Text Analysis and Ontologies

Philipp Cimiano
Institute AIFB
University of Karlsruhe

Summer School on Multimedia Semantics
September 7, 2006

Roadmap

Part I (Introduction)

Part II (Information Extraction)

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- Web-based Information Extraction
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Part III (Ontology Learning)

- Motivation
- Learning Concept Hierarchies
- Learning Relations

Part I

Introduction

SmartWeb - Goals

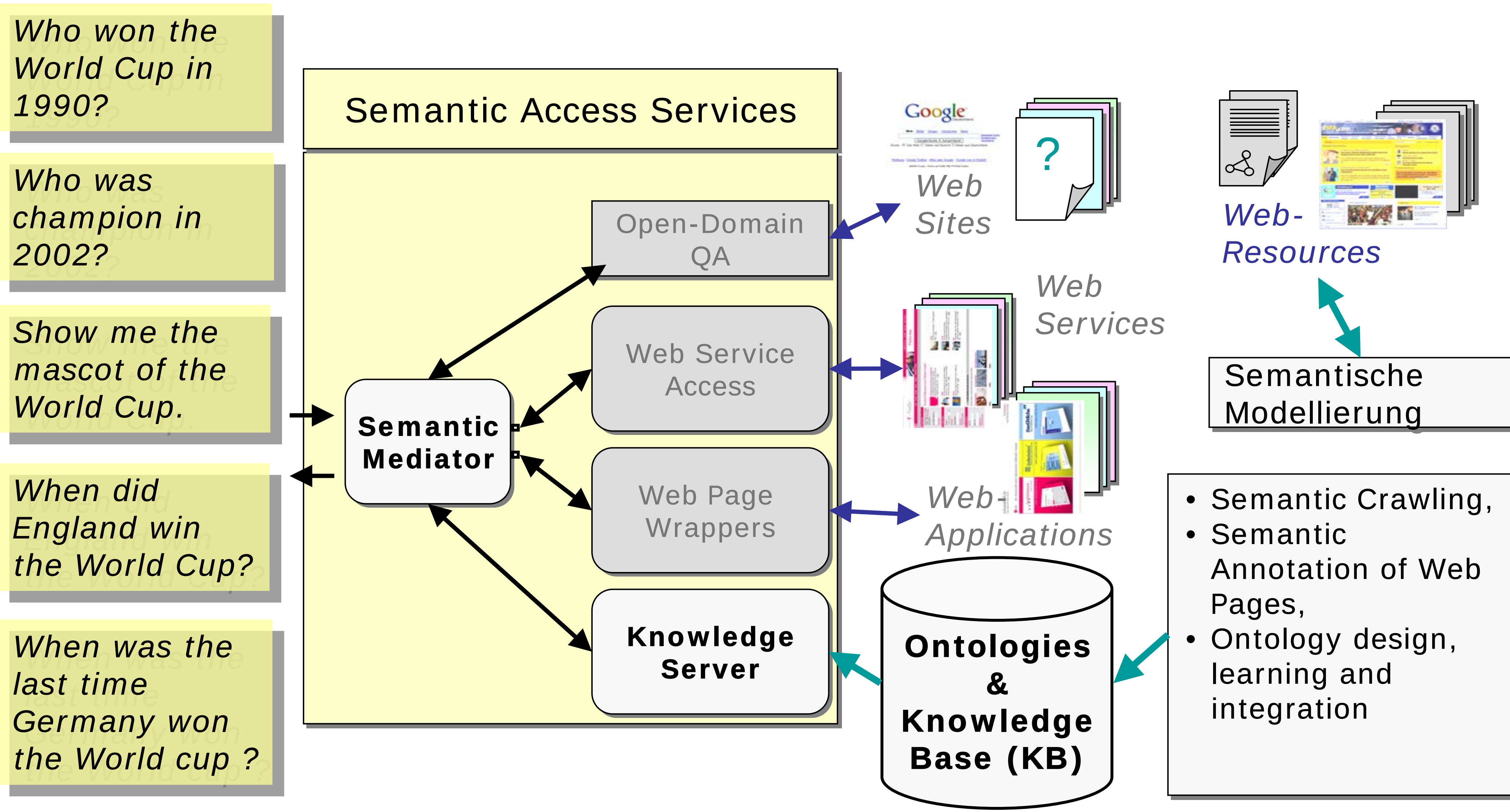
Goal: Ubiquitous and Broadband Access to the Semantic Web

Core Topics:

- Multimodality
- Question Answering
- Web Services (Matching, Composition)
- Semantic Annotation / Metadata Generation
- KB Querying / Reasoning
- Applications of Ontologies

Scenario: Question Answering for the 2006 Worldcup

The SmartWeb System



Roadmap

Part I (Semantic Karlsruhe)

Part II (SmartWeb)

- Introduction to the SmartWeb Project
- **The Role of Ontologies in SmartWeb**
- Metadata Generation in the SmartWeb System with SOBA

Part III (Reasoning)

- Motivation
- OWL DL Reasoning with KAON2
- Approximate Reasoning with Screech

Ontologies

- Computers are essentially symbol-manipulating machines.
- For applications in which meaning is shared between parties, ontologies play a crucial role.
- Ontologies fix the interpretation of symbols w.r.t some semantics (typically model-theoretic)
- Ontologies are formal specifications of a shared conceptualization of a certain domain [Gruber 93].

Ontologies in Philosophy

- A Branch of Philosophy that Deals with the Nature and Organization of Reality
- Science of Being (Aristotle, Metaphysics)
  - *What Characterizes Being?*
  - *Eventually, what is Being?*

Ontologies in Computer Science

- Ontology refers to an engineering artifact
  - a specific vocabulary used to describe a certain reality
  - a set of explicit assumptions regarding the intended meaning of the vocabulary

- An Ontology is
  - an **explicit** specification of a conceptualization [Gruber 93]
  - a **shared understanding** of a domain of interest [Uschold and Gruninger 96]

SW Ontology languages

- Nowadays, there are different ontology languages:
  - DAML + OIL
  - RDF(S)
  - OWL
  - F-Logic
- Essentially, they provide:
  - Taxonomic organization of concepts
  - Relations between concepts (with type and cardinality constraints)
  - Instantiation relations

Why Develop an Ontology?

- Make domain assumptions **explicit**
  - Easier to exchange domain assumptions
  - Easier to understand and update legacy data

- Separate **domain knowledge** from operational knowledge
  - Re-use domain and operational knowledge separately

- A **community reference** for applications

- **Shared understanding** of what information means

Applications of Ontologies

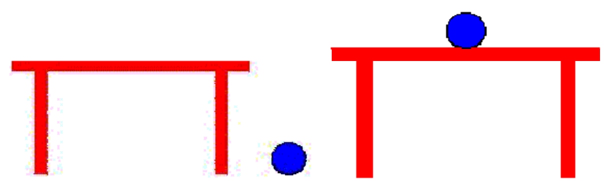
- NLP
  - **Information Extraction**, e.g. [Buitelaar et al. 06], [Stevenson et al. 05], [Mädche et al. 02]
  - **Information Retrieval (Semantic Search)**, e.g. WebKB [Martin and Eklund 00], SHOE [Hendler et al. 00], OntoSeek [Guarino et al. 99]
  - **Question Answering**, e.g. [Sinha and Narayanan 05], [Schlobach et al. 04], Aqualog [Lopez and Motta 04], [Pasca and Harabagiu 01]
  - **Machine Translation**, e.g. [Nirenburg et al. 04], [Beale et al. 95], [Hovy and Nirenburg 92], [Knight 93]

- Other
  - **Business Process Modeling**, e.g. [Uschold et al. 98]
  - **Information Integration**, e.g. [Kashyap 99], [Wiederhold 92]
  - **Knowledge Management (incl. Semantic Web)**, e.g. [Fensel 01], [Mulholland et al. 2001], [Staab and Schnurr 00], [Sure et al. 00], [Abecker et al. 97]
  - **Software Agents**, e.g. [Gluschko et al. 99], [Smith and Poulter 99]
  - **User Interfaces**, e.g. [Kesseler 96]

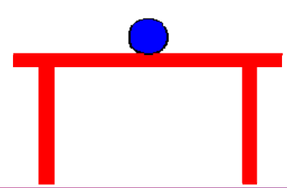
Example Semantic Image Retrieval

E.g.: Give me images with a ball on a table.

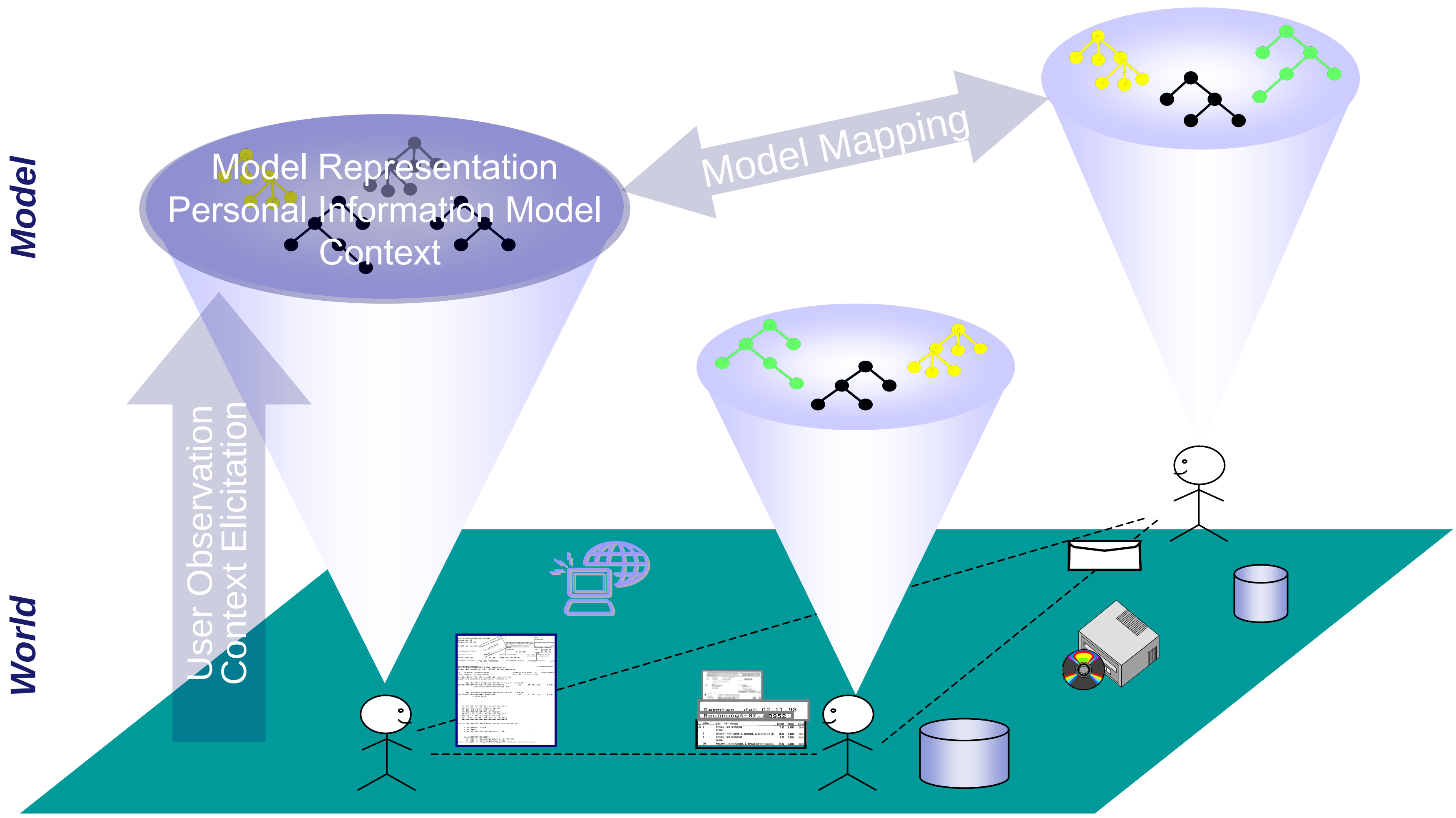
- State-of-the-art: ask Google Images for „ball on table":



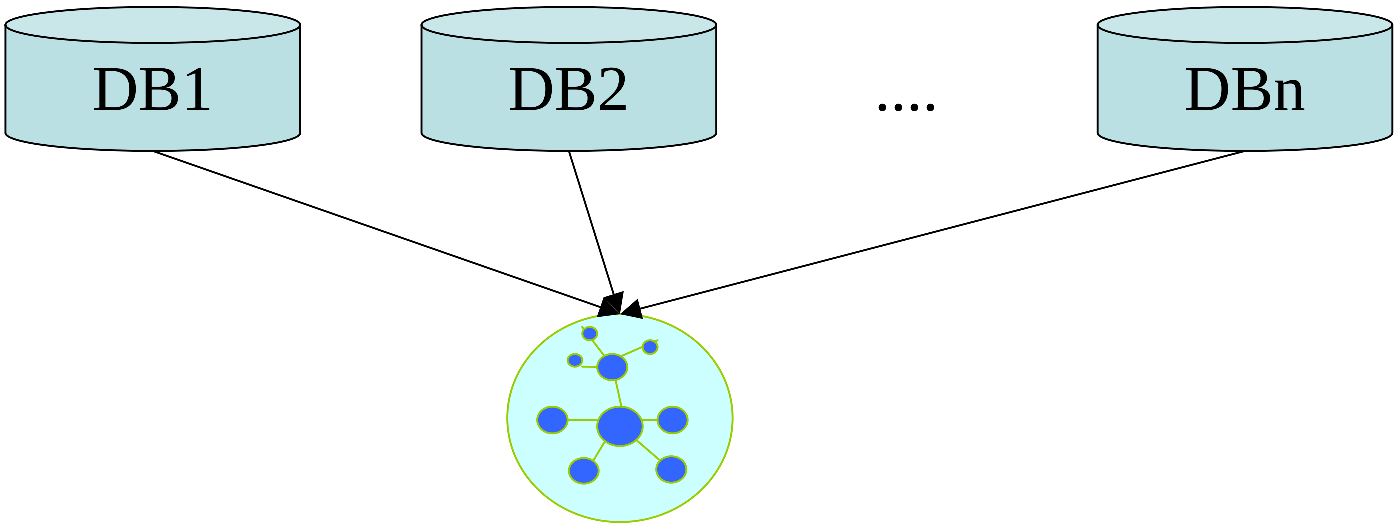
- Semantic Web: specify what you want precisely:
  - FORALL X <- X:image AND EXISTS B,T X[contains -> B] AND X[contains -> T] AND B:ball and T:table and B[locatedOn -> T].



Representation, Acquisition, and Mapping of Personal Information Models is at the heart of KM Research



Information Integration

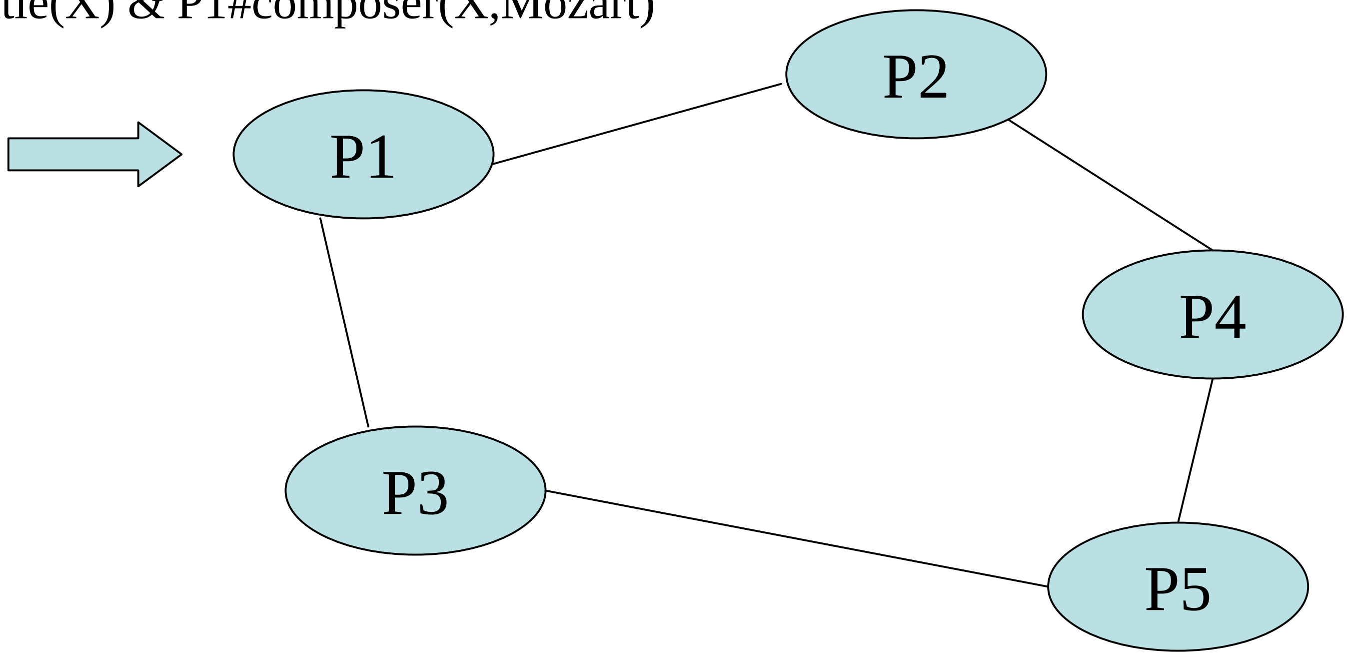


?X employee(X) & worksFor(X,salesDep)

Mapping in Distributed Systems

P1#composer(X,Y) <- P2#author(X,Y)

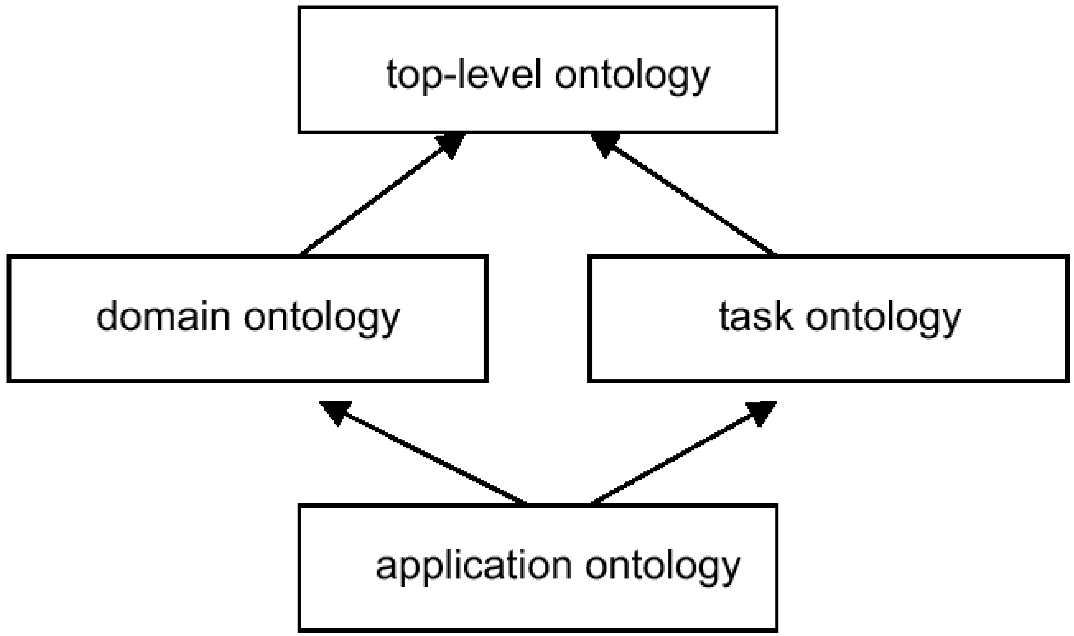
?X P1#title(X) & P1#composer(X,Mozart)



Types of Ontologies [Guarino 98]

Describe **very general concepts** like space, time, event, which are independent of a particular problem or domain.

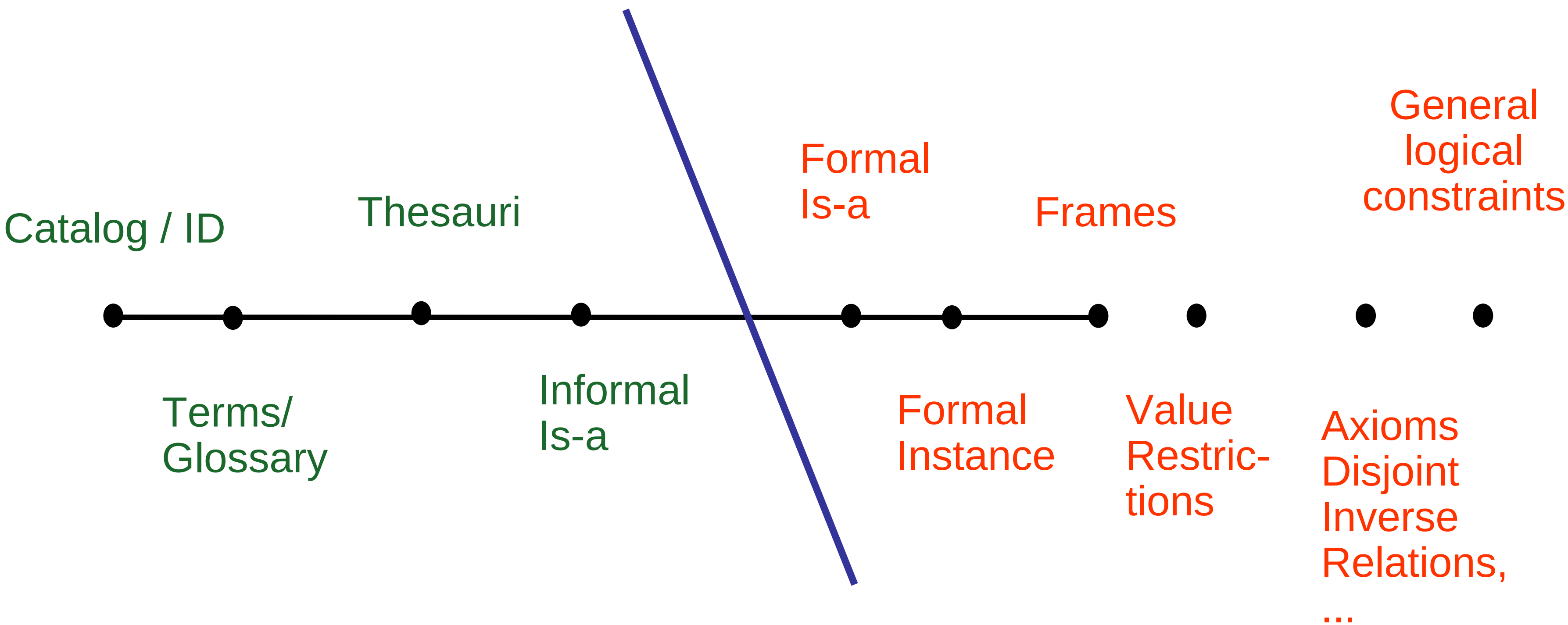
Describe the vocabulary related to a **generic domain** by specializing the concepts introduced in the top-level ontology.



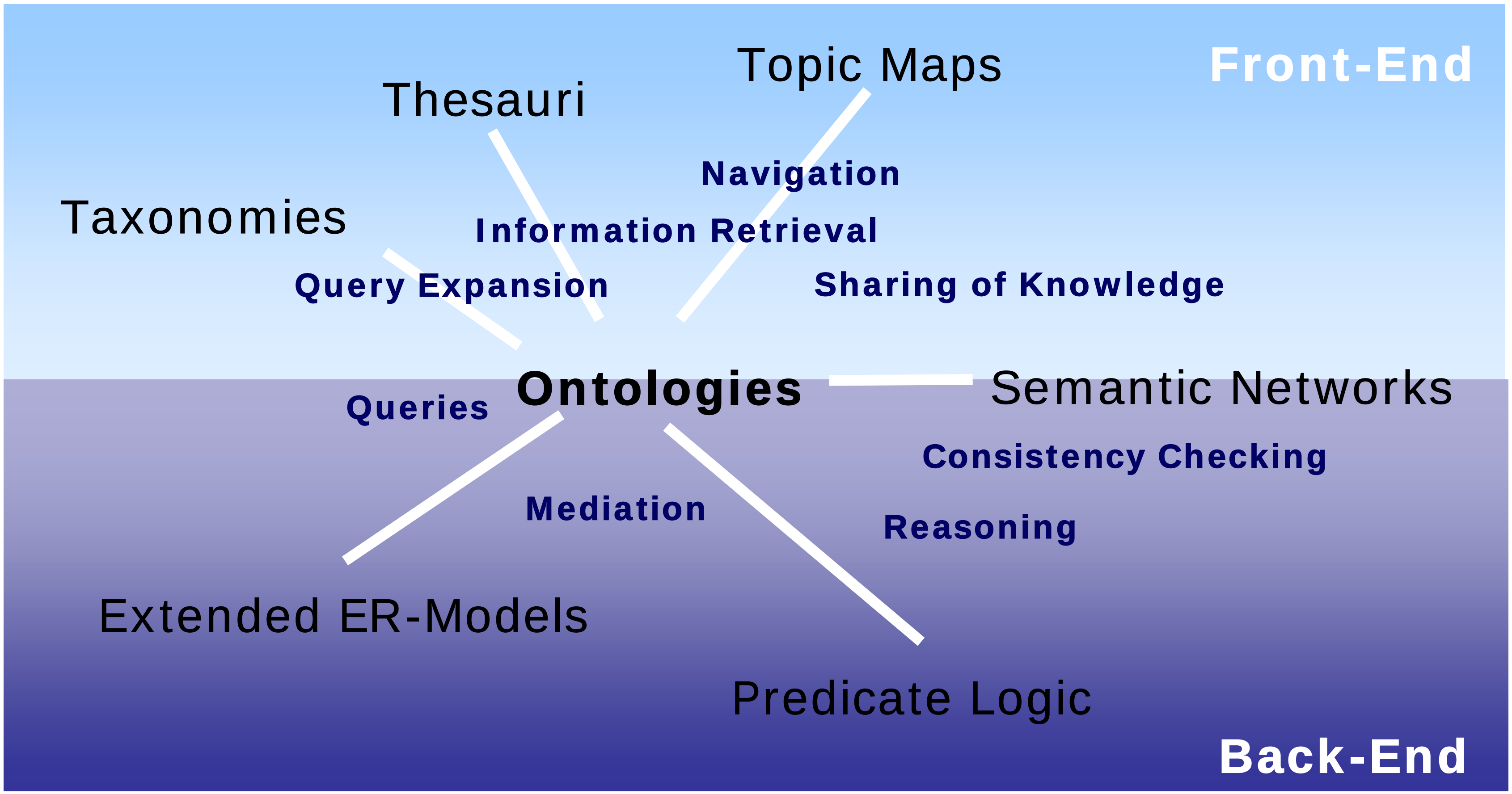
Describe the vocabulary related to a **generic task or activity** by specializing the top-level ontologies.

Concepts in application ontologies often correspond to **roles played by domain entities while performing a certain activity**.

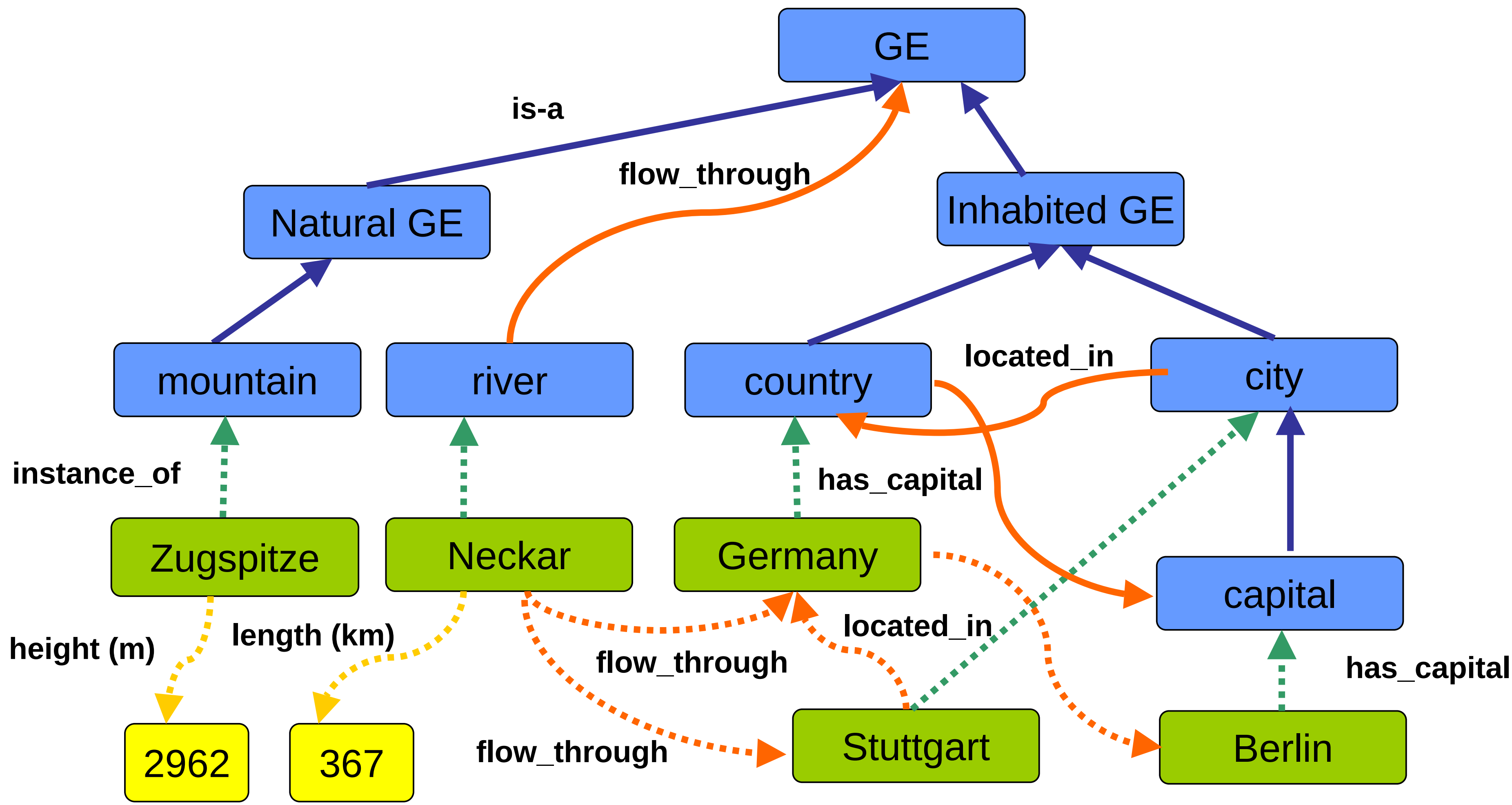
Ontologies and Their Relatives



Ontologies and Their Relatives (Cont'd)



Example: Geographical Ontology



But to be honest...

- There are not much (real) ontologies around:
  - Most SW Ontologies are RDFSed thesauri!
  - Most people don‘t think model-theoretically!
- So we have to live with:
  - Linguistic „Ontologies“ like WordNet
  - Thesauri
  - Automatically Learned Thesauri/Taxonomies/Ontologies

Example: Ontologies in SmartWeb

- Integration of Heterogeneous Sources
  - **one view** on all the data
- Clear definition of the scope of the system
  - precisely defined by ontology
- Shared understanding of the domain
  - makes communication with project partners easier
- Question Answering as a well-defined (inferencing) process
  - no “adhoc” solutions for question answering
- Inference of “implicit” relations
  - avoids redundancy in the Knowledge Base

Integration of Heterogeneous Sources

Ontology offers one view on top of:

- Manually acquired soccer facts (mainly World Cups)
- Automatically extracted metadata (FIFA Web pages)
- Semantic Web Services (e.g. Road and Traffic Conditions, Public Transport, ..)
- Open-domain Question Answering

Offline vs. online integration:

- Offline Integration
  - Ontologies with DOLCE and SmartSumo as "top level"
  - "Offline Data" (manually and automatically acquired soccer facts)

- Online Integration
  - Integration at query time (Web Service invocation, Open-domain QA)

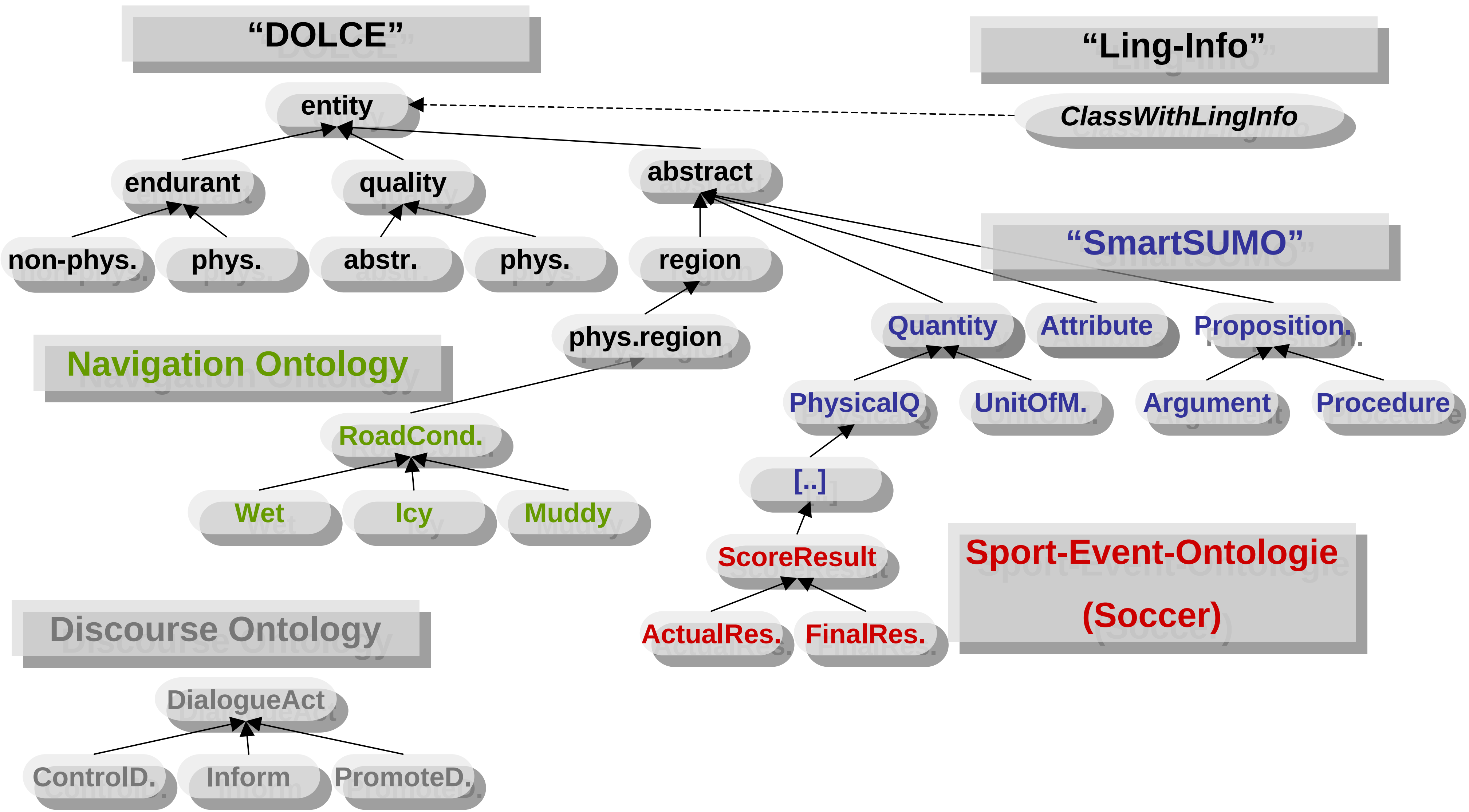
The Ontologies in the SmartWeb project

- SWIntO (SmartWeb Integrated Ontology) Components:
  - Sport-Event-Ontology (Soccer)
  - Navigation Ontology
  - Multimedia Ontology
  - Discourse Ontology
  - Linguistic Information (LingInfo)

- Integration of the above domain ontologies via:
  - DOLCE as “foundational ontology” (FO)
  - SUMO aligned to DOLCE as “upper level ontology”

- Benefits: Conceptual disambiguation and Modularisation!

