

---

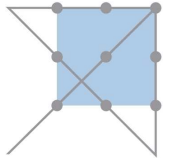
# Using Wikipedia as a Resource for Natural Language Processing (2)

**Simone Paolo Ponzetto** and **Michael Strube**  
EML Research gGmbH, Heidelberg

<http://www.eml-research.de/nlp>

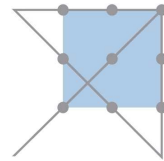
# Outline

---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. *Knowledge derived from Wikipedia*
  - (a) Semantic relatedness
  - (b) WikiRelate! Computing semantic relatedness using Wikipedia
  - (c) Knowledge bases, taxonomies
  - (d) Deriving a taxonomy from Wikipedia
5. Exploiting Wikipedia for Coreference Resolution
6. Further applications
7. Conclusions

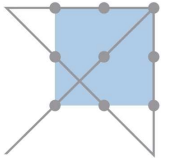
# Knowledge derived from Wikipedia



We present our work on using Wikipedia as a resource for NLP.

- ▣▣▣▣▣ WikiRelate! Computing Semantic Relatedness Using Wikipedia (AAAI '06)
- ▣▣▣▣▣ Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution (HLT-NAACL '06)
- ▣▣▣▣▣ Knowledge Derived from Wikipedia for Computing Semantic Relatedness (Journal of AI Research '07)
- ▣▣▣▣▣ Deriving a Large Scale Taxonomy from Wikipedia (AAAI '07)

# Knowledge derived from Wikipedia



All papers can be downloaded from EMLR's NLP group website!

An overview of the work can be found in

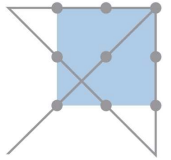
- ▣ Knowledge Derived from Wikipedia for Computing Semantic Relatedness (Journal of AI Research '07)

Core ideas:

1. derive a knowledge base (KB) from Wikipedia
2. embed it as component in NLP applications

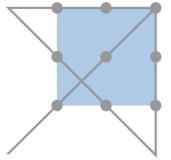
# Outline

---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. Knowledge derived from Wikipedia
  - (a) *Semantic relatedness*
  - (b) WikiRelate! Computing semantic relatedness using Wikipedia
  - (c) Knowledge bases, taxonomies
  - (d) Deriving a taxonomy from Wikipedia
5. Exploiting Wikipedia for Coreference Resolution
6. Further applications
7. Conclusions

# Semantic relatedness



- WikiRelate! Computing Semantic Relatedness Using Wikipedia (Strube and Ponzetto, AAAI-06)

<http://www.eml-research.de/~strube/papers/aaai06.pdf>



What is semantic relatedness?



Why do we need it for NLP applications?



How is it computed?

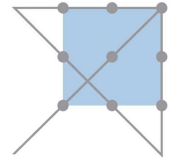


How do we compute it in Wikipedia?



How good is the performance?

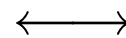
# What is semantic relatedness?



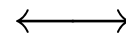
Semantic relatedness: how much words/texts are correlated in meaning to each other.

word<sub>1</sub>/text<sub>1</sub> ↔ word<sub>2</sub>/text<sub>2</sub>

football

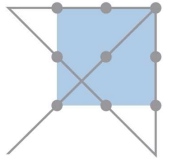


sport



# How much are words related?

---

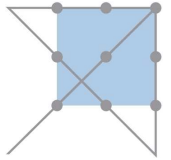


E.g. to create a humanly annotated dataset of word pairs:

- take a **set of human judges**
- take a **set of word pairs**
- for each human judge, **ask to rate each word pair with a numerical semantic similarity score**, e.g. between 1 and 10 (1 = words are unrelated, 10 = words are closely related)
- take the mean score among judges to represent the correlation between words in each pair

# How much are words related?

---

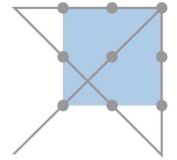


on a scale from 1 to 10 ...

<b>love</b>	<b>sex</b>	<b>6.77</b>
<b>money</b>	<b>cash</b>	<b>9.15</b>
<b>Maradona</b>	<b>football</b>	<b>8.62</b>
<b>holy</b>	<b>sex</b>	<b>1.62</b>

source: the WordSimilarity-353 Test Collection (Finkelstein et al., 2002)

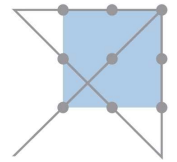
# Semantic relatedness is useful



Examples of applications of semantic relatedness measures:

- word sense disambiguation (Patwardhan, Banerjee and Pedersen, SemEval 2007): **assign a target word the sense that is most related to the senses of its neighboring words**
- query expansion: **enhance a query by including terms closely related to the original ones**
- noun compound interpretations (Kim & Baldwin, 2005)
- spelling correction (Budanitsky & Hirst, 2006)
- ...
- coreference resolution (Ponzetto & Strube, HLT-NAACL 2006)

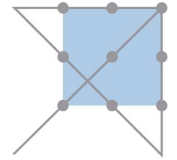
# Semantic relatedness and WSD



Assign the sense most related to the context → maximize lexical coherence:

- In 1834, Sumner was admitted to the **bar** at the age of twenty-three, and entered private practice in Boston.
  - ➡ *professional body of lawyers*
- It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every **bar**.
  - ➡ *segment of time in music*
- Jenga is a popular beer in the **bars** of Thailand.
  - ➡ *retail establishment serving alcoholics*

# Semantic relatedness and QE



Add search terms related to the original query → **include related documents in the result list.**

- ▶ i.e. relevant information about *football* could be also in the pages about *sport*. Cf. query 'football results' ...

The screenshot shows the uefa.com website. The header includes the uefa.com logo and navigation links for UEFA CHAMPIONS LEAGUE and UEFA CUP. A sidebar on the left contains a message about Macromedia flash player. The main content area lists several news items:

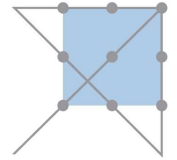
- Deschamps leaves Juventus**: Didier Deschamps has resigned as coach of Juventus just a week after securing their return to Serie A.
- Elmander fires Toulouse into Europe**: Johan Elmander scored a hat-trick to give Toulouse FC a in the UEFA Champions League qualifying stage.
- Nürnberg stun ten-man Stuttgart**: 1. FC Nürnberg won the German Cup for the fourth time as champions VfB Stuttgart after extra time.
- Živkovic to miss Croatia qualifiers**: HNK Hajduk Split defender Boris Živkovic is out of Croatia and Russia after suffering a knee injury.
- Madrid and Barça stay the course**: Real Madrid CF and FC Barcelona's Spanish title hopes are Getafe CF but Valencia CF's bid is over.

At the bottom, there are sections for 'City Guide' and 'ONLINE VIDEO'.

The screenshot shows the BBC Sport website. The header includes the BBC logo and navigation links for Home, News, Sport, Radio, TV, Weather, and Languages. The main content area features a large image of Fernando Alonso and Lewis Hamilton celebrating, with the headline 'Alonso puts Hamilton his place'. Below the image, there is a section for 'MOTORSPORT' and a 'BBC WORLD SERVICE' section with radio schedules. At the bottom, there are 'HIGHLIGHTS AND FEATURES ON THE BIG STORIES' including 'Owen delighted to be back for England' and 'Logan criticises S decisions'.

... are both relevant!

# Measures of semantic relatedness



## Two main classes of algorithms

- **Taxonomy-based** algorithms:  
based on the distance of words in a semantic network  
e.g. WordNet or ... Wikipedia!
- **Distributional algorithms**:  
compare words based on their distributional context

# Distributional measures of relatedness



Firth (1957): *You shall know a word by the company it keeps*

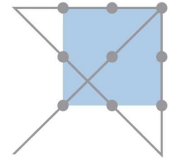
meaning  $\longleftrightarrow$  contextual occurrence

- Nida's (1975) example:
  - A bottle of **tezgüino** is on the table
  - Everybody likes **tezgüino**
  - **tezgüino** makes you drunk
  - We make **tezgüino** out of corn
- ➡ one could guess the meaning of **tezgüino** just by looking at these contexts

similar words  $\longleftrightarrow$  similar contexts

# Contextual similarity

---



- ? Why do we look at **contexts**, rather than directly at **word cooccurrence**?
- ! Words can have an *indirect association*! E.g.
- ➡ synonyms (*car, auto*) and highly similar words (*beer, wine*) do not necessarily co-occur together very often, but DO co-occur in similar contexts.

# Distributional measures of relatedness



## Three main steps

### 1. Build a vector space model for words

- ▣▣▣▣ represent each word as a **vector of co-occurrence values** with other words

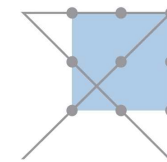
### 2. Choose a metric to weight the vector elements

- ▣▣▣▣ assign each element in the vector a value which represents **the strength of association** with the word, e.g. frequencies, probabilities or some measure of direct association such as mutual information

### 3. compute relatedness as a vector distance metric

- ▣▣▣▣ assign a relatedness score to two given words by **taking their vectors and computing their distance**, e.g. the cosine or Euclidean distance

# Vector spaces



similar words  $\longleftrightarrow$  similar contexts ( **$\sim$ word vectors**)

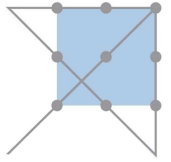
A vector space is a set of vectors with elements (scalars) from a field (such as the real or complex numbers) with two operations defined

- vector addition
- scalar multiplication (with a vector)
- certain axioms need to be satisfied for these operations (associativity, commutativity, distributivity)

How do we map words to vector spaces?!



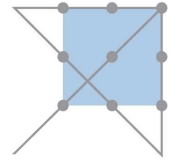
# Word vectors



- ! again, we map our **input** (i.e. *words*) to a **vectorial space** where the vectors elements are *word occurrences*

	<b>eat</b>	<b>hot</b>	<b>jazz</b>	<b>meat</b>	<b>trumpet</b>	...
?	?	?	?	?	?	...

# Word vectors

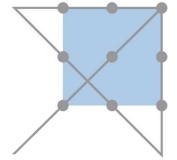


HOT-FROM-THE-OVEN MEALS: Keep

**hot food HOT**, warm isn't good enough. Set the oven temperature at 140 degrees or hotter. Use a **meat** thermometer. And cover with foil to keep food moist. **Eat** within two hours.

	<b>eat</b>	<b>hot</b>	<b>jazz</b>	<b>meat</b>	<b>trumpet</b>	...
<b>FOOD</b>	1	2	0	1	0	...

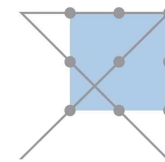
# Word vectors



“Change is always happening,” said the ebullient trumpeter, whose words tumble out almost as fast as notes from his **trumpet**. “That’s one of the wonderful things about **jazz music**.” For many **jazz** fans, Ferguson is one of the wonderful things about **jazz music** .

	<b>eat</b>	<b>hot</b>	<b>jazz</b>	<b>meat</b>	<b>trumpet</b>	...
<b>FOOD</b>	1	2	0	1	0	...
<b>MUSIC</b>	0	0	3	0	1	...

# Word association strength



We do not necessarily use *frequencies*

- conditional probability

$$\text{assoc}_{\text{prob}}(w_1, w_2) \sim p(w_1|w_2) = \frac{\text{count}(w_1, w_2)}{\text{count}(w_2)}$$

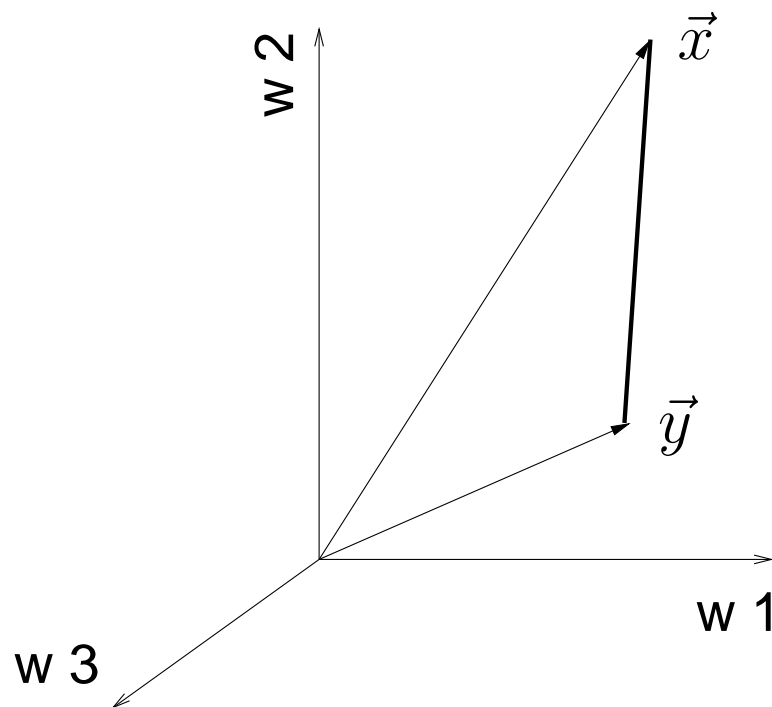
- pointwise mutual information (PMI)

$$\text{assoc}_{\text{PMI}}(w_1, w_2) \sim \text{PMI}(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

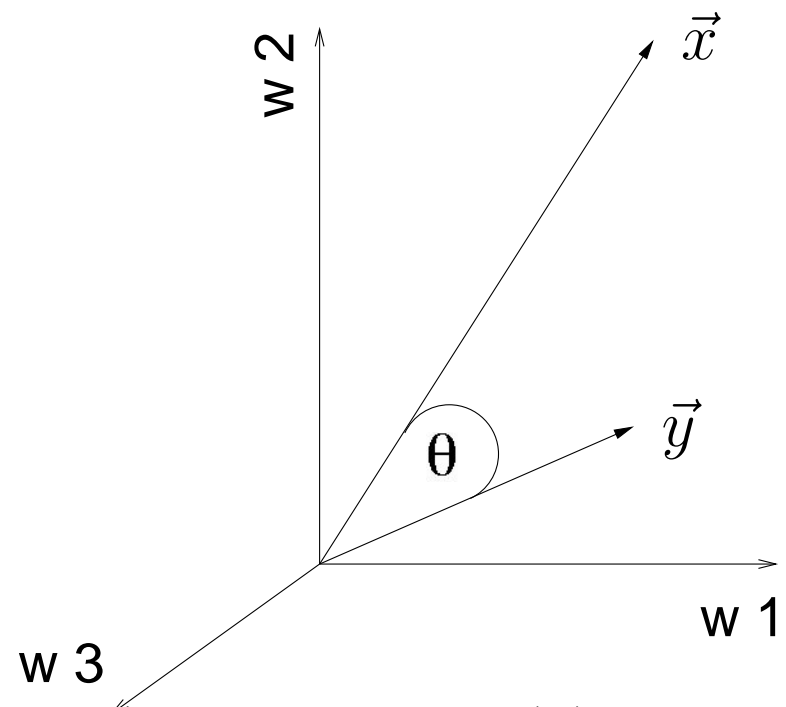
# Relatedness as word vector distance



- words are (represented as) vectors
- relatedness between words is given by the **distance between their vectors**



$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

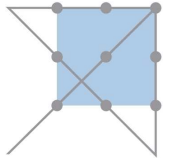


$$\theta = \arccos \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|}$$



# WordNet

---

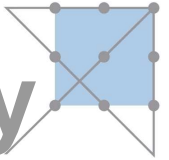


A large taxonomy is provided by WordNet

- at the lowest level, a **list of words**
- words are grouped into sets of synonyms (**synsets**)
  - ➡ *car, auto*
  - ➡ *queue, line*
- synsets are connected via **meaning-based relations**
  - ➡ *is-a*
  - ➡ *part-of*

➡ interconnected semantic network

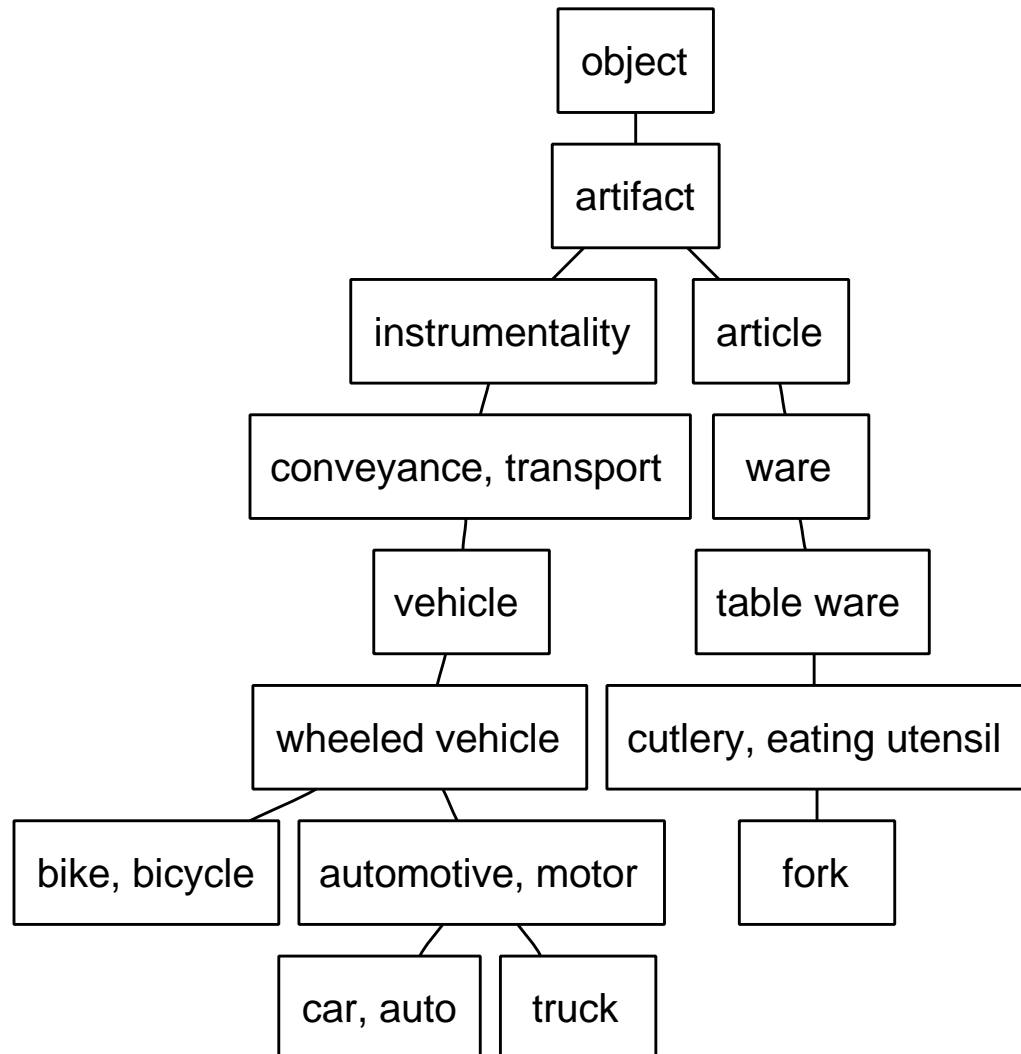
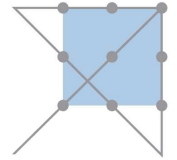
# Semantic relatedness in a taxonomy



To compute the semantic similarity/relatedness of concepts

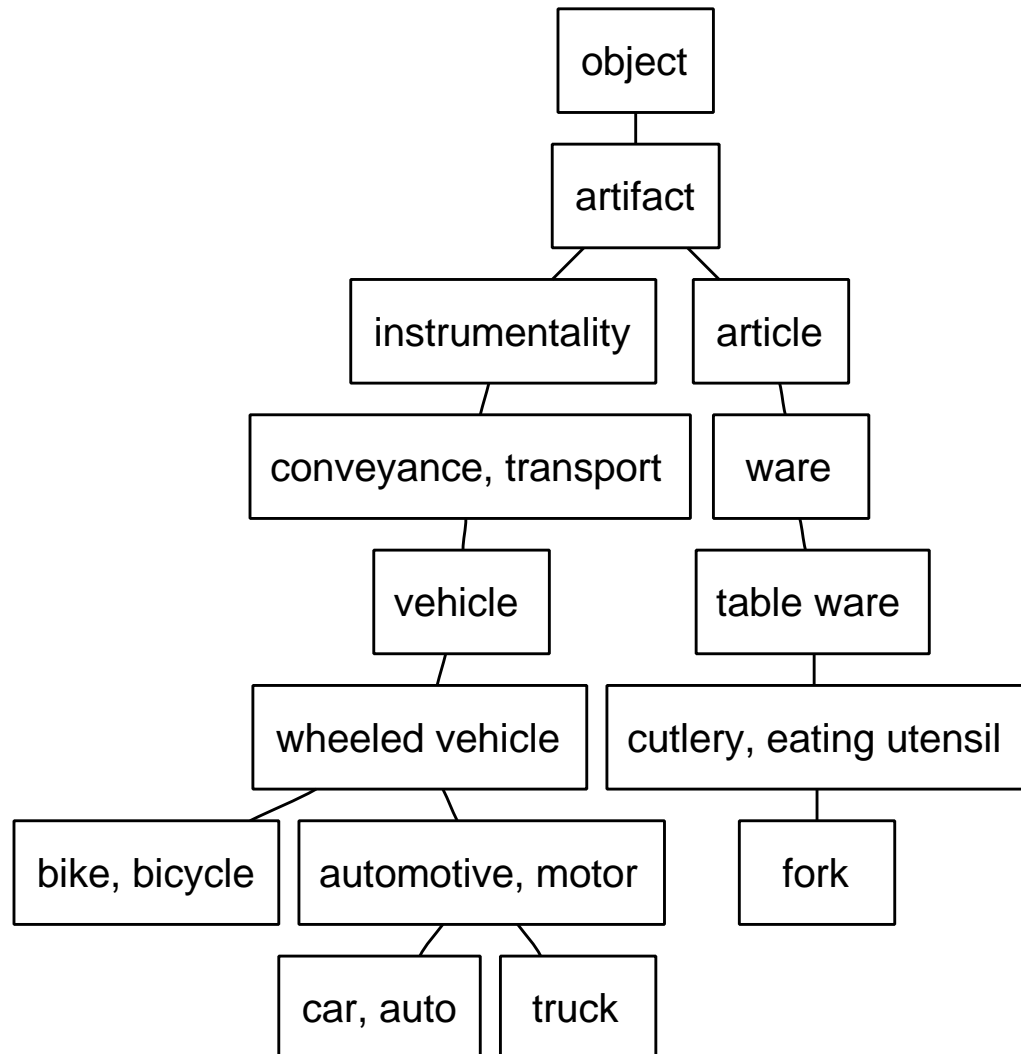
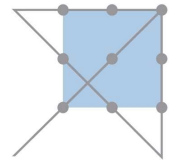
- given lexical resources  
(e.g. lexical database, semantic network)
  1. transform the resource into a **network** or **graph**
  2. compute similarity/relatedness using **paths** in it
- **similarity** → use *is-a* links only (e.g. car, bike)
- **relatedness** → use *all* relations (e.g. car, gasoline)

