Relation Extraction and Machine Learning for IE

Feiyu Xu

feiyu@dfki.de

Language Technology-Lab
DFKI, Saarbrücken

Outline

- Introduction to IE and relation extraction
- Brief history of IE
- Machine learning methods for relation extraction
  - DARE system (http://dare.dfki.de)
- Task-driven anaphora resolution for relation extraction
- References

Vision of Semantic Web

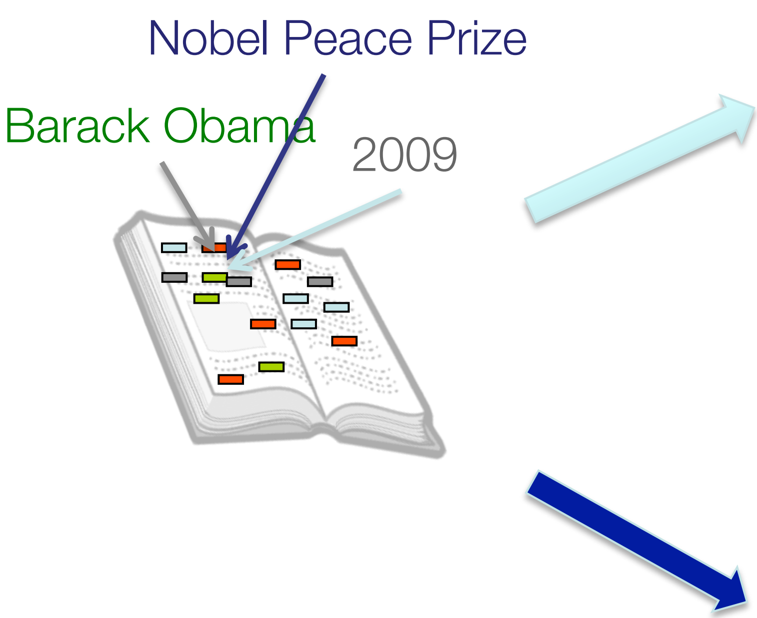
- Tim Berners-Lee defined Semantic Web as

"a web of data that can be processed directly and indirectly by machines."

What is Relation Extraction

☆ Definition

– Given an unstructured text, relation extraction tool should be able to automatically recognize and extract relations among the relevant entities or concpets that are salient to the user's needs



- Walter Kohn, Nobel, Chemistry, 1998
- J.M. Coetzee, Nobel, Literature, 2003

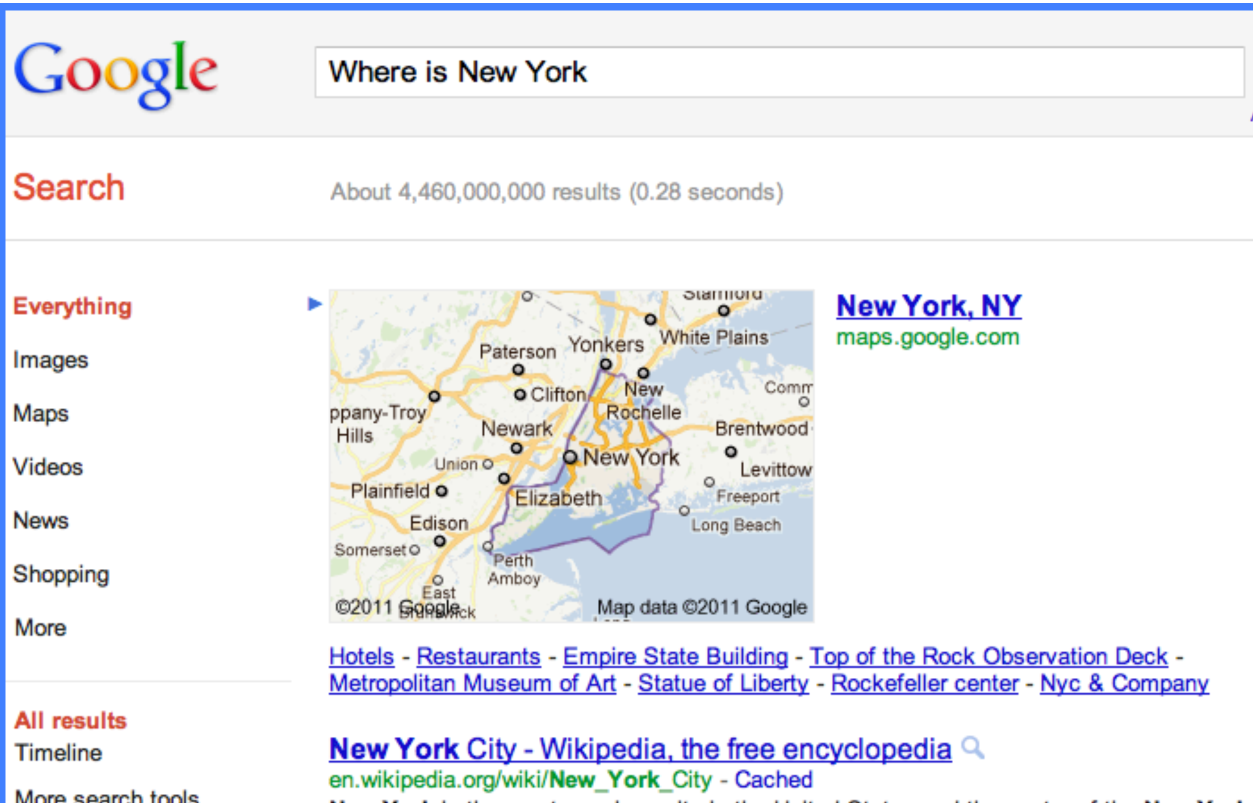
Patterns:

- *<prize>* be awarded to *<person>*
- *<person>* win *<prize>* in *<year>*
- ……

General application task 1:

☆ Information access for information finder

mapping unstructured textual queries of users to more structured formal query for search and answer engines



General application task 2:

☆ Information acquisition for information provider

extract structured information from big amount free texts to construct knowledge database

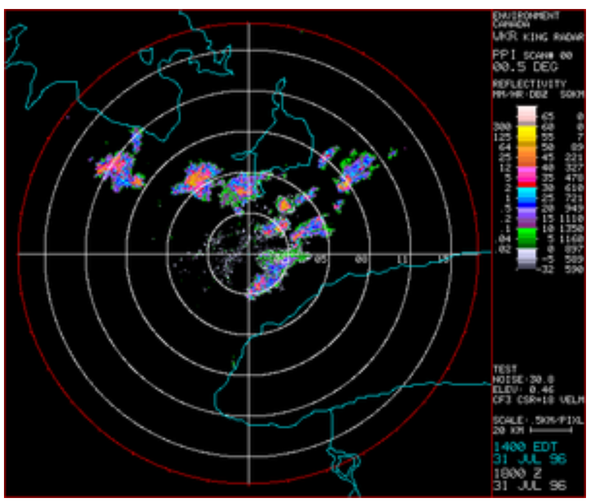
Relation Extraction

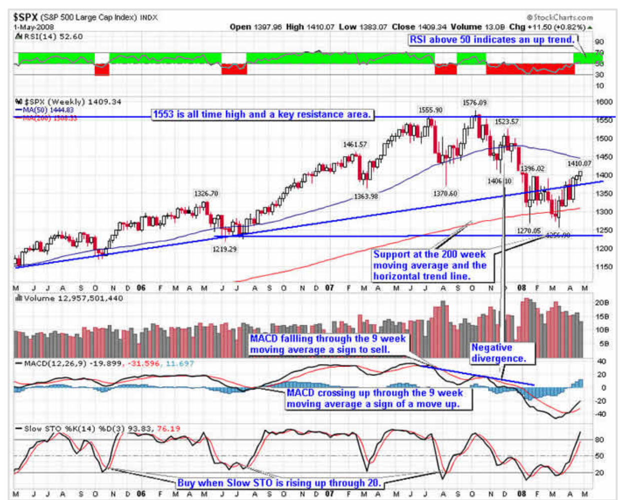
Special Case: Extensive Analyses

☆ Large-scale information monitoring

☆ Analytics: Analyses of areas, markets, trends

☆ Watch: Scanning for relevant new developments





Relation in IE

Information Extraction is …

a technology that is futuristic from the user's point of view in the current information-driven world.

Rather than indicating which documents need to be read by a user, it extracts pieces of information that are salient to the user's needs …

provided by NIST:
[http://www-nlpir.nist.gov/related_projects/muc/]

Information Extraction: A Pragmatic Approach

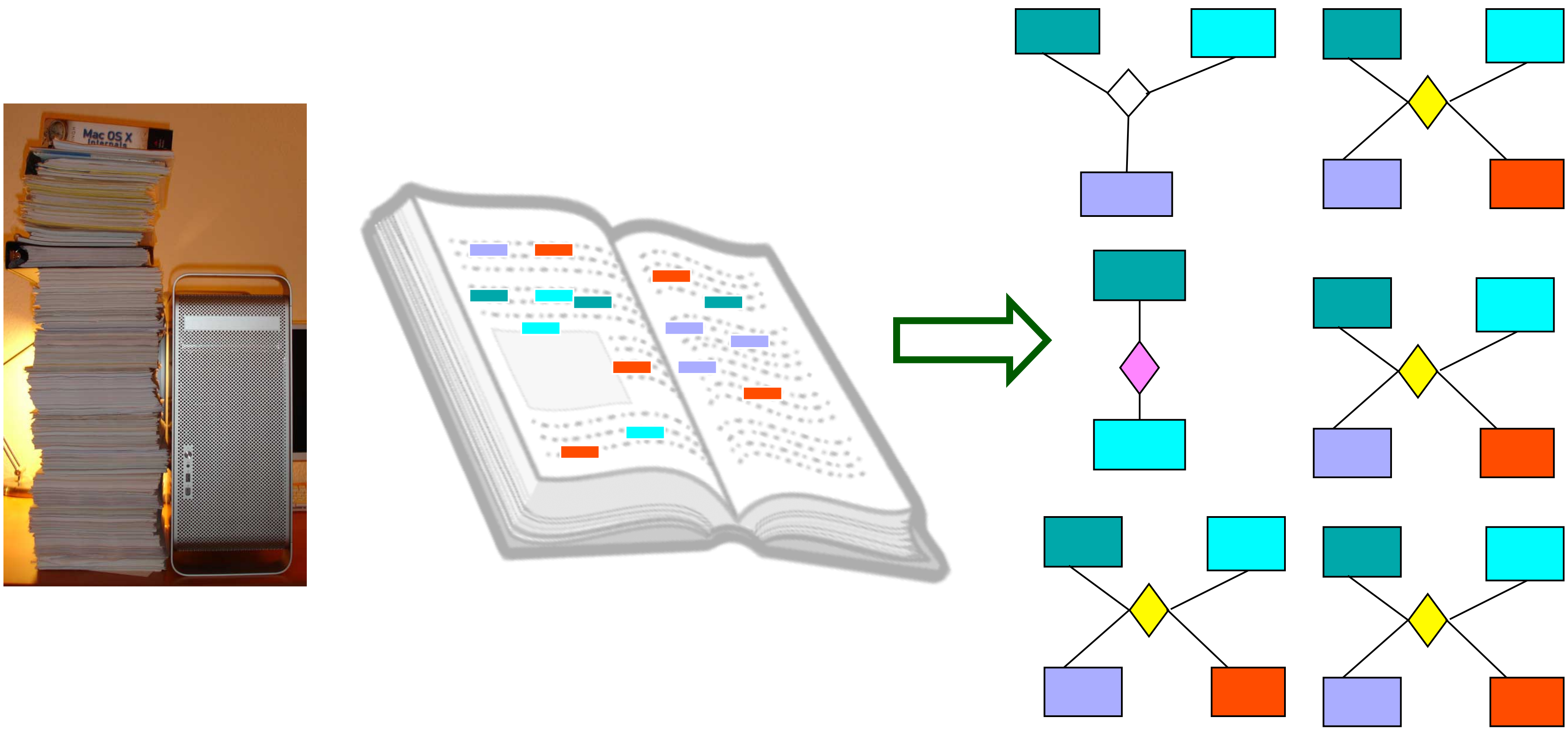
- Identify the types of entities that are relevant to a particular task
- Identify the range of facts that one is interested in for those entities
- Ignore everything else

[Appelt, 2003]

Types of Information Extraction in LT

- Topic Extraction
- Term Extraction
- Named Entity Extraction
- Binary Relation Extraction
- N-ary Relation Extraction
- Event Extraction
- Answer Extraction
- Opinion Extraction
- Sentiment Extraction

Relation Extraction is
the cover term for those IE tasks in which instances of semantic relations are detected in natural language texts.



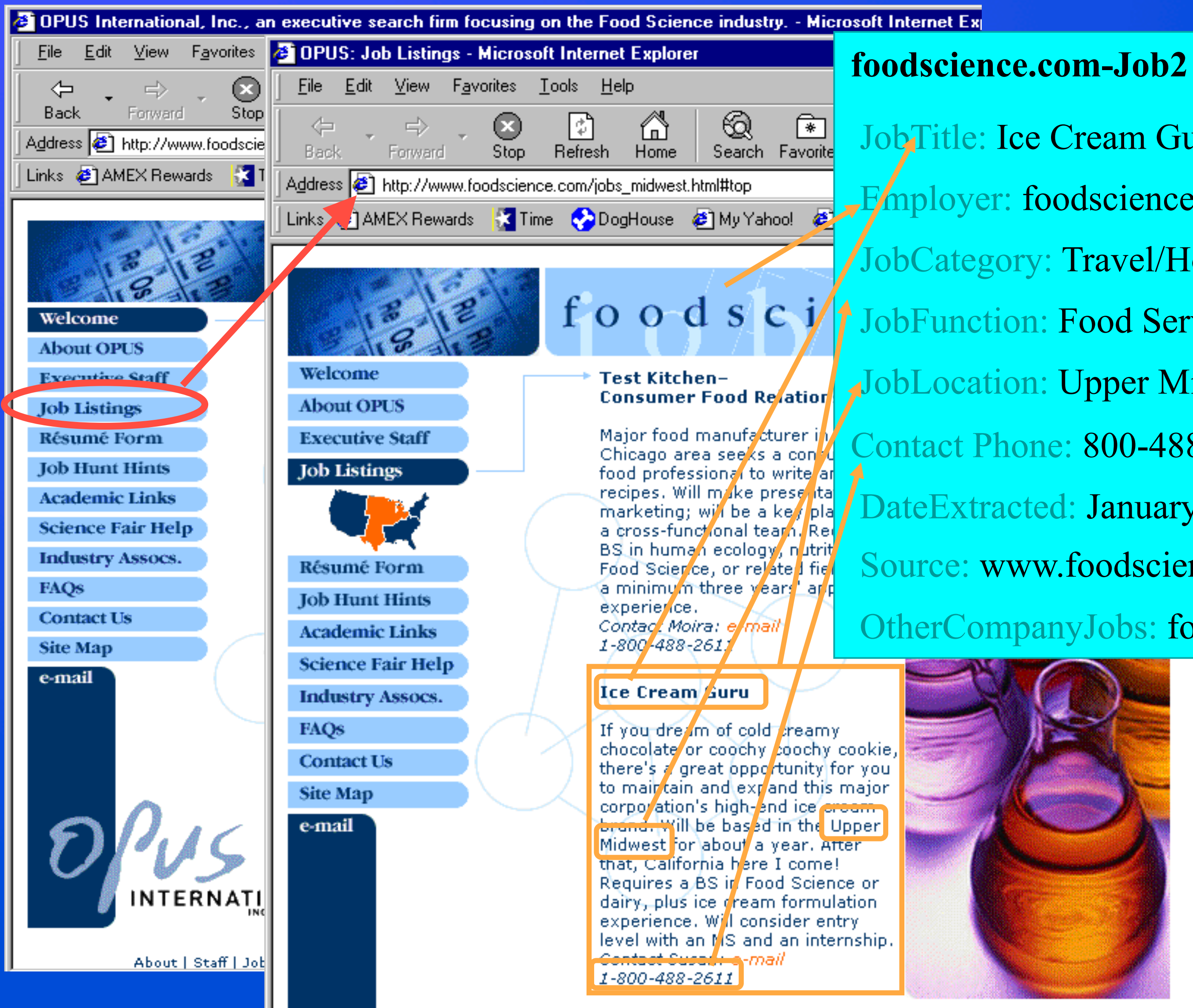
Types of Information Extraction in LT

- Topic Extraction
- Term Extraction
- Named Entity Extraction
- **Binary Relation Extraction**
- **N-ary Relation Extraction**
- **Event Extraction**
- **Answer Extraction**
- **Opinion Extraction**
- **Sentiment Extraction**

Types of Relation Extraction

Extracting Job Openings from the Web:

Semi-Structured Data



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

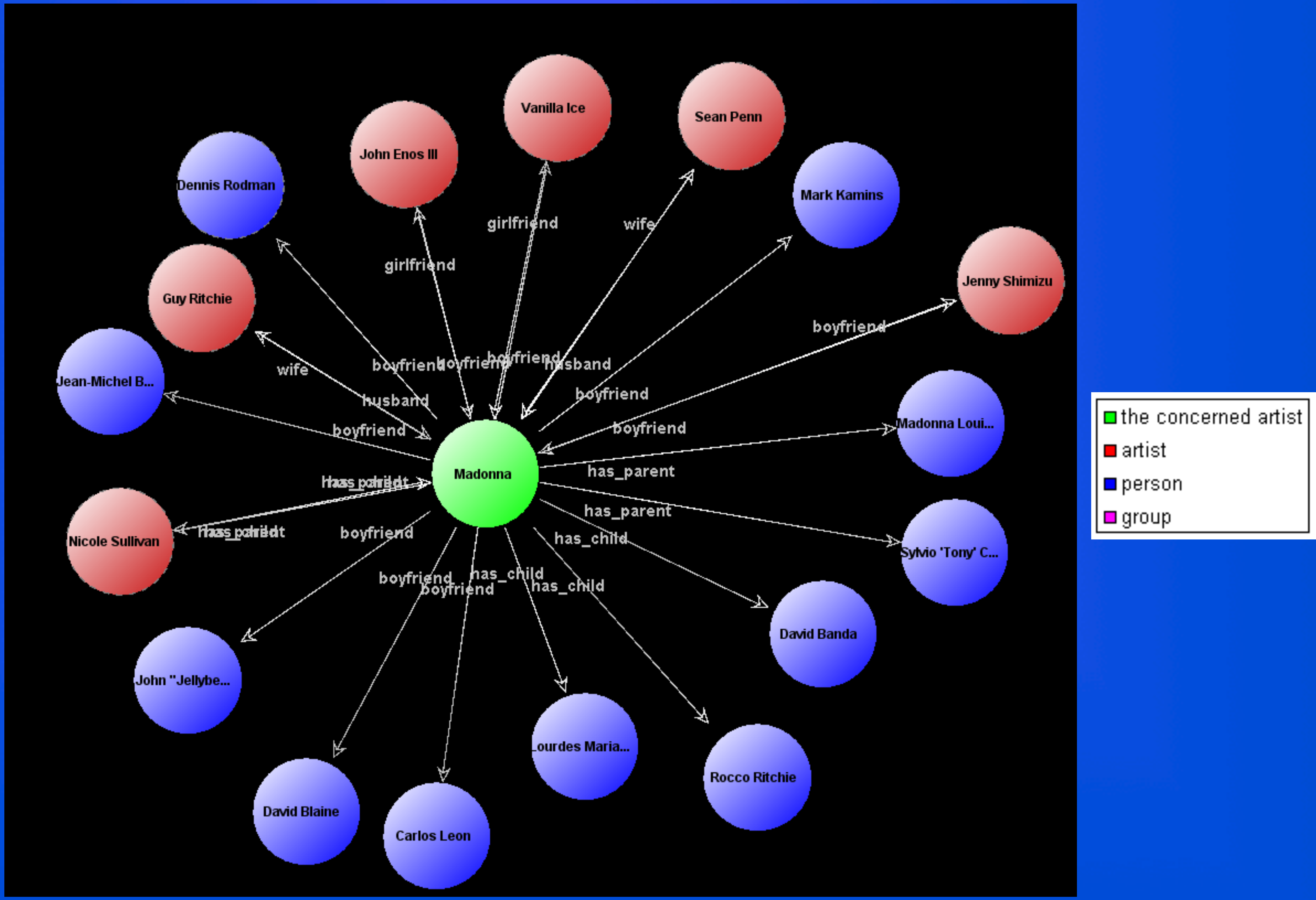
DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.html

OtherCompanyJobs: foodscience.com-Job1

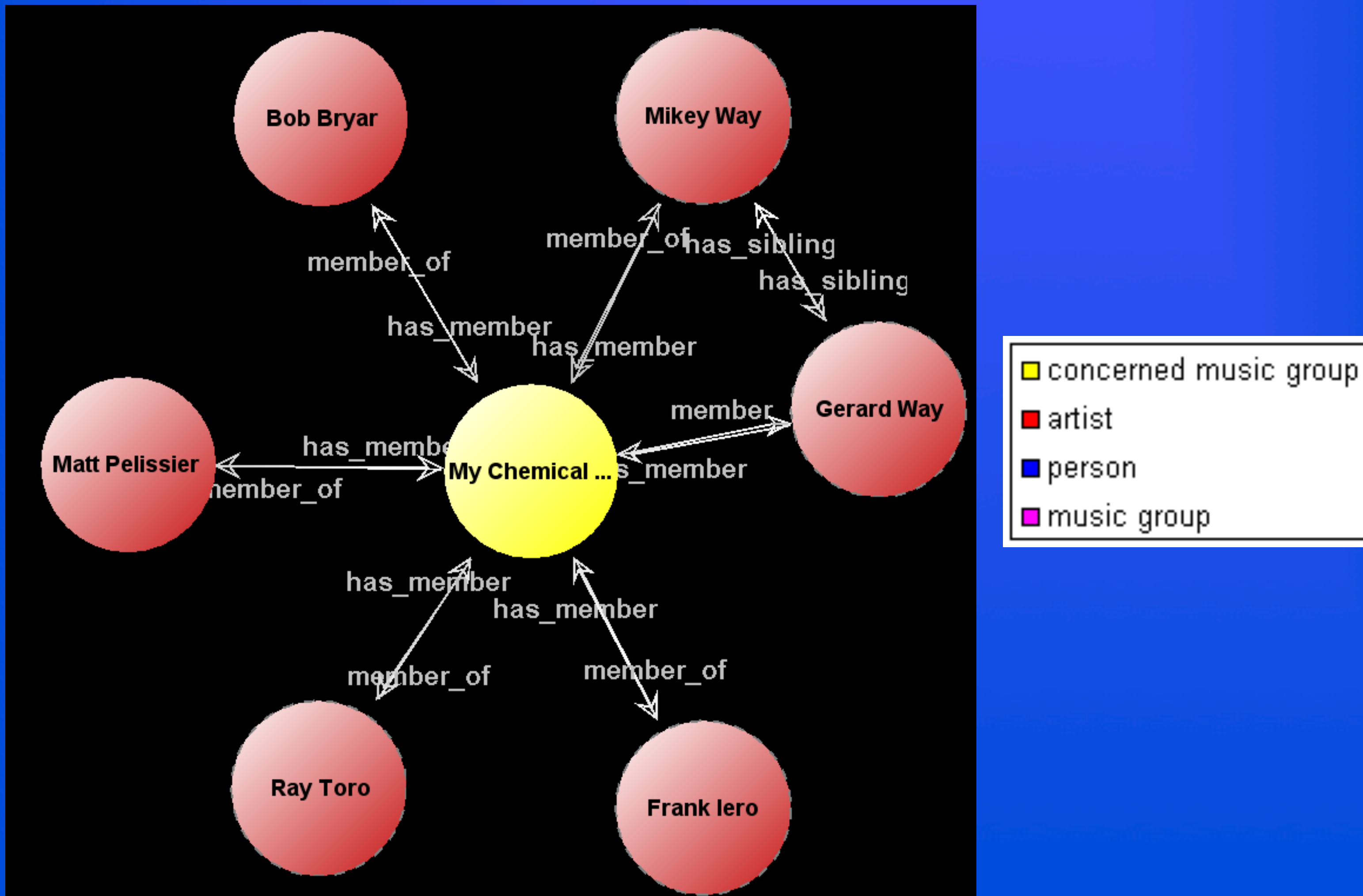
Example of Binary Social Relations

Social Network of “Madonna” (Depth = 1)



Examples of Binary Relations

Social Network of “My Chemical Romance” (Depth = 1)



Components of an IE Semantic Model (1)

- Entities - Individuals in the world *that are mentioned in a text*
  - Simple entities: singular objects
  - Collective entities: sets of objects of the same type *where the set is explicitly mentioned in the text*
- Relations – Properties that hold of tuples of entities.
- Complex Relations – Relations that hold among entities and relations
- Attributes – one place relations are attributes or individual properties

Components of an IE Semantic Model (2)

- Temporal points and intervals
- Relations may be timeless or bound to time intervals
- Events – A particular kind of simple or complex relation among entities involving a change in relation state at the end of a time interval.

Example

1. Three of the Nobel Prizes for Chemistry during the first decade **were awarded** for pioneering work in organic chemistry.
2. In **1902** **Emil Fischer** (1852-1919), then in Berlin, **was given** the prize for his work on sugar and purine syntheses.
3. Another major influence from organic chemistry was the development of the chemical industry, and a chief contributor here was Fischer's teacher, **Adolf von Baeyer** (1835-1917) in Munich, who **was awarded** the prize in **1905**.

Relations in Time

- time-dependent attribute: age(x)
- timeless two-place relation: father(x, y)
- time-dependent two-place relation: boss(x, y)

IE as a Semantic Analysis: Relating Language to the Model

- Linguistic Mention
  - A particular linguistic phrase
  - Denotes a particular entity, relation, or event
    - A noun phrase, name, or possessive pronoun
    - A verb, nominalization, compound nominal, or other linguistic construct relating other linguistic mentions
- Linguistic Entity
  - Equivalence class of mentions with same meaning
    - Coreferring noun phrases
    - Relations and events derived from different mentions, but conveying the same meaning

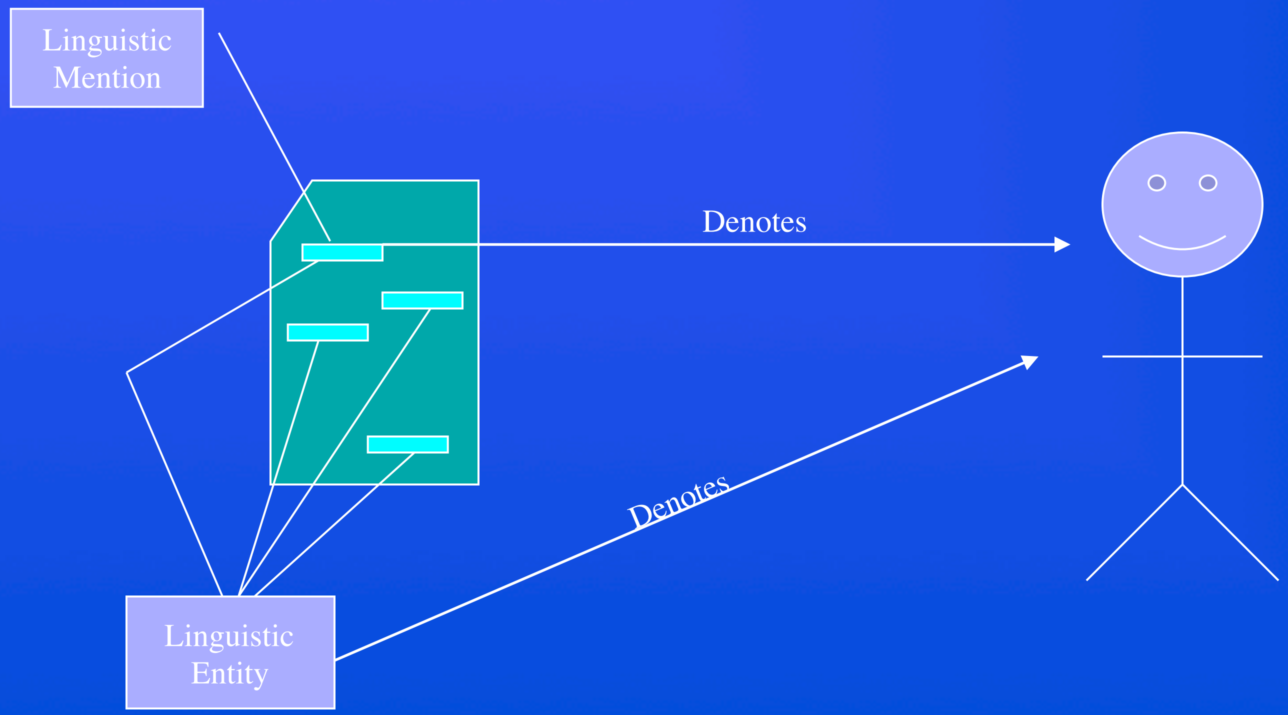
[Appelt, 2003]

The Basic Semantic Tasks of an IE System

- Recognition of linguistic entities
- Classification of linguistic entities into semantic types
- Identification of coreference equivalence classes of linguistic entities
- Identifying the actual individuals that are mentioned in an article
  - Associating linguistic entities with predefined individuals (e.g. a database, or knowledge base)
  - Forming equivalence classes of linguistic entities from different documents.

