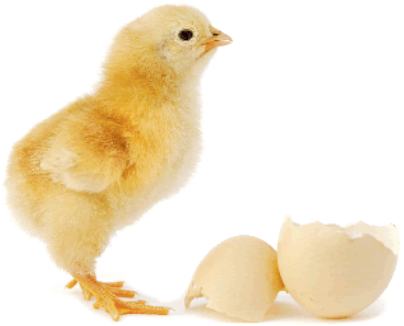
UNIVERSITÄT MANNHEIM

Ontology Learning and Information Extraction

Dr. Diana Maynard
d.maynard@dcs.shef.ac.uk
Dr. Johanna Völker
voelker@informatik.uni-mannheim.de

LREC 2010

Overview



Ontology Learning

- "*Ontology Learning 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.*"*
- "*Ontology Learning is a mechanism for semi-automatically supporting the ontology engineer in engineering ontologies.*"**
- "*Ontology Learning aims at the integration of a multitude of disciplines in order to facilitate the construction of ontologies, in particular ontology engineering and machine learning.*"***

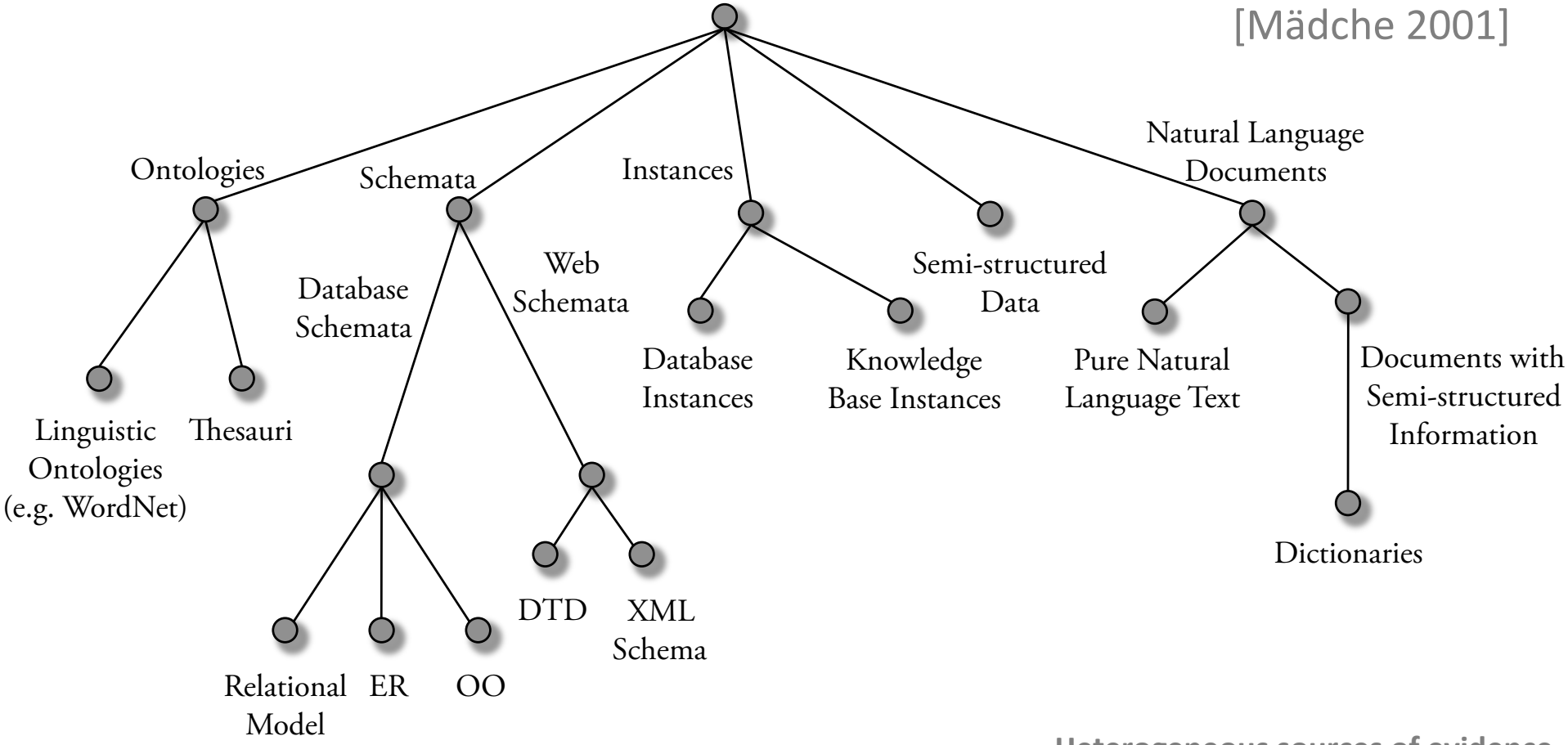
* Wikipedia 2008/12/15: *http://en.wikipedia.org/wiki/Ontology_learning*

