ID:qiqis

1. Architectural Aspect

Design Pattern:



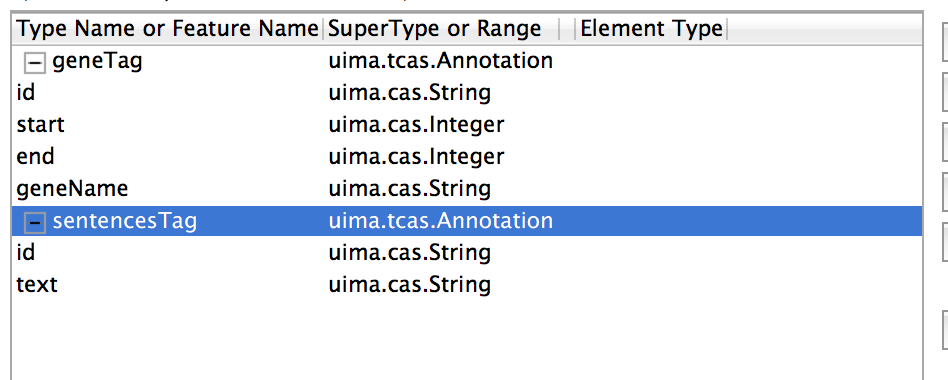
Collection Processing Engine (CPE)processes collections of artifacts (documents) through the combination of the following components: a Collection Reader, an optional CAS Initializer, Analysis Engines, and CAS Consumers. And CpeDescriptor is responsible to combine these components and set the run flow of those component.

Collection Reader is responsible for obtaining documents from the collection and returning each document as a CAS. Like all UIMA components, a Collection Reader consists of two parts — the code and an XML descriptor.

sentenceEntity is responsible for get sentence information from document and divide each sentence by its attributes. Like all UIMA components, a sentenceEntity consists of two parts — the code and an XML descriptor.

CAS Consumer receives each CAS after it has been analyzed. CAS Consumers typically do not update the CAS; they typically extract data from the CAS and output it.

Type System:



sentencesTag defines each line of sentence and store it as id and text.

geneTag defined each geneEntity in sentence and store it as id, start, end, and geneName.

The relationship between sentence and geneEntity is as follows. Each sentence can have 1 to many GeneTag.



1. Algorithm Aspect

Basically, I used Lingpipe to realize the function of Gene Name Entity Recognize.

1. machine learning techniques used:

Specifically, named entity recognition involves the supervised training of a statistical model or more direct methods like dictionary matching or regular expression matching. But I simply use a trained model: HmmChunker

1. NLP techniques/components used:

LingPipe is tool kit for processing text using computational linguistics.

1. external (marked up text) training data used

Yes, I use HmmChunker. This one is labeled by task (ne for named-entity recognition), language (en for English), genre (bio for biology) and corpus (genetag for the [GENETAG](http://www.biomedcentral.com/1471-2105/6/S1/S3) corpus), and suffixed with the name of the class of the serialized object (HmmChunker for com.aliasi.chunk.HmmChunker)

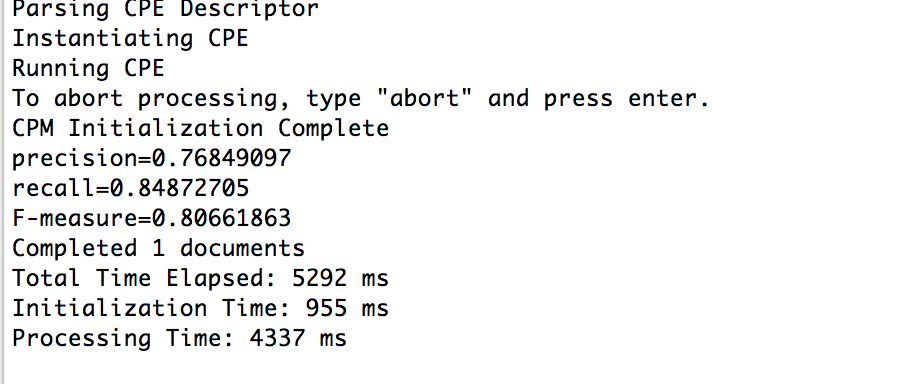
1. Please describe any rule sets used:

It is related to LingPipe rule sets.

1. The general data flow in your system:

CollectionReader first reads sentences from file(which is delivered in the function as a parameter) and delivers it to GeneEntity where sentence text are coped by LingPipe and gain the start and end position of GeneEntity. CasCustomer output the result to a file.

1. Evaluation



I implement the evaluation component in CasCustomer by Calculating the value of Precision/Recall/F-Measure. As the figure above, the F-measure is more than 0.8066 which means this algorithm and architectural implementation is efficient.