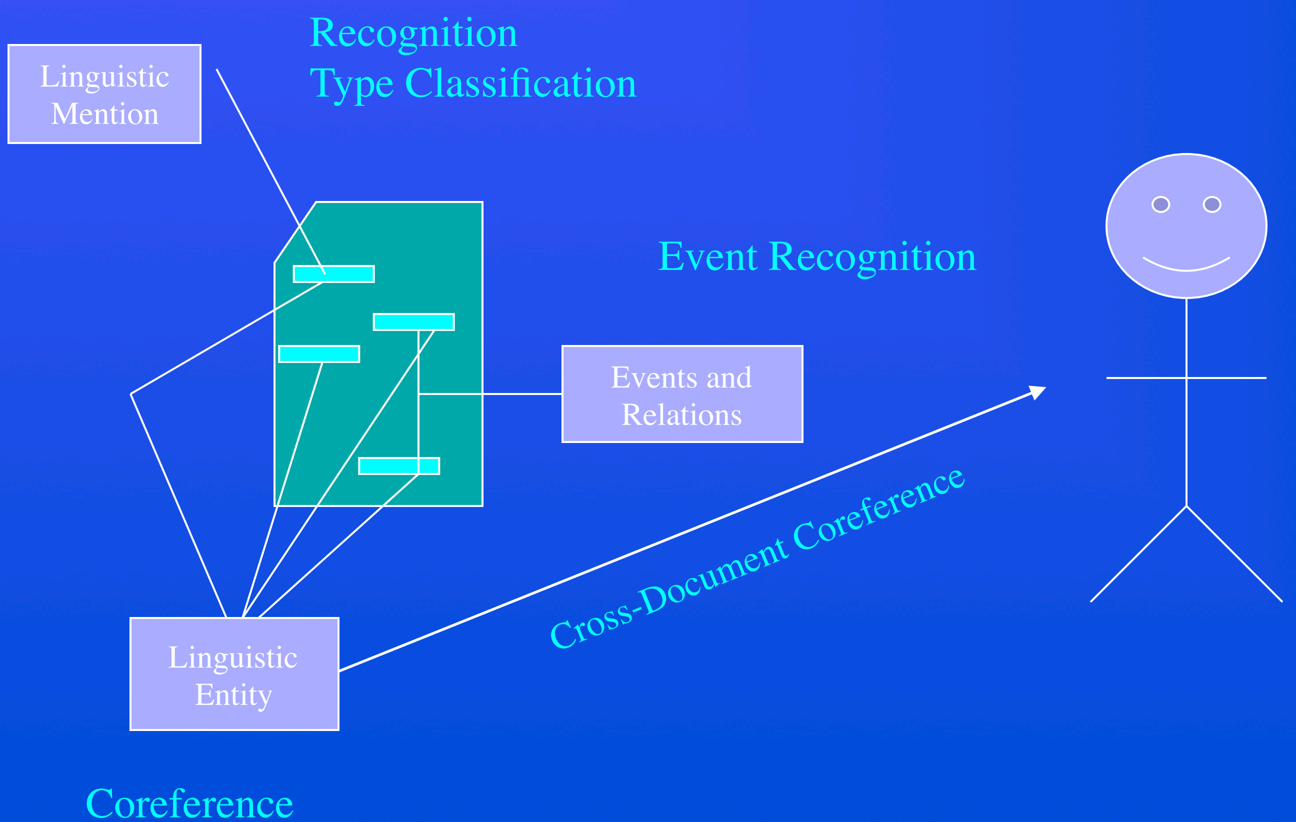
[Appelt, 2003]

Language and World Model



[Appelt, 2003]

NLP Tasks in an Extraction System



[Appelt, 2003]

Types of Linguistic Mentions

- Name mentions
  - The mention uses a proper name to refer to the entity
- Nominal mentions
  - The mention is a noun phrase whose head is a common noun
- Pronominal mentions
  - The mention is a headless noun phrase, or a noun phrase whose head is a pronoun, or a possessive pronoun

Example of Linguistic Mentions

1. Three of the Nobel Prizes for Chemistry during the first decade were awarded for pioneering work in organic chemistry.
2. In 1902 Emil Fischer (1852-1919), then in Berlin, was given the prize for his work on sugar and purine syntheses.
3. Another major influence from organic chemistry was the development of the chemical industry, and a chief contributor here was Fischer's teacher, Adolf von Baeyer (1835-1917) in Munich, who was awarded the prize in 1905.

Relation Extraction Example

A relation extraction task in the domain *management succession* (MUC-6)

< person_in, person_out, position, organisation>

- *person_in*: the person who obtained the position
- *person_out*: the person who left the position
- *position*: the job position that the two persons were involved in
- *organisation*: the organisation where the position was located

According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

According to CEO Fred Bell, <u>Acme Inc.</u> hired <u>Peter Smith</u> as <u>COO</u> yesterday, replacing <u>Hans Bloggs.</u>

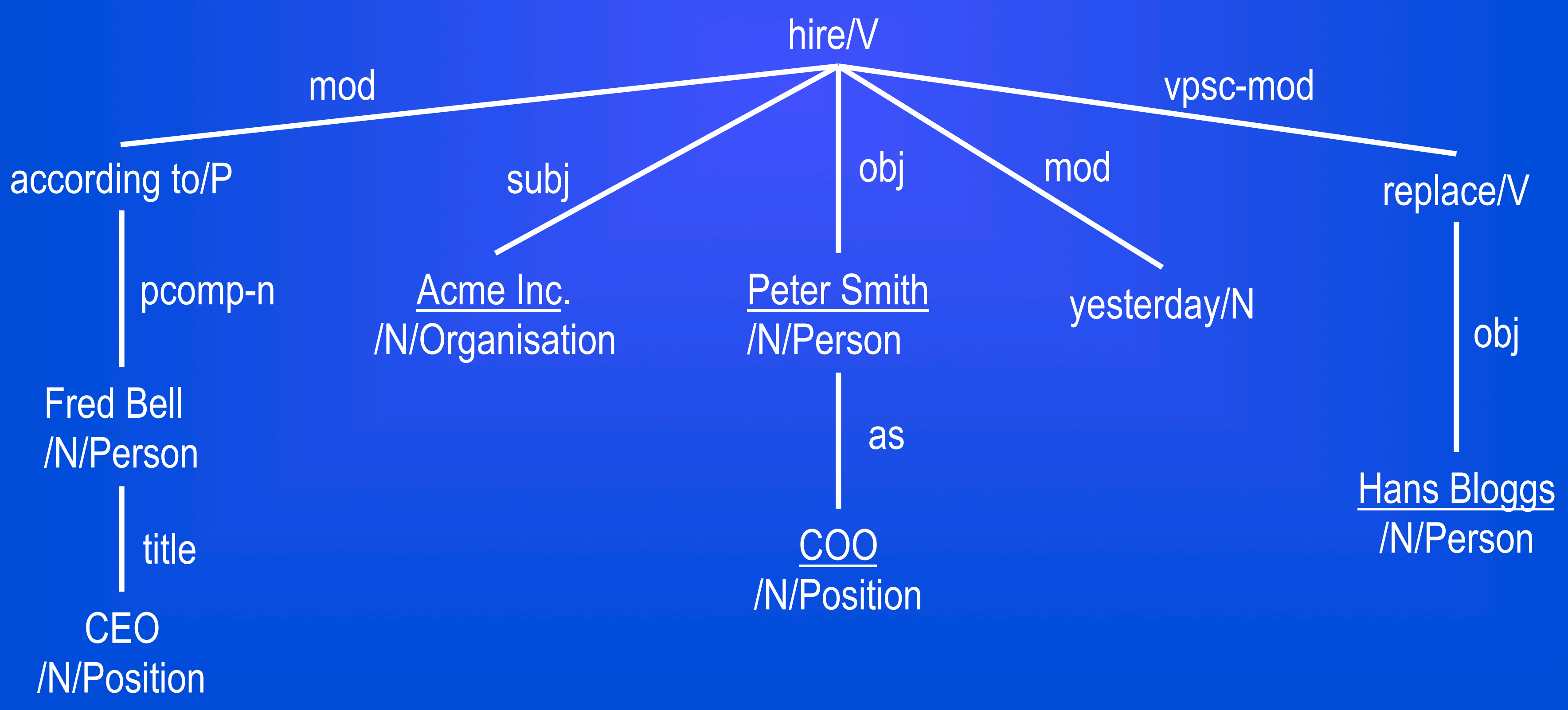
<person_in, person_out, position, organisation>

According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<person_in, person_out, position, organisation>

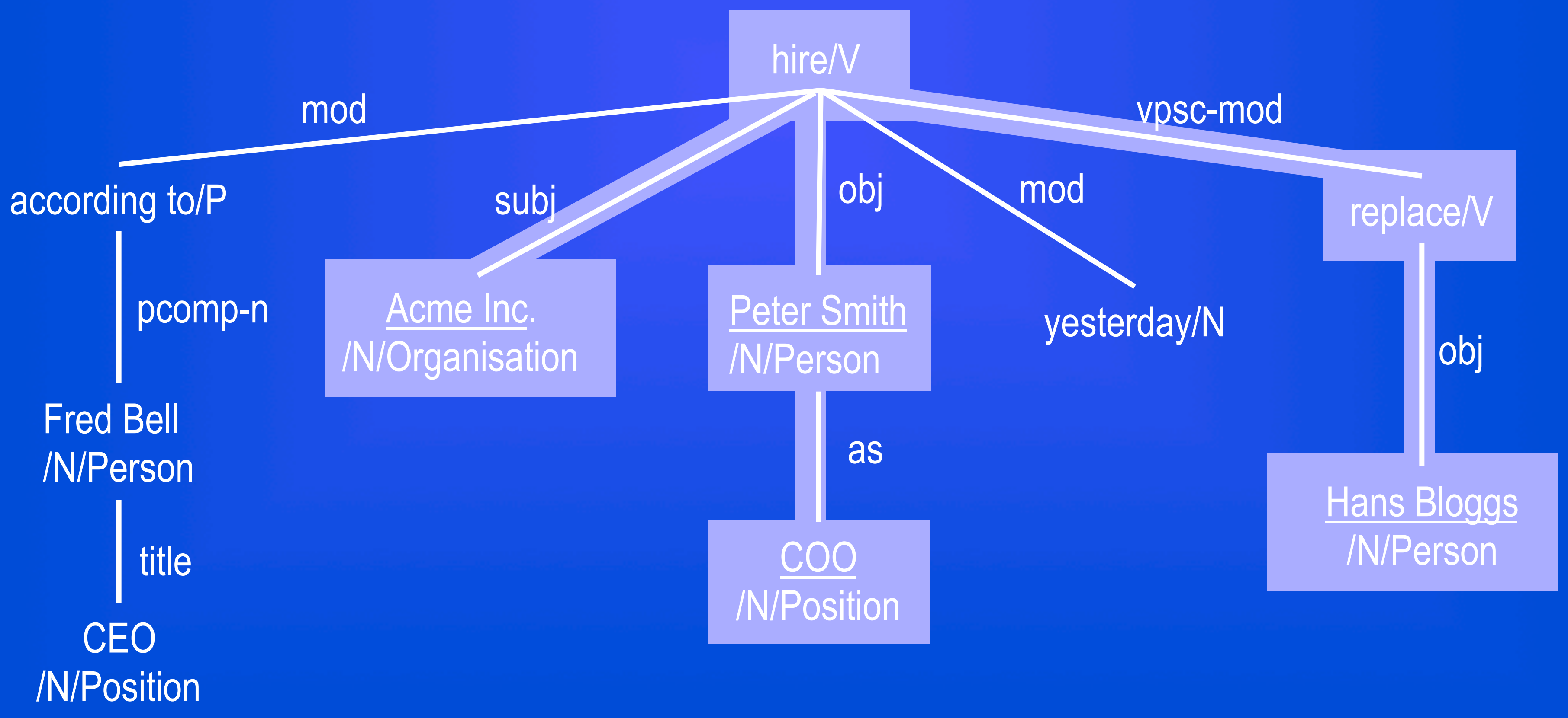
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

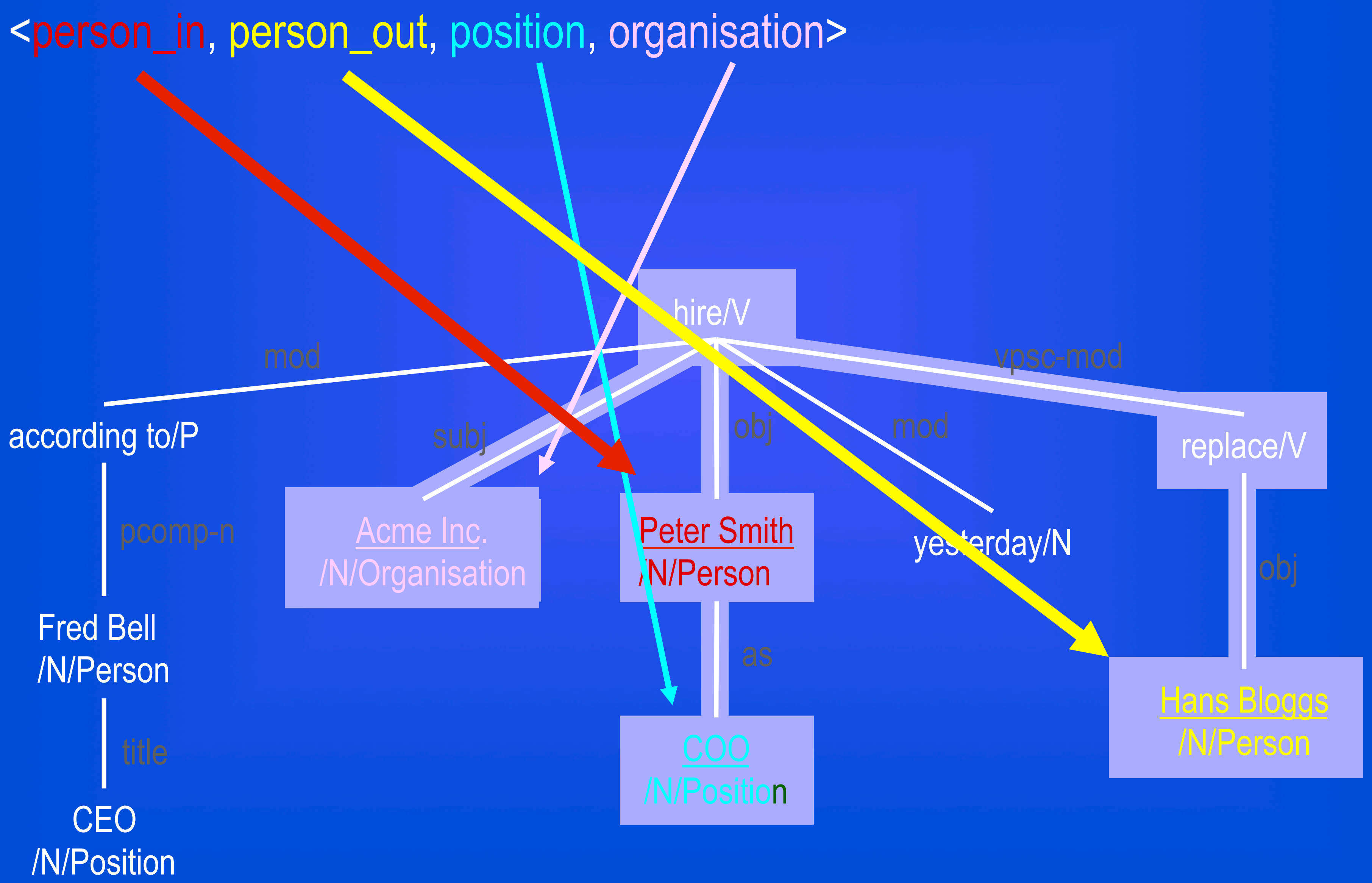
<person_in, person_out, position, organisation>



According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<person_in, person_out, position, organisation>





<person_in, person_out, position, organisation>
hire/V
mod
vpsc-mod
according to/P
subj
obj
mod
replace/V
pcomp-n
Acme Inc.
/N/Organisation
Peter Smith
/N/Person
yesterday/N
obj
Fred Bell
/N/Person
as
Hans Bloggs
/N/Person
title
COO
/N/Position
CEO
/N/Position

Explicit and Implicit Relations

- Many relations are true in the world. Reasonable knoweldge bases used by extraction systems will include many of these relations. Semantic analysis requires focusing on certain ones that are directly motivated by the text.

- Example:
  - Baltimore is in Maryland, which is in United States.

Text mentions Baltimore and United States. Is there a relation between Baltimore and United States?

Another Example

- *Prime Minister Tony Blair attempted to convince the British Parliament of the necessity of intervening in Iraq.*

Is there a role relation specifying Tony Blair as prime minister of Britain?

A test: a relation is implicit in the text if the text provides convincing evidence that the relation actually holds.

Explicit Relations

- Explicit relations are expressed by certain surface linguistic forms
  - Copular predication - Clinton was the president.
  - Prepositional Phrase - The CEO of Microsoft…
  - Prenominal modification - The American envoy…
  - Possessive - Microsoft’s chief scientist…
  - SVO relations - Clinton arrived in Tel Aviv…
  - Nominalizations - Anan’s visit to Baghdad…
  - Apposition - Tony Blair, Britain’s prime minister…

A Brief History of IE

Message Understanding Conferences [MUC-7 98]

- U.S. Government sponsored conferences with the intention to coordinate multiple research groups seeking to improve IE and IR technologies (since 1987)
- defined several generic types of information extraction tasks (MUC Competition)
- MUC 1-2 focused on automated analysis of military messages containing textual information
- MUC 3-7 focused on information extraction from newswire articles
    - terrorist events
    - international joint-ventures
    - management succession event

Evaluation of IE systems in MUC

- Participants receive description of the scenario along with the annotated *training corpus* in order to adapt their systems to the new scenario (1 to 6 months)
- Participants receive new set of documents (*test corpus*) and use their systems to extract information from these documents and return the results to the conference organizer
- The results are compared to the manually filled set of templates
(*answer key*)

Evaluation of IE systems in MUC

- precision and recall measures were adopted from the information retrieval research community

$$recall = \frac{N_{correct}}{N_{key}} \qquad precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}}$$

$$F = \frac{(\beta^2 + 1) \times precision \times recall}{\beta^2 \times precision + recall}$$

- Sometimes an *F*-meassure is used as a combined recall-precision score

Development Steps within IE Communities

- from attempts to use the methods of full text understanding to shallow text processing;
- from pure knowledge-based hand-coded systems to (semi-) automatic systems using machine learning methods;
- from complex domain-dependent event extraction to standardized domain-independent elementary entity identification, simple semantic relation and event extraction.

Machine Learning
for
Relation Extraction

Motivations of ML

- Porting to new domains or applications is expensive
- Current technology requires IE experts
  - Expertise difficult to find on the market
  - SME cannot afford IE experts
- Machine learning approaches
  - Domain portability is relatively straightforward
  - System expertise is not required for customization
  - "Data driven" rule acquisition ensures full coverage of examples

Problems

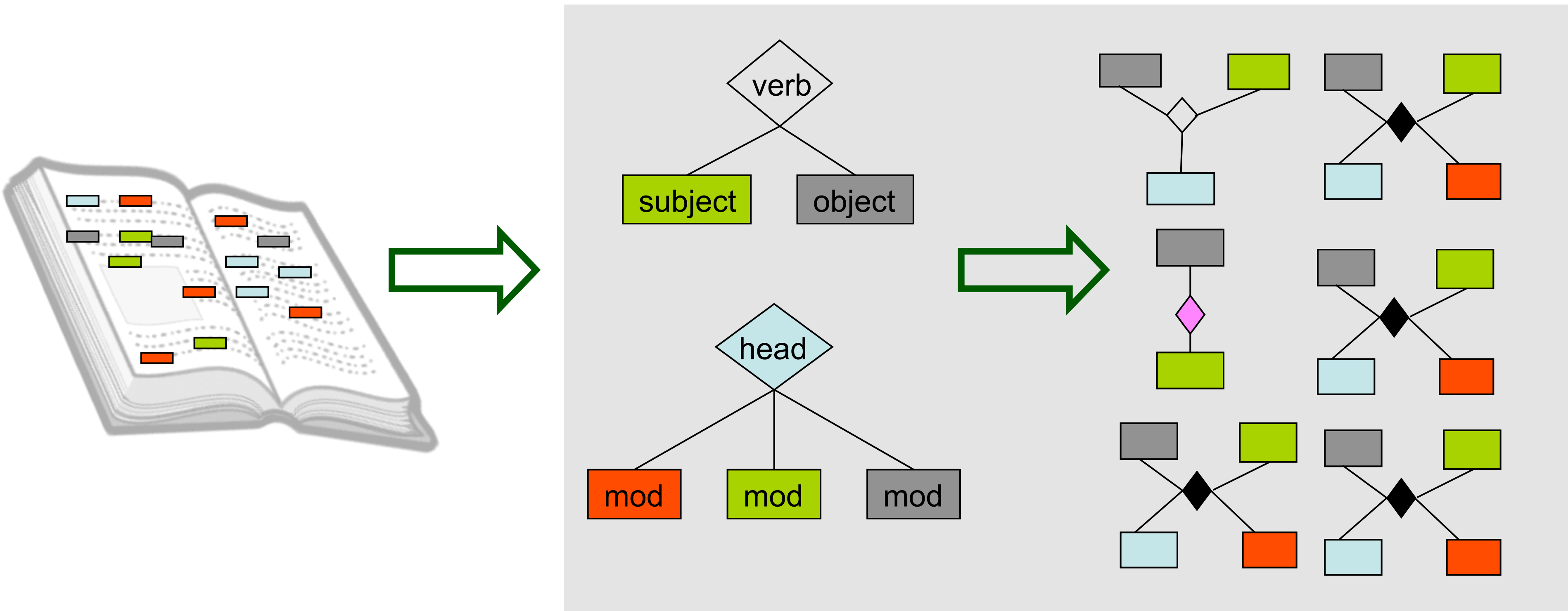
- Training data may not exist, and may be very expensive to acquire
- Large volume of training data may be required
- Changes to specifications may require reannotation of large quantities of training data
- Understanding and control of a domain adaptive system is not always easy for non-experts

Parameters of IE Real-World Tasks

- Document structure
  - Free text
  - Semi-structured
  - Structured
- Linguistic annotation
  - Shallow NLP
  - Deep NLP
- Complexity and specificity of relation
  - Unary
  - N-ary
- Depth of extraction
  - Recognition
  - Classification
  - Semantic role labelling
- Degree of automation
  - Semi-automatic
  - Supervised
  - Semi-Supervised
  - Minimally-Supervised
  - Unsupervised
- Human interaction/contribution
- Data properties
  - Domain relevance
  - Redundancy
  - Connectivity
- Evaluation/validation
  - With/without gold standard
  - Performance: recall & precision
  - Interaction among parameters

Research Goal

Development of a general framework for automatically learning mappings between linguistic analyses and target semantic relations, with minimal human intervention.