The Ontologies in the SmartWeb project

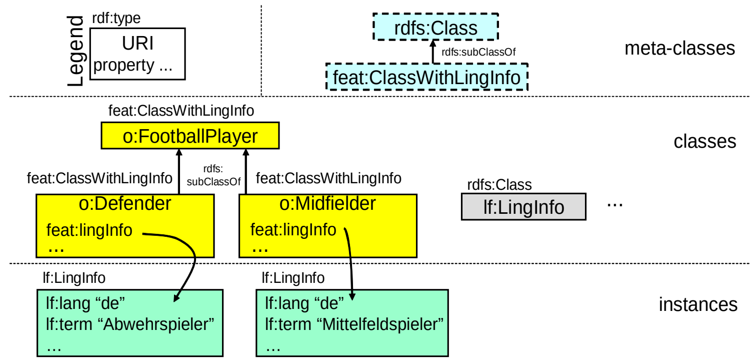


The Role of the Ontologies in SmartWeb

- **SmartSUMO (DOLCE + SUMO)**
  - Well-defined integration of the domain ontologies
  - *Descriptions & Situations* (DOLCE Extension) used for description of web services and for supporting navigation (context-modelling)
- **Sport-Event-Ontology (Soccer)**
  - Defines thematic scope of SmartWeb
- **Navigation Ontology**
  - Provides *„perdurants“* (Motions, Processes, ...), *„endurants*“ (streets, buldings, cities, etc) und *„quality regions*“ (conditions of roads) for the purpose of navigation
- **Discourse Ontology**
  - Provides Concepts for Dialog-Management, Answer-Types, Dialog-(Speech)-Acts, HCI-Aspects
- **Linginfo**
  - Provides „Grounding“ of the Ontology through natural language

SmartSUMO, Sport-Event-Ontology, Multimedia-Ontology und Linginfo in action at query time

“When did Germany win the World cup ?”



FORALL Focus <- EXISTS FocusObject, O2, O4, O3, O1, Media, FocusValue (O1:**WorldCup**[**dolce#"HAPPENS-AT**"   ->> O2:"time-interval"[**dolce#BEGINS** ->> FocusObject:"time-point"]; **winner** ->> O3:**DivisionFootballNationalTeam**[**origin** ->> O4:country[**linginfo#term** ->> ***„ Germany“]]]*** AND FocusObject[**dolce#YEAR** ->> FocusValue] AND Media[**media#shows** -> O3] AND unify(Focus, result(FocusValue, focus_media_object(O3, Media))))

Roadmap

Part I (Introduction)

Part II (Information Extraction)

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- Web-based Information Extraction
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Part III (Ontology Learning)

- Motivation
- Learning Concept Hierarchies
- Learning Relations

What is information extraction ?

- Definition: Information extraction is the task of filling certain given target knowledge structures on the basis of text analysis. These target knowledge structures are often also called ***templates***.

- Input: A collection of texts and a template schema to be filled

- Output: A set of instantiated templates.

Information Extraction vs. Natural Language Understanding

- Information Extraction **is not** Natural Language Understanding!

Natural Language Understanding (NLU)

• Aims at complete understanding of a text

• Uses deep NLP techniques (full parsing, semantic and pragmatic analysis, etc).

• Requires knowledge representation, reasoning etc.

• Is a very difficult task – AI completeness.

•There is not yet a system performing NLU to a reasonable extent.

Information Extraction (IE)

• Aims ‚only' at extracting information for filling a pre-defined schema (template)

• Typically applies shallow NLP techniques (shallow parsing, shallow semantic analysis, merging of structures, etc.)

• Is a much more restricted task than NLU and thus easier.

• There have been very succesful systems.

What do we need it for ?

- Question Answering – IE for extracting facts
- Text filtering or classification – IE facts as features
- Text Summarization – IE as preprocessing
- Knowledge Acquisition – IE for database filling

Information Extraction

- Motivation
- **Classic Information Extraction**
- Adaptive Information Extraction
- Web-based Information Extraction
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Classic Information Extraction

- Mainly sponsored by DARPA in the framework of the Message Understanding Conferences (MUC)
  - MUC-1 (1987) and MUC-2 (1989)
    - Messages about naval operations
  - MUC-3 (1991) and MUC-4 (1992)
    - News articles about terrorist attacks
  - MUC-5 (1993)
    - News articles about joint ventures and microelectronics
  - MUC-6 (1995)
    - News articles about management changes
  - MUC-7 (1997)
    - News articles about space vehicle and missile launches

MUC-7 template example

Launch Event:

- Vehicle: <VEHICLE_INFO>
- Payload: <PAYLOAD_INFO>+
- Mission_Date: <TIME>
- Mission_Site: <LOCATION>
- Mission_Type: {Military, Civilian}
- Mission_Function: {Test, Deploy, Retrieve}
- Mission_Status: {Succeeded, Failed, In_Progress, Scheduled}

Different steps at one glance

Template Merging / Fusion
Template Filling
Shallow Parsing
Part-of-Speech (POS) Tagging
Named Entity Recognition & NE Coreference
Tokenization & Normalization

Tokenization & Normalization

- **Tokenization:**
  - Good enough: white spaces indicate token boundaries
  - Full stops indicate sentences boundaries (does not always work, e.g. 1. September)

- **Normalization:**
  - Dates, e.g. 1. September 2006 -> 1.09.2006
  - Abbreviations, e.g. MS -> Microsoft

    (requires a lexicon with abbreviations!)

NER & NE Coreference

- <u>NER:</u> Recognize names of persons, organizations, companies
- <u>Methods:</u>
  - essentially lexicon lookup in so called gazetteers
  - apply trained models
    - Rule-based (transformation-based) approaches [Brill]
    - HMM-based approaches
      - bigrams, trigrams, ...
      - Probability for a tag given a certain bigram
      - Viterbi algorithm to compute most likely tag
- <u>NE Coreference:</u>
  - Detect that „Mr. Gates“, „B. Gates“ and „Bill Gates“ refer to the same entity
  - Apply heuristics!

Concrete Example

Xichang, China, Feb. 15 (Bloomberg) -- **A Chinese rocket carrying an Intelsat satellite** exploded as it was being launched **today**, delivering a blow to a group including Rupert Murdoch's News Corp. and Tele-Communications Inc. that planned to use the spacecraft to beam television signals to Latin America. ``We're in a risky business. These things happen from time to time,'' said Irving Goldstein, director general and chief executive of Intelsat. His comments came at the company's Washington headquarters, where hundreds of reporters, diplomats and industry officials gathered to watch **the launch from China** on large video screens. The China Great Wall Industry Corp. provided the **Long March 3B rocket** for **today**'s **failed launch** of a satellite built by Loral Corp. of New York for Intelsat. **It carried 40 transponders** and would have had a primary broadcast footprint that extended from southern California through Central America and from Colombia to northern Argentina in South America.

Tokenizing (CASS tokenizer)

```
a        A        \s
chinese Chinese \s
rocket  rocket  \s
carrying        carrying        \s
an      an      \s
intelsat        Intelsat        \s
satellite       satellite       \s
exploded        exploded        \s
as      as      \s
it      it      \s
was     was     \s
being   being   \s
launched        launched        \s
today   today   -
.       .       \n
```

Part-of-speech (POS) tagger (IMS Tree Tagger)

DT	a
JJ	Chinese
NN	rocket
VVG	carry
DT	an
NP	Intelsat
NN	satellite
VVD	explode
IN	as
PP	it
VBD	was
JJ	being
VVN	launch
NN	today
SENT	.

Shallow Parsing
(Steven Abney‘s CASS)

```
[nx
    [dt-a a]
    [jj Chinese]
    [nn rocket]]
  [vvg carry]
  [nx
    [dt an]]
    [np Intelsat]
    [nn satellite]]
  [vvd explode]
  [as as]
  [pp it]
  [vp
    [vx
      [be be]
      [jj being]]]
  [vvn launch]
  [today today]
  [sent .]
```

Template Extraction

[nx1:rocket] [vvg carry] [nx2:thing] =>

Vehicle: head(nx1)

Payload: head(nx2)

Mission_Date: ?

Mission_Site: ?

Mission_Type: ?

Mission_Function ?

Mission_Status: ?

A Chinese rocket carrying an Intelsat satellite exploded as it was being launched **today.** =>

Vehicle: Chinese rocket

Payload: Intelsat satellite

Mission_Date: ?

Mission_Site: ?

Mission_Type: ?

Mission_Function ?

Mission_Status: ?

Discourse Analysis / Template Merging (1)

A Chinese rocket carrying an Intelsat satellite
exploded as it was being launched **today.**

Vehicle: **Chinese rocket**
Payload: **Intelsat satellite**
Mission_Date:
Mission_Site: ?
Mission_Type: ?
Mission_Function ?
Mission_Status: ?

Vehicle:
Payload:
Mission_Date: **today**
Mission_Site: ?
Mission_Type: ?
Mission_Function ?
Mission_Status: ?

Vehicle: **Chinese rocket**
Payload: **Intelsat satellite**
Mission_Date: **14.2.1996**
Mission_Site: ?
Mission_Type: ?
Mission_Function ?
Mission_Status: ?

Discourse Analysis / Template Merging (2)

[...] hundreds of reporters, diplomats and industry officials gathered to watch **the launch from China** on large video screens.

Vehicle: **Chinese rocket**	Vehicle:	Vehicle: **Chinese rocket**
Payload: **Intelsat satellite**	Payload:	Payload: **Intelsat satellite**
Mission_Date: **14.2.1996**	Mission_Date:	Mission_Date: **14.2.1996**
Mission_Site: ?	Mission_Site: **China**	Mission_Site: **China**
Mission_Type: ?	Mission_Type: ?	Mission_Type: ?
Mission_Function ?	Mission_Function ?	Mission_Function ?
Mission_Status: ?	Mission_Status: ?	Mission_Status: ?

Discourse Analysis / Template Merging (3)

The China Great Wall Industry Corp. provided the **Long March 3B rocket** for **today**'s **failed launch of a satellite** built by Loral Corp. of New York for Intelsat.

Vehicle: **Chinese rocket**

Payload: **Intelsat satellite**

Mission_Date: **14.2.1996**

Mission_Site: **China**

Mission_Type: ?

Mission_Function ?

Mission_Status: ?

Vehicle: **Long Match 3B rocket**

Payload: **satellite**

Mission_Date: **today**

Mission_Site: ?

Mission_Type: ?

Mission_Function ?

Mission_Status: **failed**

Vehicle: **Chinese Long Match 3B rocket**

Payload: **Intelsat satellite**

Mission_Date: **14.2.1996**

Mission_Site: **China**

Mission_Type: ?

Mission_Function ?

Mission_Status: **failed**

Discourse Analysis / Template Merging (4)

It carried 40 transponders [...]

Vehicle: **Chinese Long Match 3B rocket**
Payload: **Intelsat satellite**
Mission_Date: **14.2.1996**
Mission_Site: **China**
Mission_Type: ?
Mission_Function ?
Mission_Status: **failed**

Vehicle: **Chinese Long Match 3B rocket**
Payload: **40 transponders**
Mission_Date: ?
Mission_Site: ?
Mission_Type: ?
Mission_Function ?
Mission_Status: ?

Vehicle: **Chinese Long Match 3B rocket**
Payload: {**Intelsat satellite, 40 transp.}**
Mission_Date: **14.2.1996**
Mission_Site: **China**
Mission_Type: ?
Mission_Function ?
Mission_Status: **failed**

How good does this work ?

- Information Extraction systems are typically evaluated in terms of Precision and Recall.

$$P = \frac{\text{correctly extracted facts}}{\text{extracted facts}}$$

$$R = \frac{\text{correctly extracted facts}}{\text{correct facts}}$$

$$F_1 = \frac{2PR}{P+R}$$

- This assumes a „gold standard“ specifying what is correct.
- It is typically assumed that there is a F=60% limit for IE [Appelt and Israel 1999]
  - Complex syntactic phenomena can not be handled by a shallow parser
  - Discourse processing is more than template merging and pronoun resolution
  - We need inferences, e.g.

$$\forall x, y\ launch(x) \wedge carry(x, y) \wedge TV - Satellite(y) \rightarrow MissionType(x, "civil").$$

Some reference points

- POS tagging
  - $F_1 > 95\%$
- Named Entity Recognition (Person, Company, Organization)
  - $F_1 > 95\%$
- Template Extraction
  - Best System: (MUC-7) F_1=50.79%
  - Worst System: (MUC-7) F_1=1.45%

- Have a look at:

http://www-nlpir.nist.gov/related_projects/muc/proceedings/st_score_report.html

Pros and Cons of Classic Information Extraction

PROs

- **Clearly understood technology**
- **Hand-written rules are relatively precise**
- **People can write rules with a reasonable ammount of training**

CONs

- **Rules need to be written by hand**
- **Requires experienced grammar developers**
- **Difficult to port to different domains**
- **Limits of technology (F < 70%)**

Question: Can we create more adaptive information extraction technology ?

Information Extraction

- Motivation
- Classic Information Extraction
- **Adaptive Information Extraction**
- Web-based Information Extraction
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Adaptive Information Extraction

- Why Adaptive IE ?
  - No handwriting of rules
  - Tuning to a domain by Machine Learning

- Hypothesis:
  - easier to annotate text than to write rules
  - No grammar developers needed

- Requires
  - Training set with ‚enough‘ examples for each class
  - An appropriate pattern induction technique

Principle of Adaptive IE / Lazy NLP ?

- Information extraction as a classification problem:
  - Given a text passage $w_{ij,}$ does it fill the value of some slot s, i.e.

$$f_s(w_{ij}) \rightarrow \{t, f\}$$

- Lazy NLP:
  - More information (POS-tags, Syntactic Dependencies, lexical information etc.) is only included if it help to induce ‚better‘ rules

Adaptive IE / Lazy NLP Systems

- The paradigm of IE as a classification task is implemented by a number of systems:

  – WHISK – [Soderland 1999]

  – Rapier - [Califf and Mooney 19999]

  – Boosted Wrapper Induction (BWI)– [Freitag and Kushmerick 2000]

  – Amilcare – [Ciravegna 2001]

Amilcare [Ciravegna 2001]

- Amilcare is an information extraction system based on the LP^2 rule induction algorithm
- LP^2 is a rule induction algorithm which learns patterns to extract values of a slot to be filled in a template
- It relies on a set of training data in which the values to be extracted are marked with XML-tags, e.g.
  - The seminar will start at <stime> 4 </stime> pm.
- On the basis of these annotations, rules are induced using different levels of linguistic analysis (Lazy-NLP aspect)
- It relies on word windows of a given length around the slot filler.
- An important move in LP^2 is to insert start and end tags separately, i.e. we have separate rules inserting <stime> and </stime> tags.

Rule Induction in Amilcare

- The easiest pattern corresponds to the surface word order of the example, i.e. taking a word window of 5 tokens, the simplest pattern is:

  „The seminar will start at“ -> insert <stime> tag

- This pattern has however a low recall as it captures only one example. So we want to generalize.
- As we want to move (potentially) to different levels of analysis, we specifiy that this is a pattern at the surface word level:

  w_{-5}=„The“, $w_{-4=}$„seminar“, w_{-3}=„will“, $w_{-2=}$„start“, w_{-1}=„at“
  -> insert <stime> at w_0

What generalizations could be feasible?

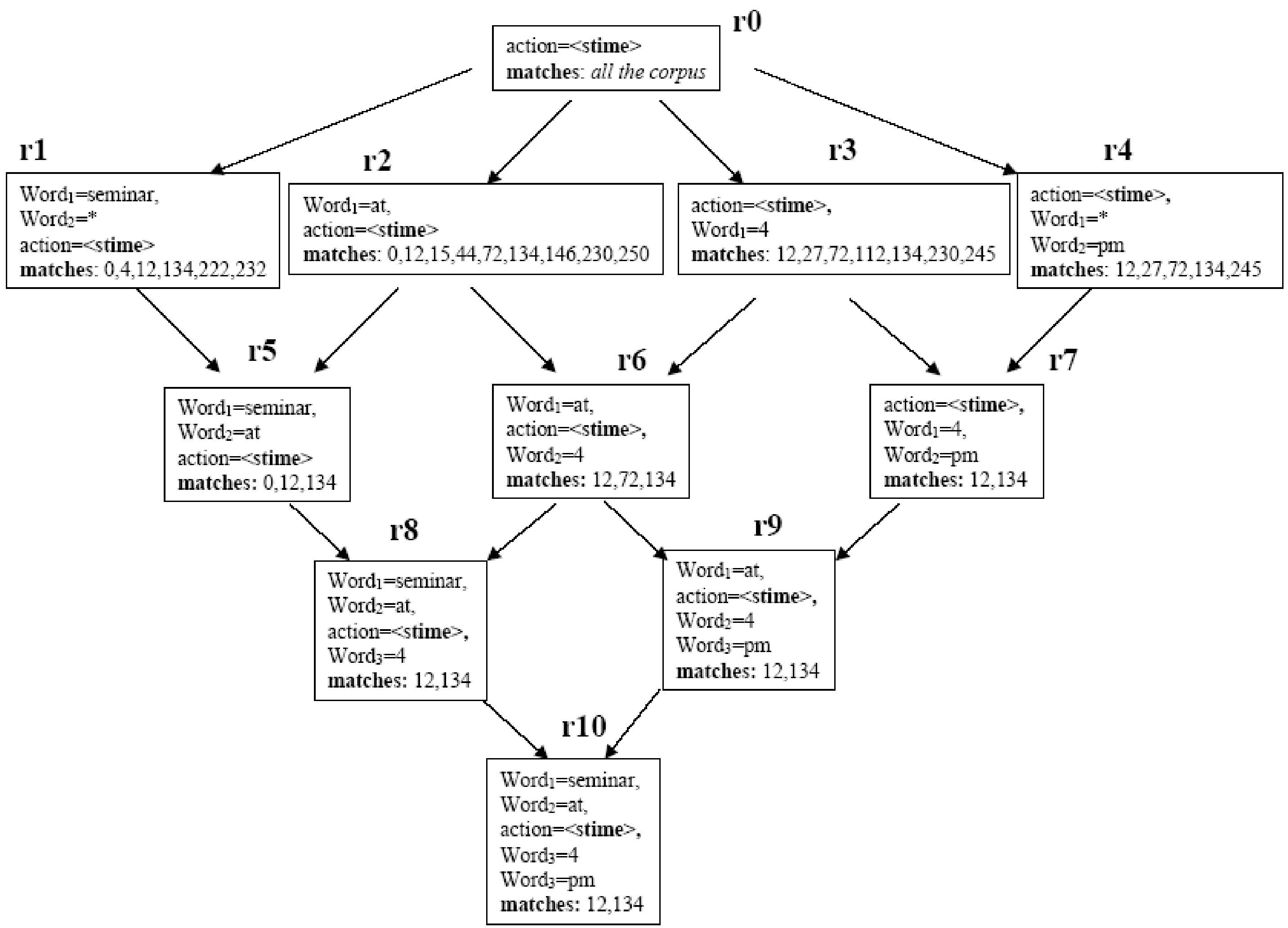
w_{-5}=„The“, w_{-4}=„seminar“, w_{-3}=„will“, w_{-2}=„start“, w_{-1}=„at“
-> insert <stime> at w_0

w_{-5}=*, pos_{-5}=DT, w_{-4}=„seminar“, w_{-3}=„will“, w_{-2}=„start“, w_{-1}=„at“
-> insert <stime> at w_0

w_{-5}=„The“, w_{-4}=„seminar“, w_{-3}=*, w_{-2}=„start“, w_{-1}=„at“
-> insert <stime> at w_0

– The search space is indeed very large as all the possible generalizations form a lattice of size 2^{f*l}.
– For each generalization, the accuracy of the rule needs to be tested to find it if this is a promising direction! This helps in reducing the search space.
– Keep always the k-best rules!

The lattice explored by Amilcare



Classic IE vs. Adaptive IE

Classical IE

+ very precise (hand-coded rules)

+ handles domain-independent phenomena (to some extent)

- **need to develop grammars**
- **expensive development & test cycle**
- **develop lexicons, gazetteers, etc.**

Adaptive IE

+ reasonable precision (rule induction)

+ higher recall

+ no need for developing grammars

- **provide training data (expensive)**
- **simplification of tasks (one template, one instance per document, etc.) (F ~ 80%)**
- **typically „overfitted“ to the domain**
- **develop lexicons, gazetteers, etc.**
- **rules can be hard to interprete**

Information Extraction

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- **Web-based Information Extraction**
  - **Instance Classification**
  - Relation Extraction
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Web- based Information Extraction

- **Problem:** Methods relying on corpora are affected by data sparseness
- **Idea:** Use the web to overcome data sparseness!

- **Advantages:**
  - Search engines have a massive coverage
  - Easy to use APIs
  - Up-to-date information

- **Disadvantages:**
  - Issuing queries to a search engine API can take a lot of time!
  - Trust (Page-rank as a solution?)
  - Commercially biased! (Any solution)

The Self-Annotating Web
- The PANKOW Approach -

- There is a huge amount of implicit knowledge in the Web
- Make use of this implicit knowledge together with statistical information to propose formal annotations and overcome the vicious cycle:

  semantics ≈ syntax + statistics?

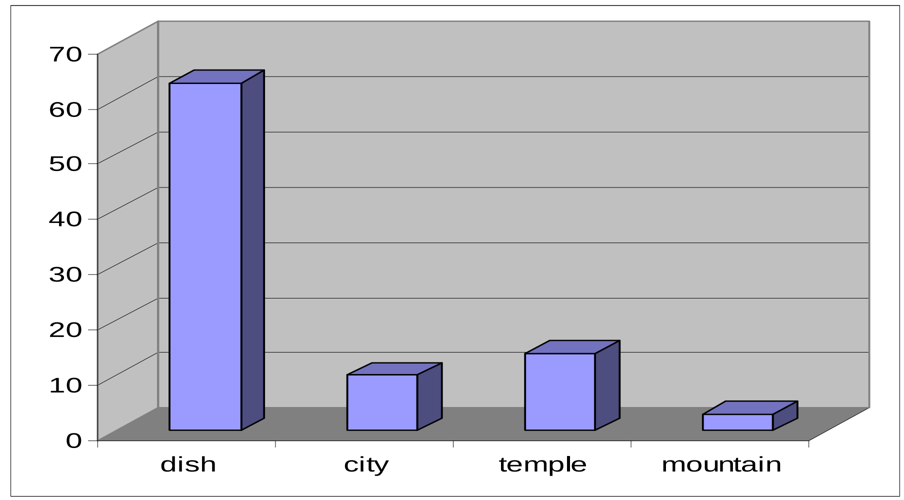
- Annotation by maximal statistical evidence

A small quiz

What is Laksa?

A: dish B: city

C: temple D: mountain



Asking Google!

- „cities such as Laksa“ 0 hits
- „dishes such as Laksa“ 10 hits
- „mountains such as Laksa“ 0 hits
- „temples such as Laksa“ 0 hits

⇒ Google knows more than all of you together!

⇒ Example of using syntactic information + statistics to derive semantic information

Patterns

- HEARST1: <CONCEPT>s such as <INSTANCE>
- HEARST2: such <CONCEPT>s as <INSTANCE>
- HEARST3: <CONCEPT>s, (especially/including) <INSTANCE>
- HEARST4: <INSTANCE> (and/or) other <CONCEPT>s

- Examples:
  - dishes such as Laksa
  - such dishes as Laksa
  - dishes, especially Laksa
  - dishes, including Laksa
  - Laksa and other dishes
  - Laksa or other dishes

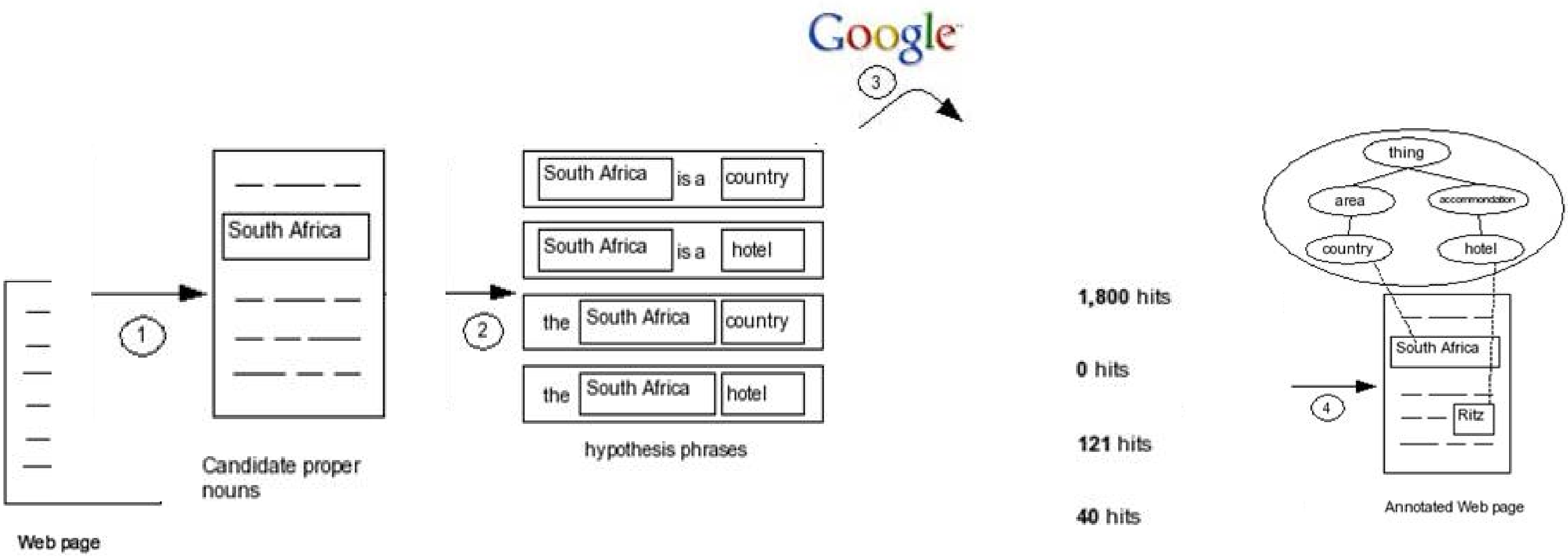
Patterns (Cont‘d)

- DEFINITE1: the <INSTANCE> <CONCEPT>
- DEFINITE2: the <CONCEPT> <INSTANCE>

- APPOSITION:<INSTANCE>, a <CONCEPT>
- COPULA: <INSTANCE> is a <CONCEPT>

- Examples:
- the Laksa dish
- the dish Laksa
- Laksa, a dish
- Laksa is a dish

PANKOW Process



Asking Google (more formally)

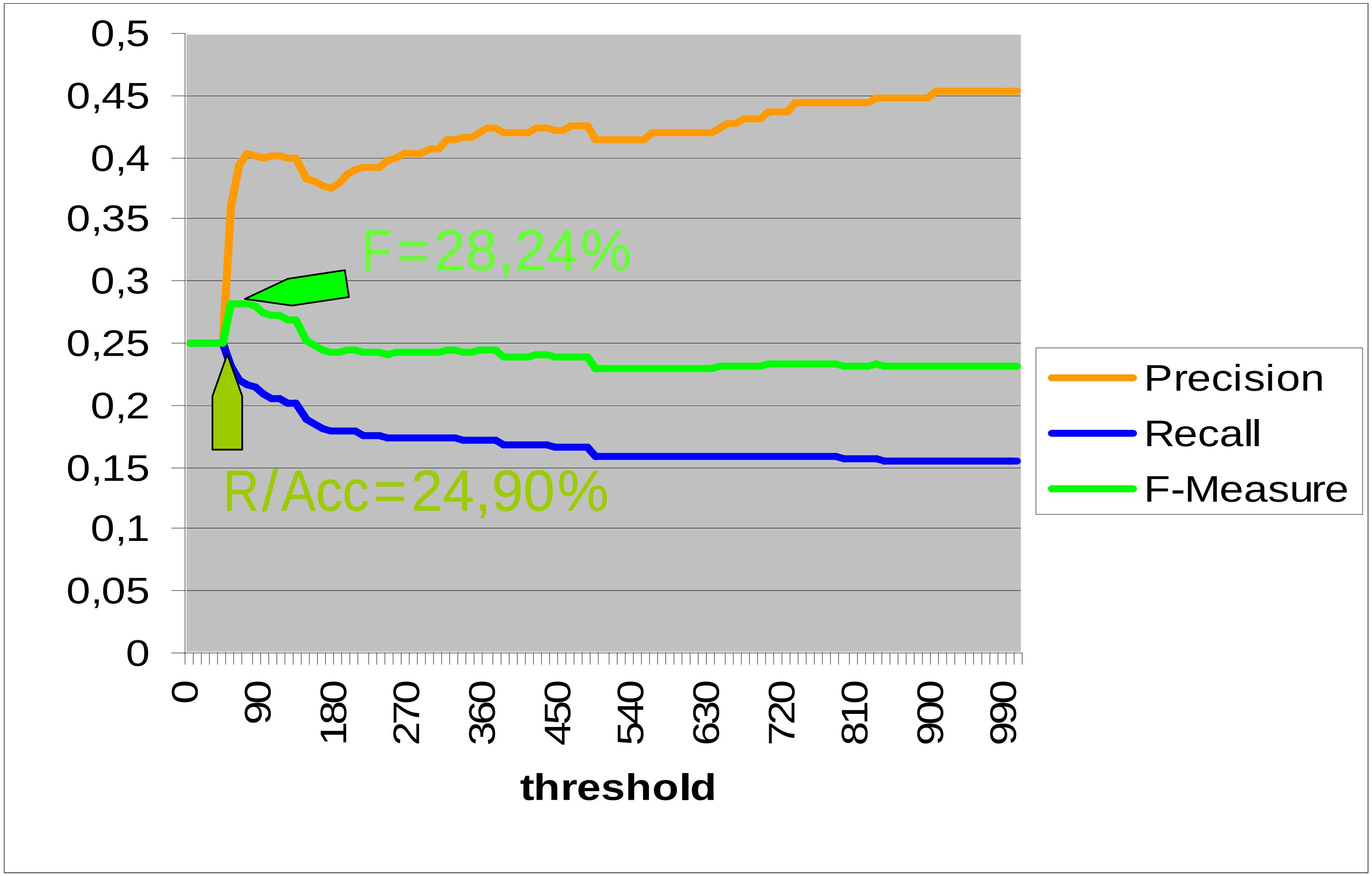
- Instance $i \in I$, concept $c \in C$, pattern $p \in$ {Hearst1,...,Copula} ***count(i,c,p)*** returns the number of Google hits of instantiated pattern

$$count(i,c) := \sum_{p} count(i,c,p)$$

- E.g. count(Laksa,dish):=count(Laksa,dish,def1)+...
- Restrict to the best ones beyond threshold θ