# Measures of semantic relatedness



*how to compute distance?!*

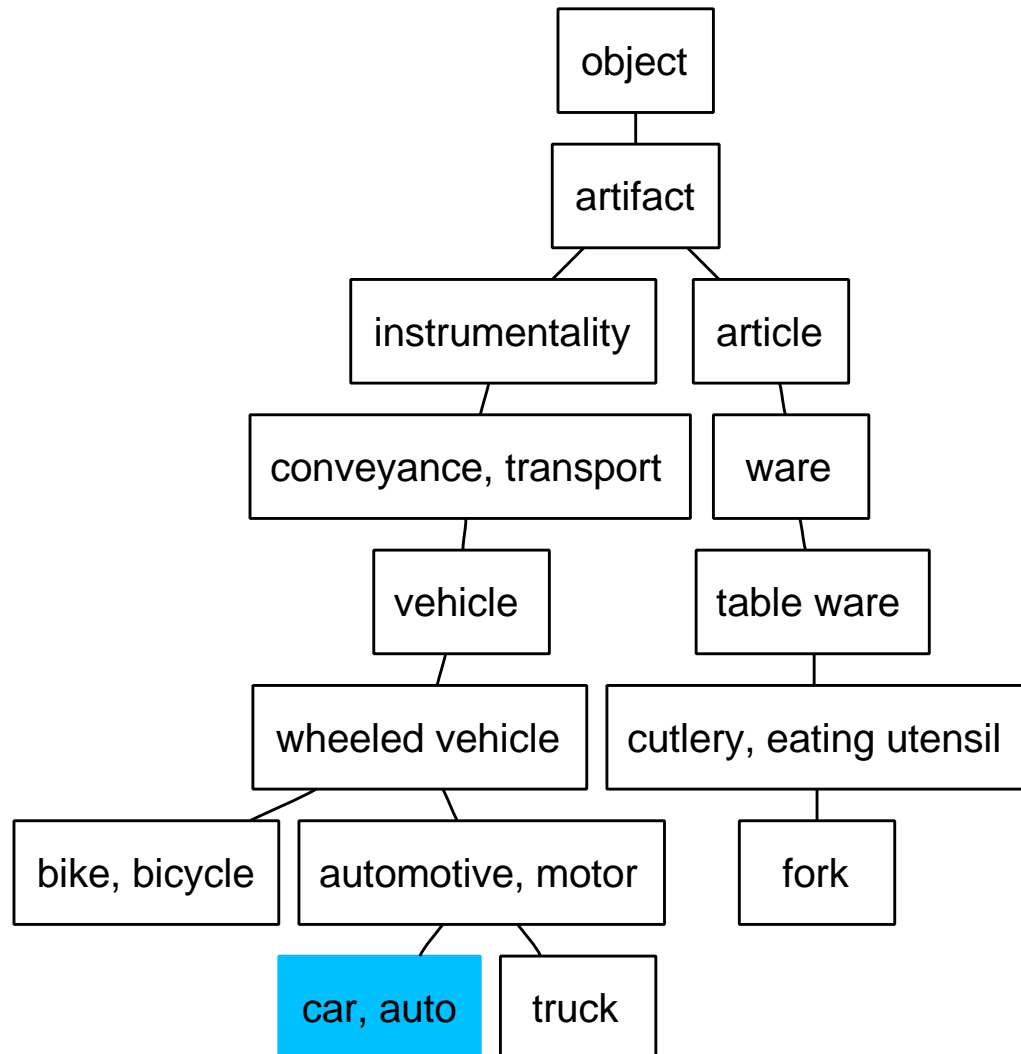
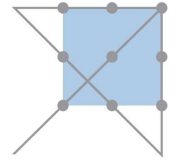
# Measures of semantic relatedness



e.g. using *node counting scheme*

$$\text{sim}(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}$$

# Measures of semantic relatedness

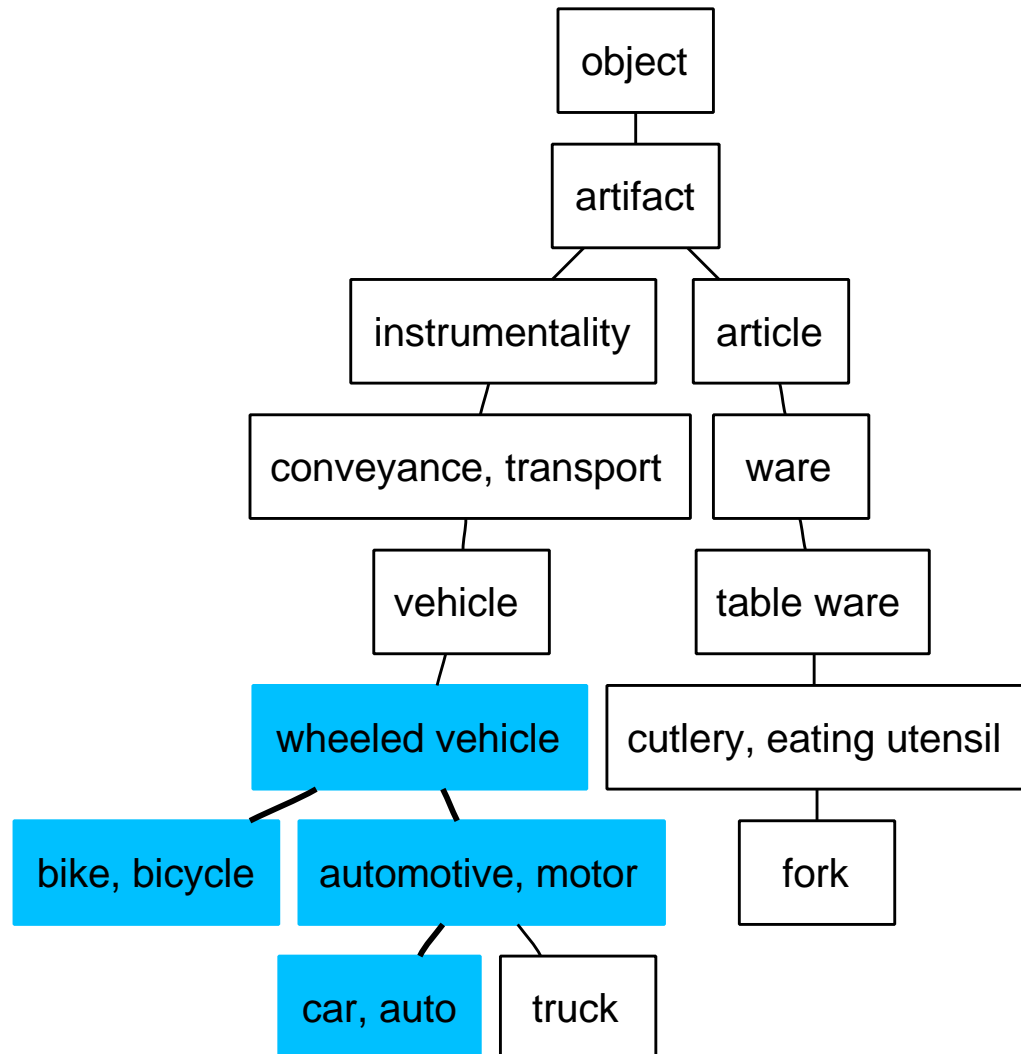
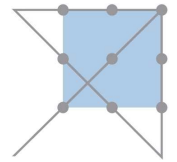


e.g. using *node counting scheme*

$$\text{sim}(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}$$

- $\text{sim}(\text{car}, \text{auto}) = 1$

# Measures of semantic relatedness

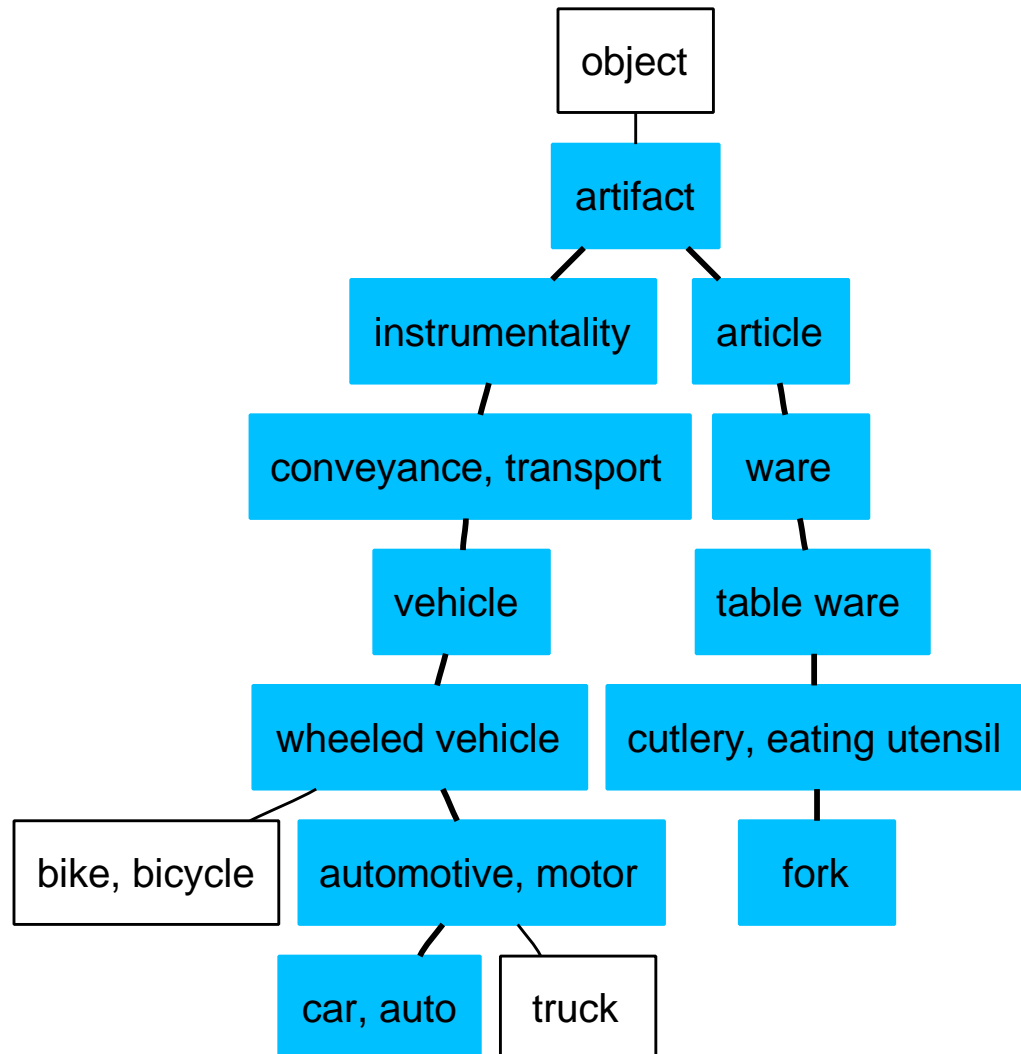
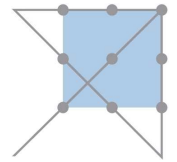


e.g. using *node counting scheme*

$$\text{sim}(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}$$

- $\text{sim}(\text{car}, \text{auto}) = 1$
- $\text{sim}(\text{car}, \text{bike}) = 0.25$

# Measures of semantic relatedness

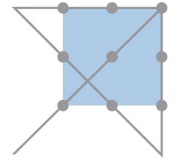


e.g. using *node counting scheme*

$$sim(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}$$

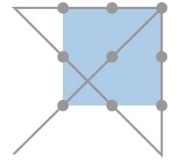
- $sim(car, auto) = 1$
- $sim(car, bike) = 0.25$
- $sim(car, fork) = 0.08$

# Measures of semantic relatedness



- the edge counting scheme (**Rada et al., 1989, p1**) assumes a uniform modeling of the hierarchy
  - BUT depth in the taxonomy yields different levels of concept granularity
  - e.g. *car* to *motor* seems closer than *transport* to *instrumentality*!
- ⇒ we want a metric which lets us represent the cost of each edge independently
- ⇒ new measures were developed to take into account node depth and concept granularity

# Path based measures



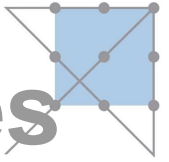
- **Leacock & Chodorow (1998, *lch*)**: a normalized path-length measure which uses the *depth of the taxonomy in which the concepts are found*.

$$lch(c_1, c_2) = -\log \frac{\text{length}(c_1, c_2)}{2D}$$

- **Wu & Palmer (1994, *wup*)**: a scaled measure which uses *the depth of the nodes together with the depth of their least common subsumer, lcs*.

$$wup(c_1, c_2) = \frac{\text{depth}(lcs_{c_1, c_2})}{\text{depth}(c_1) + \text{depth}(c_2)}$$

# Information content based measures



- **Resnik (1995, *res*)**: relatedness between the concepts as a *function of the information content* of their least common subsumer.

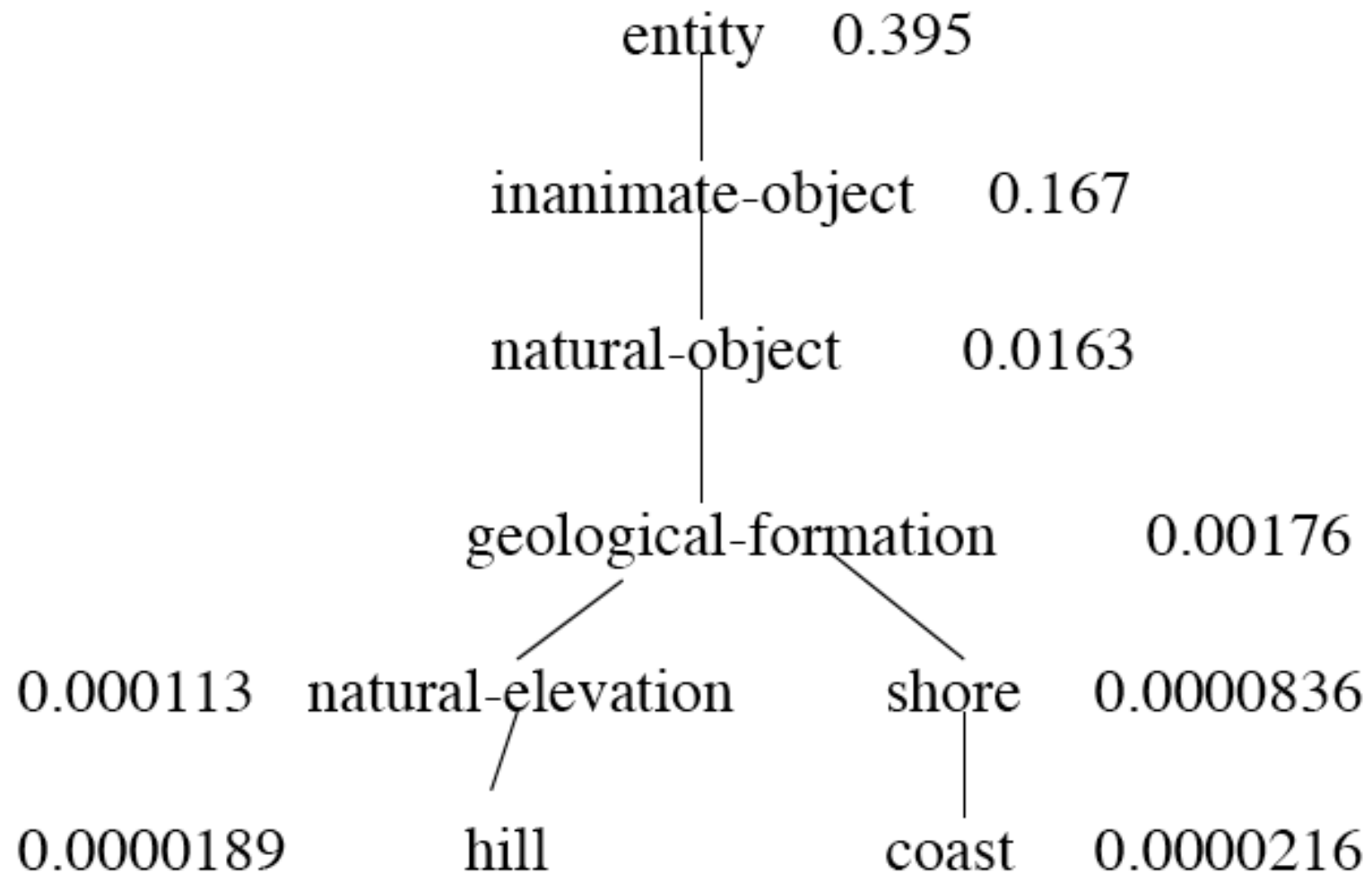
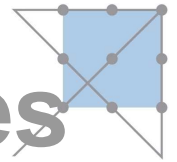
**Information content** the probability of occurrence of a concept in a corpus.

$$res(c_1, c_2) = ic(lcs_{c_1, c_2})$$

$$ic(c) = -\log(p(c))$$

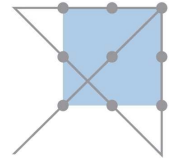
$$p(c) = \frac{\sum_{w \in words(c)} count(c)}{N}$$

# Information content based measures



# Gloss overlap measures

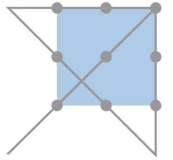
---



- **Lesk (1986)**: relatedness between two words as a *function of the shared words* (overlaps).
  - ⇒ **MEXICO**: a republic in southern North America;  
became independent from Spain in 1810
  - ⇒ **ARGENTINA**: a republic in southern South America;  
second largest country in South America
- **Banerjee & Pedersen (2003)**: *extended gloss overlap* (*lesk*) measure which computes overlaps by using all glosses of the related concepts

# WordNet::Similarity

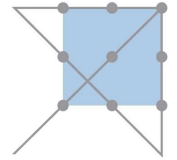
---



- software package for computing semantic relatedness/similarity using WordNet (Pedersen et al., 2004)
- available at <http://www.d.umn.edu/~tpederse/similarity.html>

# Take-home message 4

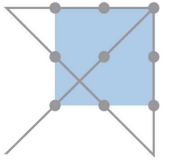
---



- distributional similarity is computed on the basis of large corpora – they are computationally expensive and do not scale well
- easy to implement but require very large corpora
- semantic relatedness is computed using semantic networks – their quality pretty much depends on the availability and size of the resource
- API for WordNet available

# Outline

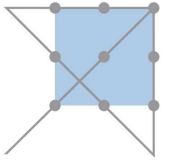
---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. Knowledge derived from Wikipedia
  - (a) Semantic relatedness
  - (b) *WikiRelate! Computing semantic relatedness using Wikipedia*
  - (c) Knowledge bases, taxonomies
  - (d) Deriving a taxonomy from Wikipedia
5. Exploiting Wikipedia for Coreference Resolution
6. Further applications
7. Conclusions

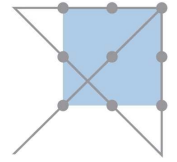
# WikiRelate!

---



- three main steps
  1. page retrieval and disambiguation
  2. category tree search
  3. relatedness measure computation

# Wikipedia page retrieval

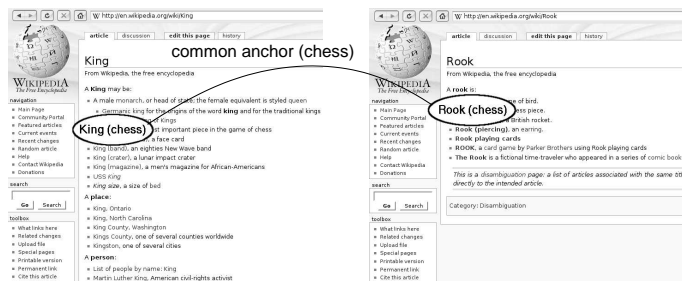


Given a pair of words, for each word

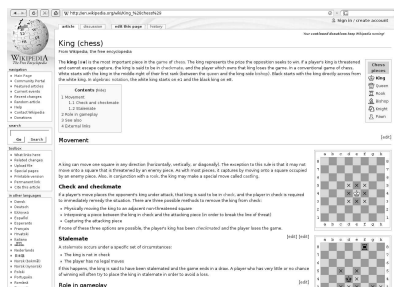
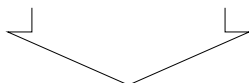
- query the page titled as the word
- if it's a **disambiguation page**, look for links overlapping the page obtained querying the other word
- we disambiguate by
  - ➡ maximizing relatedness,
  - ➡ taking advantage of text structure
  - ➡ DEMO! **King** and **Rook**

query *king*

query *rook*

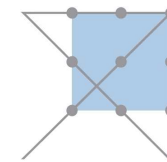


pages bootstrapped  
(no disambiguation)



disambiguated  
page  
**KING (CHESS)**

# Gloss and text overlap measures

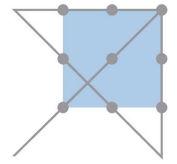


Given the text from the Wikipedia pages, we compute gloss/text overlap via a *double normalization step*

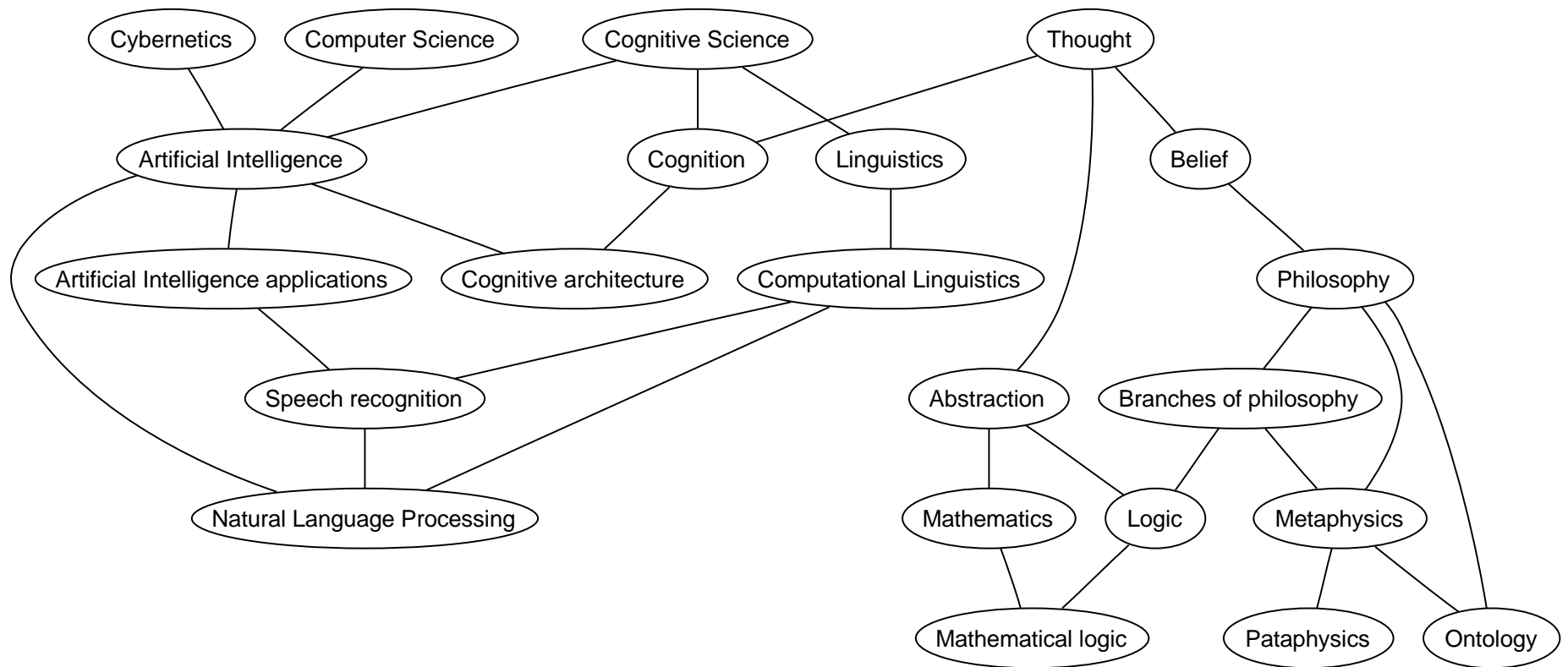
$$\text{relate}_{\text{gloss/text}}(t_1, t_2) = \tanh \left( \frac{\text{overlap}(t_1, t_2)}{\text{length}(t_1) + \text{length}(t_2)} \right)$$

where  $\text{overlap}(t_1, t_2) = \sum_n m^2$  for  $n$  phrasal  $m$ -word overlaps (Banerjee & Pedersen, 2003)

# Wikipedia category tree



- Since May 2004 Wikipedia provides a *collaboratively generated taxonomy (folksonomy)*



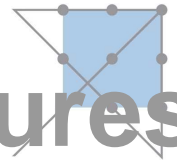
# Path and information content measures



---

- we ported the *path length* based measures from Rada et al., (1989), Wu & Palmer (1994), Leacock & Chodorow (1998), and the *information content* measure from Resnik (1995) to Wikipedia
  1. **extract categories** from the Wikipedia pages
  2. **search** for a connecting **path** along the category tree
  3. **score the paths** found
  4. **return** the one(s) satisfying the measure definitions (i.e. **shortest, most informative**)

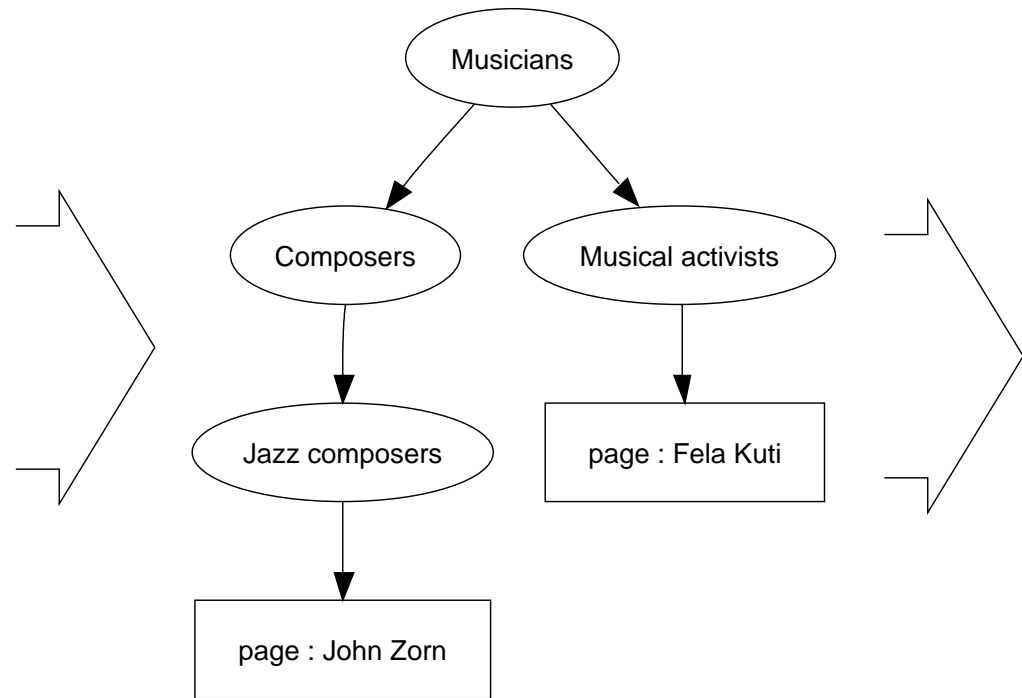
▶ DEMO



# Path and information content measures

The top screenshot shows the Wikipedia page for John Zorn. The category list at the bottom includes 'Jazz composers', which is circled in red. The bottom screenshot shows the Wikipedia page for Fela Kuti. The category list at the bottom includes 'Musical activists', which is circled in red.

page query and retrieval  
category extraction

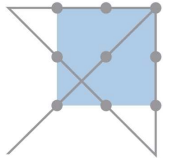


search for a connecting path  
along the category tree

relatedness measure(s) computation

# Experiments using similarity lists

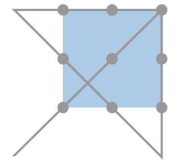
---



we experimented with the similarity lists from

- Miller & Charles (1991, M&C, 30 word pair)
- Rubenstein & Goodenough (1965, R&G, 65 word pairs)
- the WordSimilarity-353 Test Collection (Finkelstein et al., 2002, 353-TC) — full (353 word pairs) and test (153 pairs).
- Gurevych's (2005) German translation of Rubenstein & Goodenough's list.