Challenges

- Easy adaptation to new relation types with varied complexity
- Automatic learning without annotated corpus
- Exhaustive discovery of relevant linguistic patterns
- Integration of semantic role information into linguistic patterns

Example

A relation extraction task in the domain *management succession* (MUC-6)

< person_in, person_out, position, organisation>

- *person_in*: the person who obtained the position
- *person_out*: the person who left the position
- *position*: the job position that the two persons were involved in
- *organisation*: the organisation where the position was located

According to CEO Fred Bell, <u>Acme Inc.</u> hired <u>Peter Smith</u> as <u>COO</u> yesterday, replacing <u>Hans Bloggs.</u>

According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

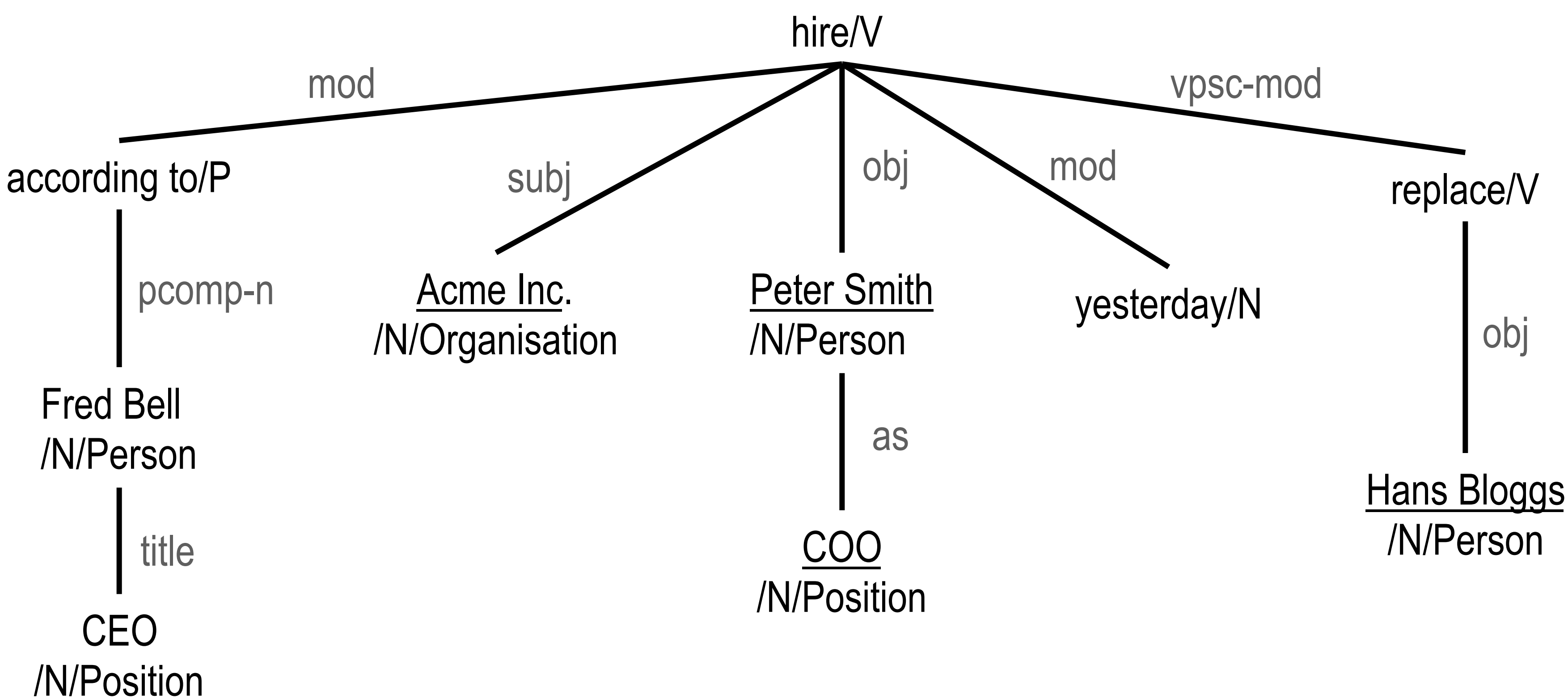
<person_in, person_out, position, organisation>

According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<person_in, person_out, position, organisation>

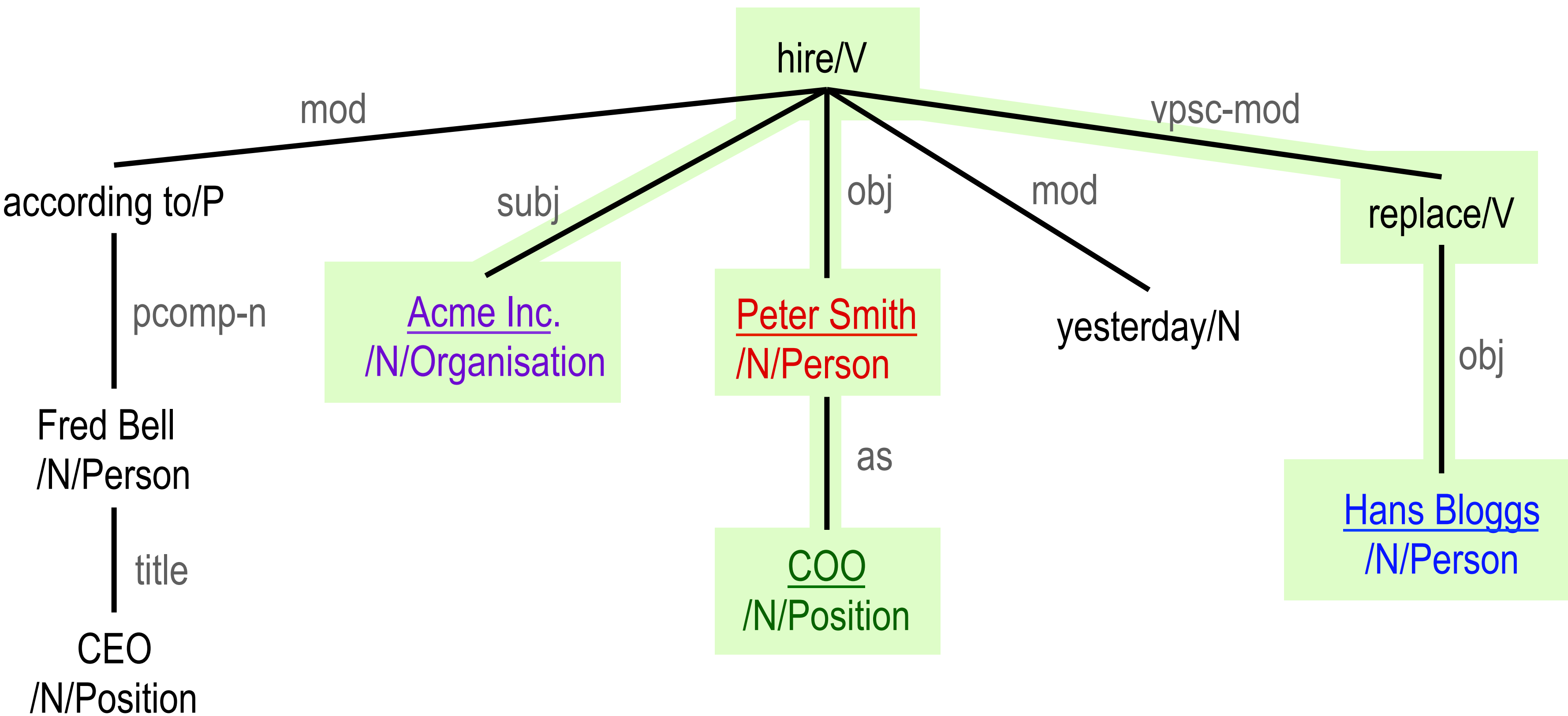
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<person_in, person_out, position, organisation>

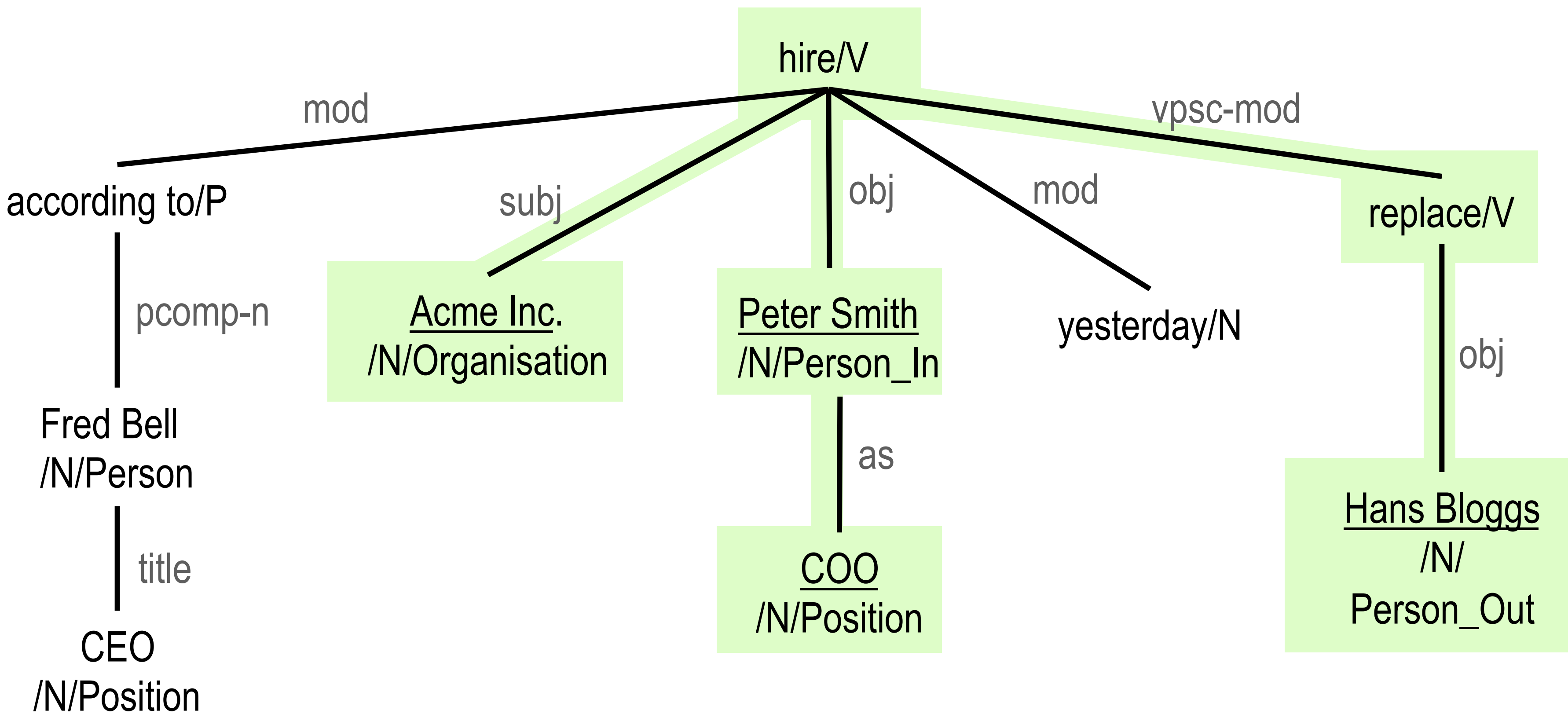


According to CEO Fred Bell, Acme Inc. hired Peter Smith
as COO yesterday, replacing Hans Bloggs.

<person_in, person_out, position, organisation>



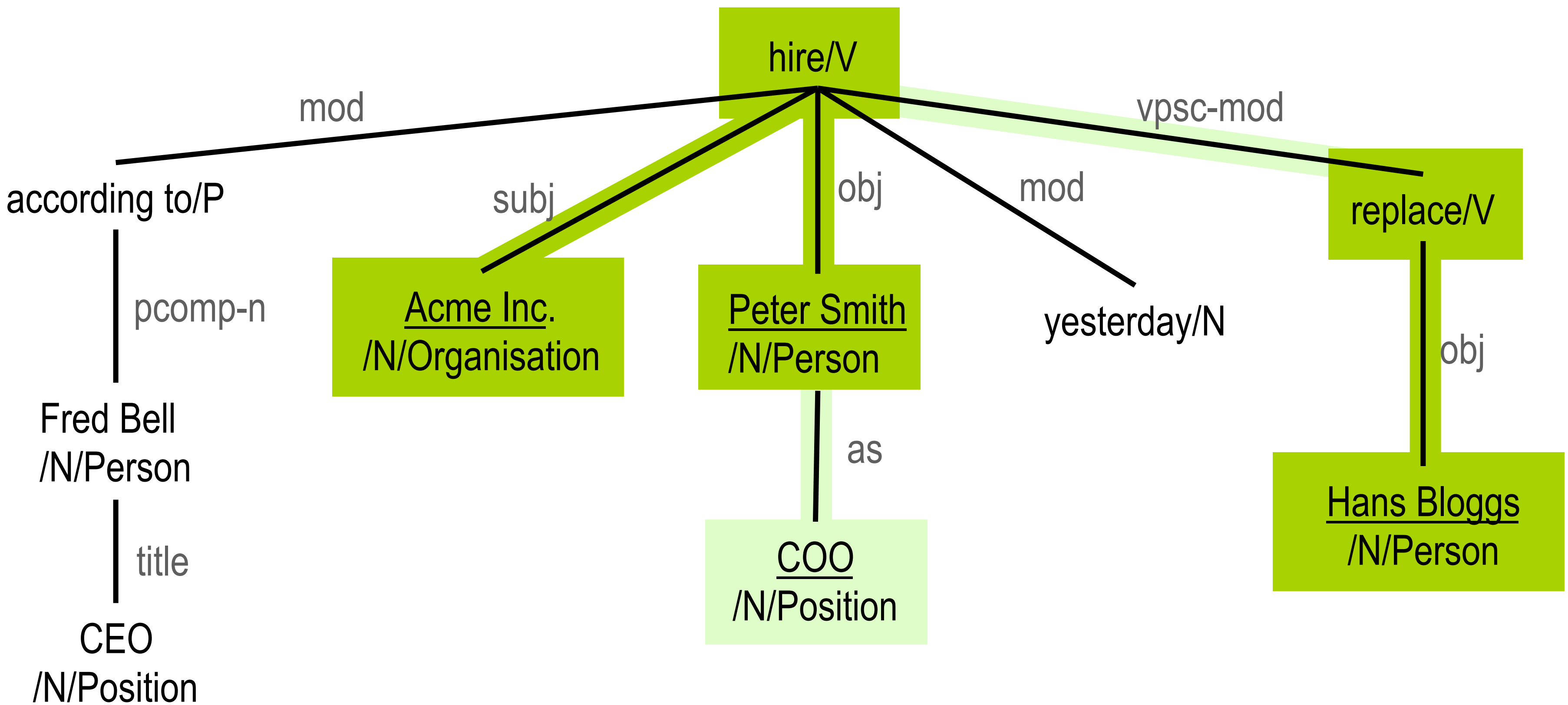
Ideal Target Pattern



Previous Work: SVO Model

Yangarber (2001)

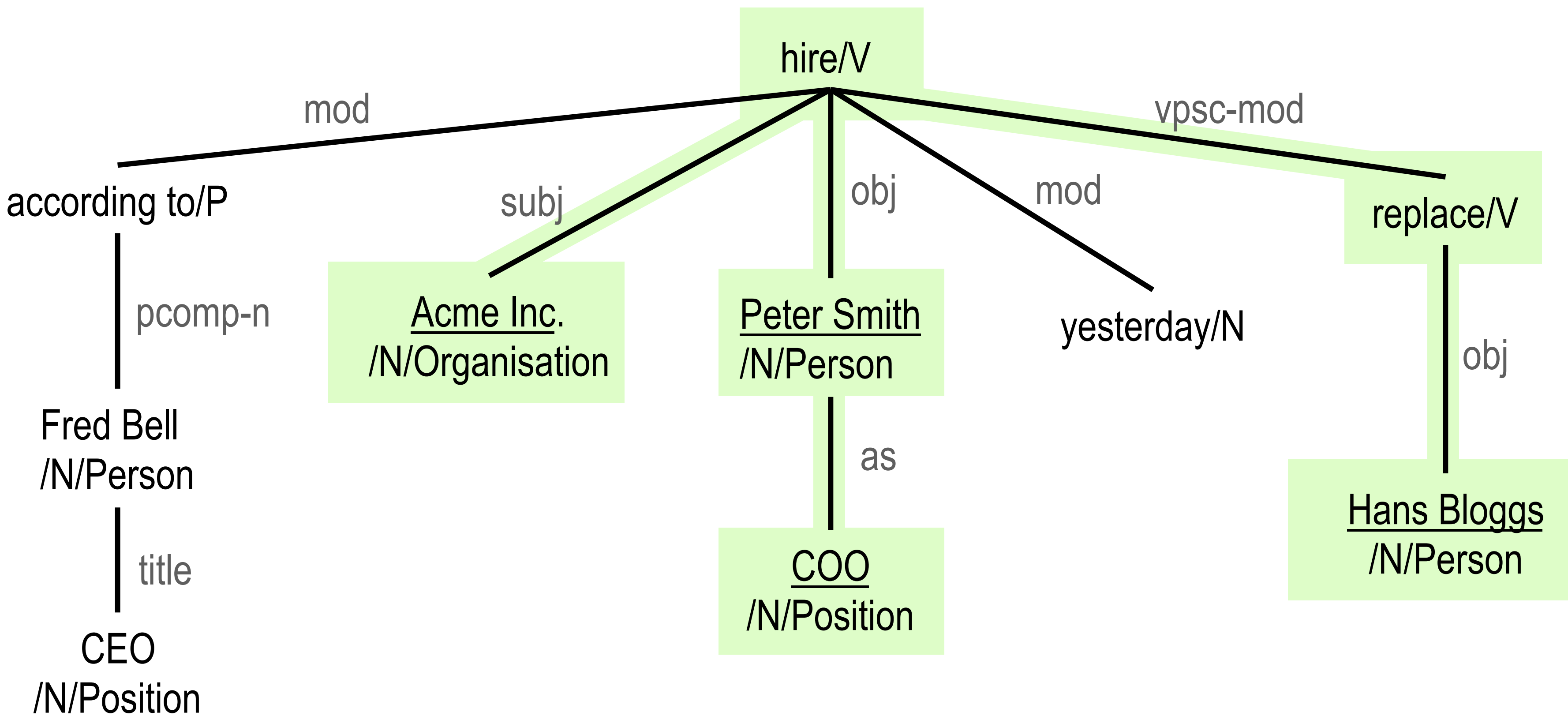
- Verb centered
- Direct relations between subject-verb-object
- Complex NP can not be extracted, e.g., the person and position relation
- The linguistic relations among patterns are not considered, e.g., hire and replace



Previous Work: Chain Model

Sudo et al. (2001)

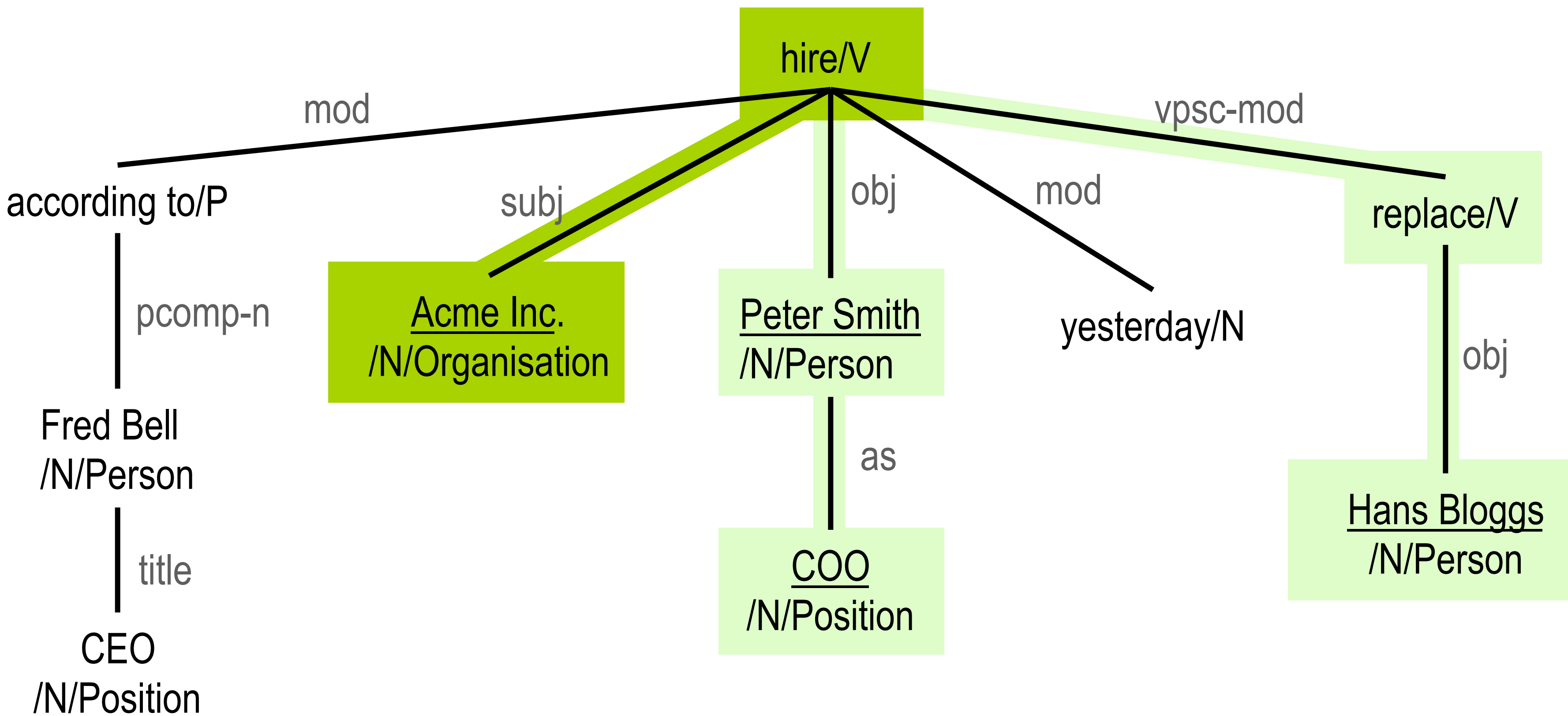
- Verb centered
- A single syntactic path dominated by a verb containing at least one relevant named entity concept



Previous Work: Chain Model

Sudo et al. (2001)

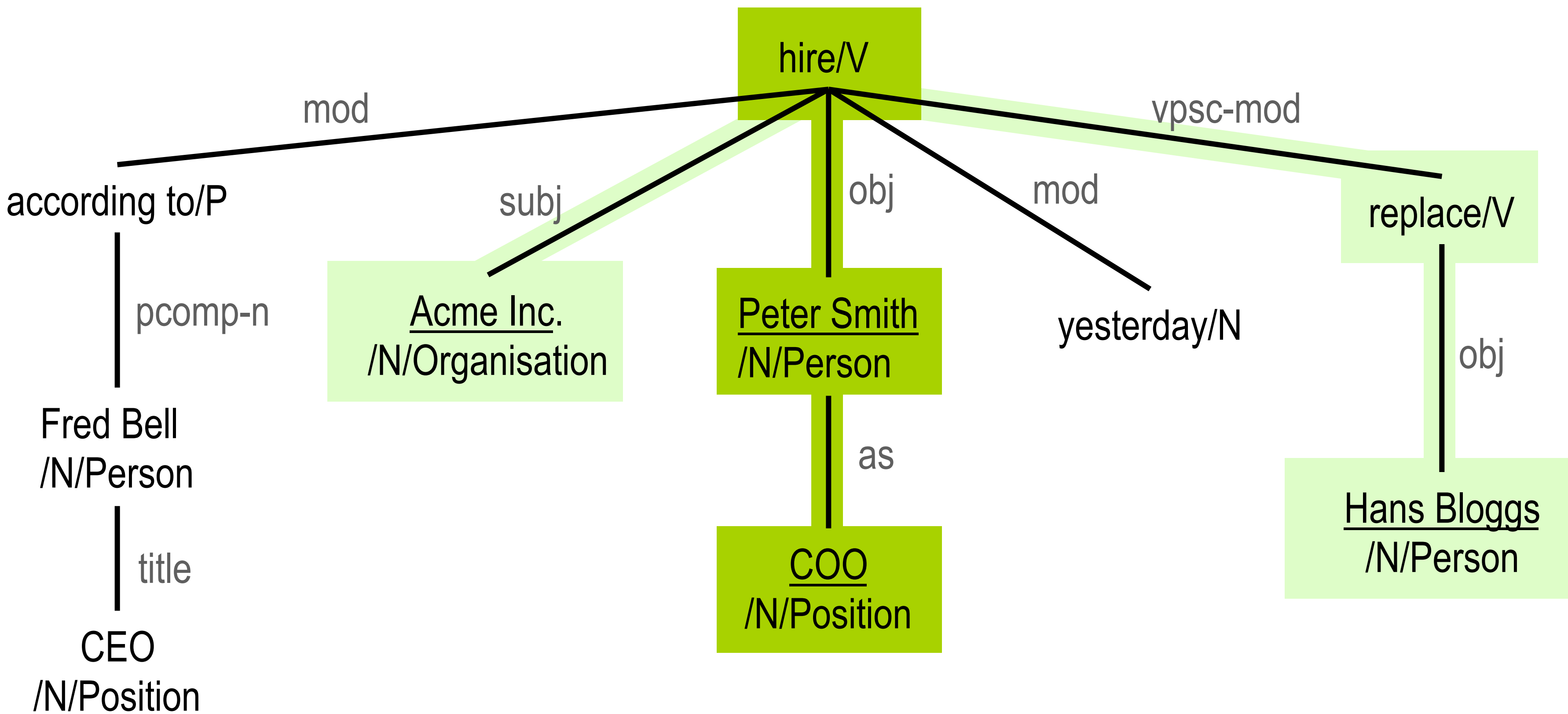
- Verb centered
- A single syntactic path dominated by a verb containing at least one relevant named entity concept



Previous Work: Chain Model

Sudo et al. (2001)

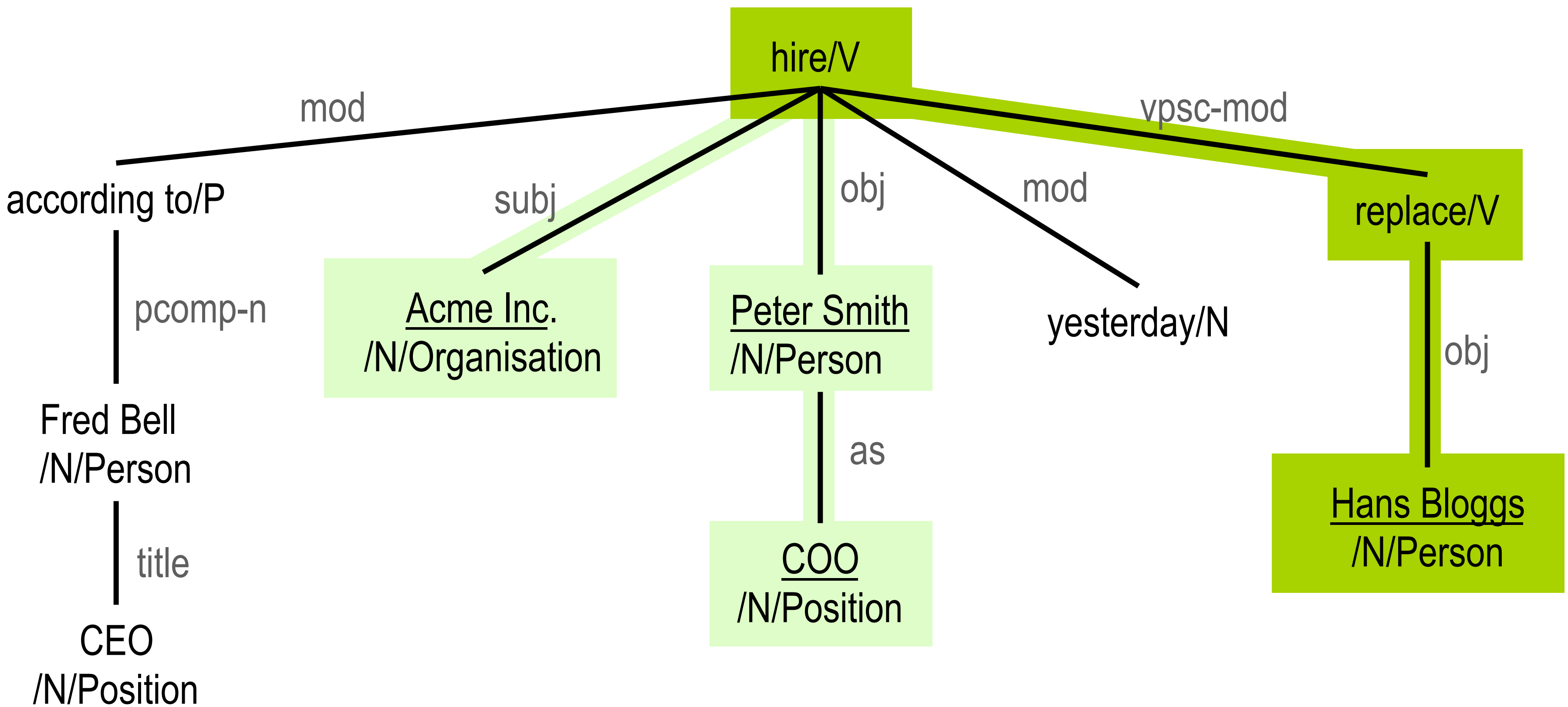
- Verb centered
- A single syntactic path dominated by a verb containing at least one relevant named entity concept



Previous Work: Chain Model

Sudo et al. (2001)