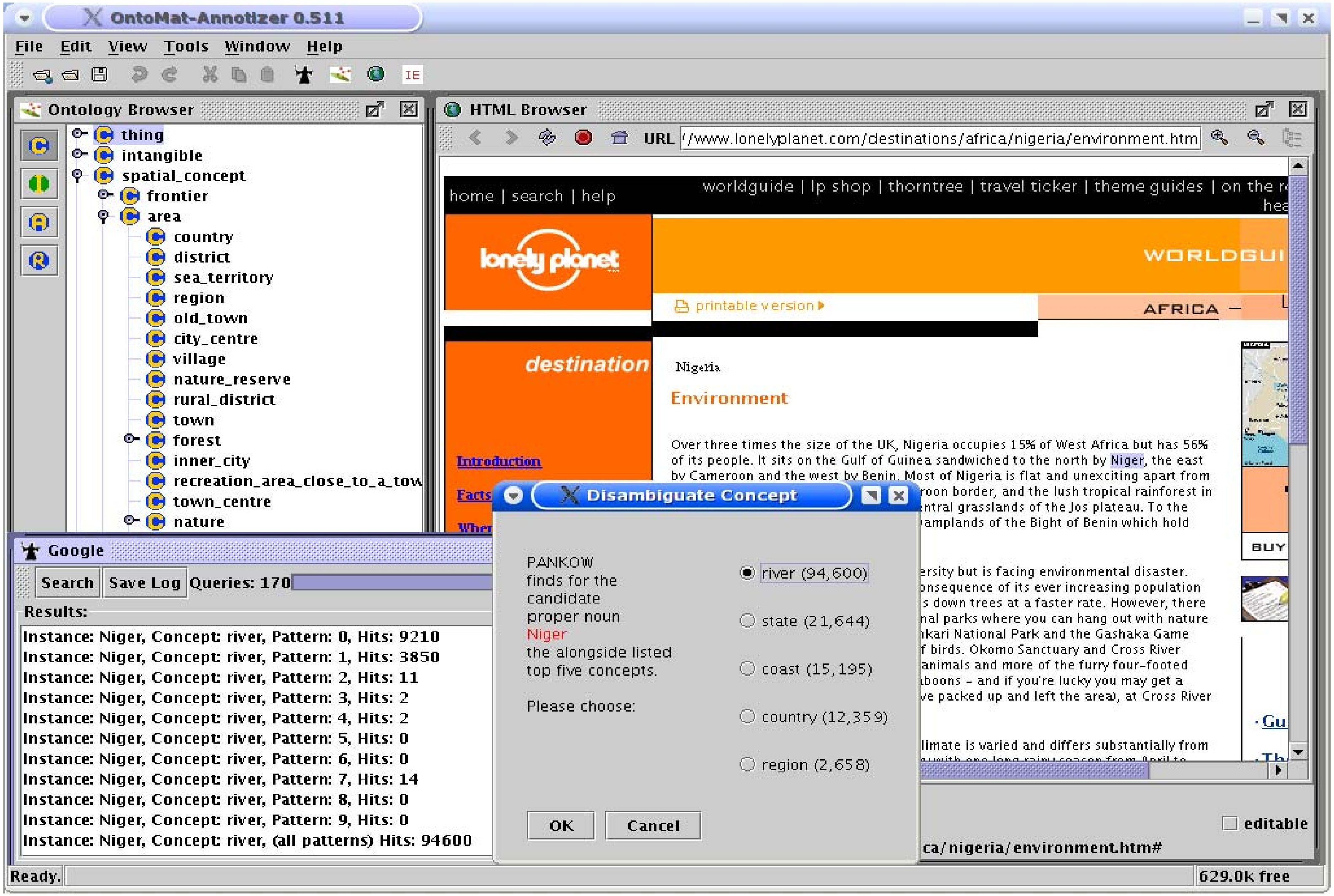
$$R_{\theta} := \left\{ (i,c_i) \mid i \in I, c_i := \arg\max_{c \in C} \left(count\,(i,c) \right) \wedge count\,(i,c) \geq \theta \right\}$$

Results

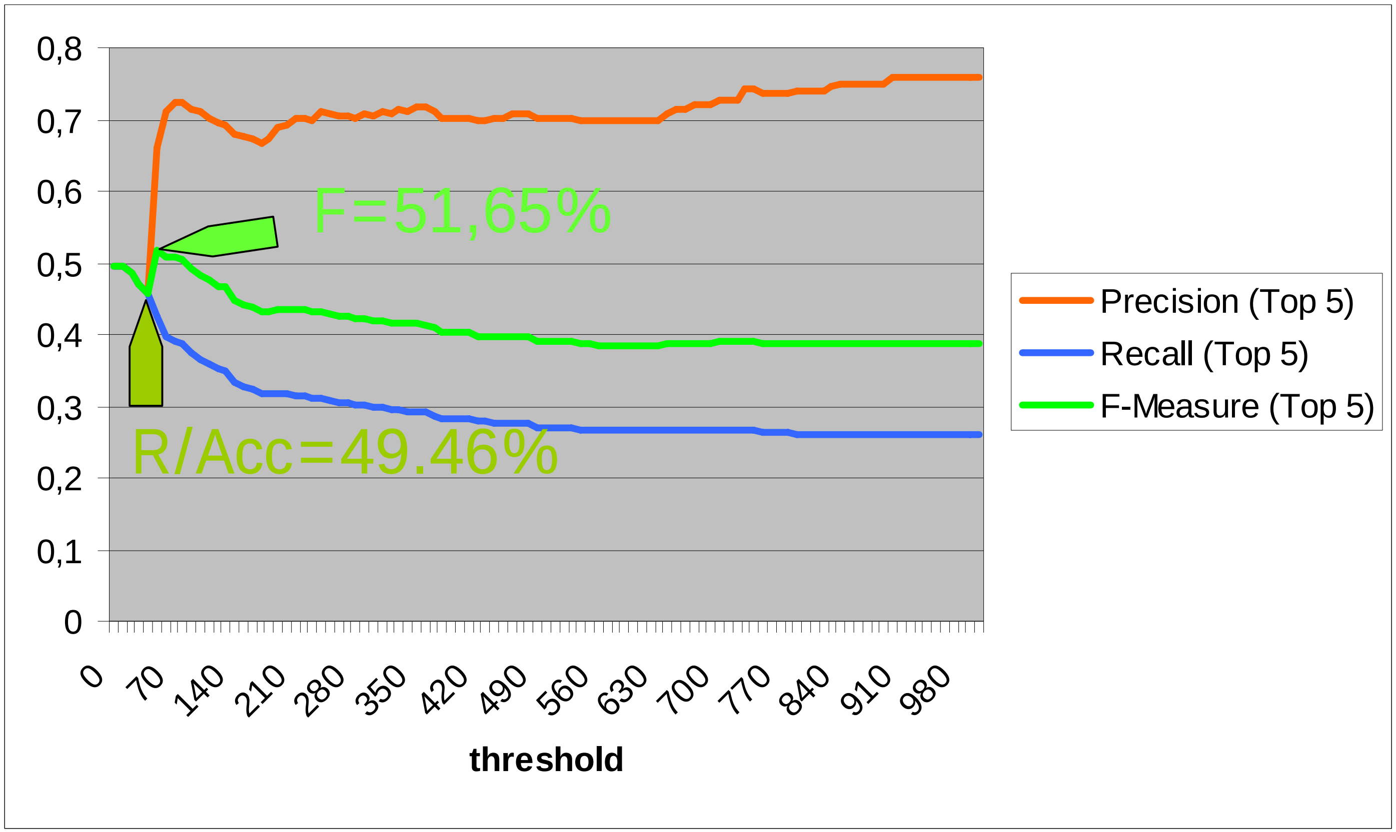


0,5
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
0
90
180
270
360
450
540
630
720
810
900
990
threshold
F=28,24%
R/Acc=24,90%
Precision
Recall
F-Measure

PANKOW & CREAM/OntoMat



Results (Interactive Mode)



Conclusion

Summary

- new paradigm to overcome the annotation problem
- unsupervised instance categorization
- first step towards the self-annotating Web
- difficult task: open domain, many categories
- decent precision, low recall
- very good results for interactive mode
- currently inefficient (590 Google queries/instance)

Challenges:

- contextual disambiguation
- annotating relations (currently restricted to instances)
- scalability (e.g. only choose reasonable queries to Google)
- accurate recognition of Named Entities (currently POS-tagger)

KnowItAll [Etzioni et al. 2004]

- KnowItAll is a search engine with the aim of `knowing it all‘
- Aims at knowing all the members of a certain class, e.g. all the actors in the world.
- It is similar in spirit to PANKOW, but can be said to work in `reverse mode´ to PANKOW
- Further, it introduces the concept of discriminators, i.e.

$$\frac{\text{Hits("John Travolta stars in")}}{\text{Hits("* stars in")}}$$

- These discriminator counts are used to train a classifier which then predicts membership to a class (e.g. the class of actors)

Information Extraction

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- **Web-based Information Extraction**
  - Instance Classification
  - **Relation Extraction**
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Relation Extraction

- Task: Given an ontological relation r as well as a set of seeds tuples S, derive patterns conveying tuples of r and derive new tuples (instances of the relation) by applying the patterns in an iterative loop

- Input: A relation r, a set of seed tuples S, e.g.

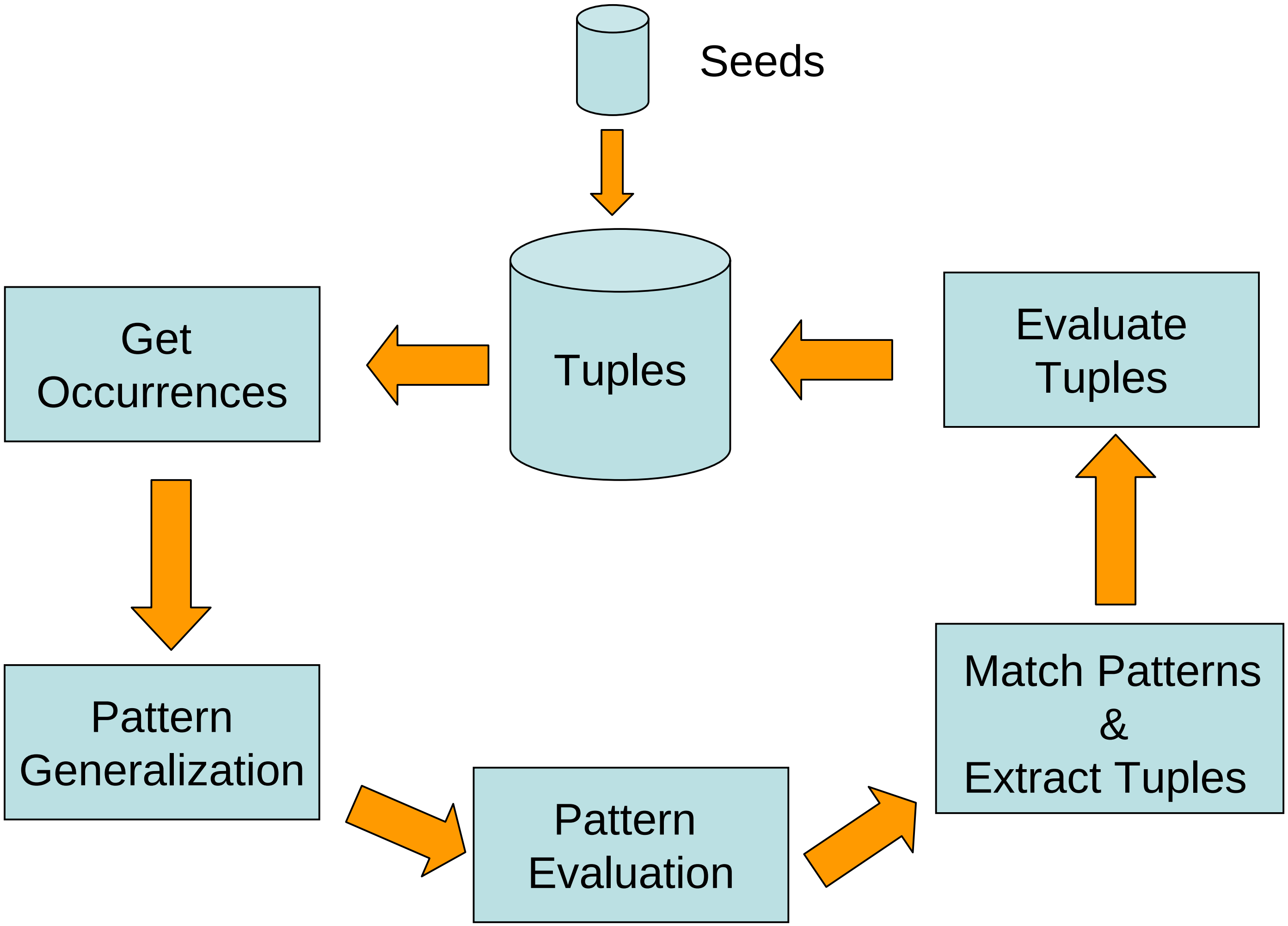
capital_of(Athens,Greece)

capital_of(Berlin,Germany)

capital_of(Madrid,Spain)

- Output: new tuples (instances of the relation r) – ideally the complete set

General Architecture



The Algorithm

```
learnTuples(Set S, Corpus C)
{
  S‘=S;
  while NOT finished
   {
      Occ = getOccurrences(S‘,C);
      P = getPatterns(Occ);
      P‘ = generalizePatterns(P);
      P‘‘ = evaluate&filter(P‘);
     S‘‘ = matchPatterns(P‘‘,C)
     S‘‘‘  =  evaluate&filter(S‘‘);
     S‘ = S‘ + S‘‘‘;
   }
}
```

Crucial Design Choices

- Problem Characterization:
  - How difficult is it to learn the relation in question ?
  - How many seed examples do we need ?
  - How many iterations ?
  - What is the precision / recall trade-off ?
- Get Occurrences:
  - What does it mean to be near each other ?
- Generalization:
  - How do we generalize patterns ?
  - One possibility: merging!
- Pattern/Tuple Evaluation:
  - How do we evaluate the patterns ?
  - How do we evaluate the tuples ?
  - Problem: we have not complete knowledge!
  - Solution: heuristics approximating the ´real´ evaluation function
- Iteration: do we keep patterns ?

Evaluation of Patterns / Tuples

- Precision/Recall: ([Agichtein and Gravano 01] - Snowball)

$$P = \frac{S''' \cap S'}{S'''}, R = \frac{S''' \cap S'}{S'}, F_1 = \frac{2PR}{P+R}$$

- PMI: ([Pantel and Penachiotti 06] - Espresso)

$$PMI(p) = \sum_{t \in T \subseteq S'} \frac{PMI(p,t)}{|T|}$$

$$PMI(p,t) = \log_2 \frac{P(t_1, p, t_2)}{P(t_1, *, t_2)\ P(*, p, *)} \approx \log \frac{|t_1, p, t_2|}{|t_1, *, t_2||*, p, *|}$$

- Evaluation of tuples: $E(t) = \sum_{p \in P'} \frac{PMI(t, p)}{|P'|}$

Open questions ?

- Which evaluation works best ?
- Does this depend on the nature of the relation considered ?
- How many patterns do we select for the matching ?
- How many tuples do we select for the next round ?

These questions are very important to ensure efficiency and effectiveness of the approach!

Web-based Information Extraction

Advantages

- relatively good results
- robustness
- Web = massive corpus (less data sparseness problems)
- search engine APIs easy to use

Disadvantages

- results dependent on the search engine (behaviour can change from one day to the other)
- trust, commercial bias of search engines
- takes al lot of time to issue queries
- ambiguity

In general: relatively new (but very promising) research field!

Information Extraction

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- Web-based Information Extraction
- **Multimedia Information Extraction**
- Merging Redundant Information – „Smushing“

Multimedia Information Extraction

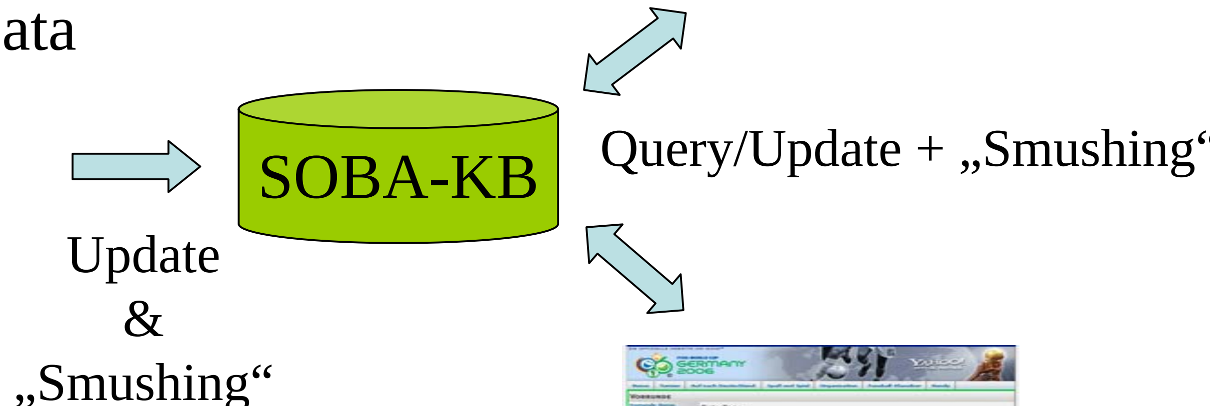
- Definition: The task here is to extract relevant information from different media types and combine them in a **reasonable** way to a **whole picture**.
- Input: Multimedia resources (images, HTML tables, text documents, videos, ...) and an ontology or template schema
- Output: A KB (with facts) representing the information extracted from the various resources, linked together in a meaningful way.
- Requires:
  - Processing different media (obvious)
  - Merging / duplicate detection
  - Detecting and handling inconsistencies

SOBA: SmartWeb Ontology-based Annotation

Goal: Generation of the SOBA-KB to support Question Answering relying on automatic semantic annotation of semi-structured data, textual reports as well as images and captions.

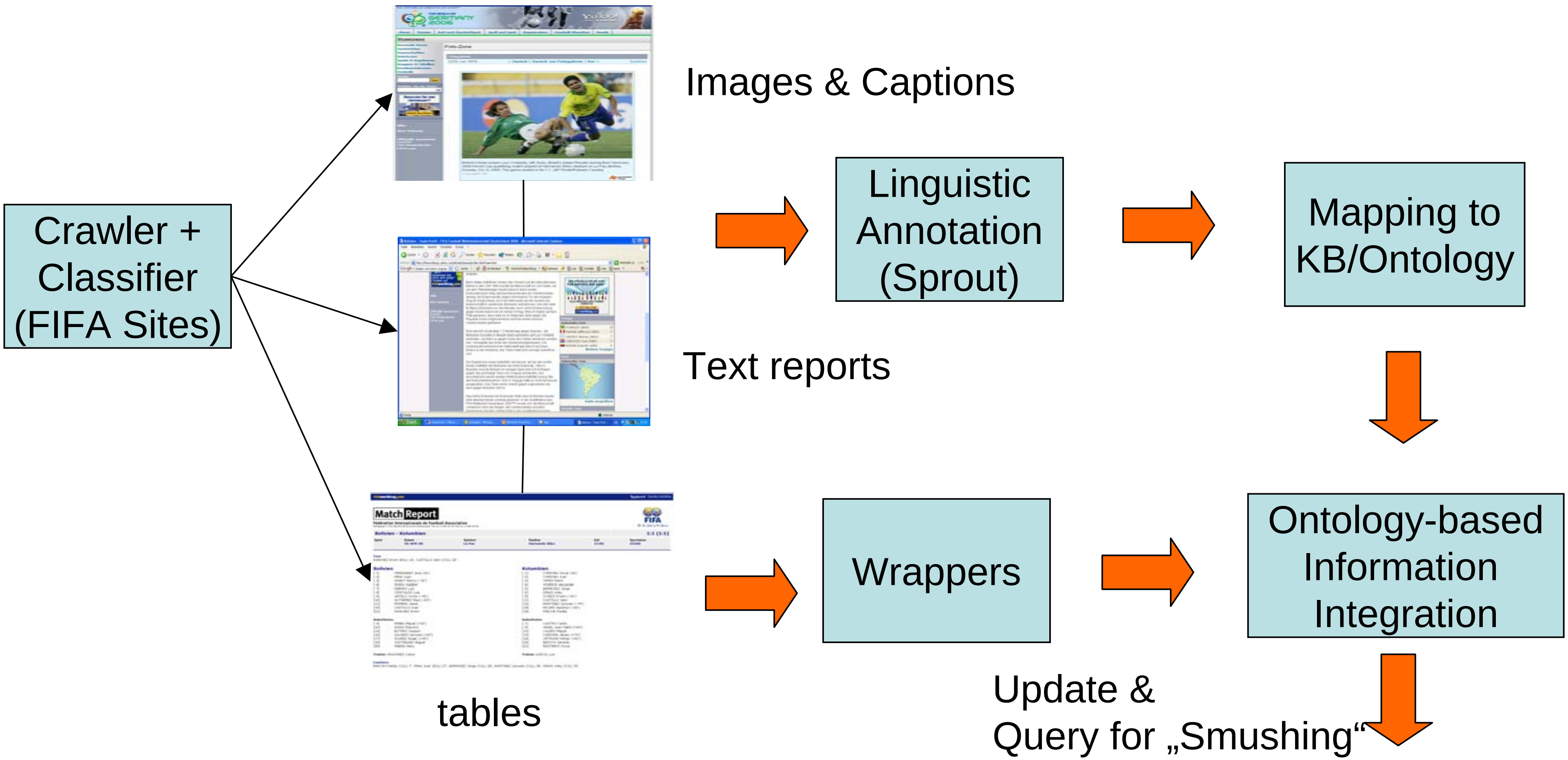
Textual Reports

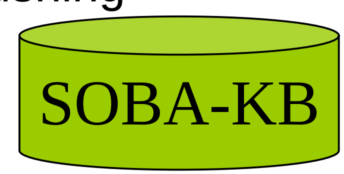
Semi-structured Data



Images and Captions

Overall SOBA Process





The Crawler downloads pages from the FIFA web site, classifies text reports and images with respect to tables, storing these references in so called „Crossref“ files.

Crossref Files

Crossref Files encapsulate all the information available about a match (text reports, tables, images)


```
<?xml version="1.0" encoding="UTF-8" ?>
<document id="cr_de_mr36712_struct" status="OK">
 <SemiStruct id="de_mr36712_struct" status="OK"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Semi_Structured/de_mr36712_struct.xml" />
 <Bodyelement id="de_050903_62g6w_0.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Text/de_050903_62g6w_0.xml" />
 <Bodyelement id="de_050903_62g6w_1.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Text/de_050903_62g6w_1.xml" />
 <Bodyelement id="de_050903_62g6w_2.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Text/de_050903_62g6w_2.xml" />
 <Bodyelem_Annotated id="de_050903_62g6w_0_struct.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Text_Annotated/de_050903_62g6w_0_struct.xml" />
 <Bodyelem_Annotated id="de_050903_62g6w_1_struct.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Text_Annotated/de_050903_62g6w_1_struct.xml" />
 <Caption id="en_i_fifa_gen_afp_20050903_i_656432235.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Caption/en_i_fifa_gen_afp_20050903_i_656432235.xml" />
 <Image id="/Images/656432235.jpg" status="new" MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Images/656432235.jpg" />
 <Caption_Annotated id="en_i_fifa_gen_afp_20050903_i_656432235_struct.xml" status="new"
  MSurl="http://localhost:8080/smartweb/media/file/local/soccer_corpus/FIFA/Caption_Annotated/en_i_fifa_gen_afp_20050903_i_656432235_struct.xml" />
</document>
```

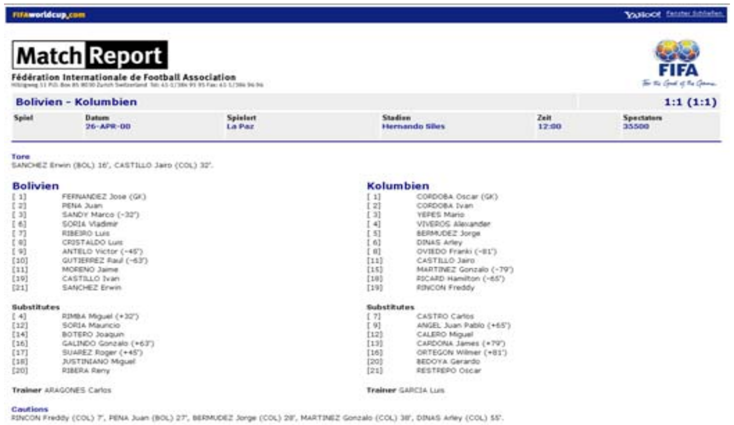
```
<?xml version="1.0" encoding="utf-8" ?>
- <document>
 - <metadata>
    <url>http://fifaworldcup.yahoo.com/06/de/050903/6/2g6w.html</url>
    <date>03. September 2005</date>
    <author>von AFP</author>
    <language>de</language>
   </metadata>
   <title>Bulgariens WM-Chancen geschrumpft - klarer Sieg der Schweden</title>
 - <body>
    <bodyelem id="de_050903_62g6w_0">Bulgarien kann eine Teilnahme bei den Weltmeisterschaften 2006 in Deutschland wohl endgültig abschreiben. Das Team
     von Bundesliga-Legionär Dimitar Berbatow hatte bei der 0:3 (0:0)-Niederlage gegen Schweden in Stockholm dem nicht den Hauch einer Chance.</bodyelem>
   </body>
</document>
```

```
?xml version="1.0" encoding="utf-8" ?>
document name="mr36712">
 <MetaData>
   <url>mr36712</url>
   <language>de</language>
 </MetaData>
 <MatchInfo>
 - <Team1>
     <Name>Schweden</Name>
     <FinalScoreResult>3</FinalScoreResult>
   </Team1>
 - <Team2>
     <Name>Bulgarien</Name>
     <FinalScoreResult>0</FinalScoreResult>
   </Team2>
   <FinalResult>3:0</FinalResult>
   <IntermediateResult>0:0</IntermediateResult>
   <HeldOn>03. September 2005 17:15</HeldOn>
   <HeldAt>STOCKHOLM Rasunda</HeldAt>
   <Attendance>35000</Attendance>
   <Referee>DE BLEECKERE Frank BEL</Referee>
 </MatchInfo>
 <Goals>
 - <Goal>
     <Scorer>LJUNGBERG Fredrik</Scorer>
     <Team>SWE</Team>
     <Minute>60</Minute>
     <CurrentScore>1:0</CurrentScore>
   </Goal>
 - <Goal>
     <Scorer>MELLBERG Olof</Scorer>
     <Team>SWE</Team>
     <Minute>75</Minute>
     <CurrentScore>2:0</CurrentScore>
   </Goal>
 - <Goal>
     <Scorer>IBRAHIMOVIC Zlatan</Scorer>
     <Team>SWE</Team>
     <Minute>90</Minute>
     <CurrentScore>3:0</CurrentScore>
   </Goal>
```



```
<?xml version="1.0" encoding="UTF-8" ?>
- <document>
   <URL>http://fifaworldcup.yahoo.com/06/de/photos/index.html?aid=114234&d=1</URL>
   <IMGURL>http://us.news1.yimg.com/us.yimg.com/i/fifa/gen/afp/20050903/i/656432235.jpg</IMGURL>
   <lang>en</lang>
   <bodyelement>Sweden's Zlatan Ibrahimovic fights for the ball with Bulgarian players Georgi Ivanov and Valentin Iliev during their 2006 World Cup qualifying match
    in Solna, 03 September 2005. AFP PHOTO/SOREN ANDERSSON</bodyelement>
</document>
```

Processing semi-structured data



HTML Wrapper

XML aligned to SWIntO

XML -> Flogic/RDF Conversion

Flogic/RDF

```
semistruct#Uruguay_vs_Bolivien_29_Maerz_2000_19:30:sportevent#LeagueFootballMatch
[
externalRepresentation@(de) ->> "Uruguay vs. Bolivien (29. Maerz 2000 19:30)";
dolce#"HAPPENS-AT" -> semistruct#"29. Maerz 2000 19:30_interval";
sportevent#heldIn -> semistruct#"Montevideo_Centenario_29_Maerz_2000_19_30_Stadium";
sportevent#team1Result -> 1;
sportevent#team2Result -> 0;
sportevent#attendance ->49811;
sportevent#team1 -> semistruct#"Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Uruguay_MatchTeam";
sportevent#team2 -> semistruct#"Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Bolivien_MatchTeam";
      (…)
]
semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Bolivien_MatchTeam:sportevent#FootballMatchTeam
[
       externalRepresentation@(de) ->> "Bolivien";
sportevent#name -> "Bolivien";
sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Jose_FERNANDEZ_PFP";
sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Juan_PENA_PFP";
sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Marco_SANDY_PFP";
sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Vladimir_SORIA_PFP";
sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Luis_RIBEIRO_PFP";
sportevent#lineup -> semistruct# Uruguay_vs_Bolivien_29_Maerz_2000_19:30_Luis_CRISTALDO_PFP";
       (...)
]
semistruct#"Uruguay_vs_Bolovien_29_Maerz_2000_19 :30_Luis_CRISTALDO_PFP":sportevent#FieldMatchFootb
          allPlayer
[
externalRepresentation@(de) ->> "Luis CRISTALDO (8)";
sportevent#number -> 8;
sportevent#impersonatedBy -> semistruct#"Luis_CRISTALDO"
].
semistruct#"Luis_CRISTALDO":dolce#"natural-person"
[
externalRepresentation@(de) ->> "Luis CRISTALDO";
dolce#"HAS-DENOMINATION" -> semistruct#"Luis_CRISTALDO_NaturalPersonDenomination"
].

semistruct#"Luis_CRISTALDO_NaturalPersonDenomination":dolce#"natural-person-denomination"
[
externalRepresentation@(de) ->> "Luis CRISTALDO";
dolce#LASTNAME -> "CRISTALDO";
dolce#FIRSTNAME -> "Luis
]
```

Semi-structured Data (Tables)

- Wrappers transform HTML tables containing basic information about matches into a XML representation.
- This XML representation is then mapped to appropriate KB structures.
- These table provide basis information about a match:
  - Basic information such as time, location (stadium), attendance, etc.
  - Name of the teams, name of the players of each team with their numbers
  - Goals together with the name of the scorer and minute
  - Yellow cards and red cards with the name of the players they were assigned
  - Semi-structured Data are crucial for SOBA:
- Represent a source of correct and basic information about each match
- Provide a background w.r.t. to interpret the text reports

Processing textual reports

Linguistic Annotation of texts with SProUT (output is SWIntO-aligned XML)


```
?xml version="1.0" encoding="UTF-8" ?>
SPROUTPUT>
 <metadata>
 <DISJ id="DI0">
 + <MATCHINFO cend="10" cstart="10" end="2" id="MI0" rule="amount" start="2">
 </DISJ>
 <DISJ id="DI1">
 - <MATCHINFO cend="63" cstart="33" end="12" id="MI1" rule="game_result_team_won_2" start="6">
   - <FS type="sprout_rule">
     + <F name="IN">
     - <F name="OUT">
       - <FS type="s_match">
         - <F name="CONFIDENCE">
             <FS type="90" />
           </F>
         - <F name="DESCRIPTOR">
             <FS type="string" />
           </F>
         - <F name="PREPOSITIONS">
             <FS type="*list*" />
           </F>
         - <F name="CEND">
             <FS type="string" />
           </F>
         - <F name="CSTART">
             <FS type="string" />
           </F>
         - <F name="SURFACE">
             <FS type="string" />
           </F>
         - <F name="VARIANT">
             <FS type="*top*" />
           </F>
         - <F name="OFFICIALS">
             <FS type="s_footballperson" />
           </F>
         - <F name="MATCHSTATISTICS">
             <FS type="string" />
           </F>
```


XML2FLogic
(semantic Integration)

```
semistruct#Uruguay_vs_Bolivien_29_Maerz_2000_19:30
[
  sportevent#matchEvents -> soba#ID11
].

soba#ID11:sportevent#Ban
[
 sportevent#commitedBy -> semistruct#Uruguay_vs_Bolivivien_(…)_Luis_CRISTALDO_PFP
].
```

Linguistic Annotation

For linguistic annotation of textual reports, SmartWeb relies on the Sprout system which:

- is part of the DFKI Heart-Of-Gold Architecture, providing a platform for grammar development,
- is a rule-based system relying on finite-state as well as unification technology to annotate text with entities specified in type a hierarchy
- has been extended in the SmartWeb project to recognize and annotated soccer-specific entities (matches, players, results, etc.)
- provides feature structures as output, e.g.

$$\begin{bmatrix} \text{Type: „PlayerAction"} \\ \text{SportActionType: „Goal"} \\ \text{CommittedBy: } \begin{bmatrix} \text{ImpersonatedBy: } \begin{bmatrix} \text{Firstname: „Michael"} \\ \text{Surname: „Ballack"} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Mapping from Feature Structures to F-Logic / RDF

Development of a declarative XML representation of the rules to transform the feature structures into KB structures, e.g.

Query the knowledge base to find out whether there is already a player in the KB (for this match) with the firstname Var1 and the surname Var2 and point to this entity as scorer!

```
<type orig="PlayerAction" target="sportevent#ScoreGoal">
  <condition attribute="SportActionType" value=„Goal">
    <output method="sportevent#matchEvents"id="http://smartweb.semanticweb.org/ontolog
   ...
      <subcase>
       <input>
         <arg orig="CommittedBy:ImpersonatedBy:First" target="VAR1"/>
         <arg orig=„CommittedBy:ImpersonatedBy:Last" target="VAR2"/>
       </input>
       <output method="sportevent#committedBy" value="q(FORALL Z <- EXIS
        ($MATCH["http://smartweb.semanticweb.org/ontology/sportevent"#te
          ...anticweb.org/ontology/sportevent";#team2->>Y]) AND
          ...eb.org/ontology/sportevent"#lineup
          ...eb.org/ontology/sportevent"#hasUpper
          ...eb.org/ontology/sportevent"#imperso
          ...eb.org/ontology/smartdolce"#"HAS-DE
          ...eb.org/ontology/smartdolce"#FIRSTNA
          ...eb.org/ontology/smartdolce"#LASTNAME
      </subcase>
      <subcase>
       <input>
         ...DBY:IMPERSONATEDBY:SURNAME" target="VAR1"/>
       </input>
      ...
  </condition>
</type>
```

If the FS (feature structure) has the Type „PlayerAction" and attribute SportActionType has the value „Goal", then fire this rule, creating a KB entity of type sportevent#ScoreGoal as output, linking this as an event of the match in question. (Thus we get a link to an existing structure in the KB).

