The Semantic Gap of Formalized Meaning

Anwendung von Semantic Web Technologien in der Domäne der Linguisitik/Textmining

Sebastian Hellmann
Betreuer: Jens Lehmann, Mole, AKSW
Als Mitarbeiter am Lehrstuhl seit Januar 2009

UNIVERSITÄT LEIPZIG

Bereiche, Lücken, Trends

- Textmining/NLP
  - Explizitierung von Wissen
- Ontology Learning
  - fragmentierte Ansätze
- Strukturiertes Wissen
  - Möglichkeit des Entity Recognition und Disambiguierung
- Machine Learning
  - Enrichment und semi-automatisches erstellen von Wissen

- Ziel: ganzheitlicher Ansatz/Referenzmodell

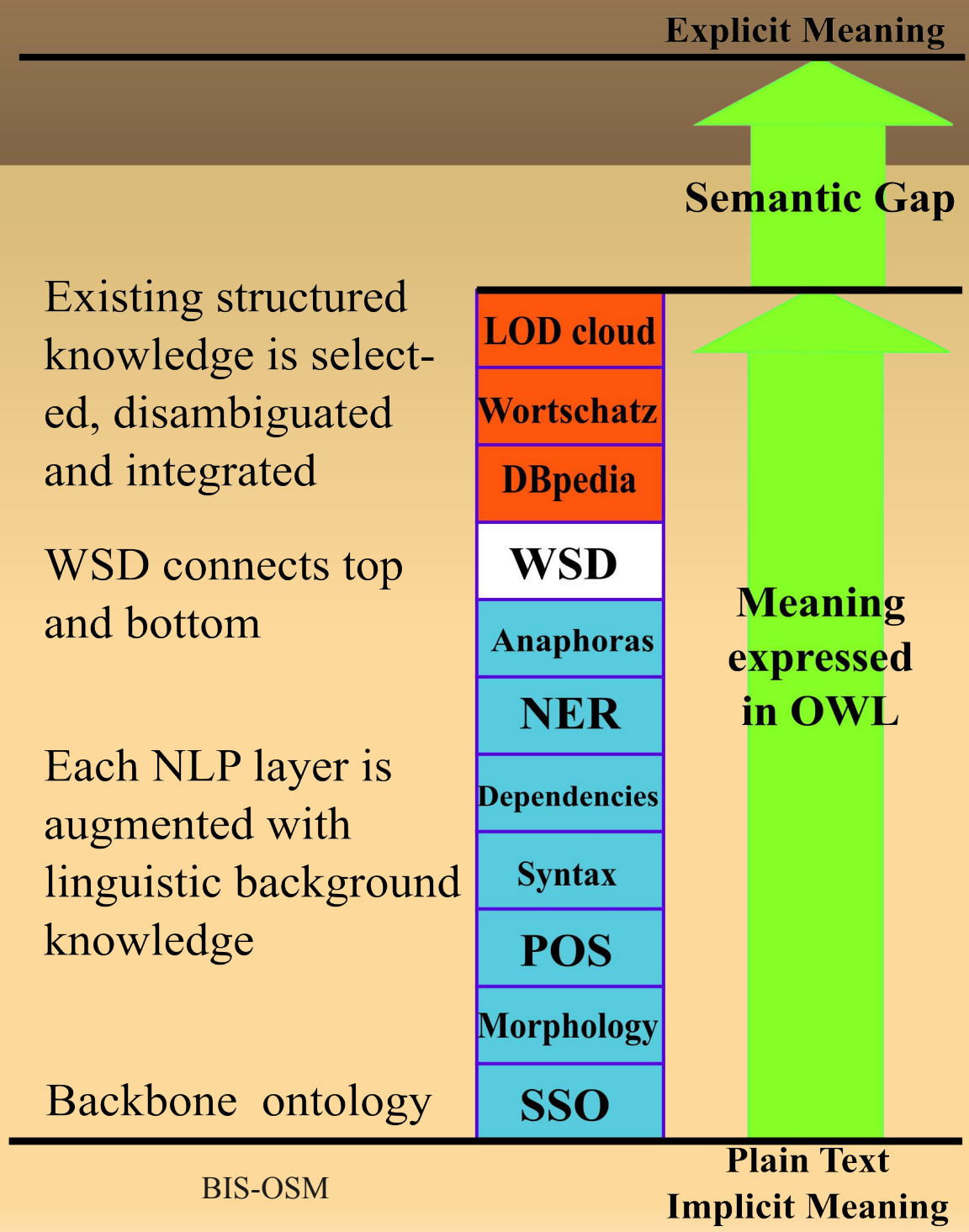
NLP2RDF stack

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



Forschungsfragen

- Definition des Semantic Gaps
- Wie kann man Fortschritt messen?
  - Benchmarking, Evaluationsframework
- Welche konkreten Aufgaben können verbessert werden?
- Ist OWL adäquat? Wo liegen die Grenzen als Meaning Representation Language?

