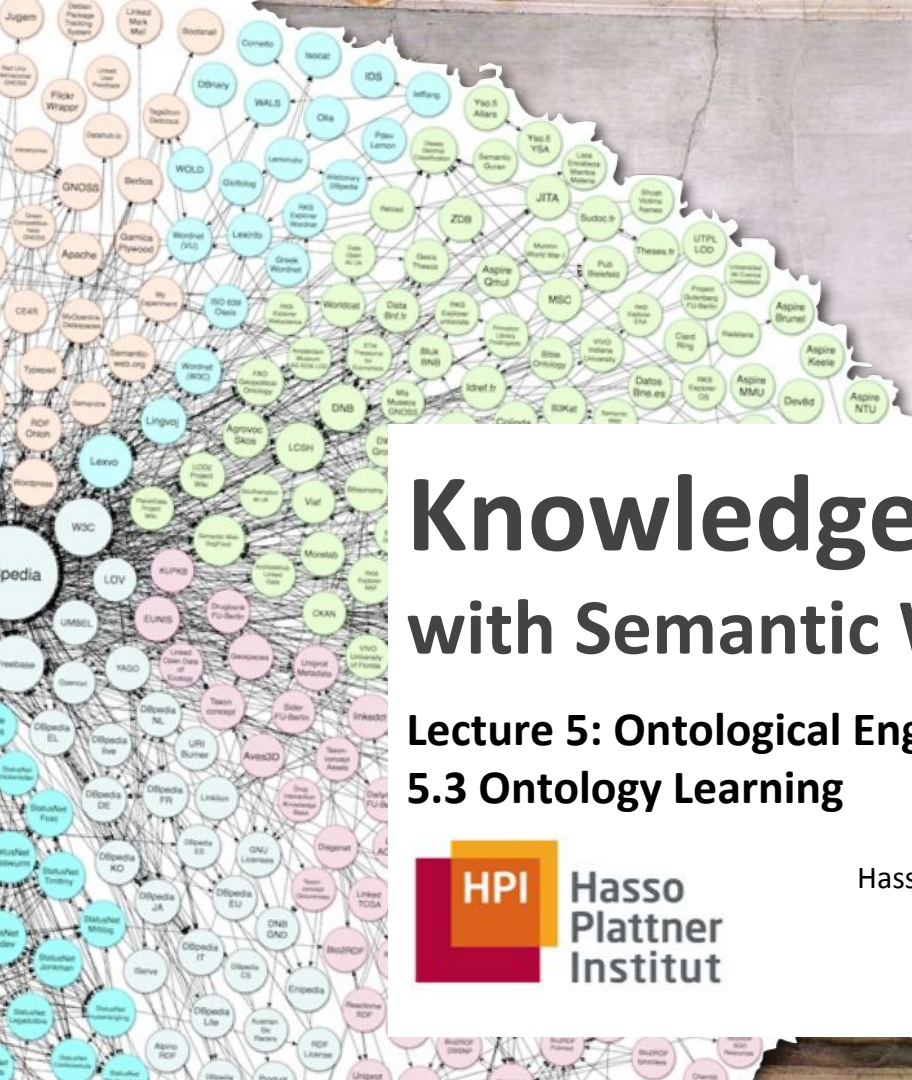


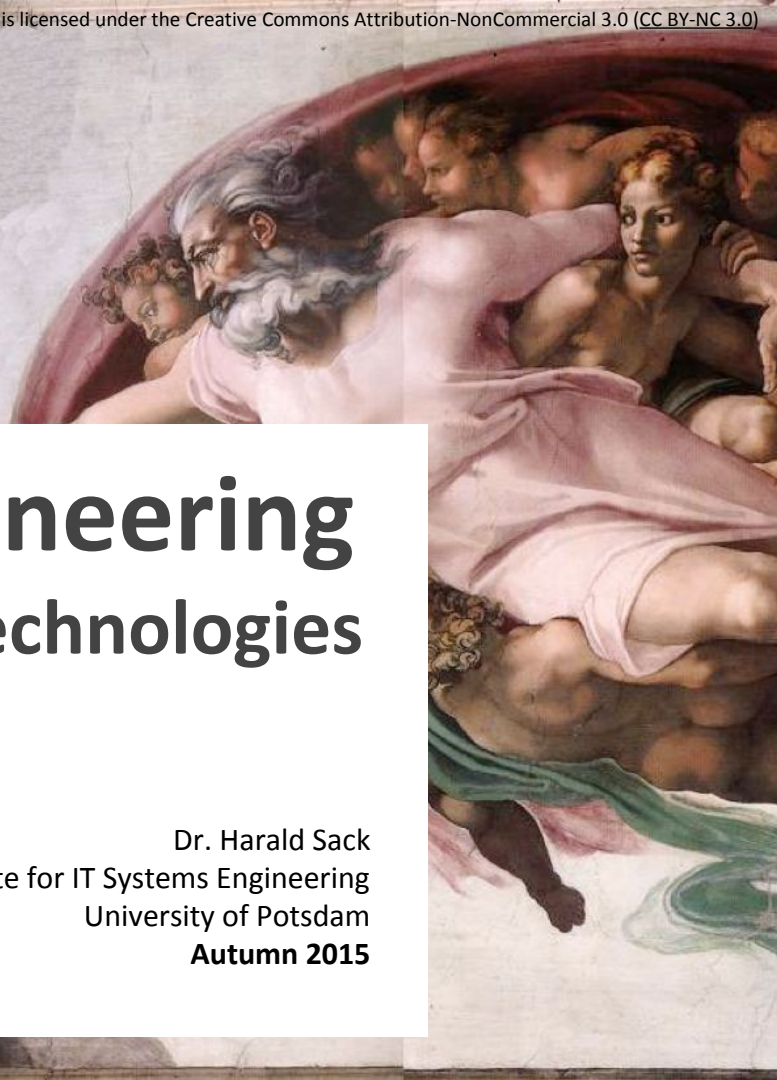


Knowledge Engineering with Semantic Web Technologies

Lecture 5: Ontological Engineering 5.3 Ontology Learning



Dr. Harald Sack
Hasso Plattner Institute for IT Systems Engineering
University of Potsdam
Autumn 2015



- Ontology Design is very expensive wrt. time and resources
 - *can we automate the process or at least some parts?*
- Ontologies can be „learned“ automatically
- **Ontology Learning** defines a set of methods and techniques
 - for **fundamental development** of new ontologies
 - for **extension or adaption** of already existing ontologies
- in a (partly) automated way from various resources.
- also referred to as *Ontology Generation, Ontology Mining, or Ontology Extraction*

- **Ontology Learning from Text**

- automatic or semi-automatic generation of lightweight ontologies by means of text mining and information extraction

- **Linked Data Mining**

- detecting meaningful patterns in RDF graphs via statistical schema induction or statistical relational learning

- **Concept Learning in Description Logics and OWL**

- learning schema axioms from existing ontologies and instance data mostly based on Inductive Logic Programming

- **Crowdsourcing Ontologies**

