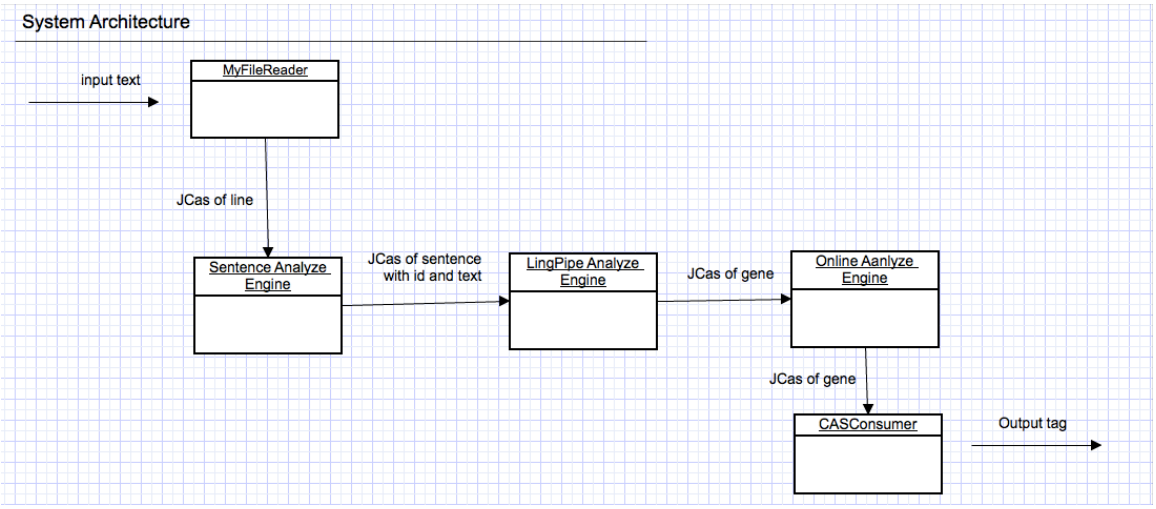
Overview

In the task, I used name entity-recognition model from LingPipe to find the gene tag with confidence, then I search the tag with low confidence in web http://bergmanlab.smith.man.ac.uk:8081 to decide whether the it is a gene tag.

Architecture:



MyFileReader: reads the input text, get sentences form text, put sentences in CAS, one sentence one CAS

Sentence Analyze Engine: for every CAS, get the sentence in it, parse the sentence into sentence and text

LingPipe Analyze Engine: for every CAS, get the sentence text in it, analyze the text using entity recognition in LingPipe, generates the gene tag with confidence

Online Analyze Engine: for every CAS, get the gene tag in it, if the confidence value is between 0.5-0.6, search it online, if found, change the confidence value to 1.0

CASConsumer: for every CAS, get the gene tag in it, if the confidence value is high that 0.6, output in the file

External resource

LingPipe:

Like many of the other modules in LingPipe, named entity recognition involves the supervised training of a statistical model or more direct methods like dictionary matching or regular expression matching. All these methods are designed to work together smoothly by using the same class for annotating the text com.aliasi.chunk.

Website: http://bergmanlab.smith.man.ac.uk:8081

We can sent gene tag to the web by HTML get, if the web can find the gene tag, it will return some information; if not found, the web will not return any information

Evaluation:

I evaluated the output by precision and recall:

Configuration: output gene tags with confidence value bigger that 0.6, search gene tags with confidence value between 0.5 ~ 0.6

Precision: 0.801476

Recall: 0.814069