# Experiments using similarity lists

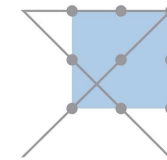


word 1	word 2	human judges	system
love	sex	6.77	?
money	cash	9.15	?
Maradona	football	8.62	?
holy	sex	1.62	?

- evaluation metric: Pearson's  $r$  correlation coefficient
- baseline: Jaccard similarity coefficient on Google page hits

$$jaccard = \frac{Hits(w1 \text{ AND } w2)}{Hits(w1) + Hits(w2) - Hits(w1 \text{ AND } w2)}$$

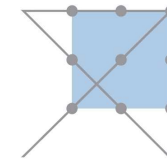
# Experiments using similarity lists



## Miller & Charles

	Google	WordNet					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>lesk</i>	
all	0.26	0.71	0.77	<b>0.82</b>	0.78	0.37	
	Google	Wikipedia					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>gloss</i>	<i>text</i>
all	0.26	0.45	0.40	0.41	0.23	<b>0.46</b>	<b>0.46</b>

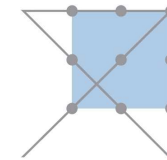
# Experiments using similarity lists



## Rubenstein & Goodenough

	Google	WordNet					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>lesk</i>	
all	0.41	0.78	0.82	<b>0.86</b>	0.81	0.34	
	Google	Wikipedia					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>gloss</i>	<i>text</i>
all	0.41	<b>0.53</b>	0.49	0.50	0.31	0.46	0.46

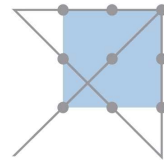
# Experiments using similarity lists



## WordSimilarity-353 TC full

	Google	WordNet					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>lesk</i>	
all	0.18	0.28	0.30	0.34	0.34	0.21	
	Google	Wikipedia					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>gloss</i>	<i>text</i>
all	0.18	0.46	0.48	0.48	0.38	0.20	0.20

# Experiments using similarity lists

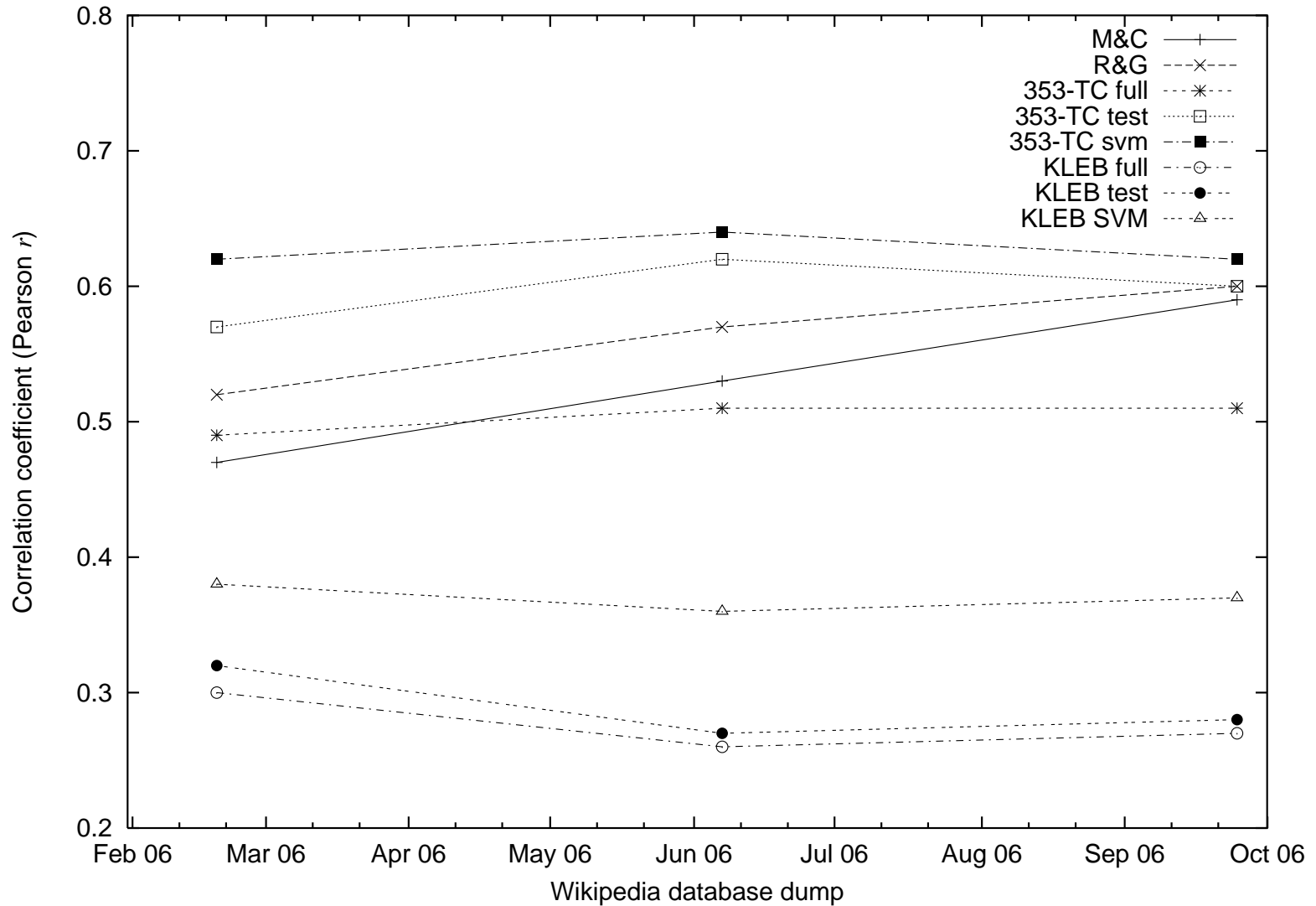
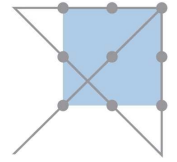


## WordSimilarity-353 TC test

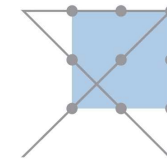
	Google	WordNet					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>lesk</i>	
all	0.27	0.29	0.28	0.35	0.38	0.21	
	Google	Wikipedia					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>gloss</i>	<i>text</i>
all	0.27	0.50	0.54	0.55	0.45	0.22	0.22

➡ best results using SVM-based combined measure **0.59**

# Experiments through time



# Experiments using German list

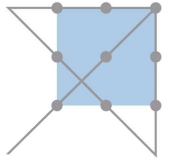


Gurevych

	Google	GermaNet					
	<i>jaccard</i>	<i>lin</i>	<i>res</i>	<i>lesk</i>			
all	0.26	<b>0.66</b>	0.64	0.49			
	Google	Wikipedia					
	<i>jaccard</i>	<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>	<i>gloss</i>	<i>text</i>
all	0.26	0.58	<b>0.65</b>	0.64	0.62	0.33	0.33

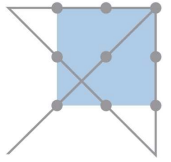
# Discussion

---



- the better performance of Wikipedia on the 353-TC dataset seems to be due to 353-TC being a *relatedness* rather than a *similarity* list
- results are stable through time: this is because **our search is focused**, i.e. we disambiguate
- results can be easily ported to other languages (e.g., German)

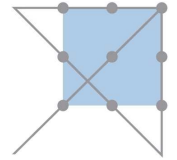
# Case study: coreference resolution



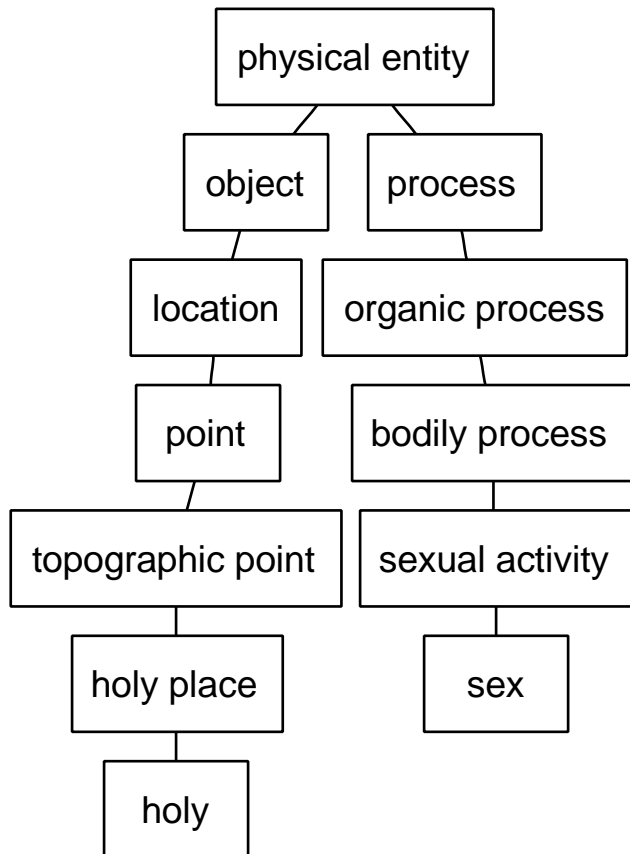
all similarity lists are rather small in size

- ⇒ we perform an **extrinsic evaluation** using the relatedness measures as *features of a machine learning based coreference resolution system*
- ⇒ this way we evaluate by computing relatedness scores for 282,658 word pairs in total!

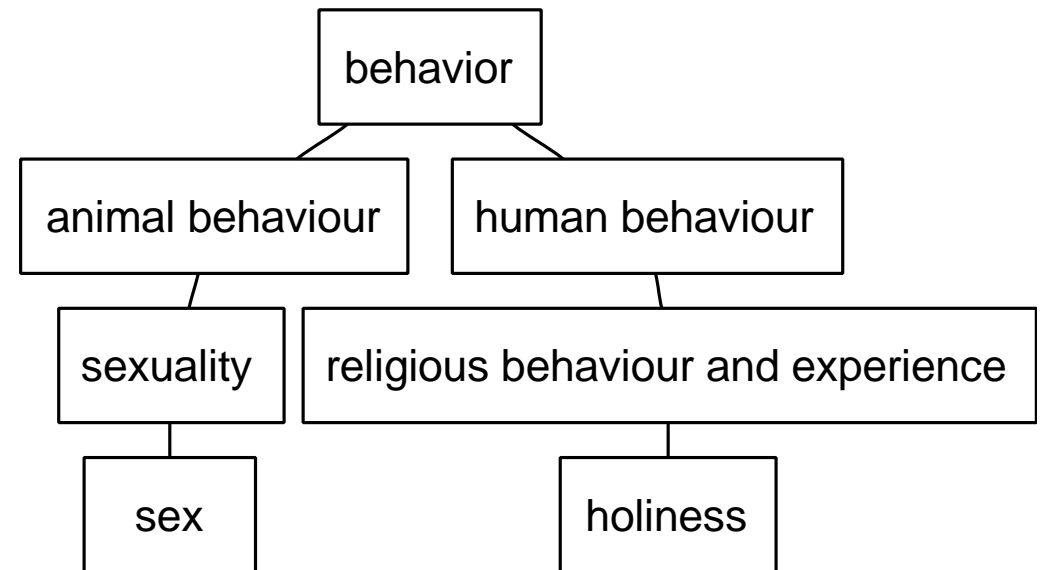
# Example: What is better?



## WordNet

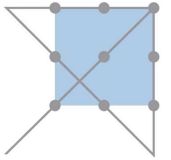


## Wikipedia



# WikiRelate! API

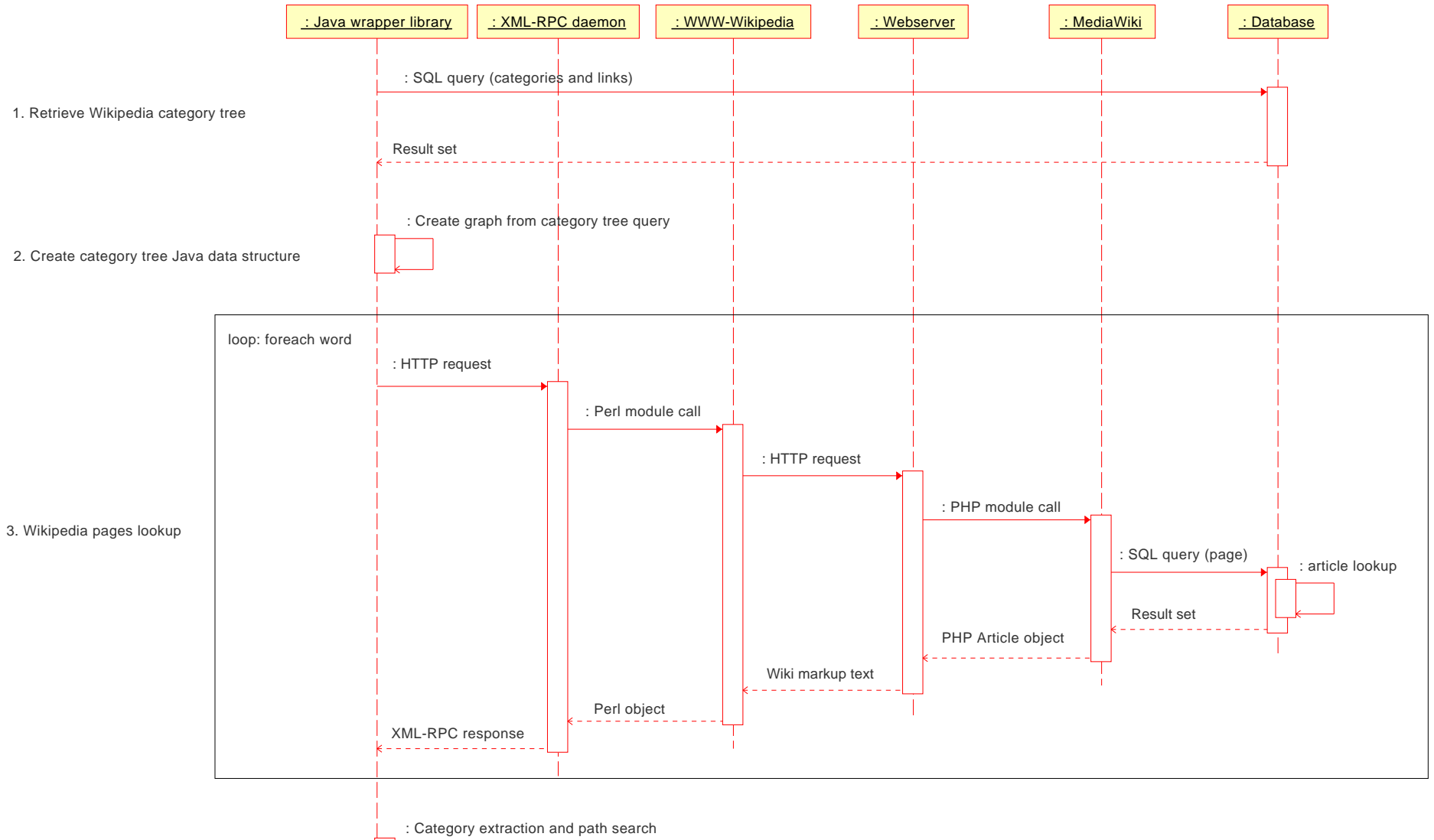
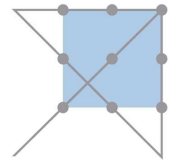
---



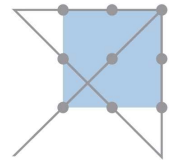
the API computes semantic relatedness by

1. taking a pair of words as input
2. retrieving the Wikipedia articles they refer to (disambiguation)
3. computing paths in the Wikipedia categorization graph
4. returning as output the set of paths found, scored according to the semantic relatedness measure used

# Software Architecture



# Take-home message 5

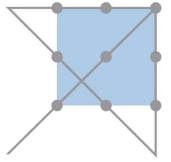


- relatedness measures computed from Wikipedia give a performance on word pair lists *competitive with WordNet*
- ⇒ Wikipedia is a promising resource
- worked 'out of the box' on a first attempt
  - can be freely downloaded and used right away
  - we used only part of it (i.e. limited use of textual content)
  - shows exponential growth

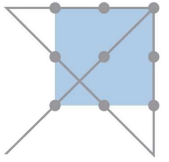
Wikipedia is a resource to be exploited for NLP applications

# Outline

---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. Knowledge derived from Wikipedia
  - (a) Semantic relatedness
  - (b) WikiRelate! Computing semantic relatedness using Wikipedia
  - (c) *Knowledge bases, taxonomies, ontologies*
  - (d) Deriving a taxonomy from Wikipedia
5. Exploiting Wikipedia for Coreference Resolution
6. Further applications
7. Conclusions

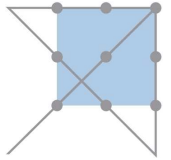


---

These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.

# Don't believe anyone

---

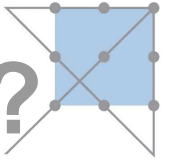


who is talking about

- ontologies
- common sense
- common sense knowledge bases
- universals, particulars and stuff like that

arbitrary (*detail*), incomplete (*coverage*) and wrong (*reliable*).  
We use them only, because they are out there – and we are not able to produce a better one.

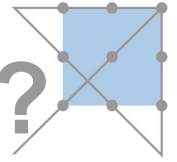
# What is a knowledge base good for?



AI applications:

- inferences!
- knowledge represented in such a way that a reasoner can be used in information processing applications
- distinction between classes/categories and instances
- relations between classes and attributes
- quantifications and restrictions on attributes

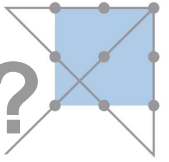
# What is a knowledge base good for?



## Database applications:

- database design
- keyword hierarchies
- consistency, vocabulary
- ...

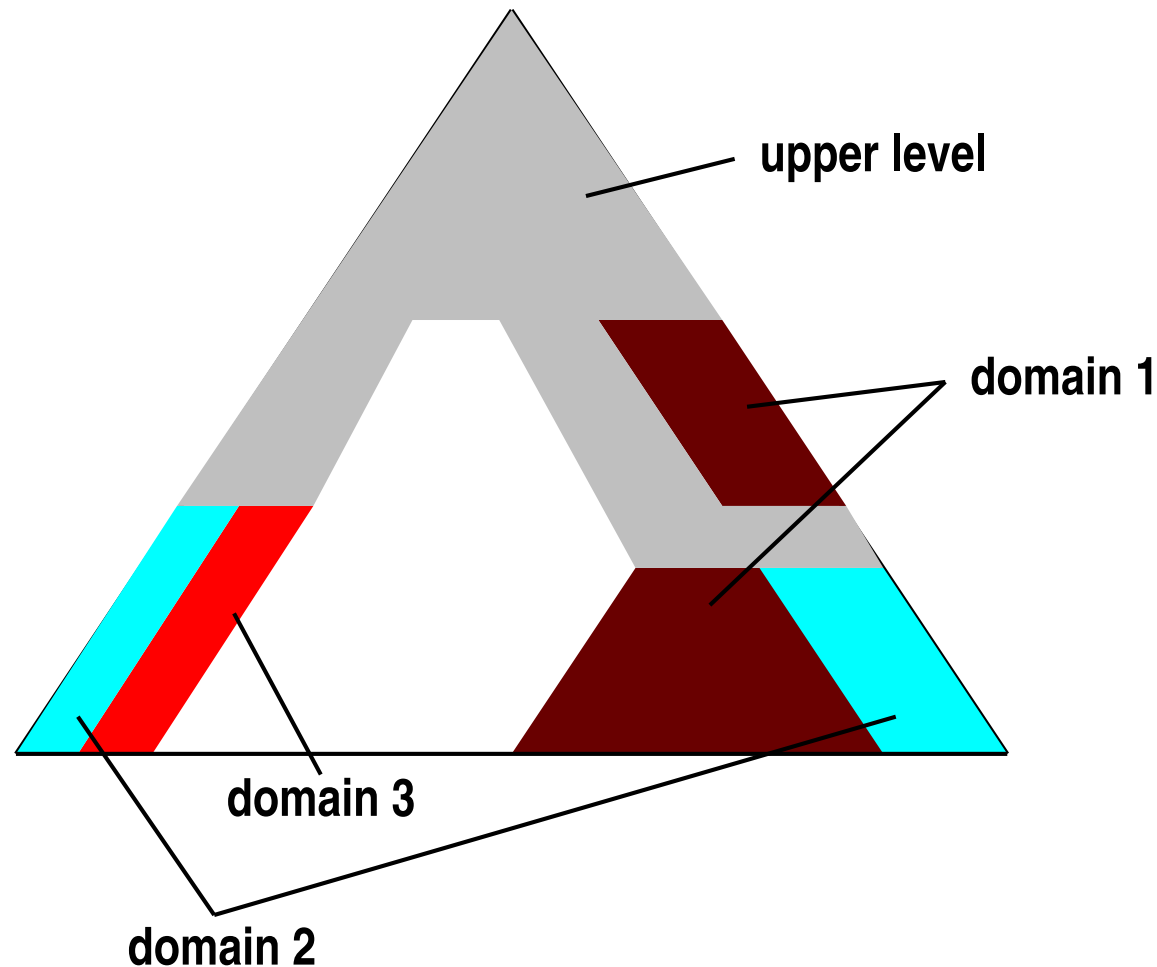
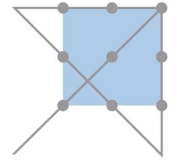
# What is a knowledge base good for?



