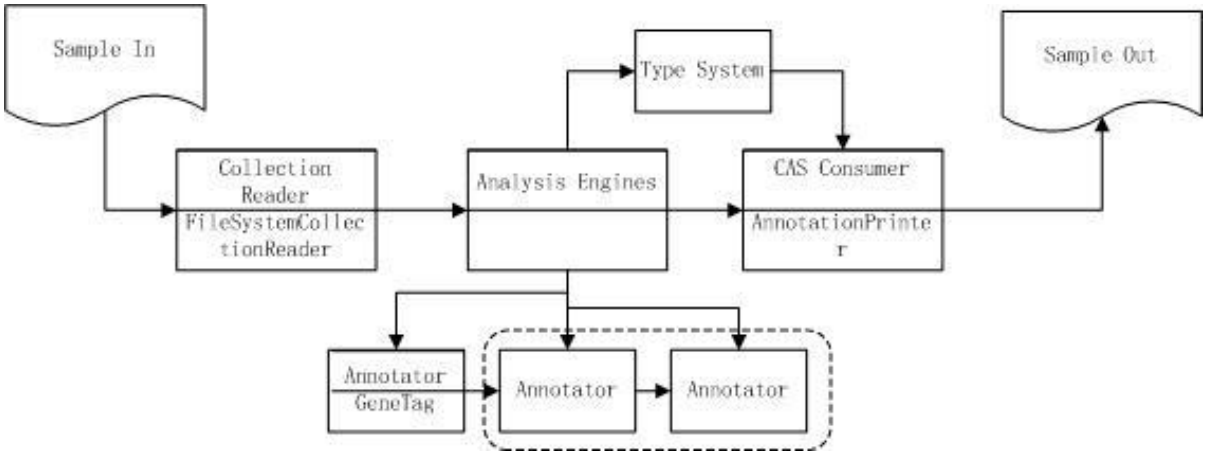
Andrew ID: Xiaoboh
Name: Xiaobo Hu

UIMA Report

--------Named Entity Recognizer

SYstem structure:



My UIMA system consists of three main parts: Collection Reader, Analysis Engines, CAS Consumer.

Collection Reader prepares an individual data item for analysis and loads it into the CAS. After the data loads into CAS, following part can analyze and process the data. The input file is /scr/main/resources/data/hw1.in

Analysis Engines is the core part of the system. In this part, the system implements the name recognition. With some algorithms, this part can catch the need words from data, and then give the result to CAS Consumer. In my system, there is only one annotator---GeneTag. Actually, this part can consist of several annotators.

CAS Consumer is responsible for output. The output file is the final work, in /scr/main/resources/outputfile/hw1-xiaoboh

Collection Reader:

My Collection Reader name is *FileSystemCollectionReader*, which come from the model in example. In this part, the route of the input file should be changed. The original route is /scr/main/resources/data. Add a new parameter to change the finding files way, point to file hw1.in instead of folder.

Text → aJCas

Type system:

I have one Type System in my project, named *TypeSystemDescriptor*. This Type system collects parameter from last Annotator, transit to next Annotator. Because I only have one Annotator, so the output of Type System transfers to CAS consumer.

There are 4 parameters in *TypeSystemDescriptor.*

- SentenceID //The ID of each sentence
- GeneName //gene name
- Begin //The position of first letter in GeneName
- End //The position of last letter in GeneName

Analysis Engines:

In my Analysis Engines, there is one Annotator GeneTag. The core function in *GeneTag* is *nBestChunks(char[],int,int,int)*, which come from Lingpipe.

"The *nBestChunks(char[],int,int,int)* method is implemented by walking over the n-best analyses generated by *nBest(char[],int,int,int)* with a maximum n-best for full analyses set to the value of *numChunkingsRescored()*, which may be changed using *setNumChunkingsRescored(int)*. For each analysis, the chunks are pulled out and their weight is incremented by the n-best analysis weight."(Reference from *http://alias-i.com/lingpipe/docs/api/com/aliasi/chunk/RescoringChunker.html*) The format of this function is nBestChunks(char[] cs, int start, int end, int maxNBest).

In my code, firstly, spilt input text into each sentence by *split()*. Then using *toCharArray* gets input char from each sentences. char[] cs is the parameter in nBestCHunks.

Main Function in GeneTag:

- Spilt(string) //spilt sentence
  Using this function to spilt the text into each sentences.

- Substring(int, int) //getting sentence ID, and spare text.
  Position from 0-14, is the sentence ID.
  The gene name come from last part of the sentences, (15 - end).

- nBestChunks // tagging gene name
  Returns the n-best chunks for the specified character slice up to the specified maximum number of chunks

Parameters *start*, *end*, *x*, *y*, are used to calculate position of gene name.

Parameters *ID*, *name* are used to record sentence and gene name.

CAS consumer:

The CAS consumer in my project is *AnnotationPrinter*, whose model comes from example. Some parameters and functions of *AnnotationPrinter* extend from *CasConsumer_ImplBase*.

Input parameters derive from Type system.

- SentenceID
- GeneName
- Begin
- End

Change the *outputfile* value to src/main/resources/outputfile/hw1-xiaoboh as final address.

Collection processing engine

Using CPE configuration builds CPE descriptors. Choosing *FileSystemCollectionReader* as Collection Reader, choosing *GeneTag* as Analyze Engines choosing *AnnotationPrinter* as CAS consumer, then save it to folder.

Accuracy

Using a text contrast program tests the accurate of the output.

Precision: 0.881345
Recall: 0.610113
F-measure: 0.721066

The accuracy is 72.1%.