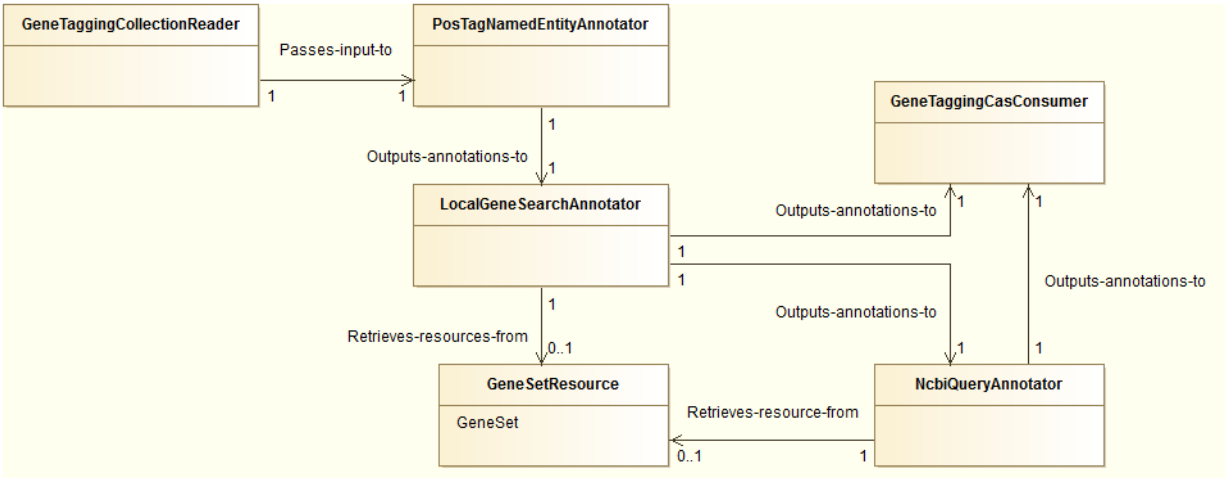
Hank Hwang
Andrew ID: hankh

Homework 1 Report

Architecture



Collection Reader

The GeneTaggingCollectionReader is responsible for reading the input file and creating a CAS with annotations defined by the type system InputTypeSystem. The Input type has two features of the sentence id and text (the actual sentence). The CAS is then passed to the PosTagNamedEntityAnnotator.

Analysis Engine

The analysis engine is composed of three annotators, PosTagNamedEntityAnnotator, LocalGeneSearchAnnotator, and NcbiQueryAnnotator.

The PosTagNamedEntityAnnotator has an input defined by the type system, InputTypeSystem, and output defined by the type system PosTagNamedEntityTypeSystem. The PosTagNamedEntity type has two features of sentence id and named entity. The annotator takes the sentence and using Stanford's coreNLP, identifies all noun phrases. The noun phrases are then stored into their own annotations in the feature namedEntity. The PosTagNamedEntityAnnotator will then pass the CAS and new annotations to the LocalGeneSearchAnnotator.

The LocalGeneSearchAnnotator has an input of PosTagNamedEntity and output type defined by the type system LocalGeneSearchTypeSystem. The output type system has two types associated with it, FoundGene and UnfoundNamedEntity. The FoundGene has two features of id and gene, and the UnfoundNamedEntity has two features of id and namedEntity. The LocalGeneSearchAnnotator retrieves a resource file from GeneSetResource that has a GeneSet, containing a local set of genes. The annotator takes the set and searches to see if any of the genes are a substring of the input namedEntity. If a namedEntity is found to have a gene, the gene is saved in a FoundGene. The FoundGene annotations are then passed to the GeneTaggingCasConsumer. If an input namedEntity never finds a gene, the

namedEntity is saved in an UnfoundNamedEntity. The UnfoundNamedEntity annotations are then passed to the NcbiQueryAnnotator.

The NcbiQueryAnnotator has an input of UnfoundNamedEntity, and output NcbiResults defined by the NcbiResultsTypeSystem. The NcibResults has two features of id and result. The NcbiQueryAnnotator also retrieves the same resource file in LocalGeneSearchAnnotator from GeneSetResource. The NcbiQueryAnnotator takes the namedEntity from UnfoundNamedEntity and uses it as a query term to the NCBI Gene database. The returned results are then parsed to produce a list of gene names. The annotator takes the gene list and searches to see if any of the genes are a substring of the input namedEntity. If a namedEntity is found to have a gene, the gene is saved to NcbiResults. The NcbiResults annotations are passed to the GeneTaggingCasConsumer. The list of gene names are saved to the GeneSet in GeneSetResource and then the resource file is saved to update the list.

CAS Consumer

The GeneTaggingCasConsumers takes the annotations produced by the LocalGeneSearchAnnotator and NcbiQueryAnnotator and produces a file of the annotations with the output of:

Sentence ID|start-offset end-offset|text

Techniques

The techniques used are Stanford's CoreNLP and NCBI Query. The StanfordCoreNLP is used to tokenize a sentence and return the noun phrases within the sentence. Each noun phrase is considered a named entity and then used for the following technique.

