11-791: Design and Engineering of Intelligent Information System

Homework 1

Zexi Mao
zexim@andrew.cmu.edu

September 24, 2014

1 Architecture

1.1 Type system

The type system for the gene mention tagging task consists of the following type:

- `types.Sentence`, which extends `org.apache.uima.jcas.tcas.Annotation`, is a type for input sentences. The text of a sentence is stored in the `org.apache.uima.jcas.JCas` object used to create the `Sentence` object. It also contains a feature `sentenceId`, which is a `org.apache.uima.cas.String` object for storing the unique ID for each sentence.
- `types.GeneEntity`, which also extends `Annotation`, is a type for the recognized gene entities. Both the start-offset and end-offset are stored in the `Annotation` directly. An extra feature in `GeneEntity` is `entityText`, which is also a `uima.cas.String` object for storing the text piece recognized as a gene entity.

1.2 Collection processing engine

This collection processing engine used in this task is `SimpleRunCPE.java`. It parses the CPE descriptor `CpeDescriptor.xml` and runs the collection reader, annotator, and CAS consumer in a pipeline.

For increasing the processing speed both `casPoolSize` and `processingUnitThreadCount` are set to 3, as 3 CPU cores are assigned to my Ubuntu virtual machine. Setting `processingUnitThreadCount` to 1, the run time on my machine is 36.0s; setting it to 3, the run time is 21.9s. Please feel free to change the two parameters for different hardware configuration.

1.3 Collection reader

The collection reader uses a `BufferedReader` to read the input file, each line (sentence) at a time and store the `sentenceId` and text into `Sentence`. The collection reader has a configuration parameter for the input file path, which is set to "hw1.in" in the release version.

1.4 Annotator

The annotator processes each `Sentence` at a time, annotates it with gene entity mentions and produces zero or more `GeneEntity` at a time.

1.5 CAS consumer

The CAS consumer takes the produced `GeneEntity` and outputs all the annotated gene entity mentions to a file. It also has a configuration parameter for the output file path, which is set to "hw1-zexim.out".

2 Method

This gene mention tagging method is based on the part-of-speech tag named entity recognizer. It first uses a tokenizer to tokenize the sentence into tokens, and then label each token with a part-of-speech tag. After that, the consecutive nouns will be grouped into "noun phrases", which will be added to the list of name entities. Finally, to reduce the amount of wrong gene mentions returned, several rules-based methods are used to clean-up the recognized named entities. For example, a name entity with a single character is discarded, a short word which is not a abbreviation is discarded, etc.