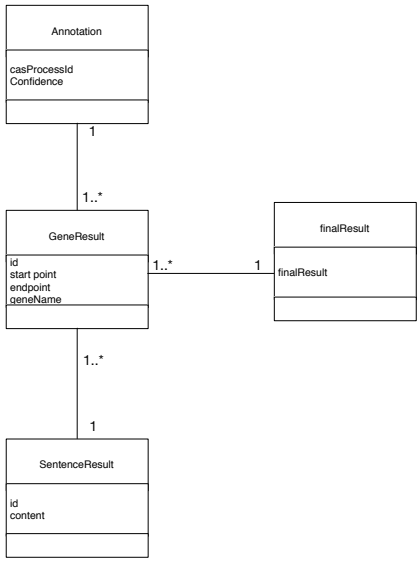
Andrew id: qiqis
October 6, 2014

Type System Design: (Inheritance)

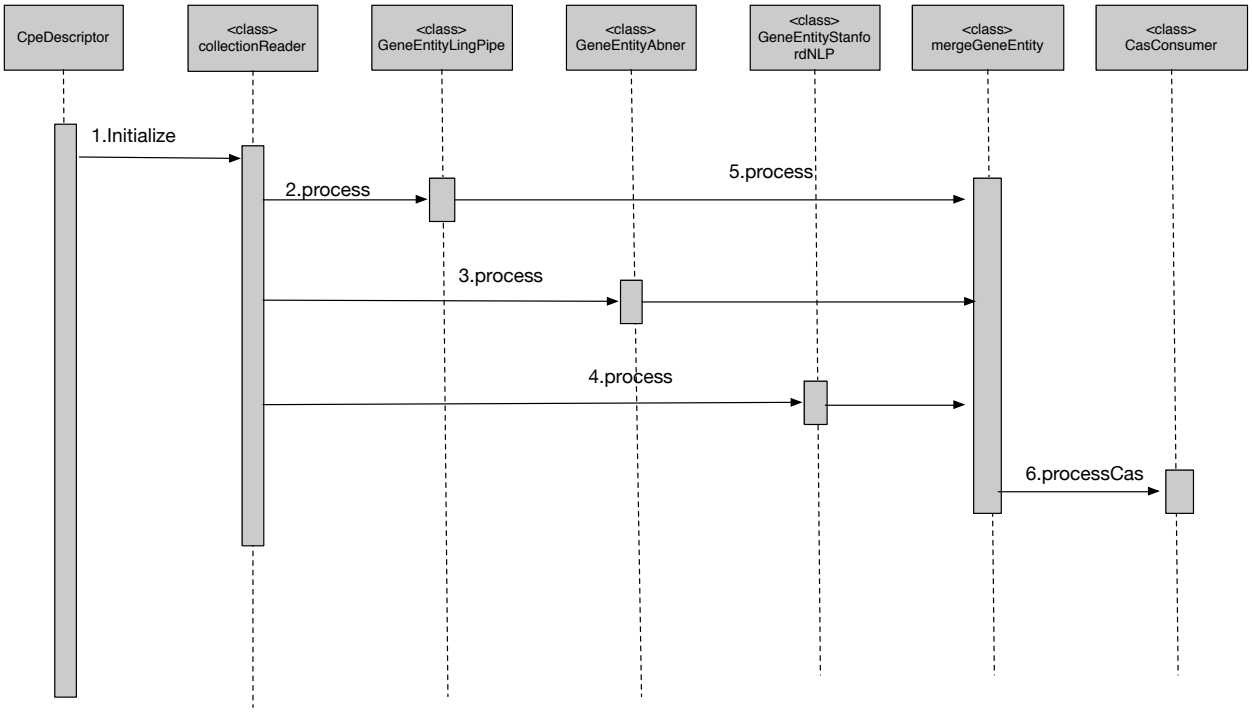


- Annotation: This is the superclass of GeneResult. Attribute casProcessId is used to mark different annotation: LingPipe, Standford NLP and Abner.
- SentenceResult: Get sentence info from input raw data.
- GeneResult:This is the GeneResult from each annotation:LingPipe, Standford NLP and Abner.
- finalResult: Finally, we merge the GeneResult from each annotation and get the final result.