The NCBI query composes of two parts; the first part is using a local list of genes to search through the named entities, and the second part is taking the named entities that are not found locally and querying the NCBI Gene database. Each gene search produces a list of genes, and each gene is checked to see if they are contained in that named entitiy. After an online query is completed, the gene list found from that query is added to the local list of genes.

The local list of genes is used to improve the speed of the search. Only using an online query produces a very slow search because of the dependence on internet queries. The local list stores the previous results and should improve the speed of searching for genes by decreasing the size of inputs for the NCBI queries.

Evaluation

Using the first 1000 sentences from the provided sample.in file, the collection processing engine produced 4389 results, while the sample.out file had 608 results. After evaluating the results from the CPE, the 608 results from sample.out were also found to be in CPE's output. So the main problem with the CPE results is the 3781 false positives.

The other evaluation point was comparing the elapsed time between running the pipeline with an empty gene file and running it again after the gene file is filled from previous searches. With an empty gene file, the elapse time was about 28 minutes, and running it a second time produced an elapsed time of 20 minutes. The results from both runs can be seen in appendix A and B.

Appendix A: 1000 samples, Empty Gene.txt – Performance Report

```
Parsing CPE Descriptor
Instantiating CPE
Running CPE
To abort processing, type "abort" and press enter.
CPM Initialization Complete
Adding annotator tokenize
Adding annotator ssplit
Adding annotator pos
Loading default properties from tagger edu/stanford/nlp/models/pos-tagger/english-
left3words/english-left3words-distsim.tagger
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-
left3words/english-left3words-distsim.tagger ... done [3.2 sec].
Completed 1 documents; 153305 characters
Total Time Elapsed: 1709548 ms
Initialization Time: 1800 ms
Processing Time: 1707748 ms


 ------------------ PERFORMANCE REPORT ------------------

Component Name: GeneTaggingCollectionReader
Event Type: Process
Duration: 345ms (0.02%)
Result: success
Component Name: GeneTaggingTAE
Event Type: Analysis
Duration: 1707358ms (99.98%)
Sub-events:
       Component Name: LocalGeneSearchAnnotator
       Event Type: Analysis
       Duration: 141ms (0.01%)

       Component Name: PosTagNamedEntityAnnotator
       Event Type: Analysis
       Duration: 9429ms (0.55%)

       Component Name: NcbiQueryAnnotator
       Event Type: Analysis
       Duration: 1697781ms (99.42%)

       Component Name: Fixed Flow Controller
       Event Type: Analysis
       Duration: 5ms (0%)

Component Name: GeneTaggingTAE
Event Type: End of Batch
Duration: 1ms (0%)
Component Name: GeneTaggingCasConsumer
Event Type: Analysis
Duration: 7ms (0%)
Component Name: GeneTaggingCasConsumer
Event Type: End of Batch
Duration: 0ms (0%)
```

Appendix B: 1000 samples, second run – Performance Report

```
Parsing CPE Descriptor
Instantiating CPE
Running CPE
To abort processing, type "abort" and press enter.
CPM Initialization Complete
Adding annotator tokenize
Adding annotator ssplit
Adding annotator pos
Loading default properties from tagger edu/stanford/nlp/models/pos-tagger/english-
left3words/english-left3words-distsim.tagger
Reading POS tagger model from edu/stanford/nlp/models/pos-tagger/english-
left3words/english-left3words-distsim.tagger ... done [3.6 sec].
Completed 1 documents; 153305 characters
Total Time Elapsed: 1222450 ms
Initialization Time: 2019 ms
Processing Time: 1220431 ms


 ------------------ PERFORMANCE REPORT ------------------

Component Name: GeneTaggingCollectionReader
Event Type: Process
Duration: 203ms (0.02%)
Result: success
Component Name: GeneTaggingTAE
Event Type: Analysis
Duration: 1219800ms (99.95%)
Sub-events:
       Component Name: LocalGeneSearchAnnotator
       Event Type: Analysis
       Duration: 96323ms (7.89%)

       Component Name: PosTagNamedEntityAnnotator
       Event Type: Analysis
       Duration: 8900ms (0.73%)

       Component Name: NcbiQueryAnnotator
       Event Type: Analysis
       Duration: 1114570ms (91.33%)

       Component Name: Fixed Flow Controller
       Event Type: Analysis
       Duration: 5ms (0%)

Component Name: GeneTaggingTAE
Event Type: End of Batch
Duration: 1ms (0%)
Component Name: GeneTaggingCasConsumer
Event Type: Analysis
Duration: 392ms (0.03%)
Component Name: GeneTaggingCasConsumer
Event Type: End of Batch
Duration: 0ms (0%)
```