- combines the speed of computers with the accuracy of humans, as e.g. taxonomy construction via Amazon Turk or games with a purpose

- **Ontology Learning from text** is the process of identifying terms, concepts, relations, and optionally, axioms from textual information and using them to construct and maintain an ontology.
- Automatisations requires help from
 - Natural Language Processing (NLP)
 - Data Mining
 - Machine Learning techniques (ML)
 - Information Retrieval (IR)

Ontology Learning from Text - Basic Approach

document corpus



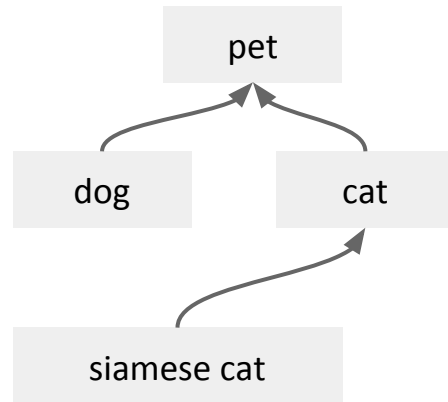
(1) term extraction

terminology

<dog> <dogs>
<cat>
<siamese cat>

(2) conceptualisation

ontology



(3) evaluation & adaption

term extractions requires linguistic processing (NLP) to identify important noun phrases and their internal semantic structure

terms: linguistic realisations of domain specific concepts

Concepts: clusters of semantically related terms

On November 3, 1954, the very first of a series of 28 Godzilla films premiered. The film focuses on Godzilla, a prehistoric monster resurrected by repeated nuclear tests in the Pacific, who ravages Japan and reignites the horrors of nuclear devastation to the very nation that experienced it first-hand. Since his debut, Godzilla has morphed into a worldwide cultural icon.