Select the values of the FS paths „CommittedBy->ImpersonatedBy->FirstName" and „CommittedBy->ImpersonatedBy -> SurName" and bind these two the variables Var1 and Var2. (These are then used in the output part).

Text Processing

Main features:

- used to extract additional facts which are not given in the semi-structured data (tables)
- Features a modularized architecture in which the mapping from linguistic structures is stored in a declarative fashion
- These mappings can thus be maintained independently of the runtime engine which applies the mappings.
- Our declarative specification of mappings can thus also be reused for other purposes or systems than SOBA
- SOBA adds new facts to the KB, paying attention to avoid creating duplicates. For this purpose, database-like „keys“ are defined for every concept to check during runtime if an corresponding entitiy already exists in the KB („smushing“)

Processing Image Captions

Linguistic Annotation

(SWIntO-aligned Feature Structures)

Sprout -> Flogic / RDF Conversion

Flogic / RDF

```
semistruct#Uruguay_vs_Bolivien_29_Maerz_2000_19:30
[
  sportevent#matchEvents -> soba#ID25
].

soba#ID25:sportevent#Foul
[
 sportevent#commitedBy ->
semistruct#Uruguay_vs_Bolivien_(…)_Luis_CRISTALDO_PFP
].

mediainst#ID67:media#Picture
[
 media#URL -> "http://fifaworldcup.yahoo.com/06/de/photos/124155.jpg";
 media#shows -> ID25
].
```

Possible Questions to the SOBA-KB

Semi-structured data:

- Who was the winner in the match between Germany and Argentina at the World cup 2006?
- Who scored a goal in the match between Italy and France in the World Cup 2006?
- Who received the most yellow cards in the World Cup 2006?
- Which German player scored the most goals in the World Cup 2006?

Textual reports:

- Who performed the most passes in the game between Germany and Costa Rica?
- Which goalkeeper saved the most shots?

Images and Captions:

- Show me an image of Michael Ballack.
- Show me images of fouls.

Conclusion: clear benefit in the extraction and combination of information contained in different media and ontology-based integration of these.

Information Extraction

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- Web-based Information Extraction
- Multimedia Information Extraction
- **Merging Redundant Information – „Smushing“**

Merging Redundancies – „Smushing“ (1)

- Motivation from the soccer domain:
  – How many goals did Ballack shoot ?

- Solution: introduce (database) keys, i.e. a goal has a match (on a certain date), a minute and a player (which identify it uniquely)

- Example from Artequakt [Kim et al. 2002]
  – System for extracting bibliographical information about artists
    - Information Extraction from Web Pages
    - Knowledge Consolidation
    - Text Generation (personalized)

Merging Redundancies „Smushing“ (2)
- Duplicate Detection -

- <u>**Problem**</u>:
  - **Rembrandt van Rijn,**
  - **Rembrandt Harmenszoon van Rijn and**
  - **Rembrandt**

  Do they refer to one and the same person ?

- <u>**Solution:**</u>
  - <u>Introduce some edit distance / similarity measure (e.g. Levensthein distance)</u>
  - <u>Check if the keys are compatible (birth date, birthplace)</u>
  - <u>Can the different entities be merged?</u>

- <u>**Merging**</u>**: Merge entities if their attributes are compatible**

- <u>**Big question**</u>**: when are their attributes compatible?**

Merging Redundancies „Smushing“ (3)

- **Consider the following examples from [Kim et al. 2002]:**
  - **Rembrandt was born in the 17th century in Leiden.**
  - **Rembrandt was born in 1606 in the Netherlands.**
  - **Rembrandt was born on July 15 1606 in Holland.**
- **<u>Conclusion</u>: - we need to consider granularity issues**
  - we need external world knowledge
- **Are these the same Philipps ?**
  - **Philipp is 176cm tall.**
  - **Philipp is 175,5 cm tall.**
  - **Philipp is 183 cm tall.**
- **<u>Conclusion</u>: we need to consider tolerable divergences for each attribute!**

Roadmap

Part I (Introduction)

Part II (Information Extraction)

- Motivation
- Classic Information Extraction
- Adaptive Information Extraction
- Web-based Information Extraction
- Multimedia Information Extraction
- Merging Redundant Information – „Smushing“

Part III (Ontology Learning)

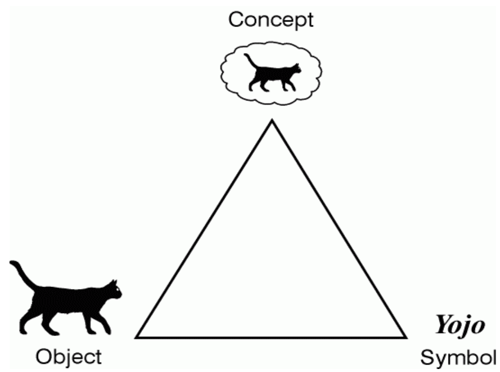
- Motivation
- Learning Concept Hierarchies
- Learning Relations

Motivation for Ontology Learning

- High cost for modeling ontologies.
- Solution: learn from existing data?
- Which data?
  - Legacy Data (XML or DB-Schema) => Lifting
  - Texts ?
  - Images ?
- In this talk we will discuss ontology learning from texts.

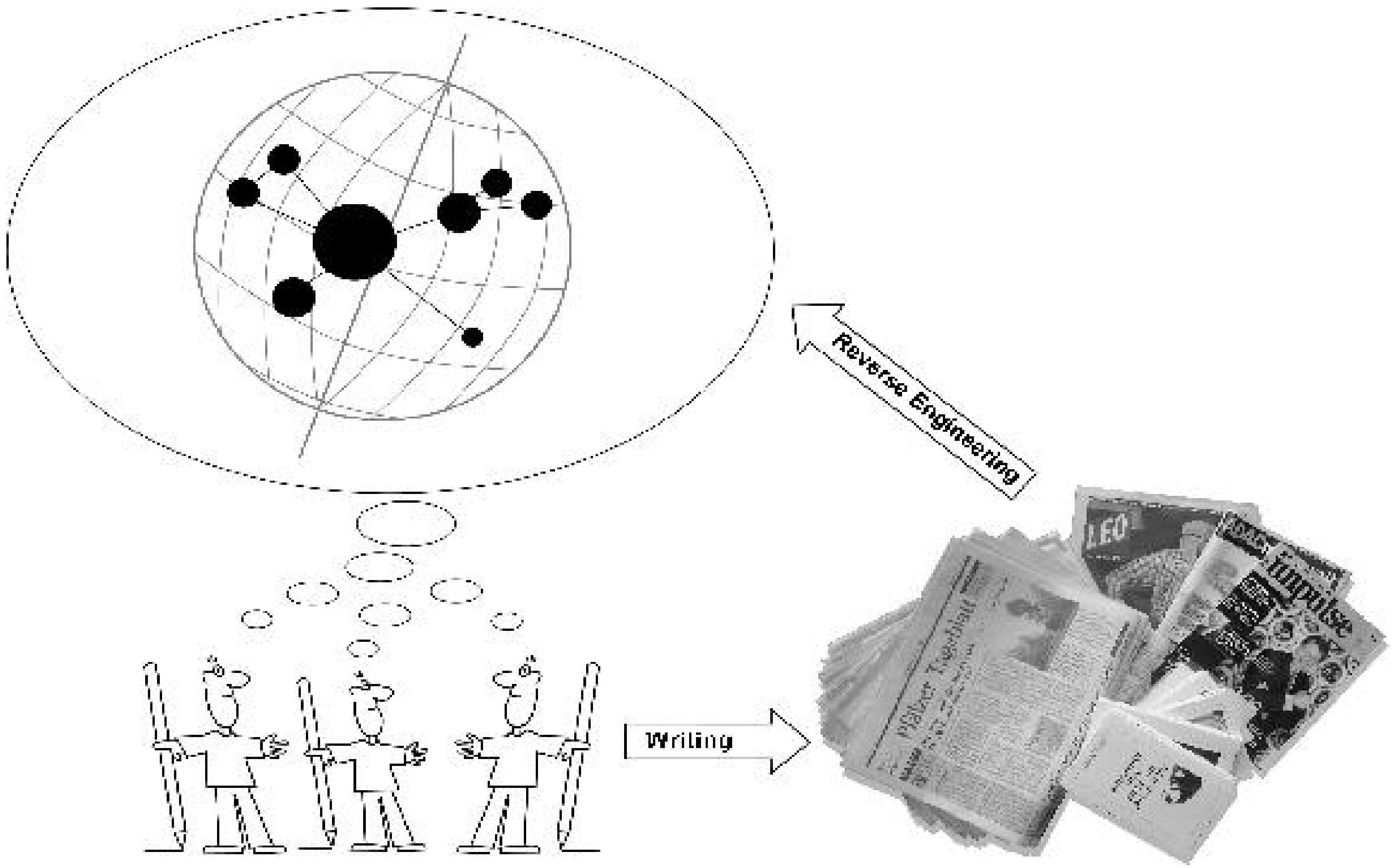
Learning ontologies from texts

- Problems:
  - Bridge the gap between symbol and concept/ontology level



  - ◆ Knowledge is rarely mentioned explicitly in texts.

OL from Text as Reverse Engineering



Ontology Learning Layer Cake

$\forall x\,(\text{country}(x) \rightarrow \exists y\ \text{capital\_of}(y, x) \wedge \forall z\,(\text{capital\_of}(z, x) \rightarrow y = z))$

$\text{disjoint}(\text{river}, \text{mountain})$

$\text{capital\_of} \leq_R \text{located\_in}$

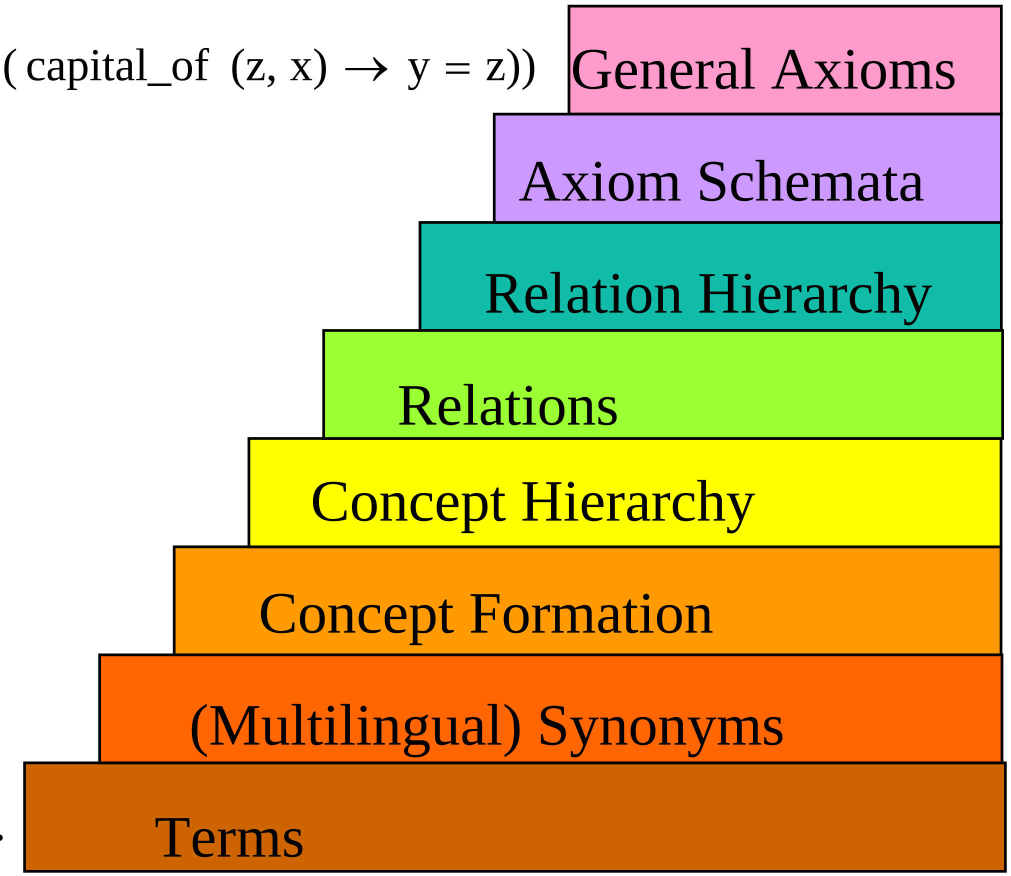
$\text{flow\_through}(\text{dom}:\text{river}, \text{range}:\text{GE})$

$\text{capital} \leq_C \text{city}, \text{city} \leq_C \text{Inhabited GE}$

$c := \text{country} := \langle i(c), \|c\|, \text{Ref}_C(c) \rangle$

$\{\text{country}, \text{nation}, \text{Land}\}$

$\text{river}, \text{country}, \text{nation}, \text{city}, \text{capital}, \ldots$



Tools

Organization	System	Ontology Learning Layers							
		Terms	*Synonyms*	*Concept Formation*	*Concept Hierarchy*	*Relations*	*Relation Hierarchy*	*Axioms Schemata*	*General Axioms*
AIFB, Univ. Karlsruhe	***Text2Onto***								
	AEON								
Amir Kabir Univ. Tehran	***HASTI***								
CNTS, Univ. Antwerpen	***OntoBasis***								
DFKI	***OntoLT / RelExt***								
Economic Univ. Prague	***TextToOnto ++***								
ISI, USC	***CBC***								
	DIRT								
Keio Univ.	***DODDLE***								
NRC-CNRC	***PMI-IR***								
Univ. de Paris-Sud	***ASIUM / Mo'k***								
Univ. di Roma	***OntoLearn***								
Univ. of Salford	***ATRACT***								
Univ. Zürich	***Parmenides***								

Ontology Learning Layer Cake

$\forall x\,(\text{country}(x) \rightarrow \exists y\ \text{capital\_of}(y, x) \wedge \forall z\,(\text{capital\_of}(z, x) \rightarrow y = z))$	General Axioms
$\text{disjoint}(\text{river}, \text{mountain})$	Axiom Schemata
$\text{capital\_of} \leq_R \text{located\_in}$	Relation Hierarchy
$\text{flow\_through}(\text{dom}: \text{river}, \text{range}: \text{GE})$	Relations
$\text{capital} \leq_C \text{city}, \text{city} \leq_C \text{Inhabited GE}$	Concept Hierarchy
$c := \text{country} := \langle i(c), \lVert c \rVert, \text{Ref}_C(c) \rangle$	Concept Formation
{country, nation, Land}	(Multilingual) Synonyms
river, country, nation, city, capital,.. .	Terms

Terms

Terms are at the basis of the ontology learning process

- Terms express more or less complex semantic units
- But what is a term?

  Huge Selection of Top Brand Computer Terminals Available for Immediate Delivery

  Because Vecmar carries such a large inventory of high-quality computer terminals, including: ADDS terminals, Boundless terminals, DEC terminals, HP terminals, IBM terminals, LINK terminals, NCR terminals and Wyse terminals, your order can often ship same day. Every computer terminal shipped to you is protected with careful packing, including thick boxes. All of our shipping options - including international - are available through major carriers.

- Extracted term candidates (phrases)

  - computer
  - terminal
  - computer terminal
  - ? high-quality computer terminal
  - ? top brand computer terminal
  - ? HP terminal, DEC terminal, …

Term Extraction

Determine most relevant phrases as terms

- Linguistic Methods
  - Rules over linguistically analyzed text
    - Linguistic analysis – Part-of-Speech Tagging, Morphological Analysis, …
    - Extract patterns – *Adjective-Noun, Noun-Noun, Adj-Noun-Noun, …*
    - Ignore *Names* (`DEC`, `HP`, …), *Certain Adjectives* (`quality`, `top`, …), etc.

- Statistical Methods
  - Co-occurrence (collocation) analysis for term extraction within the corpus
  - Comparison of frequencies between domain and general corpora
    - `Computer Terminal` will be specific to the Computer domain
    - `Dining Table` will be less specific to the Computer domain

- Hybrid Methods
  - Linguistic rules to extract term candidates
  - Statistical (pre- or post-) filtering

Statistical Analysis

Scores used in Term Extraction:

- MI (Mutual Information) – Cooccurrence Analysis
- TFIDF – Term Weighting

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

- χ^2 (Chi-square) – Cooccurrence Analysis & Term Weighting

$$X^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

- Other
  - c-value/nc-value (Frantzi & Ananiadou, 1999)
    - Considers length (c-value) and context (nc-value) of terms
  - Domain Relevance & Domain Consensus (Navigli and Velardi, 2004)
    - Considers term distribution within (DC) and between (DR) corpora

Term Extraction

Use some statistical measure to assess term relevance, e.g. tf.idf:

$$tfidf\ (w) = tf\ .\log(\frac{N}{df\ (w)})$$

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

tf(w) *term frequency (number of word occurrences in a document)*

df(w) *document frequency (number of documents containing the word)*

N *number of all documents*

tfIdf(w) *relative importance of the word in the document*

C- / NC-value ([Frantzi and Ananiadou 1999])

- Combination of:
  - C-value (indicator for termhood)
  - NC-value (contextual indicators for termhood)

$$\mathrm{C-value}(a) = \begin{cases} \log_2 |a|\, f(a) \text{ if a is not nested} \\ \log_2 |a| \left(f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b) \right) \end{cases}$$

f(a) is the frequency of a, T_a is the set of terms which contain a.

- C-value (frequency-based method sensitive to multi-word terms)

C- / NC-value

- NC-value (incorporation of information from context words indicating termhood)

$$\text{weight}(w) = \frac{t(w)}{n}$$

where $t(w)$ is the number of times that w appears in the context of a term.

- C-/NC-value

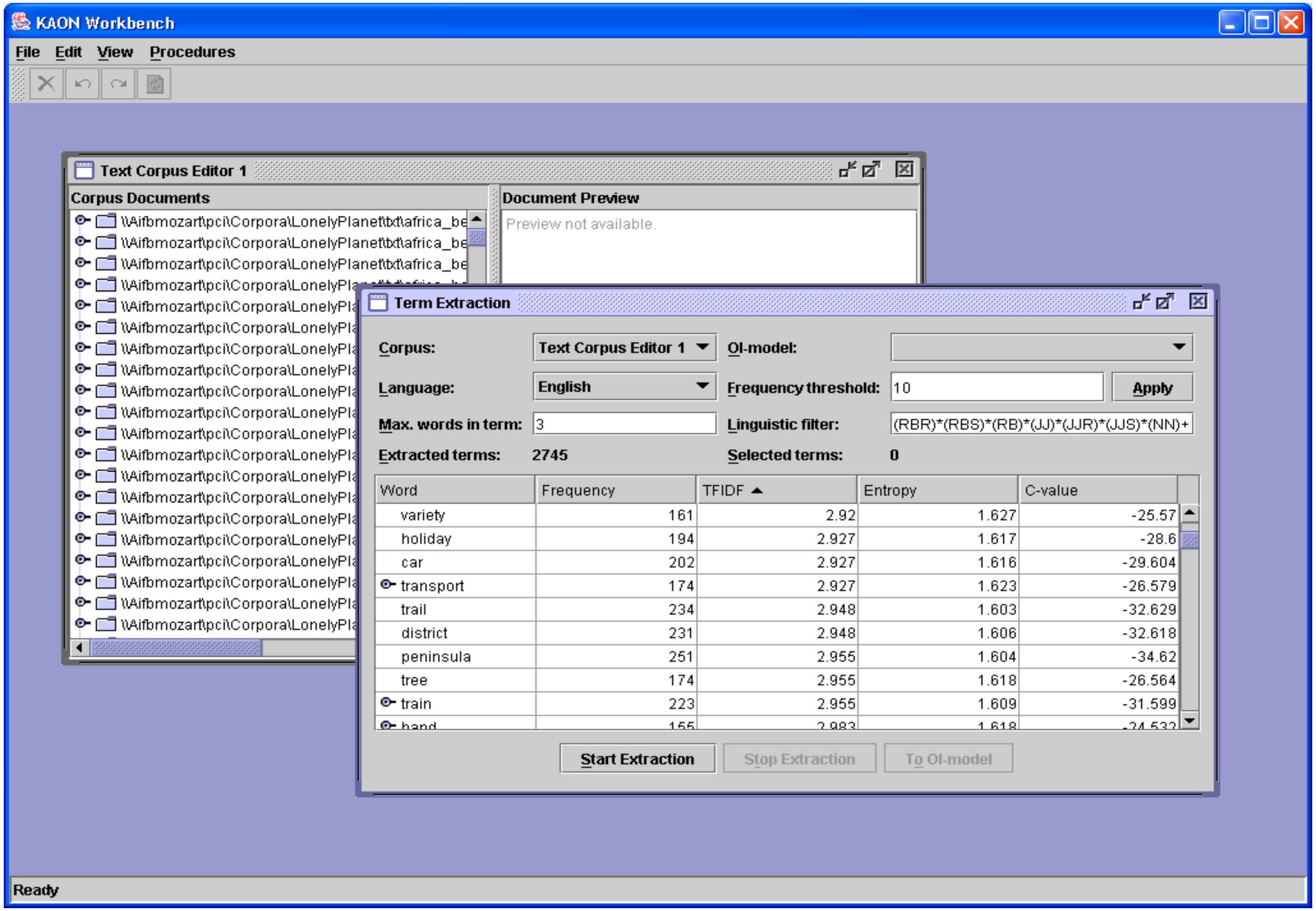
$$\text{NC-value}(a) = 0.8\,\text{C-value}(a) + 0.2 \sum_{b \in C_a} f_a(b)\,\text{weight}(b)$$

where C_a is the set of different words appearing in the context of a, $f_a(b)$ is the frequency of b in the context of a.

Terms – Tools

Organization	System	Ontology Learning Layers							
		Terms	*Synonyms*	*Concept Formation*	*Concept Hierarchy*	*Relations*	*Relation Hierarchy*	*Axioms Schemata*	*General Axioms*
AIFB, Univ. Karlsruhe	***Text2Onto***	**X**							
	AEON								
Amir Kabir Univ. Tehran	***HASTI***	**X**							
CNTS, Univ. Antwerpen	***OntoBasis***								
DFKI	***OntoLT / RelExt***	**X**							
Economic Univ. Prague	***TextToOnto ++***								
ISI, USC	***CBC***								
	DIRT								
Keio Univ.	***DODDLE***								
NRC-CNRC	***PMI-IR***								
Univ. de Paris-Sud	***ASIUM / Mo‘k***								
Univ. di Roma	***OntoLearn***	**X**							
Univ. of Salford	***ATRACT***	**X**							
Univ. Zürich	***Parmenides***	**X**							

TextToOnto



KAON Workbench
File Edit View Procedures
Text Corpus Editor 1
Corpus Documents
Document Preview
Preview not available.
\\Aifbmozart\pci\Corpora\LonelyPlanet\txt\africa_be
Term Extraction
Corpus: Text Corpus Editor 1
OI-model:
Language: English
Frequency threshold: 10
Apply
Max. words in term: 3
Linguistic filter: (RBR)*(RBS)*(RB)*(JJ)*(JJR)*(JJS)*(NN)+
Extracted terms: 2745
Selected terms: 0
Word | Frequency | TFIDF | Entropy | C-value
variety | 161 | 2.92 | 1.627 | -25.57
holiday | 194 | 2.927 | 1.617 | -28.6
car | 202 | 2.927 | 1.616 | -29.604
transport | 174 | 2.927 | 1.623 | -26.579
trail | 234 | 2.948 | 1.603 | -32.629
district | 231 | 2.948 | 1.606 | -32.618
peninsula | 251 | 2.955 | 1.604 | -34.62
tree | 174 | 2.955 | 1.618 | -26.564
train | 223 | 2.955 | 1.609 | -31.599
Start Extraction
Stop Extraction
To OI-model
Ready

Ontology Learning Layer Cake

$\forall x\,(\mathrm{country}(x) \rightarrow \exists y\ \mathrm{capital\_of}\ (y, x) \wedge \forall z\,(\mathrm{capital\_of}\ (z, x) \rightarrow y = z))$	General Axioms
$\mathrm{disjoint}(\mathrm{river}, \mathrm{mountain})$	Axiom Schemata
$\mathrm{capital\_of}\ \leq_R\ \mathrm{located\_in}$	Relation Hierarchy
$\mathrm{flow\_through}(\mathrm{dom} : \mathrm{river}, \mathrm{range} : \mathrm{GE})$	Relations
$\mathrm{capital} \leq_C \mathrm{city}, \mathrm{city} \leq_C \mathrm{Inhabited\ GE}$	Concept Hierarchy
$c := \mathrm{country} := \langle i(c), \lVert c \rVert, \mathrm{Ref}_C(c) \rangle$	Concept Formation
{country, nation, Land}	(Multilingual) Synonyms
river, country, nation, city, capital,.. .	Terms

Synonyms

- Next step in ontology learning is to identify terms that share (some) semantics, i.e., potentially refer to the same concept

- Synonyms (Within Languages)

  - ‘100% synonyms’ don’t exist – only term pairs with *similar* meanings
  - Examples from http://thesaurus.com

    - `terminal – video display – input device`
    - `graphics terminal - video display unit – screen`

- Techniques:
  - Clustering, e.g. Grefenstette
  - Significance of Co-occurrence, e.g. PMI-IR

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

Synonyms - Evaluation

- Gold Standard
  - TOEFL (Landauer – LSA: 64.45%, Turney – PMI-IR: 48-74%)
  - WordNet (problematic due to domain-independence, e.g. [Pantel and Lin 03])
    - WordNet „tuning“, e.g. [Cucchiarelli and Velardi 98], [Turcato 00], [Buitelaar and Sacaleanu 01]
- Human Evaluation
- Task-based
  - (Cross-lingual) IR/QA - e.g. Query Expansion
- Other
  - Artificial Evaluation (see [Grefenstette 94])
    - e.g. transform cell -> CELL in some contexts

Synonyms – Tools

Organization	System	Ontology Learning Layers							
		Terms	*Synonyms*	*Concept Formation*	*Concept Hierarchy*	*Relations*	*Relation Hierarchy*	*Axioms Schemata*	*General Axioms*
AIFB, Univ. Karlsruhe	***Text2Onto***	**X**	**clusters**						
	AEON								
Amir Kabir Univ. Tehran	***HASTI***	**X**							
CNTS, Univ. Antwerpen	***OntoBasis***		**clusters**						
DFKI	***OntoLT / RelExt***	**X**							
Economic Univ. Prague	***TextToOnto ++***								
ISI, USC	***CBC***		**clusters**						
	DIRT								
Keio Univ.	***DODDLE***								
NRC-CNRC	***PMI-IR***		**X**						
Univ. de Paris-Sud	***ASIUM / Mo'k***		**clusters**						
Univ. di Roma	***OntoLearn***	**X**	**X**						
Univ. of Salford	***ATRACT***	**X**	**clusters**						
Univ. Zürich	***Parmenides***	**X**							

Ontology Learning Layer Cake

$\forall x\,(\mathrm{country}(x) \rightarrow \exists y\ \mathrm{capital\_of}(y, x) \wedge \forall z\,(\mathrm{capital\_of}(z, x) \rightarrow y = z))$	General Axioms
$\mathrm{disjoint}(\mathrm{river}, \mathrm{mountain})$	Axiom Schemata
$\mathrm{capital\_of} \leq_R \mathrm{located\_in}$	Relation Hierarchy
$\mathrm{flow\_through}(\mathrm{dom}: \mathrm{river}, \mathrm{range}: \mathrm{GE})$	Relations
$\mathrm{capital} \leq_C \mathrm{city}, \mathrm{city} \leq_C \mathrm{Inhabited\ GE}$	Concept Hierarchy
$c := \mathrm{country} := \langle i(c), \lVert c \rVert, \mathrm{Ref}_C(c) \rangle$	**Concept Formation**
{country, nation, Land}	(Multilingual) Synonyms
river, country, nation, city, capital,.. .	Terms

Concepts: *Intension, Extension, Lexicon*

A term may indicate a concept, if we can define its

- Intension
  - (in)formal definition of the set of objects that this concept describes
    - *a disease is an impairment of health or a condition of abnormal functioning*

- Extension
  - a set of objects (instances) that the definition of this concept describes
    - *influenza, cancer, heart disease, …*

  <u>Discussion: what is an instance? - '*heart disease'* or '*my uncle's heart disease*'</u>

- Lexical Realizations
  - the term itself and its multilingual synonyms
    - *disease, illness, Krankheit, maladie, …*

  <u>Discussion: synonyms vs. instances – '*disease*', '*heart disease*', '*cancer*', …</u>

Concepts – *Intension*

Extraction of a Definition for a Concept from Text

- Informal Definition
  - e.g., a gloss for the concept as used in WordNet
  - *OntoLearn* (Navigli and Velardi 04; Velardi et al. 05) uses natural language generation to compositionally build up a WordNet gloss for automatically extracted concepts
    - 'Integration Strategy' : "*strategy for the integration of …*"

- Formal Definition
  - e.g., a logical form that defines all formal constraints on class membership
  - Inductive Logic Programming, Formal Concept Analysis, …

Concepts – *Extension*

Extraction of Instances for a Concept from Text

- Commonly referred to as Ontology Population
- Relates to Knowledge Markup (Semantic Metadata)
- Uses Named-Entity Recognition and Information Extraction

- Instances can be:
  - Names for objects, e.g.
    - *Person, Organization, Country, City, …*
  - Event instances (with participant and property instances), e.g.
    - *Football Match (with Teams, Players, Officials, ...)*
    - *Disease (with Patient-Name, Symptoms, Date, …)*

Concept Formation - Evaluation

- Concept Extension
  - Gold Standard
    - overlap on clusters, e.g. OntoBasis
    - overlap on set of instances w.r.t. KB (difficult)
  - Human Evaluation (e.g. OntoBasis [Reinberger et al. 2005])
  - Task Based
    - QA from KBs

- Concept Intension (in/formal definitions)
  - Gold Standard (e.g. WordNet glosses, WikiPedia)
  - Human Evaluation (e.g. WordNet glosses [Velardi et al. 05])
  - Task Based
    - Ontology Engineering
      - Understanding
      - Consistency

Concept Formation – Tools

Organization	System	Ontology Learning Layers							
		Terms	*Synonyms*	*Concept Formation*	*Concept Hierarchy*	*Relations*	*Relation Hierarchy*	*Axioms Schemata*	*General Axioms*
AIFB, Univ. Karlsruhe	***Text2Onto***	**X**	**clusters**	**int.**					
	AEON								
Amir Kabir Univ. Tehran	***HASTI***	**X**							
CNTS, Univ. Antwerpen	***OntoBasis***		**clusters**	**clusters**					
DFKI	***OntoLT / RelExt***	**X**							
Economic Univ. Prague	***TextToOnto ++***								
ISI, USC	***CBC***		**clusters**	**clusters**					
	DIRT								
Keio Univ.	***DODDLE***								
NRC-CNRC	***PMI-IR***		**X**						
Univ. de Paris-Sud	***ASIUM / Mo‘k***		**clusters**	**clusters**					
Univ. di Roma	***OntoLearn***	**X**	**X**	**int.**					
Univ. of Salford	***ATRACT***	**X**	**clusters**	**clusters**					
Univ. Zürich	***Parmenides***	**X**							

Ontology Learning Layer Cake

$\forall x\,(\text{country}(x) \rightarrow \exists y\ \text{capital\_of}(y, x) \wedge \forall z\,(\text{capital\_of}(z, x) \rightarrow y = z))$

$\text{disjoint}(\text{river}, \text{mountain})$

$\text{capital\_of} \leq_R \text{located\_in}$

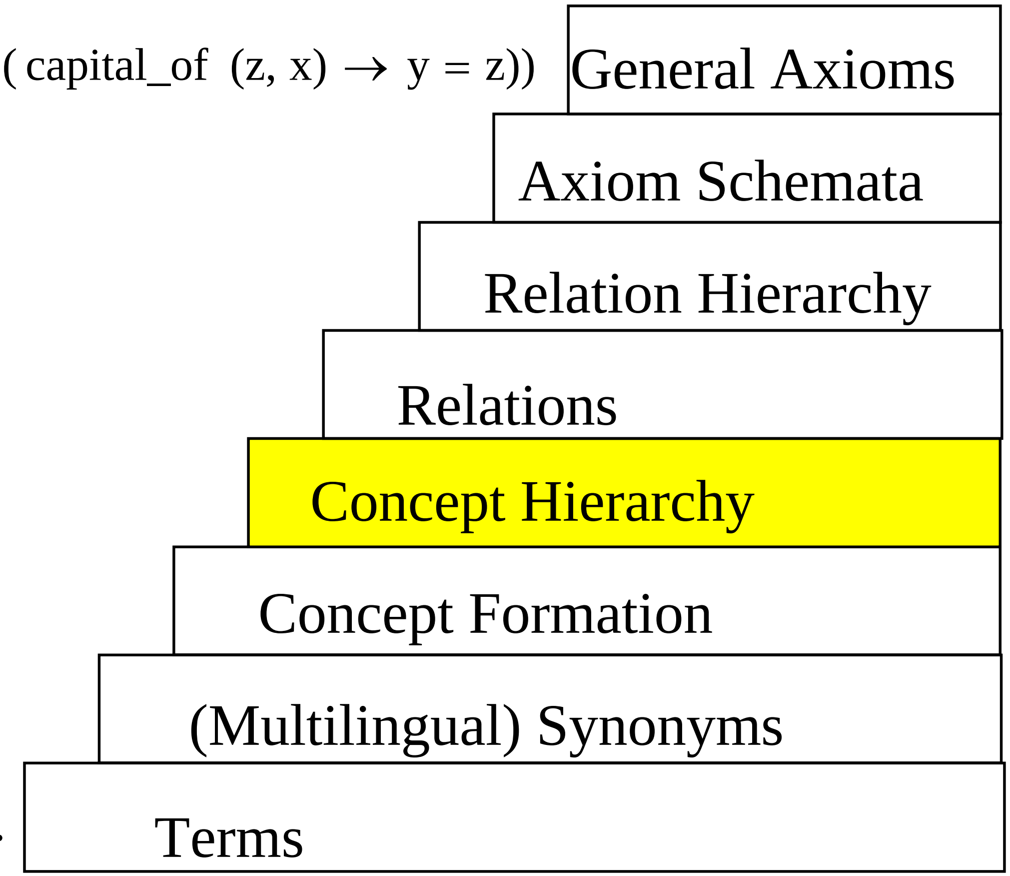
$\text{flow\_through}(\text{dom} : \text{river}, \text{range} : \text{GE})$

$\text{capital} \leq_C \text{city}, \text{city} \leq_C \text{Inhabited GE}$

$c := \text{country} := \langle i(c), \|c\|, \text{Ref}_C(c) \rangle$

$\{\text{country}, \text{nation}, \text{Land}\}$

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



Taxonomy Extraction - Overview

- **Lexico-syntactic patterns**
- Distributional Similarity & Clustering
- Linguistic Approaches
- Taxonomy Extension/Refinement
- Combination of Methods
- Evaluation
- Tools Matrix

Hearst Patterns [Hearst 1992]

Patterns to extract a relation of interest fullfilling the following requirements:

- They should occur frequently and in many text genres.
- They should accurately indicate the relation of interest.
- They should be recognizable with little or no pre-encoded knowledge.