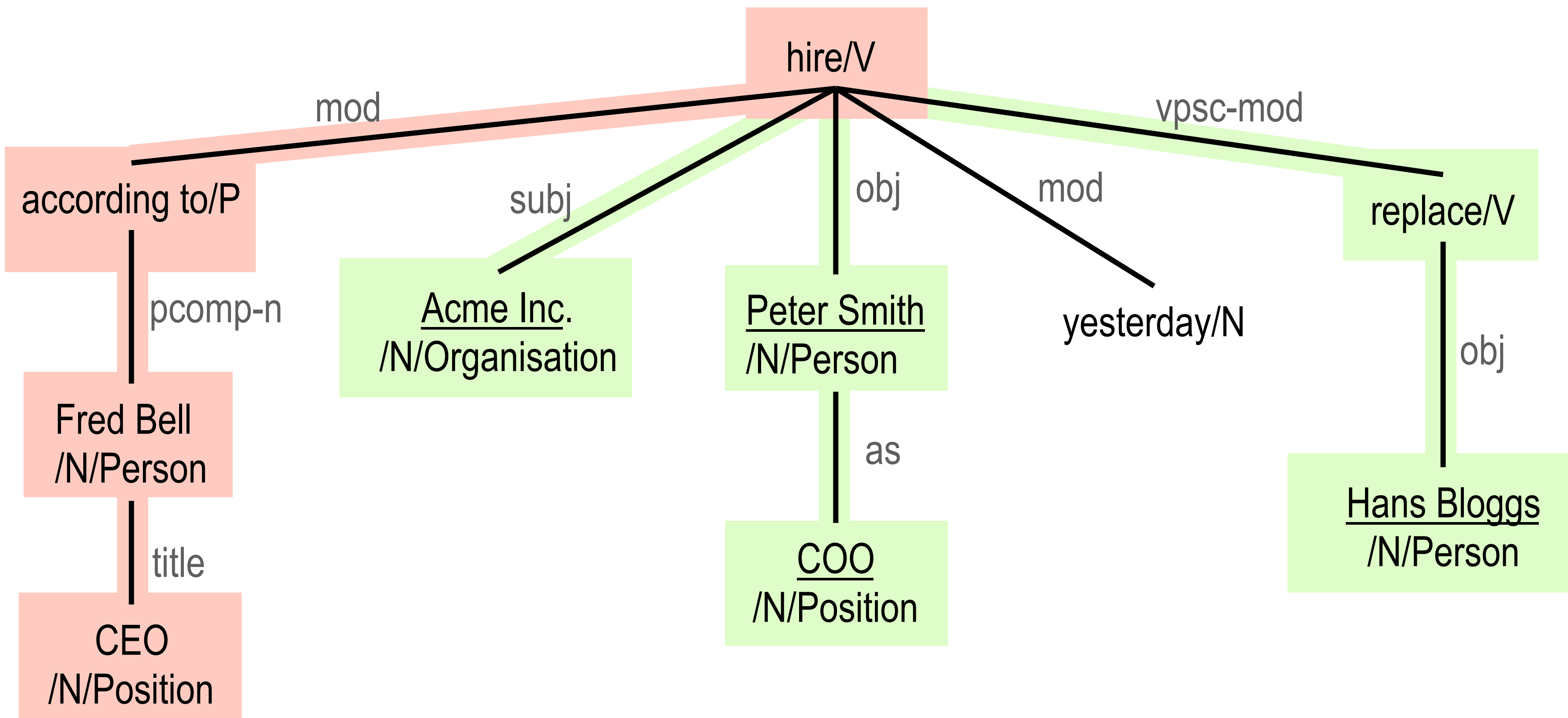
- Verb centered
- A single syntactic path dominated by a verb containing at least one relevant named entity concept



Previous Work: Chain Model

Sudo et al. (2001)

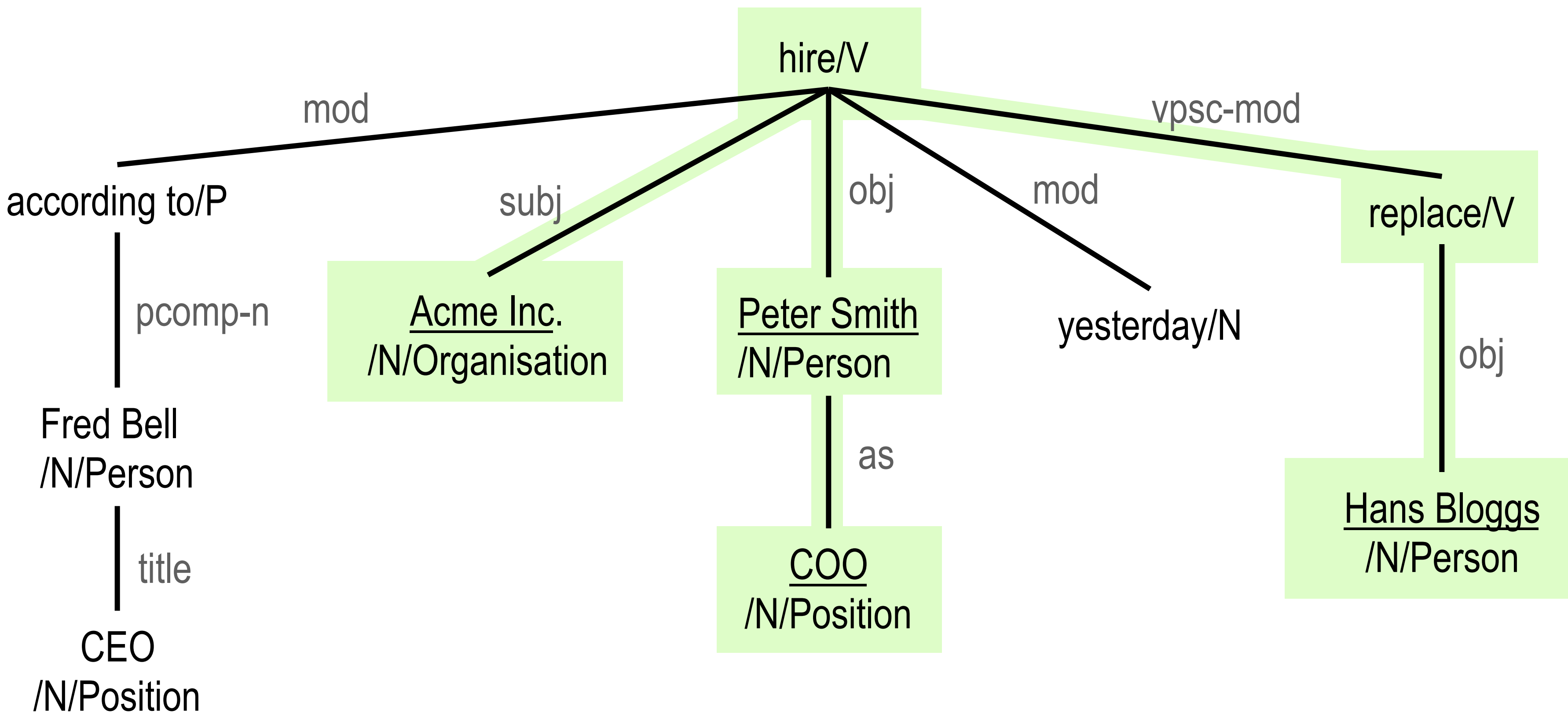
- Verb centered
- A single syntactic path dominated by a verb containing at least one relevant named entity concept



Previous Work: Linked Chain Model

Stevenson & Greenwood 2005

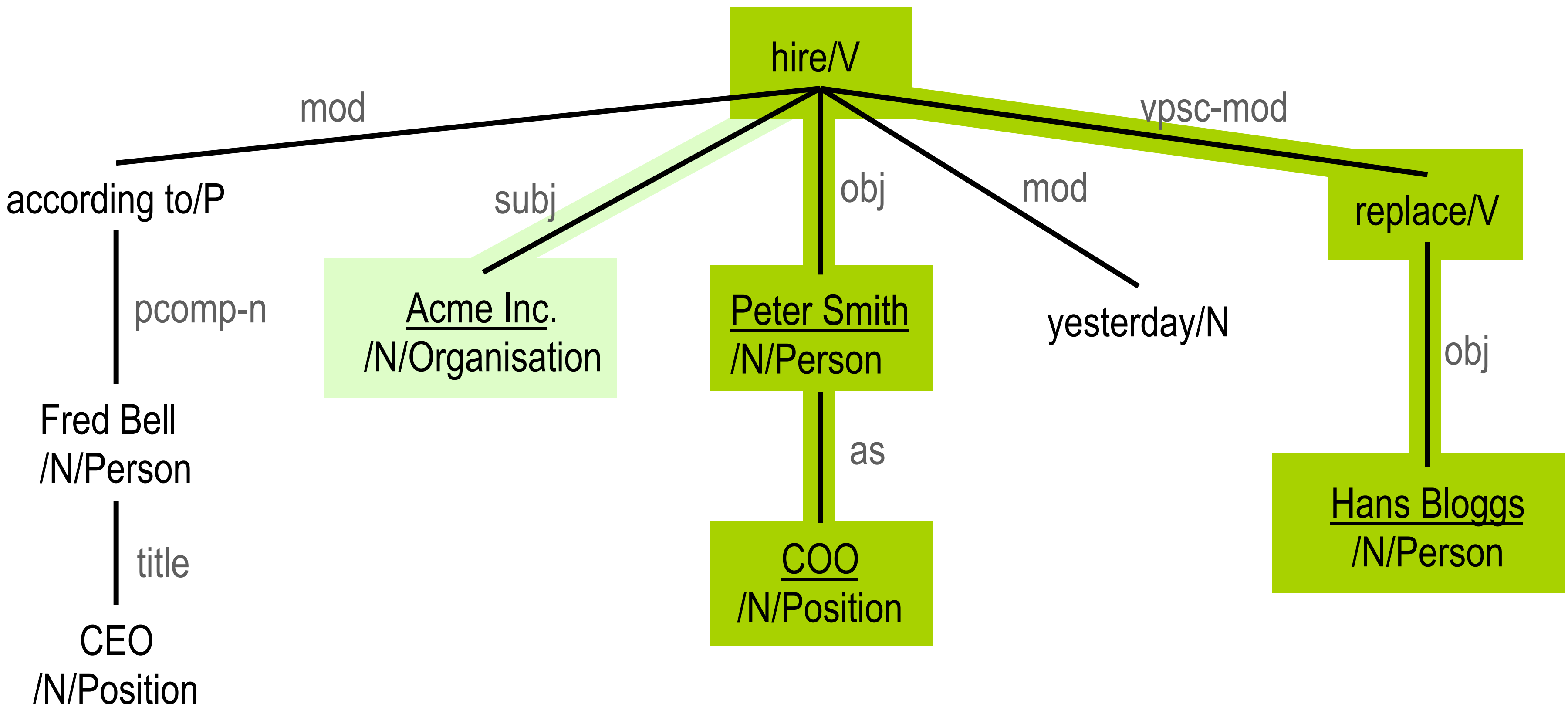
- verb centered
- pairs of chains instead of single paths



Previous Work: Linked Chain Model

Stevenson & Greenwood 2005

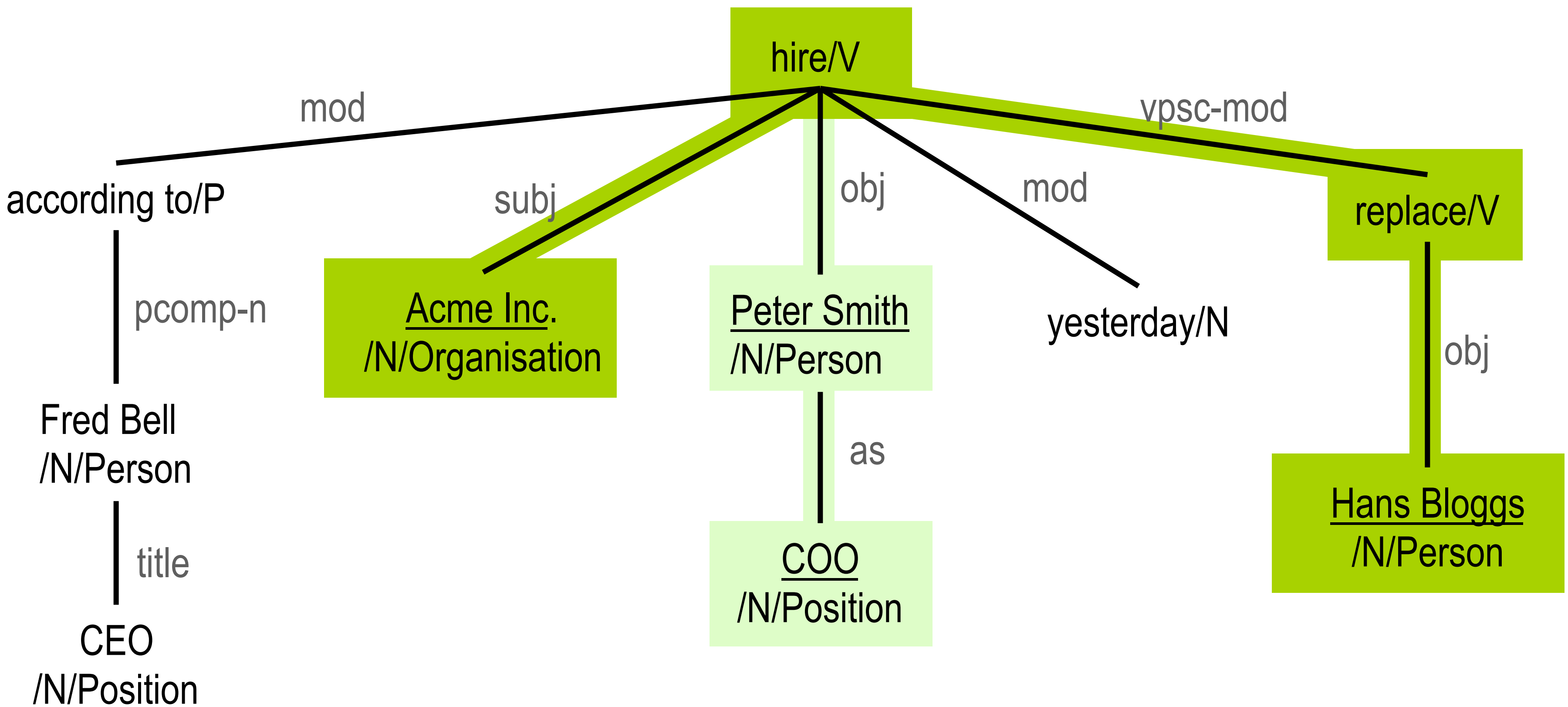
- verb centered
- pairs of chains instead of single paths



Previous Work: Linked Chain Model

Stevenson & Greenwood 2005

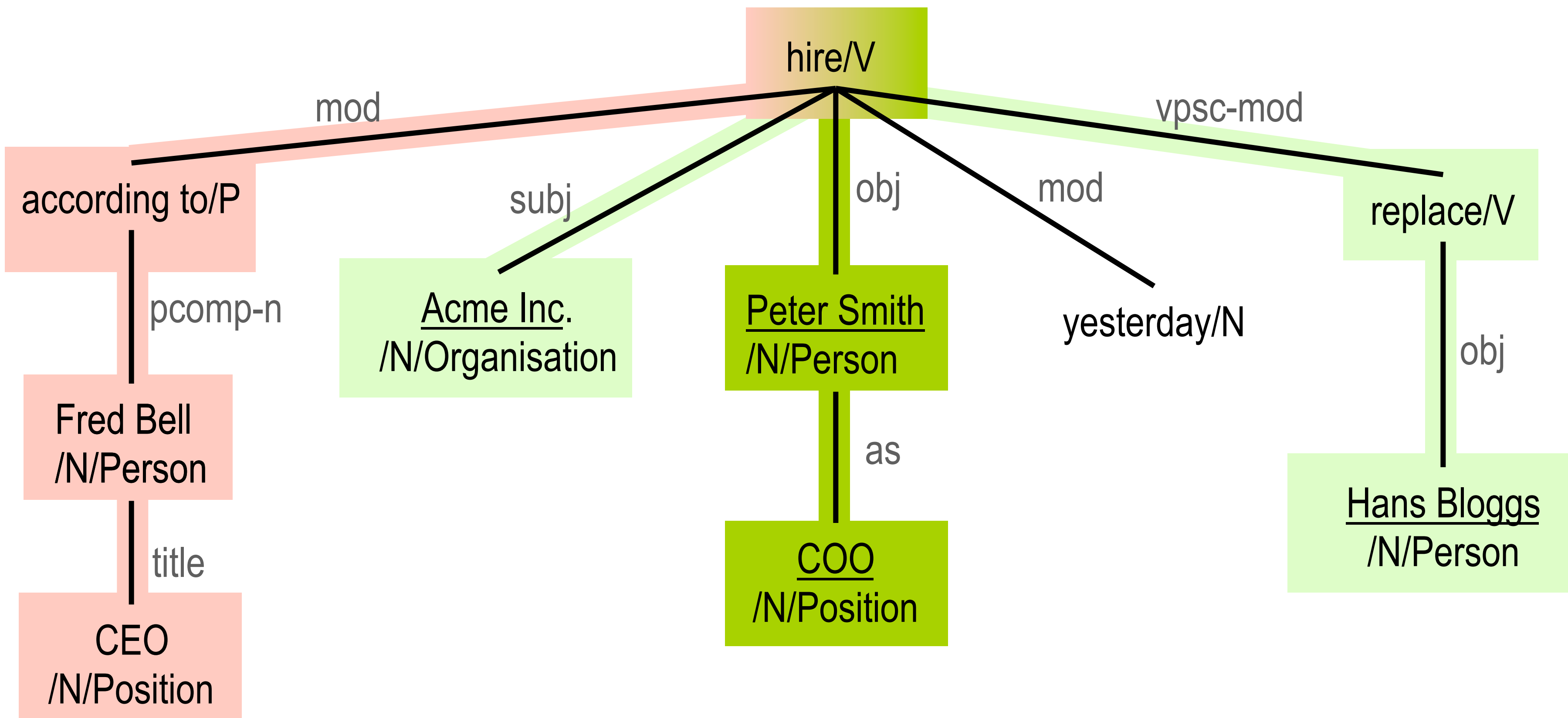
- verb centered
- pairs of chains instead of single paths



Previous Work: Linked Chain Model

Stevenson & Greenwood 2005

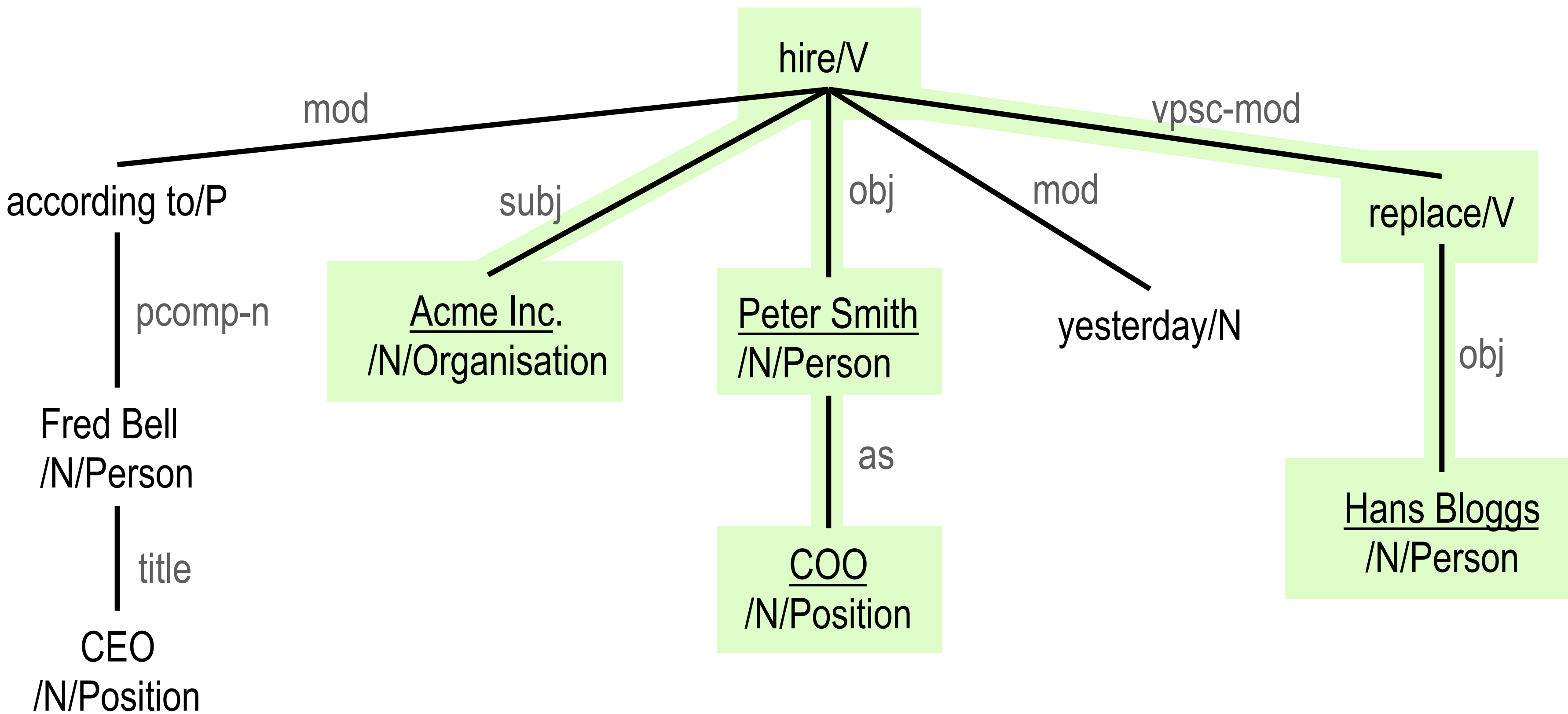
- verb centered
- pairs of chains instead of single paths



Previous Work: Subtree-Model

Sudo et al. (2003)

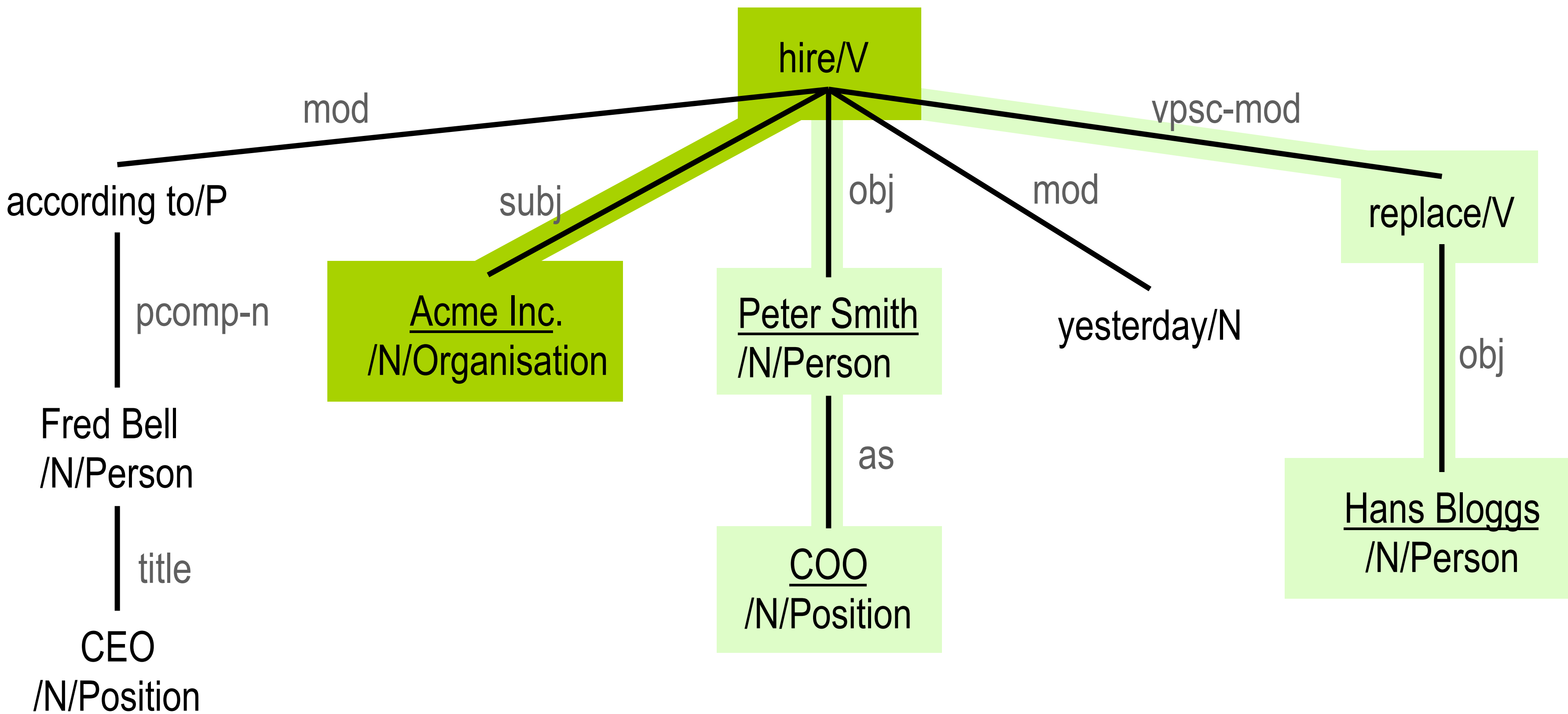
- verb centered
- All chains dominated by a verb, which contain at least one relevant named entity and their combinations



Previous Work: Subtree-Model

Sudo et al. (2003)

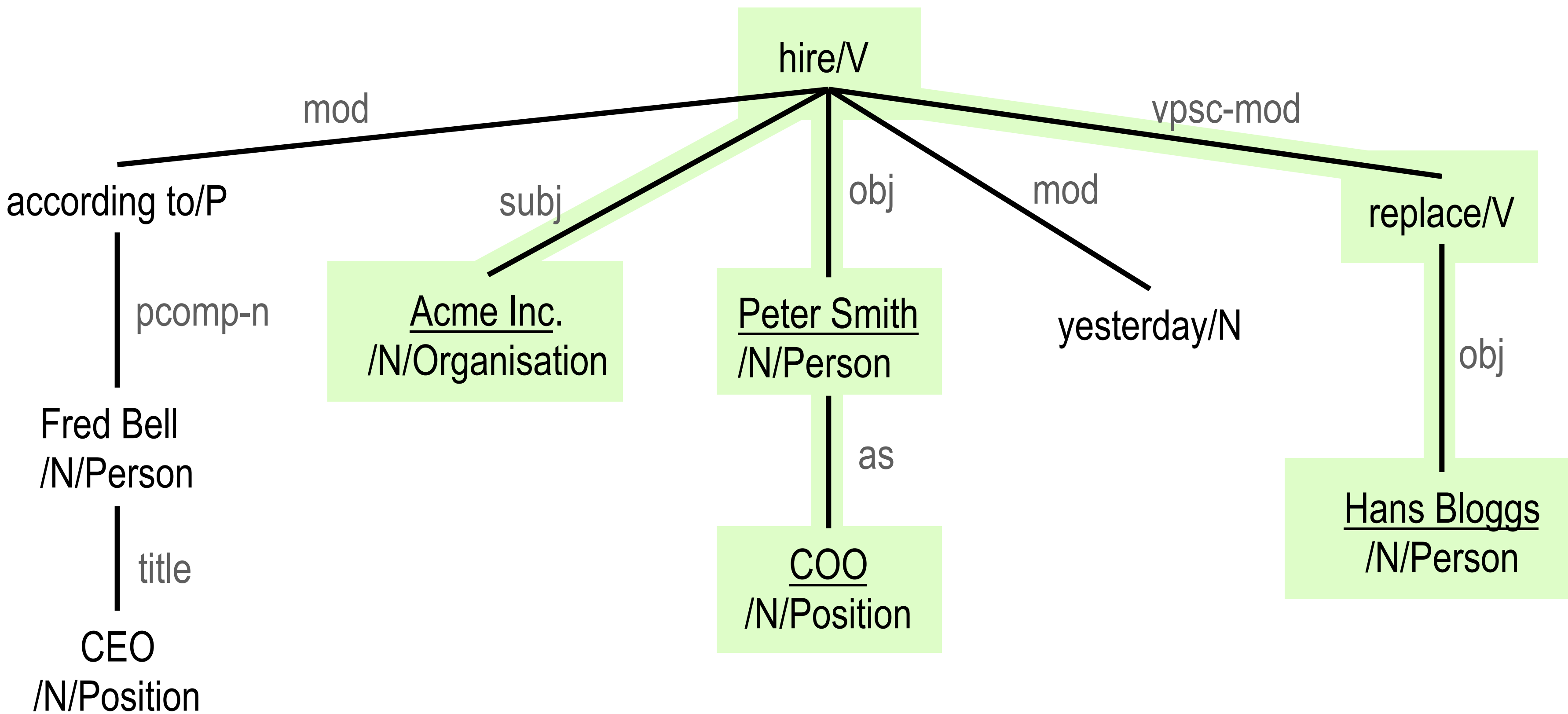
- verb centered
- All chains dominated by a verb, which contain at least one relevant named entity and their combinations



Previous Work: Subtree-Model

Sudo et al. (2003)