<http://blog.yovisto.com/godzilla/>

1. Tokenization

breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

Natural Language Processing for OL

Tokenize Text

Enter text

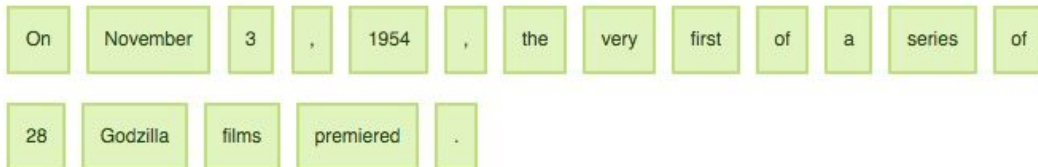
On November 3, 1954, the very first of a series of 28 Godzilla films premiered. The film focuses on Godzilla, a prehistoric monster resurrected by repeated nuclear tests in the Pacific, who ravages Japan and reignites the horrors of nuclear devastation to the very nation that experienced it first-hand. Since his debut, Godzilla has morphed into a worldwide cultural icon.

Enter up to 50000 characters

Tokenize

TreebankWordTokenizer

1.



<http://text-processing.com/demo/tokenize/>

On November 3, 1954, the very first of a series of 28 Godzilla films premiered. The film focuses on Godzilla, a prehistoric monster resurrected by repeated nuclear tests in the Pacific, who ravages Japan and reignites the horrors of nuclear devastation to the very nation that experienced it first-hand. Since his debut, Godzilla has morphed into a worldwide cultural icon.

1. Tokenization

breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

2. Morphological Analysis

analysis and description of the structure of a given language's **morphemes**, as e.g.
stemming and lemmatization

Natural Language Processing for OL

Stemming and Lemmatization with Python NLTK

This is a demonstration of **stemming** and **lemmatization** for the 17 languages supported by the **NLTK 2.0.4 stem** package.

Stem Text

Choose stemmer

English

Enter text

On November 3, 1954, the very first of a series of 28 Godzilla films premiered. The film focuses on Godzilla, a prehistoric monster resurrected by repeated nuclear tests in the Pacific, who ravages Japan and reignites the horrors of nuclear devastation to the very nation that experienced it first-hand. Since his debut, Godzilla has morphed into a worldwide cultural icon.

Enter up to 50000 characters

Stem

Stemmed Text

on novemb 3 , 1954 , the veri first of a seri of 28 godzilla film premier . the film focus on godzilla , a prehistor monster resurrect by repeat nuclear test in the pacif , who ravag japan and reignit the horror of nuclear devast to the veri nation that experienc it first - hand . sinc his debut , godzilla has morph into a worldwid cultur icon .

On November 3, 1954, the very first of a series of 28 Godzilla films premiered. The film focuses on Godzilla, a prehistoric monster resurrected by repeated nuclear tests in the Pacific, who ravages Japan and reignites the horrors of nuclear devastation to the very nation that experienced it first-hand. Since his debut, Godzilla has morphed into a worldwide cultural icon.

1. Tokenization

breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

2. Morphological Analysis

analysis and description of the structure of a given language's **morphemes**, as e.g. stemming and lemmatization

3. Part-of-Speech Tagging

marking up a word in a text (corpus) as corresponding to a particular part of speech

Natural Language Processing for OL

On November 3, 1954, the very first of a series of 28 Godzilla films premiered. The film focuses on Godzilla, a prehistoric monster resurrected by repeated nuclear tests in the Pacific, who ravages Japan and reignites the horrors of nuclear devastation to the very nation that experienced it first-hand. Since his debut, Godzilla has morphed into a worldwide cultural icon.

