ESTC2008

2ND ANNUAL EUROPEAN SEMANTIC TECHNOLOGY CONFERENCE

Getting at the Semantics of Texts

HANS USZKOREIT

DFKI AND SAARLAND UNIVERSITY

- ☆ Looking Back
- ☆ Extraction of Relation Instances
- ☆ Deep Syntactic/Semantic Processing
- ☆ Hybrid Processing Models
- ☆ Looking Ahead

Do we have Artificial Intelligence?

- ☆ AI cannot simulate a four year old child.
- ☆ But AI can beat the world champion in chess.

- ☆ Today no HLT program cannot read a given three word sentence in a regretful tone.

- ☆ But we can build an HLT program that can read texts in ten languages without major pronounciation errors and without a foreign accent.

- ☆ HLT cannot understand the full meaning of any one sentence the charta of human rights

- ☆ But HLT can find a single mentioning of some relevant event in millions of sentences within seconds.

☆ Our success may not be sweeping...

Types of Information Extraction in LT

- ☆ Topic Extraction
- ☆ Term Extraction
- ☆ Named Entity Extraction
- ☆ Binary Relation Extraction
- ☆ Nary Relation Extraction
- ☆ Event Extraction
- ☆ Answer Extraction
- ☆ Opinion Extraction
- ☆ Sentiment Extraction

Types of Information Extraction in LT

- ☆ Topic Extraction
- ☆ Term Extraction
- ☆ Named Entity Extraction
- ☆ Binary Relation Extraction
- ☆ Nary Relation Extraction
- ☆ Event Extraction
- ☆ Answer Extraction
- ☆ Opinion Extraction
- ☆ Sentiment Extraction

Types of Relation Extraction

- ☆ Even if we do not analyse every phrase or sentence, we need a grammar or an equivalent classifier to detect the relevant relations in context.

- ☆ Here we face the problem of grammar acquisition we know from language parsing.

Main Methods

- ☆ introspective intellectual development
- ☆ corpus-based intellectual development
- ☆ supervised learning from annotated data
- ☆ minimally supervised learning from examples
- ☆ completely unsupervised learning

Main Methods

☆ introspective intellectual development (expensive, inclomplete)

☆ corpus-based intellectual development (expensive, slow)

☆ supervised learning from annotated data (expensive)

☆ minimally supervised learning from examples (promising results)

☆ completely unsupervised learning (not feasible for most tasks)

☆ Unsupervised and minimally supervised automatic acquisition methods for relation extraction patterns, e.g.,

- (Brin, 1998)
- (Agichtein & Gravano, 2000)
- (Yangarber, et al., 2000), (Yangarber, 2003)
- (Sudo et al., 2003)
- (Greenwood & Stevenson, 2006)
- (Xu, Uszkoreit & Li 2007)
- (Xu, Uszkoreit & Li 2008)

☆ Pattern-oriented (e.g., ExDisco (Yangarber 2001))

– too closely bound to the linguistic representation of the seed, e.g.,

subject(company) v("appoint") object(person)

– an event can be expressed by more than one pattern and by various linguistic constructions

☆ Relation and event instances as seeds, e.g., DIPRE (Brin 1998), Snowball (Agichtein and Gravano 2000), (Xu et al. 2006, 2007)

Our Approach: *DARE*

☆ seed-driven and bottom-up rule learning in a bootstrapping framework

- starting from sample relation instances as seeds
  - complexity of the seed instance defines the complexity of the target relation
- pattern discovery is bottom-up and compositional, i.e., complex patterns are derived from simple patterns for relation projections
- bottom-up compression method to cluster and generalize rules
- only subtrees containing seed arguments are pattern candidates

Our Approach: *DARE (1)*

- ☆ seed-driven and bottom-up rule learning in a bootstrapping framework
  - – starting from sample relation instances as seeds
    - complexity of the seed instance defines the complexity of the target relation
  - – pattern discovery is bottom-up and compositional, i.e., complex patterns are derived from simple patterns for relation projections
  - – bottom-up compression method to cluster and generalize rules
  - – only subtrees containing seed arguments are pattern candidates
  - – pattern rule ranking and filtering method considers two aspects of a pattern
    - its domain relevance and
    - the trustworthiness of its origin

Compositional rule representation model

- support the bottom-up rule composition
- expressive enough for the representation of rules for various complexity
- precise assignment of semantic roles to the slot arguments
- reflects the precise linguistic relationship among the relation arguments and reduces the template merging task in the later phase
- the rules for the subset of arguments (projections) may be reused for other relation extraction tasks.

☆ Award Events

start with subdomain Nobel Prizes

reasons: good news coverage
complete list of all award events
good starting point for other award domains

☆ Management Succession Events

reason: comparison with previous work

☆ Target relation

<recipient, prize, area, year>

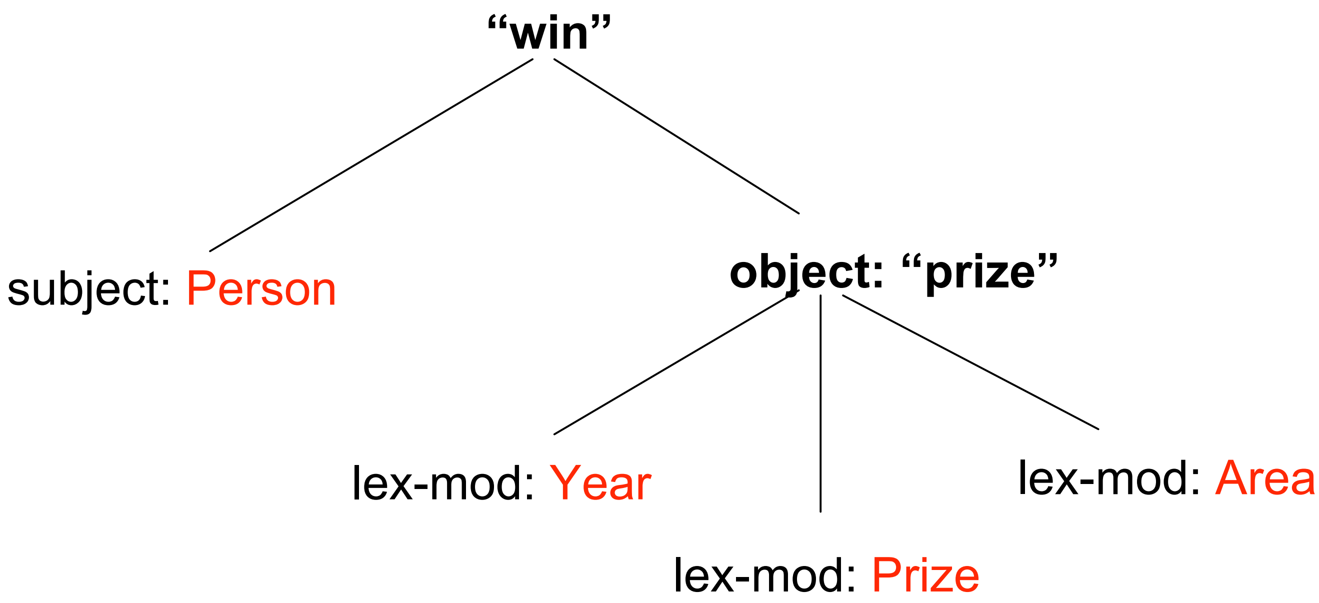
☆ Example

Seed: < "*Mohamed ElBaradei*", "*Nobel*", "*Peace*", "*2005*">

Sentence: Mohamed ElBaradei won the 2005 Nobel Peace Prize on Friday for his efforts to limit the spread of atomic weapons.

We recognize named entities with the DFKI IE-system SProUT (Piskorski, Xu, et al.)

For sentence analysis we employ MiniPar (Dekan Lin)



`Rule name::` `prize_area_year_1`

`Rule body::`

$$\left[\begin{array}{ll} \text{head} & \left[\begin{array}{ll}\text{pos} & \text{noun}\\ \text{lex-form} & \text{"prize"}\end{array}\right] \\ \text{daughters} & \left\langle \begin{array}{l} \left[\text{lex-mod} \quad \left[\text{head} \quad \boxed{3}\ \text{Year}\right]\right], \\ \left[\text{lex-mod} \quad \left[\text{head} \quad \boxed{1}\ \text{Prize}\right]\right], \\ \left[\text{lex-mod} \quad \left[\text{head} \quad \boxed{2}\ \text{Area}\right]\right] \end{array} \right\rangle \end{array}\right]$$

`Output::` $\langle \boxed{1} Prize, \boxed{2} Area, \boxed{3} Year \rangle$

`Rule name::` `recipient_prize_area_year_1`

`Rule body::`

$$\begin{bmatrix} \text{head} & \begin{bmatrix}\text{pos} & \text{verb}\\ \text{mode} & \text{active}\\ \text{lex-form} & \text{"win"}\end{bmatrix}\\ \text{daughters} & \left\langle \begin{array}{l} \begin{bmatrix}\text{subject} & \begin{bmatrix}\text{head} & \boxed{1}\ \text{Person}\\ \text{rule} & \text{recipient\_1::}\ \langle \boxed{1} Person\rangle\end{bmatrix}\end{bmatrix}, \\ \begin{bmatrix}\text{object} & \begin{bmatrix}\text{head} & \begin{bmatrix}\text{lex-form} & \text{"prize"}\end{bmatrix}\\ \text{rule} & \text{prize\_area\_year\_1::}\ \langle \boxed{2}Prize, \boxed{3}Area, \boxed{4}Year\rangle\end{bmatrix}\end{bmatrix} \end{array}\right\rangle\end{bmatrix}$$