** A. D. Mädche. *Ontology Learning for the Semantic Web.* Dissertation. Universität Karlsruhe, 2001

*** A. D. Mädche, S. Staab. *Ontology Learning*. Handbook of Ontologies in Information Systems, 2004

Classification of Ontology Learning Data

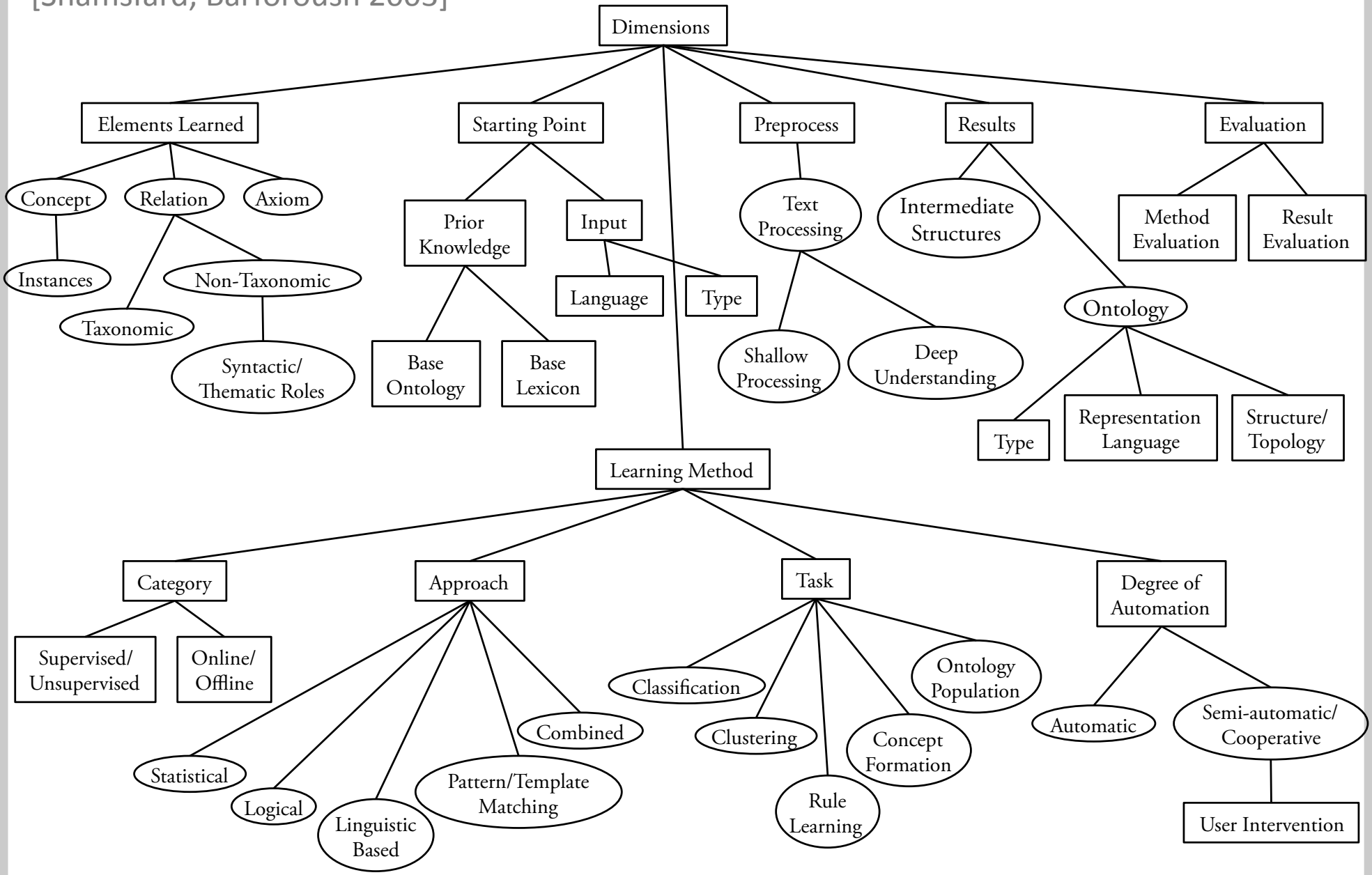
[Mädche 2001]



Heterogeneous sources of evidence

(e.g., hyponymy [Snow et al. 2006], subsumption [Cimiano et al. 2005], [Manzano-Macho et al. 2008], [Buitelaar et al. 2008], disjointness [Völker et al. 2007])

