Ontology Learning

Ontologies and Ontology Engineering

Kristian Stavåker and Dag Sonntag

VT 2011

Summary

- Wikipedia:
  - ***“Ontology learning (ontology extraction, ontology generation, or ontology acquisition) is a subtask of information extraction. The goal of ontology learning is to (semi-)automatically extract relevant concepts and relations from a given corpus or other kinds of data sets to form an ontology.”***
- Paul Buitelaar et al. / Ontology Learning from Text: An Overview [3]:
  - ***”The process of defining and instantiating a knowledge base is referred to as knowledge markup or ontology population, whereas (semi-)automatic support in ontology development is usually referred to as ontology learning.”***

Summary (2)

- A lot of the work in this area builds on work from natural language processing, artificial intelligence and machine learning.
- The *ontology layer cake* consists of the various subtasks (increasing in complexity) involved in ontology learning.

The Ontology Learning Layer Cake

$\forall x, y\ (sufferFrom(x, y) \rightarrow ill(x))$	Axioms & Rules
cure (domain:Doctor, range:Disease)	Relations
is_a (Doctor, Person)	Taxonomy (Concept hierarchies)
Disease:=<I, E, L>	Concepts
{disease, illness}	Synonyms
disease, illness, hospital	Terms

LiU EXPANDING REALITY

Outline

- Background
- Terms
- Synonyms
- Concepts
- Concept Hierarchies
- Relations
- Rules
- Conclusions

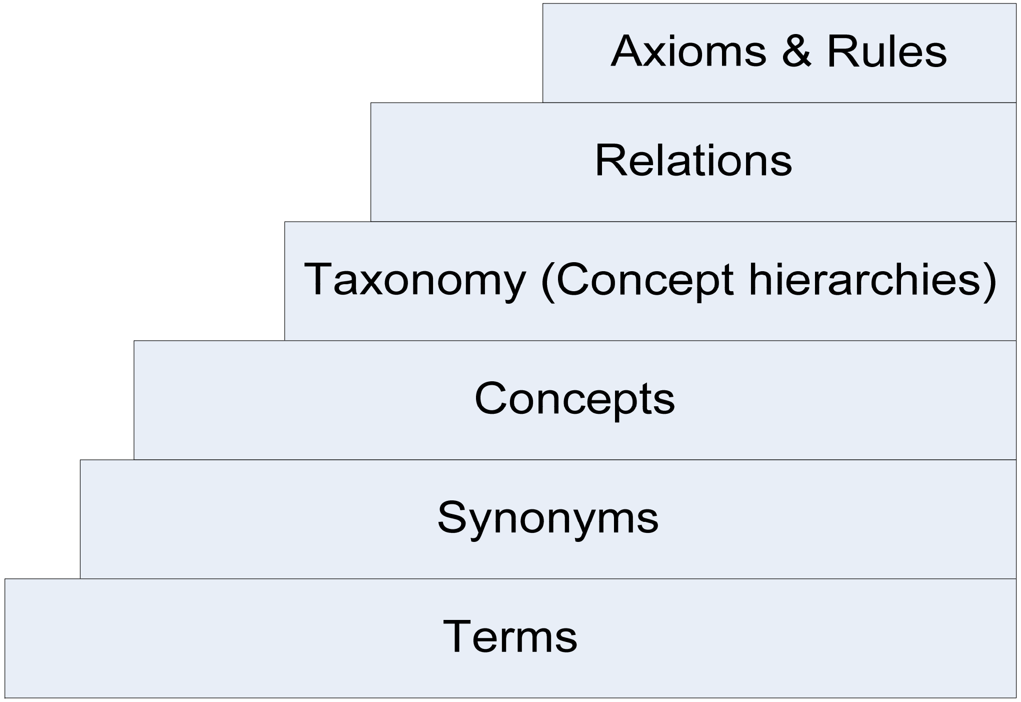
Background

- Used for building an ontology from scratch through the application of a set of methods and techniques.
- The process of identifying terms, concepts, relations and optionally axioms from textual information to form an ontology.

