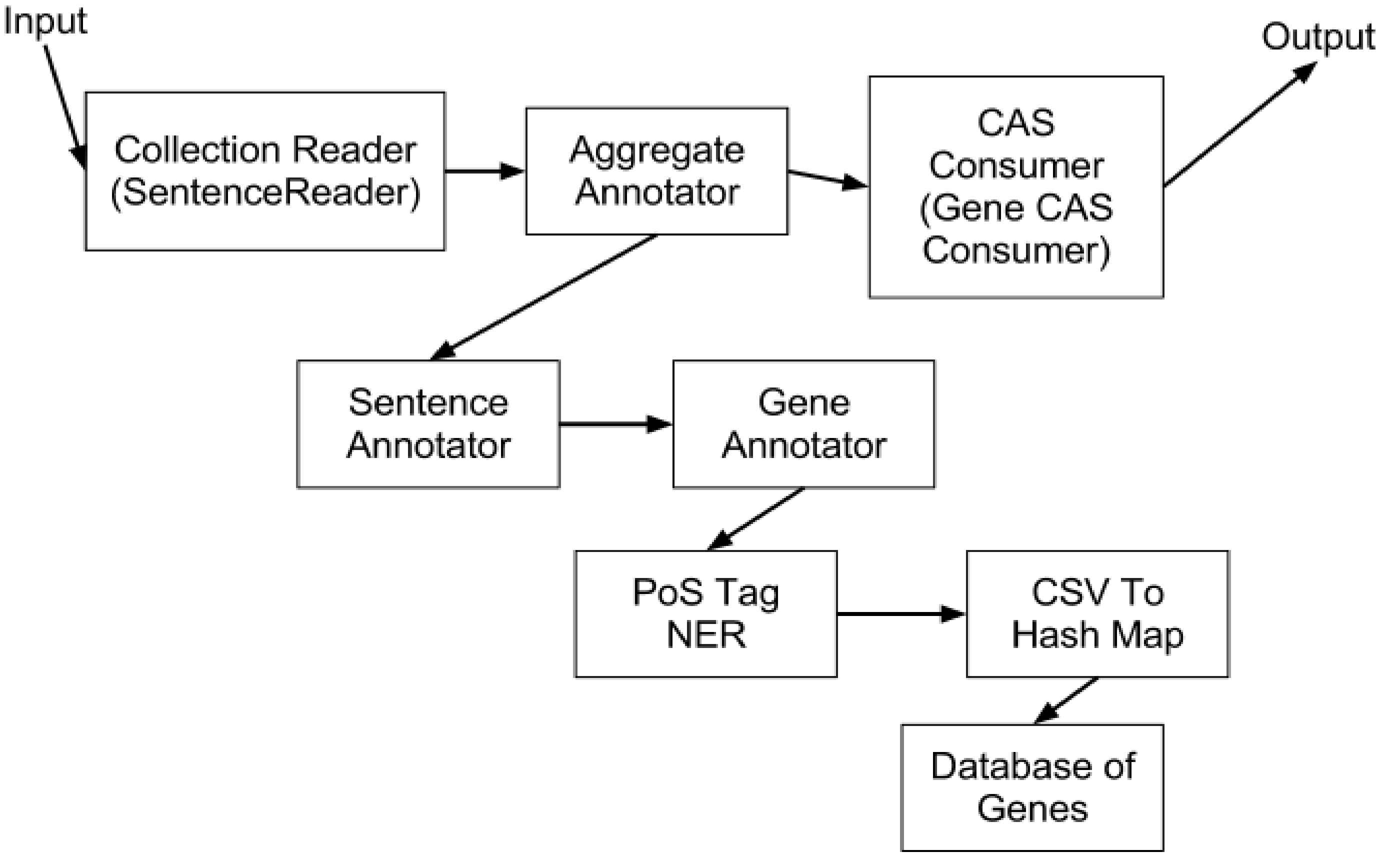
Kartik Mandaville
kmandavi

Homework 1 - Task 3 - Report

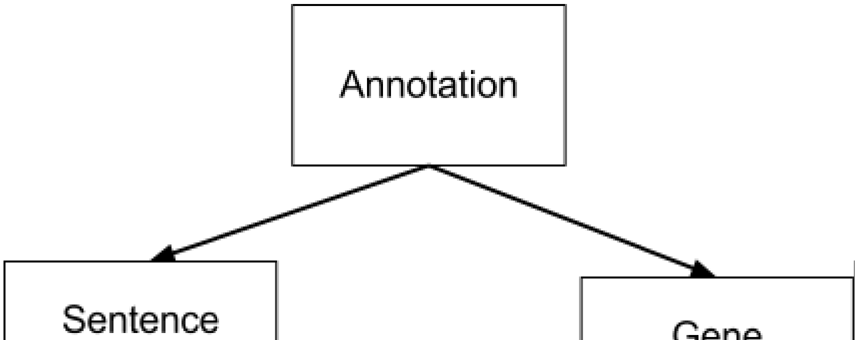
Pipeline:
Overall Architecture



Type System:

Data types are derived from the parent Annotation type which has begin, end etc.

- Sentence
  - Identifier
  - Text
- Gene
  - Identifier
  - Name



Flow:

- The input file is read by the Collection Reader(SentenceReader) and its inserted into the CAS
- The Aggregate annotator calls the Sentence Annotator which takes the CAS and add indexes for each sentence with the identifier and text
- The next annotator, GeneAnnotator iterates over the CAS on the sentence type and gets the text. Each sentence is passed to the PoS NER by Stanford and the entities returned are looked up in the Gene database through the CSVToHashMap class. If present, it is then added to the CAS with the identifier and gene name
- The CAS Consumer iterates over the CAS on the gene type and prints to the file
- The gene database is converted to a Hash Set for convenient/faster lookup by the CSVToHashMap java class. The input is in CSV format.

Techniques:

- NLP techniques used: Stanford's PoS tag Named Entity Recognition
- External data: Database of 250,000 genes from various sources: NCBI etc. (Database taken from Dr. Guy Zinman, Lane Center for Computational Biology)

Limitations

- The accuracy is very less.
- Limited by database of the genes
- Direct lookup is not accurate as a gene may have multiple names depending on the naming scheme
- Even common English words like "lack" are a gene. Hence, contextual analysis could make the accuracy much better

Future Improvements:

- Looking up the genes found in a single sentence in the Gene Ontology(GO) database and comparing the "term accession" terms returned. Ideally, there should be common terms between the different genes found in the single sentence.
- Using rule sets
- Intelligently identifying "5-nucleotidase" as "5'-nucleotidase"
- Using gene API's: ensemble (http://www.programmableweb.com/api/duplicated-genes-database)
- Confidence measure in the CAS consumer to select the one with the highest confidence