[Shamsfard, Barforoush 2003]



Overview

1 Motivation and Definition

2 Ontology Learning from Text

- 2.1 Approaches
- 2.2 Problems and Challenges

3 Reasoning for Learning and Integration

2 Ontology Learning from Text

$\forall x$ (country(x) → $\exists y$ capital_of(y,x) $\wedge \forall z$ (capital_of(z,x) → y=z))

disjoint(river, mountain)

capital_of $\leq_R$ located_in

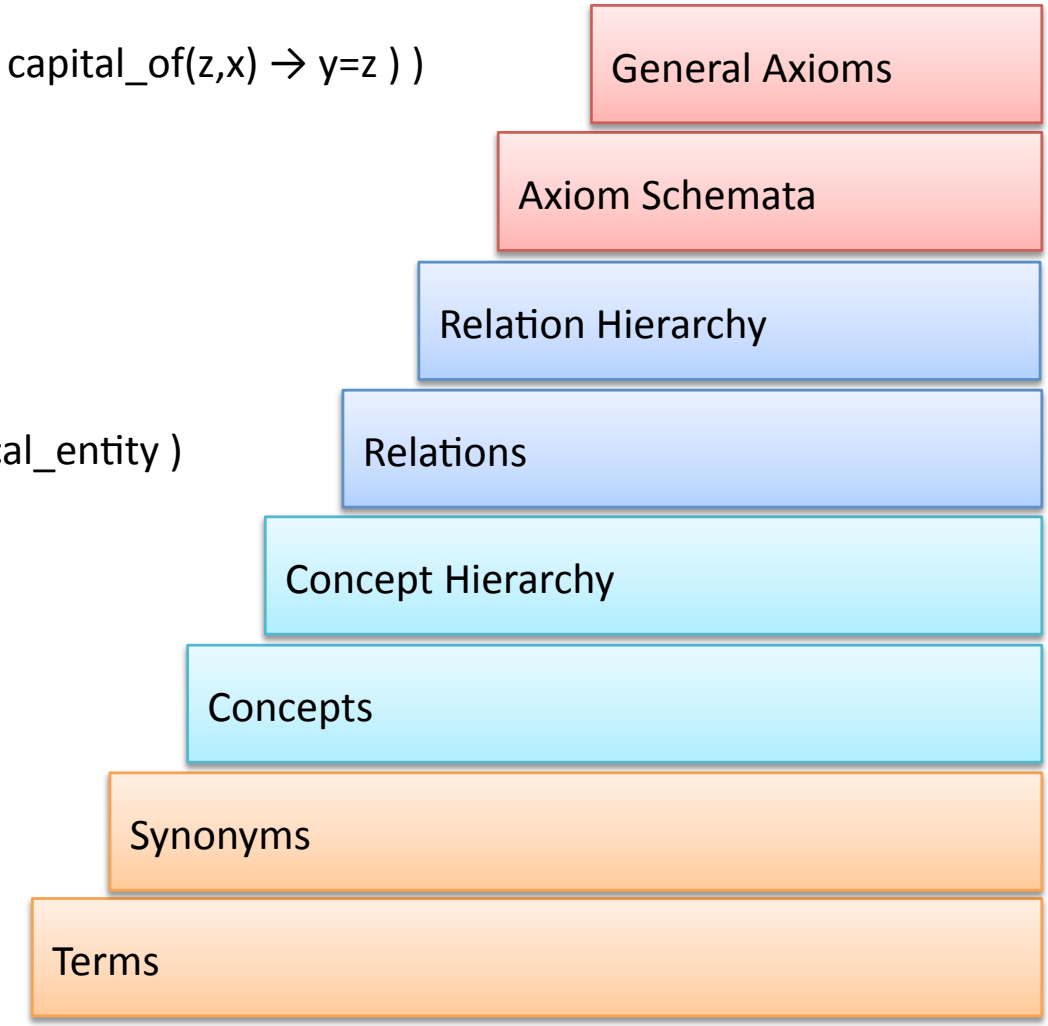
flow_through(domain:river, range:geopolitical_entity)

capital $\leq_c$ city, city $\leq_c$ geopolitical_entity

c := country := < i(c), ||c||, $Ref_c(c)$ >

{country, nation}

river, country, nation, city, capital …



2.1 Approaches

- Classification [Zhou 2007]
  - Learning units: word, term (single or multi-word units)
  - Learning targets: concept, relation, definition, axiom
  - Learning strategies: statistics-based (e.g. clustering), rule-based (e.g. ILP), hybrid (e.g. patterns)
  - Knowledge support: knowledge-rich (e.g. ontologies, WordNet), knowledge-lean (e.g. co-occurrences)
  - Data sources: document collection, web, dictionary, user interaction (e.g. games with a purpose)