Background (2)

- Unstructured sources (NLP techniques, morphological and syntactic analysis, etc.)
- Semi-structured sources (such as XML schema)
- Structured data (extract concepts and relations from for instance databases)

Terms



disease, illness, hospital

LiU EXPANDING REALITY

Terms

- Term extraction is a prerequisite for all aspects of ontology learning from text.
- Term extraction implies more or less advanced levels of liguistic processing.
- An example of extracting relevant terms is counting frequencies of terms in a given set of documents (the corpus).

Terms

- The computational linguistics community has proposed a wide range of more sophisticated techniques for term extraction.

Terms

- One common method:
  - A Part-Of-Speech (POS) tagger is run over the domain corpus
  - Possible terms are identified by constructing patterns, such as: Adj-Noun, Noun-noun, Adj-Noun-Noun, … (names are ignored)
  - Apply statistical metrics in order to identify only the relevant to the text terms

Terms

[[He SUBJ] [booked PRED] [[this] [table HEAD]NP:DOBJ:X1]…]…	Discourse Analysis
[[It SUBJ:X1] [was PRED] still available…] *[[He SUBJ] [booked PRED] [[this] [table HEAD] NP:DOBJ]S]*	Dependency Structure (S)
[[the SPEC] [large MOD] [table HEAD] NP]	Dependency Structure (Phrases)
[[the] [large] [table] NP] [[in] [the] [corner] PP]	Phrase Recognition
[work~ing V]	Morphological Analysis (stemming)
[table N:ARTIFACT] [table N:furniture]	Part of Speech & Semantic Tagging
[table] [2005-06-01] [John Smith]	Tokenization (incl. Named-Entity Rec.)

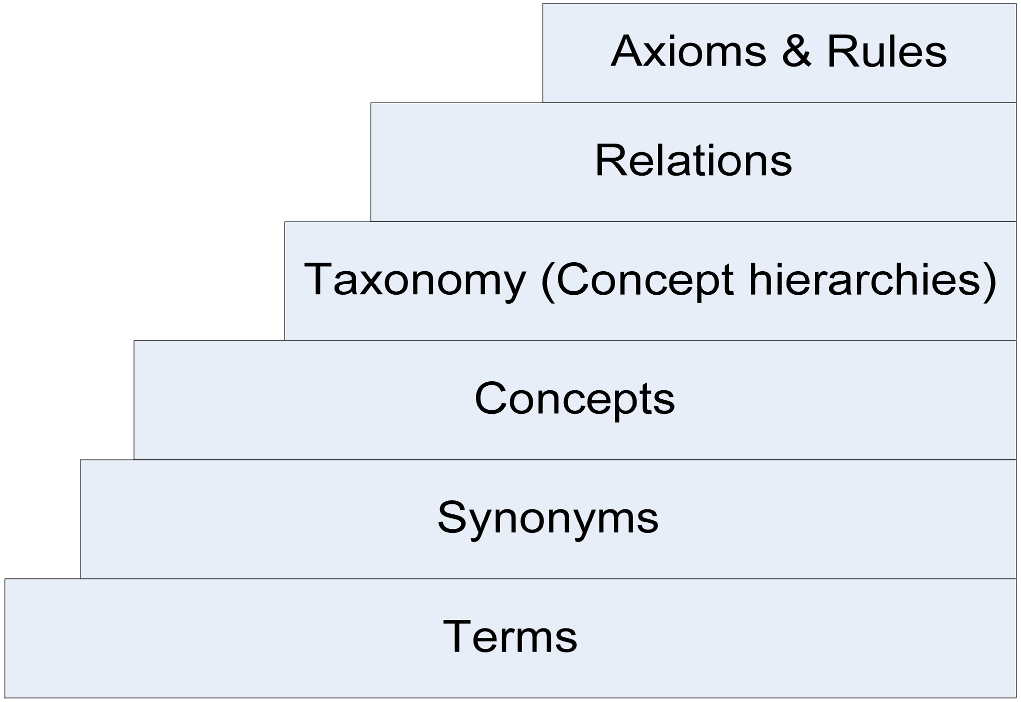
Terms

- Statistical Analysis
- Term Frequency Inverted Document Frequency (TFIDF) - a popular weighting scheme