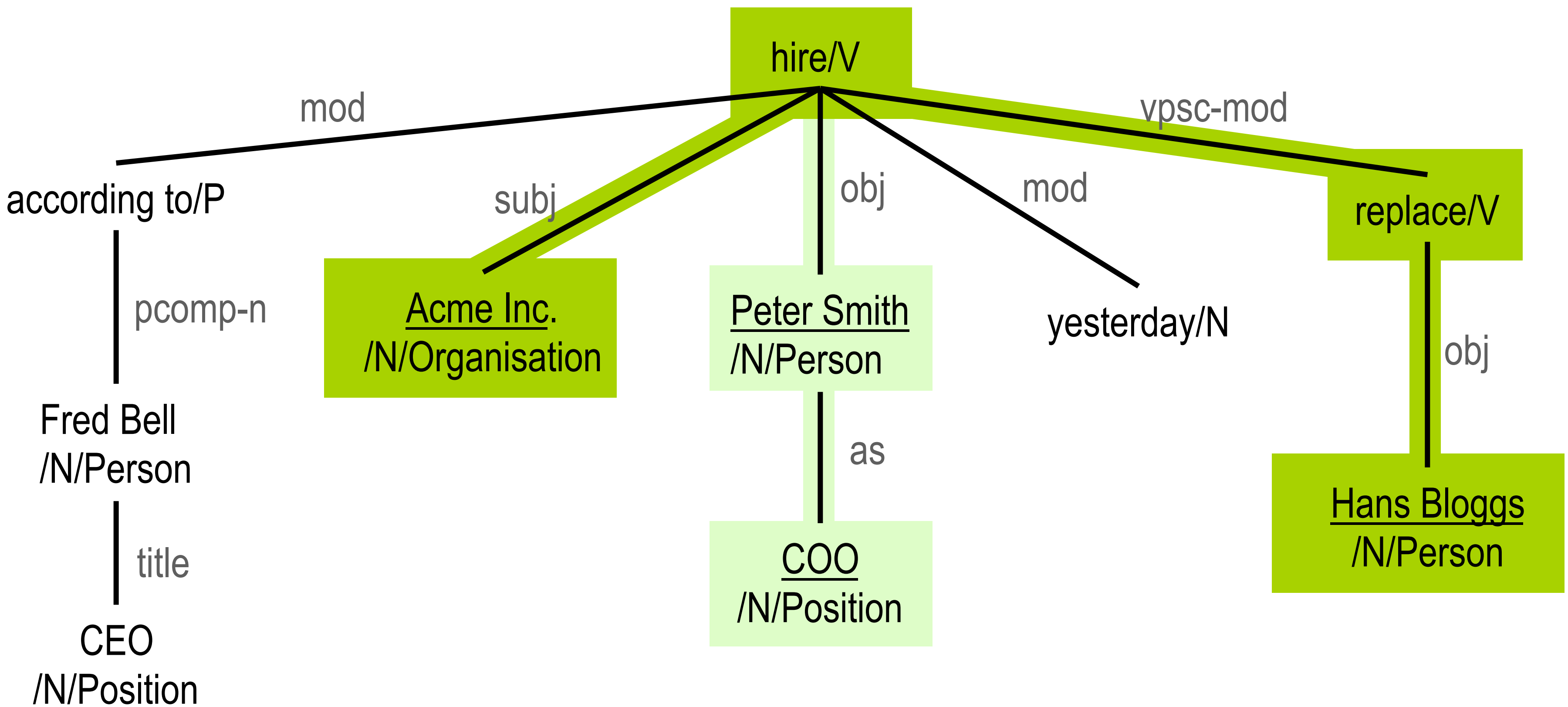
- verb centered
- All chains dominated by a verb, which contain at least one relevant named entity and their combinations



Previous Work: Subtree-Model

Sudo et al. (2003)

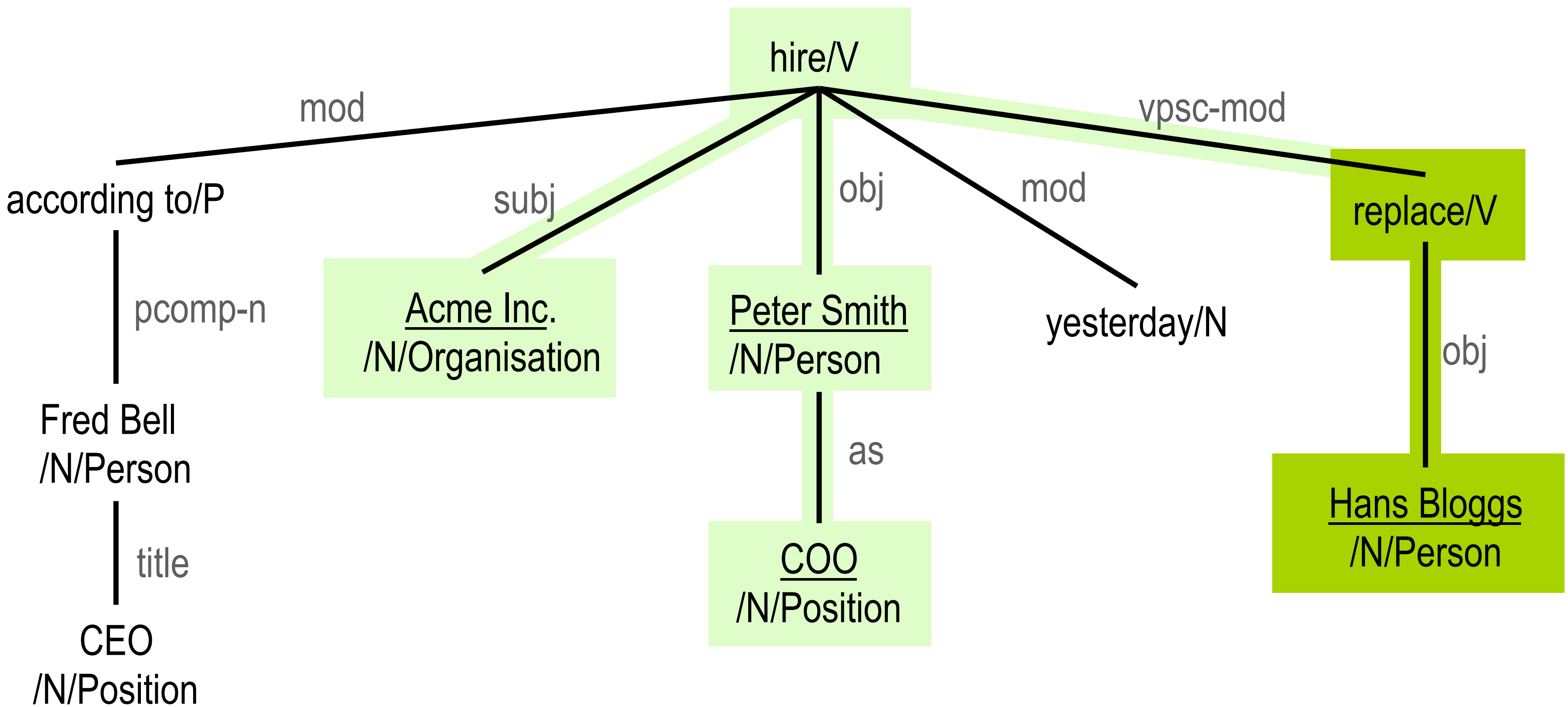
- verb centered
- All chains dominated by a verb, which contain at least one relevant named entity and their combinations



Previous Work: Subtree-Model

Sudo et al. (2003)

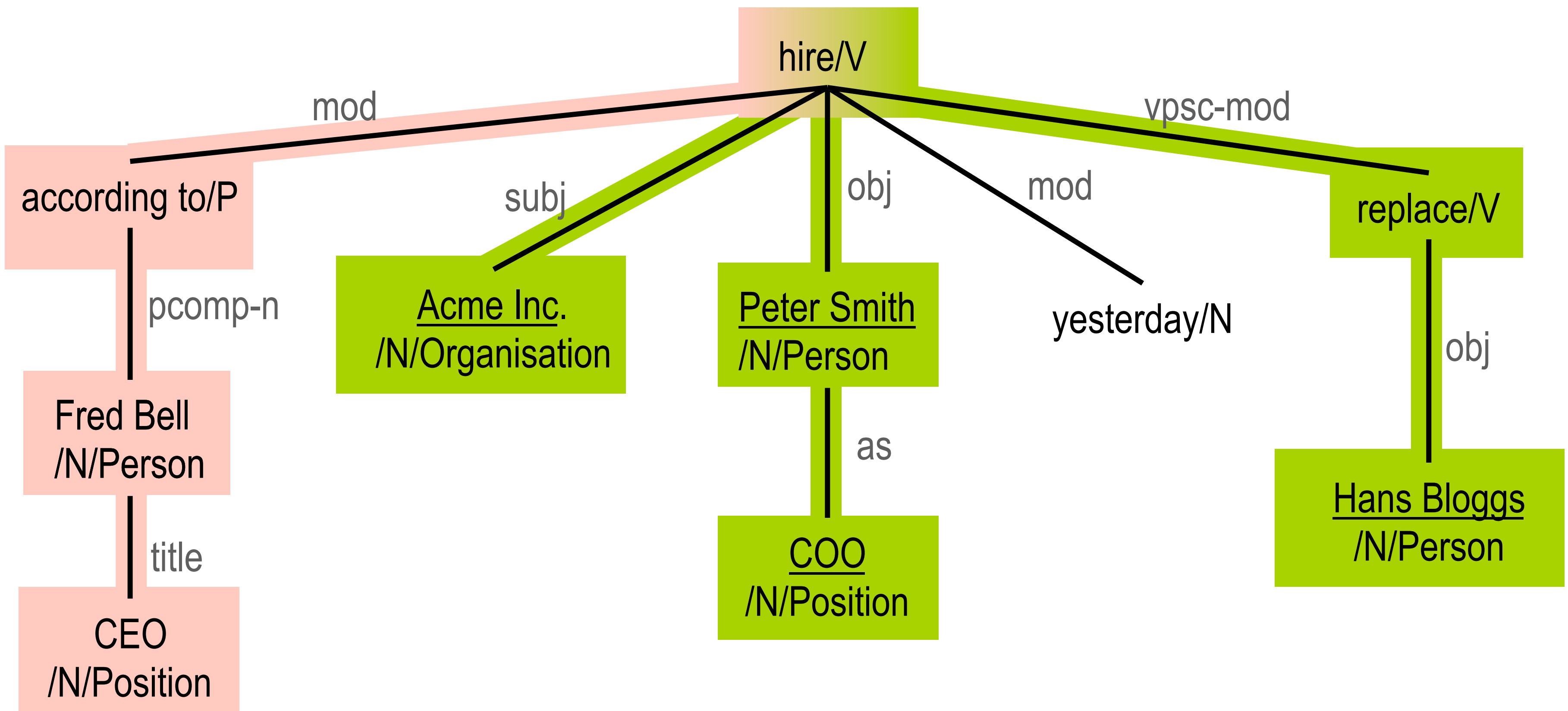
- verb centered
- All chains dominated by a verb, which contain at least one relevant named entity and their combinations



Previous Work: Subtree-Model

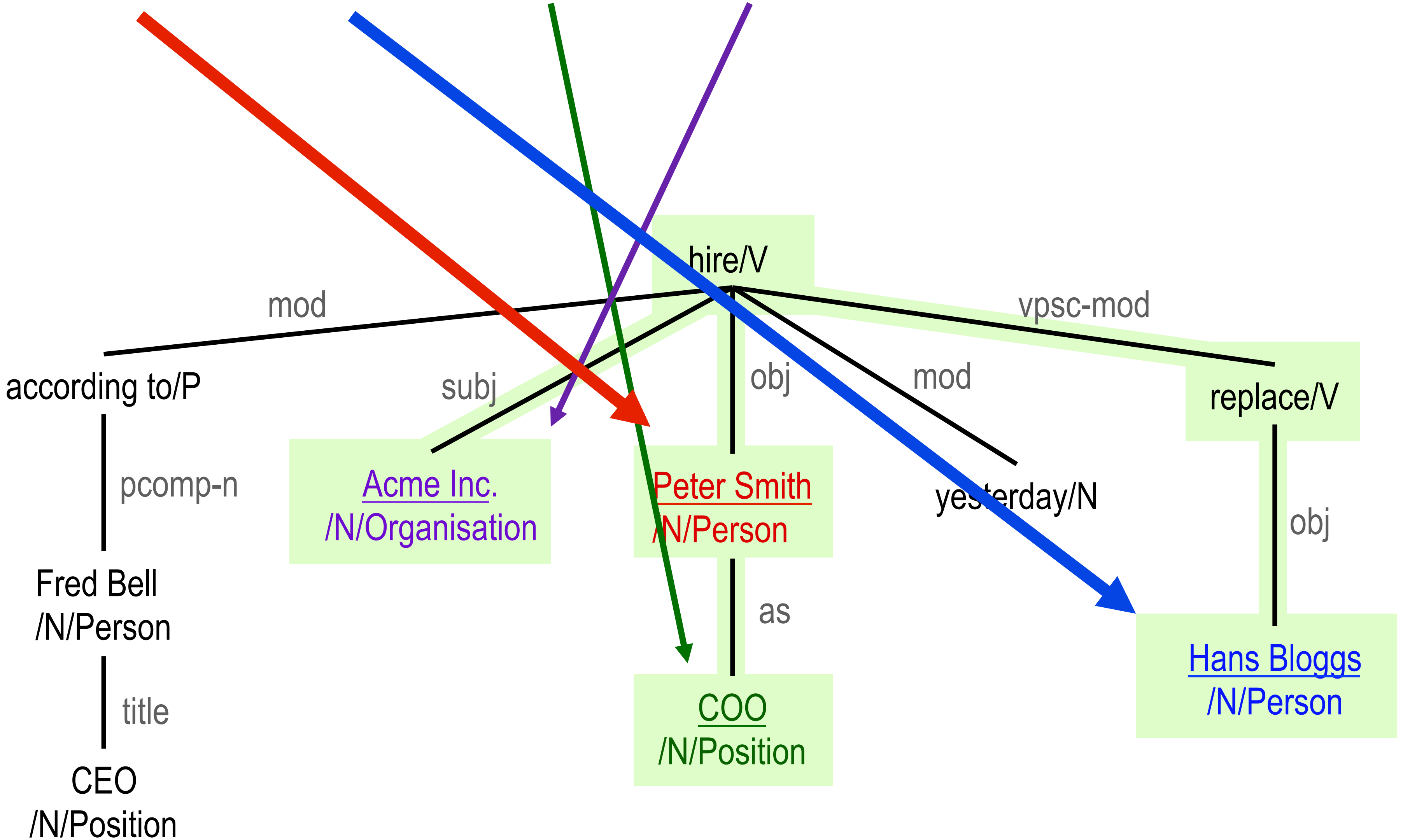
Sudo et al. (2003)

- verb centered
- All chains dominated by a verb, which contain at least one relevant named entity and their combinations



None of the existing models links the detected slot-filling candidates with their respective semantic roles

<person_in, person_out, position, organisation>



- State of the art
- Domain Adaptive Relation Extraction Framework (DARE)
- Experiments and evaluations
- Performance analysis and discussion
- Conclusion and future work

Properties of DARE

- Samples of target relation instances serve as semantic seed
- Systematic treatment of n-ary relations and their projections
- Exploitation of relation projections for pattern discovery
- Bottom-up compositional pattern discovery
- A recursive linguistic rule representation
- Rules contain semantic roles w.r.t. to target relation
- Bottom-up compression method to generalize rules
- Filtering of rule candidates by “domain relevance”

DARE: *Domain Adaptive Relation Extraction*

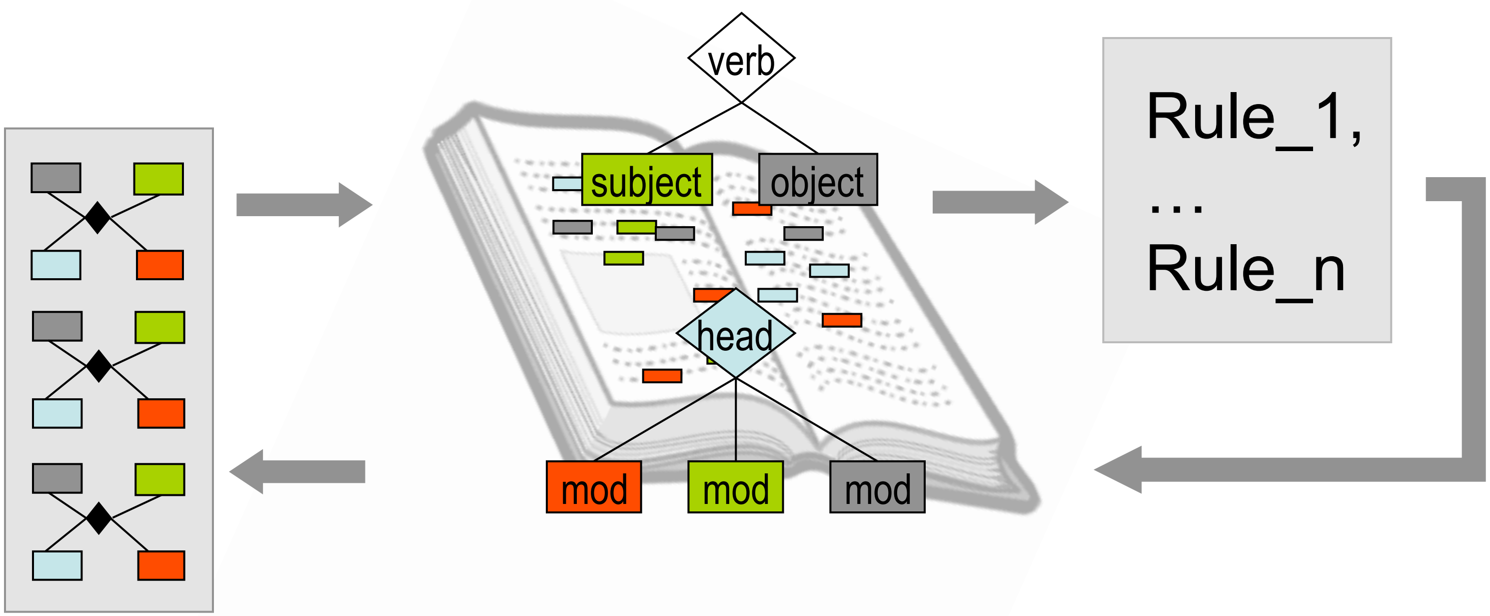
- Samples of target relation instances serve as semantic seed
- Systematic treatment of n-ary relations and their projections
- Exploitation of relation projections for pattern discovery
- Bottom-up compositional pattern discovery
- A recursive linguistic rule representation
- Rules contain semantic roles w.r.t. to target relation
- Bottom-up compression method to generalize rules

Bootstrapping Relation Extraction with Semantic Seed

Adapted from

DIPRE (Brin, 1998) and Snowball (Agichtein & Gravano, 2000)

but extended and enriched with linguistic analysis

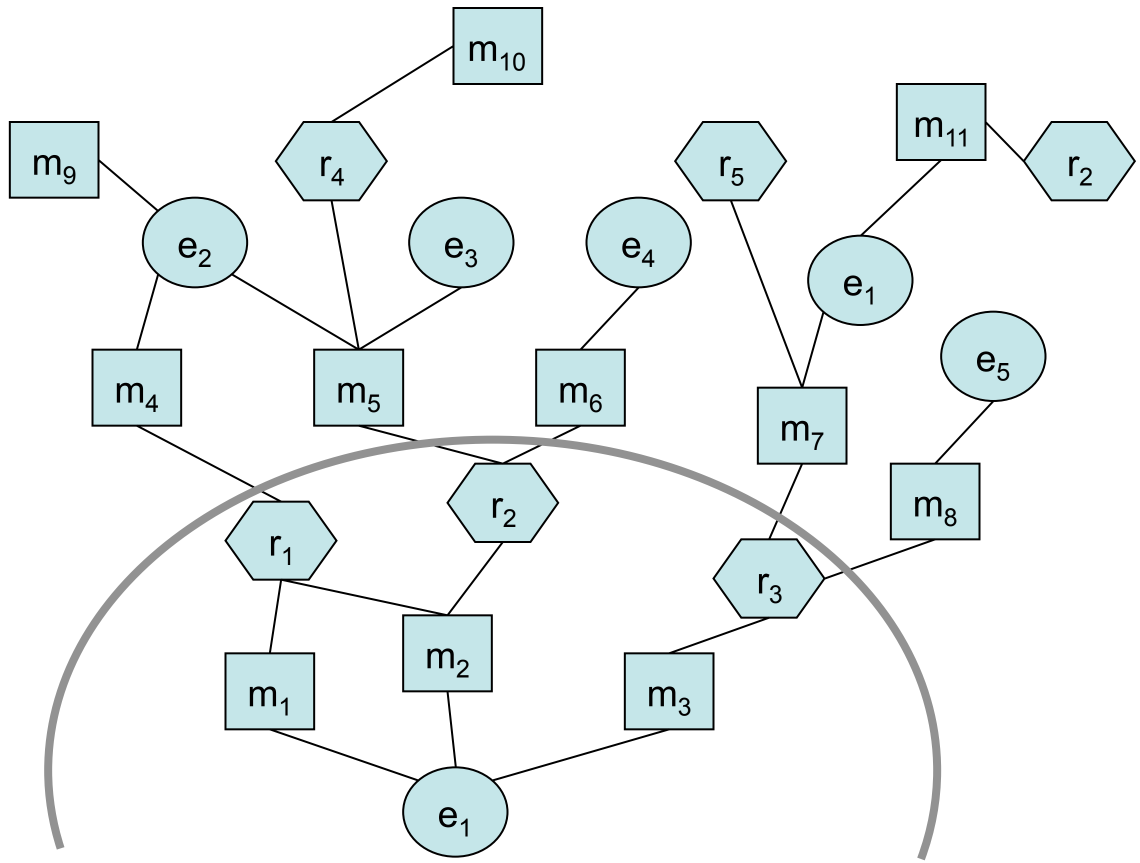


Bootstrapping Relation Extraction with Semantic Seed

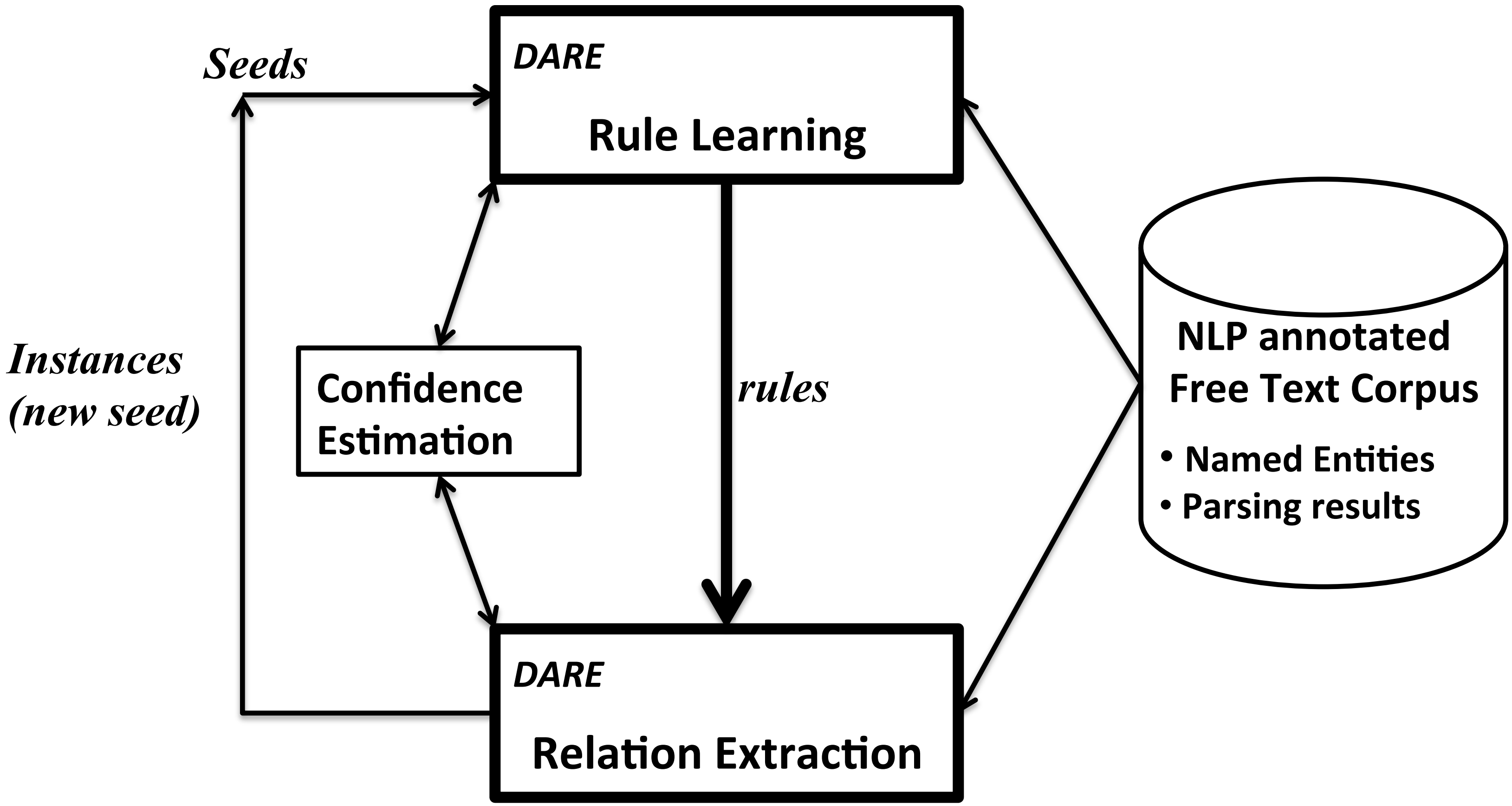
- DIPRE and Snowball
  - binary relations only, no projections, no linguistic analysis
- DARE
  - n-ary relations and their projections, deep linguistic analysis

(in the experiments I use MINIPAR by Dekan Lin 1999)

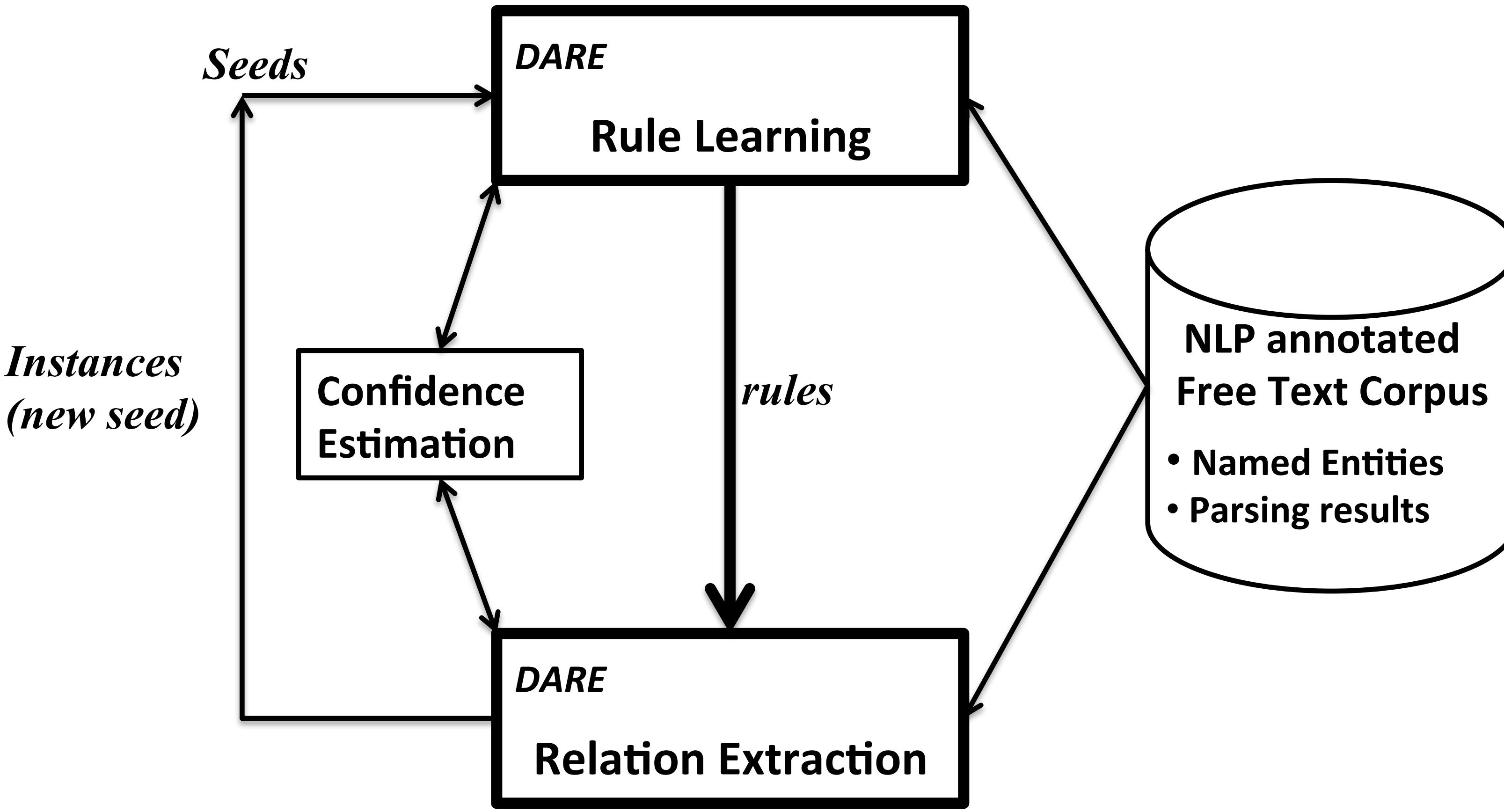
Start of Bootstrapping (simplified)



