Interfaceing linguistic analysis tools with a database for result management – workflows in sentence and text analysis

Kurt Eberle, Kerstin Eckart, Ulrich Heid
(eberle,ekhart,heid)@ims.uni-stuttgart.de

Three types of relations between analysis components – integrated environment for quality assurance in corpus based linguistic analysis

Vertical relations
- Pipeline architecture of text processing
- High level analysis, e.g., constituent trees, depends on results of lower level, e.g., word sense disambiguation
- Advantages for corpus studies:
  - Shared interest in lower levels
  - Higher level components more efficiently from results of lower levels
  - Reusability of intermediate results
- Pre-requisites:
  - Analysis tool should support pipeline architecture
  - Analyses are stored and administrated for later reuse

Horizontal relations
- Different tools producing analyses of a particular level, e.g., dependency analyses
- Taking corresponding results of the advantage into account
- Advantages to facilitate quality assurance of the annotations:
  - A analyst have to be identifiable with respect to their horizontal status, i.e., analysis level and representation format
  - Performance of analyses for non-interchangeable, e.g., into an abstract exchange format such as GAF

Temporal relations
- Analysis tool evolving over time
- Producing analyses for the same input but with different versions of a tool
- Advantages of system development:
  - Non-technical features of tool improvement or decline, or
  - Specifi city of the knowledge base
  - Identification of side-effects by comparing earlier versions of the analysis
  - Frequent changes:
    - Information about tool and component versions
    - Analyses have to be relatable to the tool or a notion of producing them

Relational database
- ID DB, implemented as a PostgreSQL database
- Type system identifies the horizontal status of an analysis
- Relating analyses and tool versions
- Displaying an annotation level and representation format
- Workflow modeling identifies vertical status of an analysis
- Relating input and output with the analysis level
- Relating tool versions that evolve over time
- Flexible queries as is conducted via SQL

Multi-level processing tool
- BI-analysis tool, based on a research prototype of the German parser of the deep machine translation product (Eberle et al., 2008)
- Adapted to collate native linguistic research
  - Pipeline where each a notation level can be extracted separately
- Modules for morphology, syntactic, semantic and text semantic parsers and analyses
- Structured analysis settings provide the complete knowledge needed by subsequent analysis steps of the pipeline
- All levels can be used at a detailed analysis
- Analyses are connected to each other by text and sentence identifiers

Use case

Task 2: specific disambiguation of German unnominalizations:
- Each PPs in combination with nominalizations of verb or direct
  - Morphological analysis
- Optional parameters to fine-tune the corresponding analysis
- General form:
  - Creating using syntax analyses from the morphological one
  - SQL:
  - Analyzing analysis(3,315,315,de,morph),
    (3,315,de,syn),
    (3,315,de,info).

Future work
- Technical extension: interface to an environment to full database
  - Capabilities and a platform independent version of tool and interface
- Architectural extension: taking into account horizontal relations and further analysis levels, such as DRS representing semantic structures

Example

Primary data
Sentence from local news (file 3, sentence 315):
He verlebte nach seiner Mitteilung in stationärer Krankenhausbehandlung.

First step: morphological analysis
- Sentence from local news (file 3, sentence 315):
  - Creating a syntactic analysis
  - Syntax analysis(3,315,syn), de,syn,
    (3,315,de,morph),
    (3,315,de,info).

Creating further steps, e.g., syntactic analyses
- Direct from sentence:
  - DB analysis sent(3,315), de,syn,
    (3,315,de,morph),
    (3,315,de,info).

Pronoun resolution, e.g. with two preceding sentences
- DB analysis sent(3,315,syn), de,syn,
  (3,315,de,morph), de,syn,
  (3,315,de,info).