Yan He (yanhe)
October 8, 2014

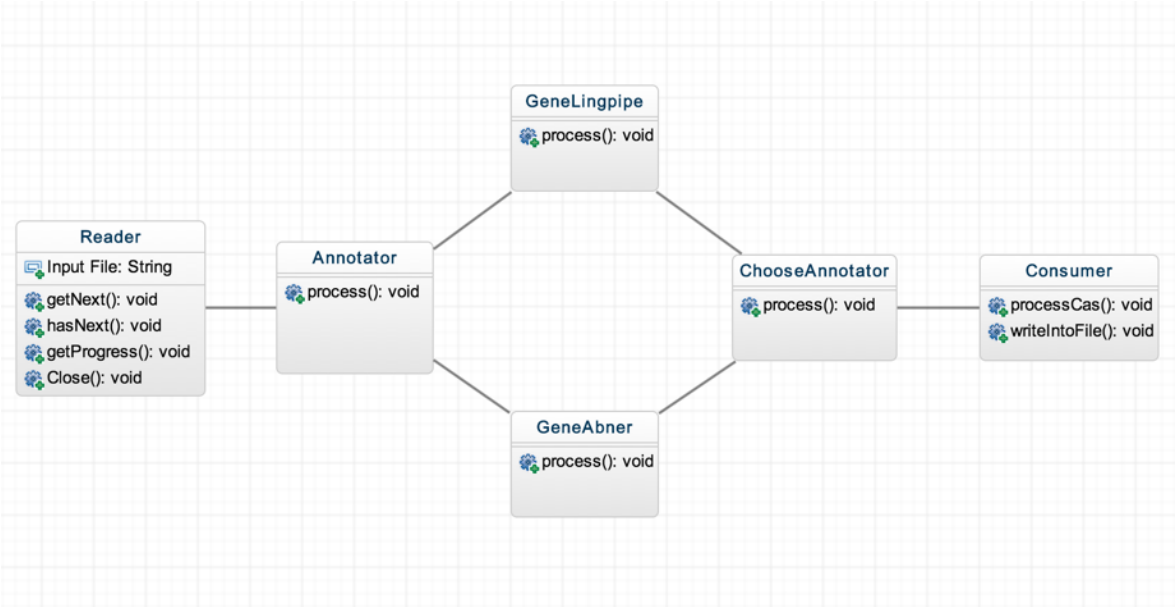
Software Methods Homework 2 Report

1. System Architecture

For this homework 2 assignment, I have implemented the logical architecture for the sample information processing task like last time. The difference is that in this assignment I made use of two different NER(Named Entity Recognizer) and chose the right gene tag from the output of the process. In details, I use both LingPipe and Abner in separate annotators and cooperate them together to recognize the gene name and extract them from the input file.

2. Data flow

The basic data flow of this system is showed as below:



3. Collection Reader (Reader)

The function of the reader is to read the input file line by line with the getNext() function, then it put each line of the file into JCas and provide it to Analysis Engine for the following process.

4. Analysis Engine

The main function of Analysis Engine is to process the content from the input file and use several annotators to get the gene related words and provide them to CAS Consumer for writing them into the output file.

4.1 Type System (deiis_types)

For this homework we use the DEIIS type system which is different from the last homework. It extends the Annotator with begin and end, and add several more parameters. The first one is ID, which stands for the ID number of each sentence. The second one is Content, which stands for the content of each sentence.

4.2 Sentence Annotator (Annotator)

The main function of Annotator is to split the content of input file into separate sentences and split the ID and Content of each sentence. These information will be stored in aCas, provide to the following annotators and participate in the following process.

4.3 LingPipe Annotator (GeneLingpipe)

The function of GeneLingpipe is to process the information inside the aCas and use the LingPipe gene name recognizer to extract the target gene name from the sentence. On the other hand, the function will calculate the confidence value of each gene name and get rid of the gene name with low confidence.

4.4 Abner Annotator (GeneAbner)

The function of GeneAbner is also to process the information inside the aCas and use the Abner gene name recognizer to find the target gene name from the sentence.

4.5 Merge Result (ChooseAnnotator)

The function of ChooseAnnotator is to merge the result of GeneLingpipe and GeneAbner. For those gene name with high confidence from GeneLingpipe, it will be selected with no doubt. And for those gene name with low confidence, it will be checked with the result of GeneAbner. If it also exits in the result of GeneAbner, it will also be selected.

5. CAS Consumer (Consumer)

Consumer build a new file and use the function of BufferedWriter to save the gene name and ID that have been selected by ChooseAnnotator.

6. External Resources

LingPipe: http://alias-i.com/lingpipe/

Abner: http://pages.cs.wisc.edu/~bsettles/abner/