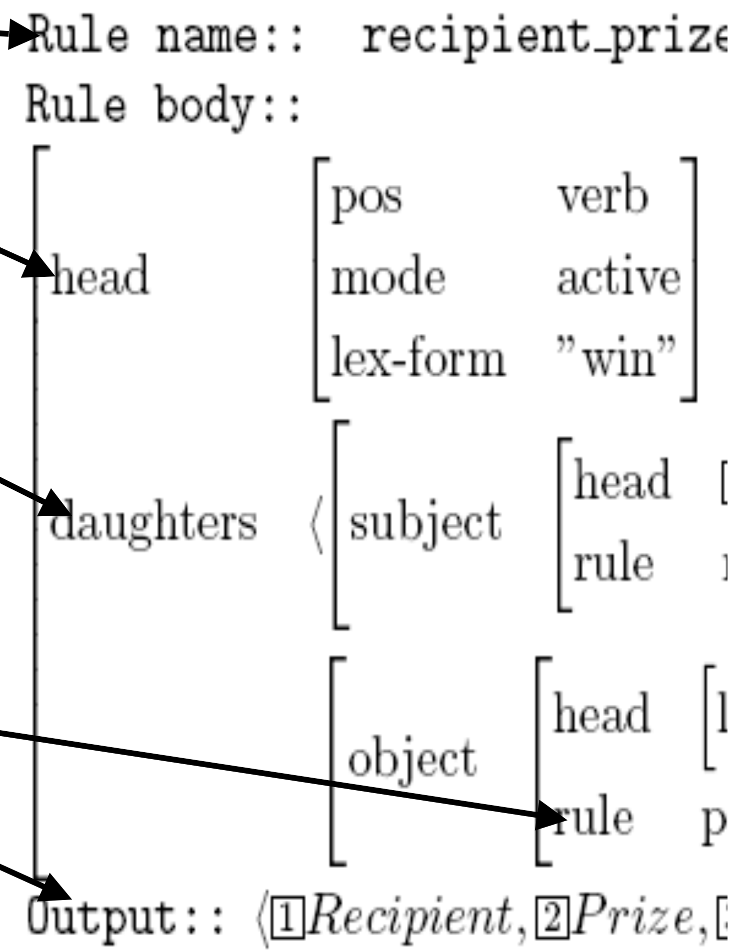
`Output::` $\langle \boxed{1}Recipient, \boxed{2}Prize, \boxed{3}Area, \boxed{4}Year\rangle$

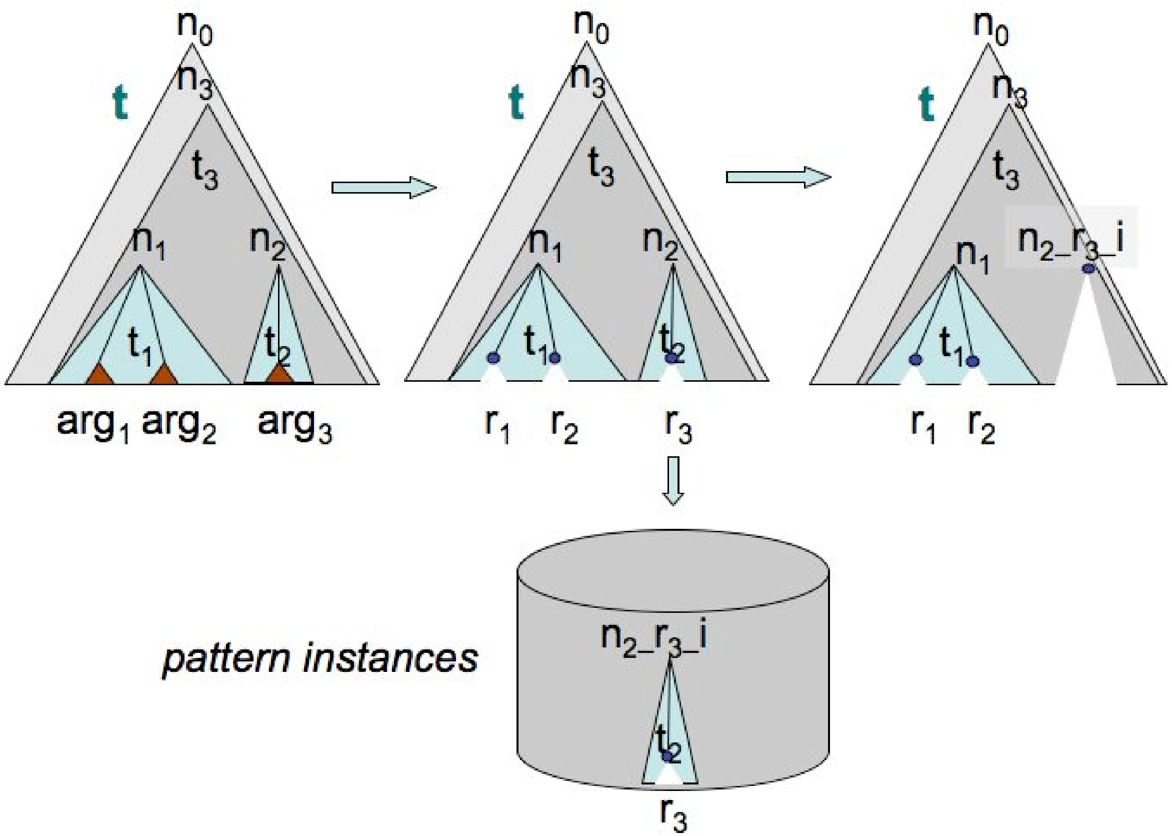
Rule Components

1. **rule name**: r_i;

2. **rule body:** in AVM format containing:
   - **head**: the linguistic annotation of the top node of the linguistic structure;
   - **daughters**: its value is a list of specific linguistic structures (e.g., subject, object, head, mod), derived from the linguistic analysis, e.g., dependency structures and the named entity information;
   - **rule**: its value is a DARE rule which extracts a subset of arguments of ***A***.

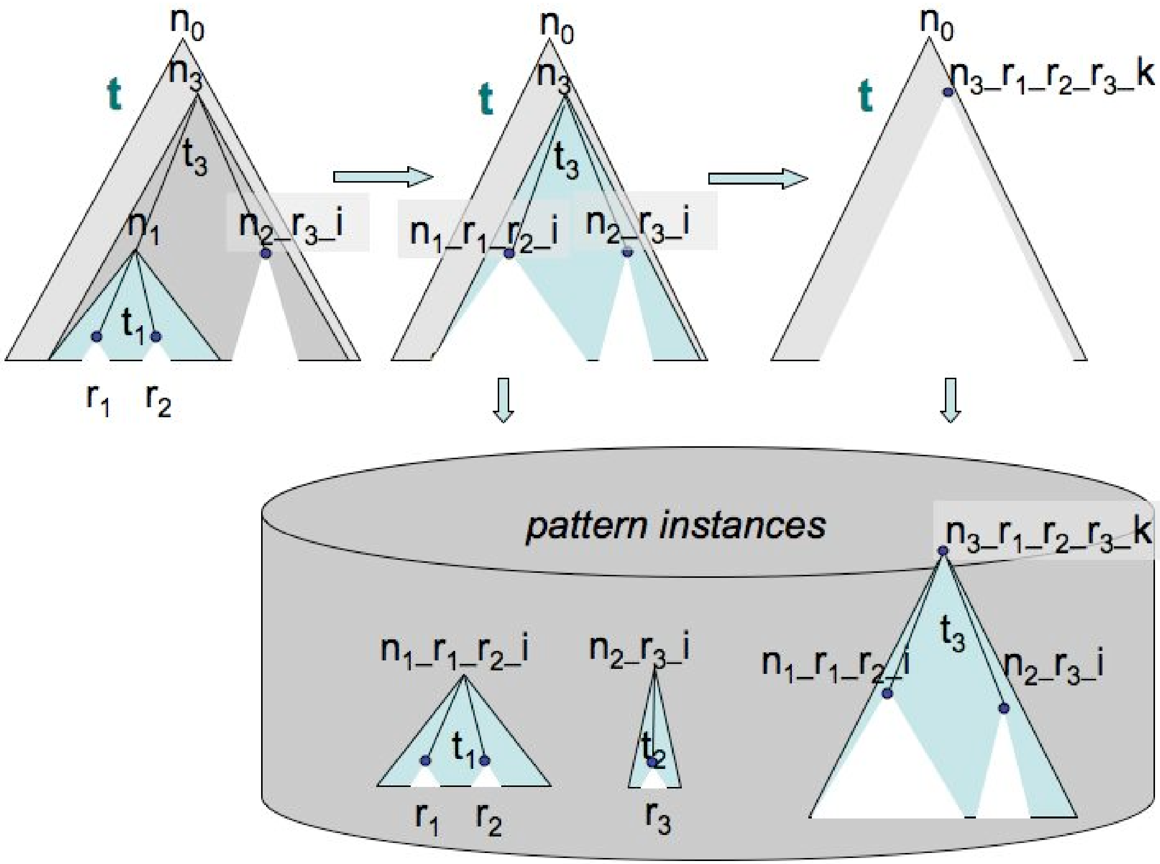
3. **output**: a set ***A*** contains the *n* arguments of the *n*-ary relation, labelled with their argument roles;





1. replace all terminal nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;
2. identify the set of the lowest nonterminal nodes N_1 in t that dominate only one argument (possibly among other nodes).
3. substitute N_1 by nodes labelled with the seed argument roles and their entity classes
4. prune the subtrees dominated by N_1 from t and add these subtrees into P. These subtrees are assigned the argument role information and a unique id.