DARE Architecture

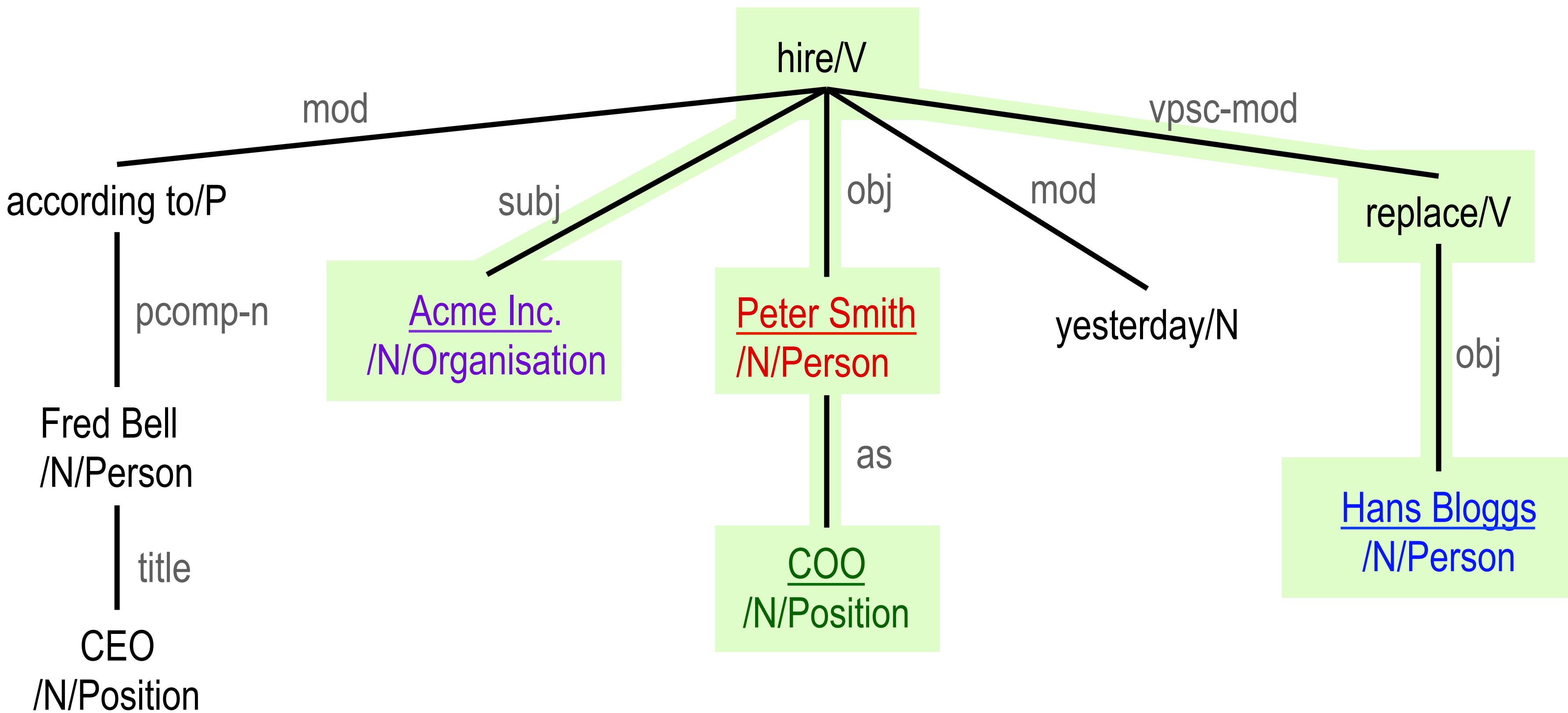


DARE Architecture



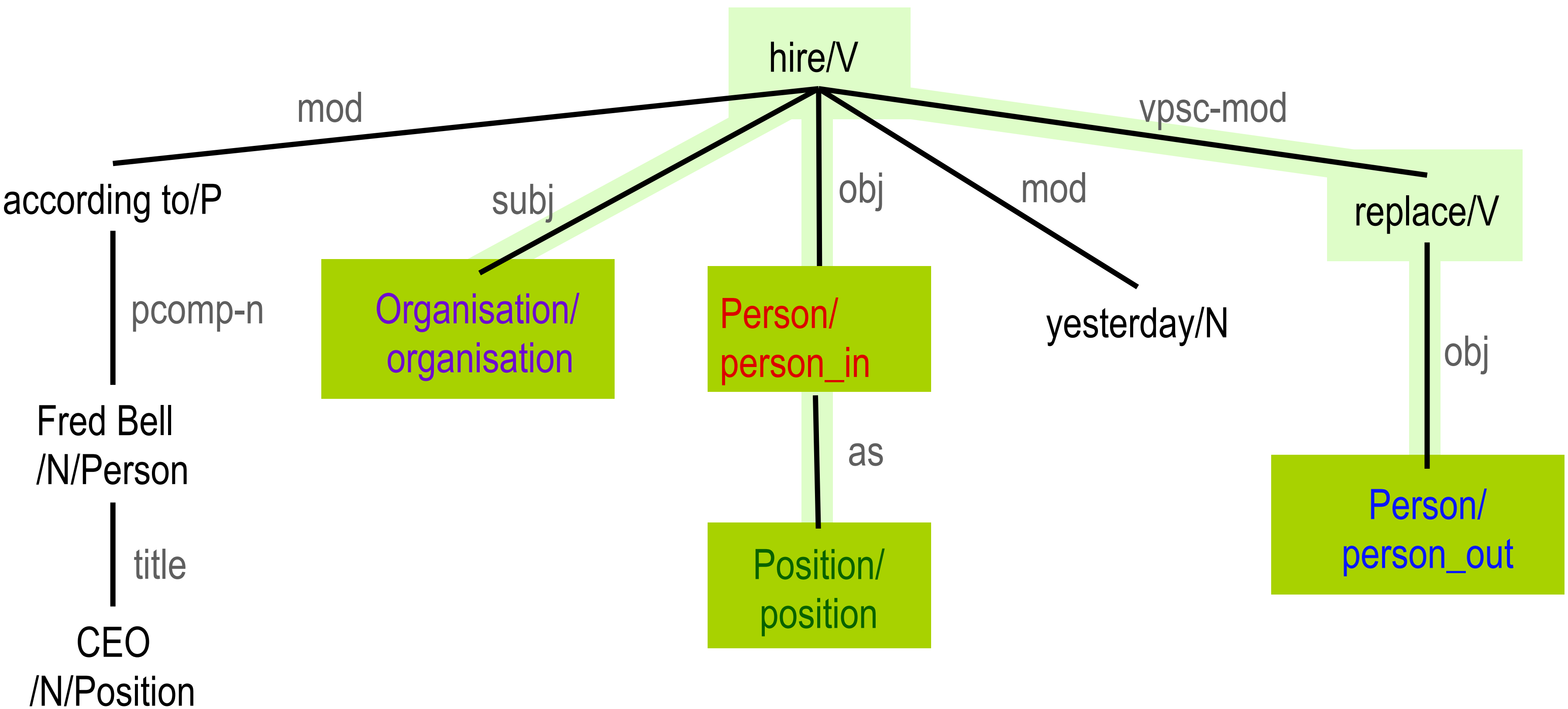
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<Peter Smith/person_in, Hans Bloggs/person_out, COO /position, Acme Inc. /organisation>



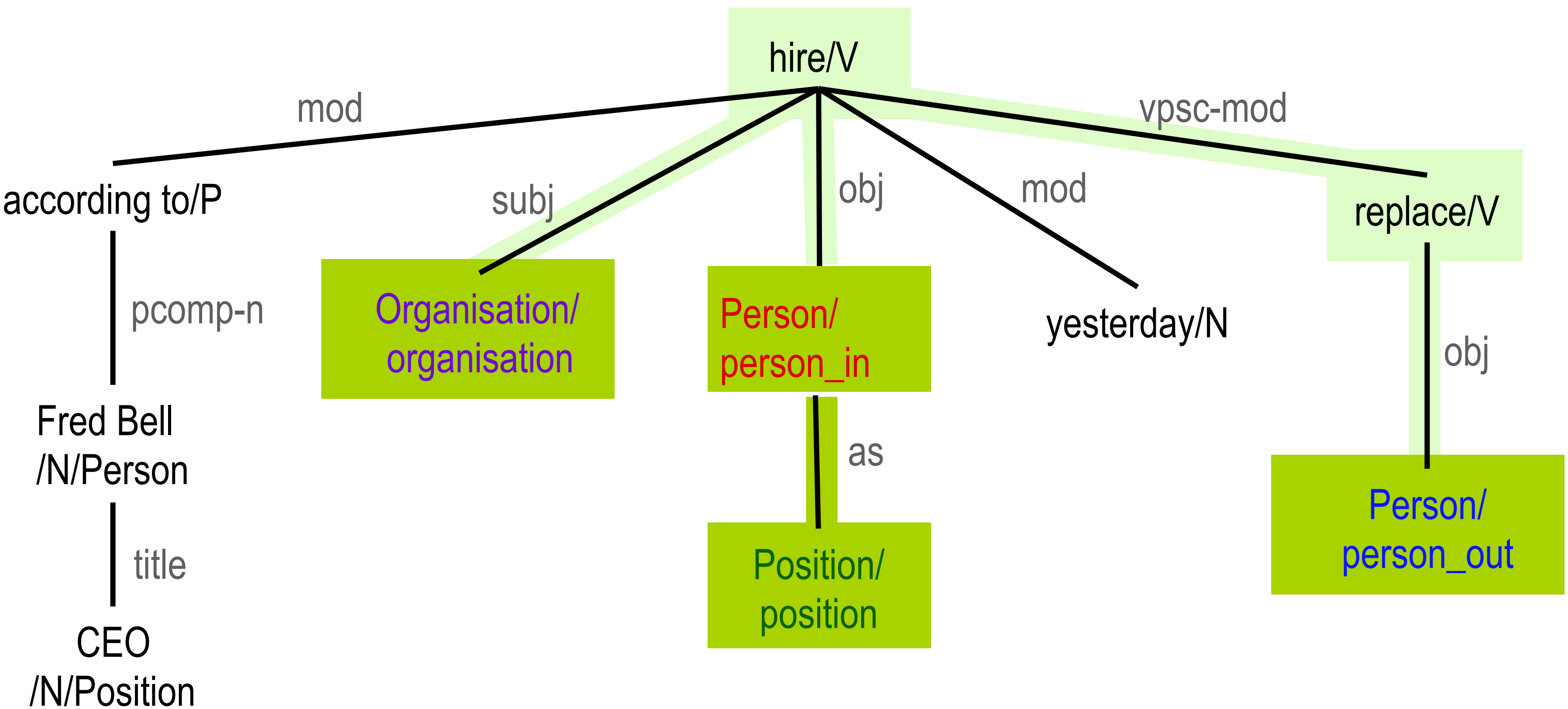
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<Peter Smith/person_in, Hans Bloggs/person_out, COO /position, Acme Inc. /organisation>



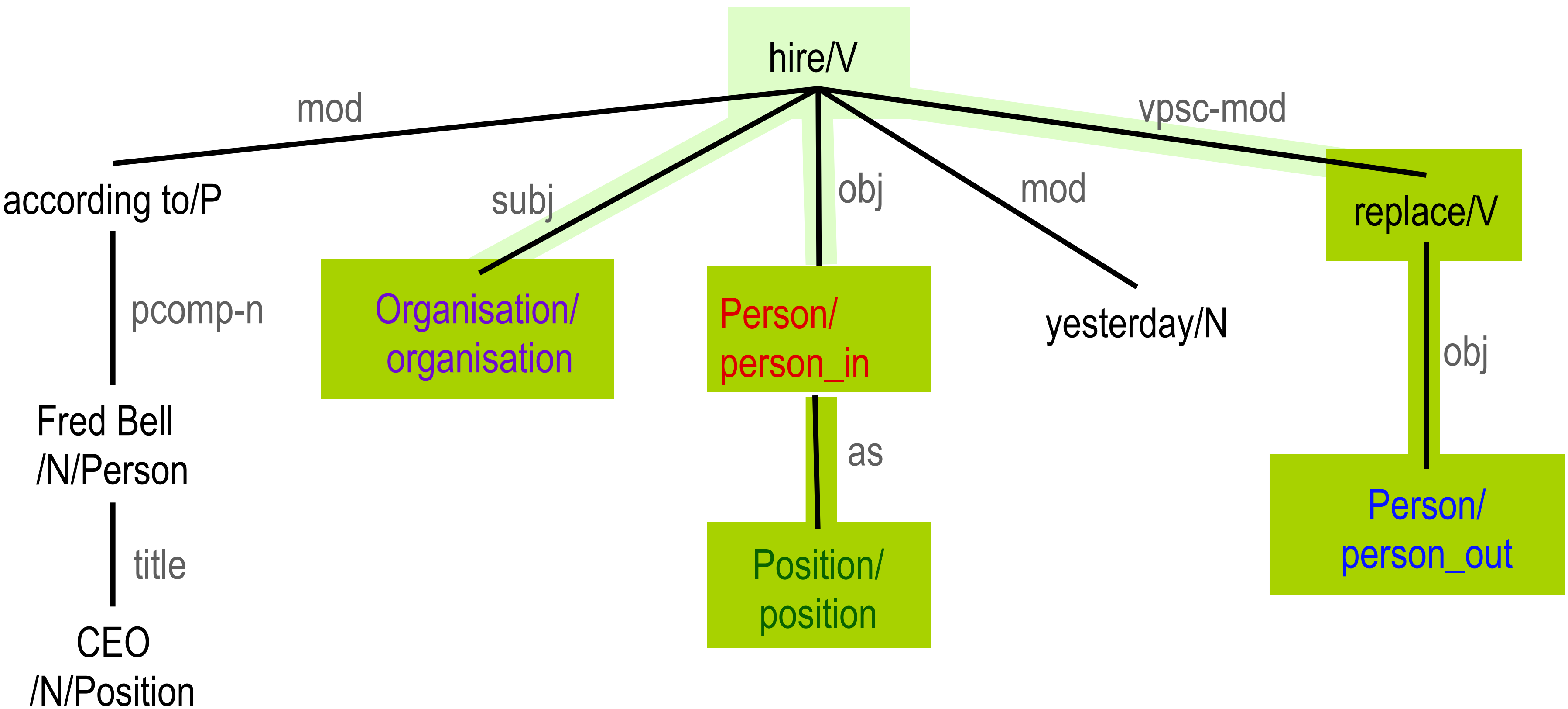
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<Peter Smith/person_in, Hans Bloggs/person_out, COO /position, Acme Inc. /organisation>



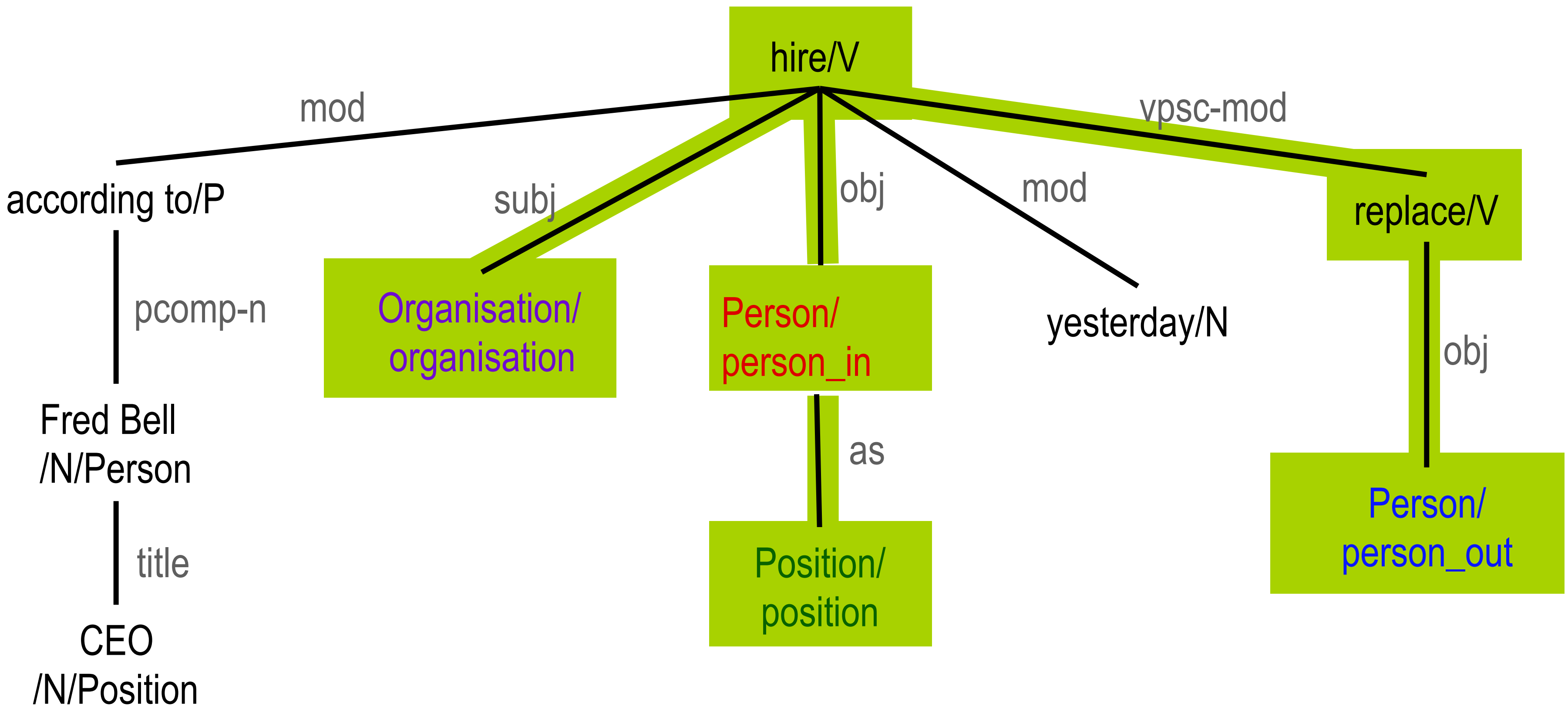
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

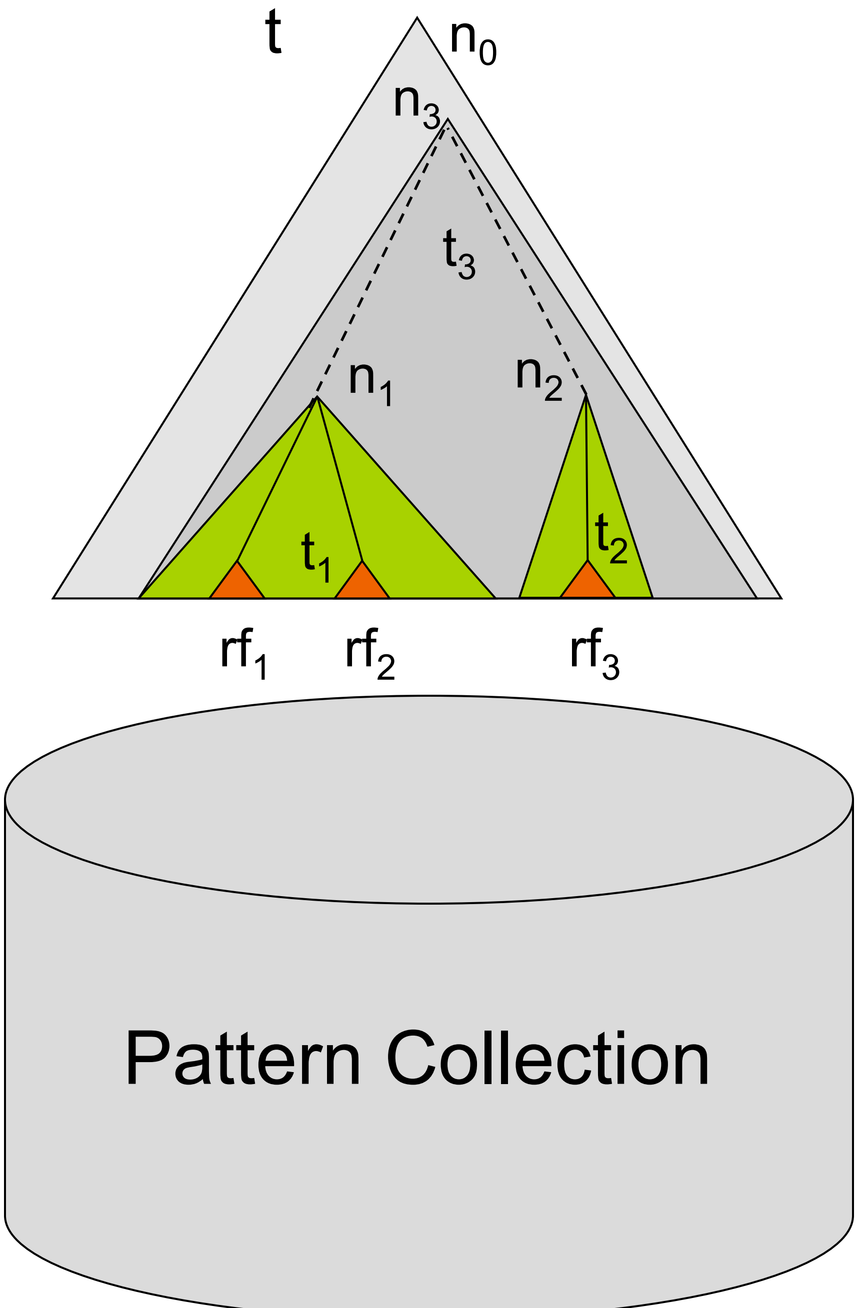
<Peter Smith/person_in, Hans Bloggs/person_out, COO /position, Acme Inc. /organisation>



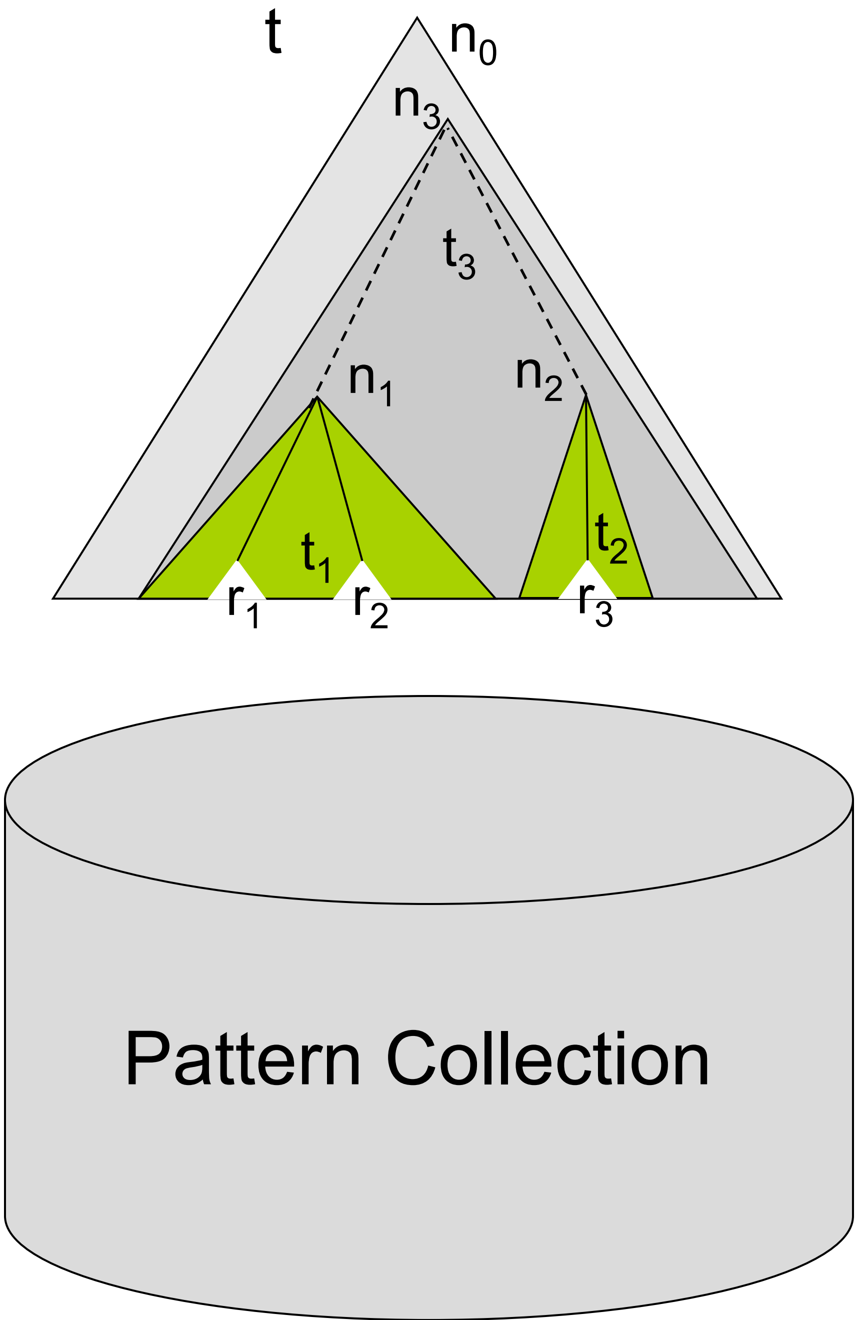
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<Peter Smith/person_in, Hans Bloggs/person_out, COO /position, Acme Inc. /organisation>