$$tfidf(w) = tf(w) \bullet \log\left(\frac{N}{df(w)}\right)$$

LiU EXPANDING REALITY

Synonyms



{disease, illness}

LiU EXPANDING REALITY

Synonyms

- Identification of terms that share semantics (potentially refer to the same concept)
- Methods for extracting synonyms
  - Based on WordNet/EuroWordNet
  - Harris' distributional hypothesis
  - Latent Semantic Indexing (LSI)
    - A NLP technique of analyzing relationships between a set of documents and the terms they contain

Synonyms

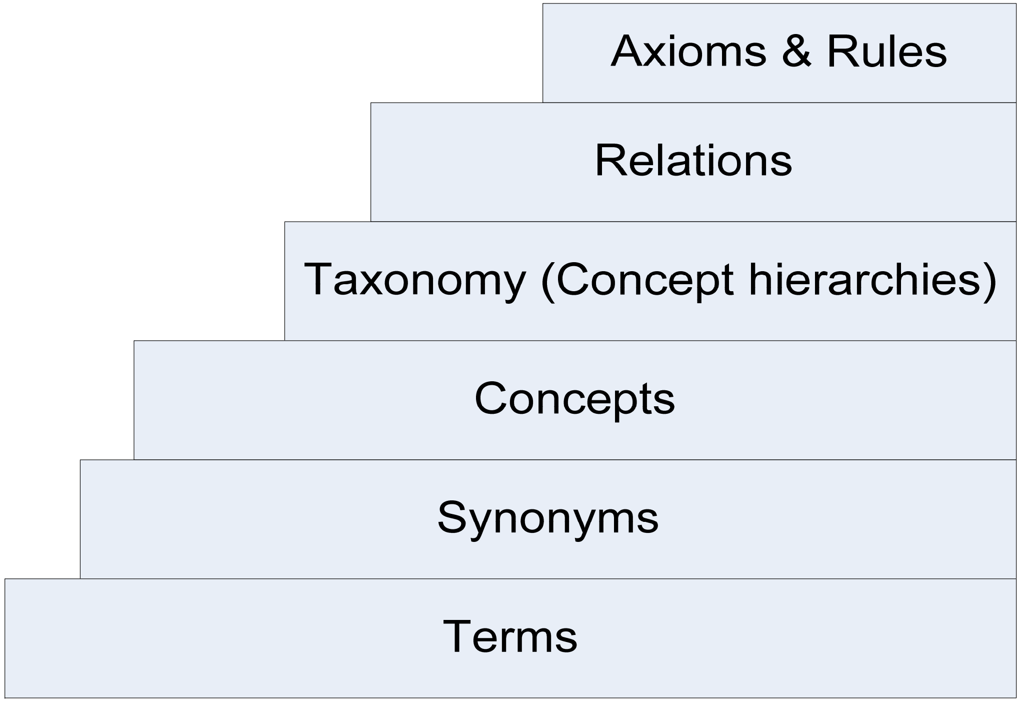
- Pointwise Mutual Information measure for extracting synonyms.

$$PMI(x,y) := \log_2 \frac{P(x,y)}{P(x)P(y)}$$

$$PMI_{Web}(x,y) := \log_2 \frac{Hits(xANDy)MaxPages}{Hits(x)Hits(y)}$$

Concepts

Disease:=<I, E, L>



Concepts

- Some confusion what extraction of concepts is since it is not clear what exactly constitutes a concept.