Submit

The Part-of-Speech tagger has automatically labeled the input in the following way.

IN/ On NNP/ November CD/ 3 , CD/ 1954 , DT/ the RB/ very JJ/ first IN/ of DT/ a NN/ series IN/ of CD/ 28 NN/ Godzilla NNS/ films VBD/ premiered . DT/ The NN/ film VBZ/ focuses IN/ on NN/ Godzilla , DT/ a JJ/ prehistoric NN/ monster VBD/ resurrected IN/ by VBN/ repeated JJ/ nuclear NNS/ tests IN/ in DT/ the NNP/ Pacific , WP/ who NNS/ ravages NNP/ Japan CC/ and VBZ/ reignites DT/ the NNS/ horrors IN/ of JJ/ nuclear NN/ devastation TO/ to DT/ the JJ/ very NN/ nation IN/ that VBN/ experienced PRP/ it JJ/ first-hand . IN/ Since PRPS/ his NN/ debut , NN/ Godzilla VBZ/ has VBN/ morphed IN/ into DT/ a JJ/ worldwide JJ/ cultural NN/ icon .

Natural Language Processing for OL

1. Tokenization

breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

2. Morphological Analysis

analysis and description of the structure of a given language's **morphemes**, as e.g. stemming and lemmatization

3. Part-of-Speech Tagging

marking up a word in a text (corpus) as corresponding to a particular part of speech