## NLP applications:

- recovering information which is not fully specified – anaphora, bridging expressions, metonymies, ...  
→ inferences using *isa*-taxonomy or meronymic relations
- setting up conceptual restrictions – semantic parsing and analysis, word sense disambiguation
- useful for NL generation – NLG does not have to deal with unrestricted input!
- useful for information retrieval and extraction – query expansion

# Coverage, Domains, Depth



WordNet, Cyc, Penman, . . . , project specific ontologies

# How to build an upper level taxonomy

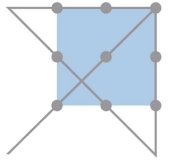


---

approches based on

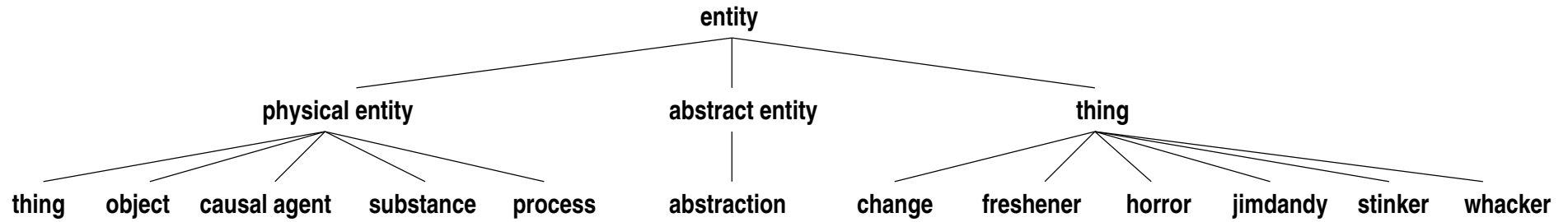
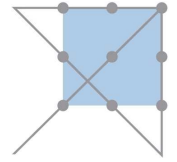
- natural language corpora (WordNet)
- introspection/knowledge engineering (Penman)
- philosophy/metaphysics (Guarino)
- common sense knowledge and inference (Cyc)

# WordNet



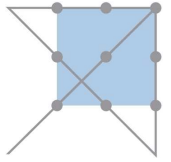
- most widely used lexical database for English
- three databases: nouns; verbs; adjectives and adverbs
- WordNet keeps information in *synsets*, a set of synonyms, a dictionary-style definition (gloss), and some example use
- also well-developed *isa*-hierarchy and some *member*-relations, *has-part*-relations

# WordNet upper levels



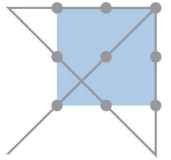
# Advantages of WordNet

---



- very good coverage of English, reasonable coverage for a range of other languages
- ready-to-be-used for NLP applications (word sense disambiguation, information retrieval and extraction, anaphora resolution, . . . )
- English *WordNet* is open source, so it's available for free
- software available for accessing, processing, . . .

# Restrictions of *WordNet*



- mostly *isa*-hierarchy, other relations less well developed
- hierarchy based on word meanings rather than on underlying concepts
- language dependent; *EuroWordNet* is based on English Wordnet but structured slightly differently (requires different software to access, process, . . . )
- EuroWordNet not free

# Knowledge Engineering à la *Penman*

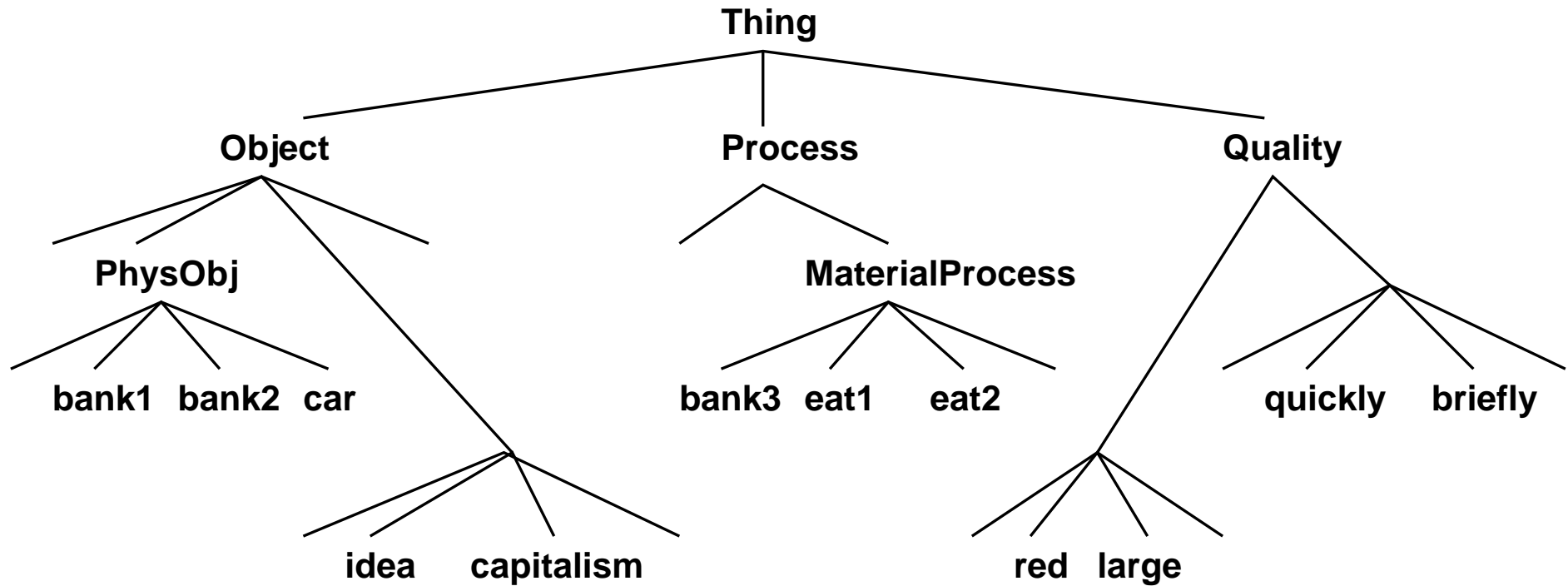
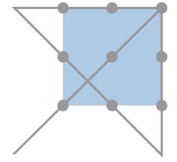


---

(Sondheimer et al., 1990; Hovy, 1993)

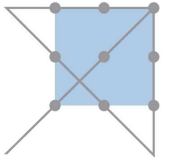
- taxonomy somewhere between philosophical investigations, common sense knowledge bases like *CYC* and lexical databases like *WordNet*
- objects in knowledge base do not have to correspond to individual lexical items, but rather to concepts that are known to be grammatically and lexically relevant
- upper level of taxonomy represents domain and language-independent notions

# Penman upper and middle models



# Philosophy/Metaphysics

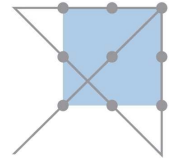
---



- based on 2500 years of research starting with Plato and Aristotle
- formal ontology may help to avoid confusion in building a upper level taxonomy

# Formal ontology (Guarino, 1998)

---



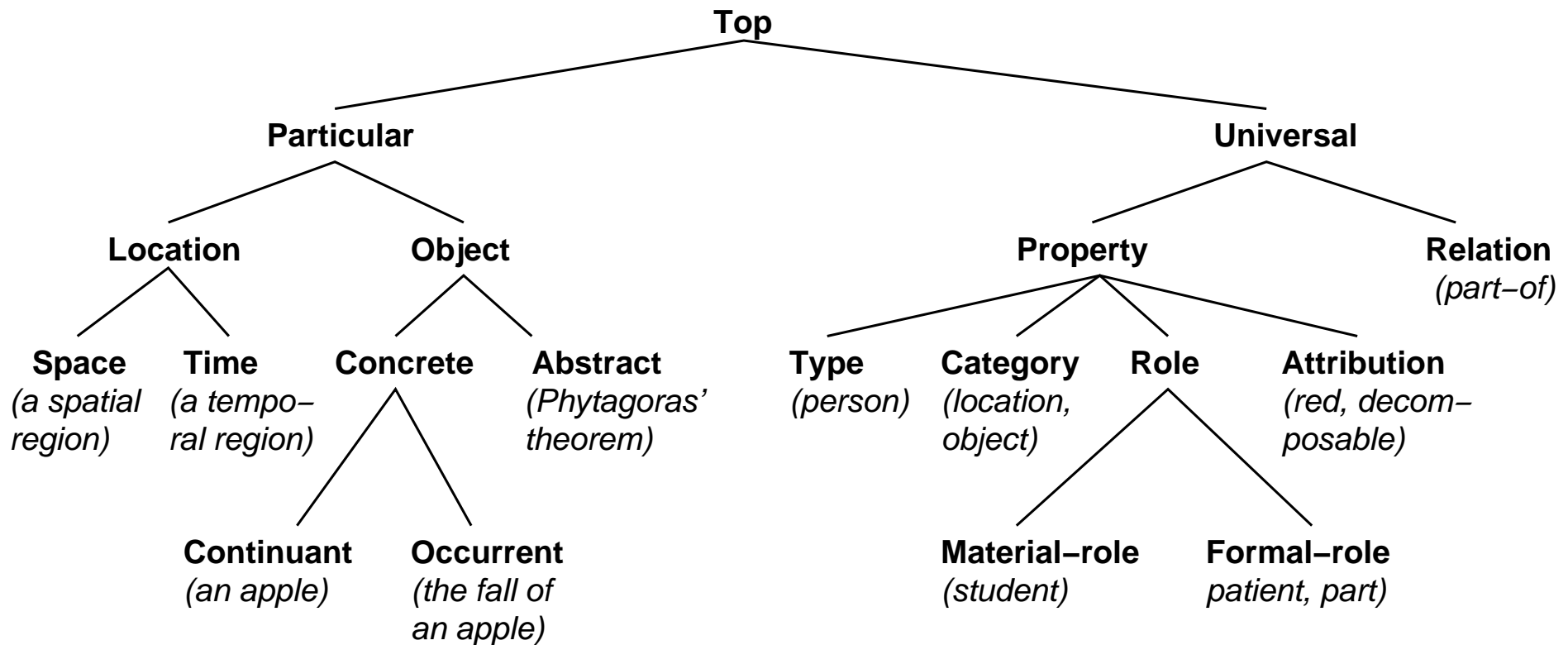
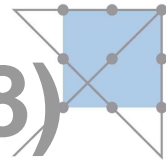
A theory of distinctions within

- the entities of the world to be included in our domain of discourse (*particulars*)
- the properties and relations used to talk about such entities (*universals*)

Methods for building a formal ontology:

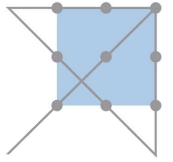
- theory of parts
- theory of wholes
- theory of identity
- theory of dependence

# Upper level ontology (Guarino, 1998)



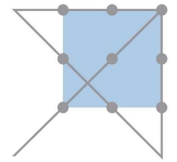
# Philosophy/Metaphysics

---



- the distinction between particulars and universals seems to be useless if one takes into account the distinction between *isa*-relations and *instance-of*-relations
- the distinction between *types* and *roles*, however, seems to be a nice tool for distinguishing between *intrinsic* properties (a building is a physical object) and *extrinsic* ones (building X is a research institute)

# Dividing the world into categories



(Russell & Norvig, 1995)

**Common properties:** categories include as members all objects having certain properties  $\implies$  taxonomy

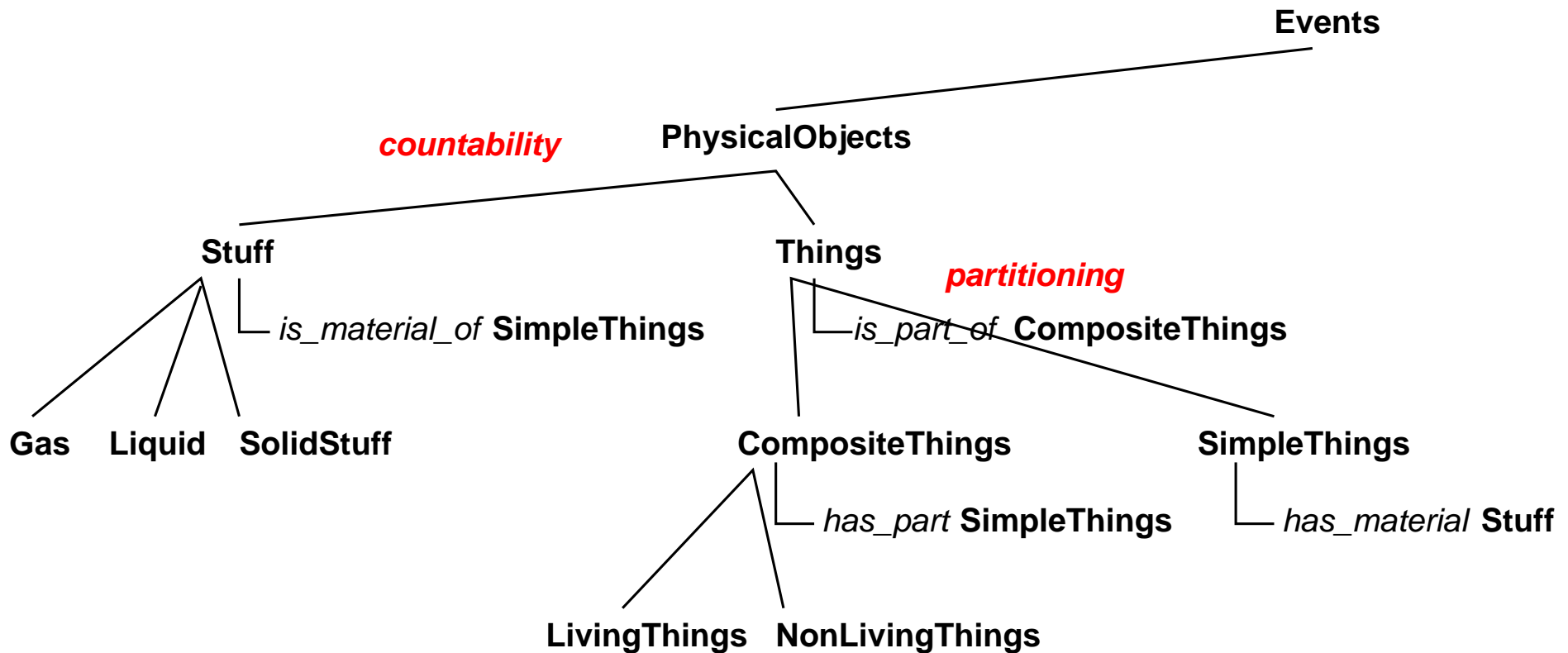
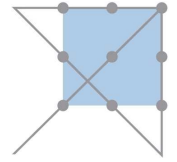
**Constituent structure:** composite objects belong to categories whose members share their structure (cars, motorcycles, bicycles, ...)

**Physical objects vs. events:** physical objects and events share their extension in space and time

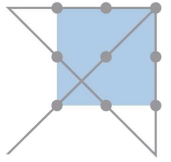
**Abstract objects:** mental objects and beliefs are not extended in space and time

**Substances, simple objects, composite objects:** physical objects may be categorized by *countability* and *partitioning*

# An example: Physical objects



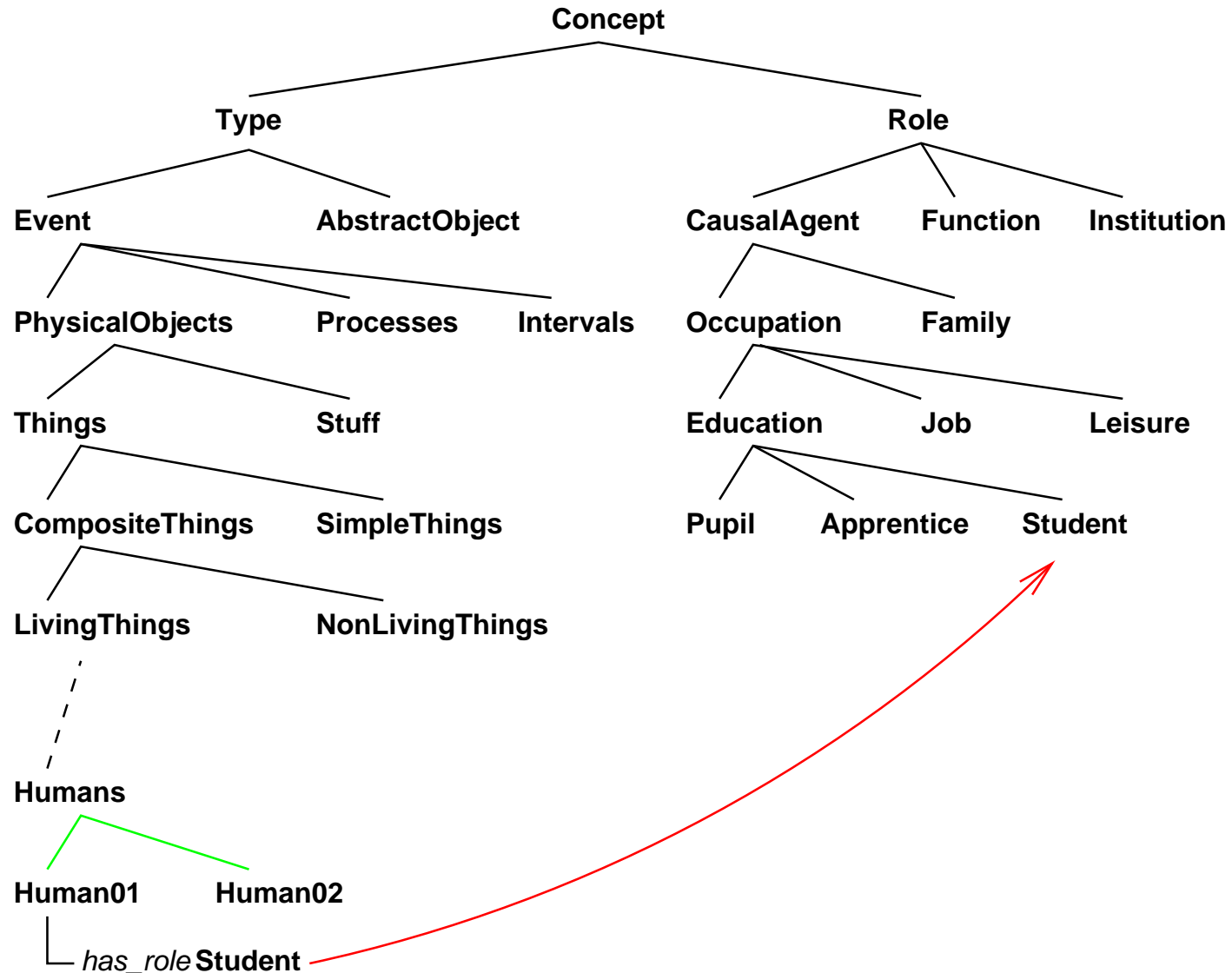
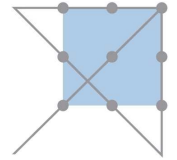
# Dividing the world into categories



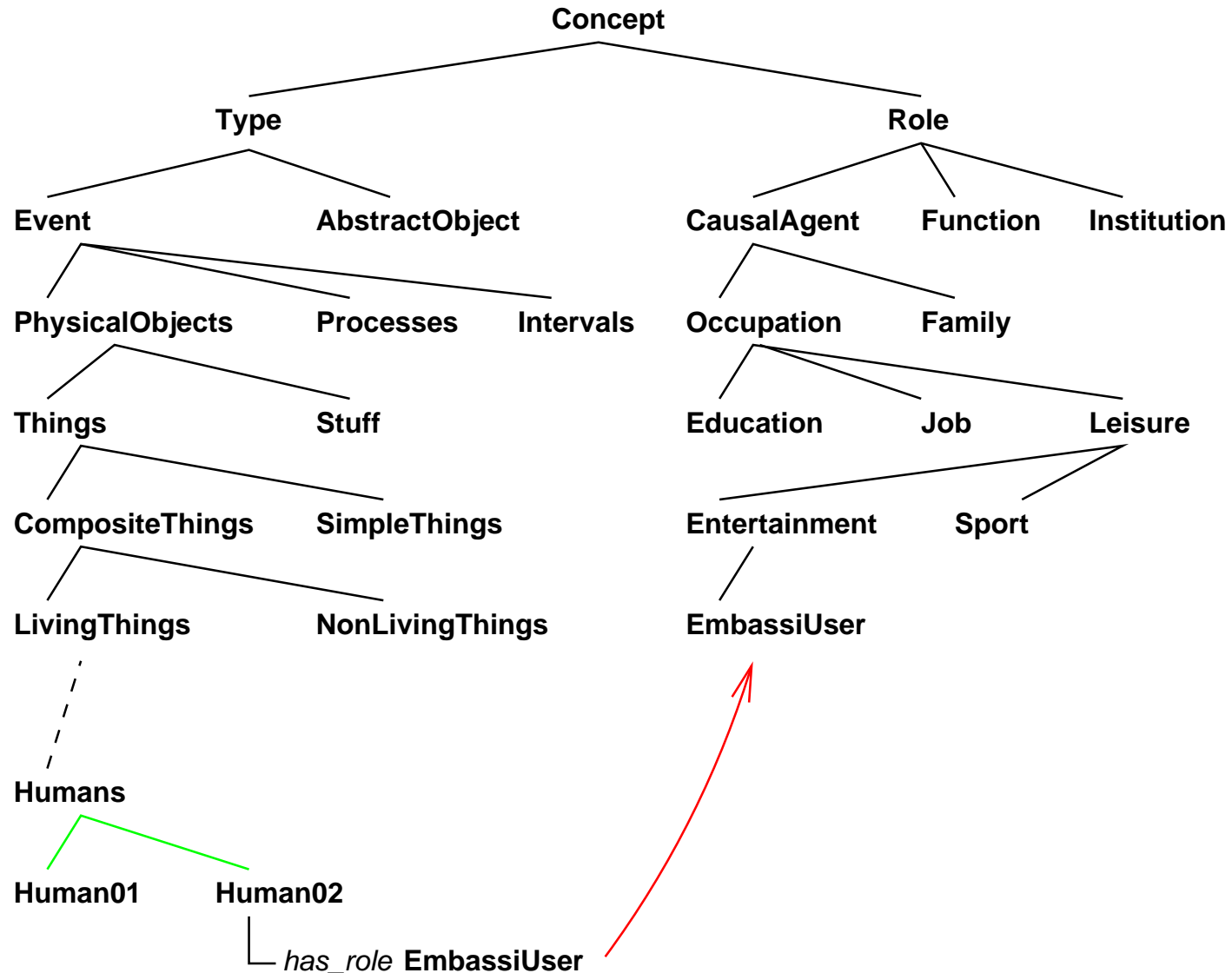
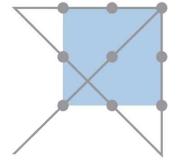
**Intrinsic vs. extrinsic properties:** objects exhibit invariable properties which can be used for categorizing them into main categories; if the properties change over time, then the property may belong to a *role* (to be a student, a city hall, a research institute)

A mixture between Guarion's (1998) proposal and Russell & Norvig (1995) may be the best compromise.