Vorgehen – Bootstrap Phase

Beginn vor ca. 6 Monaten (Dezember 2009)

Abschluss: 4. Quartal 2010

- Recherche
- Feedback – z.b. ESWC PhD Symposium (Juni 2010)
- Prototypische Implementierung (Juli 2010)
- Initiale Publikationen (ISWC und Linguistische Konferenzen)
- Kollaborationsnetzwerk

Konkrete Ziele für die erste Phase

- Konvertierung von Wortschatz nach RDF
- Tiger Corpus Navigator

- Konvertierung von Wiktionary nach RDF
- 3-4 Komponenten des Stacks und Release
- Evaluation – 2-3 Benchmarks and Tasks
  - etablierte Benchmarks und Aufgaben (z.B. Textklassifikation, WSD)
- Integration mit LEXO (aus Mannheim)

TIGER Corpus Navigator

see **here** for data license

reset

Search

Fulltext Search search

Lemma Search werden search

Search Results

hide

Positive Matches

Er tritt in die GM-Verwaltung ein und wird Großaktionär des Autokonzerns .	+	-
Als vor gut einem Jahr der 71 Jahre alte Narasimha Rao den Premierminister-Sessel in Indien erklomm , da glaubte kaum jemand , daß er mehr würde als das , wozu ihn die Führer der Congress-Partei abstempeln wollten :	+	-
Immer wieder heißt es , nun würden die Japaner und die Deutschen massiv einsteigen , aber der große Durchbruch läßt auf sich warten .	+	-
Die mächtigen Gewerkschaften haben bisher verhindert , daß unrentable Firmen geschlossen oder privatisiert werden .	+	-
Die Getreideproduktion 1992 wird voraussichtlich zehn Millionen Tonnen geringer ausfallen als geplant , während gleichzeitig die Bevölkerung um 18 Millionen Menschen steigt .	+	-
Es wäre vergleichbar mit der heutigen Situation , wenn man sich mit höheren Chargen etwa der Stasi zusammensetzen würde .	+	-

Learned Concept

(Sentence and hasToken some VVPP)

Accuracy:1

Matching

These are past participles, e.g. "gegangen", "ar

Learning Input

learn save

Positive Samples

Wo früher nur für einen abgeschotteten Markt qualitatitiv minderwertige Ware hergestellt wurde , heißt die Losung nun Wettbewerb und Export .	x
Nun werden sie umworben .	x

Negative Samples

Und ein anderer Manager vermutet , daß sich `` ein Dogmatiker wie Perot in Washington schwer tun würde , es sei denn er schafft den Kongreß ab " .	x
Viele meinen , daß Perot mit seinem Befehlston auf dem Capitol gegen eine Wand laufen würde .	x

Wissenschaftliche Zusammenarbeit

- Korrektiv – Zusammenarbeit als Indikator
- Bessere und schnellere Resultate (¾ - ¼)
-

Phase 2

Iterativ:

- Implementierung
- Evaluierung
- Publikation
- Nach ersten Resultaten: Anwendung
  - Kollaboration mit Unternehmen

Langfristige Ziele

- Engineering – Implemtierung des Stacks
- Mehr Blasen zur LOD Cloud hinzufügen
- Wissensacquisition – semi automatische Unterstützung
- Natural Language Engineering – Welche features Methoden braucht man, um gewisse Aufgaben zu lösen?

  Beispiel: Ontology Learning

  - Ontologie aus Dokumenten (Termextraktion, WSD, Domänenwissen, Syntax?, Morphologie?)

Danke

Wissenschaftlicher Kern

- Algorithmen entwickeln, integrieren und evaluieren
- Begriffsbildung Semantic Gap
- (Software Engineering)
- If features extracted by different NLP approaches (ranging from low-level morphology analysis to higher-level anaphora resolution) are explicated and combined with matching background knowledge (parser-ontology pair) in a model and if, additionally, this model is further enriched by fragments of existing knowledge bases from external sources such as DBpedia, it will be possible to reduce the Semantic Gap and improve performance on common knowledge acquisition tasks such as Ontology learning and Text understanding.

Evaluierung