Acquiring Hearst Patterns

Hearst also suggests a procedure in order to acquire such patterns from a corpus:

1. Decide on a lexical relation R of interest, e.g. hyponymy/hypernymy.
2. Gather a list of terms for which this relation is known to hold, e.g. hyponym(car, vehicle). This list can be found automatically using the Hearst patterns or by bootstrapping from an existing lexicon or knowledge base.
3. Find places in the corpus where these expressions occur syntactically near one another.
4. Find the commonalities and generalize the expressions in 3. to yield patterns that indicate the relation of interest.
5. Once a new pattern has been identified, gather more instances of the target relation and go to step 3.

Hearst Patterns - Examples

- Examples for hyponymy patterns:
  - *Vehicles **such as** cars, trucks and bikes*
  - ***Such** fruits **as** oranges, nectarines or apples*
  - *Swimming, running **and other** activities*
  - *Publications, **especially** papers and books*
  - *A seabass **is** a fish.*

Hearst Patterns (Continued)

- Use regular expression defined over syntactic categories:
  - *NP* ***such as*** *NP, NP, ... and NP*
  - ***Such*** *NP* ***as*** *NP, NP, ... or NP*
  - *NP, NP, ...* ***and other*** *NP*
  - *NP,* ***especially*** *NP, NP ,... and NP*
  - *NP* ***is*** *a NP.*
  - *...*

- Precision wrt. Wordnet: 55,46% (66/119) on the basis of New York Times corpus
  - [Cederberg and Widdows 03] report lower results: 40%

Taxonomy Extraction - Overview

- Lexico-syntactic patterns
- **Distributional Similarity & Clustering**
- Linguistic Approaches
- Taxonomy Extension/Refinement
- Combination of Methods
- Evaluation
- Tools Matrix

What does the X stand for?

„X is very nice."

„In X it is always sunny."

„We usually spend our holidays at X."

- We observe that we can group words which appear at certain contexts.
- For this purpose we need to represent the context of words.

Distributional Hypothesis & Vector Space Model

- Harris, 1986
  - *„Words are (semantically) similar to the extent to which they share similar words"*
- Firth, 1957
  - *„You shall know a word by the company it keeps"*

- Idea: collect context information and represent it as a vector:

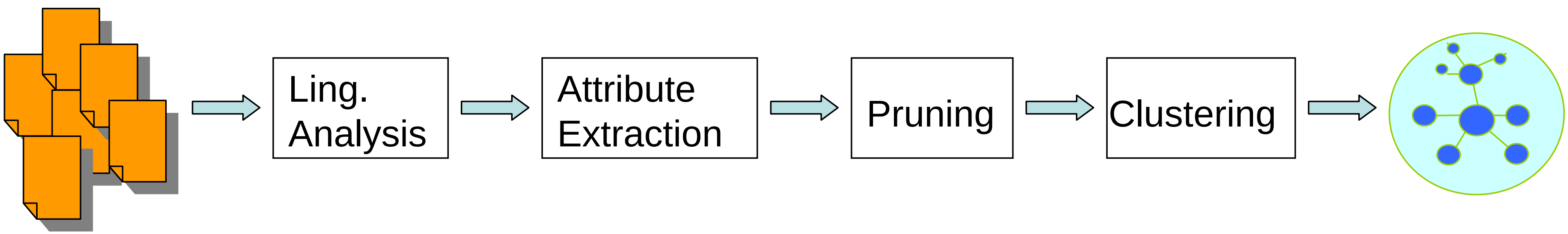
	book_obj	rent_obj	drive_obj	ride_obj	join_obj
apartment	X	X			
car	X	X	X		
motor-bike	X	X	X	X	
excursion	X				X
trip	X				X

- compute similarity among vectors wrt. a measure

Context Features

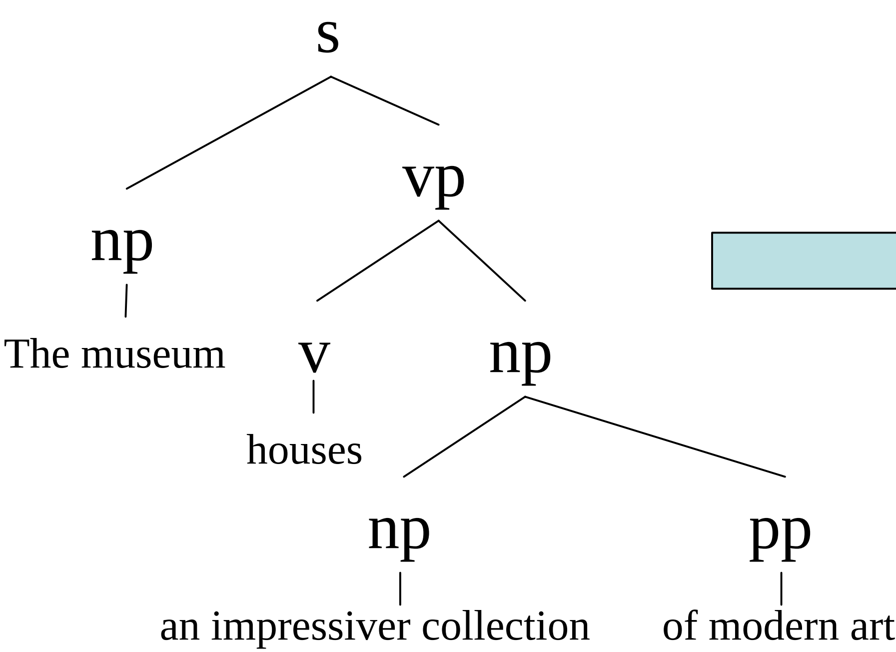
- **Four-grams** [Schuetze 93]
- **Word-windows** [Grefenstette 92]
- **Predicate-Argument relations** (SUBJ/OBJ/COMPLEMENT)
  Modifier Relations (*fast car, the hood of the car*)
  - [Grefenstette 92, Cimiano 04b, Gasperin et al. 03]
- **Appositions** (*Ferrari, the fastest car in the world*)
  - [Caraballo 99]
- **Coordination** (*ladies and gentlemen*)
  - [Caraballo 99, Dorow and Widdows 03]

Overall Process for Clustering Concept Hierarchies



Extracting contextual features

*The museum **houses** an impressive collection of medieval and modern art. The building **combines** geometric abstraction **with** classical references that **allude to** the Roman influence on the region.*



house_subj(museum)

house_obj(collection)

combine_subj(museum)

combine_obj(abstraction)

combine_with(reference)

allude_to(influence)

Pseudo-syntactic Dependencies

The museum ***houses*** *an* ***impressive*** *collection of medieval and* ***modern*** *art. The building* ***combines geometric*** *abstraction* ***with classical*** *references that* ***allude to*** *the Roman* ***influence on the region****.*

NP + verb + NP -> verb_subj / verb_obj

house_subj(museum) **house_obj(museum)** **combine_subj(museum)** **combine_obj(abstraction)** **combine_with(reference)**	+	**impressive(collection)** **geometric(abstraction)** **combine_with(reference)** **classical(reference)** **allude_to(influence)** **roman(influence)** **influence_on(region)** **on_region(influence)**

Weighting Measures

$$\mathrm{Conditional}(n, feat) = P(n \mid feat) = \frac{f(n, feat)}{f(feat)}$$

$$\mathrm{PMI}(n, feat) = \log \frac{P(n \mid feat)}{P(n)}$$

$$\mathrm{Resnik}(n, feat) = S_R(feat)\, P(n \mid feat)$$

$$\text{where } S_n(feat) = \sum_{n'} P(n' \mid feat) \log \frac{P(n' \mid feat)}{P(n')}$$

Clustering Concept Hierarchies from Text

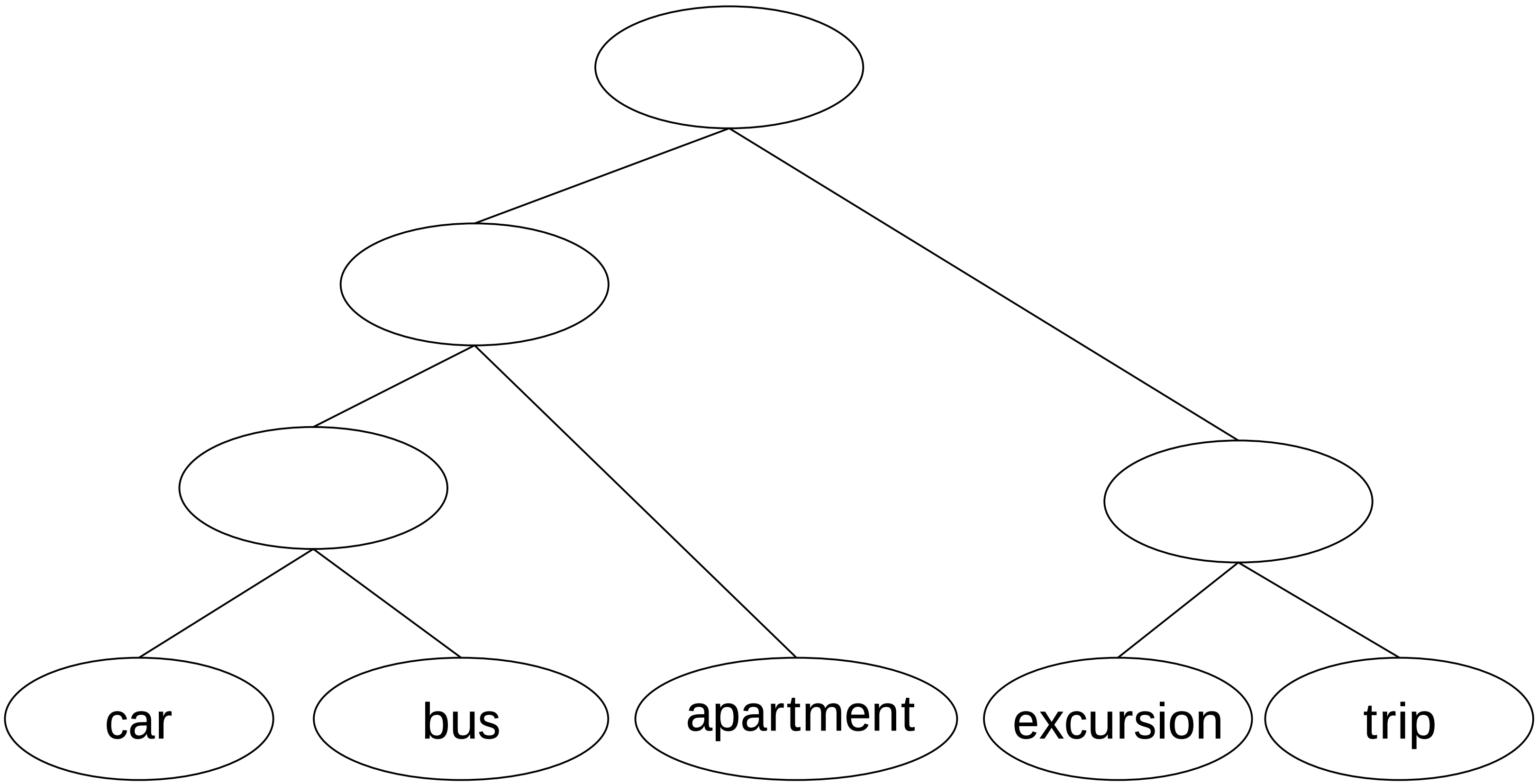
- **Similarity-based**
- Set-theoretical
- Soft clustering

Similarity-based Clustering

- Similarity Measures:
  - Binary (Jaccard, Dine)
  - Geometric (Cosine, Euclidean/Manhattan distance)
  - Information-theoretic (Relative Entropy, Mutual Information)
  - (…)

- Methods:
  - Hierarchical agglomerative clustering
  - Hierarchical top-down clustering, e.g. Bi-Section KMeans
  - (…)

Hierarchical Agglomerative Clustering



Hierarchical Agglomerative Clustering - Algorithm -

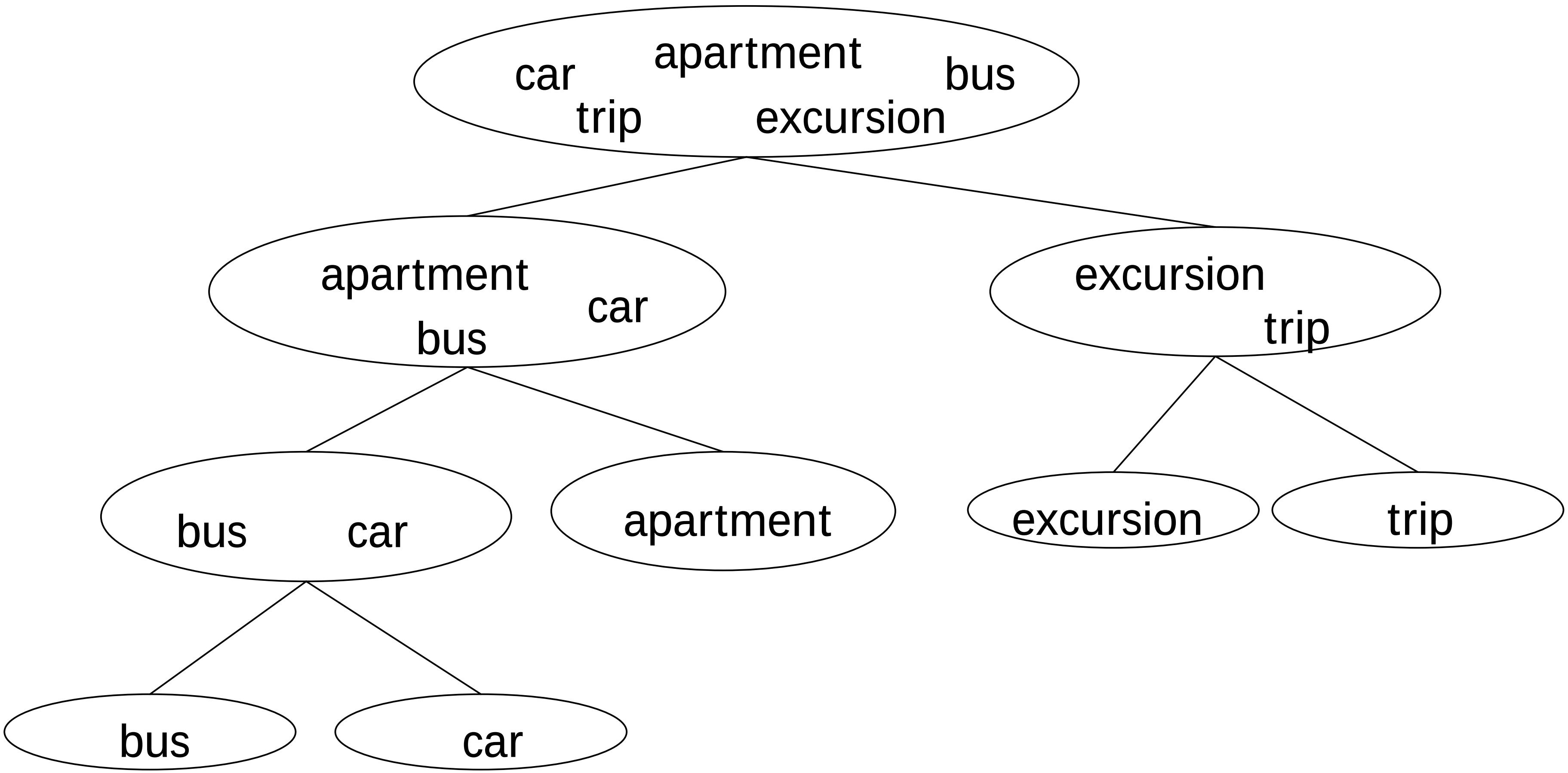
Algorithm 4 Hierarchical Agglomerative (Bottom-Up) Clustering

Input: a set $X = \{x_1, ..., x_n\}$ of objects represented by vectors $\mathbf{x_1}, ..., \mathbf{x_n} \in \mathbb{R}^m$ and a similarity function sim: $\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$

Output: a set K of $2n-1$ clusters ordered hierarchically as a binary tree (K, E) with $2(n-1)$ edges and n leaves

$\forall i\ 1 \leq i \leq n : \ k_i := \{x_i\}$
$K := K' := \{k_1, ..., k_n\}$
$E := \emptyset$
$j := n+1$
while$(|K'| > 1)$ **do**
  $(k_{u'}, k_{v'}) := \text{argmax}_{(k_u, k_v) \in K' \times K'}\ sim(k_u, k_v)$
  $k_j = k_{u'} \cup k_{v'}$
  $K' := K' \setminus \{k_{u'}\}$
  $K' := K' \setminus \{k_{v'}\}$
  $K' := K' \cup \{k_j\}$
  $K := K \cup \{k_j\}$
  $E = E \cup \{(k_{u'}, k_j), (k_{v'}, k_j)\}$
  j:=j+1
end while
return (K,E)

Bi-Section-KMeans



Bi-Section-Kmeans
- Algorithm -

Algorithm 6 Bi-Section KMeans

Input: a set $X = \{x_1, ...x_n\}$ of objects represented by vectors $\mathbf{x_1}, ..., \mathbf{x_n} \in \mathbb{R}^m$ and
a function coh : $2^{\mathbb{R}^m} \to \mathbb{R}$
a function for computing the centroid of a cluster, i.e., $\mu : 2^{\mathbb{R}^m} \to \mathbb{R}^m$

Output: a set K of clusters with $|K| = 2n - 1$ ordered hierarchically
as binary tree (K, E) with $2(n-1)$ edges and n leaves

$K = K' := \{X\}$
$E := \emptyset$
for i=1 to n-1 **do**
  choose the largest or the least coherent cluster $k_u \in K'$, i.e.
  $k_u = \text{argmax}_{k_i \in K'} |k_i|$ or $k_u = \text{argmin}_{k_i \in K'} \text{coh}(k_i)$
  choose two data points f_1 and f_2 of k_u as cluster centroids
  repeat
    assign each element in k_u to its closest centroid, i.e.
    $c_1 := \{x \in k_u \mid dist(x, f_1) \leq dist(x, f_2)\}$
    $c_2 := \{x \in k_u \mid dist(x, f_2) \leq dist(x, f_1)\}$
    recompute both centroids, i.e.
    $f_j = \mu(c_j), j \in \{1, 2\}$
  until stopping criterion is true
  $K' := K' \setminus \{k_u\} \cup \{k_1, k_2\}$
  $K := K \cup \{k_1, k_2\}$
  $E := E \cup \{(k_1, k_u), (k_2, k_u)\}$
end for
return (K, E)

Clustering Concept Hierarchies

- Similarity-based
- **Set Theoretical**
- Soft clustering

Formal Concept Analysis (FCA) [Ganter and Wille 1999]

- method used for the analysis of data

  => structure data into units (abstract concepts)

- A triple (G,M,I) is called a **formal context** if G and M are sets and $I \subseteq G \times M$ is a binary relation between G and M. The elements in G are called **objects**, those in M **attributes** and I the **incidence** of the context.

FCA in a Nutshell

- For $A \subseteq G$ and for $B \subseteq M$ we define:

$$A' = \{m \in M \mid (g,m) \in I \ \forall g \in A\}$$

$$B' = \{g \in G \mid (g,m) \in I \ \forall m \in B\}$$

- A pair (A,B) is a **formal concept** of (G,M,I)
- if and only if

$$A \subseteq G, B \subseteq M, A' = B \wedge A = B'$$

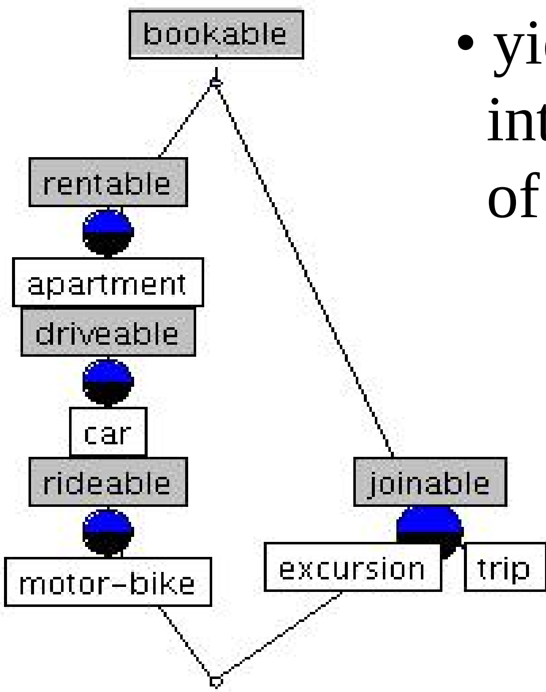
- Concepts are ordered by the **subconcept-superconcept** relation:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_2 \subseteq B_1)$$

FCA Example: Tourism Matrix

	book	rent	drive	ride	join
appartment	X	X			
car	X	X	X		
motor-bike	X	X	X	X	
excursion	X				X
trip	X				X

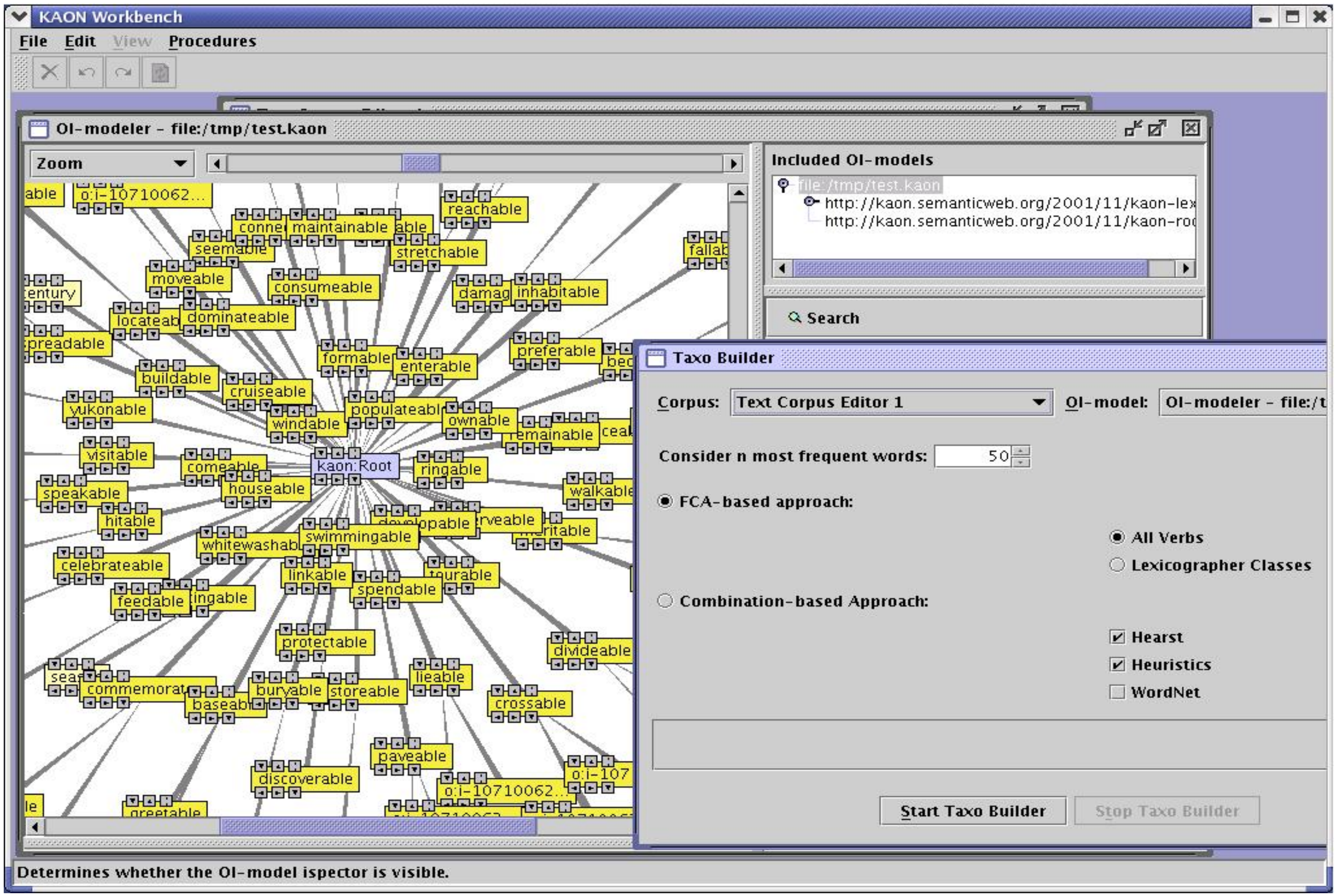
Formal Concept Analysis [Ganter, Wille 1999]



• finds ‚closed‘ sets of attributes and objects (Formal Concepts)

• yields a hierarchy with a formal interpretation in terms of subsumption of attributes

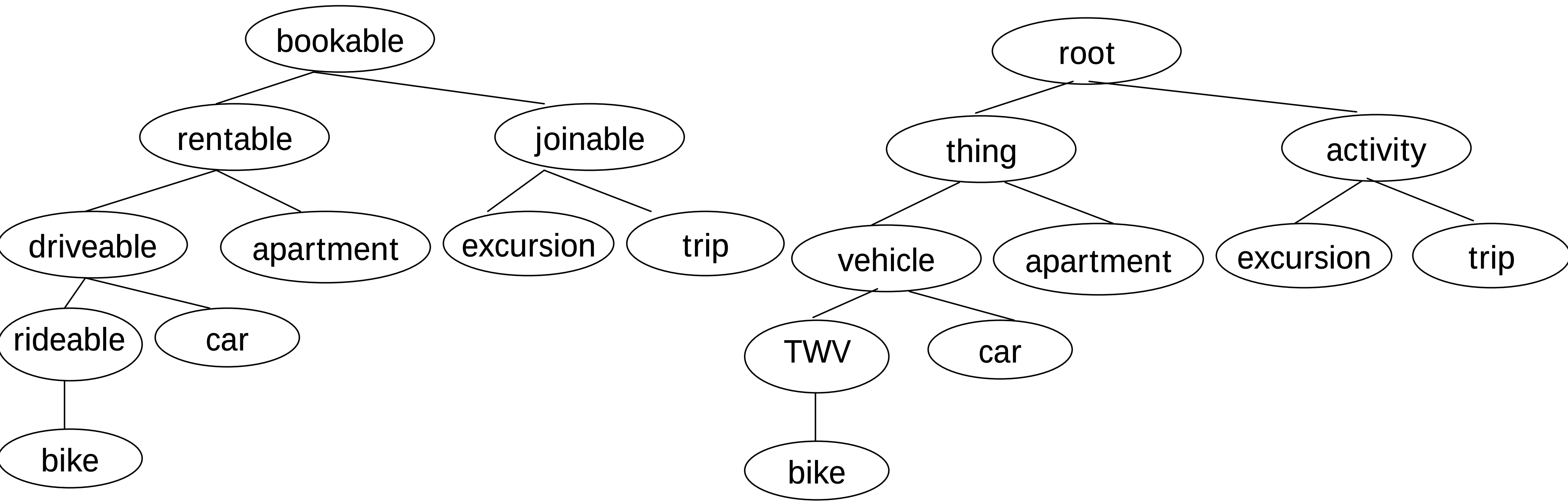
TextToOnto & FCA



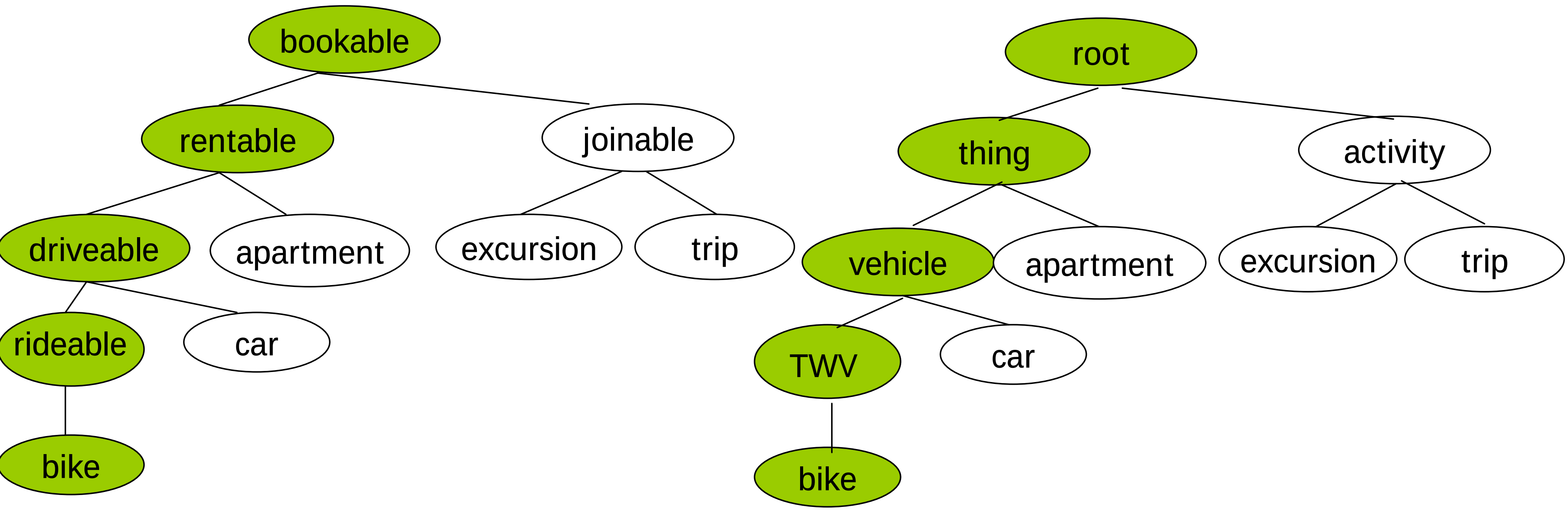
Evaluation

- Evaluation with respect to existing ontologies for a certain domain (tourism and finance)
- Quantitative comparison of agglomerative, divisive and conceptual clustering (FCA)
- Qualitative comparison: understandability, efficiency

Comparison of Hierarchies



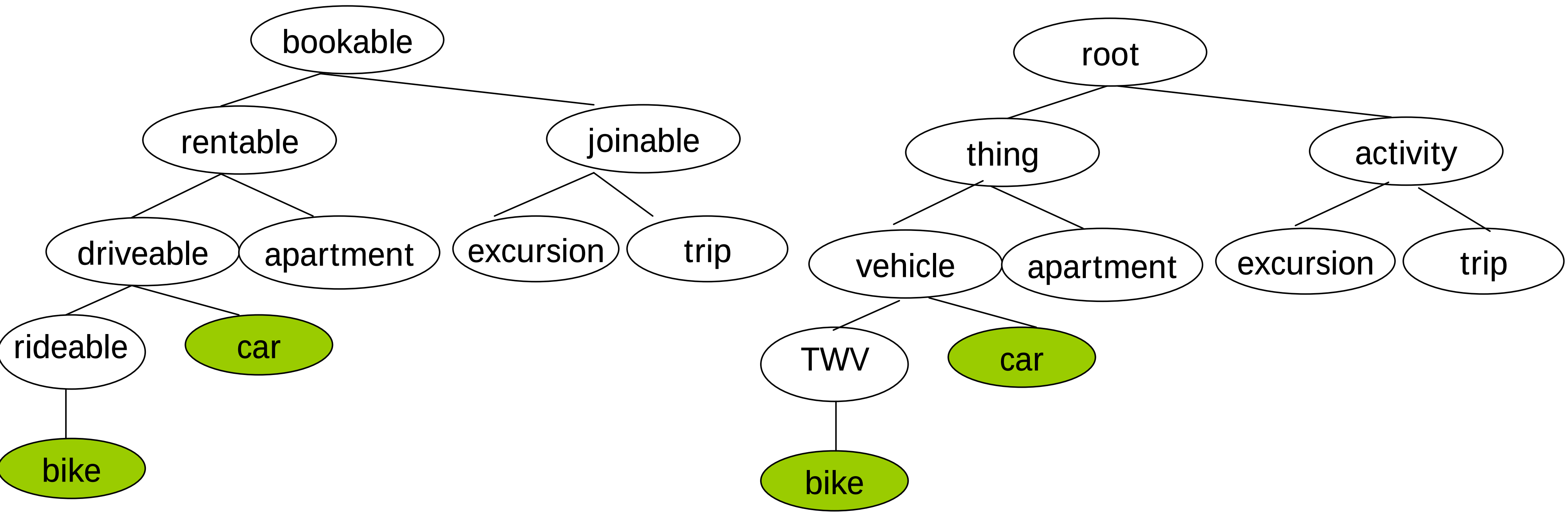
Semantic Cotopy [Maedche & Staab 02]



SC(bike)={bike,rideable,driveable,rentable,bookable}   SC(bike)={bike,TWV,vehicle,thing,root}

=> TO(bike,O_1,O_2)=1/9!!!