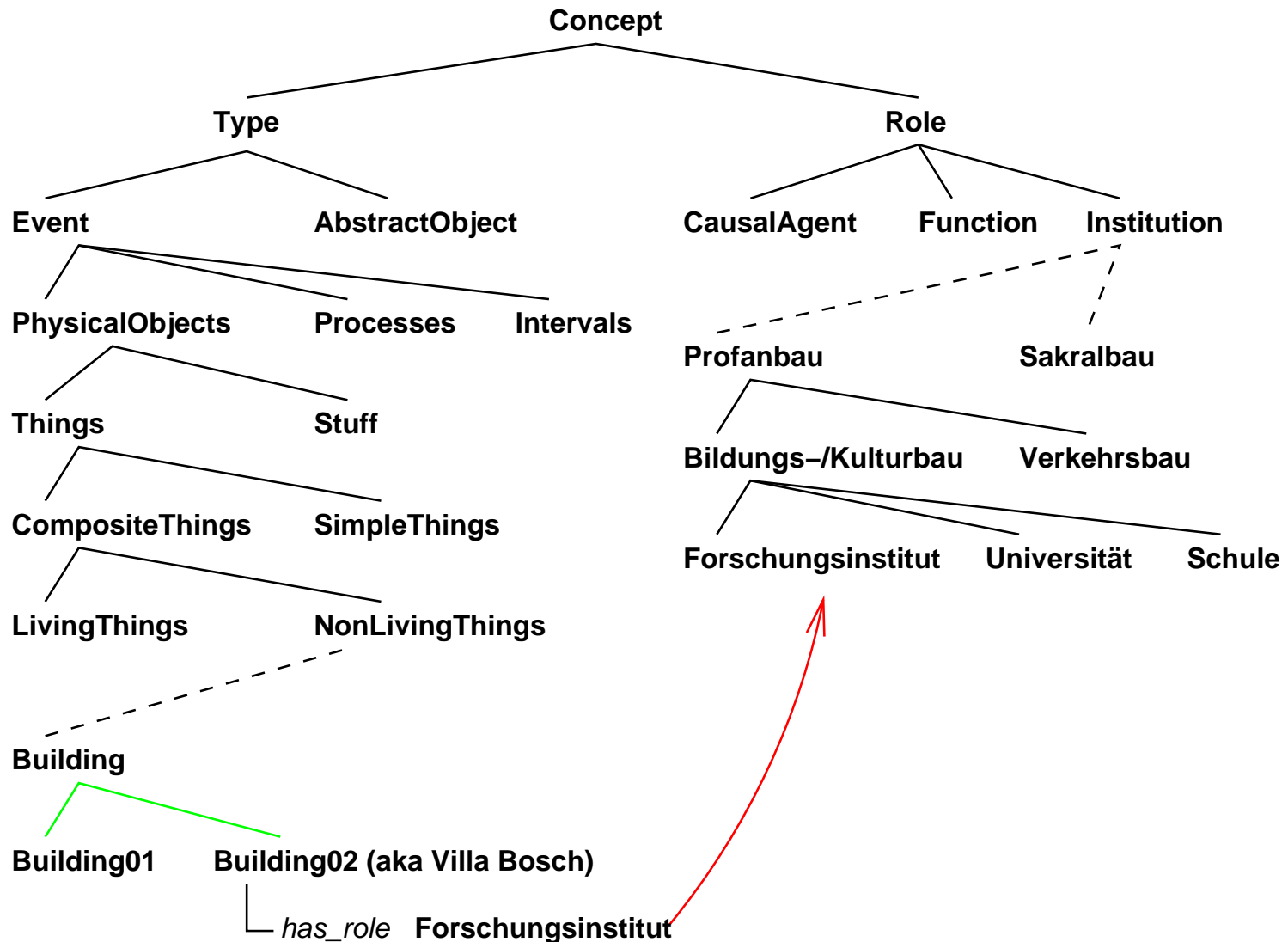
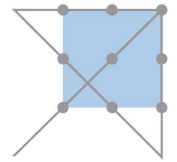
# An example: Student



# An example: EmbassiUser

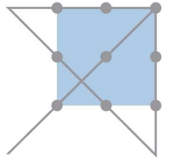


# An example: Research institute



# Preliminary conclusions

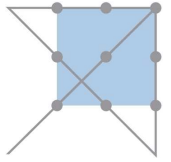
---



- in most knowledge bases *isa*-relations not modeled according to the principles mentioned above
- in a knowledge base comprising different domains most of the domains should be viewed as part of the *role*-hierarchy
- the *role*-hierarchy should not contain *instance*-relations
- the *type*-hierarchy including *instance*-relations should be the backbone; domains are connected to that via *roles*

# Outline

---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. Knowledge derived from Wikipedia
  - (a) Semantic relatedness
  - (b) WikiRelate! Computing semantic relatedness using Wikipedia
  - (c) Knowledge bases, taxonomies, ontologies
  - (d) *Deriving a taxonomy from Wikipedia*
5. Exploiting Wikipedia for Coreference Resolution
6. Further applications
7. Conclusions

# Semantic relatedness with Wikipedia



---

WikiRelate! (AAAI 2006):

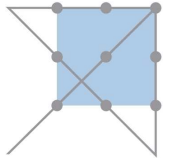
- assume Wikipedia pages represent concepts
- query the concepts and see in which categories they fall
- assume the categories to represent a semantic network

Three main steps:

1. page retrieval and disambiguation
2. category tree (network) search
3. relatedness measure computation

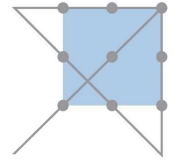
# Semantic relatedness measures

---

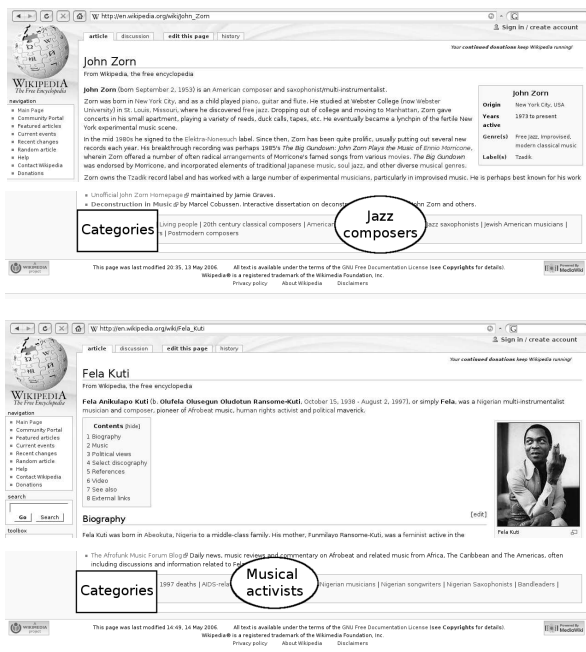


- we ported the *path length* based measures from **Rada et al., (1989, *pl*)**, **Wu & Palmer (1994, *wup*)**, **Leacock & Chodorow (1998, *lch*)**, and the *information content* measure from **Resnik (1995, *res*)** to Wikipedia

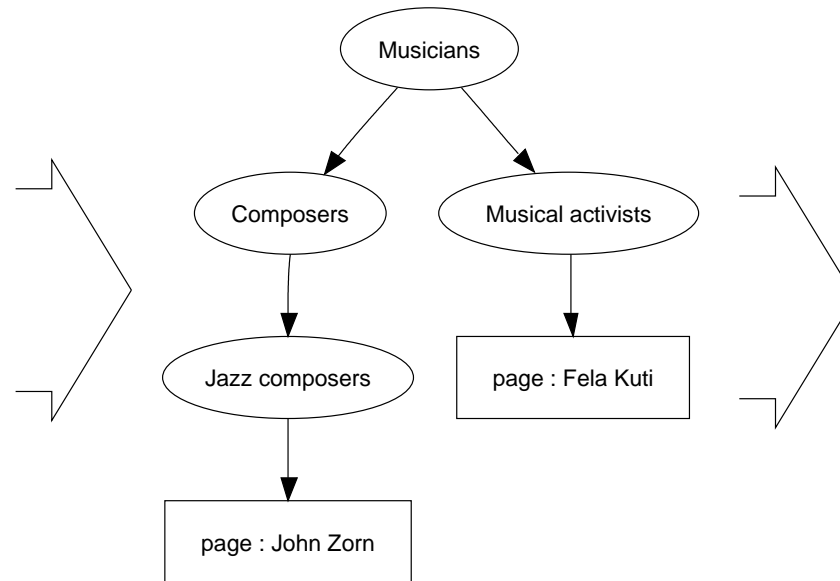
# Applying the measures



1. **extract categories** from the Wikipedia pages
2. **search** for a connecting **path** along the category network
3. **score the paths** found and return the one(s) satisfying the measure definitions



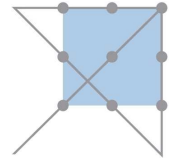
page query and retrieval  
category extraction



search for a connecting path  
along the category tree

relatedness measure(s) computation

# Experiments using similarity lists

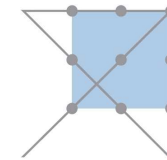


we experimented with the similarity lists from

- Miller & Charles (1991, M&C, 30 word pairs)
- Rubenstein & Goodenough (1965, R&G, 65 word pairs)
- and the relatedness list 353-TC from Finkelstein et al. (2002)

➡ evaluation metric: Pearson's  $r$  correlation coefficient

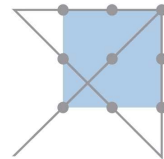
# Experiments using similarity lists



(results obtained in 2007)

<i><b>M&amp;C</b></i>		<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>
WordNet	all	0.72	0.77	<b>0.82</b>	0.78
Wikirelate!	all	<b>0.60</b>	0.53	0.58	0.30
	non-missing	<b>0.65</b>	0.61	<b>0.65</b>	0.41

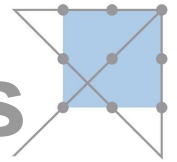
# Experiments using similarity lists



(results obtained in 2007)

<i><b>R&amp;G</b></i>		<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>
WordNet	all	0.78	0.82	<b>0.86</b>	0.81
Wikirelate!	all	0.62	0.63	<b>0.64</b>	0.34
	non-missing	0.66	0.69	<b>0.70</b>	0.42

# Experiments using relatedness lists

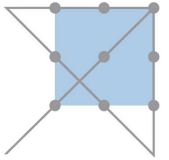


(results obtained in 2007)

<i><b>TC-353 full</b></i>		<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>
WordNet	all	0.30	0.30	<b>0.34</b>	<b>0.34</b>
Wikirelate!	all	<b>0.48</b>	0.52	0.53	0.41
	non-missing	<b>0.48</b>	0.52	<b>0.54</b>	0.41

# Discussion

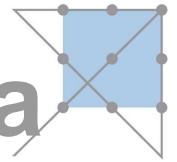
---



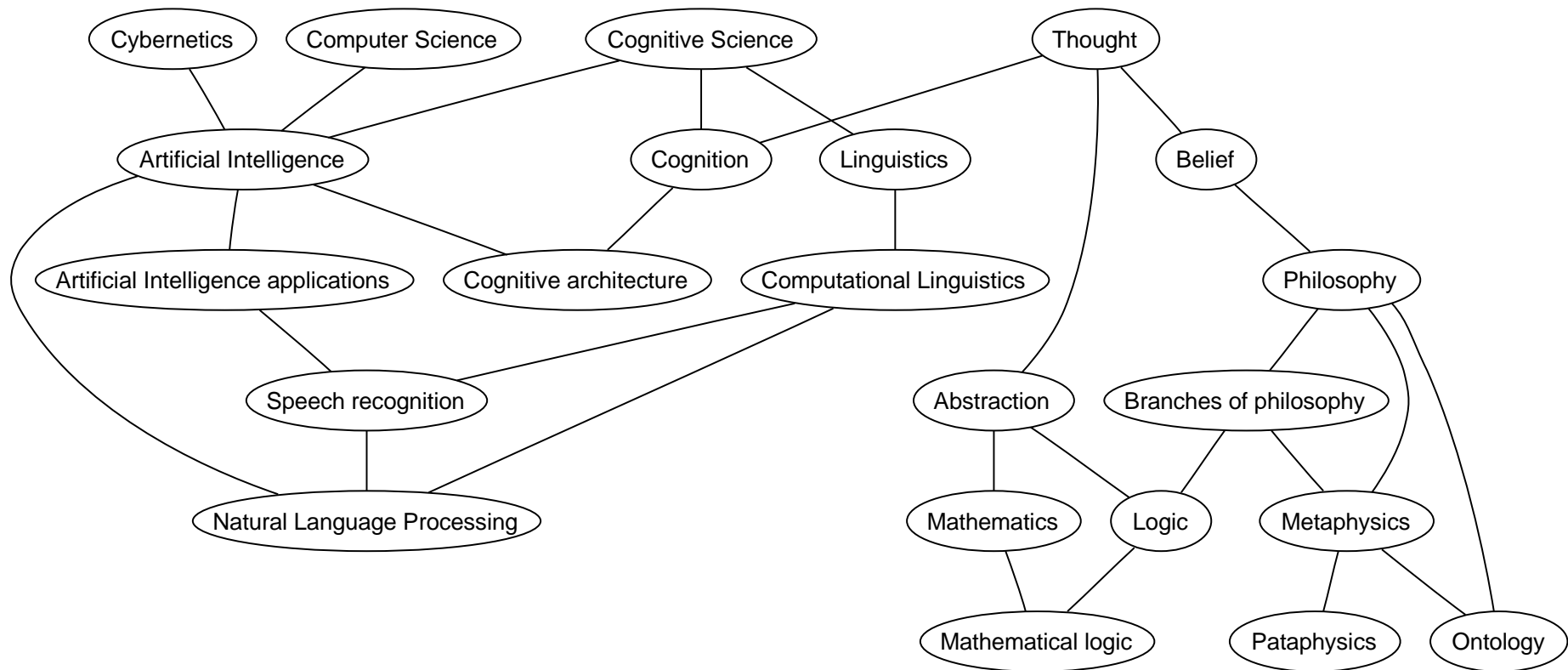
the poor performance of Wikipedia on the similarity datasets seems to be due to

- the semantic network providing information on *semantic relatedness* while the datasets were judged for *semantic similarity* (*isa* vs. all relations)
- ➡ by deriving a taxonomy and computing semantic *similarity* instead of *relatedness*, we should be able to obtain better results on the R&G and M&C datasets (and provide a more useful, informative resource)

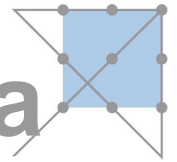
# Deriving a taxonomy from Wikipedia



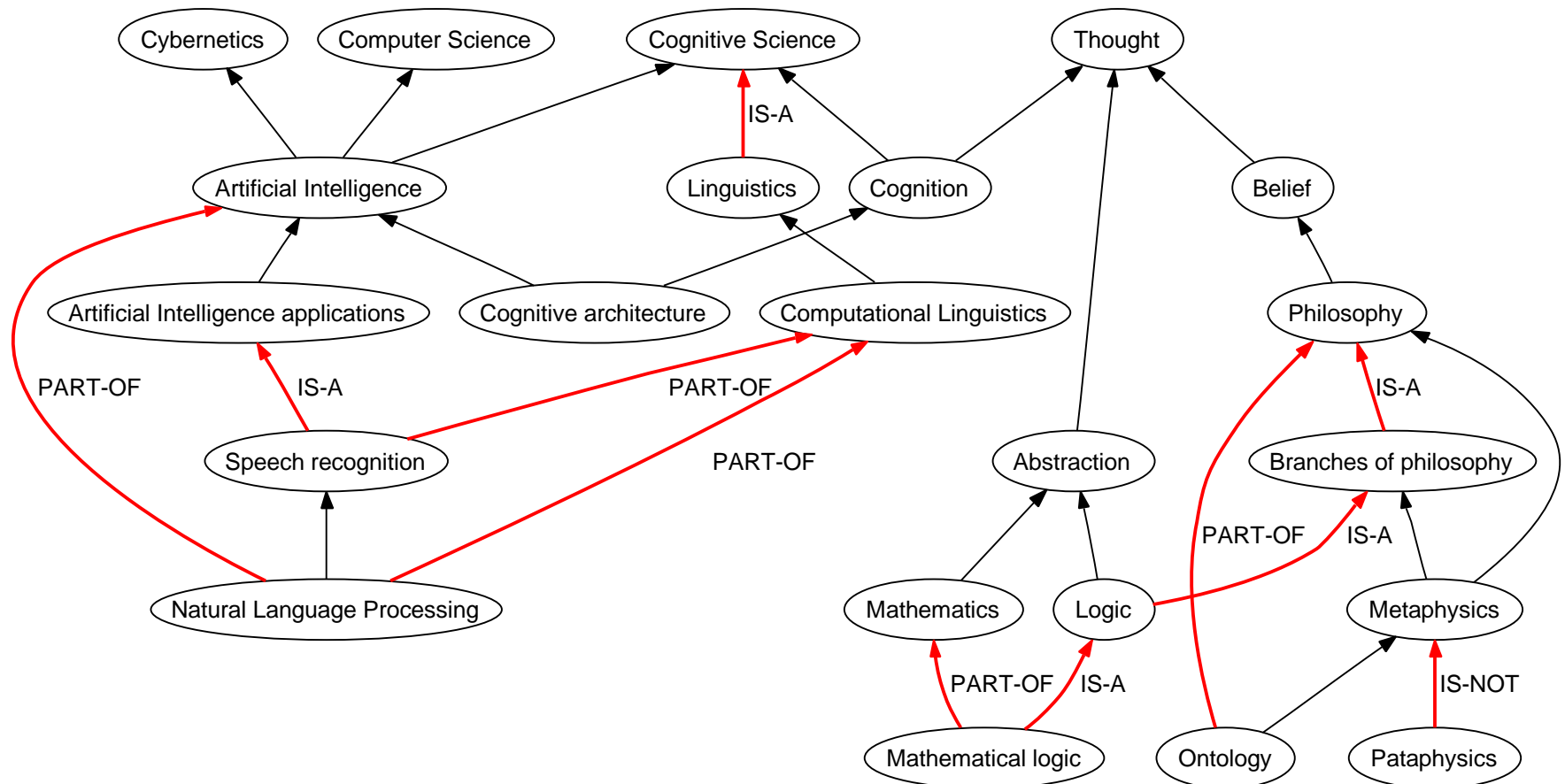
- induce **semantically-typed** category links



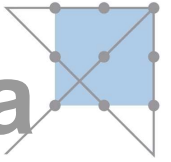
# Deriving a taxonomy from Wikipedia



- induce **semantically-typed** category links

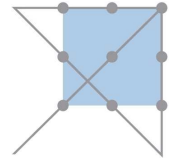


# Deriving a taxonomy from Wikipedia



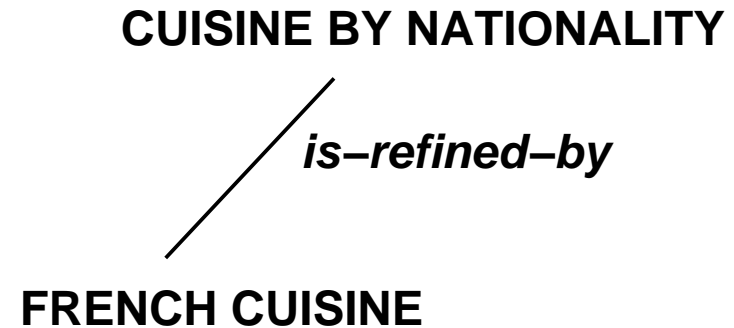
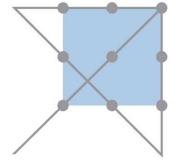
- **goal:** transform semantic network into a taxonomy by labeling *isa* and *notisa* links
- attempt to exploit Wikipedia's internal structure as much as possible
- **methods:**
  - syntactic cleanup
  - connectivity in the network
  - lexico-syntactic patterns
- we start with 165,744 categories and 349,263 links

# Category network cleanup (1)



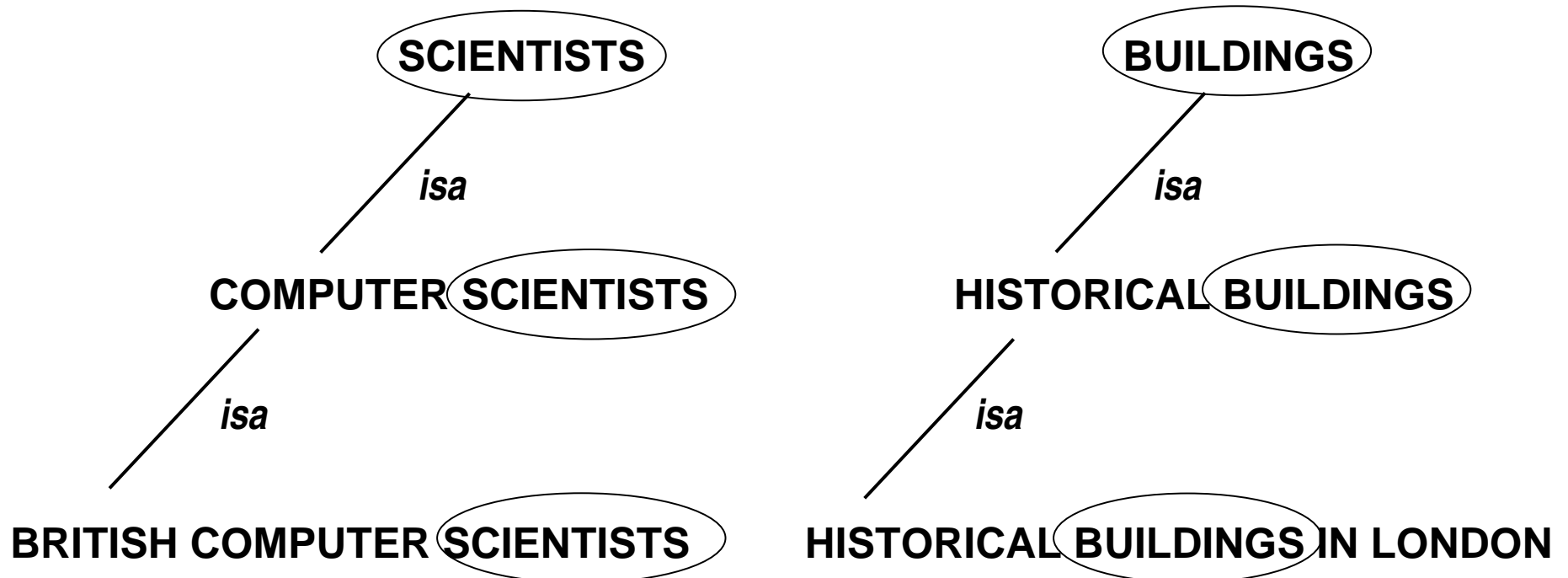
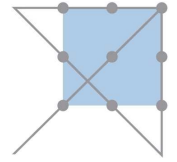
- removal of meta-categories used for encyclopedia management, e.g. categories under WIKIPEDIA ADMINISTRATION
- we remove all nodes whose labels contain any of the following strings: MEDIAWIKI, TEMPLATE, USER, PORTAL, CATEGORIES, ARTICLES, PAGES
- this leaves
  - 127,325 categories
  - 267,707 linksstill to be processed!