Pattern Extraction Step 2



For i=2 to n

1. find a set of the lowest nodes N_i in t that dominate in addition to other children only i seed arguments;
2. substitute N_i by nodes labelled with the i seed argument role combination information (e.g., $r_i\_r_j$) and with a unique id.
3. prune the set of subtrees T_i dominated by N_i in t;
4. add T_i together with the argument, role combination information and the unique id to P

☆ Which kind of sentences could represent an event?

complexity	matched sentence	event sentence	Relevant sentences in %
4-ary	36	34	94.0
3-ary	110	96	87.0
2-ary	495	18	3.6

Table 1. distribution of the seed complexity

☆ Two domains

- Nobel Prize Awards: *<recipient, prize, area, year>*
- Management Succession: *<Person_In, Person_Out, Position, Organisation>*

☆ Test data sets

Data Set Name	Doc Number	Data Amount
Nobel Prize A (1999-2005)	2296	12.6 MB
Nobel Prize B (1981-1998)	1032	5.8 MB
MUC-6	199	1 MB

☆ Conditions and Problems

- Complete list of Nobel Prize award events from online portal Nobel-e-Museum
- No gold-standard evaluation corpus available

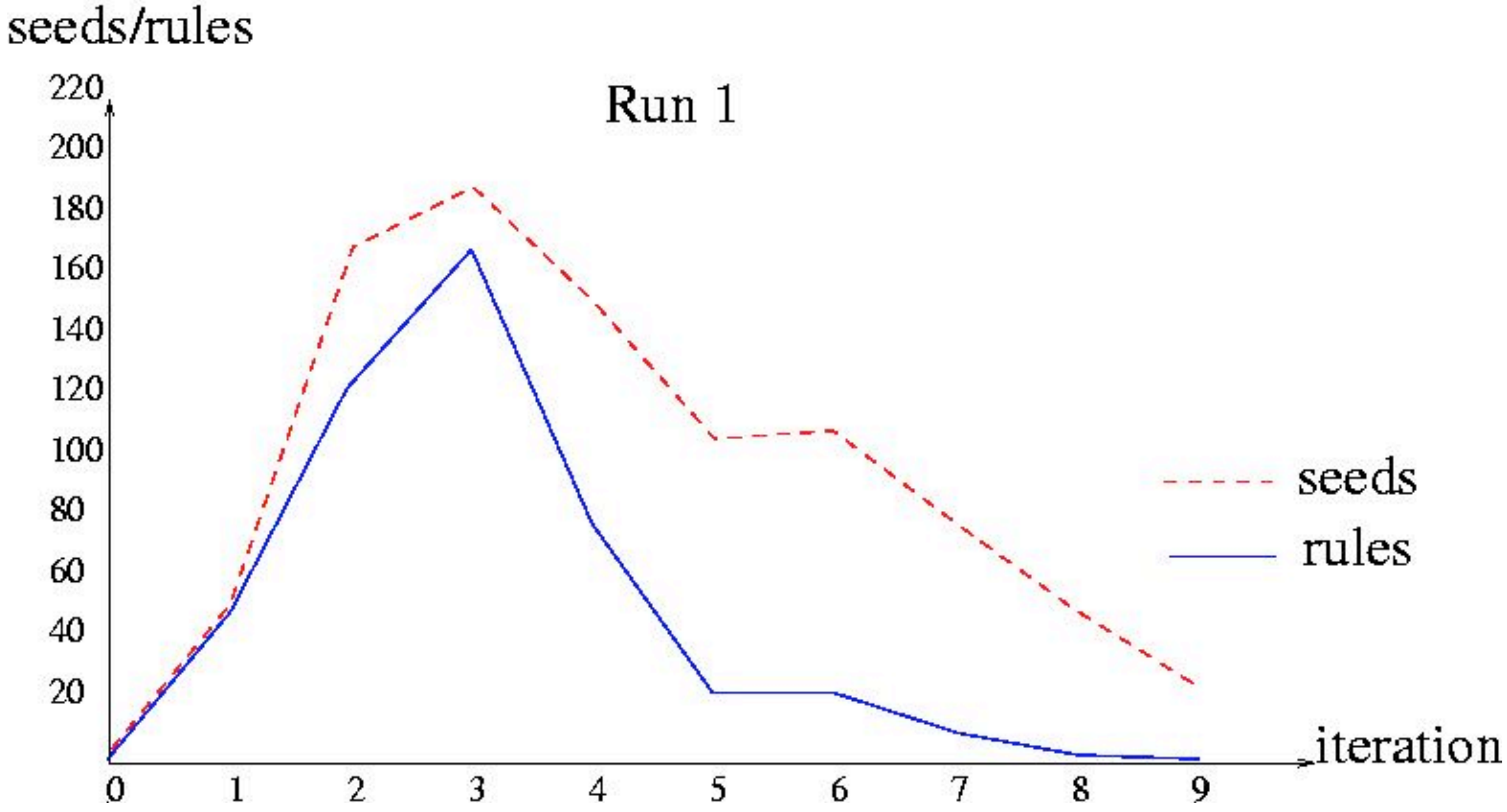
☆ Solution

- our system is successful if we capture one instance of the relation tuple or its projections, namely, one mentioning of a Nobel Prize award event. (Agichtein and Gravano, 2000)
- construction of so-called *Ideal* tables that reflect an approximation of the maximal detectable relation instances
  - The Ideal tables contain all Nobel Prize winners that co-occur with the word “Nobel” in the test corpus and integrate the additional information from the Nobel-e-Museum

Evaluation Against Ideal Tables

Data Set	Seed	Precision	Recall
Nobel Prize A	<[Zewail, Ahmed H], nobel, chemistry, 1999>	**71.6%**	**50.7%**
Nobel Prize B	<[Sen, Amartya], nobel, economics, 1998>	**87.3%**	**31.0%**
Nobel Prize B	<[Arias, Oscar], nobel, peace, 1987>	**83.8%**	**32.0%**

Iteration Behavior (Seed vs. Rule)



Initial Seed #		Precision	Recall
1	A	12.6%	7.0%
	B	15.1%	21.8%
20		48.4%	34.2%
55		62.0%	48.0%

Our result with 20 seeds (after 4 iterations)

- precision: 48.4%
- recall: 34.2%

compares well with the best result reported so far by (Greenwood and Stevenson, 2006) with the linked chain model starting with 7 hand-crafted patterns (after 190 iterations)

- precision: 43.4%
- recall: 26.5%

☆ Prize award patterns

- Detection of other Prizes such as *Pulitzer Prize*, *Turner Prize*
- Precision: 86,2%

☆ Management succession

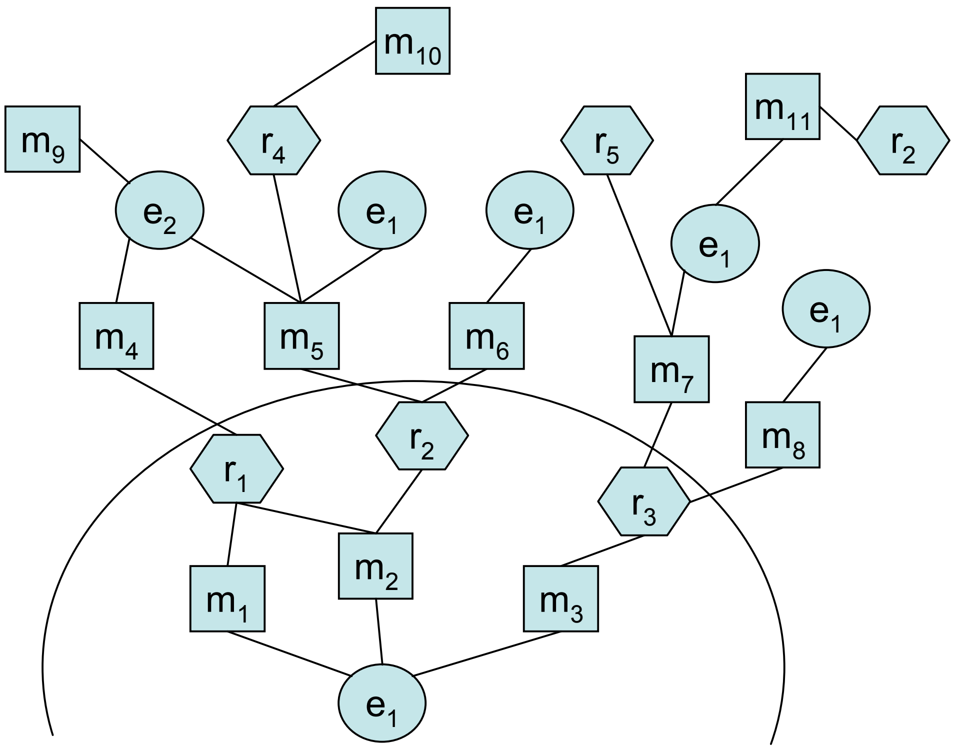
- Domain independent binary pattern rules:
  Person-Organisation, *Person-Position*
- Evaluation of top 100 relation instances
  - Precision: 98%

- ☆ Wouldn‘t it be wonderful if we could always automatically learn most or all relevant patterns of some relation from one single semantic instance!
- ☆ Or at least find all event instances. (IDEAL Tables or Completeness)
- ☆ This sounds too good to be true!

☆ As scientists we want to know:

- Why does it work for some tasks?
- Why doesn‘t it work for all tasks?
- How can we estimate the suitability of domains?
- How can we deal with less suitable domains?

