The Semantic Gap of Formalized Meaning

Employment of Semantic Web Technology in the domain of Linguistics and Textmining

Sebastian Hellmann

AKSW, Universität Leipzig

Supervisor: Sören Auer and Jens Lehmann

Content

- Overview of Areas and Trends
- OWL as a Meaning Representation Language
- First Results
- Evaluation Methodology

Overview of Areas & Trends

- **Textmining**
  - many available NLP tools
  - poor representation of output, normally Strings only (e.g. POS Tags)

Overview of Areas & Trends

- **Linguistics**
  - emerging Domain Knowledge in OWL, e.g. by Christian Chiarcos – *Ontologies of Linguistic Annotations*
  - hardly any adapted Semantic Web tools to support elicitation, creation and maintenance

Overview of Areas & Trends

- **Ontology Learning**
  - fragmented approaches, mostly only one or few NLP methods as input [1, 2]
  - preprocessing step can be optimized

1. J. Völker, P. Hitzler, and P. Cimiano. *Acquisition of OWL DL axioms from lexical resources.* In ESWC, 2007.
2. Tamas Horvath and Gerhard Paass and Frank Reichartz and Stefan Wrobel, *A Logic-Based Approach to Relation Extraction from Texts*. ILP 2009

Overview of Areas & Trends

- **LOD cloud**
  - vast amount of structured knowledge
  - concept tagging is only a first step, further enrichment is possible

Goal

- Create a holitic approach, that combines techniques and knowledge from all fields.
- Implement a general purpose preprocessing API (NLP2RDF)
- Evaluate it thouroughly based on available benchmarks and tasks

OWL as an MRL

Text represented in RDF

NLP Tools
and Linguistic Ontologies
Parser – Ontology pair

Agregated RDF/OWL Model

Knowledge from the LOD Cloud

Additional Linguistic Knowledge

NLP2RDF stack

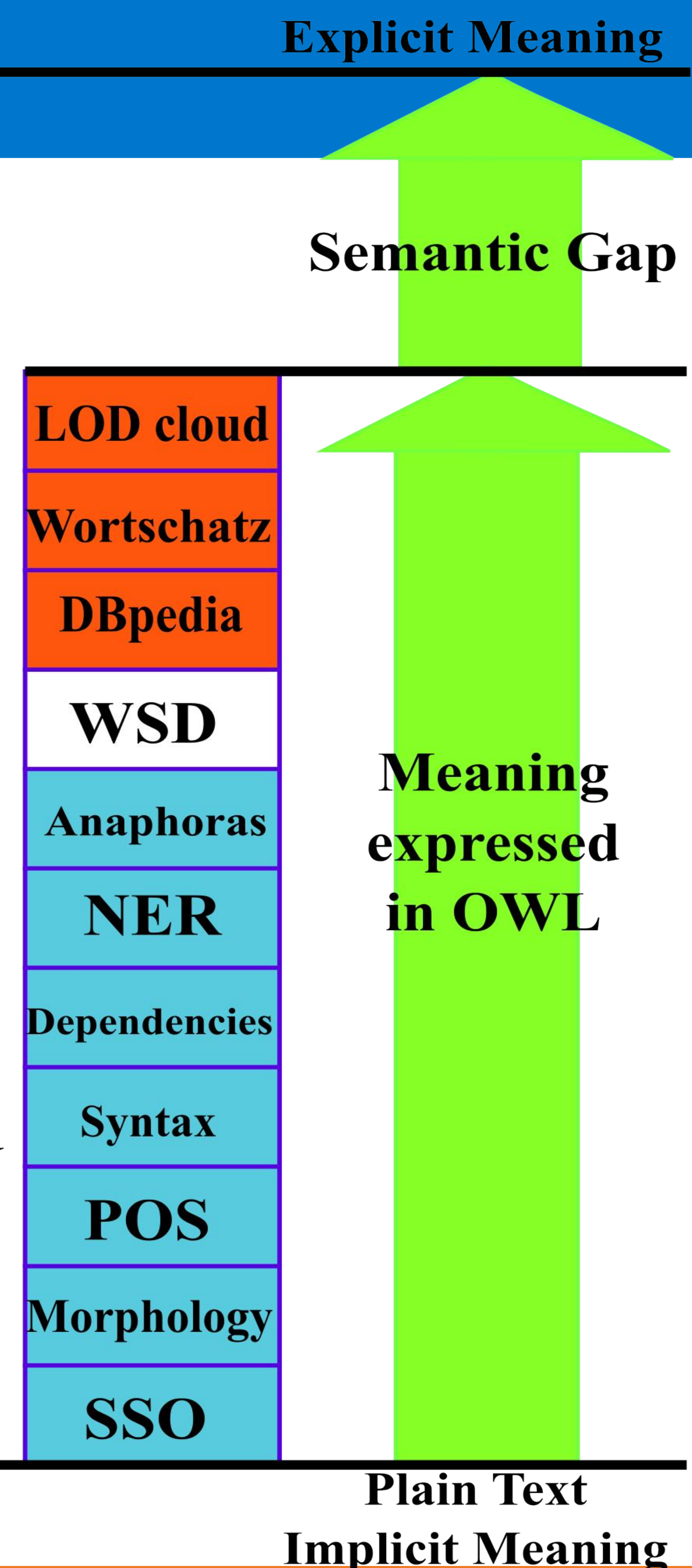
Open source implementation
http://code.google.com/p/nlp2rdf