Common Semantic Cotopy (SC´)

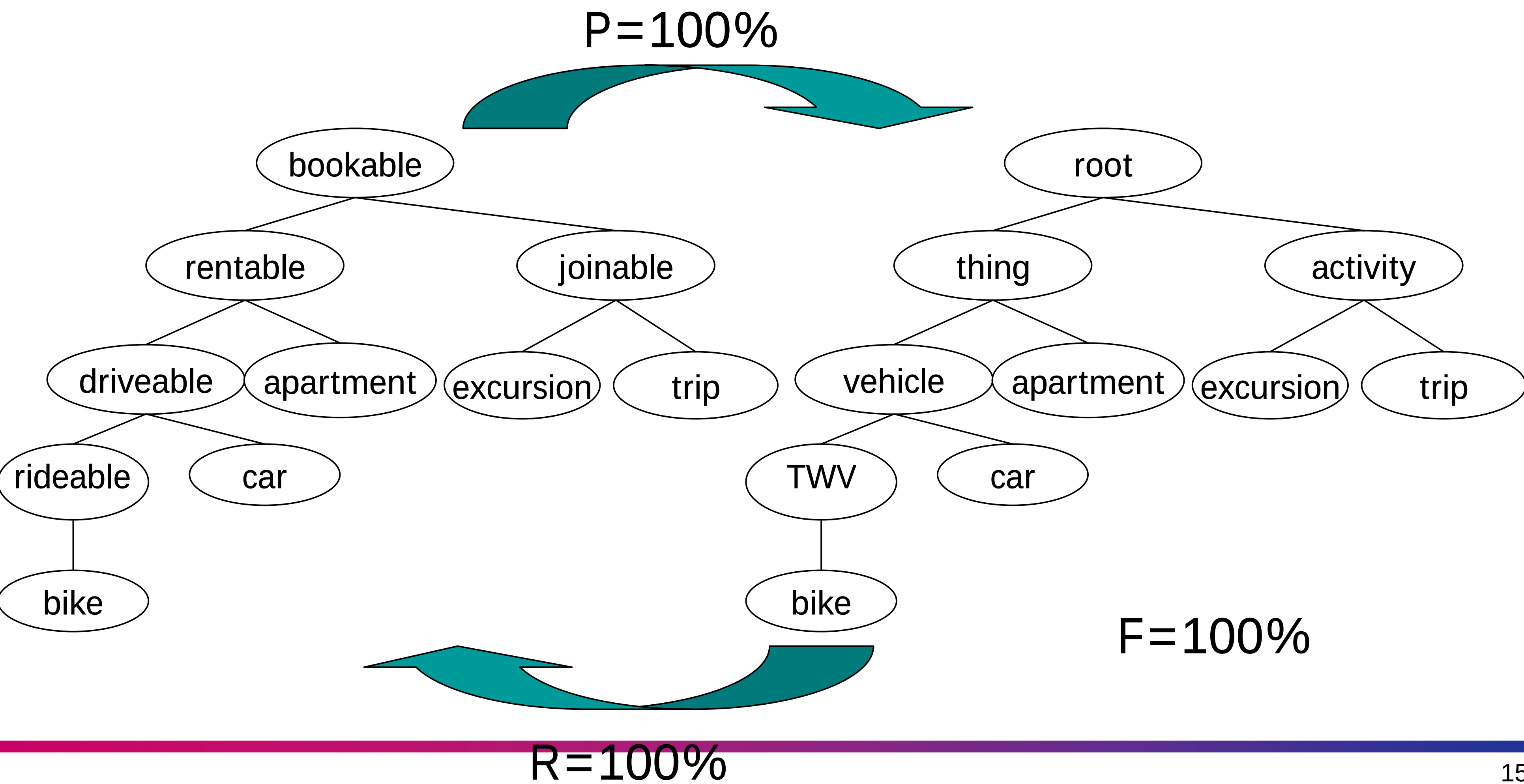


SC‘(driveable)={bike,car}

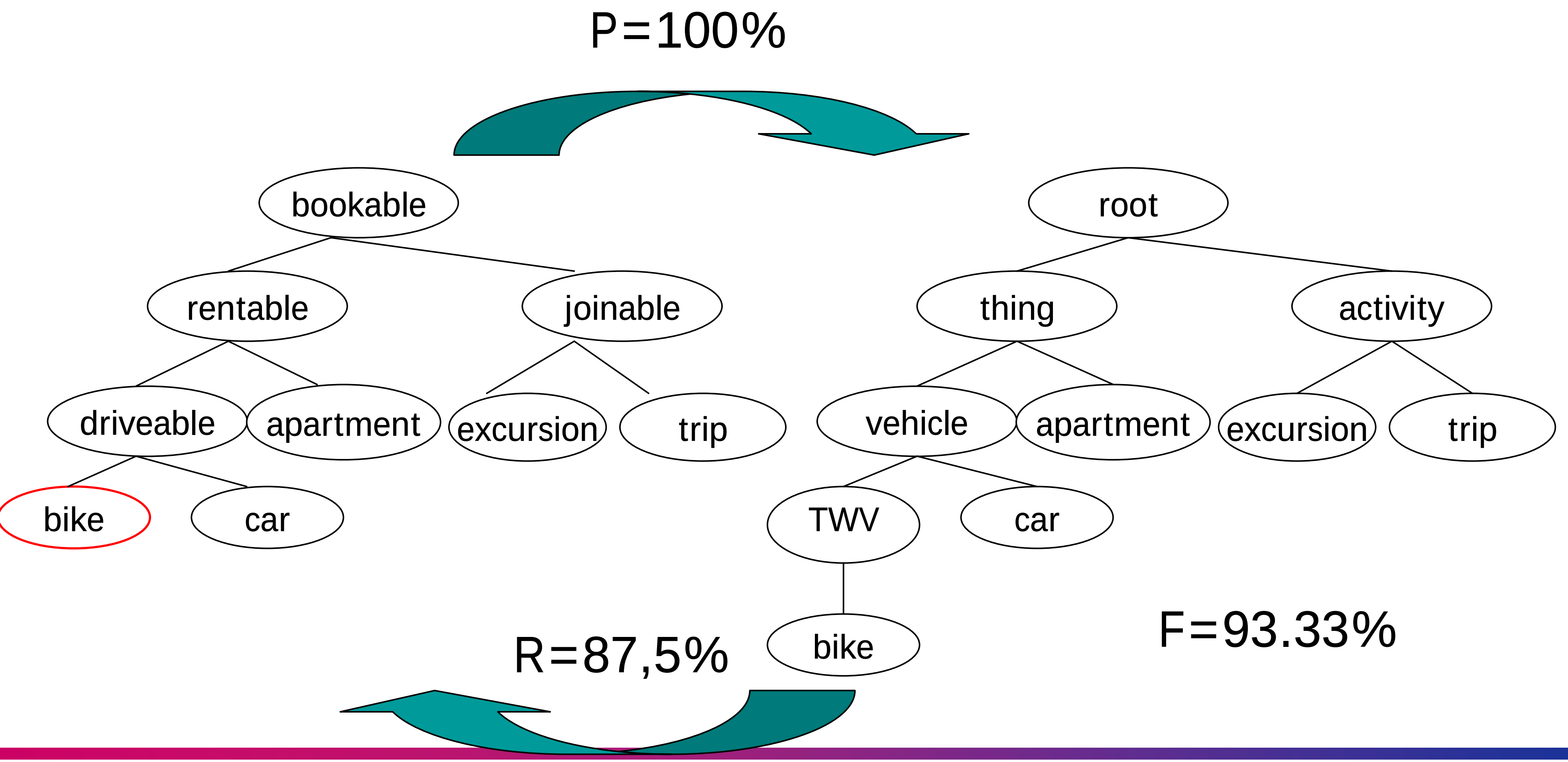
SC‘(vehicle)={bike,car}

=> TO(driveable,O_1,O_2)=1

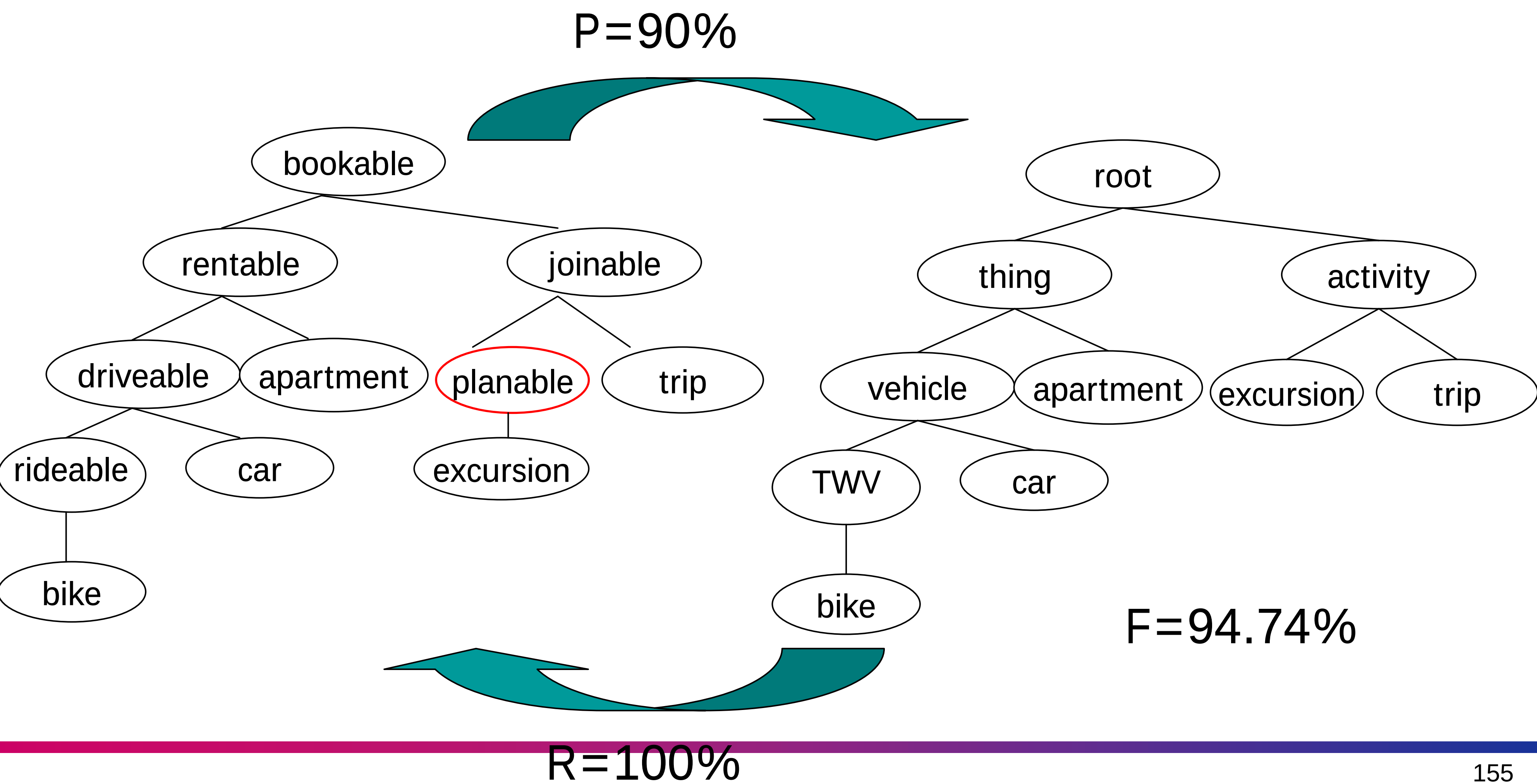
Example for Precision/Recall



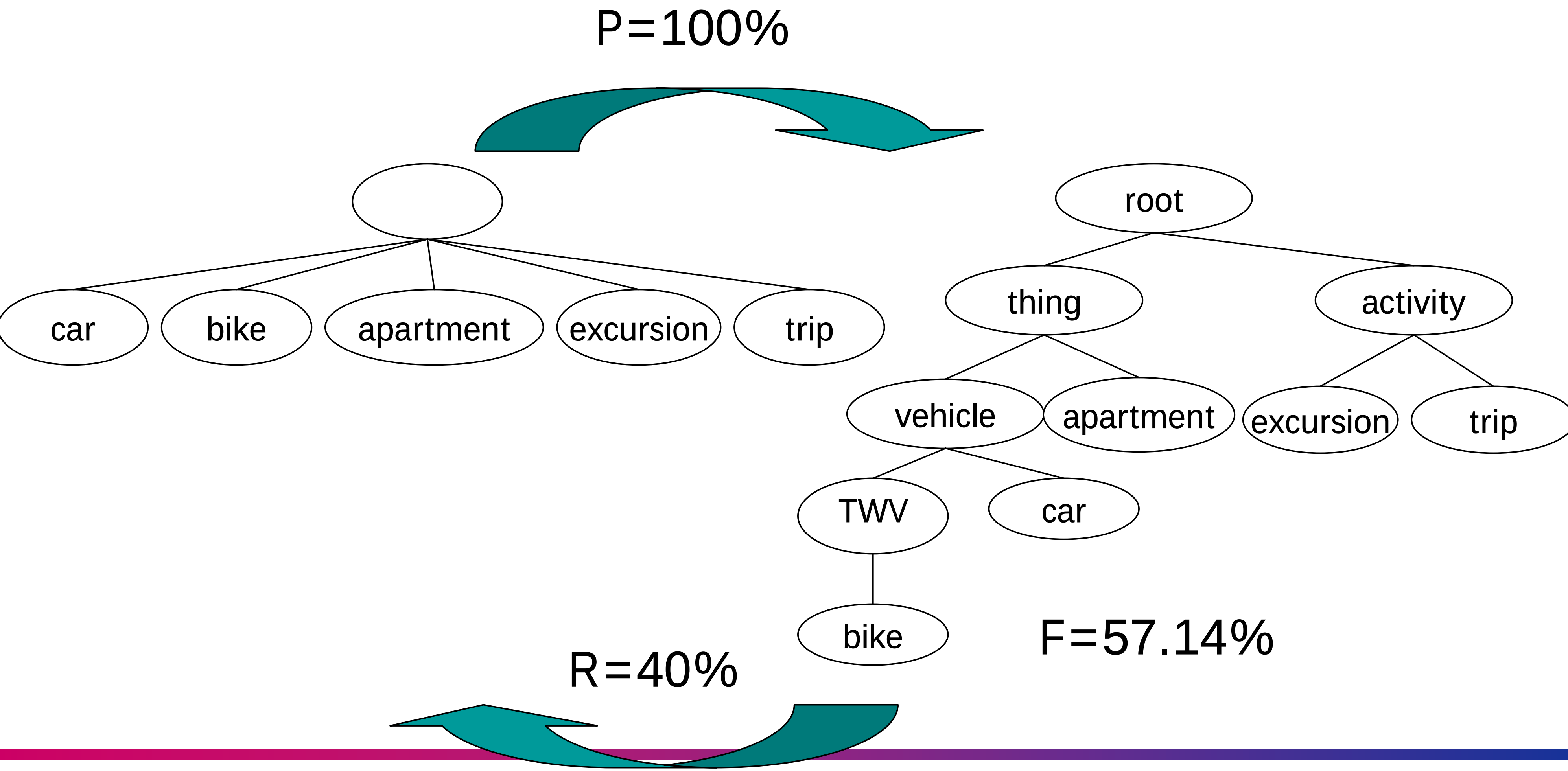
Example for Precision/Recall



Example for Precision/Recall



Trivial Concept Hierarchies



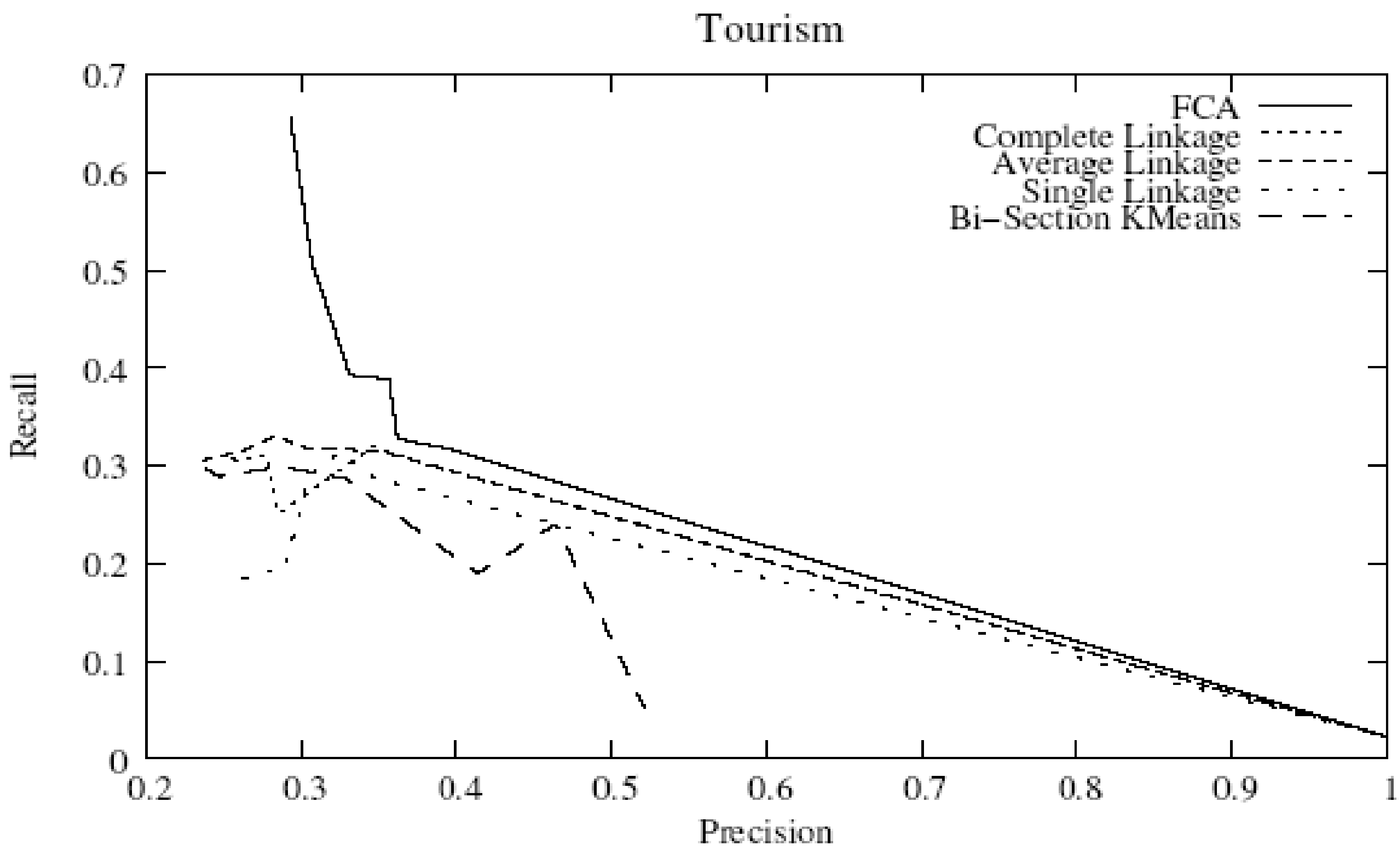
Evaluation

- Variant of the semantic cotopy
- Calculation of overlap in both directions:
  - Precision
  - Recall
  - F-Measure
- $$F' = \frac{2 \cdot F \cdot LR}{F + LR}$$

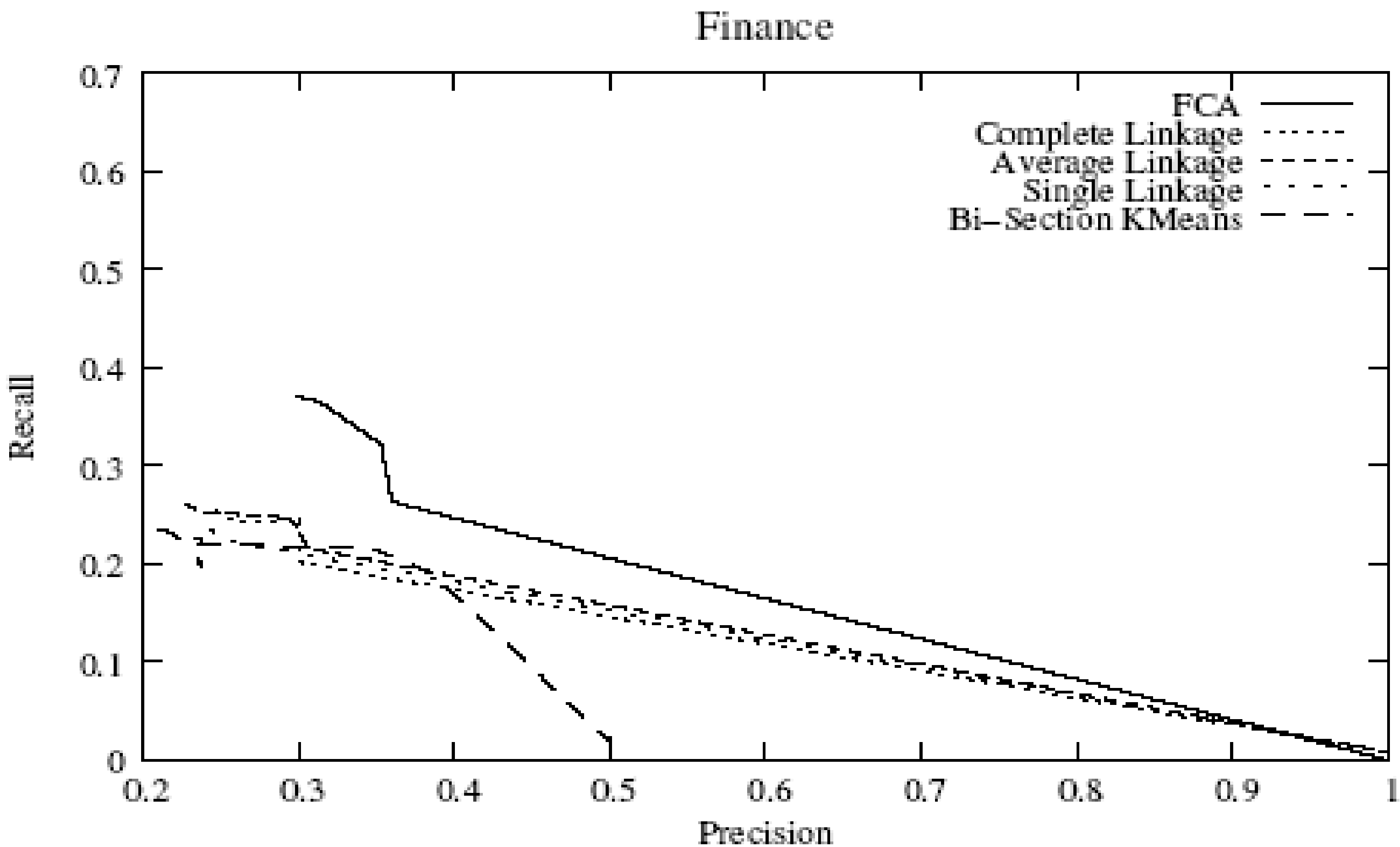
Syntactic Dependencies

Tourism				
	P_{TO}	R_{TO}	F_{TO}	F'
FCA	29.33%	**65.49%**	**40.52%**	**44.69%**
Complete Linkage	34.67%	31.98%	33.27%	36.85%
Average Linkage	**35.21%**	31.45%	33.23%	36.55%
Single Linkage	34.78%	28.71%	31.46%	38.57%
Bi-Section-KMeans	32.85%	28.71%	30.64%	36.42%
Finance				
	P_{TO}	R_{TO}	F_{TO}	F'_{TO}
FCA	29.93%	**37.05%**	**33.11%**	**38.85%**
Complete Linkage	24.56%	25.65%	25.09%	33.35%
Average Linkage	29.51%	24.65%	26.86%	32.92%
Single Linkage	25.23%	22.44%	23..75%	32.15%
Bi-Section-KMeans	**34.41%**	21.77%	26.67%	32.77%

Recall over Precision (Tourism)



Recall over Precision (Finance)



Pseudo-syntactic dependencies

Tourism				
	P_{TO}	R_{TO}	F_{TO}	F'_{TO}
FCA	27.02%	68.67%	38.78%	**48.82%**
Complete Linkage	26.44%	32.98%	29.35%	40.60%
Average Linkage	25.22%	34.68%	29.20%	40.72%
Single Linkage	40.40%	28.05%	33.08%	44.85%
Bi-Section-KMeans	22.07%	25.61%	23.66%	34.72%
Finance				
	P_{TO}	R_{TO}	F_{TO}	F'_{TO}
FCA	23.96%	33.32%	27.88%	**38.43%**
Complete Linkage	20.69%	22.98%	21.77%	32.59%
Average Linkage	19.92%	23.75%	21.66%	32.47%
Single Linkage	26.87%	19.98%	22.92%	33.86%
Bi-Section-KMeans	20.00%	21.53%	20.72%	29.53%

Summary of Results

	Effectiveness (F')		Worst Case Complexity	Traceability	Size
	Tourism	Finance			
FCA	**48.82%**	**38.85%**	$O(2^n)$	**Good**	Large
Agglomerative:					
Complete	40.60%	38.43%	$O(n^2 \log n)$	Fair	**Small**
Average	40.72%	32.92%	$O(n^2)$		
Single	44.85%	32.47%	$O(n^2)$		
Bi-Section-KMeans	36.42%	32.77%	$O(n^2)$	Weak	**Small**

Experimental results

- Formal Concept Analysis yields better concept hierarchies than similarity-based clustering algorithms.
- The results of FCA are better understandable (intensional description of concepts!)
- Bi-Section-Kmeans is most efficient ($O(n^2)$)
- Though FCA is exponential in the worst case, it shows a favorable runtime behavior (sparsely populated formal contexts)
- The more fine-grained features, the better the results!

Clustering Concept Hierarchies from Text

- Similarity-based
- Set-theoretical & Probabilistic
- **Soft clustering**

What About Multiple Word Meanings?

- **bank**: financial institute or natural object?
  - At least two clusters!
- So we need soft clustering algorithms:
  - Clustering By Committee (CBC) [Lin et al. 2002]
  - Gaussian Mixtures (EM)
  - **PoBOC** (Pole-Based Overlapping Clustering)
  - FCA
  - (...)
- Challenge: recognize multiple word meanings!

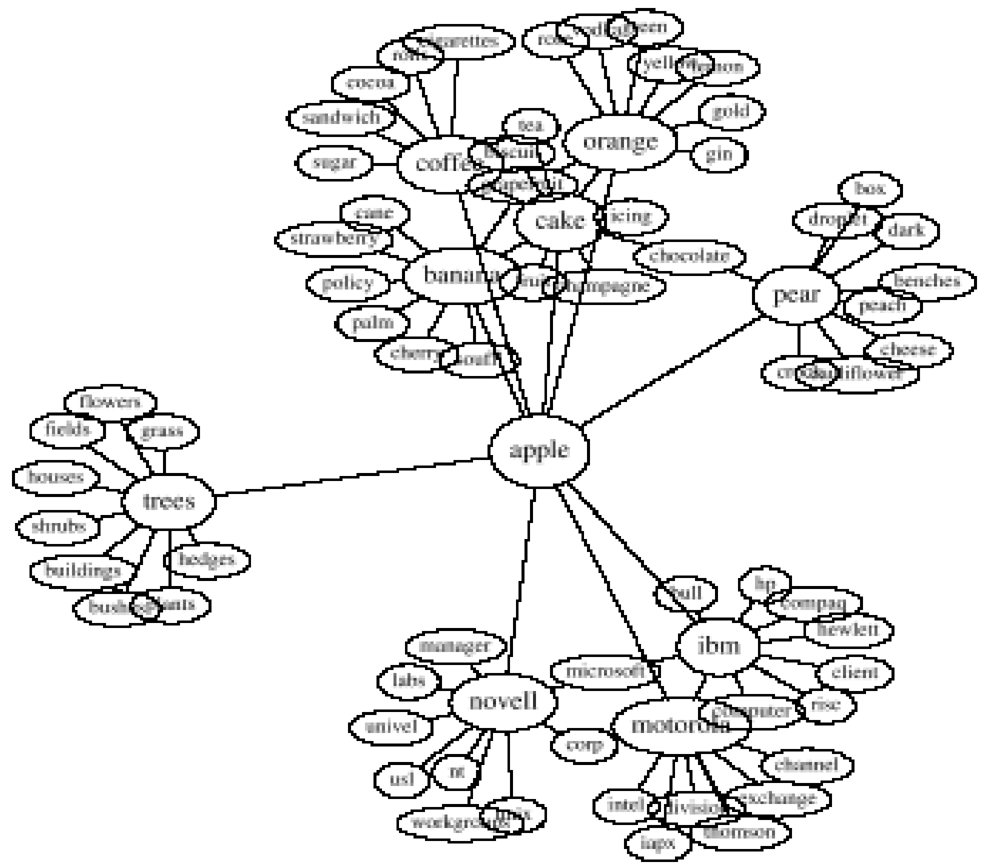
Soft clustering algorithms

- Principle underlying POBOC and CBC:
  - Construct first \`poles‘ or ´committees´ corresponding to very homogeneous groups of words, e.g. monosemous words
  - At a second step, assign words which do not form poles or committees to one or more committees; these are the ambiguous words

- Additional trick in CBC: once you assign a word to a committee, remove the overlapping features, i.e. subtract the \`meaning of the committee´

Approach by [Widdows and Dorow 2002]

- Extract shallow grammatical relations for words -> build a context vector.
- Apply LSA/LSI to reduce dimension of co-occurrence matrix.
- Calculate similarity as the cosine between the angle of the corresponding vectors.
- Senses of a word = disjoint subgraphs



Scalability

- Problem with clustering algorithms:
  - Compute at least pairwise similarity between words, i.e. $O(n^2k)$

- Idea of [Ravichandran, Pantel and Hovy]
  - Apply locality sensitive hash functions
  - i.e. approximate cosine measure by a randomized procedure

Randomly approximating the cosine measure

$$h_r(u) = \begin{cases} 1 : r \cdot u \geq 0 \\ 0 : r \cdot u < 0 \end{cases}$$

$$P[h_r(u) = h_r(v)] = 1 - \frac{\theta(u,v)}{\pi}$$

$$\cos(\theta(u,v)) = \cos((1 - P[h_r(u) = h_r(v)])\pi)$$

$$P[h_r(u) = h_r(v)] = 1 - \frac{\text{hammingDistance(u,v)}}{\text{d}}$$

where d is the number of random vectors!

Taxonomy Extraction - Overview

- Lexico-syntactic patterns
- Distributional Similarity & Clustering
- **Linguistic Approaches**
- Taxonomy Extension/Refinement
- Combination of Methods
- Evaluation
- Tools Matrix

Demos

- Similar Words:

  http://www.isi.edu/~pantel/Content/Demos/LexSem/thesaurus.htm

- CBC:

  http://www.isi.edu/~pantel/Content/Demos/LexSem/cbc.htm

Linguistic Approaches

- Modifiers:
  - Modifiers (adjectives/nouns) typically restrict or narrow down the meaning of the modified noun, i.e.
  - e.g. *isa(international credit card, credit card)*
  - Yields a very accurate heuristic for learning taxonomic relations, e.g. OntoLearn [Velardi & Navigli], OntoLT [Buitelaar et al., 2004], TextToOnto [Cimiano et al.], [Sanchez et al., 2005]

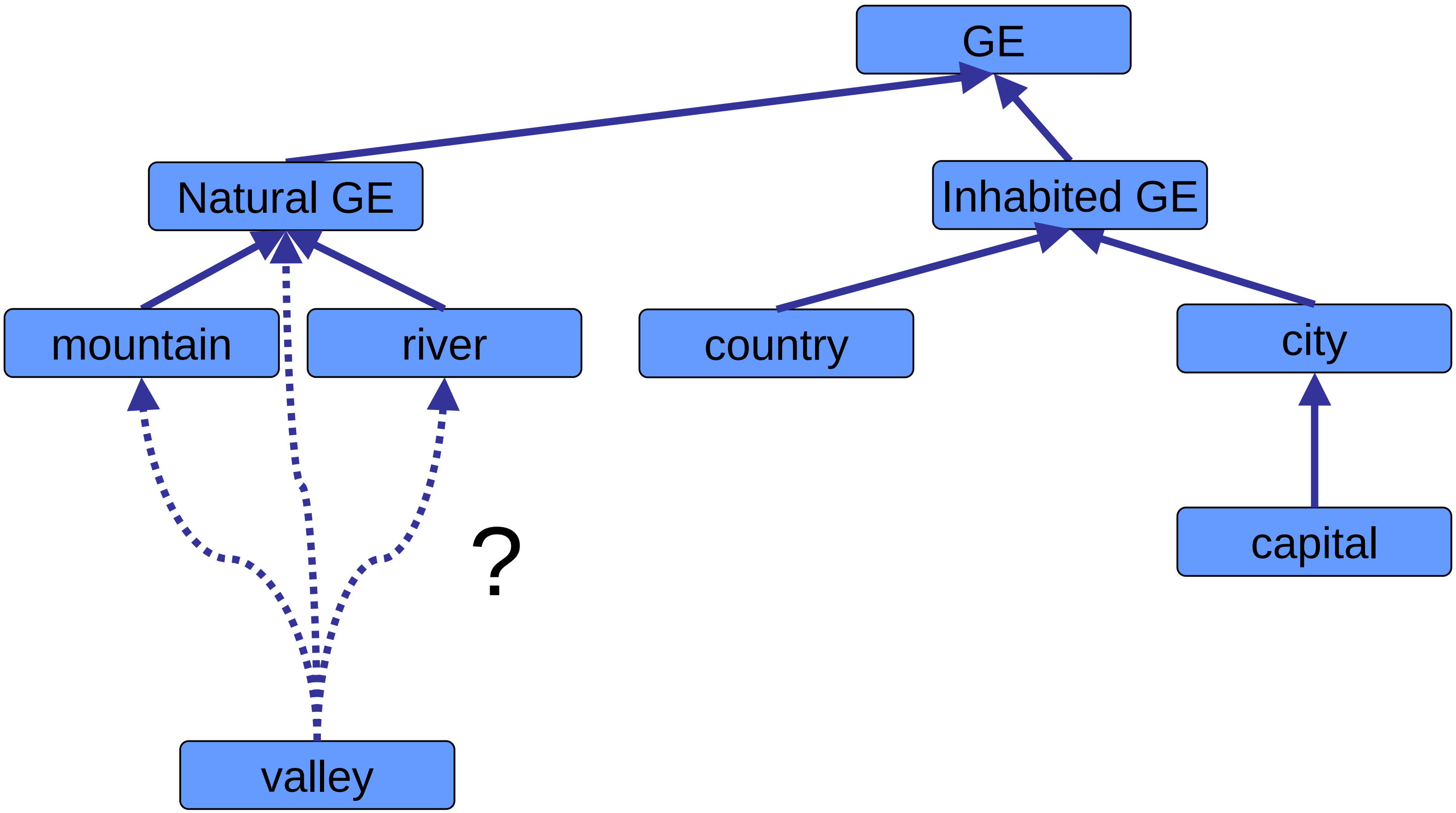
- Compositional interpretation of compounds [OntoLearn]
  - e.g. *long-term debt*
    - Disambiguate *long-term* and *debt* with respect to WordNet
    - Generate a gloss out of the glosses of the respective synsets:

      long-term debt := „*a kind of debt, the state of owing something (especially money), relating to or extending over a relatively long time*“

Taxonomy Extraction - Overview

- Lexico-syntactic patterns
- Distributional Similarity & Clustering
- Linguistic Approaches
- **Taxonomy Extension/Refinement**
- Combination of Methods
- Evaluation
- Tools Matrix

General Problem



Hearst & Schuetze 1993

- For each word *w* in WordSpace:
  - collect the 20 nearest neighbors in space using the cosine measure,
  - compute the score s_i of category *i* for *w* as the number of nearest neighbors that are in *i*, and
  - assign *w* to the highest scoring category.

