Austin Matthews (austinma)

10/17/2012

# 11-791 HW1 Write up

For this homework I built a UIMA pipeline consisting of three major components: a CollectionReader, an Annotator, and a CAS Consumer. These three components are described and link together by a CPE descriptor as is standard with the UIMA framework.

The CollectionReader is a modification of the FileSystemCollectionReader that is packaged with the UIMA SDK. The original read in all the files in a given directory, but I modified it to read in only one single file, specified in a parameter. It reads in this file, and converts its entire text into a single sofa.

The second component, and meat of the project, is the Annotator class. This piece was too loosely based on an annotator from the SDK. It has a process() method that is called once per sofa. In our case, that is once per input file. The process() method splits the input file into lines, detects the sentences' unique IDs and text, and passes the text through a gene-tagging named entity recognizer. For each named entity picked up by the recognizer, it creates an annotation object, complete with the sentence ID and span of text corresponding to the named entity. The named entity recognizer used is one freely available from LingPipe. They model the data with a Latent Dirichlet allocation, trained on BMC's GENETAG corpus.

Finally, there is a CAS consumer that was again based on an example from the SDK. The consumer simply loops over all the annotations produced and outputs them to an output file. Similar to the input mechanism, the output file is specified as a parameter of the CAS consumer. It formats them according to the specification, reading the sentence ID and text covered by the span from the annotation object.

Overall this approach seems to be quite successful. On the given test data, this approach correctly identified 15504 of the 18265 gene annotations in the reference output, giving it a recall of 84.88%. Furthermore, 15504 of the 20174 spans it tagged as genes were, in fact, in the

reference set, giving it a precision of 76.85%. Combined, this means the system's F1 score was 80.67%.