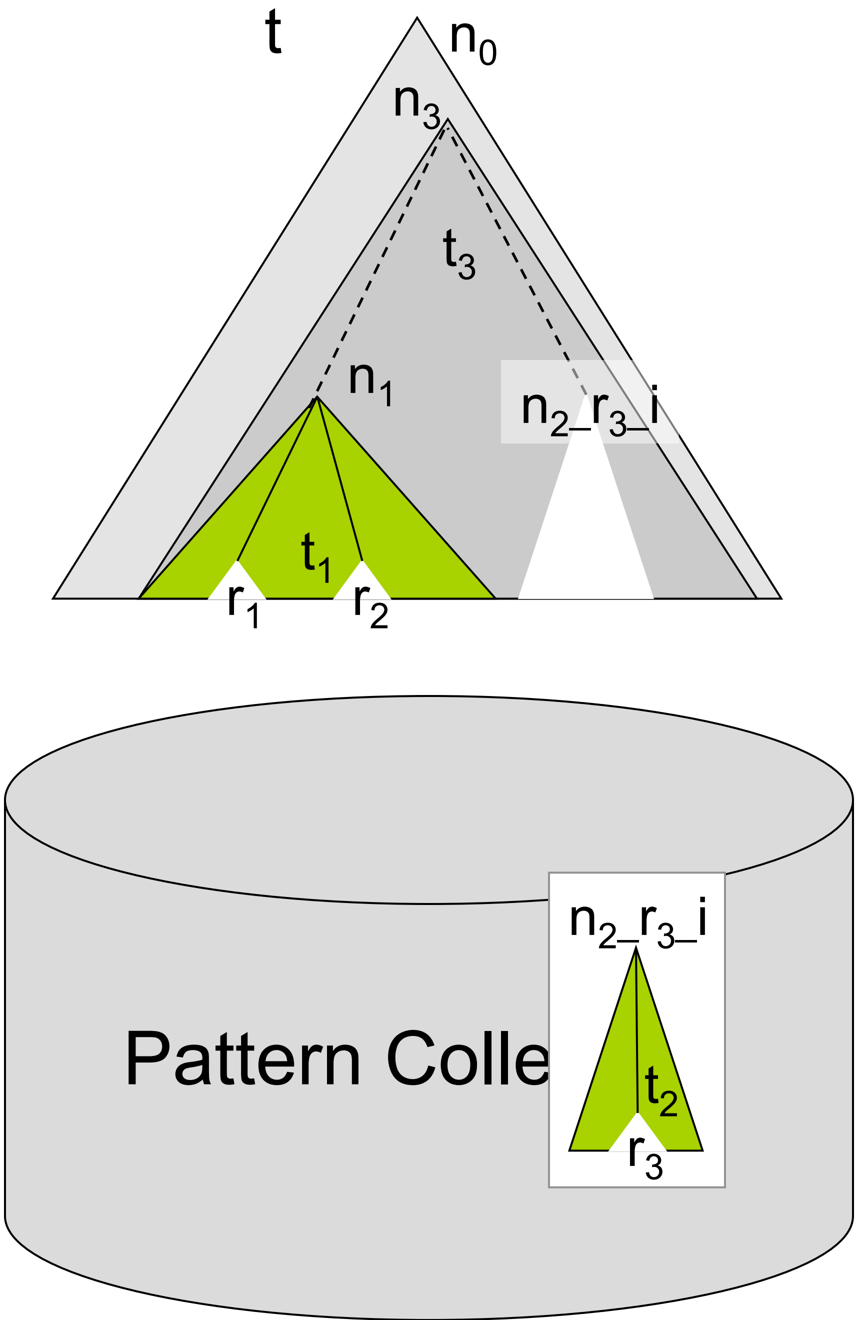




0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;



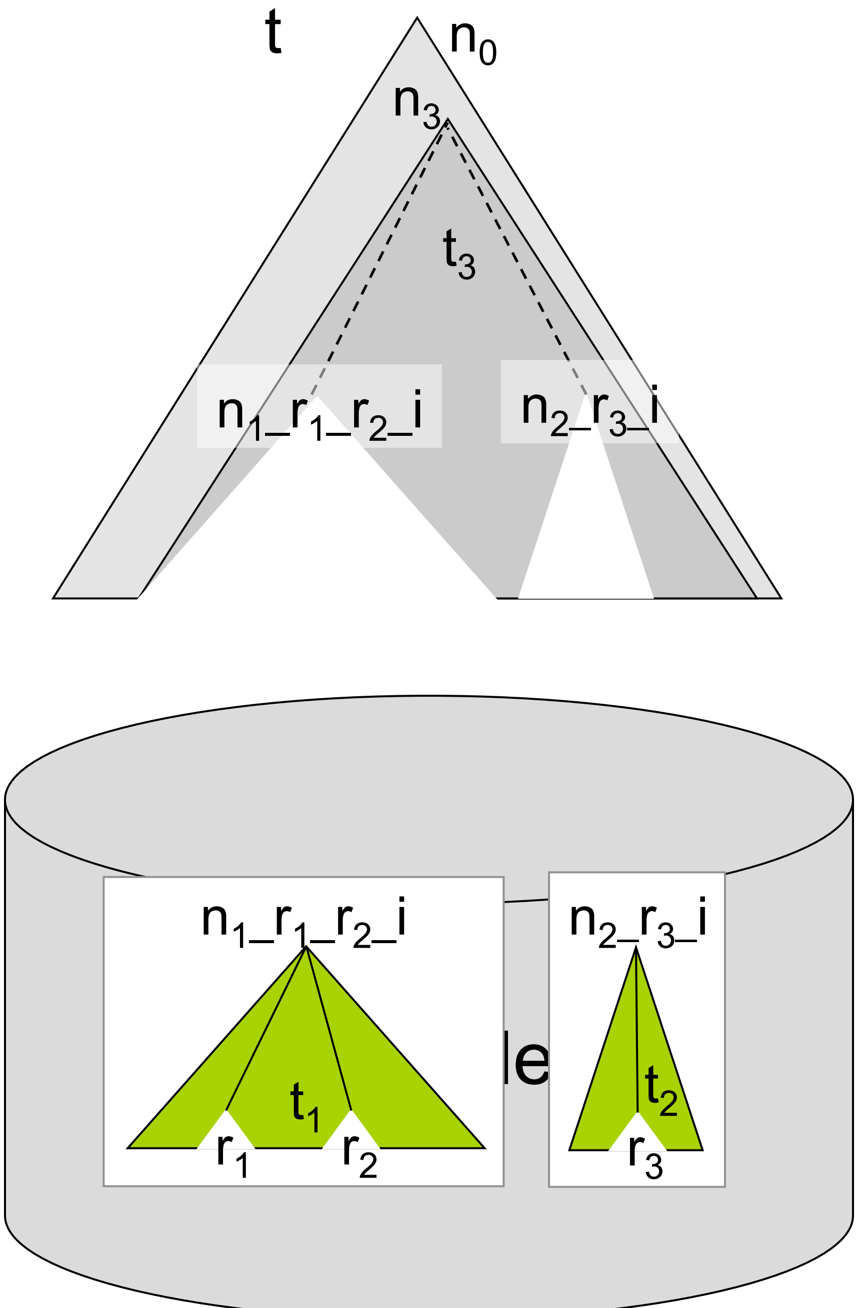
0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;



0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;

for i=1 to n

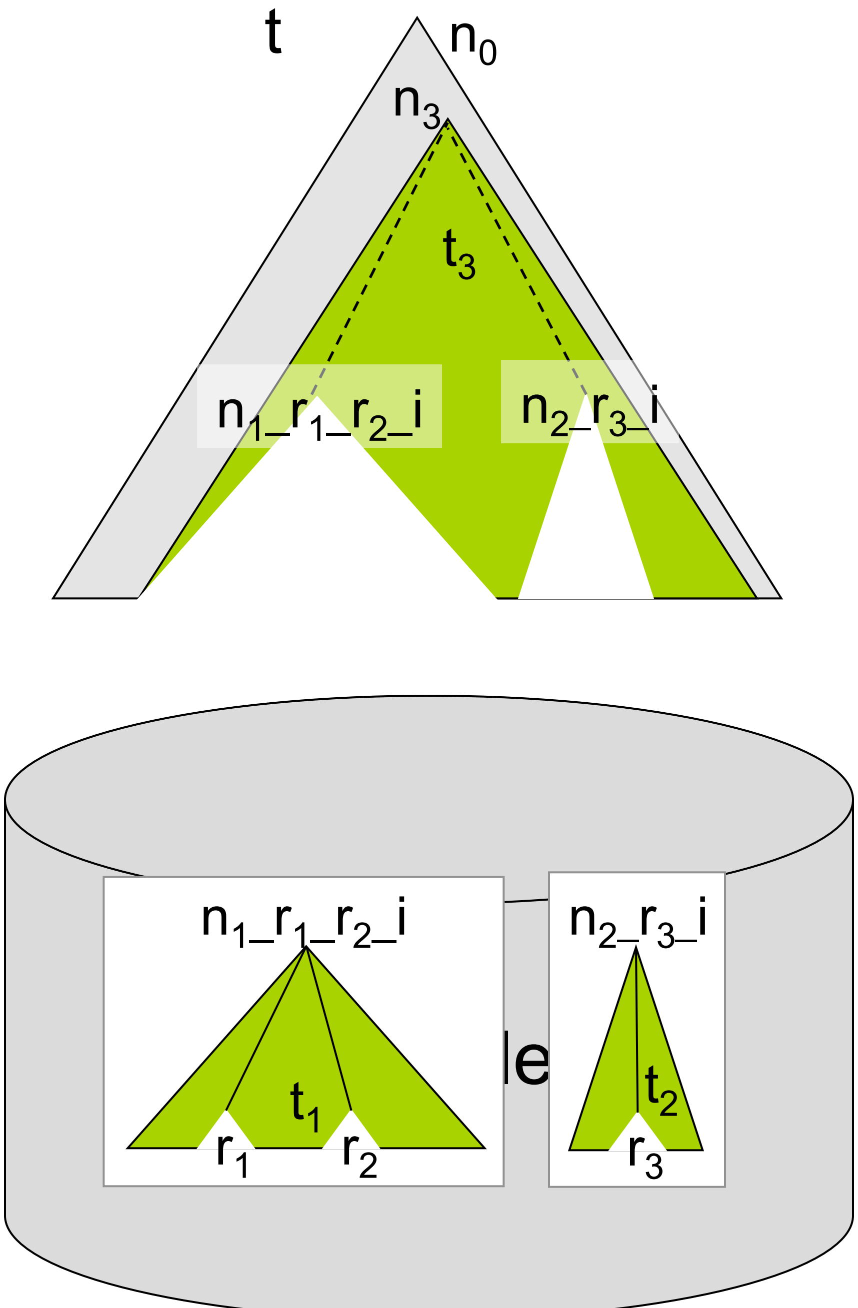
1. identify the set of the lowest non-terminal nodes N_1 in t that dominate i arguments (possibly among other nodes).
2. substitute N_1 by nodes labelled with the seed argument roles and their entity classes
3. prune the subtrees dominated by N_1 from t and add these subtrees into the pattern collection. These subtrees are assigned the argument role information and a unique id.



0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;

for i=1 to n

1. identify the set of the lowest non-terminal nodes N_1 in t that dominate i arguments (possibly among other nodes).
2. substitute N_1 by nodes labelled with the seed argument roles and their entity classes
3. prune the subtrees dominated by N_1 from t and add these subtrees into the pattern collection. These subtrees are assigned the argument role information and a unique id.



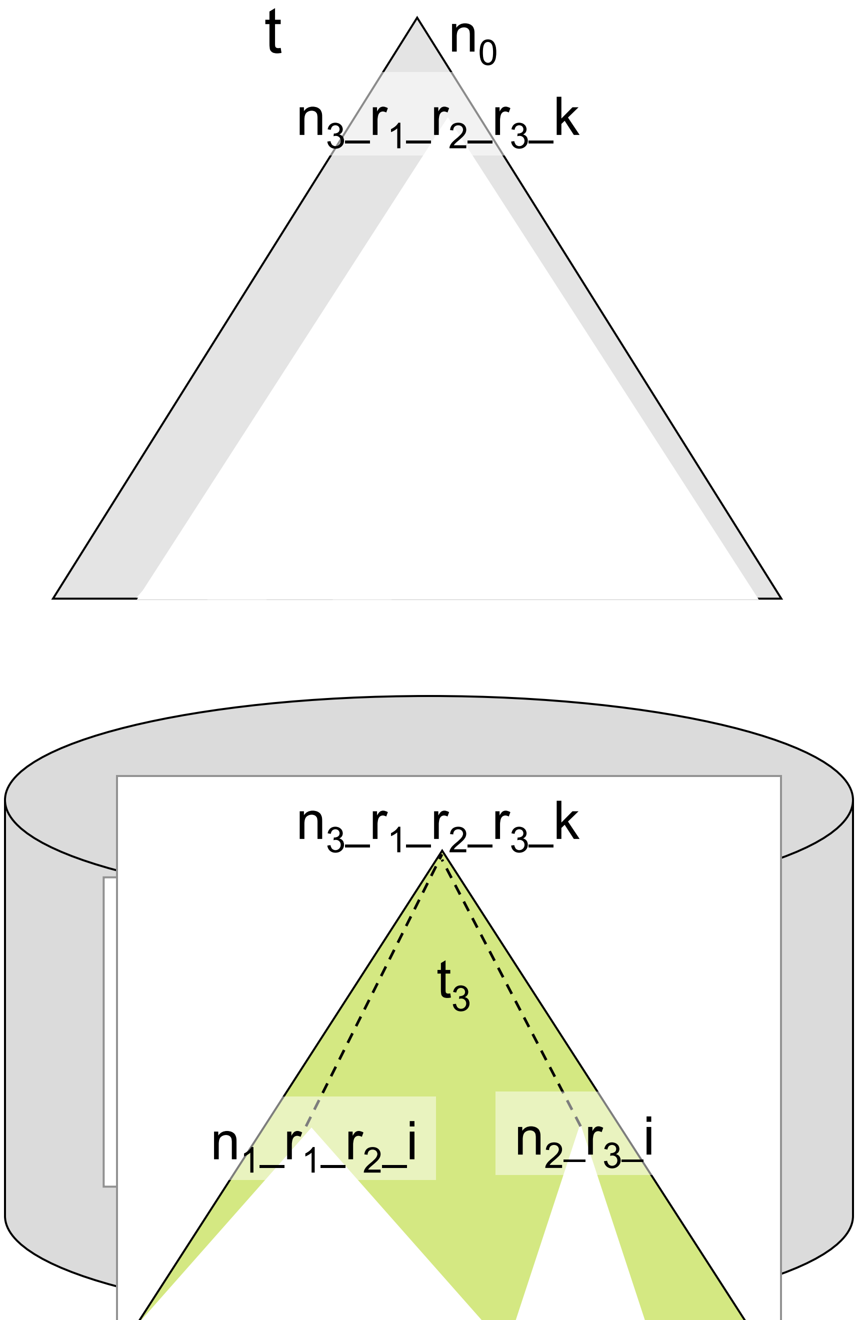
0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;

for i=1 to n

1. identify the set of the lowest non-terminal nodes N_1 in t that dominate i arguments (possibly among other nodes).

2. substitute N_1 by nodes labelled with the seed argument roles and their entity classes

3. prune the subtrees dominated by N_1 from t and add these subtrees into the pattern collection. These subtrees are assigned the argument role information and a unique id.



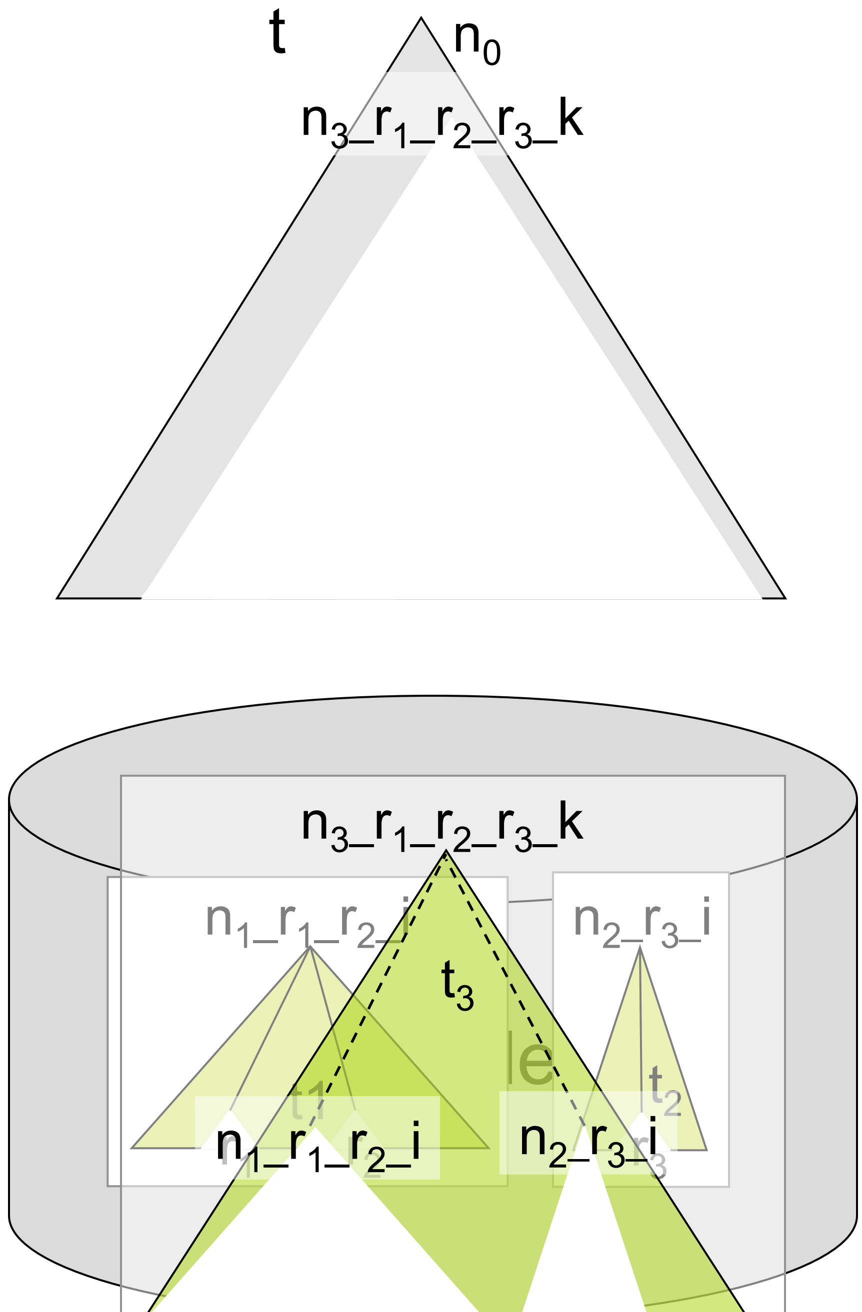
0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;

for i=1 to n

1. identify the set of the lowest non-terminal nodes N_1 in t that dominate i arguments (possibly among other nodes).

2. substitute N_1 by nodes labelled with the seed argument roles and their entity classes

3. prune the subtrees dominated by N_1 from t and add these subtrees into the pattern collection. These subtrees are assigned the argument role information and a unique id.



0. replace all nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;

for i=1 to n

1. identify the set of the lowest non-terminal nodes N_1 in t that dominate i arguments (possibly among other nodes).

2. substitute N_1 by nodes labelled with the seed argument roles and their entity classes

3. prune the subtrees dominated by N_1 from t and add these subtrees into the pattern collection. These subtrees are assigned the argument role information and a unique id.

Example in Nobel Prize Award Domain

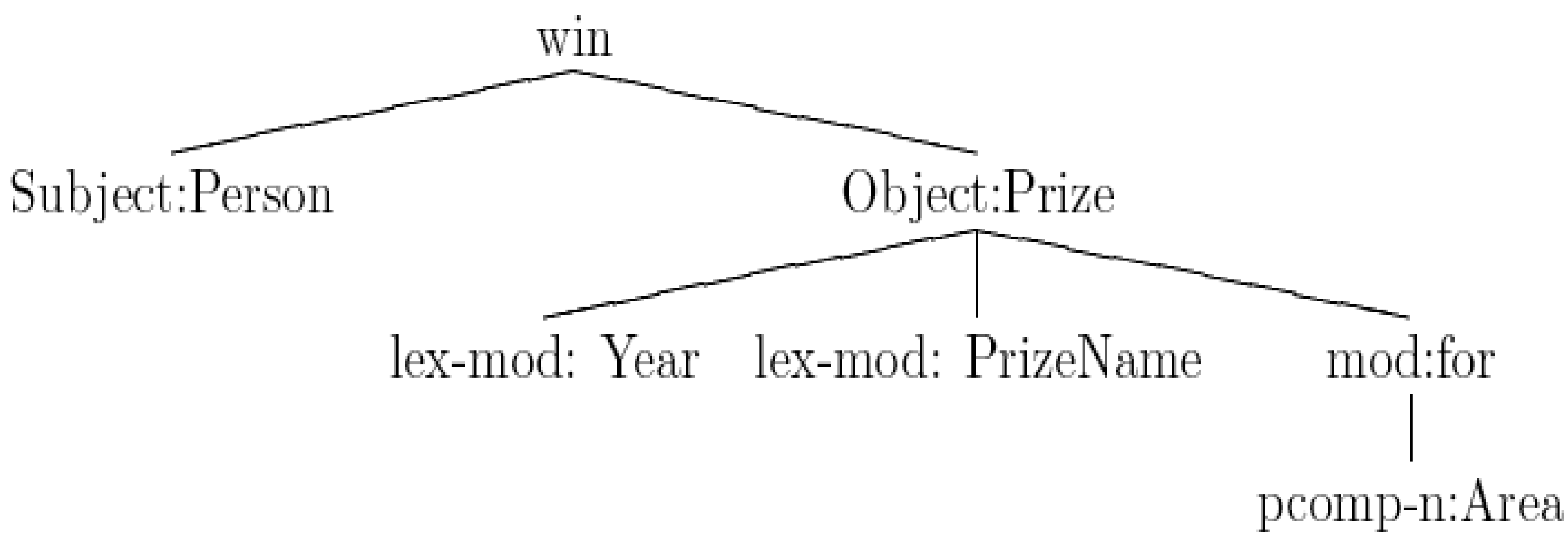
- Seed example

<Mohamed ElBaradei, Nobel, Peace, 2005>

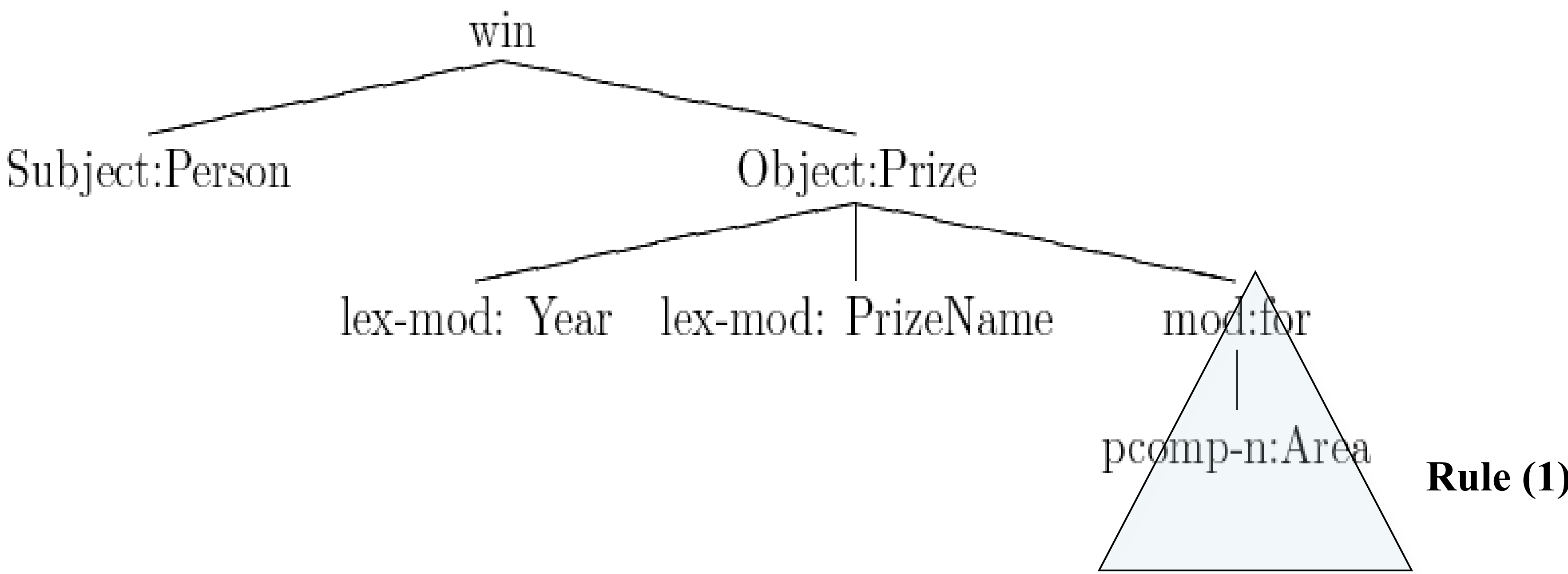
- Sentence matched with the seed

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

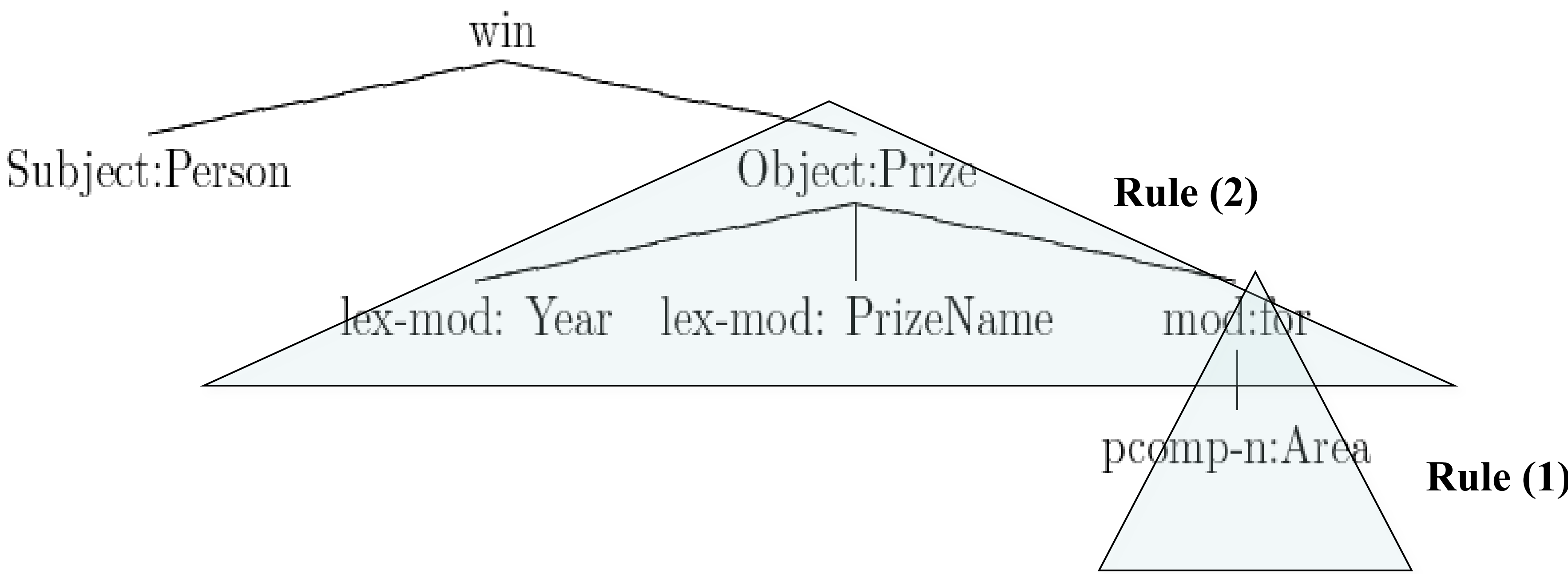
Dependency Parse Result



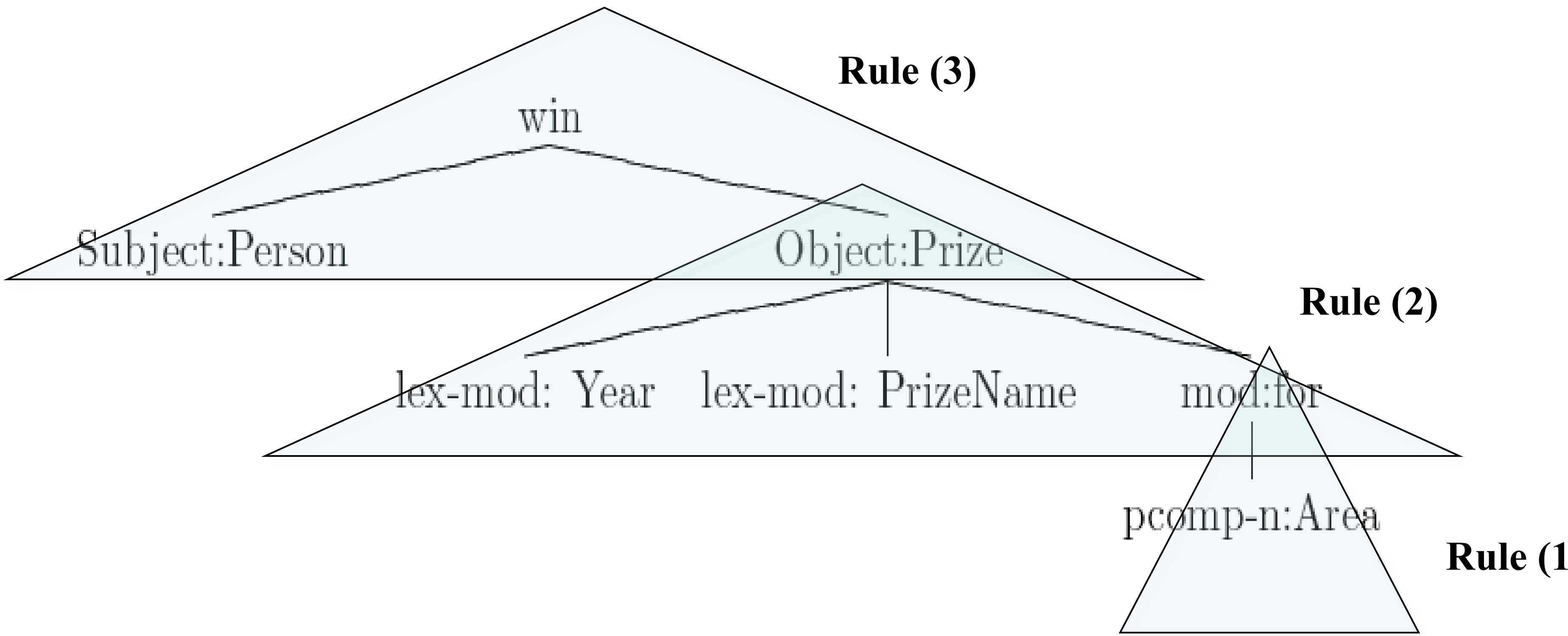
Bottom Up Rule Learning



Bottom Up Rule Learning



Bottom Up Rule Learning



Rule (1)

