# Homework 1

## 11-791 Design and Engineering of Intelligent Information System Fall 2012

## UML Design and Named Entity Recognition Implementation with UIMA SDK

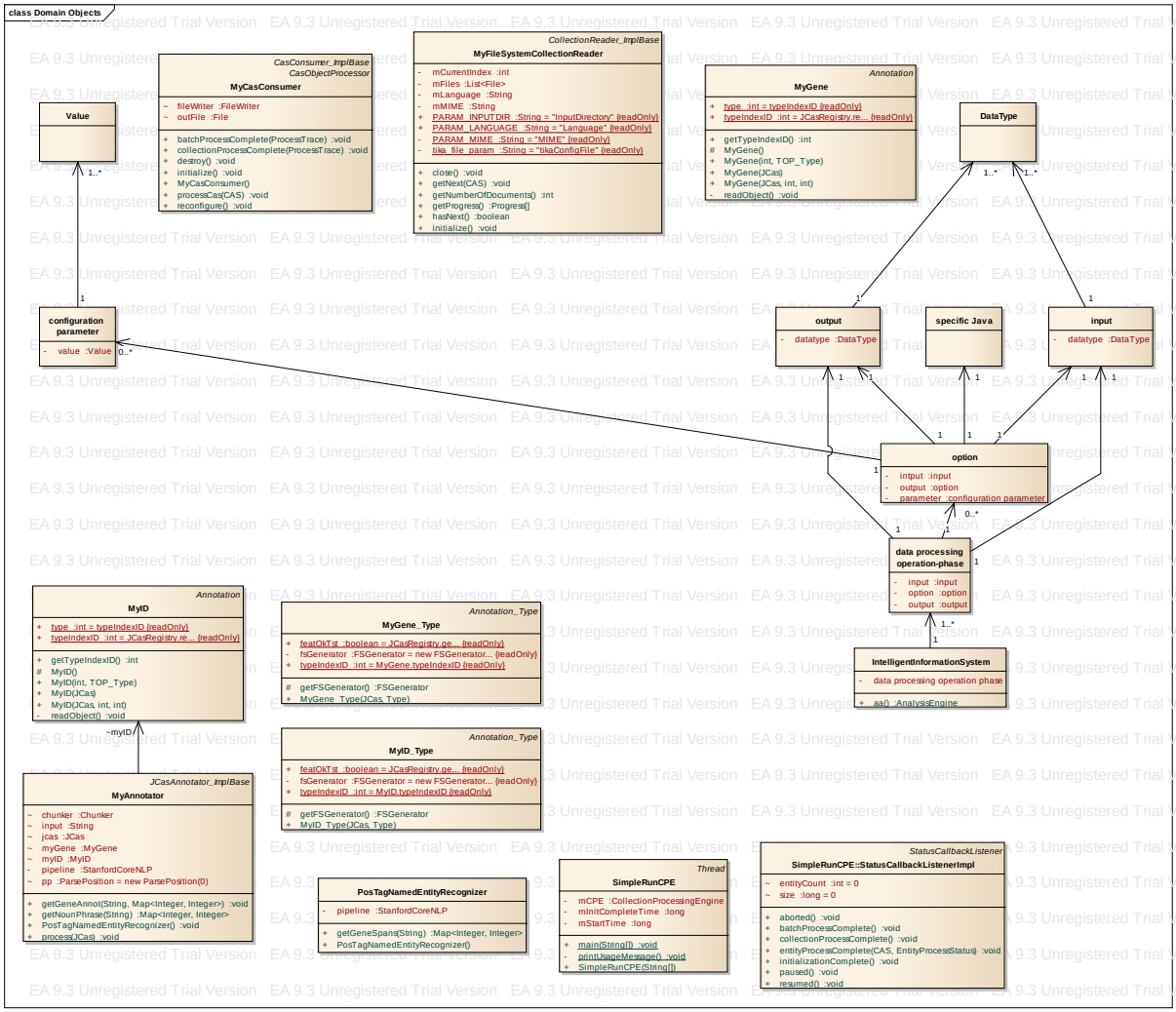

**Suyoun Kim**
**Andrew id: suyoung1**

**Language Technologies Institute**
**School of Computer Science**
**Carnegie Mellon University**
**suyoun@cs.cmu.edu**

Suyoun Kim (andrewID: suyoung1)

# 1. architectural aspects

## 1.1 UML



### 1.2 type system

For each sentences and gene tags, base annotation type is used. This type has two kinds of string feature for sentence identifier and gene name. In MyAnnotator.xml file, these types are described.

```
<typeDescription>
  <name>MyGene</name>
  <description>gene</description>
  <supertypeName>uima.tcas.Annotation</supertypeName>
</typeDescription>
<typeDescription>
  <name>MyID</name>
  <description>id</description>
  <supertypeName>uima.tcas.Annotation</supertypeName>
 </typeDescription>
```

## 2. algorithmic aspects

### 2.1 NLP techniques – Stanford NLP

In this program, the very simple named entity recognizer based on each token's part-of-speech (e.g., noun, verb, adjective, adverb, etc.) is used. At first, uses a tokenizer to tokenize the sentence into token which is labeled a part-of-speech tag.

At each sentence, it starts with the sentence-identifier, and it labeled as a noun. Therefore, first noun at each sentence, it added especially in MyID defined in typesystem.
After that, the consecutive nouns will be grouped into "noun phrases", which will be added to input stream of the other annotator.

The second annotator uses external biological database, as mentioned in 2.6.

### 2.2 external biological database used - LingPipe

The "noun phrase", the output of first the named entity recognizer, added to the input parameter of the other annotator. It uses biomedical named entity recognizer, LingPipe.

At first, it loads a biomedical named entity recognizer as an instance of the Chunker interface and then the entity which recognized as a gene name and its start/end position will be added the list of chunks.

The only detected as a gene name will be added to MyGene defined in typesystem.