Patterns [Hearst 1992]

- such NP as {NP,}* {or|and} NP
  - „such **games** as **baseball** and **cricket**“
- NP {,NP}* {,} {and|or} other NP
  - „**rabbits** and other **animals**“
  - but: „**rabbits** and other **pets**“
- NP {,} including {NP,}* {or|and} NP
  - „**fruits** including **apples** and **pears**“
- NP {,} especially {NP,}* {or|and} NP
  - „**Europeans**, especially **Italians**“
  - but: „**US presidents**, especially **democrats**“

Patterns [Ogata and Collier 2004]

- NP is a NP
  - „A **kangaroo** is an **animal** living in Australia.“
- a NP {named|called} NP
  - „Japanese people like to play a **game** called **Go**.“
- NP, NP
  - „**Sencha**, the most popular **tea** in Japan, ...“
- NP. The NP
  - „John loves his **Ferrari**. The **car** ...“
- NP and other NP
  - „**universities** and other **institutions**“
- NP such as NP
  - „**sports** such as **tennis**“
- Among NP, NP
  - „Among all **musical instruments**, **violins** are ...“
- NP {except for|other than} NP
  - „**Employees** except for **managers** suffer from ...“

Example: JAPE Rule

```
rule: Hearst_1
(
   (NounPhrase):superconcept
   {SpaceToken.kind == space}
   {Token.string=="such"}
   {SpaceToken.kind == space}
   {Token.string=="as"}
   {SpaceToken.kind == space}
   (NounPhrase):subconcept
):hearst1
-->
:hearst1.SubclassOfRelation = { rule = "Hearst1" },
:subconcept.Domain = { rule = "Hearst1" },
:superconcept.Range = { rule = "Hearst1" }
```

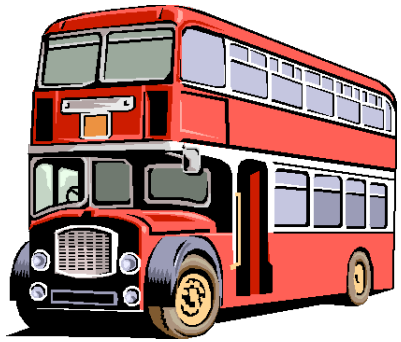
Lexical Context Similarity

- „**Columbus** is the **capital** of the **state** of **Ohio**. **Columbus** has a **population** of about 700,000 **inhabitants**.“
  - Columbus (capital (1), state (1), Ohio (1), population (1), inhabitant (1))
- City (country (2), state (1), inhabitant (2), mayor (1), attraction (1))
- Explorer (ship (1), sailor (2), discovery (1))

„most probably“: City(Columbus)

see, for example [Cimiano and Völker 2005]

Subcategorization Frames



- `Tina drives a Ford.`
  - Person(Tina). Vehicle(Ford).
- `Her father drives a bus.`
  - Father subclass-of Person
  - Bus subclass-of Vehicle
- subcat: `drive`(subj: `person`, obj: `vehicle`)

$$Person \sqsubseteq \forall drive.Vehicle$$

e.g. [Faure and Nédellec 1998], [Schutz and Buitelaar 2005]