Start of Bootstrapping (simplified)



Abstraction

bipartite graph

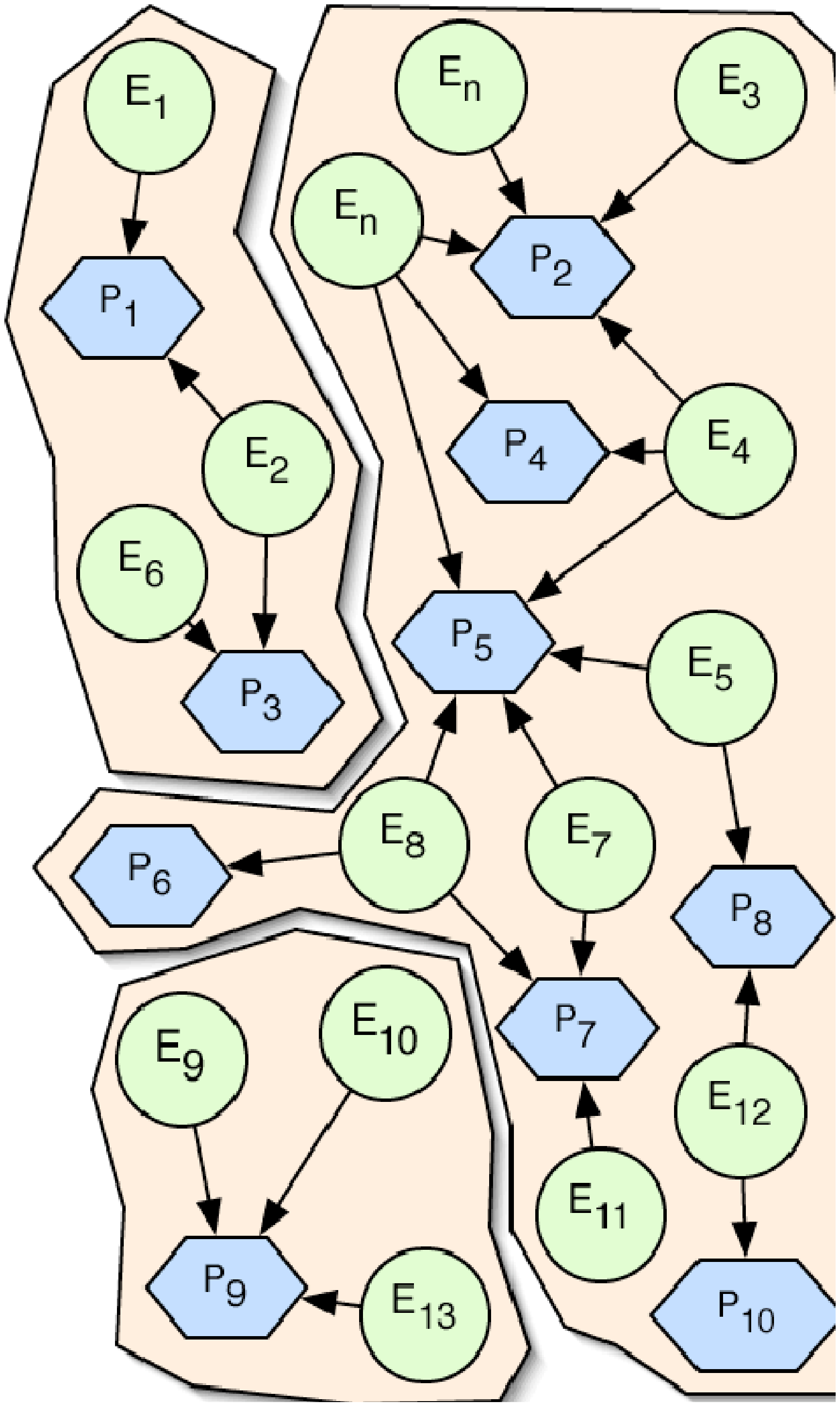
two types of vertices

E_i = event instance

P_j = linguistic pattern

relevant properties:

- ☆ two degree distributions
- ☆ connectedness
- ☆ average and maximum path lengths between events



Can we reach all events in the graph?

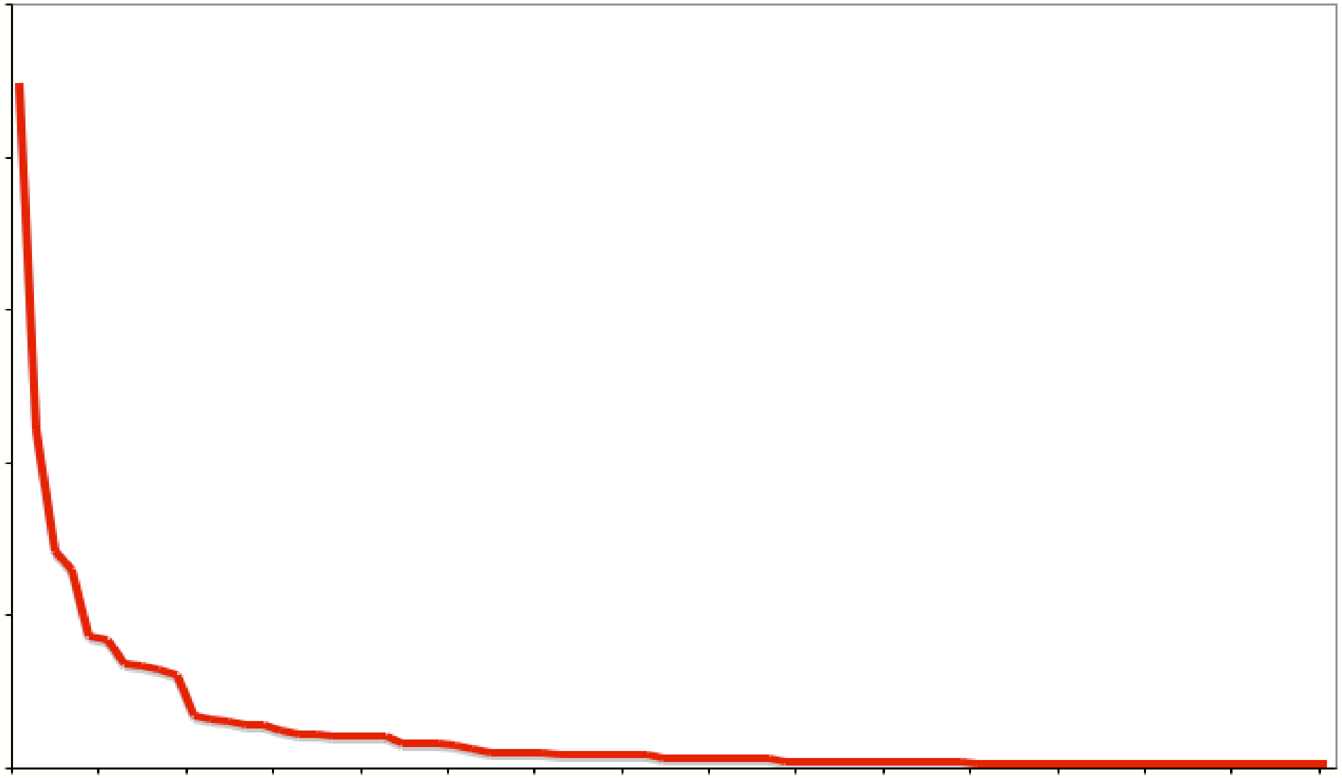
By how many steps?

From any event instance?

1. Distributions of Pattern in Texts

2. Distribution of Mentionings to Relation Instances

General distribution of patterns in texts probably follows Church's Conjecture: Zipf distribution (a heavy-tailed skewed distribution)



Distribution of Mentionings to Events

- ☆ Distribution of mentionings to relation instances (events) differs from one task to the other.
- ☆ The distribution reflects the redundancy in textual coverage of events.
- ☆ Distribution depends on text selection, e.g. number of sources (newspapers, authors, time period)

  example 1: several periodicals report on Nobel Prize events

  example 2: one periodical reports on management succession events

☆ If both degree distributions are skewed, long-tailed, we may get so called scale-free networks.

$$p(k) = \sum_{v \in V \mid \deg(v)=k} 1$$

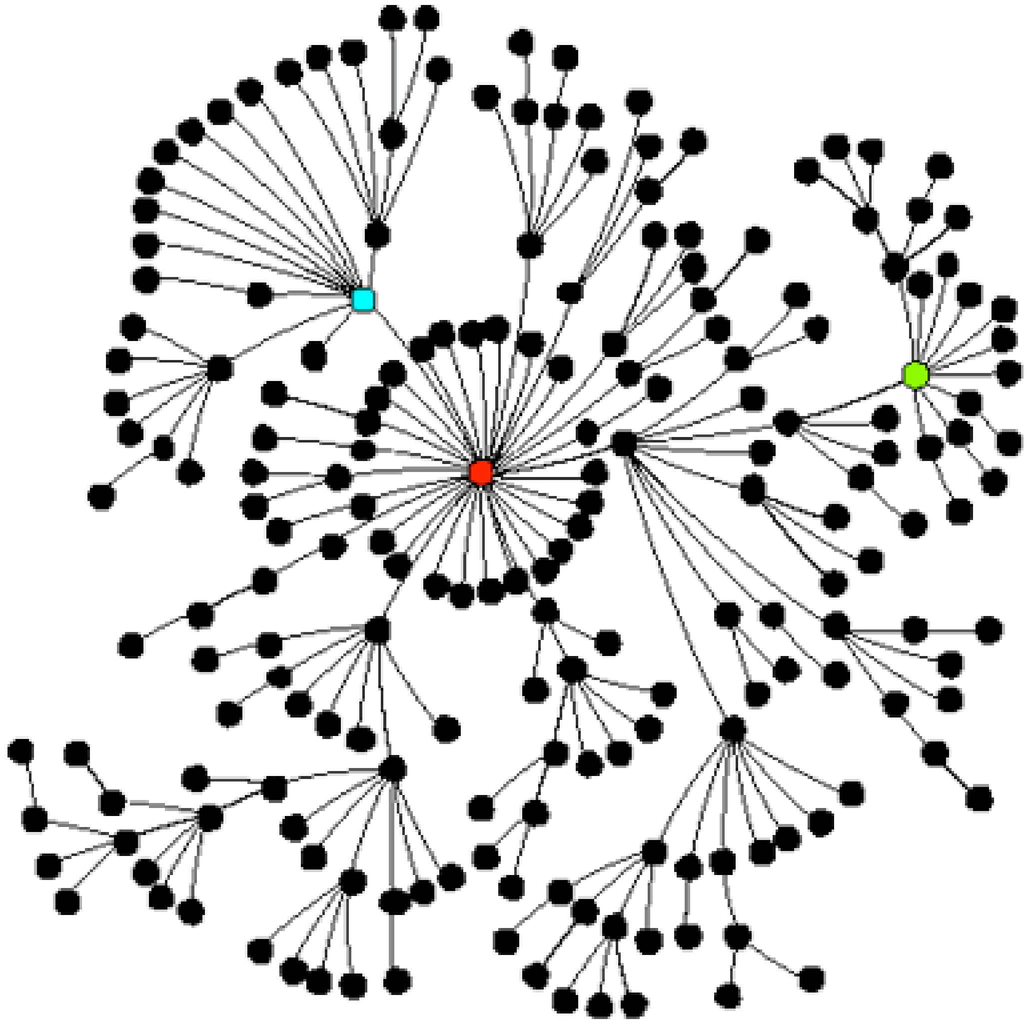
$$\mathbf{P(k) \sim k^{-\gamma}}$$

☆ with the nice properties that the usual distance between any two vertices is very short

Example of Scale-Free Nets

In scale-free networks, some nodes act as "highly connected hubs" (high degree), although most nodes are of low degree.