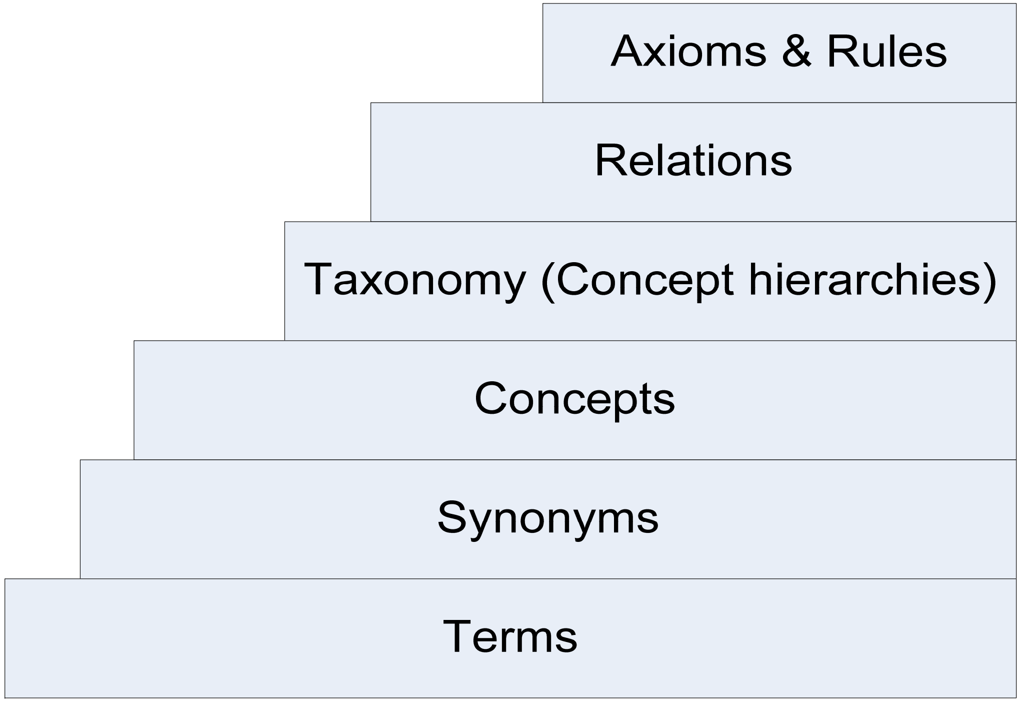
Concepts

- Intension – (in)formal definition of the set of objects that this concept describes
  - Example: sound is a mechanical wave that is an oscillation of pressure composed of frequencies within the range of hearing
- Extension – a set of objects that the definition of this concept describes
  - Example: music, noise, speech
- Lexical realizations – the term itself and its multilingual synonyms
  - Example: sound, acoustics

- Background
- Terms
- Synonyms
- Concepts
- Concept Hierarchies
- Relations
- Rules
- Conclusions

Taxonomy

is_a (Doctor, Person)



Taxonomy

- Simple ideas that work fairly well
  - Lexico-Synthatic Patterns (Hearst)
  - Formal Concept Analysis (FCA)
  - Phrase Analysis
  - WordNet
- These methods can be weightened together to get best results

Lexico-Synthatic Patterns (Hearst)

- Hearst identified the following patterns
  - Hearst1: NPhyper such as {NPhypo,}* {(and | or)} NPhypo
  - Hearst2: Such NPhyper as {NPhypo,}* {(and | or)} NPhypo
  - Hearst3: NPhypo {,NP}* {,} or other NPhyper
  - Hearst4: NPhypo {,NP}* {,} and other NPhyper
  - Hearst5: NPhyper including {NPhypo,}* NPhypo {(and | or)} NPhypo
  - Hearst6: NPhyper especially {NPhypo,}* {(and | or)} NPhypo
- Example: ”*Vehicles such as bikes and cars*”