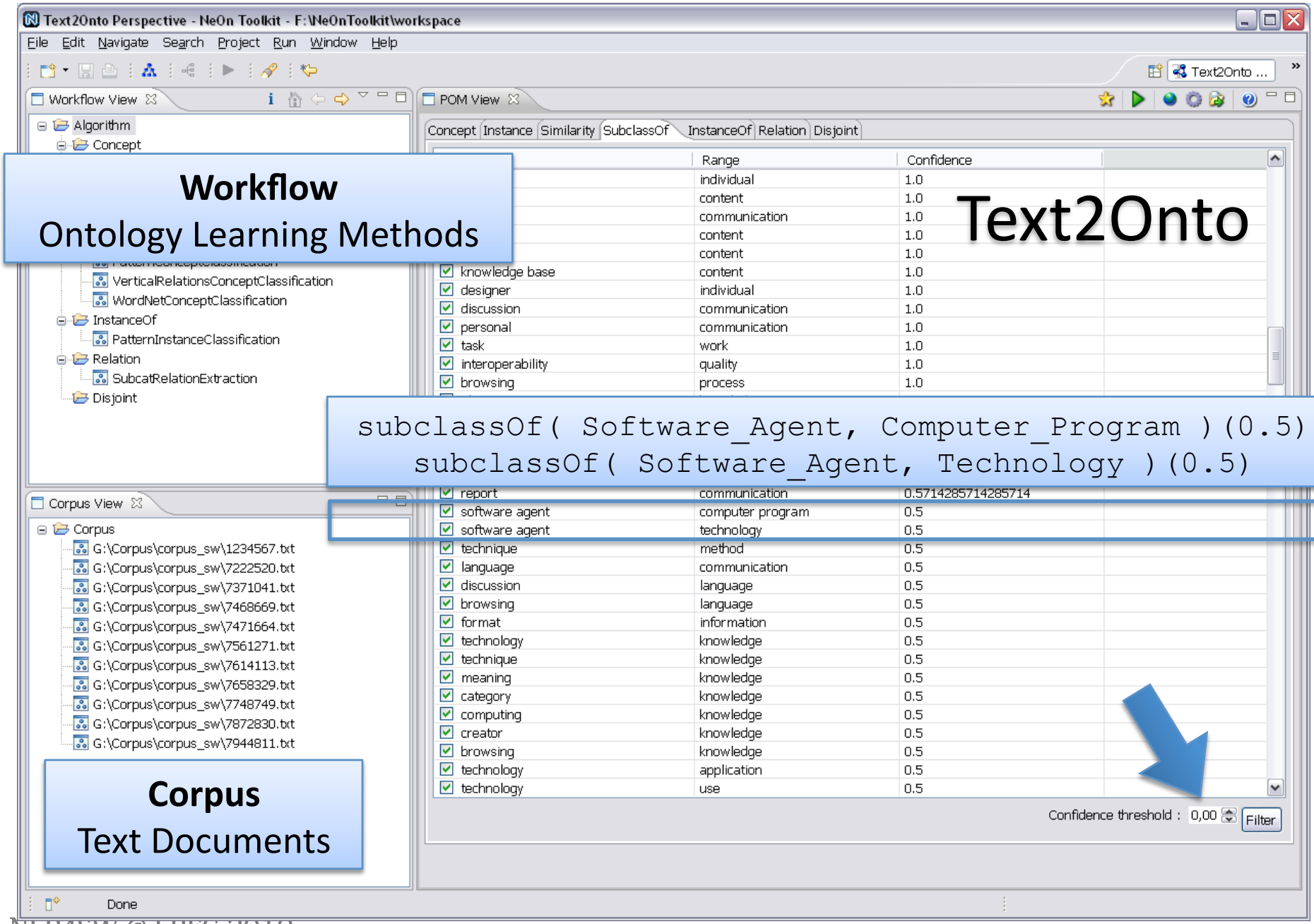
Other approaches, e.g.,

- Association rules and co-occurrence statistics
- WordNet: hyponymy ≈ subsumption
  - hyponym(`bank#1`, `institution#1`)
  - Bank subclass-of Institution
- Noun phrase heuristics
  - „`image processing` **`software`**“
- Instance clustering (e.g. `Columbus` and `Washington`)
  - Hierarchical clustering of context vectors
- Formal Concept Analysis (FCA)
  - `breathe`_subj(`animal`)
  - `breathe`_subj(`human`), `speak`_subj(`human`)
  - Human subclass-of Animal

Tools and Frameworks

Lexical ontology learning: informal or semi-formal data (e.g. texts)

Framework	Institution	Reference
ASIUM	INRIA, Jouy-en-Josas	Faure and Nedellec 1999
TextToOnto	AIFB, University of Karlsruhe	Mädche and Volz 2001
HASTI	Amir Kabir University, Teheran	Shamsfard, Barforoush 2004
OntoLT	DFKI, Saarbrücken	Buitelaar et al. 2004
DOODLE	Shizuoka University	Morita et al. 2004
Text2Onto	AIFB, University of Karlsruhe	Cimiano and Völker 2005
OntoLearn	University of Rome	Velardi et al. 2005
OLE	Brno University of Technology	Novacek and Smrz 2005
OntoGen	Institute Jozef Stefan, Ljubljana	Fortuna et al., 2007
GALeOn	Technical University of Madrid	Manzano-Macho et al. 2008
DINO	DERI, Galway	Novacek et al. 2008
OntoLancs	Lancester University	Gacitua et al. 2008


