Text-Driven Ontology Generation and Extension in the Finance Domain

Mihaela Vela
Language Technology Lab
DFKI Saarbrücken
European MUSING project

- Development of Business Intelligence tools and modules based on semantic knowledge and content systems
- www.musing.eu
Our Approach

- Ontology generation and extension on the base of textual patterns and various natural language annotation layers
- Multi-layer approach for building T-Box elements
- Corpus: German economical news articles
Processing Layers

- String processing
- Morpho-syntactic information
- Chunking and dependency information
- Accessing Semantic Resources
  - Diagonal to the other processes

21.01.2009
Processing of Text Patterns

- Extract nominals
  - Candidates for concepts
    - *E.g.: konzern, firma, chef*

- Extract compounds
  - Restricting the number of suggested concepts to those occurring in compounds
  - Candidates for relations between suggested concepts
    - *E.g.: adressdatenbanken[SuggestedConcept + suffix] ➔*
      - `objectProperty(adress, datenbanken)`
      - `adressdatenbanken[prefix + SuggestedConcept] ➔ subClassOf(adressdatenbanken, datenbanken)`

    - *E.g.: konzernchef[SuggestedConcept + suffix] ➔*
      - `objectProperty(konzern, chef)`
      - `konzernchef[prefix + SuggestedConcept] ➔ subClassOf (konzernchef, chef)`
Filtering and Consolidation

- **Reformulation of compounds**
  - SuggestedConcept + PREP[mit] + Suffix
    - *E.g.: Datenbanken mit Adressen* → `subClassOf(adressendatenbanken, datenbank)`
      - Consolidated subClassOf
      - Filter out objectProperty
  - SuggestedConcept + PREP[von|von dem|vom] + Prefix
    - *E.g.: Chef vom Konzern* → `objectProperty(konzern, chef)`
      - Consolidated objectProperty = HAS
      - Filter out subClassOf

- **More precise determination of textual patterns that can be used for OG**
  - SuggestedConcept + PREP[mit] + NOUN_MODIFIER + Suffix
    - *E.g.: Datenbanken mit Milliarden Adressen*  
      - amount of address data
  - SuggestedConcept + PREP(von|von dem|vom) + ADJ_MODIFIER + Prefix
    - *E.g.: Chef des deutschen Konzern*  
      - subclassification of bank
  - This to come to relevance for OG in next processing steps
Examples of Compounds to be Considered

- candidates for A-Box
  - needs first NE detection and recognition
    - E.g.: colonia-konzern, nokia-konzern, lornho-chef, iata-generaldirektor
- To be considered in next processing steps
Statistics

Total number of tokens: 168593

<table>
<thead>
<tr>
<th>Processing stage</th>
<th>Concept</th>
<th>Compounds</th>
<th>Reformulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept selection</td>
<td>20109</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Compound selection</td>
<td>3408</td>
<td>11163</td>
<td>-</td>
</tr>
<tr>
<td>Compound filtering</td>
<td>291</td>
<td>370</td>
<td>1093</td>
</tr>
</tbody>
</table>

*Wirtschaftswoche 1992*

Need to extend the volume of the corpus or access other corpora for searching for compound reformulations
Other related points

- Strategy for semi-automatic validation/rejection/specification
  - including when necessary/possible intervention of domain experts and ontology engineers

- Issues to be dealt with in the next steps
  - Relation transitivity
    - More structure in the ontology
      - E.g.: iata-generaldirektor + iata-konzern \( \Rightarrow x(\text{konzern, generaldirektor}) \)
  - Morphological variations
  - Fine-grainedness of relations
  - Filling A-Box
  - Consider broader context for OG
  - Domain and range
Morphology and Semantics

- Find compounds based on COMP (less restrictions)
  - E.g.: <W INFL="[17 18 19]" POS="1" STEM="chef" COMP="firmen chef" TC="22">Firmenchef</W>

- Class instantiation (A-Box) as a side effect
  - E.g.: bayer-konzern, busse design

- Deal with morphological variations
  - E.g.: firmenchef, bankenchef

- POS and semantic resources for compound components
  - E.g.: großkonzern vs. chemiekonzern
    us-konzern vs. mega-konzern
2nd Layer Conclusions and Results

- existing ontology is consolidated on T-Box and expanded on A-Box
- Context is still narrow ➔ only word analysis
Dependency Information

- Broader context
  - E.g.: [NP-Subj Er] [VG soll] [PP im Konzern] [NP-Ind-Obj Finanzchef [NE-Pers Gerhard Liener]] [VG folgen]
  - Finanzchef vs. Konzernchef
Conclusions

- Build/expand finance ontology in 3 stages
- Trackability and versioning between the different layers and runs