Towards robust multi-tool tagging: An ontology-based approach

In the realm of morphosyntactic annotations, ensemble combination techniques have been successfully applied to obtain more robust and more reliable linguistic analyses. Ensemble combination architectures employ multiple classifiers, e.g., part of speech taggers trained on different corpora, and combine their output, e.g., by means of a majority vote, phenomenon-dependent selection preferences or more elaborate techniques.

These techniques are, however, ultimately based on a string comparison of annotation values. Therefore, traditional ensemble combination architectures are restricted to classifiers that make use of the same annotation scheme. When tools trained on different annotation schemes are combined, not only an increase of robustness, but also an increase of the level of detail of the analyses can be expected, if the respective schemes differ in their granularity.

Here, I present an approach that addresses the challenge to integrate annotations from different annotation schemes. With the help of an ontology of linguistic annotations, tags are translated into sets of ontological descriptions. Using these sets, annotations produced by different tools can be compared and combined irrespectively of their original string representation.

It will be shown how annotations created by multiple NLP tools are mapped onto such tool-independent descriptions, and how a majority vote and ontological consistency constraints can be used to integrate multiple alternative analyses of the same token in a consistent way.

This approach will be illustrated for seven tools for the morphosyntactic and morphological analysis of German. For morphosyntactic (parts of speech) and morphological annotations of three newspaper corpora, the resulting merged sets of ontological descriptions are evaluated in comparison to (ontological representation of) existing reference annotations.

Finally, some experiments are described to extend the scope of this approach to syntactic annotations.