Existing structured knowledge is select-ed, disambiguated and integrated

WSD connects top and bottom

Each NLP layer is augmented with linguistic background knowledge

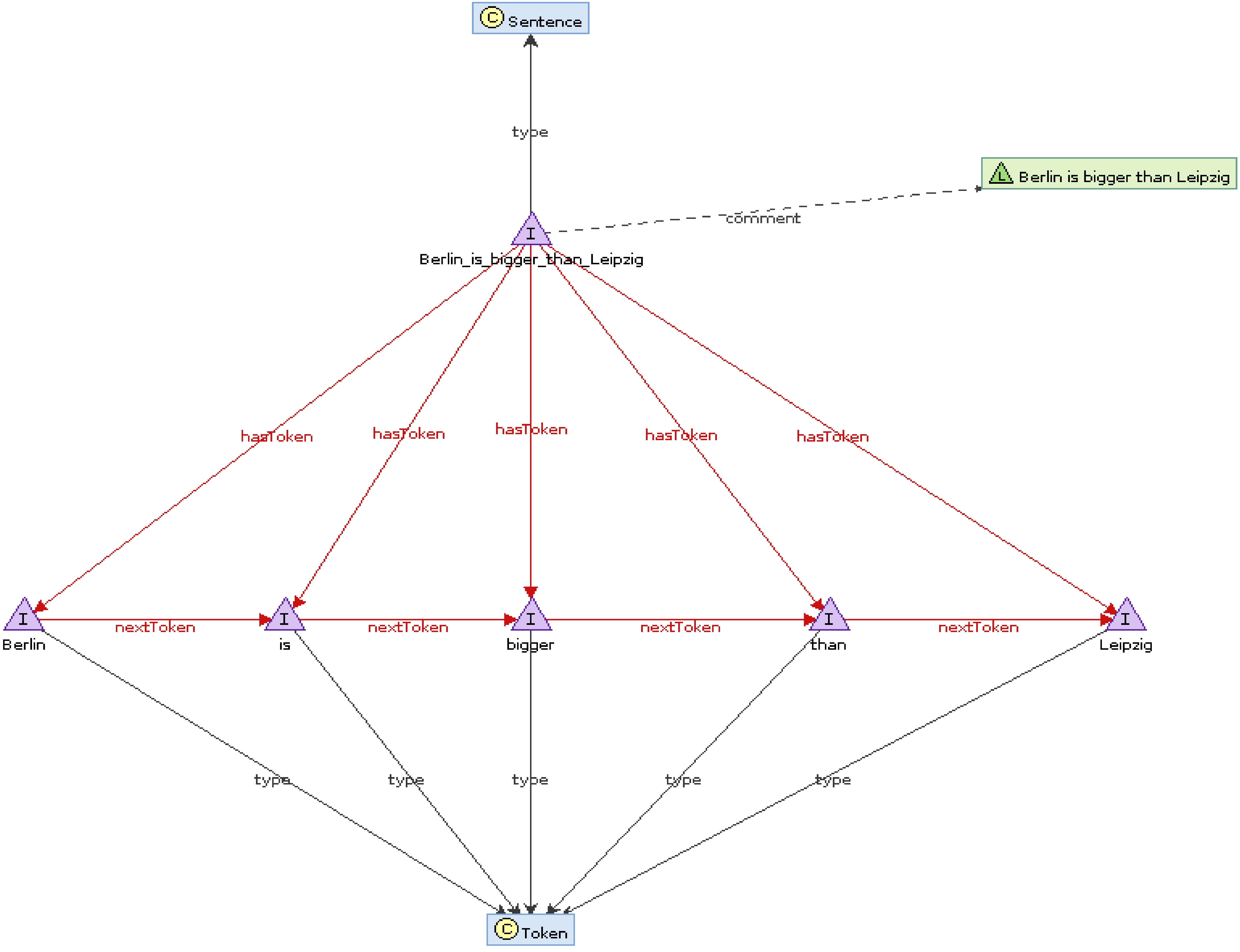
Backbone ontology



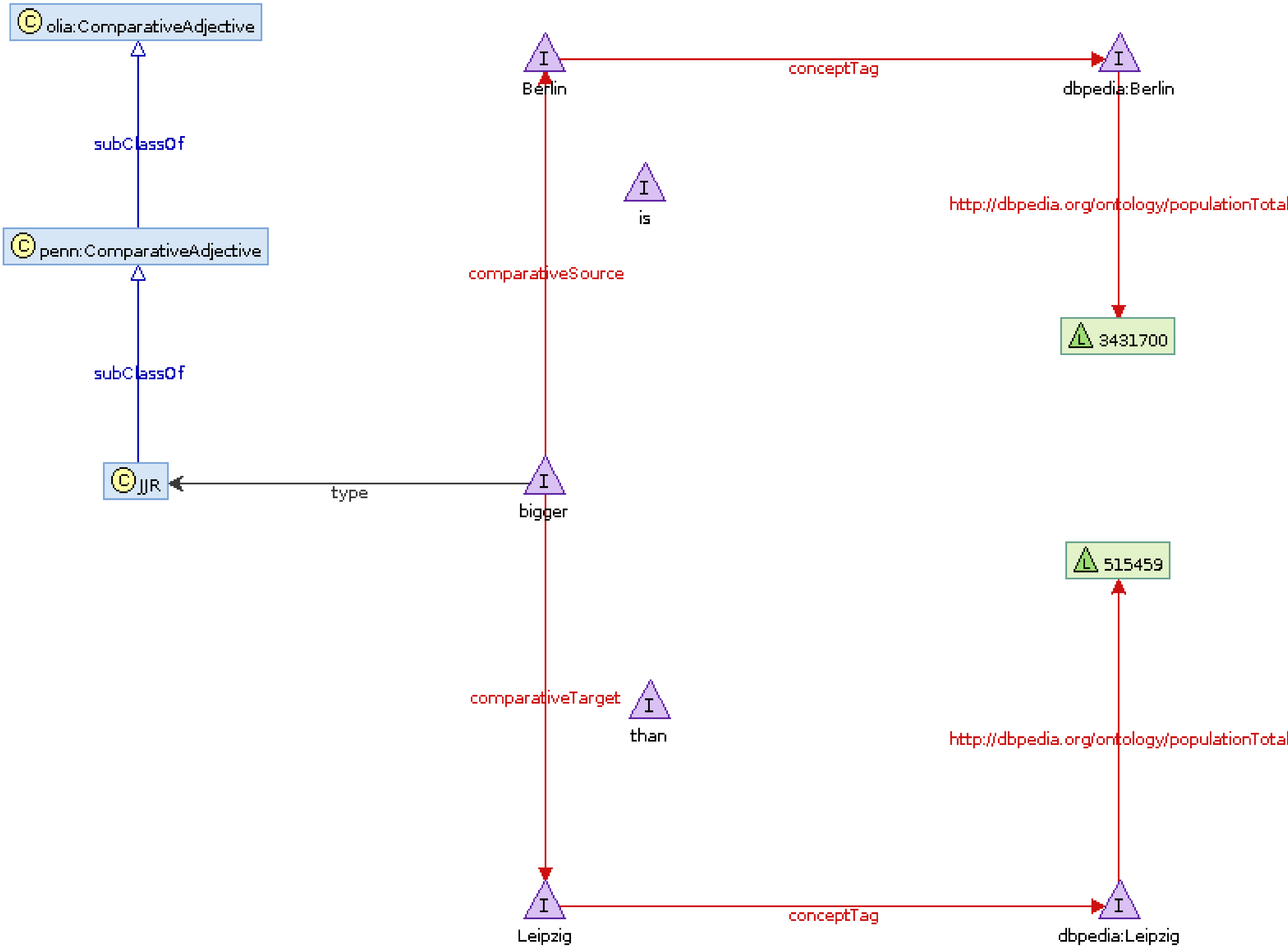
Example

Berlin is bigger than Leipzig.

Berlin | is | bigger | than | Leipzig | .



