Motivation

- Language resources:
  - Highly structured linguistic information
  - Qualitative information: "small"
  - Metadata for documentation
  - Keywords
  - Deep structure

Data Structure

- "Large" data set provided by project partner
  - NDA: data and concrete application
  - Application in the medical and pharmaceutical domain
  - Porting to other data of multimodal annotations
  - Data highly structured terminology database available for the appropriate domain

Search Grammar

\[ G = ( \Phi, T, R, \Psi ) \]

Non-terminal symbols:

- \( \Phi = \{ \text{Search}, \text{Drug}, \text{Prevention}, \text{Treatment}, \text{Side Effect}, \text{Links}, \text{etc.} \} \)

Terminal Symbols:

- \( T = \{ \text{pain killer}, \text{adult}, 10mg, \text{cancer}, \text{fever}, \text{Terms}, \text{etc.} \} \)

Rules:

- Search → Drug (Context of Disease) (Additional Drug) (Refinement)
- Context of Disease → (Drug Therapy) (Prevention) (Diagnosis) (Coexisting Disease) (Side Effect)
- Additional Drug → (Combination) (Comparison) (Interaction)
- Refinement → (Type of study) (Age group) (Sex) (Route of Administration) (Dosage) (Free Search) (Duration of Treatment)
- Drug → (⋯)

Implementation

- Standard based HTML
- XQuery
- TBX
- AJAX
- High performance
- XML-Database engine
- Rails framework

INFORMER Interface

Probabilistic Search Interface Element Ranking

- Regular grammar
- Controlled vocabulary
- Possibility of probability estimation for better interface integration

\[ P(\omega_1, \omega_2) = P(\omega_3, \omega_4) \]

- Relative Frequency (RF) used for maximum likelihood estimation (MLE)
- General equation for word sequence of length \( n \):

\[ P(\omega_n) = \frac{C(\omega_n)}{\sum(C(\omega_{1+n}))} = 0.00..1.00 \]

- Assumption: recurrent patterns of search queries
- INFORMER can be optimized by statistically modeling user behavior
- RF provides statistical joint distribution of search queries in use

Complexity

- Complexity and its reduction:
  - Front-end: Usability
  - Back-end: Computational complexity, processing time

INFORMER Interface

Synonym Search

Termbank use: TBX-Termbase
- Concept based lexical resource
- Search by synonym, hyponyms, related concept
- Language restrictions: same language, all languages, specific language

Conclusion

- Usability of resources improved
- Metadata used
- Processing complexity reduced to linear complexity
- Untrained users:
  - High precision
  - High recall
- Selection of sub-corpora for linguistic phenomena

Future work

- Advancement to more linguistic resources
- More generic approach for tailoring the interface
- Visualization of results
- Reporting for technical analysis and optimization

Contact: thorsten.trippel@uni-bielefeld.de