Machine Readable Dictionaries

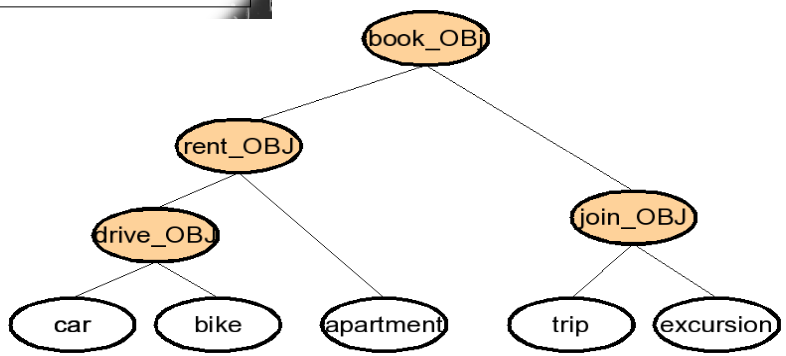
- Idea: Exploit the regularity of dictionaries like Wikepedia or Google definition
- Example: (from wikipedia)
  - Car: ”An automobile, autocar, motor car or car is a wheeled motor vehicle used for transporting passengers, which also carries its own engine or motor. “ => is(car, vehicle)

Background
Terms
Synonyms
Concepts
Concept Hierarchies
Relations
Rules
Conclusions

Formal Concept Analysis (FCA)

- Idea: Similar words share similiar attributes
- Example:

	hotel	apartment	car	bike	excursion	trip
hotel	1	0.5	0.33	0.25	0.5	0.5
apartment		1	0.67	0.5	0.33	0.33
car			1	0.75	0.25	0.25
bike				1	0.2	0.2
excursion					1	1
trip						1



Phrase Analysis

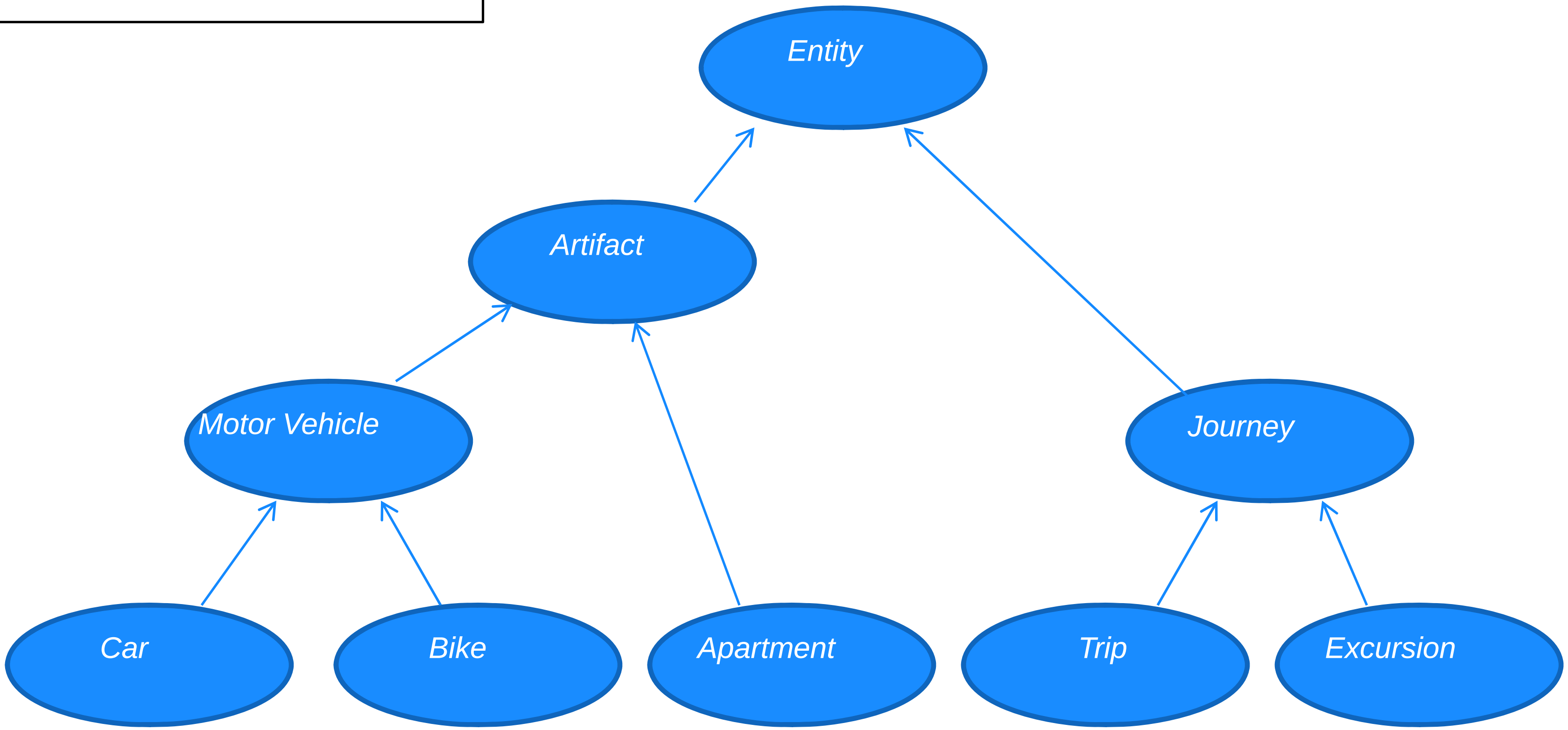
- Idea: Adjectives or Nominals in front of nouns in noun-phrases often indicate subclass
- Example: *Focal epilepsy* is a subclass of *epilepsy*

