Name: Hongfei Zhang
Andrew ID: hongfeiz

Homework 1 Final Report

1. Type system

The Type System Descriptor for my annotator is located in the file /src/main/resources/types.xml. Two types and OutputGene are described in the file.

1.1 InputSentence has two features id and text to record the id and text to put in type system.

1.2 OutputGene has two features id and geneTag to describe the id and gene mention tags of sentences after processing.

2. IIS design

2.1 In the collection reader, a line-by-line reader is designed and implemented.

2.2 In the collection reader, id and text are seperated.

2.3 In the annotator, whitespace-excluded offsets are calculated .

2.4 The input and out file directory configuration parameters are used in collection reader and cas consumer.

3. Pipeline Design

3.1 GeneCollectionReader

It reads document from a directory line by line and splits each line into two parts: id and text. The input file is in *InputDirectory*.

Initialize(), getNext(), getProgress() and hasNext() methods are overridden.

3.2 GeneAnnotator

It detects gene mention tags from input CAS using Stanford Core NLP tool. The method getGeneSpans is in PosTagNamedEntityRecognizer.java

process() method is overridden.

3.3 GeneCasConsumer

It writes recognized gene tags into a destination file, where UTF-8 encoding is used.

Initialize() and processCas() methods are overridden.

4. NLP tool

The stanford Core NLP is used in the NER systemz. Which is implemented in GeneAnnotator.java. After all gene-like candidates from a sentence are extracted, it eliminates the space characters and output the information as required to cas.