Widdows 2003

- For a target word *w*, find words from the corpus which are similar to those of *w*. Consider these corpus-derived neighbors *N(w)*

- Map the target word *w* to the place in the taxonomy where the neighbors *N(w)* are most concentrated.

- Crucial question: What does *most concentrated* mean?

Determine where they are `most concentrated´

- Maximization problem:

$$H := \bigcup_{w' \in N(w)} H(w')$$

$$\alpha(w,h) = \begin{cases} f(dist(w,h)) \text{ if } \mathrm{h} \in \mathrm{H(w)} \\ -g(w,h) \text{ if } \mathrm{h} \notin \mathrm{H(w)} \end{cases}$$

$$\max_{h \in H} \sum_{w' \in N(w)} \alpha(w',h)$$

Improving Precision and Recall of Hearst patterns

[Cederberg and Widdows 03]

Main Idea:

- Improve precision by filtering hyponym pairs using their similarity in WordSpace (error reduction by 30%, P=58%)
- Improve recall by using coordination information, i.e. A < B and coordinated(A,C) -> C < B
  - This yields a five-fold increase in recall while mantaining precision at P=54% using the WordSpace filtering technique.

Concept Hierarchy – Tools

Organization	System	Ontology Learning Layers							
		Terms	*Synonyms*	*Concept Formation*	*Concept Hierarchy*	*Relations*	*Relation Hierarchy*	*Axioms Schemata*	*General Axioms*
AIFB, Univ. Karlsruhe	***Text2Onto***	**X**	**clusters**	**int.**	**X**				
	AEON								
Amir Kabir Univ. Tehran	***HASTI***	**X**			**X**				
CNTS, Univ. Antwerpen	***OntoBasis***		**clusters**	**clusters**					
DFKI	***OntoLT / RelExt***	**X**			**X**				
Economic Univ. Prague	***TextToOnto ++***								
ISI, USC	***CBC***		**clusters**	**clusters**					
	DIRT								
Keio Univ.	***DODDLE***								
NRC-CNRC	***PMI-IR***		**X**						
Univ. de Paris-Sud	***ASIUM / Mo‘k***		**clusters**	**clusters**	**X**				
Univ. di Roma	***OntoLearn***	**X**	**X**	**int.**	**X**				
Univ. of Salford	***ATRACT***	**X**	**clusters**	**clusters**					
Univ. Zürich	***Parmenides***	**X**			**X**				

Ontology Learning Layer Cake

$\forall x\,(\mathrm{country}(x) \rightarrow \exists y\ \mathrm{capital\_of}(y, x) \wedge \forall z\,(\mathrm{capital\_of}(z, x) \rightarrow y = z))$	General Axioms
$\mathrm{disjoint}(\mathrm{river}, \mathrm{mountain})$	Axiom Schemata
$\mathrm{capital\_of} \leq_R \mathrm{located\_in}$	Relation Hierarchy
$\mathrm{flow\_through}(\mathrm{dom}: \mathrm{river}, \mathrm{range}: \mathrm{GE})$	Relations
$\mathrm{capital} \leq_C \mathrm{city}, \mathrm{city} \leq_C \mathrm{Inhabited\ GE}$	Concept Hierarchy
$c := \mathrm{country} := \langle i(c), \lVert c \rVert, \mathrm{Ref}_C(c) \rangle$	Concept Formation
{country, nation, Land}	(Multilingual) Synonyms
river, country, nation, city, capital,... .	Terms

General Relations: Exploiting Linguistic Structure

- **OntoLT**: *SubjToClass_PredToSlot_DObjToRange* Heuristic
    - Maps a linguistic subject to a class, its predicate to a corresponding slot for this class and the direct object to the range of the slot

- **TextToOnto**: Acquisition of Subcategorization Frames
    - love(man,woman)
    - love(kid,mother)
    - love(kid,grandfather)

love(person,person)

- Problem related to acquisition of *subcategorization frames* and *selectional restrictions* in Natural Language Processing
    - e.g. [Resnik 97], [Ribas 95], [Clark and Weir 02]

Finding the Right Level of Abstraction

- [Ciramita et al. 05]
  - Genia Corpus. + Genia Ontology
  - Verb-based relations
    - X activates B

- Use X^2 to decide to generalize or not (significance level)

- Results:
  - 83.3% of relations correct according to human evaluation
  - 53.1% correctly generalized

Our experiments

- Genia corpus & Genia ontology
- Extract subj-verb-obj relations using a shallow parser (Abney‘s CASS)
- Try to find the appropriate domain and range for the relations wrt. Genia
- Use different statistical measures to generalize!

Comparing different measures

- Conditional Probability
- Point-wise Mutual Information
- Chi-square test:

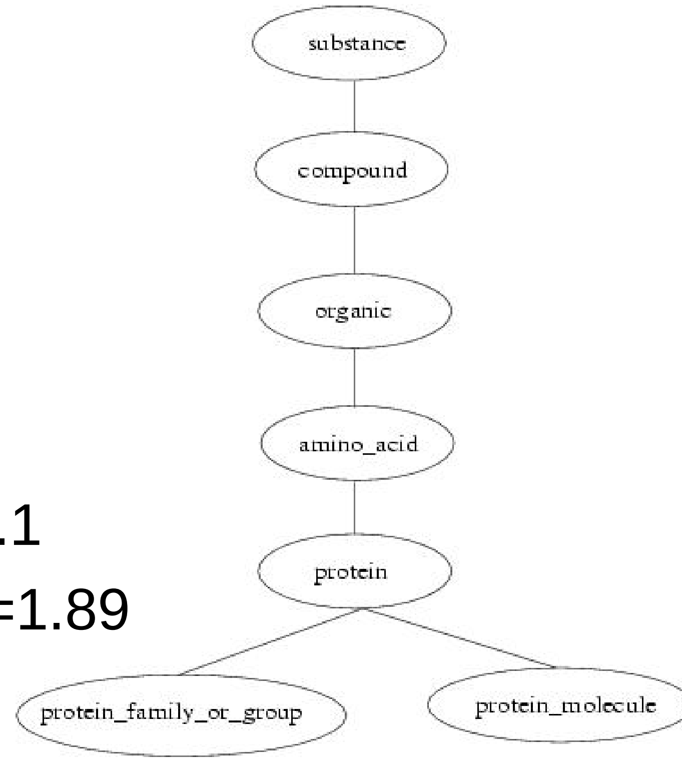
$$P(c \mid v_{\text{arg}})$$

$$\frac{P(c \mid v_{\text{arg}})}{P(c)}$$

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})}{E_{ij}}$$

An example

- Words found as objects of activate:
  - protein_molecule: 5
  - protein_family_or_group 10
  - amino_acid: 10
- Cond. Prob
  - P(protein|activate_obj)=15/25 = 0.6
  - P(amino_acid|activate_obj)=25/25 = 1
- PMI
  - PMI(protein,activate_obj)=log(0.6/0.14)= 2.1
  - PMI(amino_acid,activate_obj)=log(1/0.27)=1.89



Example (Cont'd)

	obj(activate)	~ obj(activate)
protein	15	400
~protein	35	2600

	obj(activate)	~obj(activate)
AA	25	800
~AA	25	2200

$$\chi^2(obj(activate), protein) = 11.62$$

$$\chi^2(obj(activate), AA) = 13.57$$

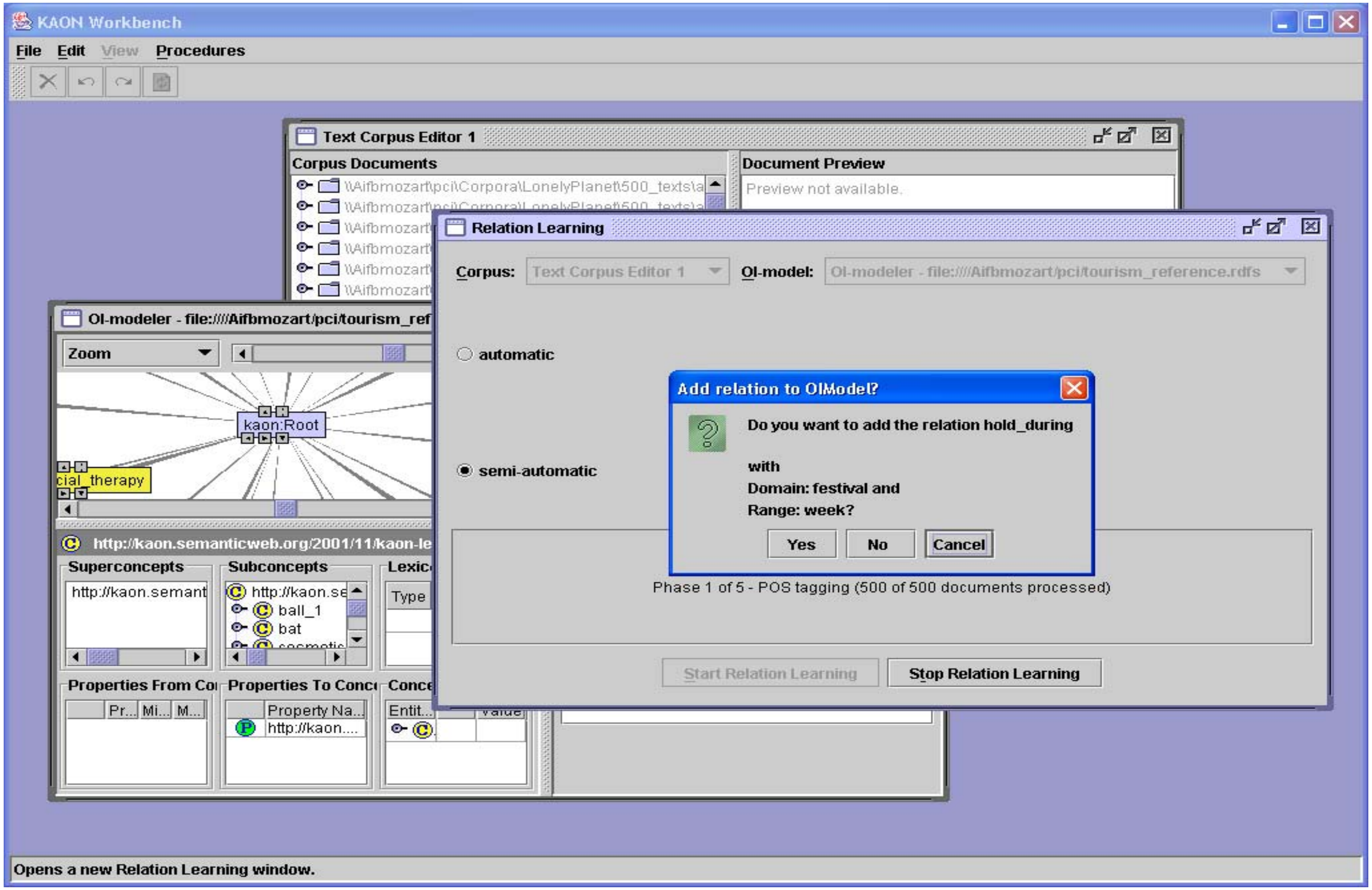
Results

- Evaluation
  – Biologist labelled 100 relations from hand by selecting the appropriate domain and range from the Genia corpus
  – Surprisingly, the conditional probability gives the best results!
  – But chi-square still works better than PMI!
- Peculiarities:
  – Genia ontology very shallow
  – Corpus semantically annotated

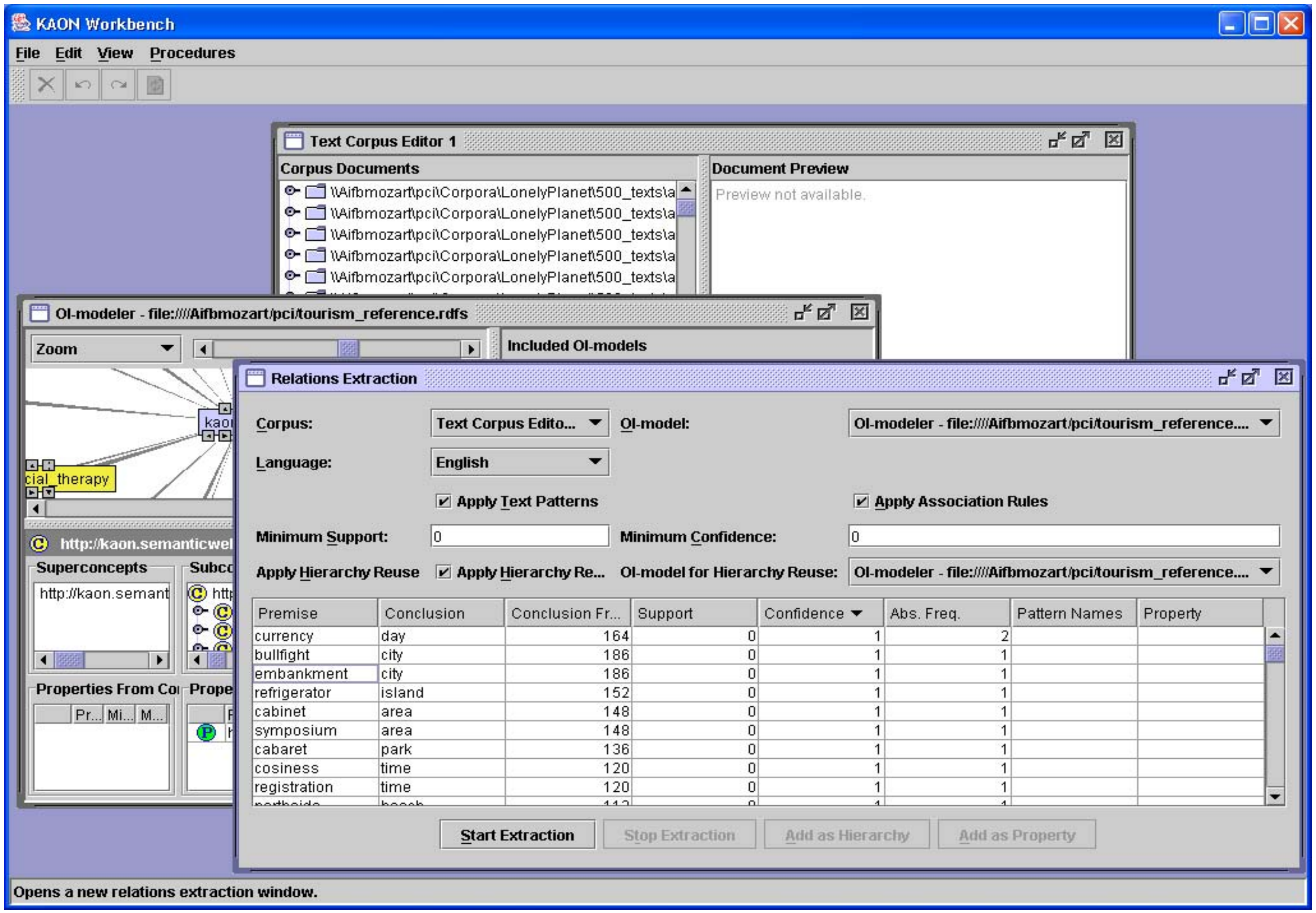
Relations – Tools

Organization	System	Ontology Learning Layers							
		Terms	*Synonyms*	*Concept Formation*	*Concept Hierarchy*	*Relations*	*Relation Hierarchy*	*Axioms Schemata*	*General Axioms*
AIFB, Univ. Karlsruhe	***Text2Onto***	**X**	**clusters**	**int.**	**X**	**X**			
	AEON								
Amir Kabir Univ. Tehran	***HASTI***	**X**			**X**	**X**			
CNTS, Univ. Antwerpen	***OntoBasis***		**clusters**	**clusters**		**?**			
DFKI	***OntoLT / RelExt***	**X**			**X**	**X**			
Economic Univ. Prague	***TextToOnto ++***					**labels**			
ISI, USC	***CBC***		**clusters**	**clusters**					
	DIRT								
Keio Univ.	***DODDLE***					**X**			
NRC-CNRC	***PMI-IR***		**X**						
Univ. de Paris-Sud	***ASIUM / Mo‘k***		**clusters**	**clusters**	**X**	**X**			
Univ. di Roma	***OntoLearn***	**X**	**X**	**int.**	**X**	**X**			
Univ. of Salford	***ATRACT***	**X**	**clusters**	**clusters**					
Univ. Zürich	***Parmenides***	**X**			**X**				

TextToOnto & Relations



TextToOnto - Relations



Ontology Learning Layer Cake

$\forall x\,(\text{country}(x) \rightarrow \exists y\ \text{capital\_of}(y, x) \wedge \forall z\,(\text{capital\_of}(z, x) \rightarrow y = z))$

$\text{disjoint}(\text{river}, \text{mountain})$

$\text{capital\_of} \leq_R \text{located\_in}$

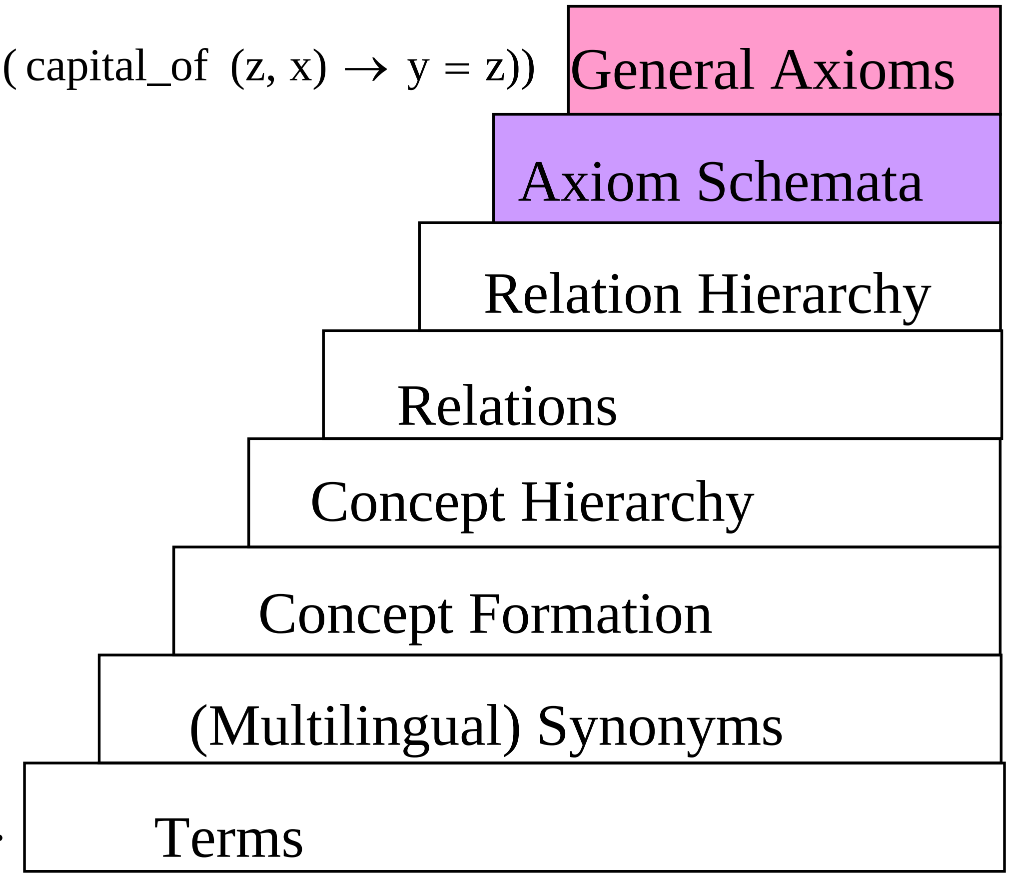
$\text{flow\_through}(\text{dom}: \text{river}, \text{range}: \text{GE})$

$\text{capital} \leq_C \text{city}, \text{city} \leq_C \text{Inhabited GE}$

$c := \text{country} := \langle i(c), \|c\|, \text{Ref}_C(c) \rangle$

$\{\text{country}, \text{nation}, \text{Land}\}$

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



Summary

- Terms: use some statistical measure to assess relevance wrt. to a corpus
- Concept Hierarchies:
  - Formal Concept Analysis & Clustering
  - Hearst Patterns
- Relations: use NLP techniques to extract verbs and their argument structure (Generalize!)

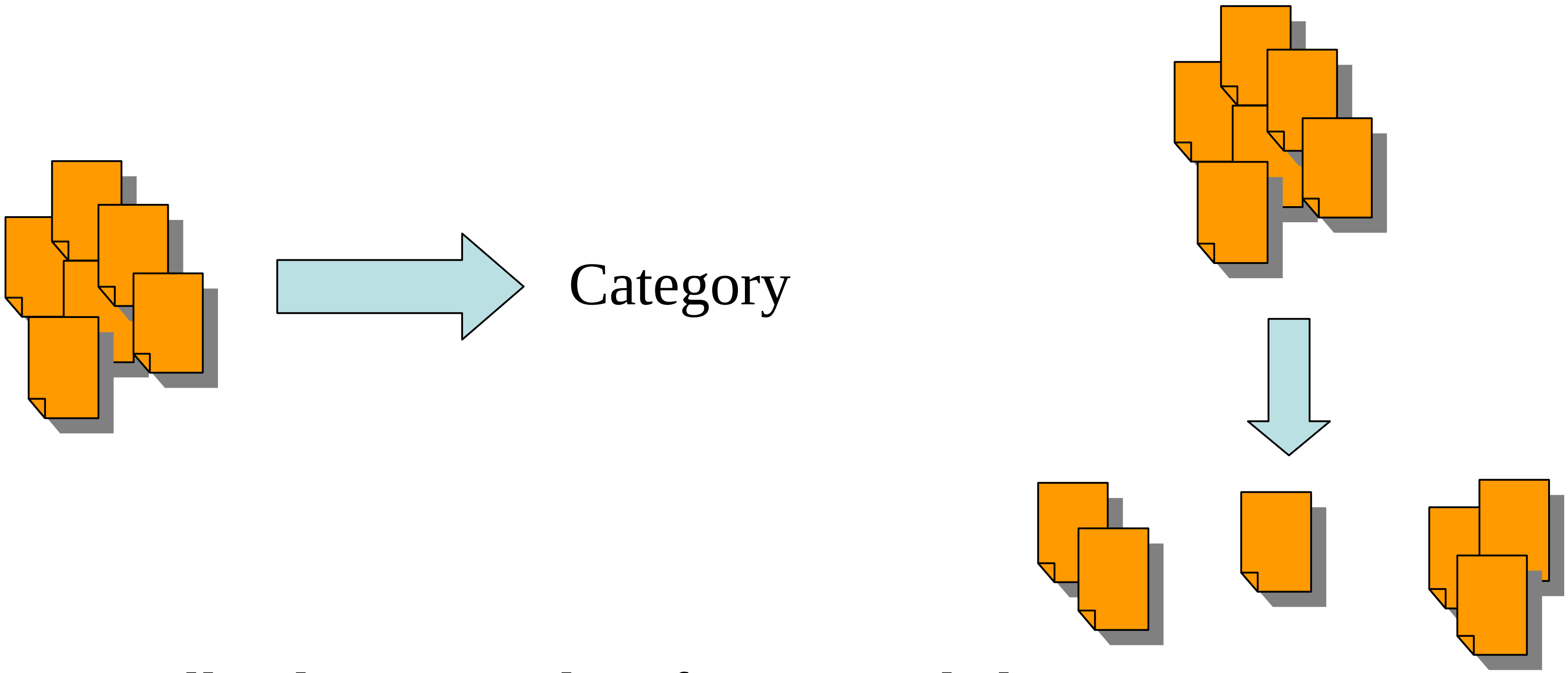
Agenda

- Ontologies
- Motivation
- Ontology Learning
  - Layer Cake
  - Term Extraction
  - Concept Hierarchies
  - Relations
- **Applications**
- Conclusion

Applications

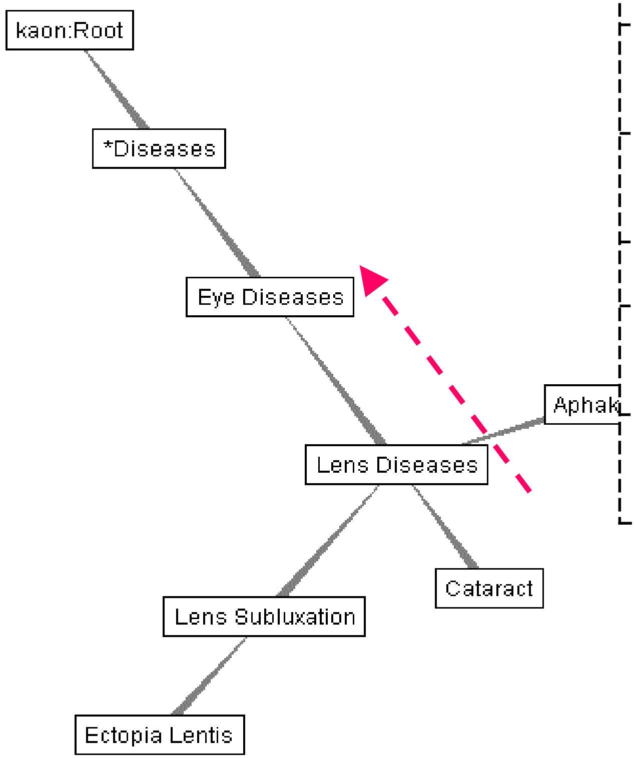
- Information Retrieval:
  - Query Expansion
  - Document Similarity (IR)
- Natural Language Processing
  - Word Sense Disambiguation
- Text Mining:
  - Enhanced bag-of-word model

Classification and Clustering of Texts



- Typically, document classification and clustering methods rely on the bag-of-words model.
- Recently, the bag-of-words model has been enhanced to also contain conceptual features derived from a domain ontology [Bloehdorn et al. 2005].

Generalization

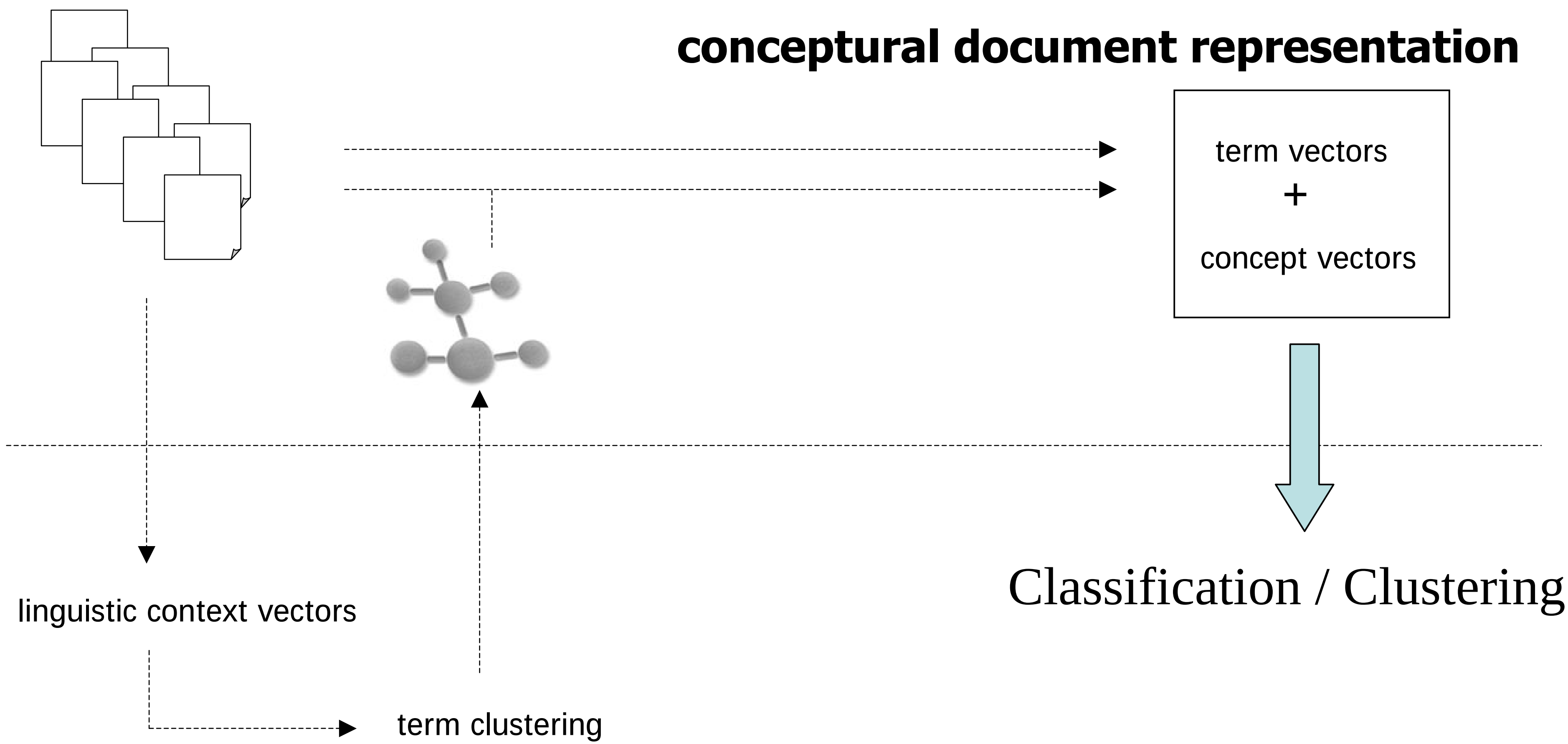


	term: catar	term: eye	term: disease	concept: cataract	concept: eye disease
doc_1	1	0	0	N/A	N/A
doc_1 (1)	1	0	0	**1**	0
doc_1 (2)	1	0	0	1	**1**
doc_2	0	1	1	N/A	N/A
doc_2 (1)	0	1	1	0	**1**
doc_2 (2)	0	1	1	0	**1**

h=2

Relative Improvement w.r.t. bag-of-words model between 2 and 7%.

Using automatically learned ontologies



Results

- Automatically learned ontologies achieve comparable results to hand-crafted ontologies wrt. clustering and classification tasks.
- Best Algorithm: Bi-Section KMeans
- Unclear how many levels one has to move up!
- Conclusion: For some applications automatically generated ontologies are ‚good enough‘.