Wordnet

- Idea: Use already created ontologies
- WordNet contains ≈155000 words
  - Type (noun, verb and so on)
  - Their definition
  - Semantic synsets like
    - Hypernyms, hyponyms, meronyms, synonyms…

Background
Terms
Synonyms
Concepts
Concept Hierarchies
Relations
Rules
Conclusions

Wordnet

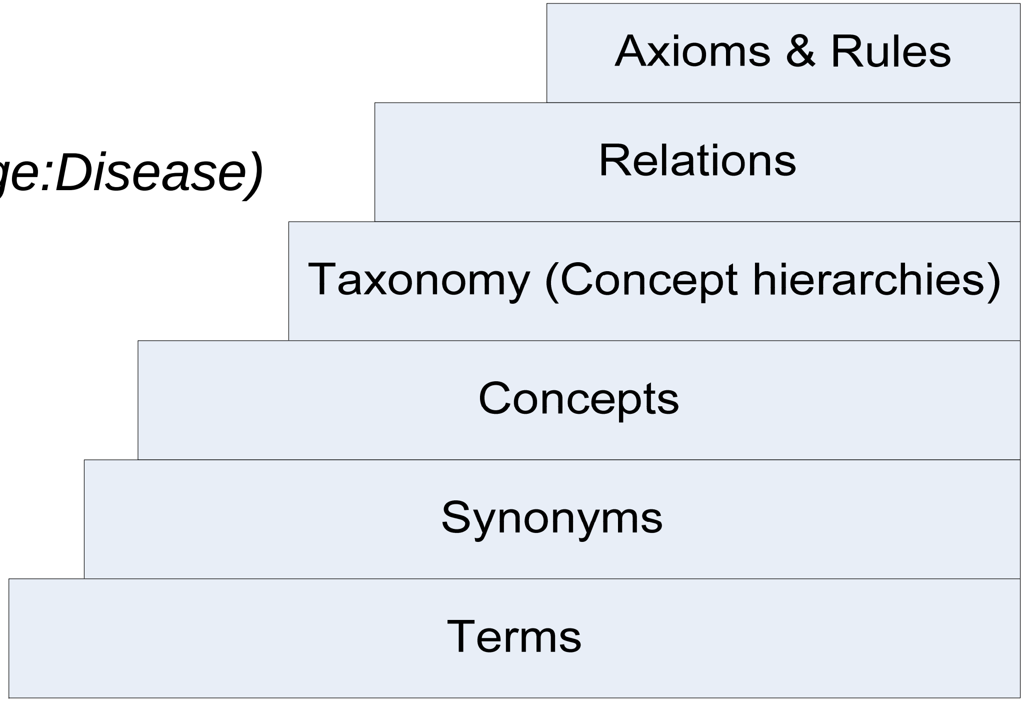


LiU EXPANDING REALITY

- Background
- Terms
- Synonyms
- Concepts
- Concept Hierarchies
- Relations
- Rules
- Conclusions

Relations

cure (domain:Doctor, range:Disease)



Relations

- Can be found similar as concept hierarchies (is relation)
  - Lexico-Synthatic Patterns
  - Collocation Discovery
  - WordNet
- Specific relations
  - Part of (meronym)
- Attributes
  - Adjectives, e.g. color, weight and so on…

