Analyzing German Noun Compounds using a Web-Scale Dataset – Report Week 10

UIMA Software Project WS 2010/2011
Jens Haase

1 Work done in the last week

1.1 UIMA component

To use the generated code of the last weeks in real applications a UIMA component was developed. This component can later be used in different projects. It consist of two annotator and one Annotation Type. The two components are nearly identically. They require a index path and a token class as parameter. As optional parameter the ranking algorithm can be set (default is the frequency algorithm). The difference of the two components is that one uses the ranking algorithms on a tree and the other on a list, as we saw last week.

Before using the component in an analysis engine a tokenizer is required. The component iterates over the all token and tries to split each one. As result a set of SplittedToken Annotation will be added to the CAS. Each Annotation contains a StringArray with the list of individual words. In each individual word morphemes are marked with brackets. For example the token “Aktionsplan” is splitted in “Aktion(s)+plan” than the resulting string array is [“aktion(s)”, “plan”].

1.2 Code cleanup and Refactoring

For the final delivery a lot of code cleanup and refactoring was done. Javadoc was added to all classes and command line tools are created.

1.3 Evaluation on a larger dataset

At least the evaluation ran on a larger dataset. Currently 10,000 entries were evaluated. A evaluation of the the complete dataset takes several days or requires a faster machine. Here are the results:

Algorithm	Recall tree	Recall list
Frequency Based	0.5226 (0.7204)	0.5078 (0.7029)
Probability Based	0.1815 (0.2518)	0.1843 (0.2558)
Mutual Information Based	0.2952 (0.4415)	0.3681 (0.542)

1.4 Time Tracking

Date	Task	Needed time	Planned time
2011-01-30	Javadoc	2	2
2011-01-31	Refactoring and Cleanup	6	8
2011-02-02	Uima Component	4	4
2011-02-02	Report Writing	2	3