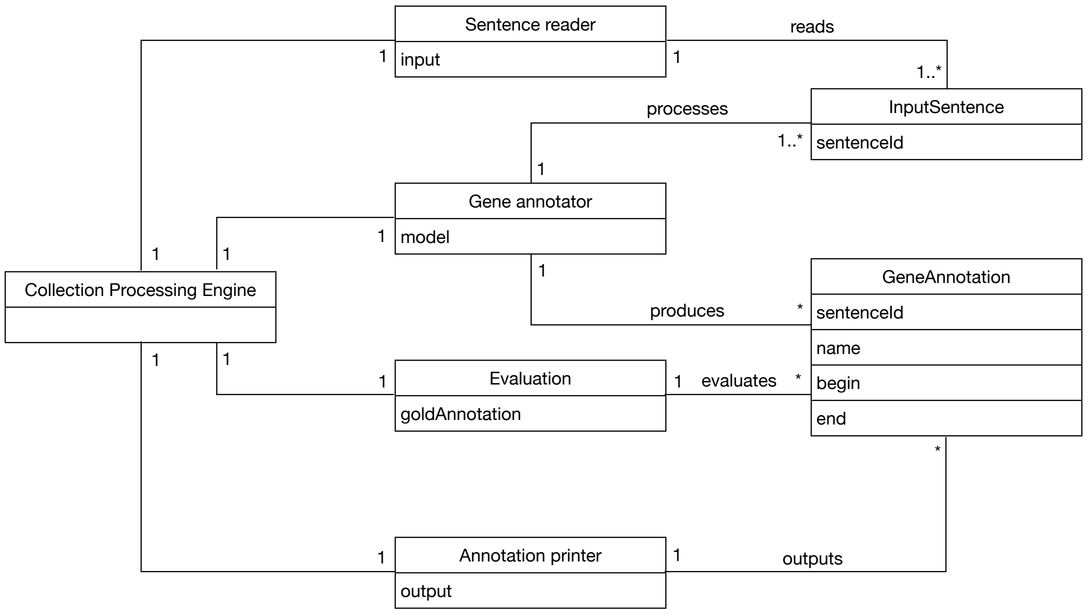
Named entity recognition with UIMA

Victor Chahuneau

System architecture

Global design



Collection Reader

A collection reader implemented as the `SentenceReader` class reads input sentences line by line from the file specified as the `input` parameter. It updates the `documentText` of the CAS and produces `InputSentence` instances containing a sentence ID.

Analysis Engine

Then, an analysis engine implemented as the `GeneAnnotator` class retrieves and processes each `InputSentence` instance in the CAS using the HMM chunker described in the following section to produce `GeneAnnotation` instances representing a segment recognized as a named entity. Each segment is identified by the `begin` and `end` attributes defined as required by the annotation task guidelines (number of non-whitespace characters). Annotations are stored in the CAS.

CAS Consumers

Finally, the `GeneAnnotation` instances can be processed by several endpoint CAS consumers:

- The `AnnotationPrinter` produces an output file with the path specified as the `output` parameter containing the required output format.
- The `AnnotationEvaluator` reads gold annotations from the file specified as the `goldAnnotation` parameter and evaluates the output of the analysis engine by aggregating the necessary sufficient statistics: number of true/false positive/negatives. When the entire CAS has been processed, it outputs the precision/recall/F1 results.
  Evaluation is optional: if the `goldAnnotation` parameter is not set, the `AnnotationEvaluator` is inactive.

Named entity recognition engine

Machine learning techniques

We rely on the LingPipe library for the implementation of the NER annotator. We use the HMM chunker with character language model emission probabilities described in Carpenter 2007, LingPipe for 99.99% Recall of Gene Mentions. LingPipe took part in the Biocreative II annotation task and obtained decent performance with this model; therefore we use directly the trained model available on their website as a component of our annotation pipeline.

Preliminary evaluation

We conducted an evalutation on the provided development data and obtained the following results:

Precision	76.85%
Recall	84.88%
F-measure	80.67%