Scale-free networks' structure and dynamics are independent of the system's size N, the number of nodes the system has. In other words, a network that is scale-free will have the same properties no matter what the number of its nodes is.



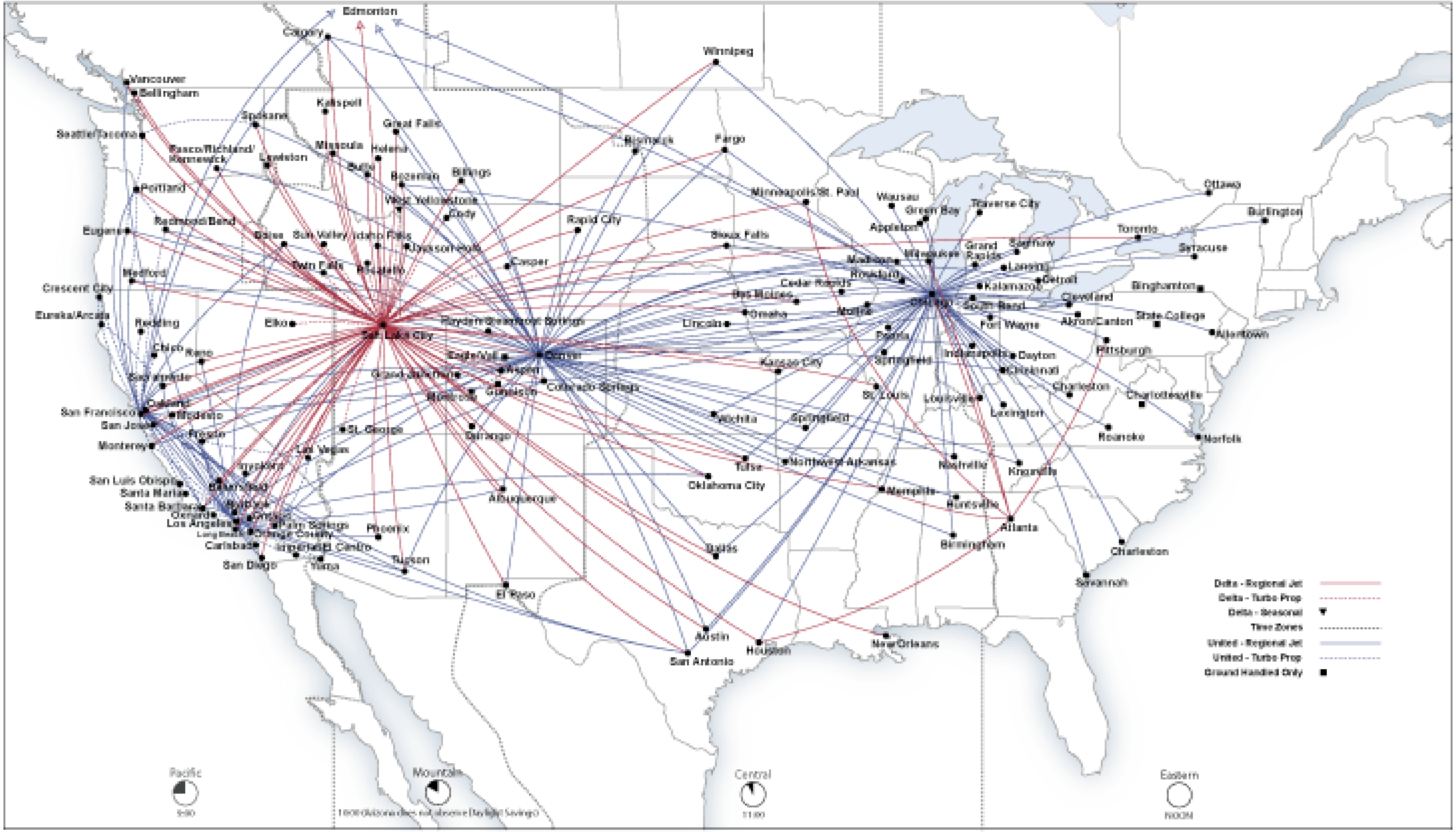
Networks exhibiting the small-world property

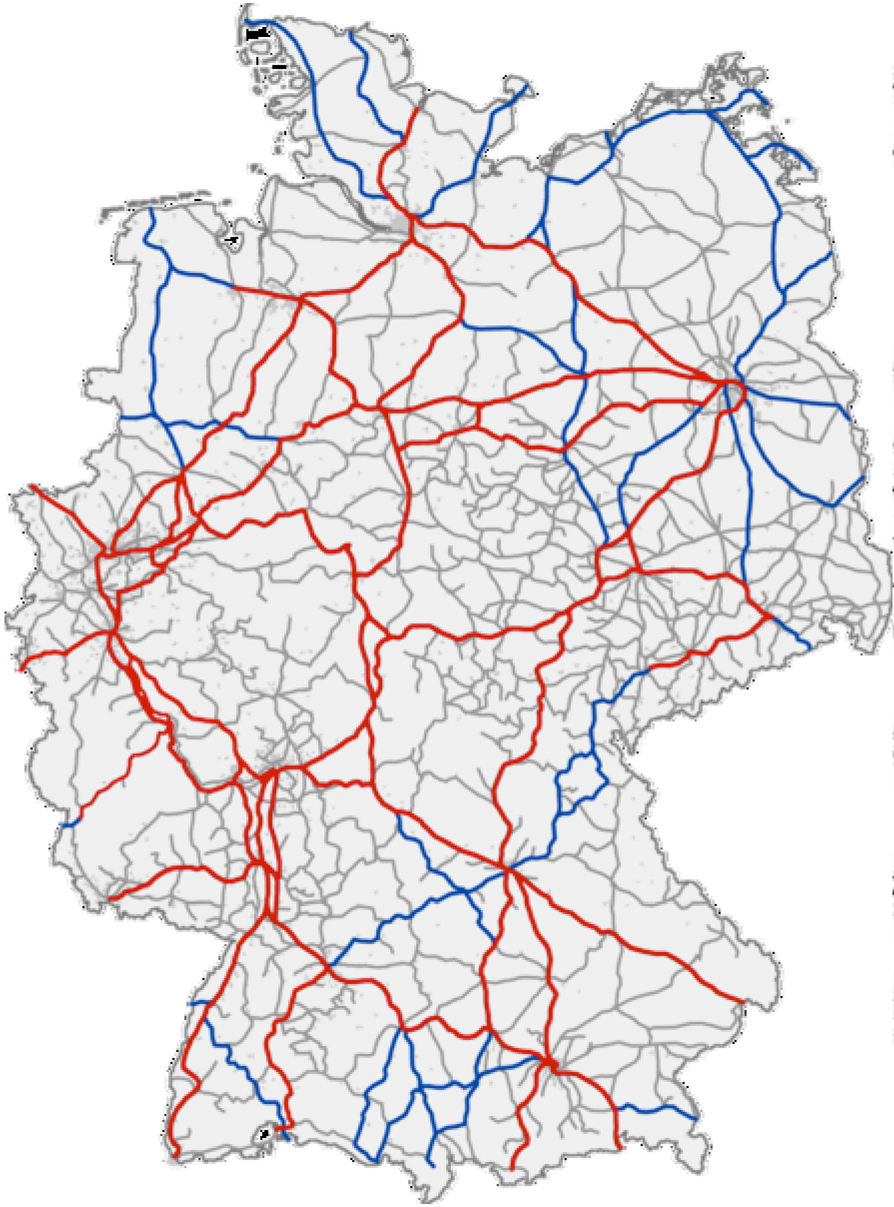
- social networks (max path-length 5-7)
- co-authorship networks (Erdös number)
- Internet
- WWW
- air traffic route maps (max. 3 hops)