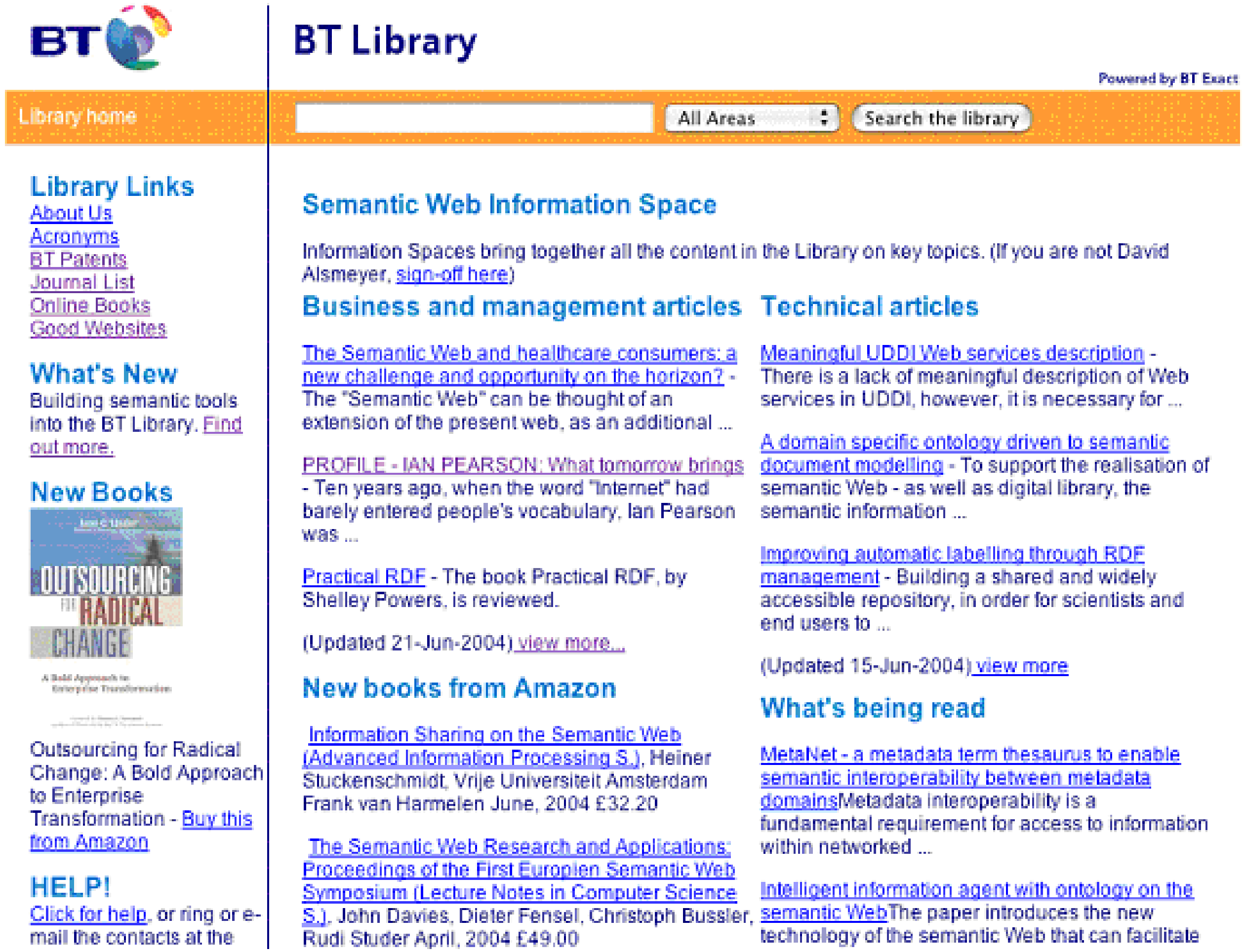
SEKT Case Studies

- BT case study
- Legal case study

BT (British Telecom) Case Study

- Digital Library (since 1994)
- Single interface for accessing multiple databases with content from different publishers
- More than 1 million technical articles and papers from 12000 publications, about 1000 business and management magazines
- Main features:
  - Information spaces: collections of documents about ‚interesting' topics
  - Searching and browsing
  - Personalization: alerts, bookmarks, annotations, private information spaces

BT Case Study ‚Semantic Web' Information Space



BT Library

Powered by BT Exact

Library home

All Areas | Search the library

Library Links

About Us
Acronyms
BT Patents
Journal List
Online Books
Good Websites

What's New

Building semantic tools into the BT Library. Find out more.

New Books

Outsourcing for Radical Change: A Bold Approach to Enterprise Transformation - Buy this from Amazon

HELP!

Click for help, or ring or e-mail the contacts at the

Semantic Web Information Space

Information Spaces bring together all the content in the Library on key topics. (If you are not David Alsmeyer, sign-off here)

Business and management articles

The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon? - The "Semantic Web" can be thought of an extension of the present web, as an additional ...

PROFILE - IAN PEARSON: What tomorrow brings - Ten years ago, when the word "Internet" had barely entered people's vocabulary, Ian Pearson was ...

Practical RDF - The book Practical RDF, by Shelley Powers, is reviewed.

(Updated 21-Jun-2004) view more...

New books from Amazon

Information Sharing on the Semantic Web (Advanced Information Processing S.), Heiner Stuckenschmidt, Vrije Universiteit Amsterdam Frank van Harmelen June, 2004 £32.20

The Semantic Web Research and Applications: Proceedings of the First Europien Semantic Web Symposium (Lecture Notes in Computer Science S.), John Davies, Dieter Fensel, Christoph Bussler, Rudi Studer April, 2004 £49.00

Technical articles

Meaningful UDDI Web services description - There is a lack of meaningful description of Web services in UDDI, however, it is necessary for ...

A domain specific ontology driven to semantic document modelling - To support the realisation of semantic Web - as well as digital library, the semantic information ...

Improving automatic labelling through RDF management - Building a shared and widely accessible repository, in order for scientists and end users to ...

(Updated 15-Jun-2004) view more

What's being read

MetaNet - a metadata term thesaurus to enable semantic interoperability between metadata domainsMetadata interoperability is a fundamental requirement for access to information within networked ...

Intelligent information agent with ontology on the semantic WebThe paper introduces the new technology of the semantic Web that can facilitate

BT Case Study Ontology Learning Scenario

- Learn fine-grained topic hierarchy from each information space
- Why?
  - Visualization of information spaces
  - Searching and browsing information spaces (Query Refinement)
  - Topic discovery
- Integrated with a Query Refinement Tool

Evaluation Setting

- Corpus: 1700 abstracts from ‚knowledge management' information space
- 5 human annotators, domain experts
- For each type of ontology element …
  - Each annotator was given the top 50 ontology learning results (regarding confidence / relevance)
  - Rating scale ranging from 1 (completely wrong) to 5 (perfectly correct)

Algorithms

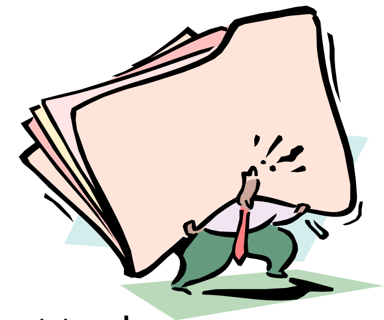
- Concept and Instance Extraction:
  - TFIDF (discussed)
- Subclass relations
  - Combination of Hearst Patterns + WordNet + Linguistic Heuristics (partially discussed)
- Instance-of relations
  - Hearst Patterns (discussed)
- Non-taxonomic relations
  - Analysis of verb structure (discussed)
- Subtopic relations
  - Sanderson and Croft algorithm
- Disjointness Axioms
  - Analysis of enumerations, e.g. men and women

Evaluation Results
Conclusion

- Promising evaluation results
- Problems due to evaluation procedure and human perception
- High disagreement among human annotators
  - ‚What is a topic?'
  - ‚Which score do I have to assign if I do not know a concept / instance or if the label is ambiguous?'
  - ‚How can you talk about disjointness of concepts which do not have a set theoretic interpretation?'

Legal Case Study

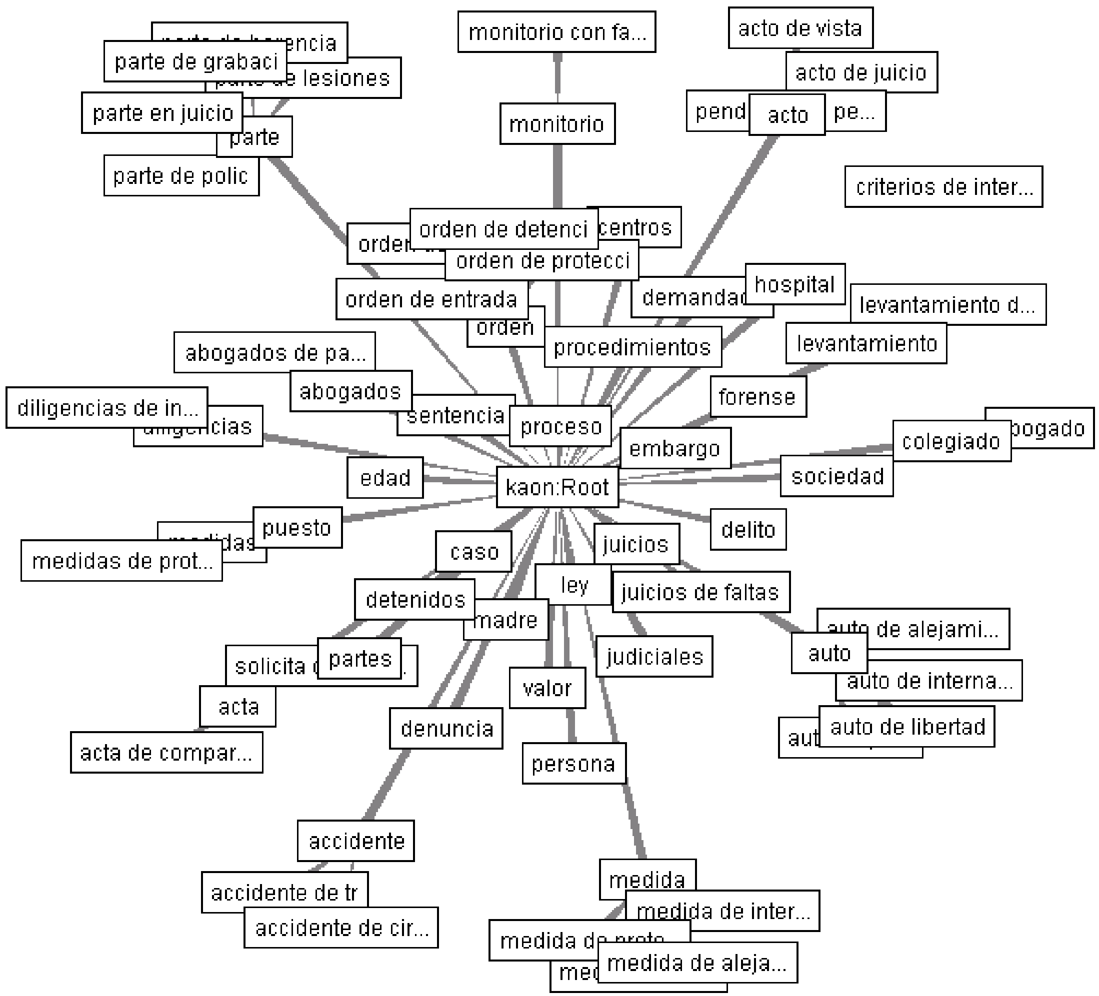
- In General:
  - Complaint about diligence of legal administration.
  - The Judges are overworked.
- In Particular:
  - New Judges
  - A lot of theoretical knowledge, but few practical knowledge
  - On Duty.
    - When they are confronted with situations in which they are not sure what to do
    - “Disturb” experienced judges with typical questions.
    - Usually his/her former tutor (Preparador)
- Existing Technology
  - Legal Databases
    - Essential in their daily work
    - Based on keywords and boolean operators
    - A search retrieves a huge number of hits



Description of the Problem: Legal Domain

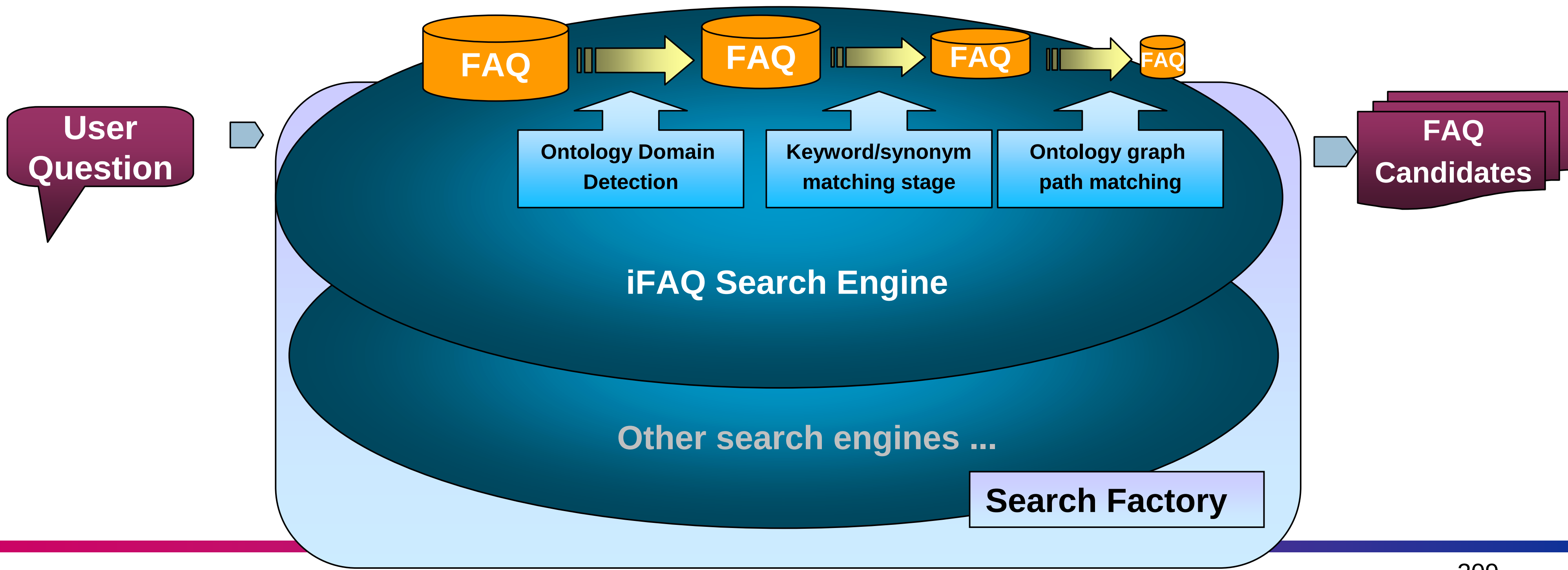
- Solution:
  - Design an intelligent system to help new judges with their typical problems.
  - Extended FAQ system using Semantic Web technologies
  - Connect the FAQ system with the exiting jurisprudence.
    - Search Jurisprudence using Semantic Web technologies.

Learning Concept Hierarchies with the Spanish version of TextToOnto



Expert Knowledge Retrieval

- Use automatically learned ontologies for computation of similarity between question and FAQ database (consider synonyms, etc.)



Applications in IR

- Query Refinement:
  - Use corpus-derived synonyms
  - Use corpus-derived subconcepts
- Query Interpretation:
  - Headache medicine
    - Cure or cause ?
- See OntoQuery project

Take-home Message

- Powerful Methods:
  - Matching of lexico-syntactic patterns
  - Distributional Similarity:
    - Use any similarity measure of your choice
    - Yields similar words (near synonyms)
- Very promising applications:
  - Information retrieval
  - Text Mining in general

References

[Abecker et al. 1997] - A. Abecker, S. Decker, K. Hinkelmann, U. Reimer. In: Proceedings of the International Workshop on Knowledge-Based Systems for Knowledge Management in Enterprises at the German AI Conference (KI-97), 1997.

[Agichtein and Gravano, 2000] - E. Agichtein, L. Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries (ACM DL), pp. 85-94, 2000.

[Alani et al. 2003] - H. Alani, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, N. R. Shadbolt. *Automatic Ontology-Based Knowledge Extraction from Web Documents*. IEEE Intelligent Systems, 18(1), pp. 14-21, 2003.

[Appelt and Israel 1999] – Tutorial Notes of the IJCAI Tutorial on Information Extraction Technology, 1999.

[Beale et al.1995] - S. Beale, S. Nirenburg, K. Mahesh. Semantic Analysis in the Mikrokosmos Machine Translation Project. In: Proceedings of the 2nd Symposium on Natural Language Processing, pp. 297-307, 1995.

[Bloehdorn et al. 2005] – S. Bloehdorn, P. Cimiano, A. Hotho, Learning Ontologies to Improve Text Clustering and Classification, In From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the German Classification Society (GfKl), 2005.

[Bisson et al. 2000] - G. Bisson, C. Nedellec, L. Canamero. Designing clustering methods for ontology building - The Mo'K workbench. In: Proceedings of the ECAI Ontology Learning Workshop, pp. 13-19, 2000.

References

[Brin 1998] – S. Brin, *Extracting patterns and relations from the world wide web*. In Proceedings of the WebDB at EDBT'98, 172--183.

[Buitelaar, Sintek 2004] – P. Buitelaar, M. Sintek. OntoLT Version 1.0: Middleware for Ontology Extraction from Text. In: Proceedings. of the Demo Session at the International Semantic Web Conference (ISWC), 2004.

[Buitelaar et al. 2004b] – P. Buitelaar, D. Olejnik, M. Hutanu, A. Schutz, T. Declerck, M. Sintek. Towards Ontology Engineering Based on Linguistic Analysis. In: Proceedings of LREC, 2004.

[Buitelaar et al . 2004c] - P. Buitelaar, D. Olejnik, M. Sintek. A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In: Proceedings of the 1st European Semantic Web Symposium (ESWS), 2004.

[Buitelaar et al., 2006a] – P. Buitelaar, T. Eigner, G. Gulrajani, A. Schutz, M. Siegel, N. Weber, P. Cimiano, G. Ladwig, M. Mantel and H. Zhu, Generating and Visualizing a Soccer Knowledge Base, Demo Proceedings of EACL, 2006.

[Buitelaar et al., 2006b] – P. Buitelaar, P. Cimiano. S. Racioppa, M. Siegel, Ontology-based information extraction with SOBA, Proceedings of LREC 2006.

Buitelaar P., Sacaleanu B. Ranking and Selecting Synsets by Domain Relevance. In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001

[Califf and Mooney 1999] - M. Califf and R. Mooney. *Relational learning of patternmatch rules for information extraction*. In Proc. 16th Nat. Conf. Artifical Intelligence, 1999.

References

[Caraballo 1999] – S.A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 120-126, 1999.

[Cederberg and Widdows 2003] – S. Cederberg, D. Widdows. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: Proceedings of the Conference on Natural Language Learning (CoNNL), 2003.

[Ciramita et al. 2005] - M. Ciramita, A. Gangemi, E. Ratsch, J. Saric, I. Rojas. Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology. In. Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI), 2005.

[Cimiano et al. 2004] - P. Cimiano, S. Handschuh, S. Staab. Towards the Self-Annotating Web. IN: Proceedings of the 13th World Wide Web Conference, pp. 462-471, 2004.

[Cimiano et al. 2004b] – P. Cimiano, A. Hotho, S. Staab. Comparing Conceptual, Partitional and Agglomerative Clustering for Learning Taxonomies from Text
In: Proceedings of the European Conference on Artificial Intelligence (ECAI'04), pp. 435-439. IOS Press, 2004.

[Cimiano and Staab 2004] - P. Cimiano, S. Staab. Learning by Googling, SIGKDD Explorations, 6(2), 2004.

[Cimiano et al. 2005] - P. Cimiano, G. Ladwig, S. Staab. Gimme, The Context: Context-driven automatic semantic annotation with C-PANKOW, IN: Proceedings of the 14th World Wide Web Conference, 2005.

References

[Cimiano et al. 2005c] – P. Cimiano and S. Staab, Learning Concept Hierarchies from Text with a Guided Agglomerative Clustering Algorithm. In: Proceedings of the ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods. 2005.

[Cimiano and Hartung 2005] - P. Cimiano, M. Hartung, Finding the Appropriate Level of Generalization for Relations Extracted from the Genia Corpus. In: Proceedings of the International Lexical Resources and Evaluation Conference (LREC), 2006.

[Ciravegna 2001] – F. Ciravegna, Adaptive Information Extraction from Text by Rule Induction and Generalisation, In: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), 2001.

[Clark and Weir 2002] - S. Clark, D.J. Weir. Class-Based Probability Estimation Using a Semantic Hierarchy. Computational Linguistics, 28(2), pp. 187-206, 2002.

[Cleuziou et al. 2004] - G. Cleuziou, L. Martin, C. Vrain. PoBOC: An Overlapping Clustering Algorithm, Application to Rule-Based Classification and Textual Data. In: Proceedings of the European Conference on Artificial Intelligence (ECAI), pp. 440-444, 2004.

[Copestake et al. 1992] - Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni, E. Marinai. Multilingual Lexical Representation. ESPRIT BRA-3030 ACQUILEX - WP No. 043, 1992.

[Cucchiarelli and Velardi 1998] – A. Cucchiarelli and P. Velardi 1998. Finding a domain-appropriate sense inventory for semantically tagging a corpus. *Natural Language Engineering*, 4(4):325–344.

References

[Dorow and Widdows 2003] – B. Dorow, D. Widdows. Discovering Corpus-Specific Word Senses. In: Proceedings of EACL, pp. 79-82, 2003.

[Downey et al. 2004] - D. Downey, O. Etzioni, S. Soderland, D. Weld. Learning Text Patterns for Web Information Extraction and Assessment. In: Proceedings of the AAAI Workshop on Adaptive Text Extraction and Mining, 2004.

[Etzioni et al. 2004] - O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Web-Scale Information Extraction in KnowItAll (Preliminary Results), In: Proceedings of the 13th World Wide Web Conference, pp. 100-109, 2004.

[Etzioni et al. 2005] - O. Etzioni, M. Cafarella, D. Downey, A-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, Unsupervised Named-Entity Extraction from the Web: An Experimental Study. Artificial Intelligence, 165(1), pp. 91-134, 2005.

[Faure and Nedellec, 1998] – D. Faure, C. Nedellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: Proceedings of LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, 1998.

[Freitag and Kushmerick 2000] D. Freitag and N. Kushmerick. Boosted wrapper induction. In Proc. of the 17th National Conference on Artificial Intelligence AAAI-2000, pages 577--583, 2000.

[Fensel 2001] - D. Fensel, Ontologies: Silver bullet for knowledge management and electronic commerce, Springer, 2001.

[Firth 1957] - J. Firth, A synopsis of linguistic theory 1930-1955, Longman, Studies in Linguistic Analysis, Philological Society, 1957.

References

[Freitag and Kushmerick 2000] – D. Freitag, N. Kushmerick, Boosted Wrapper Induction, In: Proceedings of the 17th National Conference on Artificial Intelligence (AAAI), pp. 577-583.

[Frantzi and Ananiadou, 1999] – K.T. Frantzi, S. Ananiadou.The C-Value/NC-Value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3):145-179,1999.

[Ganter and Wille 1999] – B. Ganter, R. Wille. Formal Concept Analysis – Mathematical Foundations, Springer Verlag, 1999.

[Gasperin et al. 2001] - C. Gasperin, P. Gamallo, A. Agustini, G. Lopes and V. de Lima, Using Syntactic Contexts for Measuring Word Similarity. In: Proceedings of the ESSLLI Workshop on Semantic Knowledge Acquisition and Categorization, 2001.

[Gaizauskas et al. 1995] - R. Gaizauskas, T. Wakao, K. Humphreys, H. CunninghamY. Wilks. *Description of the LaSIE system as used for MUC*-6. In Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufmann, California, 1995.

[Girju et al. 2002] - R. Girju, D. Moldovan, Text Mining for Causal Relations, In: Proceedings of the FLAIRS Conference, pp. 360-364, 2002.

[Gluschko et al. 1999] - R. Gluschko and J. Tenenebaum, B. Meltzer. An XML Framework for Agent-based E-Commerce. In: Communications of the ACM 42(3):106-114, 1999.

[Grefenstette, 1992] - Grefenstette. Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In: Proceedings of the *30th Annual Meeting of the Association for Computational Linguistics,* Newark, Delaware, 28 June - 2 July 1992.

References

[Grefenstette 1994] – G. Grefenstette. Explorations in Automatic Thesaurus Discovery, Kluwer Academic Publishers, 1994.

[Grefenstette 1998] – G. Grefenstette. Cross-Language Information Retrieval, Kluwer Academic Publishing, 1998.

[Gruber 1993] - T.R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Formal Analysis in Conceptual Analysis and Knowledge Representation, Kluwer, 1993.

[Guarino et al. 1999] - N. Guarino, C. Masolo, G. Vetere. OntoSeek: Content-Based Access to the Web. In: IEEE Intelligent Systems, 14(3), 70--80, 1999.

[Hearst 1992] - M.A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, pp. 539-545, 1992.

[Hearst and Schütze 1993] – M.A. Hearst, H. Schütze. Customizing a lexicon to better suit a computational task. In: Proceedings of the ACL SIGLEX Workshop on Acquisition of Lexical Knowledge from Text, 1993.

[Hendler 2000] - J. Heflin, J. Hendler. Searching the Web with SHOE, In: Papers from the AAAI Workshop on Artificial Intelligence for Web Search, pp. 35-40, 2000.

[Kashyap 1999] - V. Kashyap. Design and Creation of Ontologies for Environmental Information Retrieval. Proceedings of the 11th European Workshop on Knowledge Acquisistion, Modeling,and Management (EKAW), 1999.

[Kesseler 1996] - M. Kesseler. A Schema Based Approach to HTML Authoring. In: World Wide Web Journal 96(1), O'Reilly, 1996.

References

[Kim et al. 2002] – S. Kim, H. Alani, W. Hall and P. Lewis and D. Millard and N. Shadbolt and M. Weal

Artequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web. In *Proceedings Semantic Authoring, Annotation and Knowledge Markup Workshop in the 15th European Conference on Artificial Intelligence*, 2002.

[Knight 1993] – K. Knight. Building a Large Ontology for Machine Translation, In Proceedings of the DARPA Human Language Conference, 1993.

[Landauer and Dumais 1997] – T. Landauer, S. Dumais, A solution to Plato's problem: The latext semantic analysis theory of acquisition, induction and representaion of knowledge, Psychological Review 104, pp. 211-240, 1997.

[Lin and Pantel 2001] - D. Lin, P. Pantel, DIRT - Discovery of Inference Rules from Text. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 323--328, 2001.

[Lin and Pantel 2001] - D. Lin, P. Pantel, Discovery of Inference rules for Question Answering. Natural Language Engineering, 7(4), pp. 343-360, 2001.

[Lopez and Motta 2004] – V. Lopez, E.Motta. Ontology-Driven Question Answering in AquaLog. In: Proceedings of NLDB, pp. 89-102. 2004.

[Mädche 2002] – A. Maedche. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002.

[Mädche and Staab 2002] – A. Maedche, S. Staab, Measuring similarity between ontologies. In: *Proceedings of the 13th Conference on Information and Knowledge Management* (EKAW), 2002.

References

[Mädche and Staab, 2000] – A. Maedche, S. Staab. Semi-automatic Engineering of Ontologies from Text. In: Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering, 2000.

[Mädche et al. 2002] - A. Maedche, G. Neumann, S. Staab. Bootstrapping an Ontology-Based Information Extraction System, Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, 2002.

[Mädche et al. 2002] - A. Maedche and V. Pekar and S. Staab. Ontology Learning Part One - On Discovering Taxonomic Relations from the Web. In: Web Intelligence, pp. 301-322, Springer, 2002.

[Martin and Eklund 2000] – P. Martin and P. Eklund. *Knowledge Indexation and Retrieval and the Word Wide Web.* In: IEEE Intelligent Systems, Special Issue "Knowledge Management and Knowledge Distribution over the Internet", 2000.

[Mulholland et al. 2001] – P. Mulholland, Z. Zdrahal, J. Domingue, M. Hatala, A. Bernardi. A Methodological Approach to Supporting Organizational Learning. International Journal of Human-Computer Studies, 55 (3), 337-367, 2001.

[Navigli and Velardi, 2004] - R. Navigli, P. Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Websites, Computational Linguistics (30-2), MIT Press, 2004.

[Navigli and Velardi 2004b] – R. Navigli and P. Velardi. Structural Semantic Interconnection: a Knowledge-Based Approach to Word Sense Disambiguation. In *Proceedings of ACL SENSEVAL-3 Workshop on Sense Evaluation*, pp. 179-182. 2004.

[Nedellec and Nazarenko 2005] – C. Nedellec and A. Nazarenko. Ontology and Information Extraction: A Necessary Symbiosis, In: Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.

References

[Nirenburg and Raskin, 2004] – S. Nirenburg and V. Raskin. Ontological Semantics SERIES: Language, Speech, and Communication, MIT Press, 2004.

[Ogden and Richards, 1923] – C.K. Ogden, I. A. Richards. The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism. 8th ed. 1923. Reprint, New York: Harcourt Brace Jovanovich, 1946.

[Pantel and Lin 2003] - P. Pantel, D. Lin, Automatically Discovering Word Senses. In: Proceedings of HLT-NAACL, 2003.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. To appear in *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*. Sydney, Australia

[Pasca and Harabagiu 2001] - M. Pasca, S. Harabagiu, The Informative Role of WordNet in Open-Domain Question Answering. In: Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pp. 138-143, 2001.

[Ravichandran et al. 2005] - D. Ravichandran, P. Pantel, E. Hovy. Randomized Algorithms and NLP: Using Locality Sensitive Hash Functions for High Speed Noun Clustering. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2005.

References

[Reinberger et al., 2004] - M.-L. Reinberger, P. Spyns, A.J. Pretorius, and W. Daelemans, Automatic initiation of an ontology, in R. Meersman, Z. Tari et al. (eds.), On the Move to Meaningful Internet Systems, LNCS 3290 , Springer, 600–617, 2004.

[Riloff, 1993] - E. Riloff, W. Lehnert. Automated Dictionary Construction for Information Extraction from Text. In: *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications*. IEEE Computer Society Press. pp. 93–99, 2003.

[Rinaldi et al., 2005] - Fabio Rinaldi, Elia Yuste, Gerold Schneider, Michael Hess, David Roussel. Exploiting Technical Terminology for Knowledge Management. In: P.Buitelaar, P. Cimiano, B. Magnini (eds.), Ontology Learning and Population, IOS Press, 2005.

[Resnik 1993] - P. Resnik. Selection and Information: A Class-Based Approach to Lexical Relationships. PhD Thesis, University of Pennsylvania, 1993.

[Resnik 1998] – P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11, pp. 95-130, 1998.

[Ribas 95] - F. Ribas. On learning more appropriate selectional restrictions. In: Proceedings of the 7th Conference of the European chapter of the Association for Computational Linguistics (EACL), pp. 112-118, 1995.

[Sanchez, Moreno, 2005] - D. Sanchez and A. Moreno, Web-scale taxonomy learning, In: Proceedings of the ICML Workshop on Extending and Learning Lexical Ontologies using Machine Learning, 2005.

References

[Saussure 1916] – Ferdinand de Saussure. Cours de linguistique générale. Ed. Charles Bally and Albert Sechehaye in collaboration with Albert Riedlinger. Paris: Payot, 1916.

[Schlobach et al. 2004] - S. Schlobach, M. Olsthoorn,M. de Rijke. Type checking in open-domain question answering. In: Proceedings of the European Conference on Artificial Intelligence (ECAI), 2004.

[Schutz and Buitelaar, 2005] – A. Schutz, P. Buitelaar RelExt: A Tool for Relation Extraction in Ontology Extension. In: Proceedings of the 4th International Semantic Web Conference, 2005.

[Schütze 1993] – H. Schütze. Word space, Advances in Neural Information Processing Systems 5, pp. 895-902, 1993.

[Sintek et al. 2004] – M. Sintek, P. Buitelaar, D. Olejnik. A Formalization of Ontology Learning from Text. In: Proceedings of the ISWC Workshop on Evaluation of Ontology-based Tools (EON2004), 2004.

[Sinha and Narayanan 2005] – S. Sinha and S. Narayanan. Model Based Answer Selection. In: Proceedings of the AAAI Workshop on Textual Inference in Question Answering, 2005.

[Smith and Poulter 1999] - H. Smith, K. Poulter. Share the Ontology in XML-based Trading Architectures. In: Communications of the ACM 42(3):110-111, 1999.

References

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. Machine Learning

[Staab and Schnurr 2000] - S. Staab and H.-P. Schnurr. Smart Task Support through Proactive Access to Organizational Memory. Journal of Knowledge-based Systems}, Elsevier, 2000.

[Stevenson 2005] – M. Stevenson, M. Greenwood. A Semantic Approach to IE Pattern Induction, In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005.

[Sure 2003] – Y. Sure, Methodology. Tools and Case Studies for Ontology based Knowledge Management. PhD Thesis, University of Karlsruhe, Institute AIFB, 2003.

[Sure et al. 2000] - Y. Sure and A. Maedche and S. Staab. Leveraging Corporate Skill Knowledge -- From ProPer to OntoProPer, In: Proceedings of PAKM, pp. 1-9, 2000.

[Turcato et al. 2000] – D. Turcato, F. Popowich, J. Toole, D. Fass, D. Nicholson and G. Tisher. Adapting a synonym database to specific domains. In: Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval, 2000.

[Turney 2001] – P. Turney, Mining the Web for synonyms: PMI-IR vs LSA on Toeffl, In: Proceedings of the 12th European Conference on Machine Learning(ECML), pp. 401-502, 2000.

[Uschold and Grunninger 1996] - M. Uschold, M. Grunninger. Ontologies. Principles, Methods and Applications. *Knowledge Engineering Review* 11, 1996.

[Uschold et al. 1998] - M. Uschold and M. King and S. Moralee and Y. Zorgios. The Enterprise Ontology, In: Knowledge Engineering Review, 13(1), pp. 31-89, 1998.

References

[Velardi et al., 2005] - P. Velardi, R. Navigli, A. Cucchiarelli, F. Neri. Evaluation of OntoLearn, a Methodology for Automatic Learning of Domain Ontologies, In: P.Buitelaar, P. Cimiano, B. Magnini (eds.), Ontology Learning and Population, IOS Press, 2005.

[Widdows 2003] - D. Widdows. Unsupervised method for developing taxonomies by combining syntactic and statistical information. In: Proceedings of HLT/NAACL, pp. 276-283, 2003.

[Wiederhold 1992] - G. Wiederhold. Mediators in the architecture of future information systems. In: IEEE Computer 25(3):38-49, 1992.