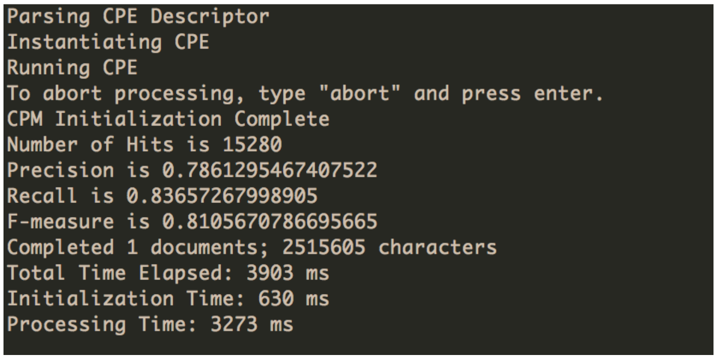
Xinyun (Victor) Zhao

HW1 Final Report

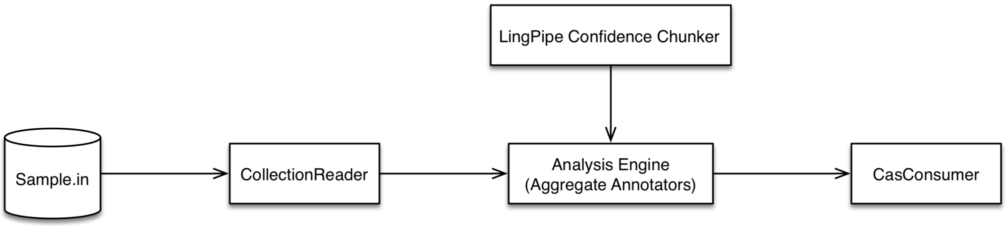
Running Result



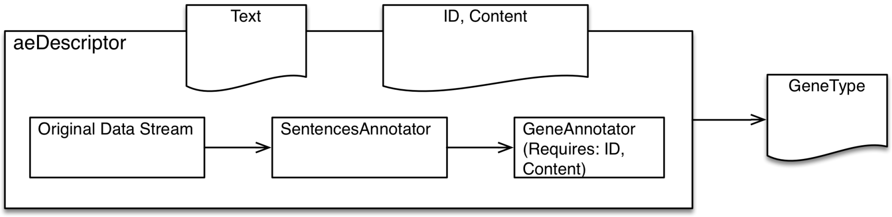
```
Parsing CPE Descriptor
Instantiating CPE
Running CPE
To abort processing, type "abort" and press enter.
CPM Initialization Complete
Number of Hits is 15280
Precision is 0.7861295467407522
Recall is 0.83657267998905
F-measure is 0.8105670786695665
Completed 1 documents; 2515605 characters
Total Time Elapsed: 3903 ms
Initialization Time: 630 ms
Processing Time: 3273 ms
```

The photo clip above shows the processing result of NER.

NER Design and Architecture



NER(CPE) Components



Aggregate Analysis Engine

The diagram above 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 SentencesAnnotator and GeneAnnotator.

Xinyun (Victor) Zhao

Type System

The pipeline has a pipeline of two 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. On the other hand, the features of GeneType could be created based on the features of SentenceType. Therefore, the Spelling feature, which is same to the name of specific gene, is extracted from content feature. Since both the start-offset and end-offset don't include non-whitespace character, there are two more features added in GeneType, which are BeginWithoutSpace and EndWithoutSpace corresponding to the two offsets.

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 GeneAnnotator 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 GeneAnnotator, the candidates with confidence greater than 0.5 will be selected as output. In the end, these selected items will have a GeneType as an annotation. Comparing to the implementation of PosTagNameEntityRecognizer, the outcome of F-measure is much higher.

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 sampleout file.

Miscellaneous

Package com.victorzhao.archive contains the PreGene Annotator, which makes use of PosTagNameEntityRecognizer, and PreGene type. In the previous version of this system, PreGene Annotator has been set between SentenceAnnotator and GeneAnnotator. Due to the worse performance of PosTagNameEntityRecognizer, the implementation related to PreGene have been discarded.

Reference:

Alias-i, L. (2008). 4.1. 0. *URL http://alias-i. com/lingpipe.*