# Refinement link identification (2)



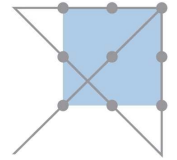
- patterns such as **Y X and X BY Z**
- their purpose is to better structure and simplify the categorization network
- we assume this represents *is-refined-by*-relations
- this labels 54,504 category links *notisa* and leaves 213,203 relations to be analyzed

# Syntax-based methods (3)



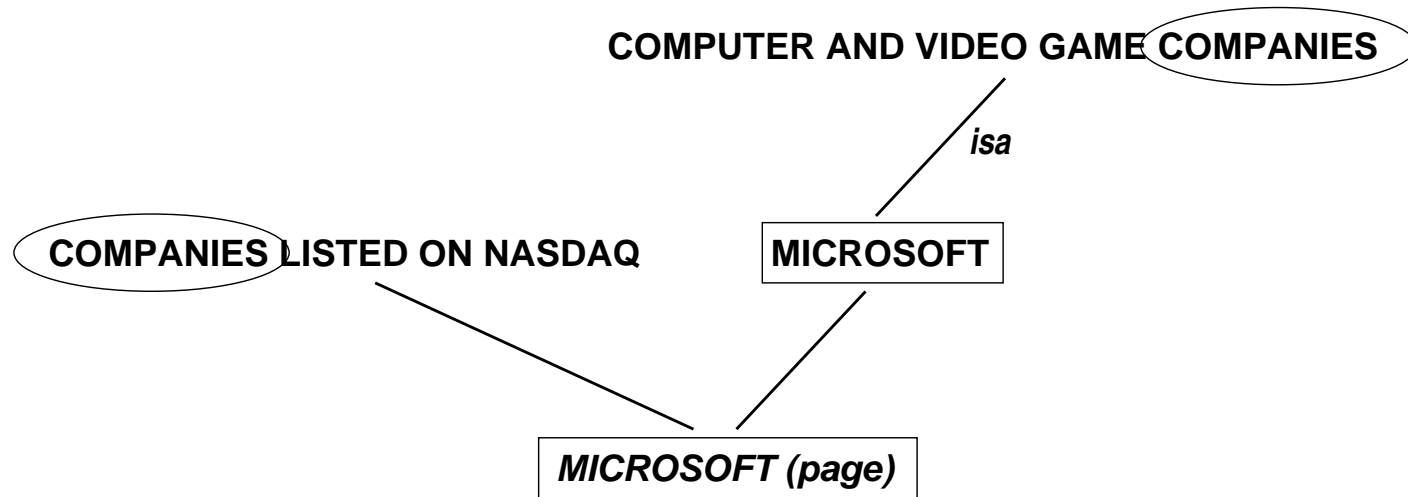
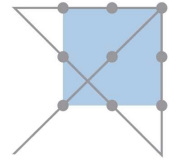
- **head matching** labels pairs of categories sharing the **same lexical head word (or lemma)**
- we identify lexical heads using the *Stanford parser* (Klein & Manning, 2003) and lemmata using `morpha` (Minnen et al., 2004)

# Syntax-based methods (3)



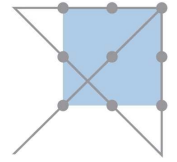
- **modifier matching** labels pairs as *notisa*, if the stem of the lexical head of one of the categories occurs in non-head position in the other category, e.g. **CRIME COMICS** and **CRIME** or **ISLAMIC MYSTICISM** and **ISLAM**
- these methods identify 72,663 *isa* relations and 37,999 *notisa* relations. NOTE:
  - ▣➔ relatively 'simple' (→ **baseline**)
  - ▣➔ still *large coverage*

# Connectivity-based methods (4)



- *instance categorization* uses the idea that relations between entities (Wikipedia pages) and classes (categories) can be labeled as *instance-of* (cf. Suchanek et al., WWW 2007)

# Connectivity-based methods (4)

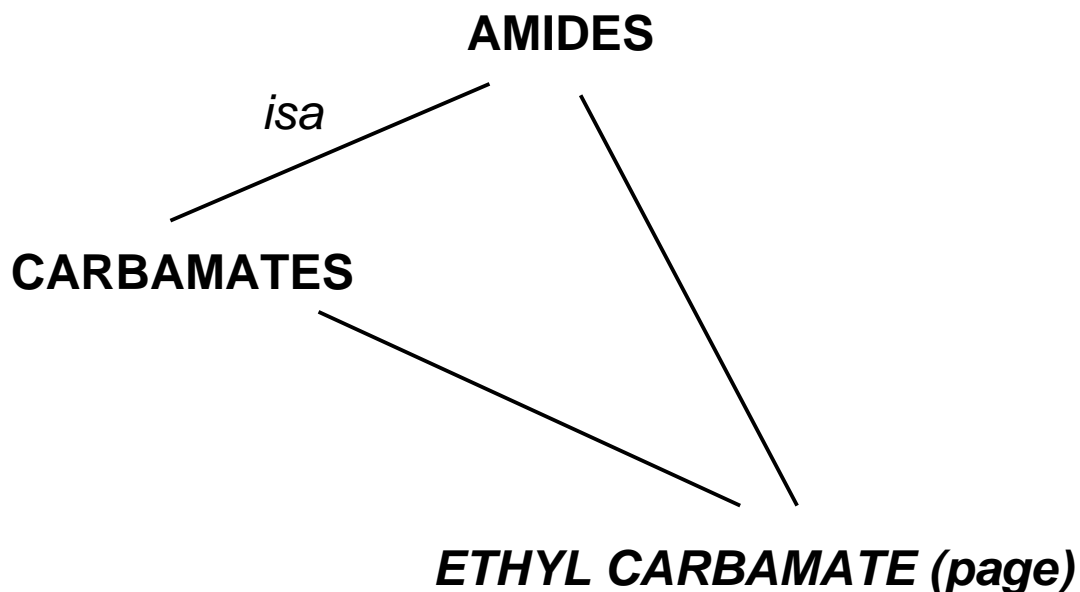
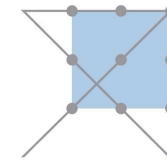


For each category  $c$ ,

1. find the page titled as the category or its lemma, for instance the page **MICROSOFT** for the category **MICROSOFT**;
2. collect all the page's categories whose lexical head is a plural noun  $CP = \{c_1, c_2, \dots, c_n\}$ ;
3. for each  $c$ 's supercategory  $sc$ , we label the relation between  $c$  and  $sc$  as *isa*, if the head lemma of  $sc$  matches the head lemma of at least one category  $cp \in CP$ .

⇒ identifies 9,890 *isa* relations

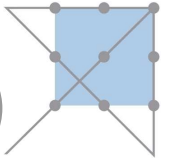
# Connectivity-based methods (4)



- *redundant categorization* if users redundantly categorize we take this as evidence for *isa* relations, cf. **ETHYL CARBAMATE**
- ➡ identifies 11,087 *isa* relations

we are left with 81,564 unclassified relations ...

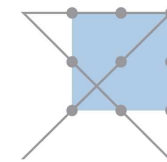
# Lexico-syntactic based methods (5)



- we apply lexico-syntactic patterns to sentences in large text corpora to identify *isa* relations (Hearst, 1992; Caraballo, 1999) and patterns to identify *notisa* relations
- ➡ we assume that patterns used for identifying *meronymic relations* (Berland & Charniak, 1999) indicate that the relation **is not** an *isa* relation
- corpora used: TIPSTER ( $2.5 \times 10^8$  words) and the English Wikipedia ( $5 \times 10^8$  words)

# Preprocessing: Corpora

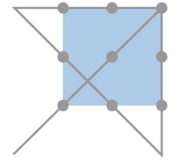
---



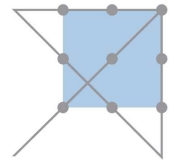
- remove SGML (Tipster) and Wiki markup (Wikipedia)
- split into reasonable chunks for further processing (1 MB files, spread over several directories)
- split into sentences (Perl script obtained from Univ. of Illinois at Urbana-Champaign)
- POS tagging with TnT (Brants, 2000) which is fast
- NP chunking using a SVM based chunker (Kudoh & Matsumoto, 2000)
- result overall about 15GB data (5GB plain text, 5GB POS tags, 5GB NP chunks)
- further processing done on 2 servers with 4 processors and 8GB RAM each

# Further technical details

---



- most categories are plural, hence convert to singular – mapping taken from XTAG grammar
- convert to lower case
- during processing convert to plural again, if pattern requires it
- cascaded processing:
  - build index using Lucene (you want to search only those files which actually contain the two words)
  - if file contains words, search for sentences which contain both words
  - if sentence contains words, apply all patterns and count

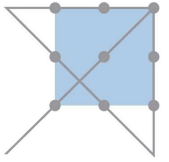


# Examples of ISA patterns:

---

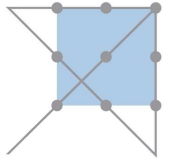
1. *NP2, ? (such as |like|, especially) NP\* NP1*  
*a stimulant such as caffeine*
2. *such NP2 as NP\* NP1*  
*such stimulants as caffeine*
3. *NP1 NP\* (and|or|,like) other NP2*  
*caffeine and other stimulants*
4. *NP1, one of det\_pl NP2*  
*caffeine, one of the stimulants*
5. *NP1, det\_sg NP2 rel\_pron*  
*caffeine, a stimulant which*
6. *NP2 like NP\* NP1*  
*stimulants like caffeine*

# Examples of NOTISA patterns



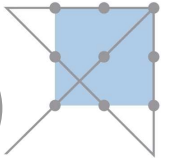
1. *NP2's NP1*  
*car's engine*
2. *NP1 in NP2*  
*engine in the car*
3. *NP2 with NP1*  
*a car with an engine*
4. *NP2 contain(s|ed|ing) NP1*  
*a car containing an engine*
5. *NP1 of NP2*  
*the engine of the car*
6. *NP1 are? used in NP2*  
*engines used in cars*
7. *NP2 ha(s|ve|d) NP1*

# Sample output: Result



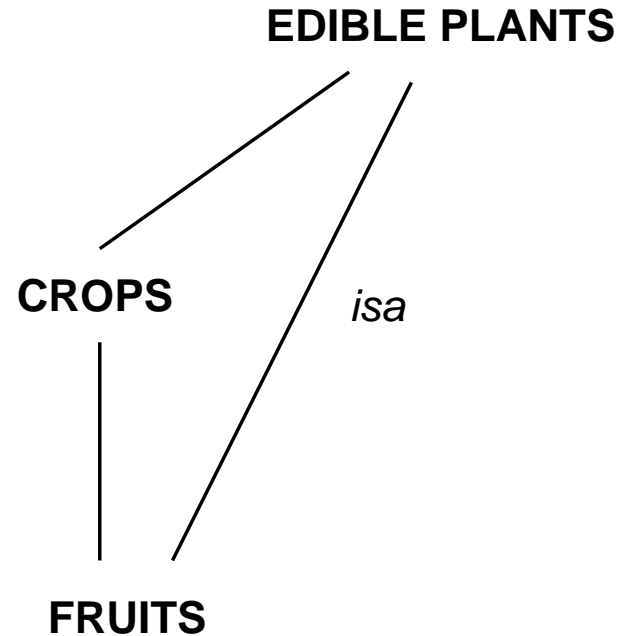
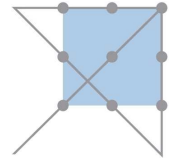
```
0: aachen ISA Westphalia
0: aachen NOTISA Westphalia
2: aardvark ISA animal
0: aardvark NOTISA animal
0: abbasid ISA people
0: abbasid NOTISA people
2: abbot ISA clergy
0: abbot NOTISA clergy
0: abbot ISA history
0: abbot NOTISA history
1: abbot ISA monk
1: abbot NOTISA monk
0: abbot ISA people
0: abbot NOTISA people
0: abbreviation ISA reference
0: abbreviation NOTISA reference
0: abbreviation ISA word
19: abbreviation NOTISA word
```

# Lexico-syntactic based methods (5)



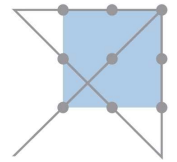
- majority voting strategy between *isa* and *notisa* patterns
- this method identifies 15,055 *isa* relations
- we apply this method also to the relations identified in step (4) and filter out 3,277 previously identified *isa* relations

# Inference-based methods (6)



- propagate previously found relations via multiple inheritance and transitivity

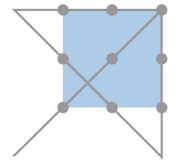
# Evaluation: Semantic similarity



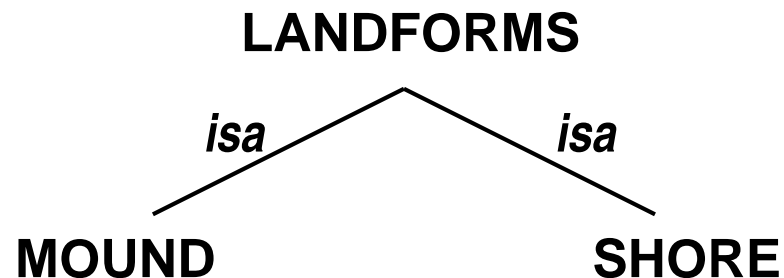
<i><b>M&amp;C</b></i>		<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>
WordNet	all	0.72	0.77	<b>0.82</b>	0.78
Wikirelate!	all	<b>0.60</b>	0.53	0.58	0.30
	non-missing	<b>0.65</b>	0.61	<b>0.65</b>	0.41
Wikirelate! <i>isa-only</i>	all	0.67	0.65	0.67	<b>0.69</b>
	non-missing	0.71	0.70	0.72	<b>0.74</b>

➡ performance **improvement lower than expected**

# Intermediate error analysis

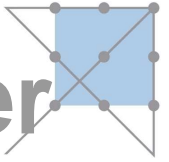


- experiments on development data revealed a performance **improvement lower than expected**
- error analysis revealed that many dissimilar pairs received a high score because of **coarse-grained over-connected categories with a large number of pages**, e.g.



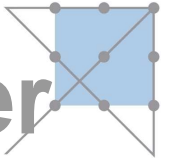
➡ *how to filter these categories out?*

# PageRank as a coarse category filter



- a way to model the categories' connectivity is to compute their authoritativeness, i.e.
- we assume that **over-connected, semantically coarse categories** will be the **most authoritative ones**
- since the Wikipedia categorization network is a directed acyclic graph, we can find authoritative categories by computing their **centrality scores** using the PageRank algorithm (Brin & Page, 1998)

# PageRank as a coarse category filter



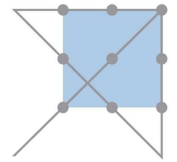
- PageRank scores are computed recursively for each category vertex by the formula

$$PR(v) = (1 - d) + d \sum_{v' \in I(v)} \frac{PR(v')}{|O(v')|}$$

where  $d \in (0, 1)$  is a dumping factor (we set it to the standard value of .85),  $I(v)$  is the set of nodes linked to  $v$  and  $|O(v')|$  the number of outgoing links of node  $v'$

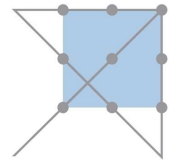
- filtering boils down to removing the top- $n$  categories with the highest PageRank score, i.e. FUNDAMENTAL, SOCIETY, KNOWLEDGE, PEOPLE, SCIENCE, ACADEMIC DISCIPLINES
- we removed the top 200 categories (experiments on development data)

# Evaluation: Semantic similarity



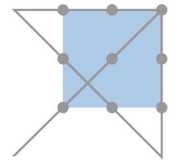
<i><b>M&amp;C</b></i>		<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>
WordNet	all	0.72	0.77	<b>0.82</b>	0.78
Wikirelate!	all	<b>0.60</b>	0.53	0.58	0.30
	non-missing	<b>0.65</b>	0.61	<b>0.65</b>	0.41
Wikirelate! <i>isa-only</i>	all	0.67	0.65	0.67	<b>0.69</b>
	non-missing	0.71	0.70	0.72	<b>0.74</b>
Wikirelate! PageRank filter	all	0.68	<b>0.74</b>	0.73	0.62
	non-missing	0.72	<b>0.79</b>	0.78	0.68
Wikirelate! <i>isa + PageRank</i>	all	0.73	0.79	0.78	<b>0.81</b>
	non-missing	0.76	0.84	0.82	<b>0.86</b>

# Evaluation: Semantic similarity



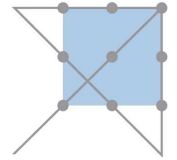
<b><i>R&amp;G</i></b>		<i>pl</i>	<i>wup</i>	<i>lch</i>	<i>res</i>
WordNet	all	0.78	0.82	<b>0.86</b>	0.81
Wikirelate!	all	0.62	0.63	<b>0.64</b>	0.34
	non-missing	0.66	0.69	<b>0.70</b>	0.42
Wikirelate! <i>isa-only</i>	all	0.67	0.69	<b>0.70</b>	0.66
	non-missing	0.70	<b>0.73</b>	<b>0.73</b>	0.70
Wikirelate!	all	0.67	<b>0.74</b>	0.73	0.58
PageRank filter	non-missing	0.70	<b>0.79</b>	0.77	0.63
Wikirelate!	all	0.69	0.75	0.74	<b>0.76</b>
<i>isa</i> + PageRank	non-missing	0.72	0.79	0.77	<b>0.80</b>

# Comparison with ResearchCyc



- for each relation we map each category to its Cyc concept using Cyc's internal *lexeme-to-concept denotational mapper*
- we query the full concept (Alan Turing), if this is not found we fall back to the lexical head only (hardware for IBM hardware)
- 85% of the category pairs found have corresponding concepts in Cyc
- we query Cyc whether the concept denoted by the Wikipedia subcategory is either an *instance of* (`#$isa`) or *is generalized by* (`#$genls`) the concept denoted by its superclass
- evaluation in terms of precision, recall, F-measure

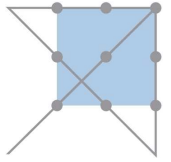
# Comparison with ResearchCyc



	$R$	$P$	$F_1$
baseline (methods 1-3)	73.7	100.0	84.9
+ connectivity (methods 1-4, 6)	80.6	91.8	85.8
+ pattern-based (methods 1-3, 5-6)	84.3	91.5	87.7
all (methods 1-6)	89.1	86.6	87.9

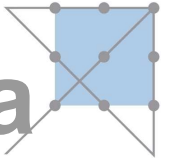
# Discussion

---



- inspection of random sample of 200 errors revealed that about 50% of those were correct indeed but could not be found in Cyc, because Cyc is missing
  - ➡ the relations (BRIAN ENO *isa* MUSICIAN)
  - ➡ the concepts (BEE TRAIN *isa* ANIMATION STUDIOS, query for heads TRAIN and STUDIOS)

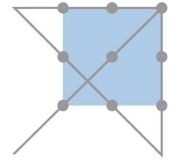
# Deriving a taxonomy from Wikipedia



- ➡ Wikipedia categories provide a semi-structured resource that makes it possible to generate a large scale taxonomy, e.g. we can **generate semantic relations from the connectivity in the network**
- ➡ **relatedness measures** computed from Wikipedia could not be applied successfully to similarity datasets
- ➡ by inducing a *taxonomy*, we are able to compute **semantic similarity** and compete with WordNet
- ➡ comparison with ResearchCyc showed that we derived a high quality taxonomy from Wikipedia

# Current and future work

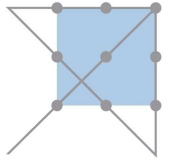
---



- Bring the idea to **complete systems**, e.g.
  - put it in our JHU summer workshop system
  - our DUC-2008 system (NER, SRL, coreference resolution, topic segmentation, lexical chains for summarization, . . . )
- ⇒ perform an **extrinsic evaluation** using the similarity measures as *features of a machine learning based coreference resolution system*  
Cf. Ponzetto & Strube (HLT-NAACL 2006)

# Current and future work

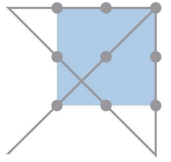
---



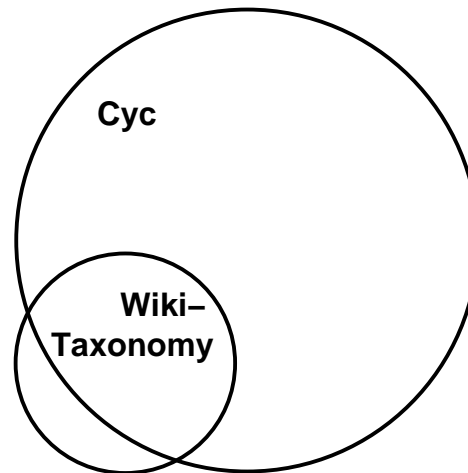
- distinguish between *instances* and *classes*
- apply methodology to other languages (first language will be French)
- derive an ontology by labeling **other semantic relations as well**, e.g. meronymic relations (e.g. the relations identified by Girju et al. (2006))
- populate/enlarge the taxonomy by including all pages (not only categories)

# Size of the taxonomy

---

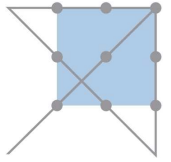


- decision (*isa vs. notisa*) for 127,325 categories by looking only at the category network (we did not include the pages themselves)

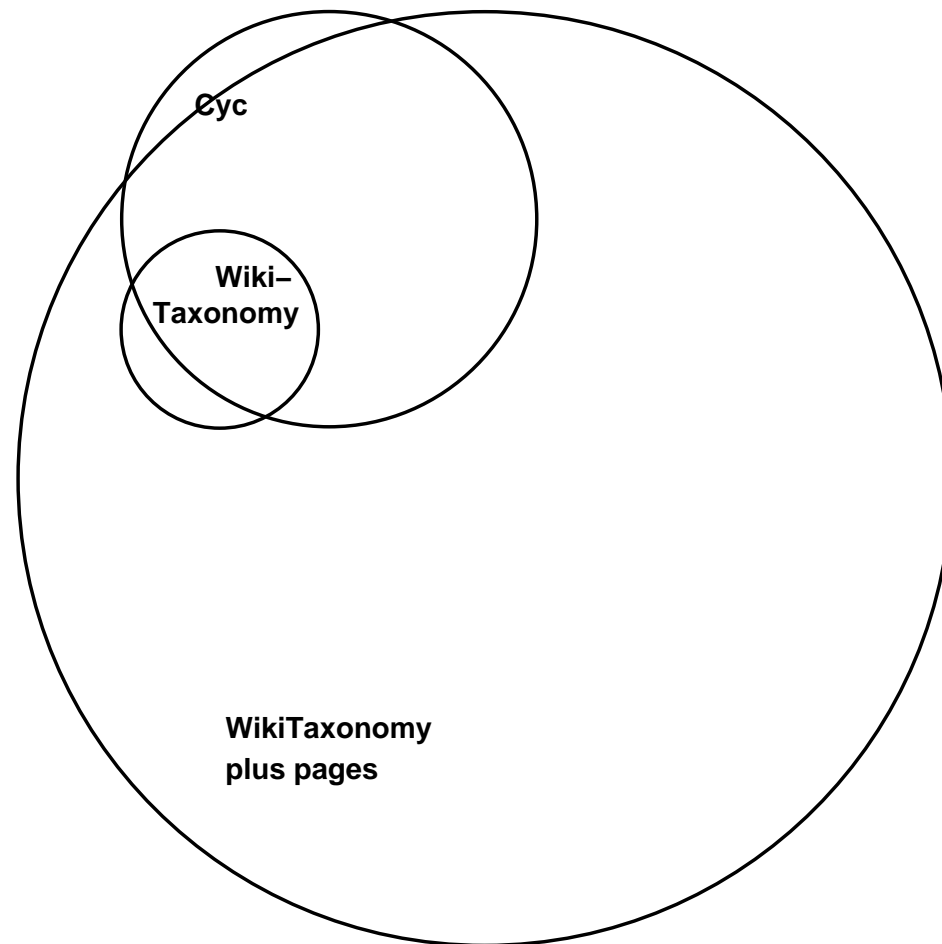


# Size of the taxonomy

---

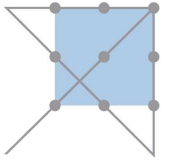


- after including the pages as well it will rather look like this



# Future work

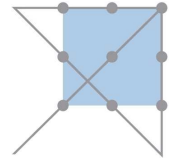
---



- generate language aligned knowledge bases
  - increases size and coverage of resource
  - resource for multilingual applications (e.g. multilingual QA)
  - resource for stepping into semantics/knowledge-based statistical MT

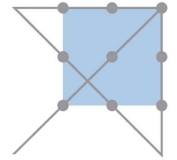
***Wanna get involved? – Drop me a line!***

# When we are done, we get ...



These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.

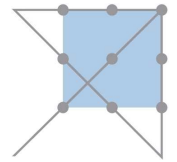
# Take-home message 6



- ➡ it could be that a folksonomy is exactly what we need for AI and NLP applications
- ➡ achieving a *large coverage* and *robust* taxonomy by **collaboration** of many users
- ➡ generated by those users whose behavior we are trying to model in our 'intelligent' applications
- ➡ we stake out a middleground between completely manual ontology creation and completely automatic ontology learning

# Outline

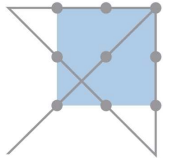
---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. Knowledge derived from Wikipedia
  - (a) Semantic relatedness
  - (b) WikiRelate! Computing semantic relatedness using Wikipedia
  - (c) Knowledge bases, taxonomies, ontologies
  - (d) Deriving a taxonomy from Wikipedia
5. *Exploiting Wikipedia for Coreference Resolution*
6. Further applications
7. Conclusions

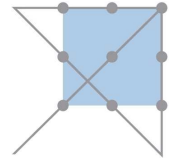
# The Problem – Brown cn15

---



- (1) **Donovan** snatched **Greg's chute** from **him** with a belligerent motion and almost ran to **the plane** with **it**.
- (2) **His face** was dark as the sky above **it** as **he** stood on **the wing** and waited for **his pilot**.
- (3) **Greg** climbed into **the cockpit** feeling as if **he** had never been in one before.
- (4) But **his** hands and those of **Donovan** moved automatically adjusting and arranging in **the check-out procedure**.

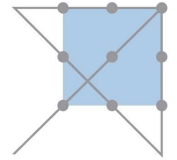
# The Problem – Brown cn15



- (1) **Donovan** snatched **Greg**'s **chute** from **him** with a belligerent motion and almost ran to **the plane** with **it**.
- (2) **His face** was dark as the sky above **it** as **he** stood on **the wing** and waited for **his pilot**.
- (3) **Greg** climbed into **the cockpit** feeling as if **he** had never been in one before.
- (4) But **his** hands and those of **Donovan** moved automatically adjusting and arranging in **the check-out procedure**.

▣▣▣▣▶ **pronouns, named entities**

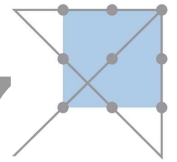
# The Problem – Brown cn15



- (1) **Donovan** snatched **Greg's chute** from **him** with a belligerent motion and almost ran to **the plane** with **it**.
- (2) **His face** was dark as the sky above **it** as **he** stood on **the wing** and waited for **his pilot**.
- (3) **Greg** climbed into **the cockpit** feeling as if **he** had never been in one before.
- (4) But **his** hands and those of **Donovan** moved automatically adjusting and arranging in **the check-out procedure**.

