RULE-BASED AUTOMATIC MULTI-WORD TERM EXTRACTION AND LEMMATIZATION

Cvetana Krstev, Ranka Stanković and Duško Vitas

University of Belgrade

A pipe-line for MWT processing in Serbian (WG1, WG2)

- Domain specific corpus preparation;
- Multi-word term extraction using rule-based approach (finite-state transducers and lexical resources)
- Multi-word term lemmatization and rejection of (some) parasite candidates (resource based)
- Ranking of candidates according to various association measures (unithood and termhood, domain specific and general corpora)
- Human evaluation
- Dictionary entry production