Name: Yan Zhao

Andrew ID: yanzhao2

11791 HW2 Report

Logical Architecture and UIMA Analysis Engines Design & Implementation

Running Result

Precision: 0.8609822733337569

Recall: 0.7419107582808651

F1_Score: 0.7970238795435831

Completed 1 documents; 2515605 characters

Initialization Time: 2007 ms

Processing Time: 130322 ms

NER Design and Architecture

The diagram below displays the whole structure of this Collection Processing Engine, This system has three main components, which are CollectionReader, Analysis Engine, and CasConsumer. The Analysis Engine is a aggregate annotator, which consists of one SentenceAnnotator, two GeneAnnotators using lingpipe and abner, and one MergeAnnotator.

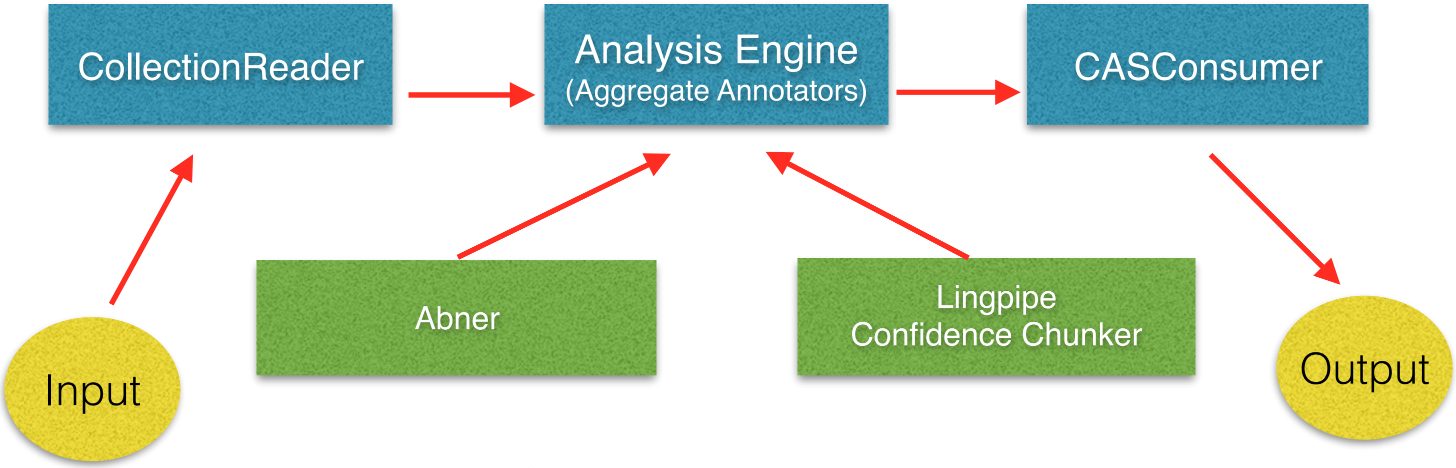

figure 1 NER(CPE) Components

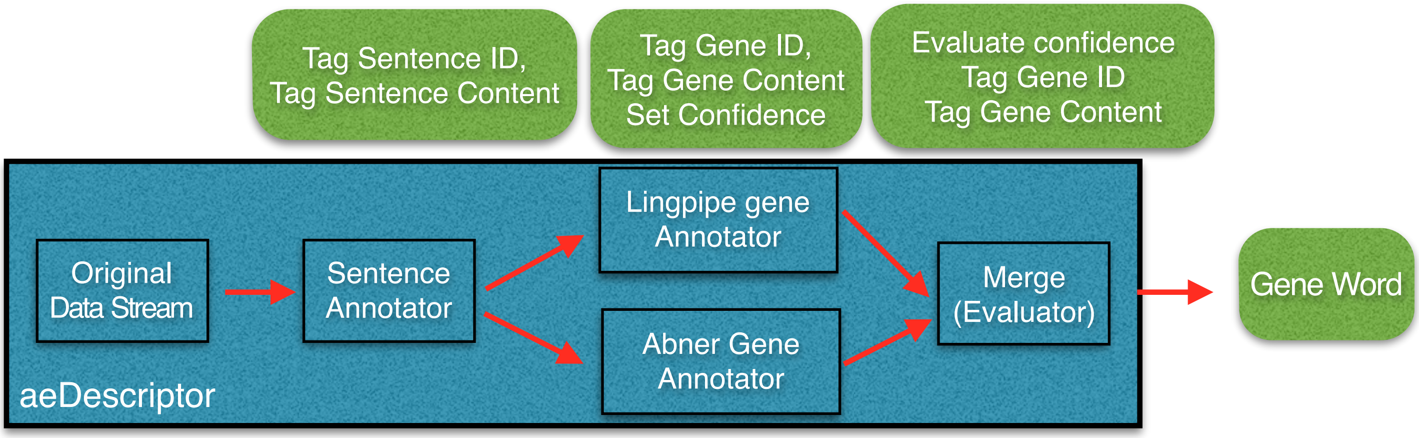

figure 2 Aggregate Analysis Engine

Type System

The pipeline has a pipeline of three phases.

The first one extracts a line/sentence feature from the paragraph and add to one type call SentenceType. This SentenceType also stores the content of specific line as a string in the content feature.

The second phase is to recognize gene words. This phase includes two ways using Lingpipe and Abner. The features of GeneLingpipeType is created based on the features of SentenceType, as is the same with features of GeneAbnerType. Therefore, the Spelling feature, which is same to the name of specific gene, is extracted from content feature. Since an evaluation of these two features will be added in the next phase, there is one more feature named Confidence added in GeneLingpipeType and GeneAbnerType, which reflects the probability of correctness. Besides, cause both the start-offset and end-offset don't include non-whitespace character, there are another two features added in GeneLingpipeType and GeneAbnerType, which are Begin and End corresponding to the two offsets.

The third phase extract features from GeneLingpipeType and GeneAbnerType, and add the final gene word with features of Id, Content, Begin and End into ResultType.

Collection reader