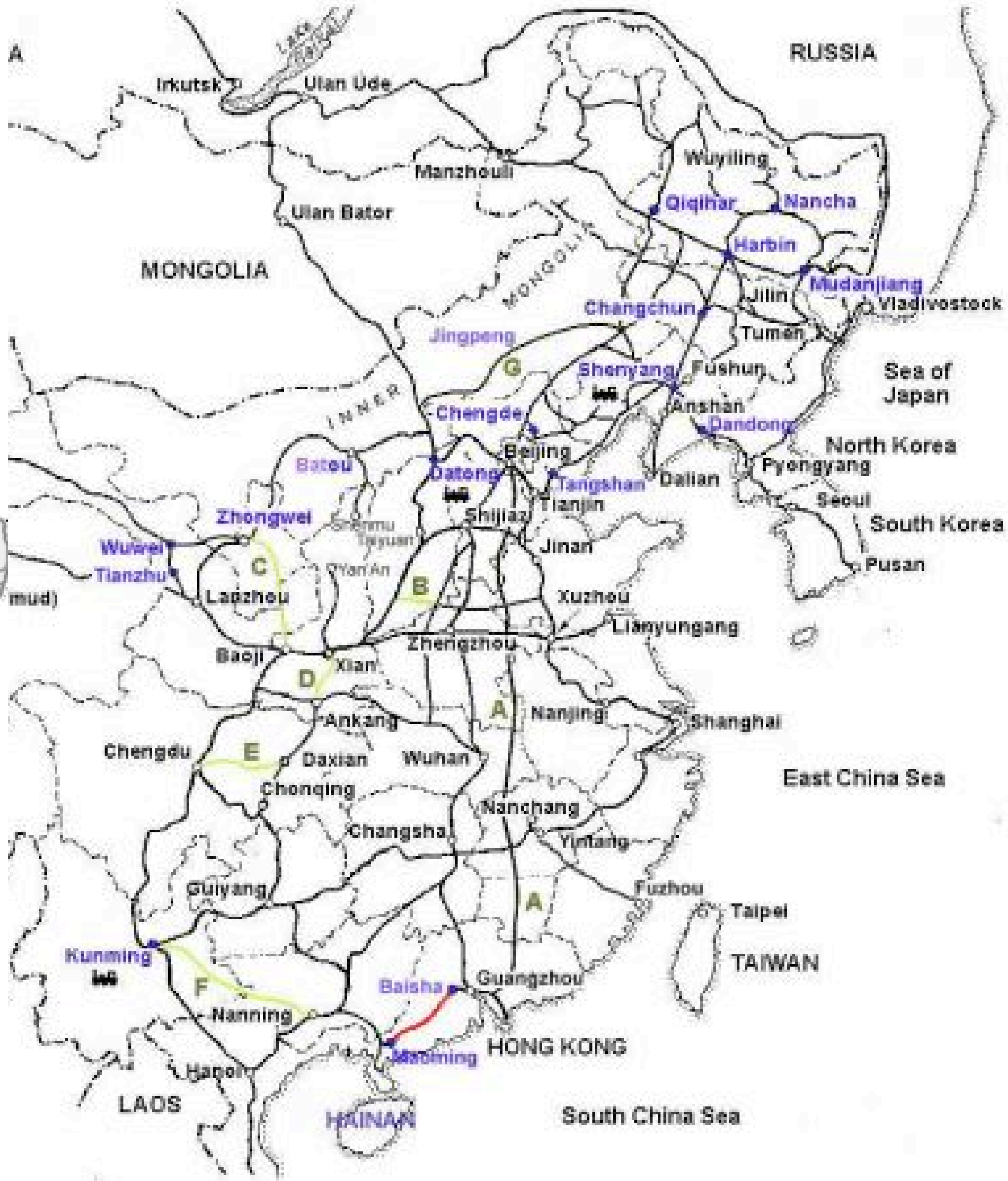
Networks that do not exhibit the small-world property

- road networks
- railway networks
- kinship networks

Airline Route Networks





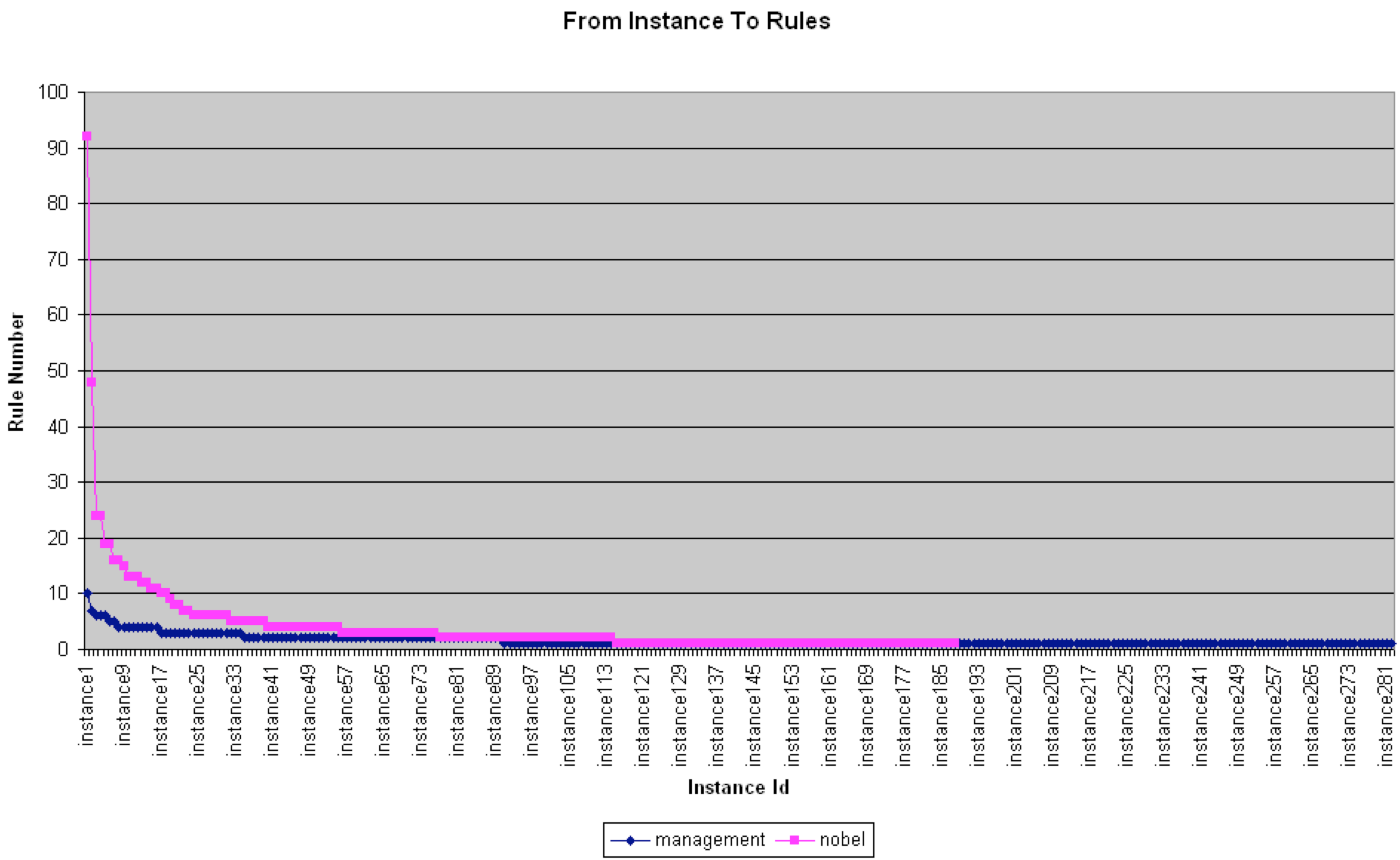

RUSSIA
Irkutsk
Ulan Ude
Manzhouli
Ulan Bator
MONGOLIA
INNER MONGOLIA
Wuyiling
Qiqihar
Nancha
Harbin
Mudanjiang
Jilin
Vladivostock
Changchun
Tumen
Jingpeng
Shenyang
Fushun
Sea of Japan
Anshan
Chengde
Dandong
North Korea
Pyongyang
Seoul
Beijing
Dalian
Baotou
Datong
Tangshan
Tianjin
South Korea
Zhongwei
Shijiaz
Jinan
Pusan
Wuwei
Tianzhu
Lanzhou
Xuzhou
Lianyungang
Zhengzhou
Baoji
Xian
Nanjing
Shanghai
Ankang
Chengdu
Daxian
Wuhan
East China Sea
Chonqing
Nanchang
Changsha
Yintang
Guiyang
Fuzhou
Taipei
TAIWAN
Kunming
Baisha
Guangzhou
Nanning
HONG KONG
Maoming
Hanoi
LAOS
HAINAN
South China Sea
A
B
C
D
E
F
G

