Software Engineering for Information Systems

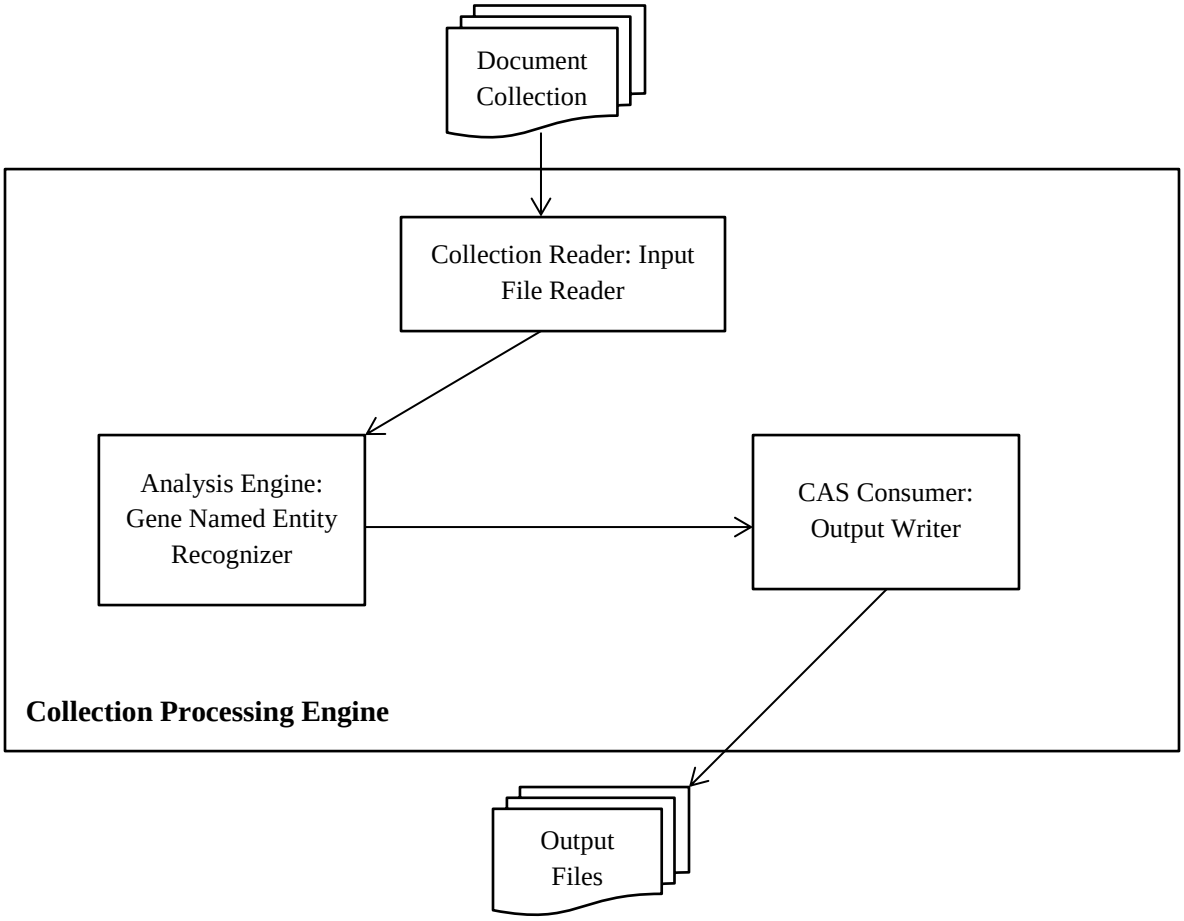
Homework 1 Report

Sushant Kumar
sushantk@andrew.cmu.edu

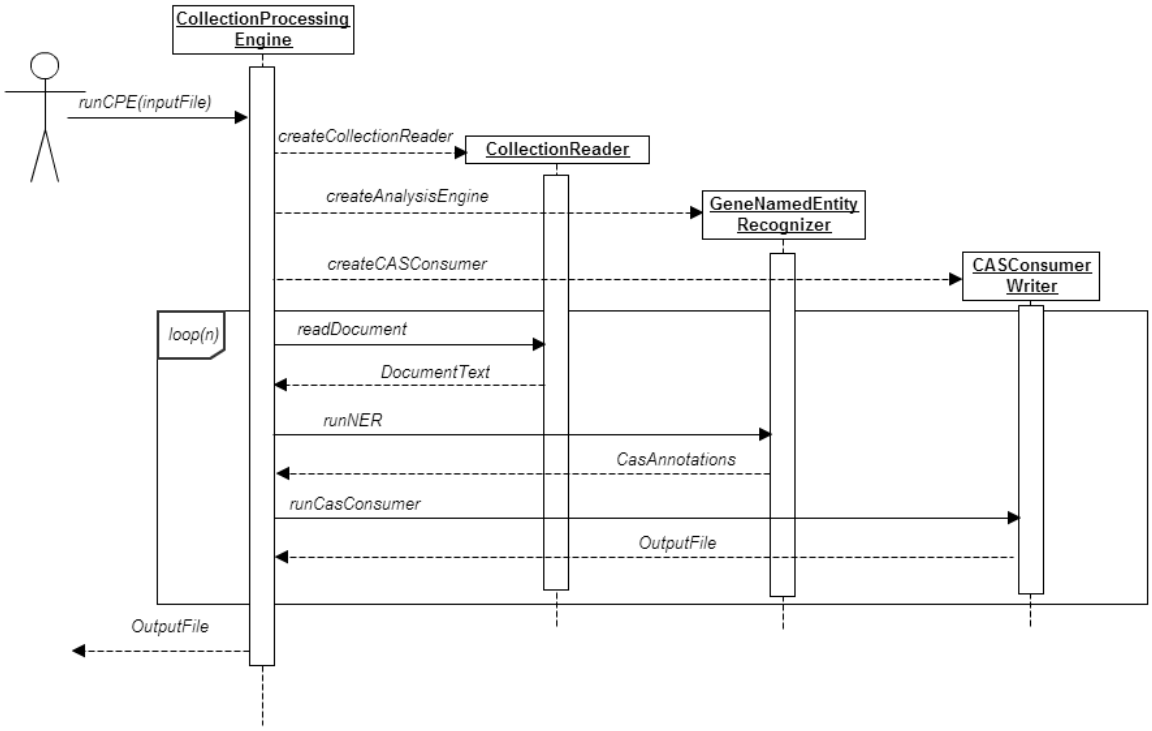
1 System Architecture

We present the architecture design of the system which shows how various components in the Collection Processing Engine interact. We show the pipeline design in Section 1.1 to show the UIMA workflow. The sequence diagram in Section 1.2 shows how the messages/data flow occur between the different classes.

1.1 Pipeline Design

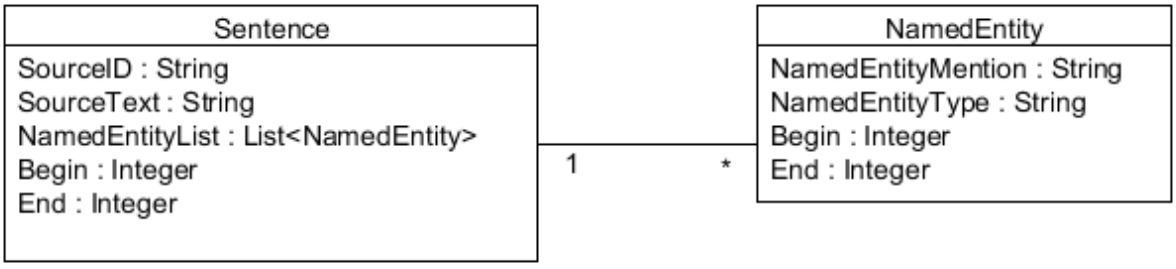


1.2 Sequence Diagram



2 Type System

The system defines two basic type system classes – Sentence and Named Entity. A sentence is basically the input sentence which consists of an ID and the text. A named entity consists of the information for an identified named entity and its type. Clearly, one sentence can have zero or more named entities.



3 Named Entity Recognizer

The system uses LingPipe[1] NLP toolkit to identify named entities in the given document. LingPipe supports a trained named entity recognizer for identifying biological gene entities. It is based on a supervised statistical approach involving a hidden markov model to identify the Gene mentions in the text.

Other than LingPipe, I also attempted to use Stanford's named entity recognizer to obtain Gene mentions, but due to difficulties in generating a trained model file for the system, I had to drop it midway.

4 Experiments and Analysis

I experimented with the output obtained from LingPipe and comparing it with the sample output file provided with the homework source release. Using the sample output, I obtained the Precision-Recall-FMeasure of the system output. These metrics help in comparing performance of different experiments with the system. LingPipe, for some odd reason, was not giving the confidence score of the named entity extracted which could have been used to prune some named entities for which the confidence is low, and help improve the precision of the system.

The results obtained from LingPipe NER, as compared with the sample output provided are
Precision = 0.77
Recall = 0.85
F-measure = 0.81

Looking at some of the Gene mentions which were missed by the system like 'E2 protein', 'ferric uptake regulation protein', 'POU-homeodomain protein', 'ICE inhibitor protein', 'BTB protein' etc. it seems that some simple pattern based rules such as "X Y Z protein" can be used to identify more Gene mentions. Similarly some other patterns can be learned from a training corpus and applied to improve the recall of the system. Also some of the mentions were missed due to incorrect span (begin-end) which can be handled by a better token combiner or else relaxing the named entity matching for such cases. We can also use a dictionary of biological entities to help identify more entities.

[1] http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html