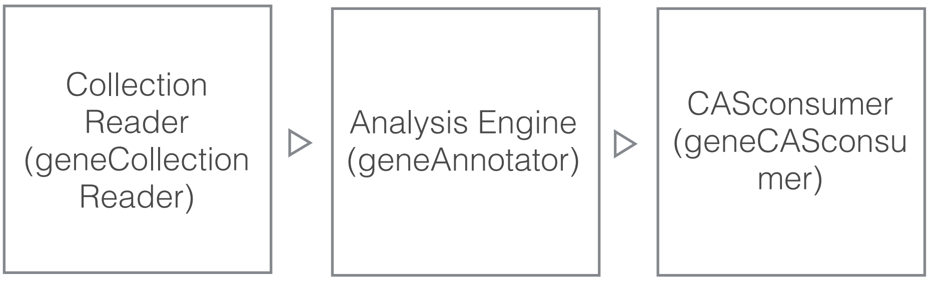
Report for 11791 HW1

Name: Chao-Hung Chen

Andrew ID: chaohunc

System Overview


Collection Reader (geneCollection Reader)
Analysis Engine (geneAnnotator)
CASconsumer (geneCASconsu mer)

Type System

Type or Feature	SuperType	
geneObj (Type)	uima.tcas.Annotation	
ID (Feature)	uima.cas.String	used to record ID
geneName (Feature)	uima.cas.String	recognized gene name from NER
posStart (Feature)	uima.cas.Integer	record the start position of gene in the sentence
posEnd (Feature)	uima.cas.Integer	record the end position of gene in the sentences
confidence (Feature)	uima.cas.Double	record the confidence scores of gene from NER

Although the Annotation object already native variables to record the begin positions and start positions, I think it’s more appropriate to store both global positions (the document) and local positions (the sentence). Therefore, I set posStart and posEnd as features to record positions of gene name in the sentences.

Collection Reader (geneCollectionReader)

As the collection reader, geneCollectionReader first need to read the file. Since the input file in this homework contains sentences, which contain one to several gene in the sentence, I then decided to read the file line-by-line and then combined those lines to put them as the input of geneAnnotator.

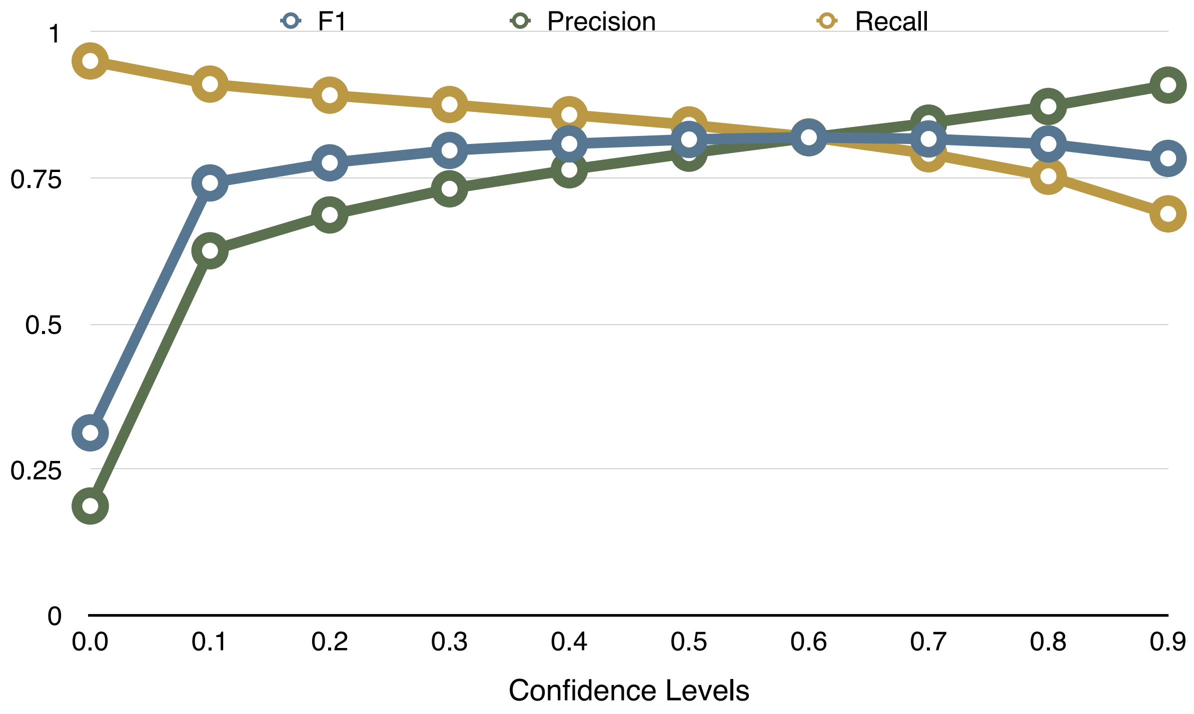
Analysis Engine (geneAnnotator)

geneAnnotators played the important role as the analysis engine in the system. At first, I tried to use the Stanford NLP tagging tools as core of analysis engine. However, it ran very slow. Therefore, I tried to use the other model, provided by lingpipe, as core of analysis engine. It used HMM (Hidden Markov Model) to train a model of name entity recognition. For each sentence in the CAS, we put it in the model and check if there exists any possible gene. Based on the model, we could get the possible gene and its confidence of score. Then, we put the all of possible genes into the CAS.

CASconsumer (geneCASconsumer)

geneCASconsumer was used to process CAS file and output the result of CAS. The output file contains several sentences. For each sentence, it outputs the sentence ID, positions of genes in the sentence and gene name. I then extracted the potential gene from CAS and put them in the output sentences.

Experiment of parameters for confidence levels



Confidence Levels	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
F1	0.3128	0.7419	0.7764	0.7977	0.8090	0.8165	**0.8203**	0.8174	0.8088	0.7840
Precision	0.1872	0.6254	0.6870	0.7318	0.7645	0.7930	0.8195	0.8446	0.8731	**0.9098**
Recall	**0.9513**	0.9116	0.8925	0.8765	0.8590	0.8415	0.8210	0.7919	0.7533	0.6888

The experiment result shows that when we set the parameter at lower confidence level, we will get high recall and low precision, and vice versa. Therefore, we try to use F1 measurement to pick our parameter, which is 0.6.