Text2Onto Perspective - NeOn Toolkit - F:\NeOnToolkit\workspace
File Edit Navigate Search Project Run Window Help
Workflow View
Algorithm
Concept
VerticalRelationsConceptClassification
WordNetConceptClassification
InstanceOf
PatternInstanceClassification
Relation
SubcatRelationExtraction
Disjoint
Workflow
Ontology Learning Methods
Text2Onto
POM View
Concept Instance Similarity SubclassOf InstanceOf Relation Disjoint
Range Confidence
individual 1.0
content 1.0
communication 1.0
content 1.0
content 1.0
knowledge base content 1.0
designer individual 1.0
discussion communication 1.0
personal communication 1.0
task work 1.0
interoperability quality 1.0
browsing process 1.0
subclassOf(Software_Agent, Computer_Program)(0.5)
subclassOf(Software_Agent, Technology)(0.5)
report communication 0.5714285714285714
software agent computer program 0.5
software agent technology 0.5
technique method 0.5
language communication 0.5
discussion language 0.5
browsing language 0.5
format information 0.5
technology knowledge 0.5
technique knowledge 0.5
meaning knowledge 0.5
category knowledge 0.5
computing knowledge 0.5
creator knowledge 0.5
browsing knowledge 0.5
technology application 0.5
technology use 0.5
Confidence threshold : 0,00 Filter
Corpus View
Corpus
G:\Corpus\corpus_sw\1234567.txt
G:\Corpus\corpus_sw\7222520.txt
G:\Corpus\corpus_sw\7371041.txt
G:\Corpus\corpus_sw\7468669.txt
G:\Corpus\corpus_sw\7471664.txt
G:\Corpus\corpus_sw\7561271.txt
G:\Corpus\corpus_sw\7614113.txt
G:\Corpus\corpus_sw\7658329.txt
G:\Corpus\corpus_sw\7748749.txt
G:\Corpus\corpus_sw\7872830.txt
G:\Corpus\corpus_sw\7944811.txt
Corpus
Text Documents
Done

Tools and Frameworks

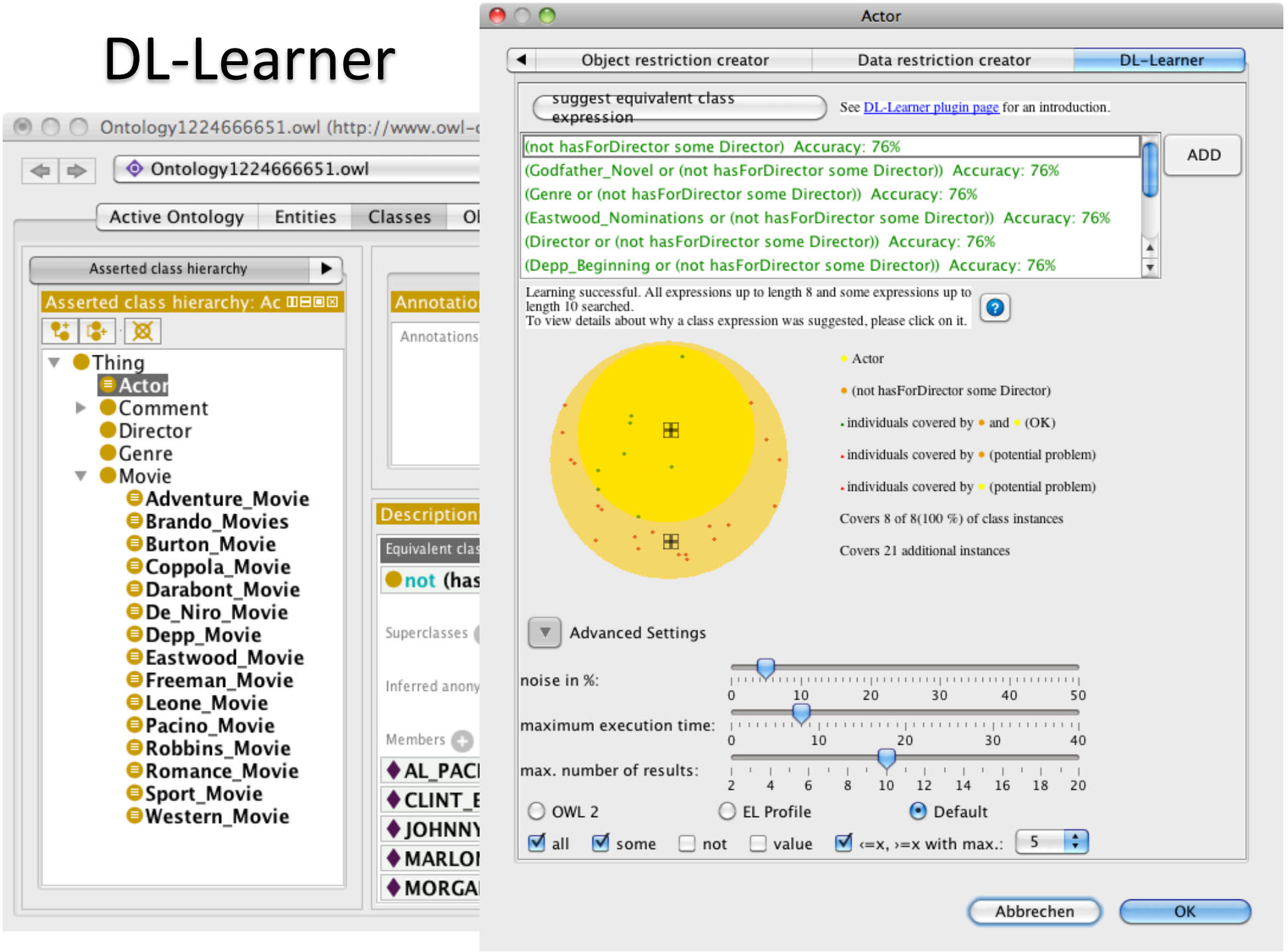
Logical ontology learning: formal data (e.g. ontologies)