Sentence
type
Berlin is bigger than Leipzig
comment
Berlin_is_bigger_than_Leipzig
hasToken
hasToken
hasToken
hasToken
hasToken
Berlin
nextToken
is
nextToken
bigger
nextToken
than
nextToken
Leipzig
type
type
type
type
type
Token



olia:ComparativeAdjective
subClassOf
penn:ComparativeAdjective
subClassOf
JJR
type
Berlin
conceptTag
dbpedia:Berlin
is
http://dbpedia.org/ontology/populationTotal
comparativeSource
3431700
bigger
515459
comparativeTarget
than
http://dbpedia.org/ontology/populationTotal
Leipzig
conceptTag
dbpedia:Leipzig

```
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
```

```
SELECT * WHERE {
<http://dbpedia.org/resource/Berlin> ?p ?o1.
<http://dbpedia.org/resource/Leipzig> ?p ?o2 .
Filter (?o1 < ?o2 || ?o1 > ?o2 ).
Filter (?p LIKE <http://dbpedia.org/ontology/%>) .
Filter (xsd:int (?o1) || xsd:double(?o1))
}
```

Results: Browse Go! Reset

SPARQL results:

p	o1	o2
dbpedia:ontology/areaCode	"030"@en	"0341"@en
dbpedia:ontology/areaTotal	891820000	297600000
dbpedia:ontology/populationTotal	3431700	515459

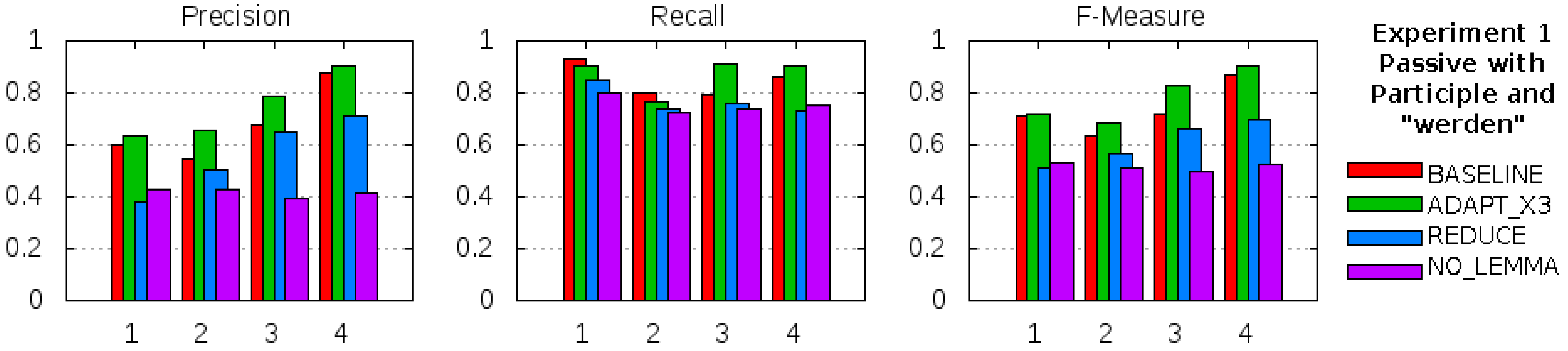
First results

- **Tiger Corpus Navigator**
  - TIGER is a collection of 50,000 German sentences
  - Conversion to RDF yielded 9 Million triples
  - Navigation with Active Machine Learning
  - http://tigernavigator.nlp2rdf.org/

Demo Tiger Navigator

http://tigernavigator.nlp2rdf.org/

Benchmarking



Find an OWL class description which covers 6300 passive sentences (of 50,000)

currently only POS tags are used

Benchmarking

- Creation of a benchmark suite with tasks from Textmining and Ontology Learning (like the 6300 passive sentence)
- NLP2RDF produces input for machine learning algorithms such as DL-Learner (as seen in the TCN)
- Which features are necessary for which tasks, how do they need to be represented?
- Improvement can be measured directly.

Next steps

- Planned benchmarks
  - ACE - 2003, Reuters ...
- Conversion of more linguistic corpora
  - Penn, Susanne, Wiktionary (DBictionary)
- Implementation of the NLP2RDF stack and framework
  - POS tags, Syntax and Disambiguation partially finished
- Using Ontology Learning on top
  - Integration with LExO (Johanna Völker)

Thank you

Collaboration

- good indicator of right direction
- better and faster results
- Currently:
  - Dr Christian Chiarcos, SFB 632, Potsdam
  - Dr Johanna Völker, KR & KM Research Group, Mannheim
  - Christian Meyer, UKP Darmstadt

Scientific core

- Development of algorithms (including evaluation)
- Creation and collection of benchmarks for Natural Language Engineering

Usefullness

- Creation and Enrichment of Linguistic Resources (Add bubbles to LOD cloud)
- Tools for Linguists (Search and Ontology Engineering)
- Ready-to-use API (NLP2RDF)
- Hopefully improvement on Textmining and Ontology Learning tasks