Small Worlds for Bootstrapping

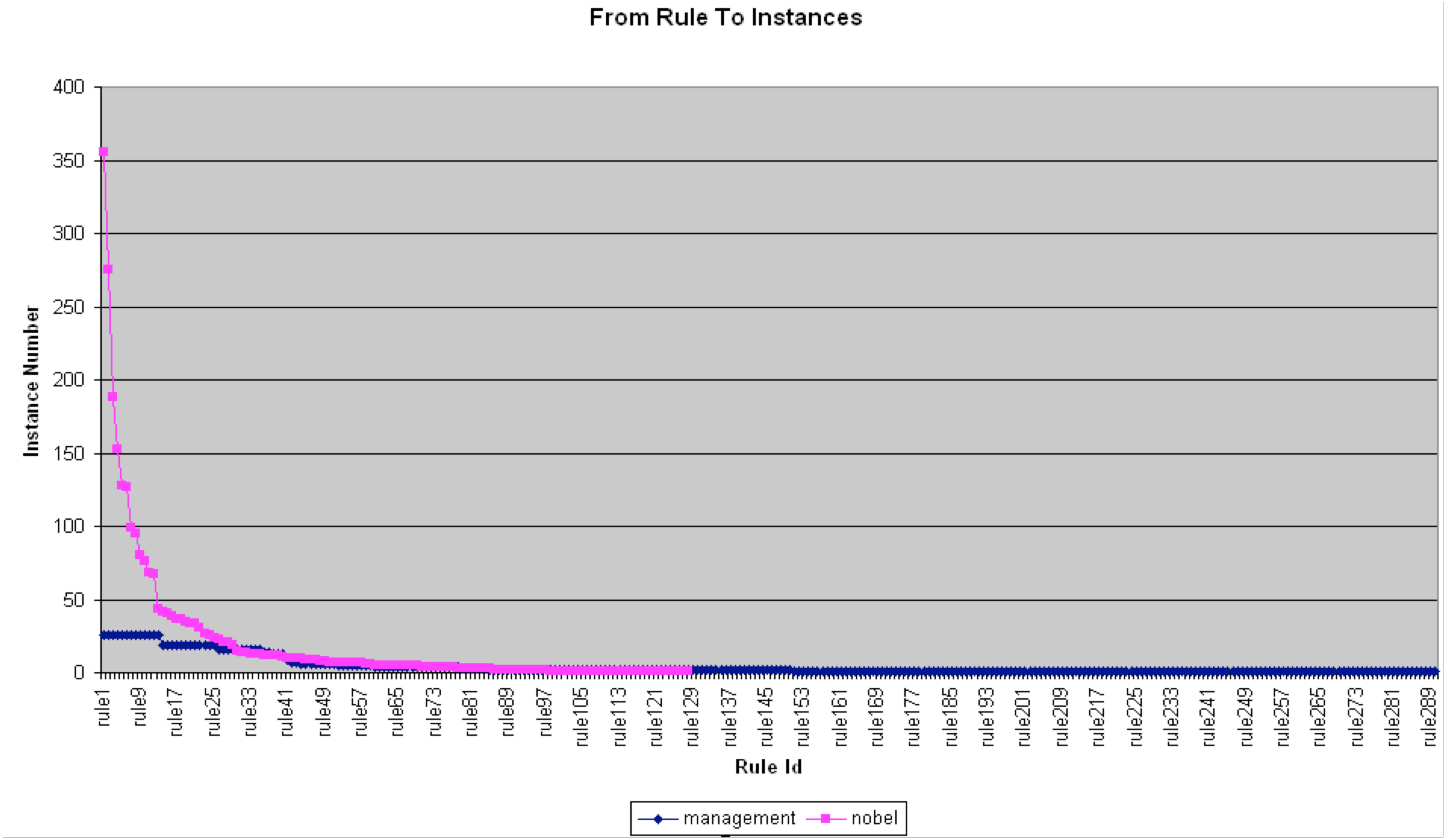
☆ If both distributions follow a skewed distribution and if the distributions are independent from each other, then we get a scale-free network in the broader sense of the term.

☆ For each type of vertices we get strong hubs. This leads to very short paths (for most connections).

Instance to Pattern



Rules to Instances



If we find a large world with continents and islands...

- ☆ What can we do if our domain or our data do not fit the nicely connected small world picture?

- ☆ Then we can give up and search for another domain...

- ☆ ...or we can try to change the data or relations in order to get benevolent learning graphs.

☆ Enlarging the domain

Pulitzer Prize —> all Prizes

☆ selecting Carrier Domains (parallel learning domains)

Pulitzer Prize —> Nobel Prize
Ernst Winter Preis —> Nobel Prize
Fritz Winter Preis —> Nobel Prize

Other Discovered Award Events

Academy Award
actor % (Cannes Film Festival's Best Actor award)
American Library Association Caldecott Award
American Society
award
Blitzker
Emmy
feature % (feature photography award)
first % (the first Caldecott Medal)
Francesca Primus Prize
gold % (gold medal)
Livingston Award
National Book Award
Newbery Medal
Oscar
P.G.A

PEN/Faulkner Award
prize
reporting % (the investigative reporting award)
Tony
Tony Award
U.S. Open

But also:
nomination
$1 million
$29,000
about $226,000
praise
acclaim
discovery
doctorate
election

Further Approaches

☆ enlarging the text base for finding seeds and patterns

- New York Times MUC data –> general press corpora
- New York Times MUC data –> WWW

☆ enlarging the text base for finding new seeds

- New York Times MUC data –> WWW
- German Press Data –> English Press Data

Next Steps

- ☆ Go beyond the sentence.
- ☆ Investigate properties of relations w.r.t. data.
- ☆ Try to describe them as graph properties.
- ☆ Try out auxiliary data sets (such as the Web).
- ☆ Extend to deep processing: extract patterns from RMRS with extended ERG (first tests by Zhang Yi 80% coverage for Nobel prize sentences, 61% for management succession)

☆ We tried to apply the method to another domain Pop Artist Gossip in an EU funded project RASCALLI

☆ First experiment: use the learned patterns for detecting other prize winning events such as Grammy and Music Awards

- So far, we were missing patterns in which some role fillers are expressed as pronouns or as noun phrase anaphora
  - *He won the prize for his earth-shaking discoveries in genetic sequencing.*
  - *In the same year, the two biologists received the Nobel Prize in Medicine.*
- We included sentences in which potential role fillers occured some sentences before or after the pattern.
- We then used the domain ontology to determine whether the anaphorical phrase constituted a semantically suitable candidate for the relation and the coreference.

Improving Precision

- ☆ more than 40% or our errors can be attributed to the shortcomings of MINIPAR, the robust dependency parser

- ☆ Tempting alternative: Use a more precise deep parser.

- ☆ However, we do not want to give up the robustness.

- ☆ inference is computationally intractable
- ☆ inference is too inefficient for practical use
- ☆ too much reliance on human knowledge engineering
  - – > specifications of practical scale cannot be achieved by human engineering
  - – > even experts cannot achieve formal specifications of knowledge domains by introspection that are correct, complete, consistent

All ontologies leak

Problems with Formal Grammars

- ☆ deep parsing is computationally intractable
- ☆ deep parsing is too inefficient for practical use
- ☆ too much reliance on human knowledge engineering
  - – > specifications of practical scale cannot be achieved by human engineering
  - – > even experts cannot achieve formal specifications of languages by introspection that are correct, complete, consistent

All grammars leak (Edward Sapir, 1921)

Sour Grapes (Fable by Aesop 620 - 560 BC)

- ☆ A very hungry fox walked into a vineyard where there was an ample supply of luscious looking grapes.
- ☆ However, the grapes hung higher than the fox could reach.
- ☆ He jumped and stretched and reached and jumped some more trying to get those yummy grapes, but to no avail.
- ☆ "Those grapes surely must be sour," he finally said "I wouldn't eat them if they were served to me on a silver platter."
- ☆ Moral of the story: It is easy to hate what you cannot have. (Denial of Desire)

A Contradiction

? Why is it good to drop deep processing in favour of shallow approaches?

! Because in this way we can build useful applications today.

? Why is it good to continue with deep processing?

! Because this is the ultimate goal of computational linguistics and the key to many additional applications.

What has changed for Knowledge Processing?

☆ tractable subsets of first order logic

- The German School of Description Logics (1988-98)
- Detailed catalogue of complexity of family

☆ much more efficient inferencing technologies

- Complete decidable algorithms using tableaux methods (1991-1992)

☆ more intuitive notations and editing tools

☆ better knowledge engineering methods

☆ a stronger, application-driven demand

What has changed for Deep Processing?

☆ computationally more benign grammar formalisms

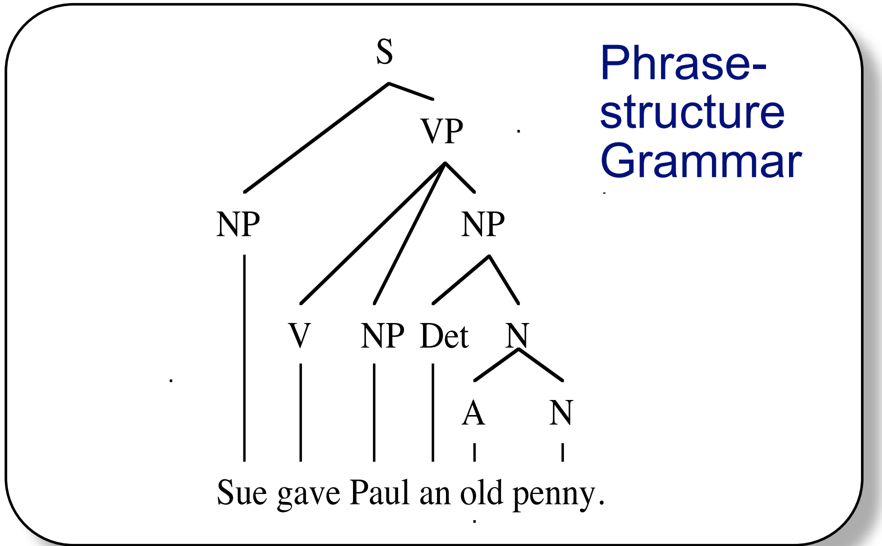
☆ much more efficient parsing technologies

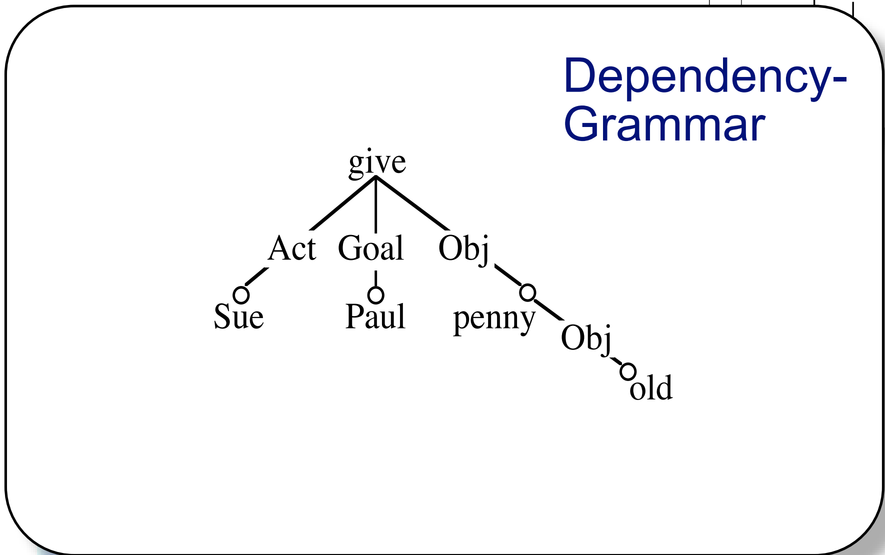
☆ more intuitive notations and editing tools