Framework	Institution	Reference
YINGYANG	University of Bari	Iannone 2006
DL Learner	University of Leipzig	Lehmann 2006
RELExO	AIFB, University of Karlsruhe	Völker and Rudolph 2008
RoLExO	AIFB, University of Karlsruhe	Völker and Rudolph 2008
OntoComp	University of Dresden	Sertkaya 2008

Hybrid implementations

Framework	Institution	Reference
LeDA	AIFB, University of Karlsruhe	Völker et al. 2007
SOFIE	MPI, Saarbrücken	Suchanek et al. 2009
…	…	…

DL-Learner


Actor
Object restriction creator
Data restriction creator
DL-Learner
suggest equivalent class expression
See DL-Learner plugin page for an introduction.
(not hasForDirector some Director) Accuracy: 76%
(Godfather_Novel or (not hasForDirector some Director)) Accuracy: 76%
(Genre or (not hasForDirector some Director)) Accuracy: 76%
(Eastwood_Nominations or (not hasForDirector some Director)) Accuracy: 76%
(Director or (not hasForDirector some Director)) Accuracy: 76%
(Depp_Beginning or (not hasForDirector some Director)) Accuracy: 76%
ADD
Learning successful. All expressions up to length 8 and some expressions up to length 10 searched.
To view details about why a class expression was suggested, please click on it.
Actor
(not hasForDirector some Director)
individuals covered by and (OK)
individuals covered by (potential problem)
individuals covered by (potential problem)
Covers 8 of 8(100 %) of class instances
Covers 21 additional instances
Advanced Settings
noise in %:
maximum execution time:
max. number of results:
OWL 2
EL Profile
Default
all
some
not
value
<=x, >=x with max.:
5
Abbrechen
OK
Ontology1224666651.owl (http://www.owl-
Ontology1224666651.owl
Active Ontology
Entities
Classes
Asserted class hierarchy
Asserted class hierarchy: Ac
Thing
Actor
Comment
Director
Genre
Movie
Adventure_Movie
Brando_Movies
Burton_Movie
Coppola_Movie
Darabont_Movie
De_Niro_Movie
Depp_Movie
Eastwood_Movie
Freeman_Movie
Leone_Movie
Pacino_Movie
Robbins_Movie
Romance_Movie
Sport_Movie
Western_Movie
Annotations
Description
Superclasses
Members

2.2 Problems and Challenges

- Homonymy and polysemy e.g. [Ovchinnikova et al. 2006]
  - "Peter is sitting on the **bank** in front of the **bank**."
  - "An interesting **book** is lying on the table."
- Semantics of adjectives
  - "**red** flower", "**small** elephant", "**false** friend"
- Empty heads e.g. [Völker et al. 2005], [Cimiano and Wenderoth 2005]
  - "**Tuna** is a **kind** of fish. The **Southern Bluefin** is **one** of the most endangered **types** of Tuna."
- Ellipsis and underspecification
  - "Mary **started** the book."
- Anaphora (e.g. pronouns) e.g. [Cimiano and Völker 2005]
  - "There is an apple on the table. **It** is red."
- Metaphors and analogies
  - "**Live** is a **journey**."

2.2 Problems and Challenges (ctd.)

- Near-synonymy (e.g. `human` and `person`)
- Transitivity of lexical and conceptual relations
  - "`The button is part of the elevator. The elevator is part of the house.`"
- Opinions, quotations and reported speech
  - "`Tom thinks that` **`dolphins`** `are` **`mammals`**`.`"
- Uncertainty and imprecision e.g. [Haase and Völker 2008]
  - "`Mannheim is a` **`city`**`. (...) Mannheim is` **`big`**`.`"
- What should be represented as an individual? e.g. [Zirn et al. 2008]
  - "`The` **`kangaroo`** `is an animal living in Australia.`"
- What should be represented as a property?
  - "`All elephants are` **`grey`**`.`"
- Object property or datatype property?
  - "`Easter monday is a` **`national holiday`**`.`"
- Knowledge is changing e.g. [Zablith et al. 2009]
  - "**`Pluto`** `is a` **`planet`**`.`"

Overview

1 Motivation and Definition
2 Ontology Learning from Text
  2.1 Approaches
  2.2 Problems and Challenges
3 Reasoning for Learning and Integration