Lexico-Synthatic Patterns

- Finds relations by searching after patterns of words
- E.g. The car has wheels =>, part-of(wheel, car)
  The car is red => is(car, red) or color(car,red)
  The car consists of wheels, engine, …
  => part-of(engine, car)…
- Hard to model every possible combination

Collocation Discovery

- Idea: find words that occur together in a statistically significant manner
- Similar to the PCA-approach for the taxonomy
- The type of relation can then later be found by linguistic approaches
- Example: www-search for two concepts K1 and K2 using the Jaccard coefficient:

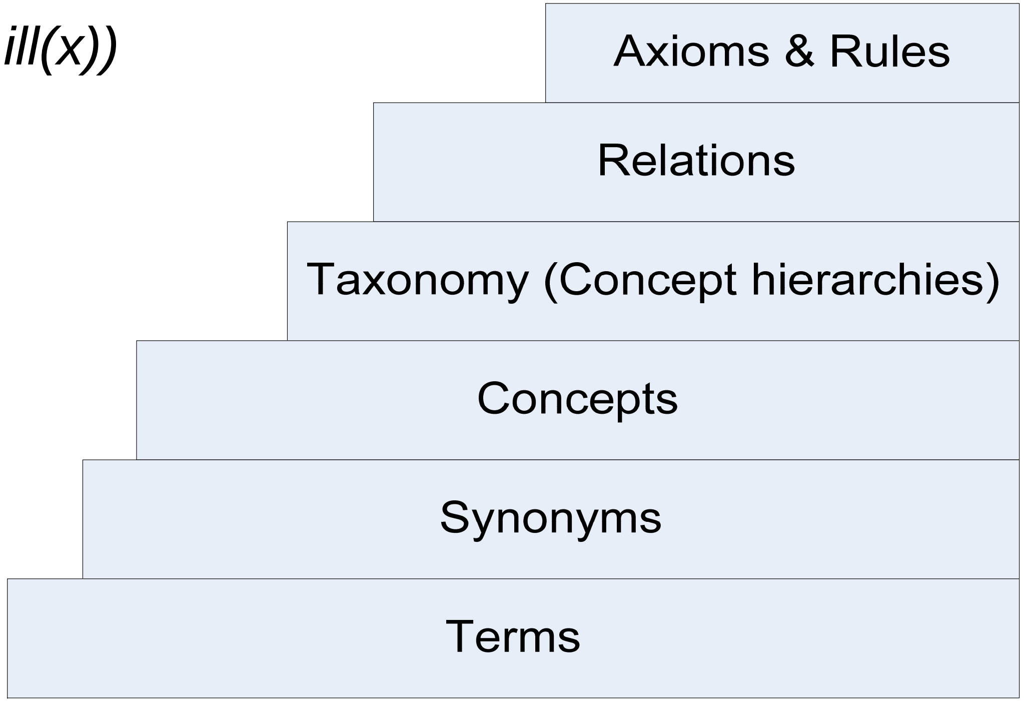
$$\frac{\text{GoogleHits(Keyword1,Keyword2)}}{\text{GoogleHits(K1)} + \text{GoogleHits(K2)} - \text{GoogleHits(K1,K2)}}$$

