Exercise 4 – Documentation

Florian Eckerstofer ()

Haichao Miao (0726810)

Group C

CFPExtractor.java:

This class contains the entire code. In general a controller is initilized and the Batch Learning PR creole plugin is added into it. The Batch Learning PR is running twice. First in Training mode, where classes are created for Machine Learning model. The Second time is in Application mode, where we apply our Machine Learning model. The resulting Annotations will be saved into an XML-file.

Batch Learning PR

This Creole Plugin offers different runmodes, that is sufficient for our purpose. It uses the machine-learning approach to learn patterns from our already annotated training corpus. First we have to configure through an configuration file (see ml-config.xml).

Training:

The training corpus will be presented to the machine learning algorithm. The data must be well defined for the machine learning algorithm.

1. instances: These are objects, that are considered and learned by the learning algorithm. Every instance influences what the algorithms learns. Tokens are very convenient here.
2. attributes: The fraction of information of the instance. In our case the features of a Token. e.g. String, Length.
3. class: What we want to learn.

Example:

The **"International School on Information Extraction"** (Token-Feature: String) is a Instance of the Class **conferencename**.

Application:

When Batch Learning PR is applied under the Application Runmode, ut creates new Annotations on our test corpus. The new Annotations are saved in the AnnotationSet "machine_learned". They can now compared and evaluated against the original markups, to see how well the machine learner worked.

Problems during Implementation:

1. Although the Gate-API is very well documented only few information could be found the the configuration on the Creole Plugin "Batch Learning PR".