⇒ definite Noun Phrases (NPs), bridging

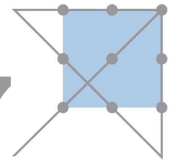
# The Problem – Switchboard sw3117



- A.63 I think **it** really depends a lot on the child, because our daughter is, was just a lot more levelheaded about her proc-, the process.
- B.64 Luckily I still have twelve more years to worry about **it**.
- A.65 Yeah [laughter]  
...
- B.96 ... the University of Virginia. How much did **it** en-, end up costing?
- A.97 ... I would just make it a rough figure of about, uh, with, with the travel expenses and so on, although she didn't come home that much, uh, actually.

▣▶ **vague anaphors, discourse-deictic anaphors**

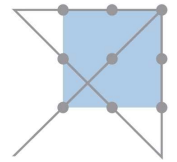
# The Problem – Switchboard sw3117



- A.63 I think **it** really depends a lot on the child, because **our daughter** is, was just a lot more levelheaded about her proc-, the process.
- B.64 Luckily I still have twelve more years to worry about **it**.
- A.65 Yeah [laughter]  
...
- B.96 ... the University of Virginia. How much did **it** en-, end up costing?
- A.97 ... I would just make it a rough figure of about, uh, with, with the travel expenses and so on, although **she** didn't come home that much, uh, actually.

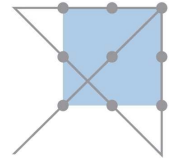
▣▣▣▣▶ **long-distance pronouns, messy spoken language**

# The Problem – Multimodal Dialog



- what does *that one* refer to?!
- ⇒ multimodal dialogues contain deictic pronouns which have a **reference to an extra-linguistic context!**

# Referring Expressions



A.63 I think it really depends a lot on the child, because our daughter was a lot more levelheaded about ...

...

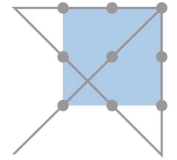
...

B.96 ... the University of Virginia.

How much did it en-, end up costing?

A.97 Uh [breathing], I think, uh, on a yearly basis, I'm trying to think. I would just make it a rough figure of about, uh, with the travel expenses and so on, although she didn't come home that much ...

# Referents



A.63 I think it really depends a lot on the child, because our daughter was a lot more levelheaded about ...



...

...

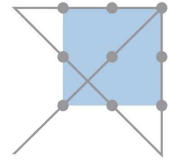
B.96 ... the University of Virginia.

How much did it en-, end up costing?

A.97 Uh [breathing], I think, uh, on a yearly basis, I'm trying to think. I would just make it a rough figure of about, uh, with the travel expenses and so on, although she didn't come home that much ...



# Coreference (cospecification)



A.63 I think it really depends a lot on the child, because our daughter was a lot more levelheaded about ...

...

...

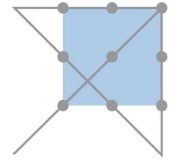
B.96 ... the University of Virginia.

How much did it en-, end up costing?

A.97 Uh [breathing], I think, uh, on a yearly basis, I'm trying to think. I would just make it a rough figure of about, uh, with the travel expenses and so on, although she didn't come home that much ...



# Anaphora



A.63 I think it really depends a lot on the child, because our daughter was a lot more levelheaded about ...

...

...

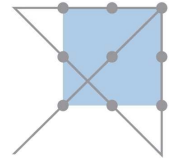
B.96 ... the University of Virginia.

How much did it en-, end up costing?

A.97 Uh [breathing], I think, uh, on a yearly basis, I'm trying to think. I would just make it a rough figure of about, uh, with the travel expenses and so on, although she didn't come home that much ...



# Terminology: (Co-)Reference, ...



**Referring Expression:** NL expression used by discourse participants to refer to entities.

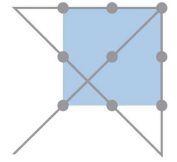
**Referent:** Entity that is referred to.

**Coreference:** two referring expressions that are used to refer to the same referent are said to corefer.

**Anaphora:** Reference to an entity that has been previously introduced in the discourse; the referring expression used is called anaphoric.

**Note:** The speaker is performing the act of referring to an entity by uttering the referring expression Sidner (1983).

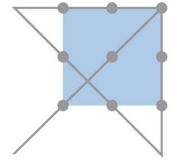
# Bridging



- ! The notion of anaphora can be generalized to **relations other than identity**.
- ➡ So-called **bridging references** Clark (1975) are expressions that refer to objects only related to their antecedent by generic knowledge.

Land looked back toward **the dilapidated house**. He thought he saw a pale face at **a window**. (Brown cl04)

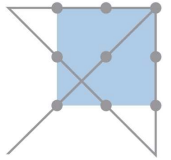
# Referring Expressions



- indefinite noun phrases (*a car, a house*)
- definite noun phrases (*the king of France*)
- proper names (*George W. Bush*)
- pronouns
  - **personal pronouns** (*I, you, he, him, ...*)
  - **possessive pronouns** (*my, your, his, ...*)
  - **reflexive pronouns** (*myself, yourself, himself, ...*)
  - **demonstrative pronouns** (*this, that, ...*)

**Note:** I do not want to talk about plural pronouns with multiple singular antecedents, singular pronouns co-specifying with a member of a set, generics, ...

# Indefinite Noun Phrases



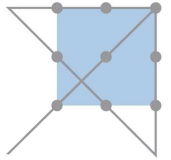
Indefinite NPs introduce entities that are new to the hearer **into the discourse context.**

... America has already made **great contributions in the past two years to the world's fund of knowledge of astrophysics and space science.** (Brown cg35)

But indefinite NPs can also be used for bridging expressions.

Land looked back toward **the dilapidated house.** He thought he saw a pale face at **a window.** (Brown cl04)

# Definite Noun Phrases



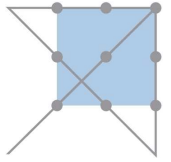
Definite NPs are used to refer to entities which are identifiable to the hearer **in the given discourse context.**

I believe that **the industrial countries** are ready to participate actively in supplementing the efforts of the developing nations to achieve progress. (Brown cg35)

... America has already made **great contributions in the past two years to the world's fund of knowledge of astrophysics and space science. These discoveries** are of present interest chiefly to the scientific community;  
... (Brown cg 35)

# Proper Names

---

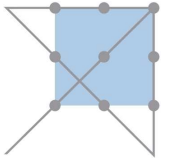


Proper names are used to refer to individuals. Can be used for referring to *both new and identifiable* entities.

**The United States** is always ready to participate with **the Soviet Union** in serious discussion . . . (Brown cg35)

When I interviewed **Kirby, who as a boy picked up pears in the Borden yard**, I asked if anybody else in the household besides Lizzie and Morse had been under any suspicion at the time of the murders. (Brown cf31)

# Pronouns

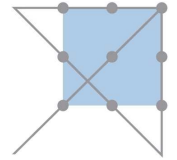


Like definite NPs, pronouns are used to refer to entities which are identifiable to the hearer.

- They require that the **referent is highly salient**.
- ➡ There are also **syntactic, semantic and pragmatic constraints** on the use of pronouns.

“I’m mad”, shouted **Payne**, as **he** ran out into the hall. “I’m mad”, and only wished **he** had been. (Brown ck05)

# What is it good for?



- NL Dialog Systems
  - dialog systems need to consider discourse phenomena
- ⇒ **anaphora are one of the most important discourse phenomena**

*Dave Bowman:* Hello, HAL do you read me, HAL?

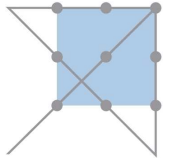
*HAL:* Affirmative, Dave, I read you.

*Dave Bowman:* Open the pod bay doors, HAL.

*HAL:* I'm sorry Dave, I'm afraid I can't do that.

# What is it good for?

---



- **Information Extraction**

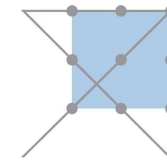
- **⇒ e.g. relation extraction**

- **Microsoft Corporation is a multinational computer technology corporation.**

- **It is headquartered in Redmond, Washington, USA.**

→ **MICROSOFT *located-in* REDMOND**

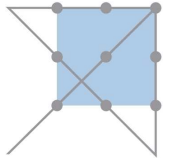
# What is it good for?



- Summarization
- summarization systems often just **determine important sentences** and **extract them**
- ⇒ if these sentences contain pronouns but not their antecedents, the pronouns have to be resolved
  - Presidential elections were held yesterday in France.
  - ? **They** said that it has been a long night.
- ⇒ the choice of important sentences **is affected by resolving anaphors**
  - **TOPIC**: significant events for N. Sarkozy in 2007.
  - **He** was elected on May 16, 2007.

# Coreference Resolution

---



- linguistically motivated approaches based on focusing/centering (Sidner, 1983; Brennan et al., 1987; Strube & Hahn, 1996, 1999; Tetreault, 2001; ...)
- heuristics (Hobbs, 1978; Lappin & Leass, 1994; Kennedy & Boguraev, 1996; Baldwin, 1997; ...)
- ML-based approaches (McCarthy & Lehnert, 1995; Aone & Bennett, 1995; Ge et al., 1998; Soon et al., 2001; ...)

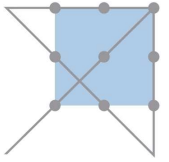
# Soon et al. (2001): A baseline system



---

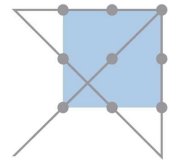
1. MUC-style coreference resolution
2. Evaluation
3. Coreference resolution as a machine learning task
4. (No) knowledge

# MUC-style Coreference Resolution



- **task definition:** find the *discourse entities (DE)* a set of *referring expressions (RE)* (are used to) refer to
- includes all kind of NPs
  - pronouns (*p*)
  - common nouns (*cn*)
  - proper names (*pn*)

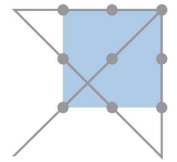
# MUC-style Coreference Resolution



## (1) raw text

But frequent visitors say that given the sheer weight of the country's totalitarian ideology and generations of mass indoctrination, changing this country's course will be something akin to turning a huge ship at sea. Opening North Korea up, even modestly, and exposing people to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as his people begin to get a clearer idea of the deprivation they have suffered, especially relative to their neighbors. “This is a society that has been focused most of all on stability, [...]”.

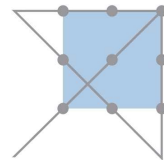
# MUC-style Coreference Resolution



## (2) RE identification

But frequent visitors say that given the sheer weight of the country's totalitarian ideology and generations of mass indoctrination, changing this country's course will be something akin to turning a huge ship at sea. Opening North Korea up, even modestly, and exposing people to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as his people begin to get a clearer idea of the deprivation they have suffered, especially relative to their neighbors. “This is a society that has been focused most of all on stability, [...]”.

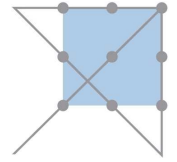
# MUC-style Coreference Resolution



(3) **DE identification** → create *coreference chains*

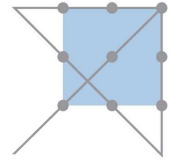
But frequent visitors say that given the sheer weight of the country's totalitarian ideology and generations of mass indoctrination, changing this country's course will be something akin to turning a huge ship at sea. Opening North Korea up, even modestly, and exposing people to the idea that Westerners – and South Koreans – are not devils, alone represents an extraordinary change. [...] as his people begin to get a clearer idea of the deprivation they have suffered, especially relative to their neighbors. “This is a society that has been focused most of all on stability, [...]”.

# MUC-style Coreference Resolution



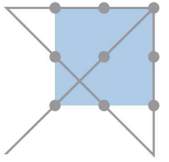
- MUC: Message Understanding Conference held until 1997
- evaluation by **MUC scoring program** Vilain et al. (1995):  
precision, recall, F-measure
- standard partitions: 30 docs training, 30 docs testing (MUC 6); 30 docs training, 20 docs testing (MUC 7)

# ML based coreference resolution



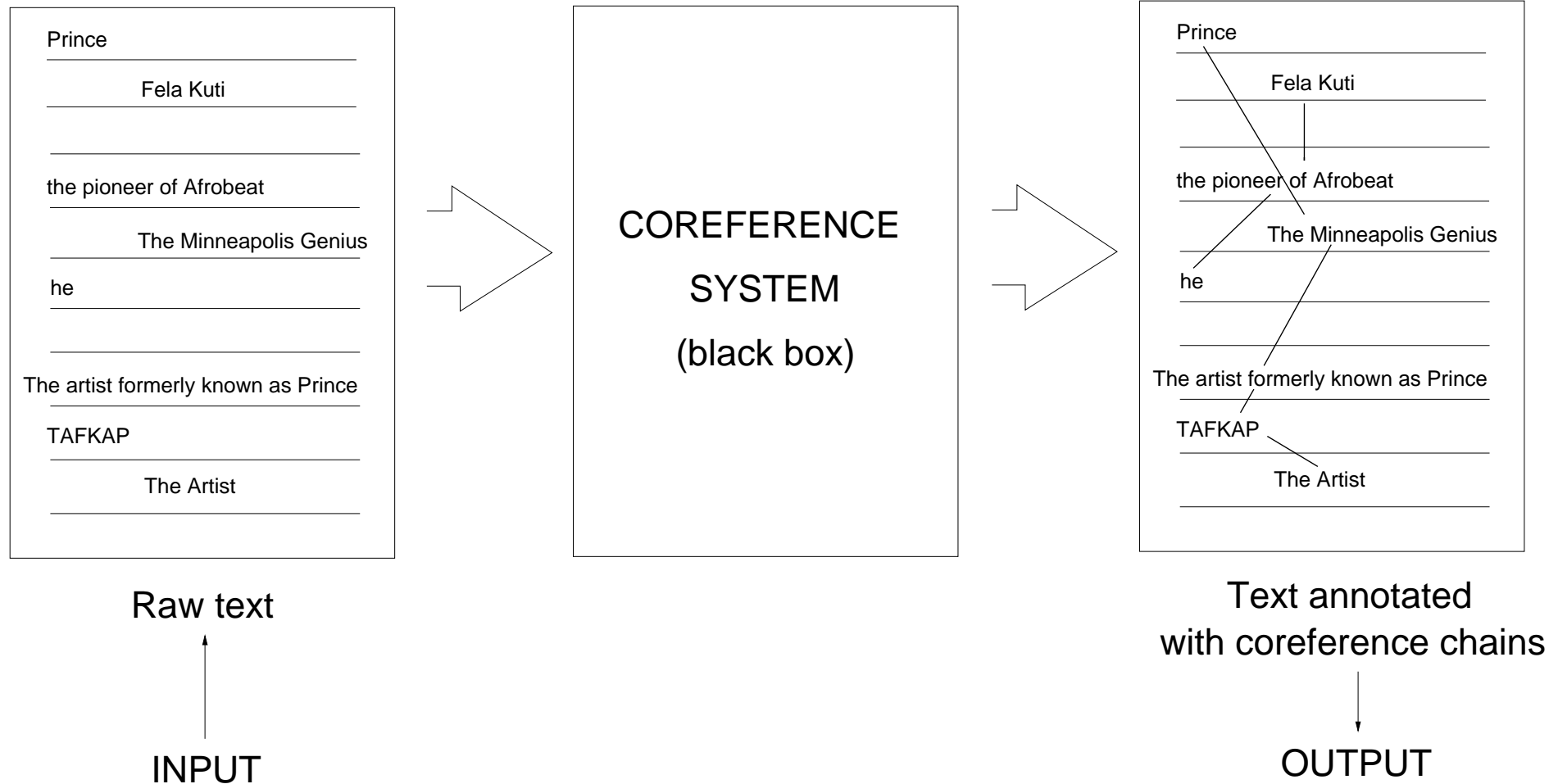
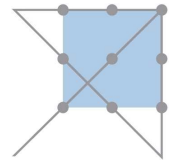
- most of the approaches model coreference resolution as a **classification task** (Soon et al., 2001)
- further work on **exploring different kinds of data representations, task definitions or machine learning techniques** – cf. Ng & Cardie (2002), Yang et al. (2003), Luo et al. (2004), Ng (2005)

# ML based coreference resolution

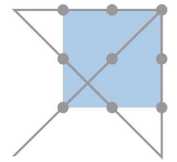


- BUT semantics has been pointed out as being relevant since seminal work, e.g. Charniak (1973) and Hobbs (1978)
- we concentrate therefore on including semantic knowledge into the model!
- we follow previous work on including semantics in ML coreference models
  - Harabagiu et al. (2001): *WordNet*
  - Kehler et al. (2004): *predicate-argument statistics*
  - Yang et al. (2005): *semantic compatibility information*
- we show that semantics helps coreference resolution!

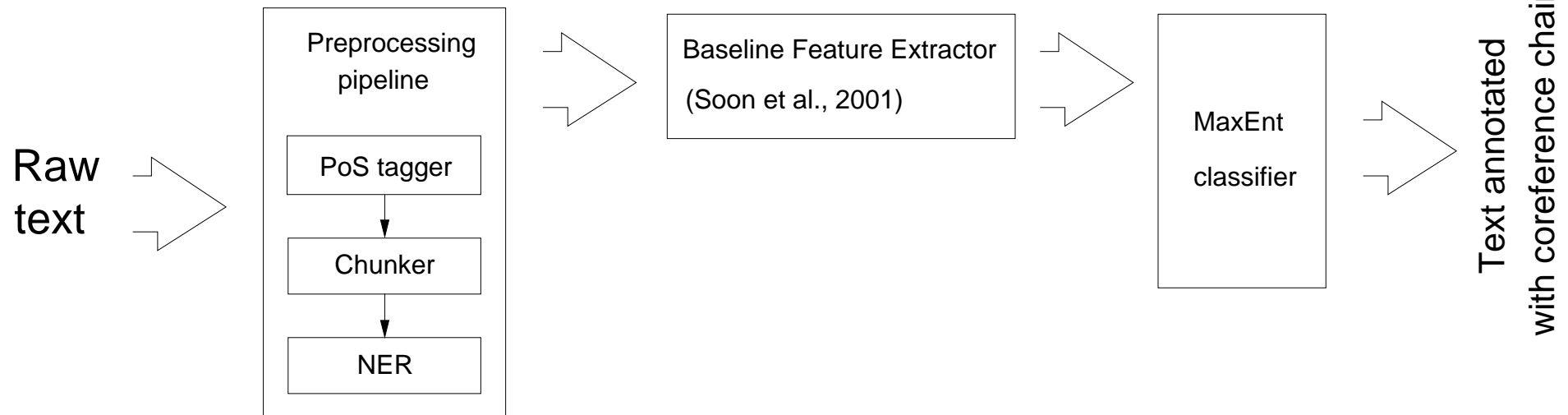
# System outline



# System outline

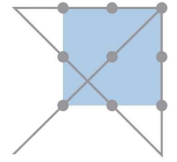


## 1. use the system from Soon et al. as *baseline*

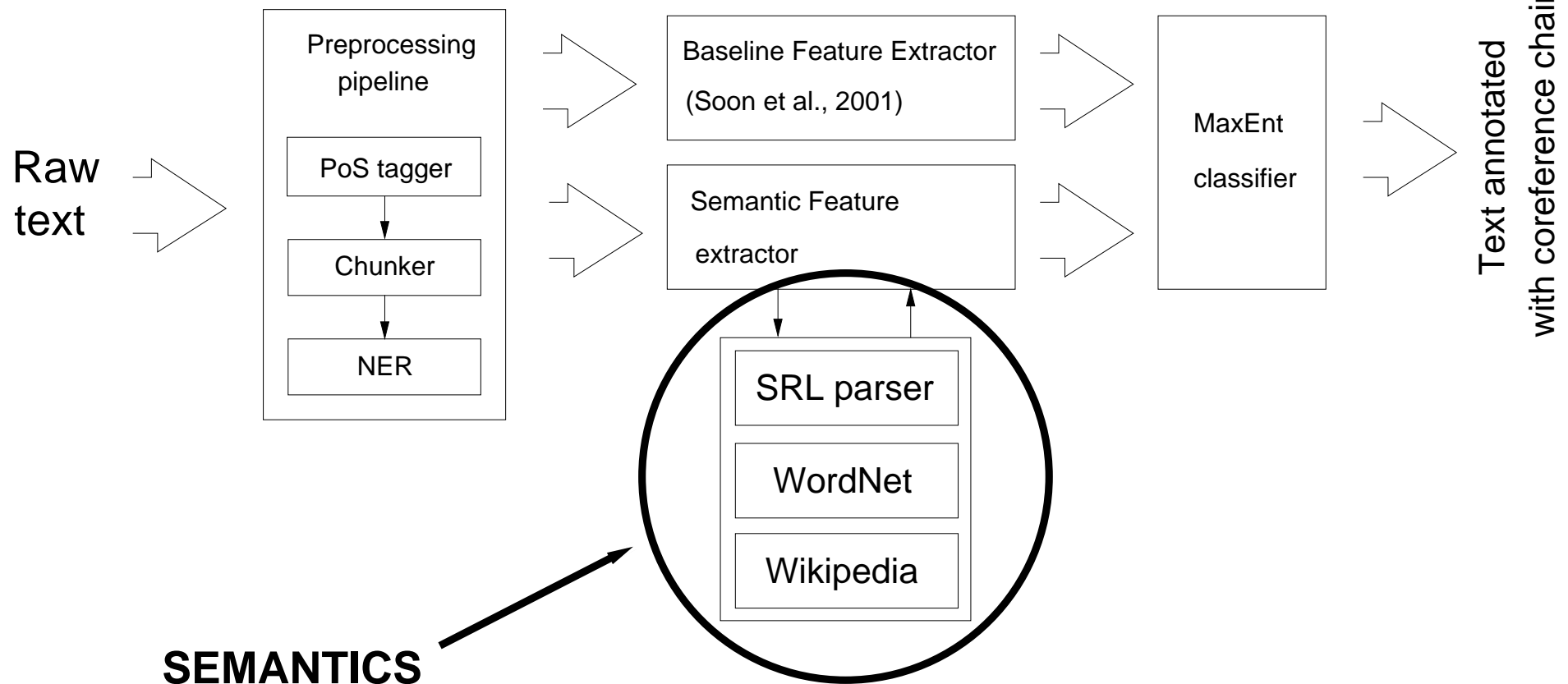


- **PoS tagger:** SVMTool (Giménez & Màrquez, 2004)
- **Chunker:** YamCha (Kudoh & Matsumoto, 2000)
- **NER:** *Alias-1 LingPipe* Named Entity Recognizer
- **MaxEnt toolkit:** MALLET (McCallum, 2002)

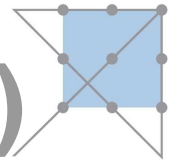
# System outline



1. use the system from Soon et al. as *baseline*
2. enlarge the feature set with features generated from *semantic knowledge sources*



# Baseline features (Soon et al., 2001)



The learner builds a “local model”

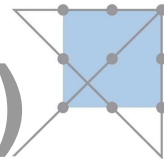
- learning *instances* are pairs of REs
- the classifier has to learn whether they are coreferent or not

Given a pair of candidate referring expressions  $RE_i$  and  $RE_j$

(a) Lexical features

- **STRING\_MATCH** T if  $RE_i$  and  $RE_j$  have the same spelling, else F
- **ALIAS** T if one RE is an alias of the other; else F

# Baseline features (Soon et al., 2001)

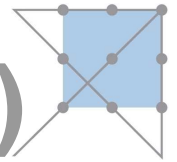


Given a pair of candidate referring expressions  $RE_i$  and  $RE_j$

## (b) Grammatical features

- **I\_PRONOUN** T if  $RE_i$  is a pronoun; else F
- **J\_PRONOUN** T if  $RE_j$  is a pronoun; else F
- **J\_DEF** T if  $RE_j$  starts with *the*; else F
- **J\_DEM** T if  $RE_j$  starts with *this, that, these, or those*; else F
- **NUMBER** T if both  $RE_i$  and  $RE_j$  agree in number; else F
- **GENDER** U if either  $RE_i$  or  $RE_j$  have an undefined gender. Else if they are both defined and agree T; else F
- **PROPER\_NAME** T if both  $RE_i$  and  $RE_j$  are proper names; else F
- **APPOSITIVE** T if  $RE_j$  is in apposition with  $RE_i$ ; else F

# Baseline features (Soon et al., 2001)



Given a pair of candidate referring expressions  $RE_i$  and  $RE_j$

(c) **Semantic features**

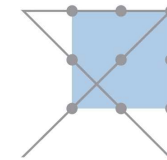
- **WN\_CLASS** U if either  $RE_i$  or  $RE_j$  have an undefined WordNet semantic class. Else if they both have a defined one and it is the same T; else F

(d) **Distance features**

- **DISTANCE** how many sentences  $RE_i$  and  $RE_j$  are apart

⇒ except for (c) they are all **surface features!**

# Results

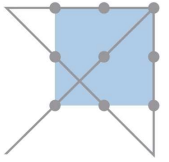


	MUC-6			MUC-7		
original	R	P	$F_1$	R	P	$F_1$
Soon et al.	58.6	67.3	62.3	56.1	65.5	60.4
duplicated baseline	64.9	65.6	65.3	55.1	68.5	61.1

slight increase in performance due to

- **more recent preprocessing components**, in particular named entity recognition (Alias-I)
- **different classifier** (MaxEnt instead of C5.0)

# (No) Knowledge



- except for the WordNet feature **only surface features** and even this feature checks only for WordNet classes
- among the other features the **string match features works best** (for proper names and common nouns)
- **room for improvement**
  - ➡ **including semantic, conceptual knowledge**

# Semantics for coreference resolution



- in the area of anaphora resolution there is a **large body of work** on the necessity of semantics and knowledge
  - ML-based work mostly neglected this issue
- ! very simple surface features seem to work as well
- ▣▣▣▣ *distance features* for pronoun resolution
  - ▣▣▣▣ *string-based features* for common nouns
  - ▣▣▣▣ *gazetteers* for proper names
- ▣▣▣▣ however, researchers also noticed a plateauing of performance on a rather low level (e.g. Kehler et al. (2004))
- ! with the advance of NLP techniques, it may now be the right time to check out semantics and ontologies again

# Semantics for coreference resolution



---

*which* kind of **semantics** is useful for coreference resolution?

