Design and Engineering of Intelligent Information Systems

Alok Kothari

September 24, 2014

1 Report

I used a Hidden Markov Models- based named entity recognizer (lingpipe) for this task. The lingpipe used a model file "English Genes: GeneTag". I did not have to use any external training data, lexical resources or rule sets. I did not connect to any biological database either.

General data flow The Collection Reader (NewCollectionReader.java) reads the input file. The name of the input file is supplied as a parameter: "InputString". It uploads each sentence as a CAS. This then read by the Annotator, which then annotates the named-entities in the sentence. The annotator uses lingpipe. The model file for lingpipe is supplied to the annotator through a parameter: ModelFileName. The type system is as follows: for NEAnnotate (The Annotator) the features/attributes are begin, end, NamedEntity ? used by Annotator. SentenceAnnotation has features Sentence, sentid ? used in the collection reader.

The annotated entities are pushed to CAS again. The CAS consumer reads this from the CAS and writes the output file. The name of the output file is supplied through the parameter 'outputfilename?.