4. Regular Expression Matching

define regular expressions and match them in text

5. chunker / n-gram analysis

detect larger coherent structures

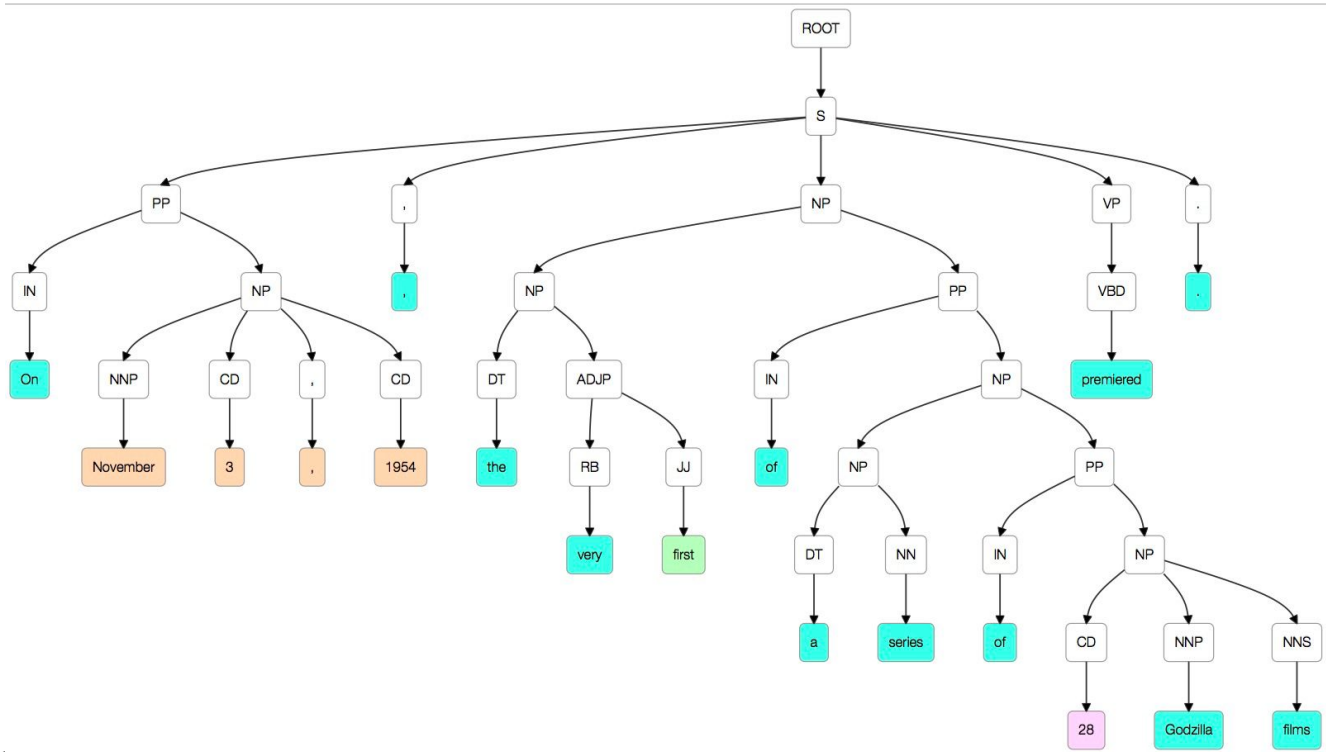
6. Syntax Tree Parsing

determine the full syntactic structure of a sentence

Natural Language Processing for OL

On November 3, 1954, the very first of a series of 28 Godzilla films premiered.

Person Date Organization Location Ordinal Number



Machine Learning for OL

1. Association Rule Discovery

- discovery of interesting associations between terms

2. (Hierarchical) Clustering

- unsupervised learning, in particular clustering of terms

3. Classification

- classification of new concepts into an existing hierarchy
- e.g. with Support Vector Machines (SVM), Naive Bayes, knn

4. Inductive Logic Programming

- induction of rules from data, i.e. discovery of new concepts from extensional data

5. Conceptual Clustering

- Formal Concept Analysis, learning concepts and concept hierarchies

The Ontology Learning Layer Cake

Country $\sqsubseteq \leq 1$ hasCapital. $\top$

General Axioms

River $\sqcap$ Mountain $\sqsubseteq \perp$

Axiomatic Schemata

capitalOf $\sqsubseteq$ locatedIn

Relation Hierarchies

flowThrough(dom:River, range:GeoEntity)

Relations

Capital $\sqsubseteq$ City , City $\sqsubseteq$ InhabitedGeoEntity

Concept Hierarchies

c:=country:=<description(c), uri(c)>

Concept Description

{country, nation, land}

Multilingual Synonyms

river, country, nation, city, capital, ...

Terms

1. Ontology Creation

- Design of an ontology from the scratch by a team of experts
- *Maschine Learning* (ML) supports the experts during the design phase by
 - Suggestions of well suited relations among concepts
 - Integrity / consistency checking of the designed ontology

2. Ontology Schema Extraction

- Extraction of ontology schemata from heterogeneous documents
- *Machine Learning* uses input data and meta ontology to create full-fledged domain ontologies (with the help of human experts)

3. Extraction of Ontology Instances (Populate Ontologies)

- Extraction of ontology instances from semi-structured / unstructured data to populate already existing ontology schemata with individuals
- applies technologies from *Information Retrieval* and *Data Mining*

4. Ontology Integration and Navigation

- Reconstruction of existing knowledge bases and navigation in existing knowledge bases,
- as e.g. translation of an existing knowledge base from FOL to OWL 2 DL

5. Ontology Update

- Extension, reconstruction and adaption of already existing ontologies, as e.g. adation of an ontology to a changed domain
- relates to parts of ontologies that have been created in the way that they can be changed

6. Ontology Enrichment

- (also Ontology tuning) relates to automated update of smaller parts of existing ontologies
- doesn't changes concepts and relations, but refines them (more precise)
- in difference to ontology update only parts of the ontology are considered that usually shouldn't be changed

Challenges in Ontology Learning

- **Heterogeneity**
 - data on the web differs largely, , e.g., wrt. formats, languages, domains and quality
- **Uncertainty**
 - low-quality data, which is hard to interpret by computational means, as well as inherently imperfect methods for learning ontologies
- **Reasoning**
 - applications relying on reasoning need consistent ontologies, must be explicitly supported by ontology learning
- **Scalability**
 - knowledge extraction from growing amounts of web data requires scalable ontology learning
- **Quality**
 - ontology evaluation to enable formal correctness, completeness and consistency
- **Interactivity**
 - human involvement increases the quality of learned ontologies



04: Ontology Alignment

OpenHPI - Course Knowledge Engineering with Semantic Web Technologies

Lecture 5: Ontological Engineering