3 Reasoning for Learning and Integration

- Ontologies and rules can express formal constraints on the interpretation of entities and relations (e.g. domain and range restrictions, class disjointness)
- Checks for logical contradictions can help to detect errors, e.g., in automatically extracted information

see, for example, [Navigli and Velardi 2006], [Welty and Murdock 2006], [Suchanek et al. 2008], [Meilicke et al. 2008], [Carlson et al. 2010]

Suchanek et al. 2009

$$R(a,b) \wedge type(R, functional) \wedge b \neq c \Rightarrow \neg R(a,c)$$

Example: type(hasCapital, functional)

„Washington is the capital of the US. (...)
New York is the US capital of fashion."

hasCapital(US, New York)

hasCapital(US, Washington)

$$Washington = New\ York$$

$$Washington \neq New\ York$$

Welty and Murdock 2006

$$KB = \left\{ \begin{array}{l} \exists\ \mathrm{R}.\top \sqsubseteq \mathrm{C} \\ \top \sqsubseteq \forall\ \mathrm{R}.\mathrm{D} \\ \mathrm{D} \sqsubseteq \neg\ \mathrm{E} \end{array} \right\}$$

$$\exists\ \mathrm{bornIn}.\top \sqsubseteq \mathrm{Person}$$

$$\top \sqsubseteq \forall\ \mathrm{bornIn}.\mathrm{Location}$$

$$\mathrm{Person} \sqsubseteq \neg\ \mathrm{Location}$$

„Washington was born in Westmoreland County, Virginia on February 22, 1732."

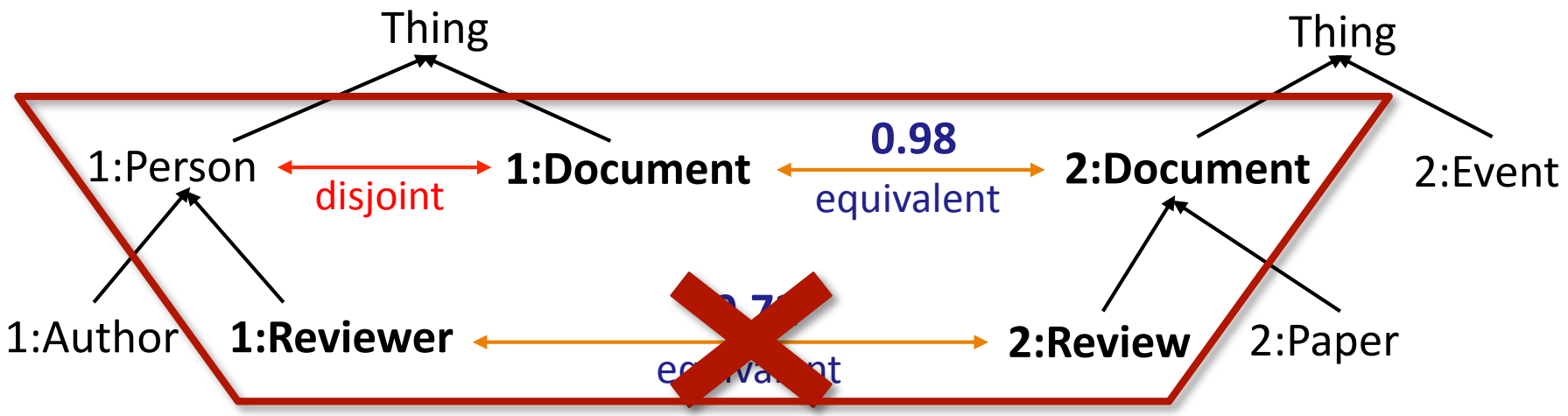
$$\mathrm{Location}(\ \mathrm{Washington}\ )$$

$$\mathrm{bornIn}(\ \mathrm{Washington},\ \mathrm{Virginia}\ )$$

$$\mathrm{Person}(\ \mathrm{Washington}\ )$$

$$\neg\ \mathrm{Location}(\ \mathrm{Washington}\ )$$

Meilicke et al. 2008



$$O_1 \cup_M O_2 \models \text{1:Person} \sqsubseteq \neg\ \text{1:Document}$$

$$O_1 \cup_M O_2 \models \text{1:Reviewer} \sqsubseteq \text{1:Person}$$

$$O_1 \cup_M O_2 \models \text{1:Reviewer} \sqsubseteq \text{1:Document}$$

~~$O_1 \cup_M O_2 \models \text{1:Reviewer} \sqsubseteq \bot$~~

Summary

- Ontology learning and population
- Automatic and semi-automatic approaches to ontology learning from text
- Various types of input resources (formality, structure etc.)
- Challenges, e.g., ambiguity of natural language
- Sometimes, inference helps!