HW1- 11-791 Date: 21st September 2014

(HW1) UML Design and Named Entity Recognition Implementation with UIMA SDK- Report

Name: Niloy Gupta AndrewID: niloyg

1. Please identify/describe any machine learning techniques used:
A Hidden Markov Model (HMM) was used to the train the named entity classifier. The HMM chunker is part of the LingPipe package, which is used for tagging. According to the documentation it uses a character model for each tag and a "maximum likelihood bigram transition model".

2. Please identify/describe any NLP techniques/components used:
We use the N-best search algorithm.
General NLP techniques use a top-down search that for each word consult a knowledge source to determine the probability for which words comes next. Many of these knowledge sources contain a long-distance effect (association between words that are fart apart) the search space can be quite large. N-best algorithm ensures that the knowledge sources are applied in the proper order whereby the search space is reduced. It employs the most accurate and computationally efficient knowledge sources to generate the top N hypotheses. These N hypothesis are evaluated by further knowledge sources.

4. Please identify/describe any external (marked up text) training data used:
The GENETAG (a tagged corpus for gene/protein named entity recognition) dataset was used for training the Named Entity algorithm. Particularly the file genetag.tag was used.

5. If your system interacts with or uses data from any biological database(s), please describe:

There was an initial attempt to use the Entrez Gene database provided by the National Center for Biotechnology Information (NCBI), however it was later removed due limitations of the avaialble Java wrapper.

6. Please describe the general data flow in your system:

The Collection Processing Engine (CPE) beings by instantiating the collection reader (GeneERCollectionReader). The collection reader reads the input data line by line. Each line is a sentence comprising is an excerpt from biological literature prefixed by a line identifier. Each line is marshalled into a UIMA CAS (Common Analysis Structure) object. Each CAS object is passed through the analysis engine. In this implementation the Analysis Engine comprises only of one annotator (GeneAnnotator).
GeneAnnotator uses pre-trained Hidden Markov Model (HMM) which was built using the LingPipe package. The trained model is kept under resources/data. This trained model is loaded during the initialize phase in the collection reader.
The LingPipe chunker returns the identified named entity i.e. genes. These are indexed inside the CAS.
The CAS is then forwarded to the CAS Consumer, which prints the named entity (gene) in the specified format ("sentence-identifier-2|start-offset-1 end-offset| gene-name").

UML Class Diagram of the system: