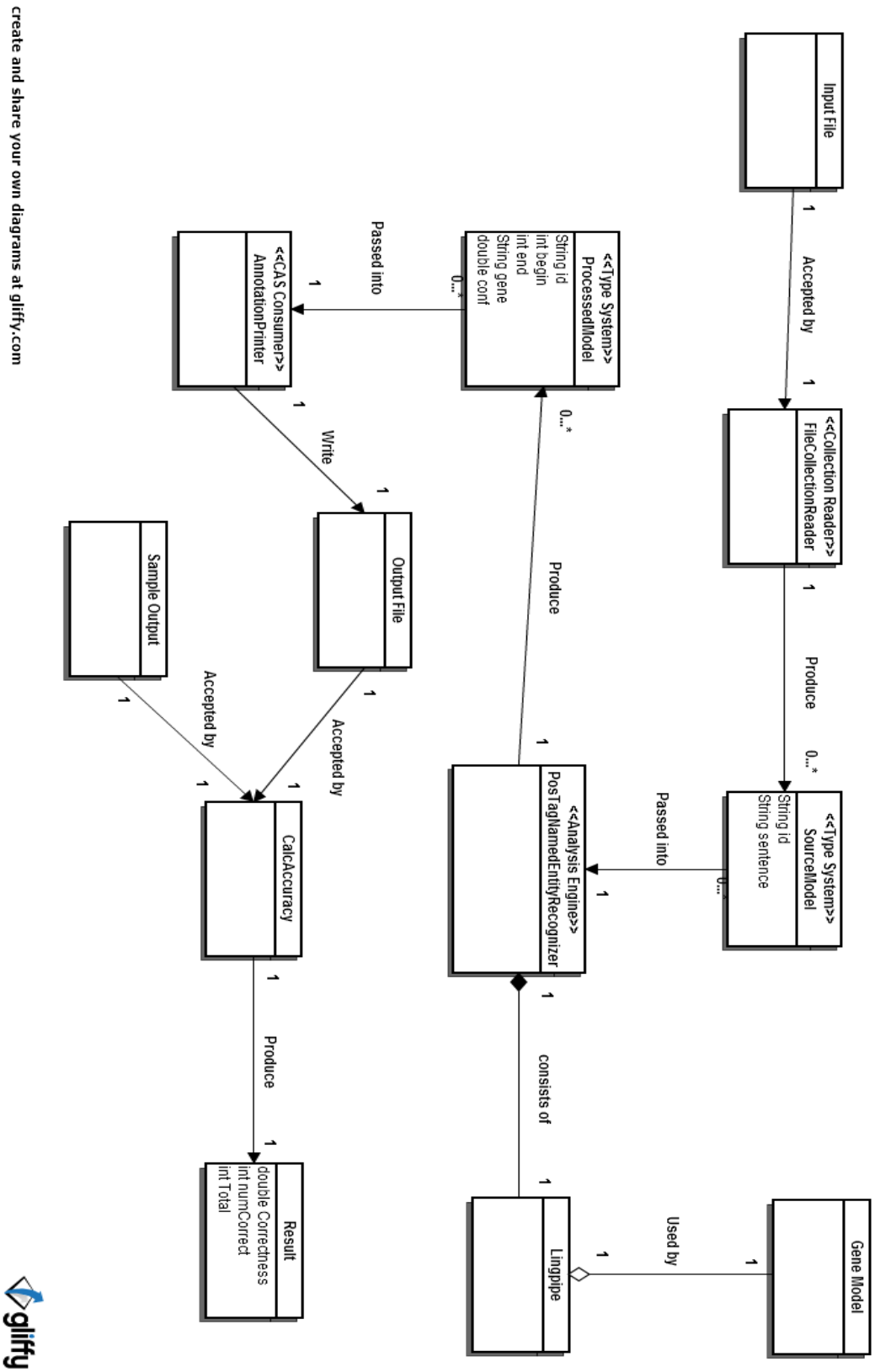
2012

Carnegie Mellon
University

Yang Sun
<yksun@cs.cmu.edu>

[11791 HW1 PROJECT REPORT]

General Data Flow / UML Domain Model



Type System

Two kinds of UIMA types are implemented in my program.

SourceModel

SourceModel is the data type that CollectionReader (FileCollectionReader) produces and is passed into Analysis Engine (PosTagNameEntityRecognizer). There are two features in this type: id and sentence, representing the sentence id and the content of the sentence.

ProcessedModel

ProcessModel is the data type that Analysis Engine (PosTagNameEntityRecognizer) produces and is passed into CAS Consumer (AnnotationPrinter). There are 3 self-defined features: id, gene and conf plus 2 default features: begin and end. Id feature is the same as the one defined in SourceModel. Gene represents the gene name mention and conf stands for the confidence of the NER (lingpipe) to its determination of the classification. Beigin and end are the offsets described in the project handout.

Collection Processing Engine (CPE)

The CPE defines the data flow of the project, including the collection reader and casProcessors. From the performance perspective, I set the casPoolSize as 2 and processingUnitThreadCount as 1. The casProcessor is integrated with the main process.

Collection Reader

The Collection Reader takes a String of file path as the input, read each line of the file by using Scanner. Then it separates the sentence id from the actual content of the sentence. Finally, the collection reader generates many SourceModel instances and proceeds to the next phase.

Analysis Engine / Intermediate Annotator

The PosTagNamedEntityRecognizer receives many SourceModel instances from the collection reader and processes each of them by using lingpipe with gene mention model. Lingpipe will provide a confidence level for each determination, where I set a threshold of 0.65 to maximize the correctness. The PosTagNamedEntityRecognizer will finally output many ProcessedModel to the next phase to proceed.

Confidence Threshold

There are some restrictions on setting the confidence threshold. By using Lingpipe's confidence mechanism, it is possible to generate more than one gene names with overlapped offset range, which is not favorable. However, this will not be a problem if the threshold is set to 50+%. Because all gene names with overlapped offset range will share up to 100% confidence, 50+%

confidence threshold guarantees that there can be at most one gene name produced within the same offset range.

Cas Consumer

The AnnotationPrinter takes a String of the file path as the input, write each ProcessedModel passed from Analysis Engine to that file. The output line format has been defined in the ProcessedModel. The output file path parameter is defined in the corresponding descriptor xml file.

Performance / Evaluation

I created a separate class from the entire project to evaluate the correctness of the analysis, named CalcAccuracy. This class takes two files: project output file and sample output file as the input, maps each line into an independent HashMap. That means, there will be two Maps, representing the content of two files in the class. Then CalcAccuracy will go through each of them to accumulate the total and the number of matched items. The correctness is calculated as (number of matched items) / total. The following is the correctness of my current program analysis:

```
Correct Number: 29420
Total Number: 36000
The correctness rate is 81.7222222222223%
```

How to Run

In Eclipse, create two new Java Applications in Run Configuration. One is for SimpleRunCPE and the other is for CalcAccuracy.

SimpleRunCPE:
Main class: SimpleRunCPE
Arguments: src/main/resources/CpeDescriptor.xml

CalcAccuracy:
Main class: CalcAccuracy
Arguments: sample.out hw1-yksun.out