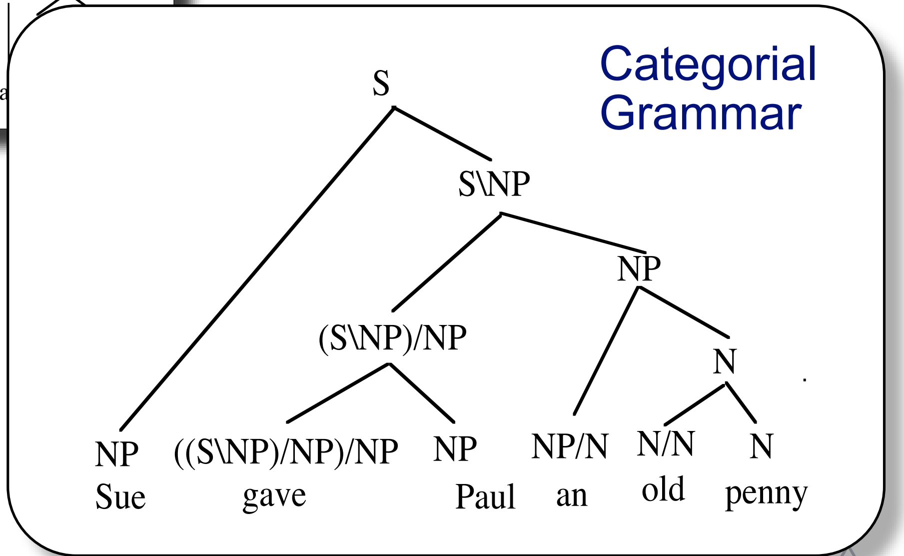
☆ better grammar engineering methods

☆ a stronger, application-driven demand

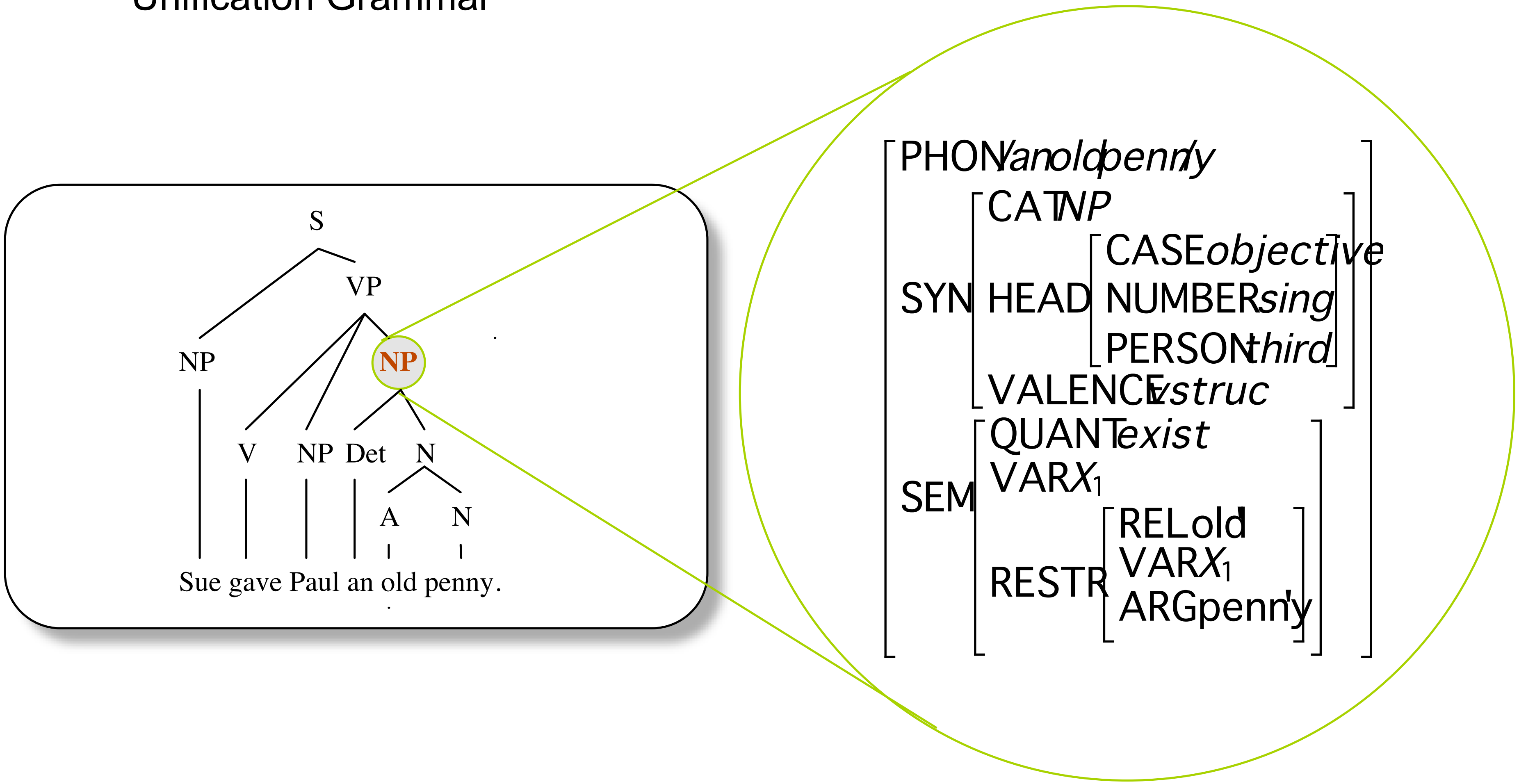
THREE TRADITIONS







Unification Grammar



- ☆ Nearly all shallow processing systems including statistical processing are restricted to one or several applications.

- ☆ They do not model the human language faculty.

- ☆ Humans can acquire new types of linguistic performance (applications) exploiting the same basic linguistic knowledge:
  - translating
  - summarizing
  - writing
  - etc.
  - skimming
  - question answering
  - singing

The Dream is Living On

- ☆ The dream of reusable linguistic knowledge has not been given up
- ☆ Even if the great majority of papers at large conferences is dedicated to shallow (including statistical and table-lookup systems), some researchers are still trying to solve the much harder problem
- ☆ However, the problem is so complex that it takes large efforts and international cooperation to achieve progress
- ☆ The LFG group at PARC is running an international cooperation called PARGRAM
- ☆ Parts of the HPSG community have formed a large international cooperation to boost deep processing

The DELPH-IN Initiative

- ☆ Cambridge University (UK), Computer Laboratory
- ☆ DFKI Saarbrücken (Germany), Language Technology Lab (co-founder)
- ☆ LORIA Nancy (France)
- ☆ NTT Communication Science Laboratory (Japan), Machine Translation Research Group
- ☆ Norwegian University of Science and Technology (Norway), Lingvistisk Institutt
- ☆ Saarland University (Germany), Department for Computational Linguistics
- ☆ Stanford University (US), LinGO Laboratory at CSLI (co-founder)
- ☆ Tokyo University (Japan), Tsujii Laboratory
- ☆ University Linköping (Sweden)
- ☆ University of Lisbon (Portugal)
- ☆ University of Oslo (Norway), MT Research Group
- ☆ University Pompeu Fabra Barcelona (Spain)
- ☆ University of Seoul (Korea)
- ☆ University of Sussex (UK), School of Cognitive and Computing Sciences
- ☆ University of Washington (US), Computational Linguistics Laboratory

- ☆ Head-Driven Phrase Structure Grammar by Pollard and Sag
- ☆ Uniform formalism: typed feature structures
- ☆ High degree of lexicalization: very few PS-rules, rich lexicon structure
- ☆ Ontological structure: Multiple inheritance type hierarchy

Stanford

- **HPSG Group at CSLI**
- Sag, Flickinger, Copestake, Malouf, Carroll (now Sussex),...

Saarbrücken

- **LT Lab at DFKI and Dept. of CL**
- Oepen, Callmeier, Krieger, Kiefer, Ciortuz, Müller,...

Tokyo

- **Tsujii Lab at the University of Tokyo**
- Tsujii, Torisawa, Ninomiya, Taura, Yoshida, Mitsoishi,...

Efficiency Problem Solved

☆ By combining methods from participants:

☆ Drastic efficiency boost -- by a factor of 2000 and more

Still not Robust...

☆ So we are still left with the problem of insufficient coverage.

☆ Two remedies:

- combining the accurate deep processing with robuts shallow processing
- improving coverage through data-intensive learning methods

- ☆ We combined the deep processing with a number of shallow processing systems, among them simple finite state parsers for Named Entity Recognition.

- ☆ The architecture for combining the processing components is called Heart of Gold (HoG) 2007 Ph.D. Thesis by Ulrich Schäfer

- ☆ All components use multilevel stand-off annotation.

- ☆ The language for annotation and thus the interface language of the hybrid architecture is Minimal Recursion Semantics (MRS).

☆ Zhang Yi - 2007 PhD Thesis

☆ Extending the coverage by about 20% through data induced type prediction (learning of lexical types from corpora).

Conclusion and Outlook

- ☆ Yes, there has been progress. Even if it has been slow.
- ☆ Doing things right and not giving up the more demanding pricipled ways is paing off.
- ☆ Both in the area of knowledge technologies and linguistic technologies, the high hanging grapes actually are quite sweet.
- ☆ But the harvest season has not even started yet.

Thank you for your attention!