WordNet

- Idea: Use already created ontologies
- Previously found relations can be used to train other machine learning techniques (e.g. decision trees, association rule learning, neural networks)

Axioms & Rules

$$\forall x, y\ (sufferFrom(x, y) \rightarrow ill(x))$$



Axioms & Rules

- Mostly Lexico-Synthatic Patterns
- Example: LExO

Rule	Natural Language Syntax	OWL Axioms
Disjunction	NP_0 or NP_1	$X \equiv (NP_0 \sqcup NP_1)$
Conjunction	NP_0 and NP_1	$X \equiv (NP_0 \sqcap NP_1)$
Determiner	Det_0 NP_0	$X \equiv NP_0$
Intersective Adjective	Adj_0 NP_0	$X \equiv (Adj_0 \sqcap NP_0)$
Subsective Adjective	Adj_0 NP_0	$X \sqsubseteq NP_0$
Privative Adjective	Adj_0 NP_0	$X \sqsubseteq \neg NP_0$
Transitive Verb Phrase	V_0 $NP(obj)_0$	$X \equiv \exists V_0.NP_0$
Verb with Prep. Compl.	V_0 $Prep_0$ $NP(pcomp\text{-}n)_0$	$X \equiv \exists V_0\_Prep_0.NP_0$
Noun with Prep. Compl.	NP_0 $Prep_0$ $NP(pcomp\text{-}n)_1$	$X \equiv (NP_0 \sqcap \exists Prep_0.NP_1)$
Prepositional Phrase	$Prep_0$ NP_0	$X \equiv \exists Prep_0.NP_0$

Data: Facts that result from measurements or observations.

$$\text{Data} \equiv (\text{Fact} \sqcap \exists \text{result\_from}.(\text{Measurement} \sqcup \text{Observation}))$$

LiU EXPANDING REALITY

Background
Terms
Synonyms
Concepts
Concept Hierarchies
Relations
Rules
Conclusions

Conclusions

$\forall x, y\ (sufferFrom(x, y) \rightarrow ill(x))$	Axioms & Rules
cure (domain:Doctor, range:Disease)	Relations
is_a (Doctor, Person)	Taxonomy (Concept hierarchies)
Disease:=<I, E, L>	Concepts
{disease, illness}	Synonyms
disease, illness, hospital	Terms

Conclusions

- Mainly three methods used:
  - Natural language processing (NLP)
  - Statistical methods
  - Using other existing ontology

Natural Language Processing

- Positive:
  - Easy to find corpus
  - High success-rate for certain patterns
- Negative:
  - Hard to model every kind of pattern
  - Ambiguity in natural text

Statistical Methods

- Positive:
  - Needed in some parts of the layer cake, like term extraction
  - Can give confidence to relations (or synonyms) found by other methods
- Negative:
  - Even if a statistical similarity if found, the relation binding the words together is unknown

Use Other Existing Ontology

- Positive:
  - Very easy to find relations and definitions
  - If a relation or similar is found the probability-rate of it being correct is very high
  - Can be combined in a successful way with other methods
- Negative:
  - Can be hard to find ontologies in specialized areas
  - Words can have multiple meaning

LiU EXPANDING REALITY

References

- [1] Ontology Learning – Philipp Cimiano, Alexander Mädche, Steffen Staab, Johanna Völker, 2009.
- [2] Ontology Learning – Alexander Maedche, Steffen Staab.
- [3] Ontology Learning from Text: An Overview – Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, 2003.
- [4] Ontology Learning from Text: A Look back and into the Future – Wilson Wong, Wei Liu, Mohammed Bennamoun, 2011.
- [5] Aquisition of OWL DL Axioms from Lexical Resources – Johanna Völker, Pascal Hitzler and Philipp Cimiano

Linköping University
expanding reality

www.liu.se