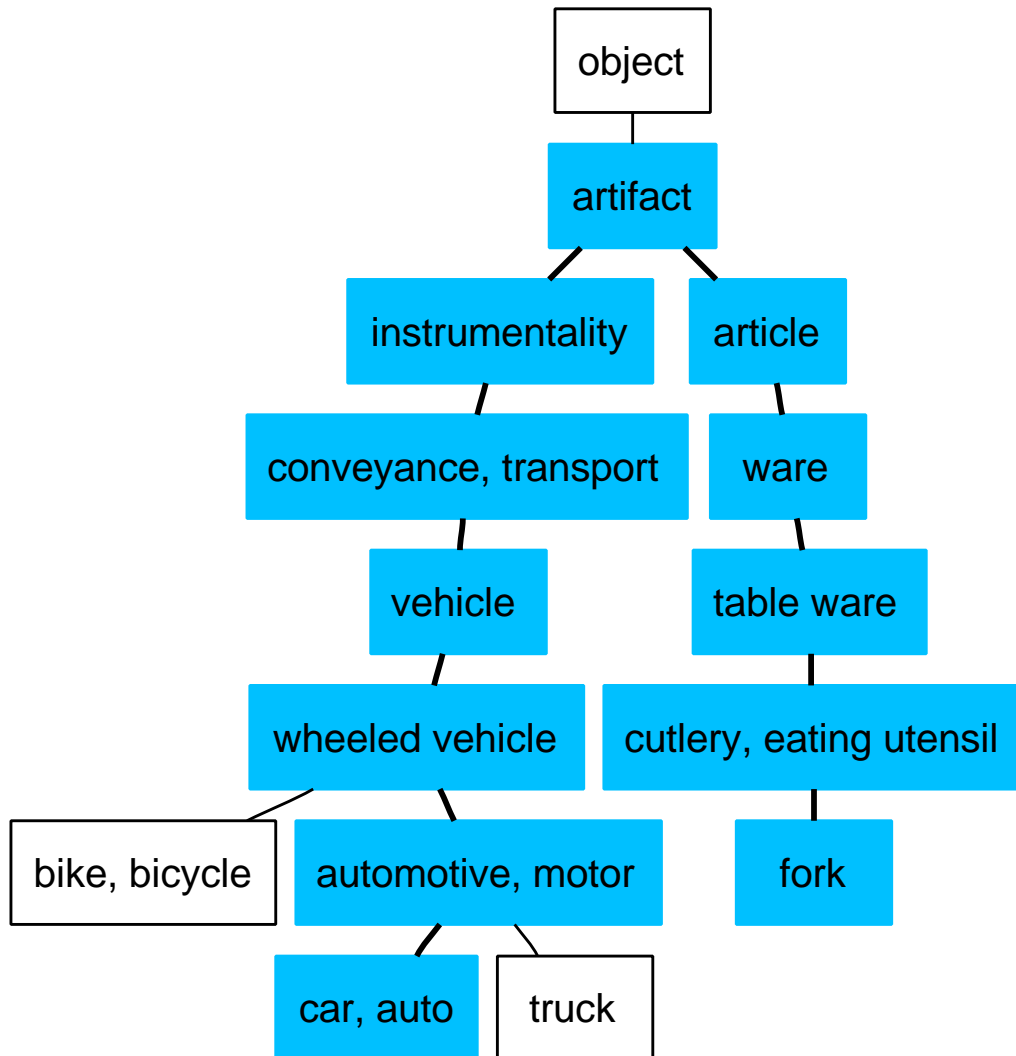
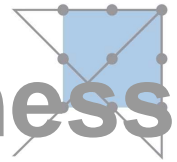
Core idea from Ponzetto & Strube (2006):

▣ include semantic similarity/relatedness from

- **WordNet**
- **Wikipedia**

! into the set of features made available to the classifier

# Taxonomy-based semantic relatedness

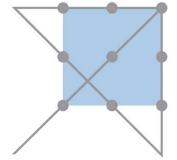


e.g. using *node counting scheme*

$$\text{sim}(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}$$

- $\text{sim}(\text{car}, \text{auto}) = 1$
- $\text{sim}(\text{car}, \text{bike}) = 0.25$
- $\text{sim}(\text{car}, \text{fork}) = 0.08$

# WordNet similarity features



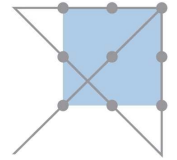
- if it's not a named entity, identify the lexical heads of  $RE_i$  and/or  $RE_j$

a young star	⇒	star
Miles Davis	⇒	Miles Davis
a Meinl-Weston trumpet	⇒	trumpet

- collect all available senses  $SENSE_{RE_i}$  and  $SENSE_{RE_j}$  of (the heads of)  $RE_i$  and  $RE_j$

star	⇒	a celestial body
	⇒	a plane figure with 5 or more points
	⇒	an actor who plays a principal role

# WordNet similarity features

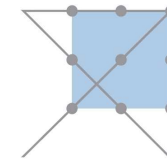


- compute similarity of sense pairs using *path length* based measures (Rada et al., 1989; Wu & Palmer, 1994; Leacock & Chodorow, 1998), as well as *information content* based measures (Resnik, 1995; Jiang & Conrath, 1997; Lin 1998)
- for each measure, introduce two features

(f) WN similarity features

- **WN\_SIMILARITY\_BEST** the *highest* similarity score from all  $\langle \text{SENSE}_{RE_i, n}, \text{SENSE}_{RE_j, m} \rangle$  synset pairs.
- **WN\_SIMILARITY\_AVG** the *average* similarity score from all  $\langle \text{SENSE}_{RE_i, n}, \text{SENSE}_{RE_j, m} \rangle$  synset pairs.

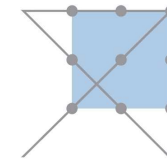
# Evaluation WN Similarity Features



ACE 2003 – BNEWS section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	46.7	86.2	60.6	36.4	10.5	44.0
+WordNet	<b>54.8</b>	86.1	<b>66.9</b>	<b>36.8</b>	<b>24.8</b>	<b>47.6</b>

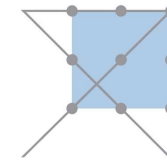
# Evaluation WN Similarity Features



ACE 2003 – NWIRE section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	56.7	88.2	69.0	37.6	23.1	55.6
+WordNet	<b>61.3</b>	84.9	<b>71.2</b>	<b>38.9</b>	<b>30.8</b>	55.5

# Evaluation WN Similarity Features



ACE 2003 – MERGED section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	54.5	88.0	67.3	34.7	20.4	53.1
+WordNet	<b>56.7</b>	87.1	<b>68.6</b>	<b>35.6</b>	<b>28.5</b>	49.6

# Semantic Relatedness Using Wikipedia

---

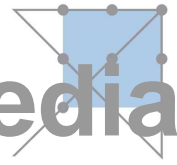
Basic idea (Strube & Ponzetto, 2006):

- almost every article is categorized
- **categories** are related in a taxonomic fashion and *can be used as a semantic network*

Three main steps

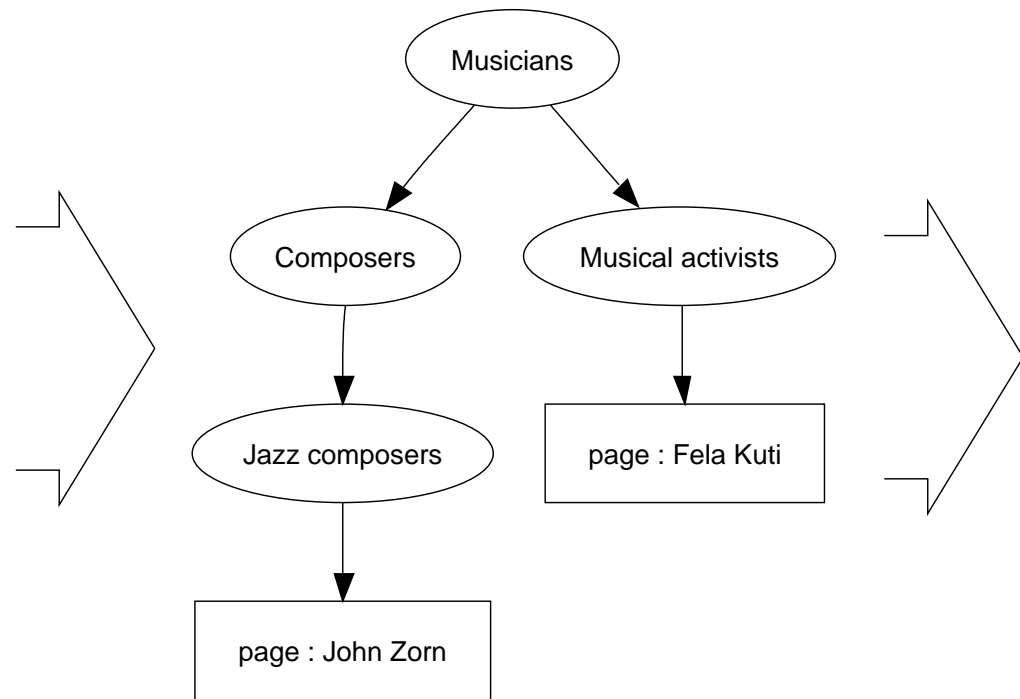
1. page retrieval and disambiguation
2. category tree search
3. relatedness measure computation

# Semantic Relatedness Using Wikipedia



The top screenshot shows the Wikipedia page for John Zorn. In the 'Categories' section, 'Jazz composers' is circled. The bottom screenshot shows the Wikipedia page for Fela Kuti. In the 'Categories' section, 'Musical activists' is circled.

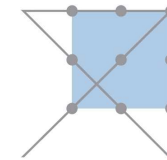
page query and retrieval  
category extraction



search for a connecting path  
along the category tree

relatedness measure(s) computation

# Wikipedia page based features

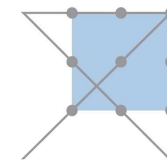


Given a candidate coreference pair  $RE_{i/j}$  and the Wikipedia pages  $P_{RE_{i/j}}$  they point to, obtained by querying pages titled as  $T_{RE_{i/j}}$ , we extract the following features:

(g) Wikipedia page features

- **I/J\_GLOSS\_CONTAINS** U if no Wikipedia page titled  $T_{RE_{i/j}}$  is available. Else T if the first paragraph of text of  $P_{RE_{i/j}}$  contains  $T_{RE_{j/i}}$ ; else F.
- **I/J\_RELATED\_CONTAINS** U if no Wikipedia page titled as  $T_{RE_{i/j}}$  is available. Else T if at least one Wikipedia hyperlink of  $P_{RE_{i/j}}$  contains  $T_{RE_{j/i}}$ ; else F.

# Wikipedia page based features

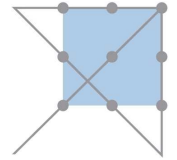


Given a candidate coreference pair  $RE_{i/j}$  and the Wikipedia pages  $P_{RE_{i/j}}$  they point to, obtained by querying pages titled as  $T_{RE_{i/j}}$ , we extract the following features:

(g) Wikipedia page features

- **I/J\_CATEGORIES\_CONTAINS** U if no Wikipedia page titled as  $T_{RE_{i/j}}$  is available. Else T if the list of categories  $P_{RE_{i/j}}$  belongs to contains  $T_{RE_{j/i}}$ ; else F.
- **GLOSS\_OVERLAP** the overlap score between the first paragraph of text of  $P_{RE_i}$  and  $P_{RE_j}$ .

# Wikipedia relatedness features

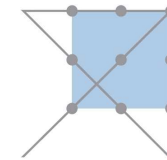


- collect categories  $CAT_{RE_i}$  and  $CAT_{RE_j}$  of the pages for the heads of  $RE_i$  and  $RE_j$
- compute relatedness of category pairs: for each measure, introduce two features

(h) Wikipedia relatedness features

- **WIKI\_RELATEDNESS\_BEST** the *highest* relatedness score from all  $\langle CAT_{RE_i,n}, CAT_{RE_j,m} \rangle$  category pairs.
- **WIKI\_RELATEDNESS\_AVG** the *average* relatedness score from all  $\langle CAT_{RE_i,n}, CAT_{RE_j,m} \rangle$  category pairs.

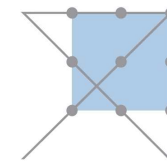
# Evaluation of Wikipedia Features



ACE 2003 – BNEWS section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	46.7	86.2	60.6	36.4	10.5	44.0
+WordNet	54.8	86.1	66.9	36.8	24.8	47.6
+Wikipedia	<b>52.7</b>	<b>86.8</b>	<b>65.6</b>	36.1	<b>23.5</b>	<b>46.2</b>

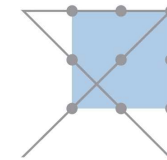
# Evaluation of Wikipedia Features



## ACE 2003 – NWIRE section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	56.7	88.2	69.0	37.6	23.1	55.6
+WordNet	<b>61.3</b>	84.9	<b>71.2</b>	<b>38.9</b>	<b>30.8</b>	55.5
+Wikipedia	<b>60.6</b>	83.6	<b>70.3</b>	<b>38.0</b>	<b>29.7</b>	55.2

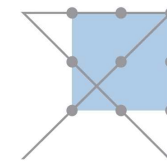
# Evaluation of Wikipedia Features



## ACE 2003 – MERGED section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	54.5	88.0	67.3	34.7	20.4	53.1
+WordNet	56.7	87.1	68.6	35.6	28.5	49.6
+Wikipedia	<b>55.8</b>	87.5	<b>68.1</b>	<b>34.8</b>	<b>26.0</b>	50.5

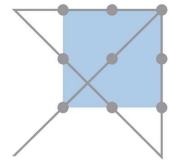
# Joining Forces Pays Off



## ACE 2003 – BNEWS section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	46.7	86.2	60.6	36.4	10.5	44.0
+SRL	53.3	85.1	65.5	37.1	13.9	46.2
+WordNet	54.8	86.1	66.9	36.8	24.8	47.6
+Wikipedia	52.7	86.8	65.6	36.1	23.5	46.2
→ all features	<b>59.1</b>	84.4	<b>69.5</b>	<b>37.5</b>	<b>27.3</b>	<b>48.1</b>

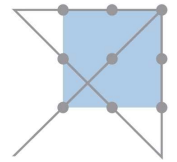
# Joining Forces Pays Off



## ACE 2003 – NWIRE section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	56.7	88.2	69.0	37.6	23.1	55.6
+SRL	58.0	89.0	70.2	38.3	25.0	56.0
+WordNet	61.3	84.9	71.2	38.9	30.8	55.5
+Wikipedia	60.6	83.6	70.3	38.0	29.7	55.2
→ all features	<b>63.1</b>	83.0	<b>71.7</b>	<b>39.8</b>	<b>31.8</b>	52.8

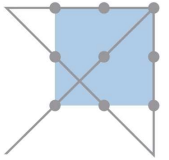
# Joining Forces Pays Off



## ACE 2003 – MERGED section

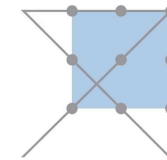
	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	54.5	88.0	67.3	34.7	20.4	53.1
+SRL	56.3	88.4	68.8	38.9	21.6	51.7
+WordNet	56.7	87.1	68.6	35.6	28.5	49.6
+Wikipedia	55.8	87.5	68.1	34.8	26.0	50.5
→ all features	<b>61.0</b>	84.2	<b>70.7</b>	<b>38.9</b>	<b>29.9</b>	51.2

# Feature selection (BNEWS section)



Feature set	$F_1$
baseline (Soon w/o DISTANCE)	58.4%
+WIKI_WU_PALMER_BEST	+4.3%
+J_SEMROLE	+1.8%
+WIKI_PATH_AVG	+1.2%
+I_SEMROLE	+0.8%
+WN_WU_PALMER_BEST	+0.7%

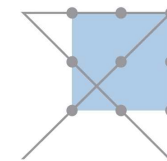
# Latest Results (JAIR paper)



## ACE 2003 – MERGED section

	R	P	$F_1$	$A_p$	$A_{cn}$	$A_{pn}$
baseline	54.5	85.4	66.5	40.5	30.1	73.0
+WordNet	<b>60.6</b>	79.4	<b>68.7</b>	<b>42.4</b>	<b>43.2</b>	66.0
+Wikipedia	<b>59.4</b>	82.2	<b>68.9</b>	38.9	<b>41.4</b>	<b>74.5</b>

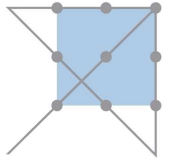
# Latest Results (JAIR paper)



		BNEWS	NWIRE	MERGED
<b>backward feature selection</b>	starting	WN_CLASS		
	features	PROPER_NAME	J_DEM	DISTANCE
	removed	J_DEM		
<b>forward feature selection</b>	WordNet features added	<i>jcn</i> average <i>jcn</i> shortest <i>pl</i> shortest <i>wup</i> average	<i>jcn</i> shortest <i>lin</i> average <i>pl</i> shortest	<i>lch</i> shortest <i>pl</i> shortest <i>pl</i> average
	Wikipedia features added	<i>wup</i> average <i>pl</i> shortest <i>pl</i> average <i>gloss</i>	<i>wup</i> average <i>lch</i> shortest	<i>pl</i> shortest <i>pl</i> average

# Conclusions – Results

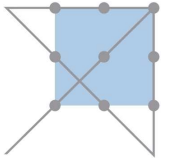
---



- ▶ empirical results show that **coreference resolution benefits from semantic features**
  - WordNet and Wikipedia features increase performance on common nouns and proper names
  - features induced from Wikipedia are competitive with the ones from WordNet

# Scientific Methodology

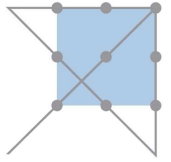
---



- first establish a **competitive baseline** (system)
- only then get **bright ideas**
- change only **one variable at a time**
- perform **significance tests**
- perform **feature analysis**

# Take-home message 7

---

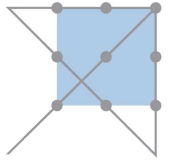


Coreference resolution and conceptual knowledge:

- everyone agrees that coreference resolution should take more semantic and conceptual knowledge into account
- Wikipedia is a strong contender as a ontological resource in NLP and **may outperform WordNet** soon in certain tasks
- WordNet and Wikipedia combined should always provide **better results**

# Outline

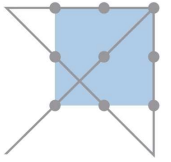
---



1. NLP and encyclopedic knowledge
2. Why Wikipedia?
3. NLP applications using Wikipedia: Question Answering, Explicit Semantic Analysis, Word Sense Disambiguation
4. Knowledge derived from Wikipedia
  - (a) Semantic relatedness
  - (b) WikiRelate! Computing semantic relatedness using Wikipedia
  - (c) Knowledge bases, taxonomies, ontologies
  - (d) Deriving a taxonomy from Wikipedia
5. Exploiting Wikipedia for Coreference Resolution
6. Further applications
7. *Conclusions*

# Take-home message 1

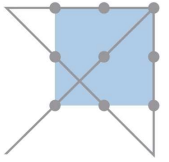
---



- Wikipedia text is **semi-structured**
- ⇒ we can take the structure to **represent (implicit) information**

# Take-home message 2

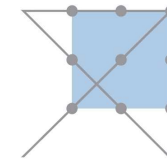
---



- Wikipedia articles can be taken as **concepts**
- ⇒ we can use Wikipedia as a **semantic space**

# Take-home message 3

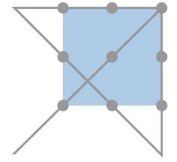
---



- Wikipedia hyperlinks can be used as **sense/entity annotated data**
- ⇒ we can use Wikipedia as a **data resource for complex NLP tasks**

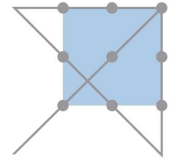
# Take-home message 4

---



- distributional similarity is computed on the basis of large corpora – they are computationally expensive and do not scale well
- easy to implement but requires very large corpora
- semantic relatedness is computed using semantic networks – their quality pretty much depends on the availability and size of the resource
- API for WordNet available

# Take-home message 5

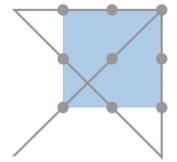


- Wikipedia categories can be taken as semantic network
  - relatedness measures computed from Wikipedia give a performance on word pair lists *competitive with WordNet*
- ⇒ Wikipedia is a promising resource
- worked 'out of the box' on a first attempt
  - can be freely downloaded and used right away
  - we used only part of it (i.e. limited use of textual content)
  - shows exponential growth

Wikipedia is a resource to be exploited for NLP applications

# Take-home message 6

---

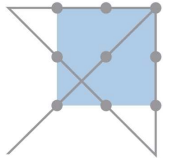


transform semantic network into a large scale, high-quality taxonomy

- ➡ it could be that a folksonomy is exactly what we need for AI and NLP applications
- ➡ achieving a *large coverage* and *robust* taxonomy by **collaboration** of many users
- ➡ generated by those users whose behavior we are trying to model in our ‘intelligent’ applications
- ➡ we stake out a middleground between completely manual ontology creation and completely automatic ontology learning

# Take-home message 7

---

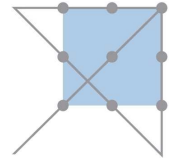


Coreference resolution and conceptual knowledge:

- everyone agrees that coreference resolution should take more semantic and conceptual knowledge into account
- Wikipedia is a strong contender as a ontological resource in NLP and **may outperform WordNet** soon in certain tasks
- WordNet and Wikipedia combined should always provide **better results**

# Conclusions

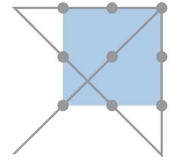
---



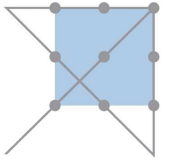
- building large scale knowledge bases has been unsuccessful (with the exception of WordNet and Cyc)
- the OpenMind ([www.openmind.org](http://www.openmind.org)) and MindPixel ([www.mindpixel.com](http://www.mindpixel.com)) projects rely on volunteers to build knowledge bases – however, entrance barrier is too high
- the Semantic Web also did not take off
- Wikipedia may allow to derive a high quality ontological resource

# Conclusions

---



- Wikipedia can be used as corpus and as resource
- Wikipedia may turn out as one of the most useful resources for NLP, because
  - it is a lot of text
  - it is structured (paragraphs, tables, articles, hyperlinks, categories, lists, . . . )
  - it is multilingual
  - is is multimedial
- and we only scratched the surface . . .



---

More information – papers, software, semantic network and taxonomy – and an updated version of these slides – available at

<http://www.eml-research.de/~strube>

and

<http://www.eml-research.de/nlp>

Download annotation tool MMAX2 at

<http://mmax2.sourceforge.net>

***Wanna get involved? – Drop me a line!***