HOMEWORK 1

Rahul Goutam, rgoutam 09/24/2014

Task 1: Report

The architecture of my IIS is shown in the figure below. The architecture consists of three main components, collection reader, analysis engine and CAS consumer.

The function of the collection reader is to read one sentence at a time from the text document, initialize the CAS with the sentence to be analyzed and send the CAS to the analysis engine.

The analysis engine contains the meat of the IIS. In this case, the analysis engine has a single component - the Named Entity Analysis component. The analysis engine receives the CAS from the collection reader, processes it to tag named entities, adds the named entity data to cas and sends it to the CAS consumer.

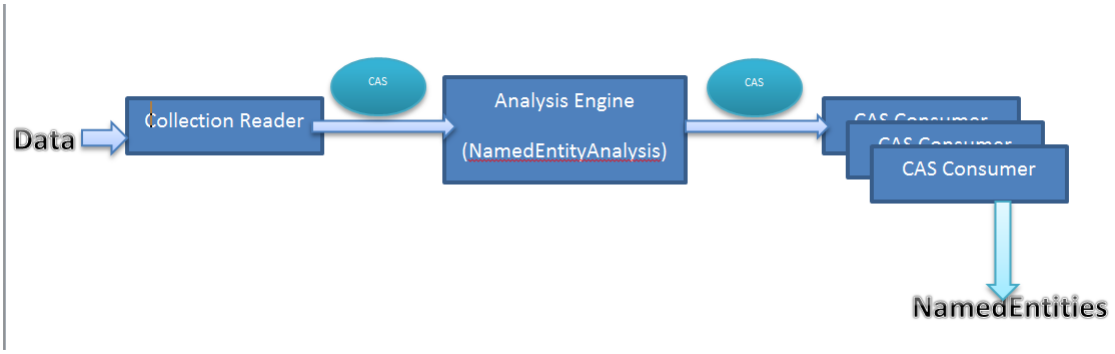
The CAS consumer receives the CAS with analysis data from analysis engine, extracts relevant data from CAS and writes to file.

The analysis engine contains the Named Entity Analysis component. The named entity engine is a statistically trained model that has been trained on the GENETAG corpus. This model is provided by lingpipe. GENETAG corpus is a tagged corpus for gene/protein named entities. The model is basically a hidden markov model (HMM) that sequentially tags words into named entity categories. Each named entity can span across multiple contiguous. A word in the beginning of the named entity is tagged as B-<NER> and word in the middle of the named entity is tagged as I-<NER> where <NER> stands for the symbol of the specific named entity.

1. Machine learning technique used : Hidden markov models for the sequential tagging task of Named Entity Recognition. The HMM model was trained on the GENETAG corpus which is a corpus of tagged gene/protein named entities.

2. NLP technique/component used : Named Entity Recognition is a NLP technique to tag named entities in text. The analysis engine contains the NER engine.

3. No external training data has been used.



4. No external lexical resources used.
5. No rule sets used
6. No interaction with any other biological database
7. Relevant resources used : I have used a pre-trained model that has been trained on the GENETAG corpus. This model is publicly available on the lingpipe website.
8. Data flow : The data consists of sentence ids and sentences. This is read from the input file by the collection reader. Each sentnece and the corresponding sentence id are added to a CAS and sent to the analysis engine. The analysis engine extracts the named entities from the sentences and adds them to the cas. The cas consumer writes the sentence ids and the named entities into a file.