Sequence Diagram:

This is the sequence diagram of each component:

- CollectionReader: read data from hw2.in file.
- GeneEntityLingPipe: use LingPipe to cope data to get the result.
- GeneEntityAbner: use Abner to cope with the raw data.
- GeneEntityStanfordNLP:use StanfordNLP to cope with the raw data
- MergeGeneEntity： use merge algorithm to merge all three results and get the final ones.



Merge Algorithm:

Bootstrap aggregating, often abbreviated as bagging, involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set. As an example, the random forest algorithm combines random decision trees with bagging to achieve very high classification accuracy.

In this algorithm, we assign three GeneEntity Recognizer: LingPipe, Abner and StanfordNLP the same weight and vote according to the separate results. If the weight of the final result is bigger than 2, we assume that the Gene entity is true.

Regarding the LingPipe Gene Entity recognizer, we assume that if the confidence is bigger than 0.6, the result is the gene entity recognized by LingPipe. And 0.6 is calculated by training the sample.out to get the maximum F score.

Performance:

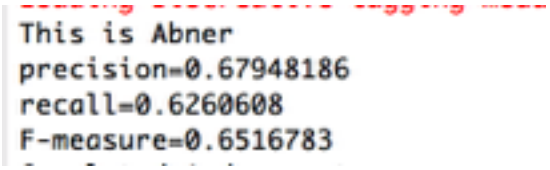
performance of LingPipe:

The following figure shows that the F score of LingPipe is 81% which means that this is a good performance of LingPipe. And the precision and recall of LingPipe are kind of good. However, there are some drawbacks that there is possibility that this data is fit for Lingpipe but another data may not fit for this. Therefore, we need other ways to accomplish Gene Entity recognize.

```
This is LingPipe
precision=0.81440383
recall=0.81970984
F-measure=0.8170482
```

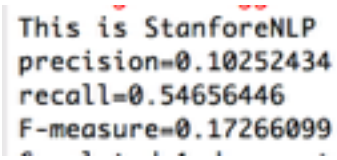
performance of Abner:

The following figure shows that the F score of Abner is 65% which means that this is a good performance of LingPipe. And the precision and recall of Abner are kind of good. Besides, abler can distinguish protein, DNA etc. However, there are some drawbacks that Abner can not recognize all the right answer. Therefore, we need other ways to accomplish Gene Entity recognize and combine them together.



performance of StanfordNLP:

The following figure shows that the F score of StanfordNLP is 17% which means that this is not a good performance of StanfordNLP. And the precision and recall of StanfordNLP are kind of bad. Besides, StanfordNLP can only distinguish noun. However, StanfordNLP can be used as a way to detect noun as the basic recognize. Therefore, we combine all of them together.



Total performance

After each GeneEntity recognized the Gene Name in the sentence, the F-measure of each Annotator is as follows. And we calculate the final Merge score of three annotation and get the score of 76%. This merge algorithm allows to sum three different algorithm up and avoid drawbacks of each algorithm.

	LingPipe	Abner	Standford NLP	Merge
F-measure	81%	65%	17%	76%

Problem and Solution

- **Merge algorithm:** At first, I came up a more complex Bagging algorithm to solve this problem. First, I assign the same weight to three GeneEntity Recognizer and we

change the weight according the f score of each recognizer. And we get a new sample output according the new weight. And we calculate the new f score. Then we do the same thing to change the weight according the new f score. We change specific times and the weight of each one will convergent to specific number.

- **Maven Release:** I have encountered the javadoc problem when releasing. And I solved the problem according the solution provided by Fei. And by update the version of javadoc
- **Grade. sh:** I have encountered path problems, text not found problem and all of these problems have been solved according to the solutions provided by TA and partners.