The reader will read all of the information into system at one time.

Aggregate Annotators

The SentenceAnnotator is to separate paragraph or input stream into lines. When a line break has been detected, a line/sentence content has been added SentenceType as an annotation.

The LingpipeAnnotator has integrated a confidence named entity chunking function from LingPipe. By using a GeneTag model provided by LingPipe, the Confidence Named Entity Chunking will return a set of results with their confidence feature. In this annotator, the candidates with confidence less than 0.3 will be dropped. In the end, the selected items will store in GeneLingpipeType for further extraction.

The AbnerAnnotator is quite like LingpipeAnnotator, it is another biological NER. In this annotator, confidence feature is added manually, by detecting the length of a content. In the end, the selected items will store in GeneAbnerType for further extraction.

The MergeAnnotator is to extract features from GeneLingpipeType and GeneAbnerType, using Content and Confidence from GeneAbnerType to build two hash maps, and compare the Content from GeneLingpipeType with these two hash maps. One item is stored in ResultType for three rules.

1) When the Confidence of Content from GeneLingpipeType is greater than 0.85

2) When the Confidence of Content from GeneLingpipeType is between 0.6 and 0.85, and GeneAbnerType has the same Content.

3) When the Confidence of Content from GeneLingpipeType is less than 0.6, but GeneAbnerType has the same Content with high Confidence.

CAS Consumer

CASConsumer will write all of the gene that the system analyzed into one single file according to the given format. In addition, the precision, recall and F-measure can also be calculated according to the given hw2.gold_stand file.

Future Improvements

1. The way I evaluate lingpipe and abner contents is simple, so can not avoid overfitting efficiently.
2. I could add some Machine Learning techniques myself, to fulfill my project.
3. I could use some external training data to help to improve my performance.