Named-Entity Recognition Project Documentation

Natural Language Processing and the Web

Florian Schneider, Julian Betz December 2, 2017

1 Newly implemented classes

The following classes were newly implemented in `de.unihamburg.informatik.nlp4web.tutorial.tut5`.

1.1 `writer.NERWriter`

An Analysis Engine that processes a CAS to generate the evaluation file and to calculate statistics for the `NEIOBAnnotations` generated earlier in the pipeline. It finds all `NEIOBAnnotations` that have an attached prediction value and searches for the `NEIOBAnnotation` that has the corresponding gold standard value. Each gold standard / prediction pair, along with the corresponding token, is printed to a text file that can be used as input for the evaluation scripts.

For each gold standard value, the number of predictions of each named-entity type is shown in the output. As an aggregate result, the absolute and relative amounts of correct classification are given. Furthermore, a table is generated that shows the number of classifications of a token as a named entity or non-named entity. Finally, the absolute and relative amounts of correct classification of tokens that are named entities according to the gold standard are given. (This number may be of interest as there are far more non-named entities than named entities in the data, according to the gold standard.)

1.1.1 Configuration parameters

The configuration parameter `PARAM_FILENAME` is the filename of the evaluation file to be generated.

The configuration parameter `PARAM_NULL_TYPE` determines which string is used for marking non-named entities in the input. (Set to "O" in our case.)

If the configuration parameter `PARAM_VERBOSE` is set to `true`, all incorrect predictions are printed out to the log before printing the statistics.

The configuration parameter `PARAM_EXPECTED_ENTITY_TYPE_NUM` is used for the initialization of data structures and only affects efficiency, but not functionality.

2 Adapted classes

The following classes were adapted in `de.unihamburg.informatik.nlp4web.tutorial.tut5`.

2.1 `ner.ExecuteNER`

The Analysis Engine `writer.NERWriter` was added to the end of the pipeline in `classifyTestFile`.