2005 Nobel Prize for Peace

Rule name:: area_1

$$\texttt{Rule body::}\ \begin{bmatrix} \text{head} & \begin{bmatrix} \text{pos} & \text{preposition} \\ \text{lex-form} & \text{``for''} \end{bmatrix} \\ \text{daughters} & \left\langle \begin{bmatrix} \text{pcomp-n} & \begin{bmatrix} \text{head} & \boxed{1}\ \text{Area} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

Output:: $\langle \boxed{1} Area \rangle$

Rule (2)

2005 Nobel Prize <for Peace>

`Rule name::  year_prize_area_1`

`Rule body::`

$$\begin{bmatrix} \text{head} & \begin{bmatrix} \text{pos} & \text{noun} \\ \text{lex-form} & \text{“prize”} \end{bmatrix} \\ \text{daughters} & \left\langle \begin{bmatrix} \text{lex-mod} & \begin{bmatrix} \text{head} & \boxed{1}\ \text{Year} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{lex-mod} & \begin{bmatrix} \text{head} & \boxed{2}\ \text{Prize} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{mod} & \begin{bmatrix} \text{head} & \begin{bmatrix} \text{pos} & \text{preposition} \\ \text{lex-form} & \text{“for”} \end{bmatrix} \\ \text{rule} & \text{area\_1::}\ \langle \boxed{3} Area \rangle \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

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

Rule (3)

`Rule name::  recipient_prize_area_year_1`

`Rule body::`

$$
\begin{bmatrix}
\text{head} & \begin{bmatrix} \text{pos} & \text{verb} \\ \text{mode} & \text{active} \\ \text{lex-form} & \text{“win”} \end{bmatrix} \\
\text{daughters} & \left\langle \begin{bmatrix} \text{subject} & \begin{bmatrix} \text{head} & \boxed{1}\ \text{Person} \end{bmatrix} \end{bmatrix}, \right. \\
 & \left. \begin{bmatrix} \text{object} & \begin{bmatrix} \text{head} & \begin{bmatrix} \text{pos} & \text{noun} \\ \text{lex-form} & \text{“prize”} \end{bmatrix} \\ \text{rule} & \text{year\_prize\_area\_1::}\ \langle \boxed{4} Year, \boxed{2} Prize, \boxed{3} Area \rangle \end{bmatrix} \end{bmatrix} \right\rangle
\end{bmatrix}
$$

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

DARE Rule Components

`Rule name:: recipient_prize_area_year_1`

`Rule body::`

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

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

Two Domains

- 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

Experiments

- 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	3328	18.4 MB
MUC-6	199	1MB

Relation Extraction without Coreference Resolution

domain	data size	initial seed no.	precision	recall
Nobel Prize	18.4 MB	1	86.5%	50.7%
MUC-6	1 MB	55	62%	48%

Management Succession Domain

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

The Dream

- 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.
- This sounds too good to be true!

Research Questions

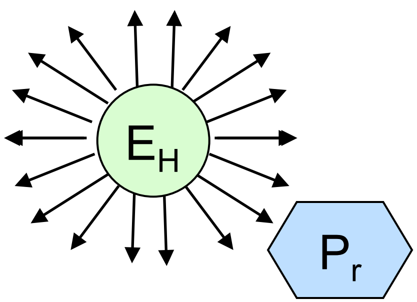
As scientists we want to know

- 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?

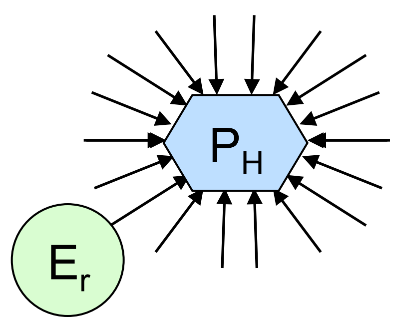
Careful analysis confirmed the following assumption:

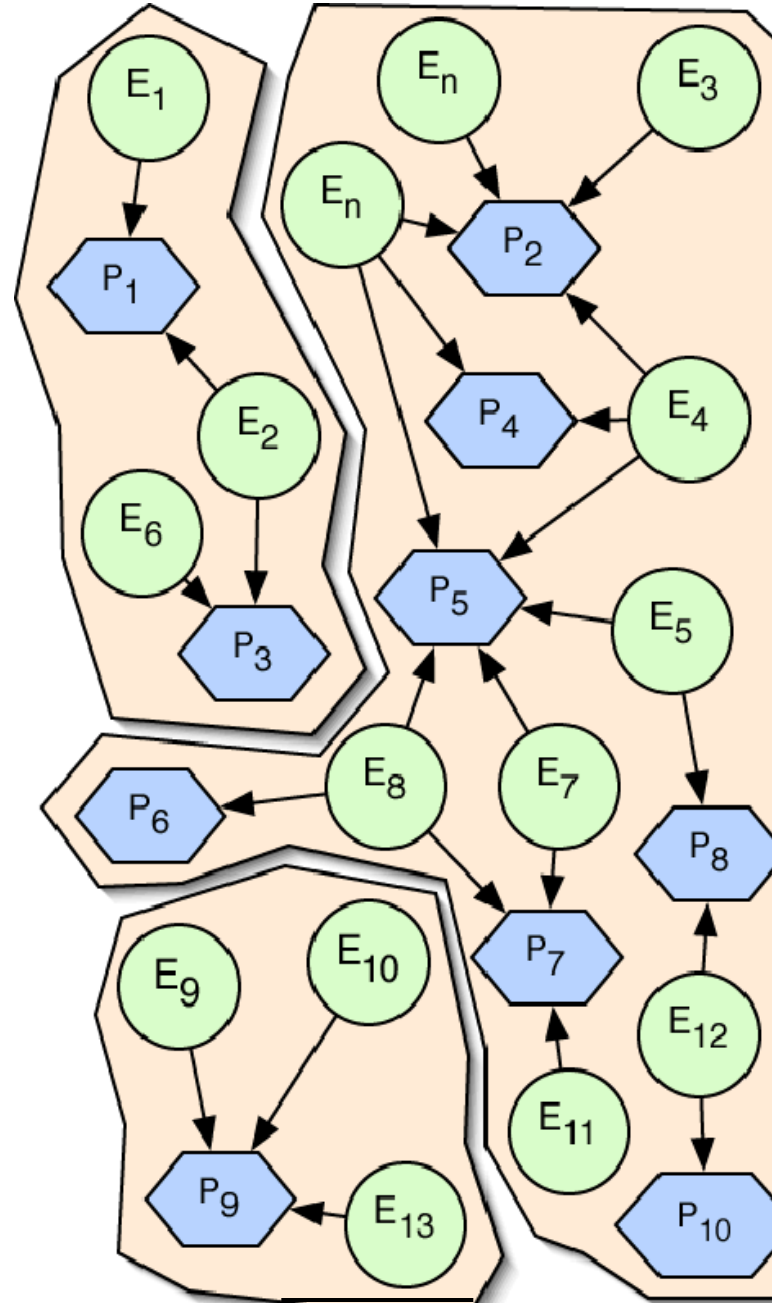
redundancy, both on patterns and event mentions, helps.

Frequently reported events make rare patterns reachable

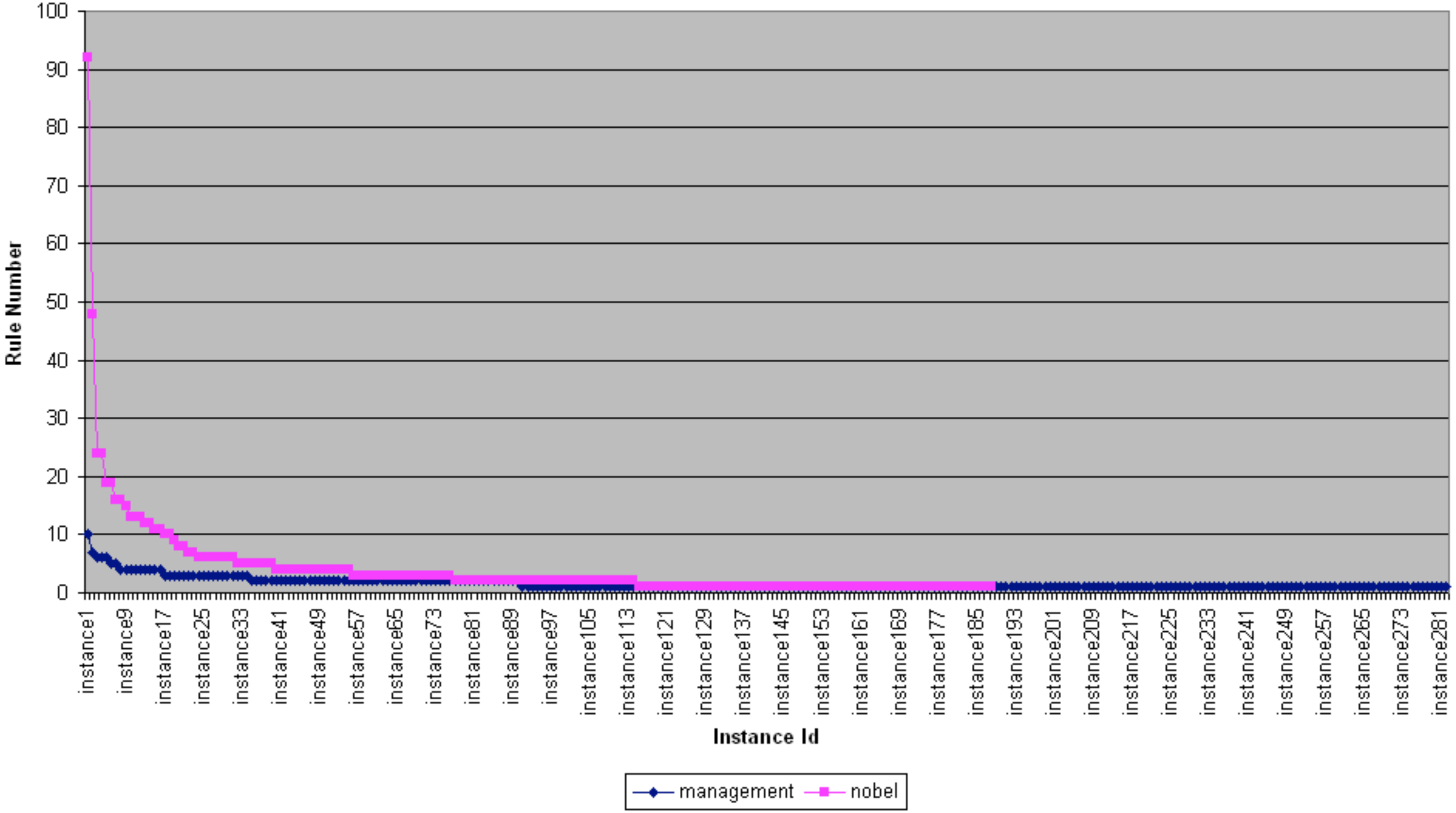


Popular patterns help to reach rarely mentioned events

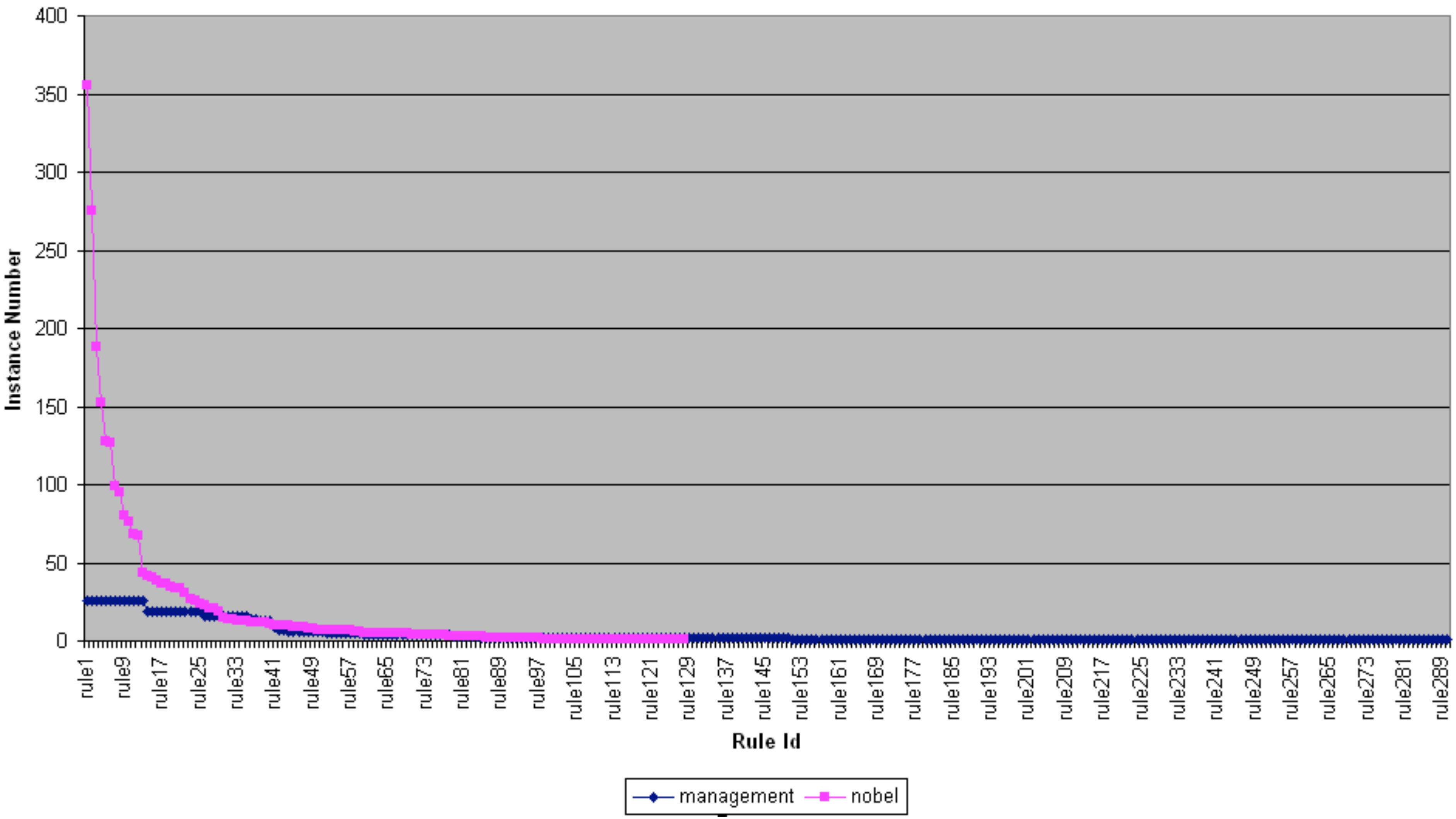




Instance to Pattern
Nobel Prize vs. Management Succession



Rule to Instances
(Nobel Prize vs. Management Succession)



Insights

- Results from graph theory help to understand the requirements on data.

  Example: small world property

- For data sets with continents and islands, we can sometimes exploit additional data or auxiliary domains to bridge the islands by learning rare patterns.

  Example: use of Nobel prize domain for learning patterns for events concerning less popular prizes (many other prizes could be detected)

Conclusion

- DARE is the first approach to combine the idea of bootstrapping IE systems with a linguistic grammar

- This can be illustrated by a simple formula:

    reusable generic linguistic knowledge
  + raw data
  + a few examples (seed)
  = domain specific relation extraction grammar

- In addition to the obvious practical advantages, the approach offers theoretical benefits: It supports a view of IE as a systematic gradual approximation of language understanding.

Overcoming Obstacles

- **Obstacles to Recall**
  - **missing bridges between islands/continents ....use of auxiliary data**
  - **overly specific rules................................ better rule generalization**
  - **spread over several sentences**
    - **missing coreferences.......................... coreference resolution**
- **Obstacles to Precision**
  - **intrusion of other relations........................ learning of negative rules**
  - **modality contexts.................................. learning of negative rules**
  - **Integration of more linguistic context and structures ...... deep NLP**

Reality in IE Projects

- Our IE users are often not domain experts
- IE experts have to develop methods and strategies for
  - Prospecting a domain
  - Proposing relevant relations
  - Finding relevant and suitable data

Task-Driven Anaphora Resolution

Example

1. Three of the Nobel Prizes for Chemistry during the first decade **were awarded** for pioneering work in organic chemistry.
2. In **1902** Emil Fischer (1852-1919), then in Berlin, **was given** the prize for his work on sugar and purine syntheses.
3. Another major influence from organic chemistry was the development of the chemical industry, and a chief contributor here was Fischer's teacher, Adolf von Baeyer (1835-1917) in Munich, who **was awarded the prize** in **1905**.

Anaphora in Texts

He/The scientist won the 2005 Nobel Prize for Peace on Friday for his efforts to limit the spread of atomic weapons.

↓

<?PERSON, Nobel, Peace, 2005>

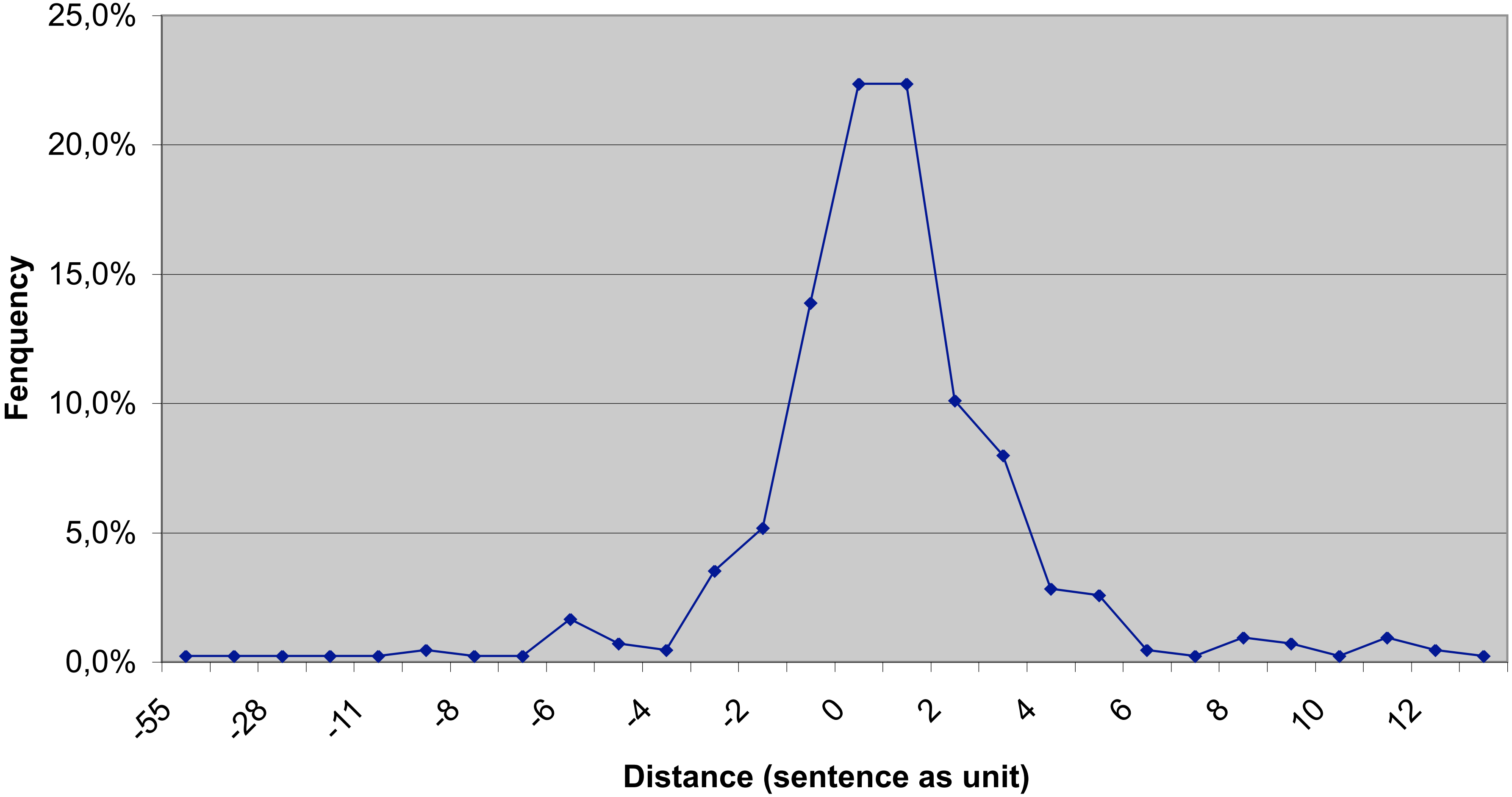
Coreference Relations and Indicators

- Complex linguistic phenomena, influenced by lexical, syntactic, semantic and discourse constraints

- The indicators shared by many approaches are
  - Distance: coreference expressions are often close to each other in the surface structure;
  - Syntactic: pronominal resolution constraints within sentence
  - Semantic: same or compatible semantic category, agreement in number, gender and person;
  - Discourse: parallelism, repetition, apposition, name alias.

Receny Indicator in Nobel Prize Domain

- News reports from New York Times, online BBC and CCN (18.4 MB, 3328 documents)



- Two Americans have won the 2002 Nobel Prize in Economic Sciences.
- The two scientists, Daniel Kahneman and Vernon L. Smith, received the honour on Wednesday for their work using psychological research and laboratory experiments in economic analysis.

- Egypt honours its Nobel Prize chemist.
- President Hosni Mubarak of Egypt has awarded the country's most prestigious prize - the Nile Necklace - to the Egyptian-born chemist Ahmed Zewail.

Repetition and Elaboration

- Cohension indicator *repetition* is often used as indictor for semantic similarity and semantic consistency, e.g.,
  - „two Americans“ and „two scientists“
  - „chemist“ and „chemist“

- Elaboration phenomena are normal in newspaper texts

S1 is an Elaboration of S0 if a proposition P follows from the assertions of both S0 and S1, but S1 contains a property of one of the elements of P that is not in S0 (Hobbs, 1979)

Relation Argument as a Complex Semantic Object

- A complex noun phrase contains often more than one property about an argument: e.g.

  Egyptian-born chemist Ahmed Zewail

- Relevant properties of a winner in Nobel Prize domain
  - Nationality/origin/inhabitant: e.g., two Americans, the Egyptian-born, a Dutch
  - Profession/occupation: e.g., novelist, chemist, scientist, researcher
  - Title/position: e.g., professor, president
  - Domain description: e.g., recipient, winner, Nobel Laureate
  - General description: e.g., the man, a woman, the team

„two Americans“

$$\begin{bmatrix} sentence\_id : i \\ number : \begin{bmatrix} type : plural \\ amount : 2 \end{bmatrix} \\ definite : indef \\ grammarrole : subject \\ semantics : \begin{bmatrix} nationality : american \end{bmatrix} \\ \end{bmatrix}$$

„the two scientists, ...“

$$\begin{bmatrix} sentence\_id : i+1 \\ number : \begin{bmatrix} type : plural \\ amount : 2 \end{bmatrix} \\ definite : def \\ grammarrole : subject \\ semantics : \begin{bmatrix} class : person \\ profession : scientist \end{bmatrix} \\ names : \langle name1 \quad name2 \rangle \end{bmatrix}$$

Unification of
„*two Americans*“ and „the *two scientists,...*“

- **Recency and valide chain**
- **Parallel subject**
- **Repetition: number**
- **Semantic compatibility**

$$\begin{bmatrix} number : \begin{bmatrix} type : plural \\ amount : 2 \end{bmatrix} \\ semantics : \begin{bmatrix} class : person \\ nationality : american \\ profession : scientist \end{bmatrix} \\ names : \langle name1 \quad name2 \rangle \\ \end{bmatrix}$$

References

1. N. Kushmerick. Wrapper induction: Efficiency and Expressiveness, Artificial Intelligence, 2000.
2. I. Muslea. Extraction Patterns for Information Extraction. AAAI-99 Workshop on Machine Learning for Information Extraction.
3. Riloff, E. and R. Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99) , 1999, pp. 474-479.
4. R. Yangarber, R. Grishman, P. Tapanainen and S. Huttunen. Automatic Acquisition of Domain Knowledge for Information Extraction. In Proceedings of the 18th International Conference on Computational Linguistics: COLING-2000, Saarbrücken.
5. F. Xu, H. Uszkoreit and Hong Li. Automatic Event and Relation Detection with Seeds of Varying Complexity. In Proceedings of AAAI 2006 Workshop Event Extraction and Synthesis, Boston, July, 2006.
6. F. Xu, D Kurz, J Piskorski, S Schmeier. A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In Proceedings of LREC 2002.
7. W. Drozdzyski, H.U. Krieger, J. Piskorski, U. Schäfer and F. Xu. Shallow Processing with Unification and Typed Feature Structures -- Foundations and Applications. In KI (Artifical Intelligence) journal 2004.
8. Feiyu Xu, Hans Uszkoreit, Hong Li. A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity. In Proceeedings of ACL 2007, Prague
9. http://www.dfki.de/~neumann/ie-esslli04.html

DARE and Extensions (http://dare.dfki.de)

- Xu, Feiyu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. ACL 2007
- Xu, Feiyu. 2007. Bootstrapping Relation Extraction from Semantic Seeds. Phd-thesis.
- Feiyu Xu, Hans Uszkoreit, Hong Li. Task driven coreference resolution for relation extraction.

  ECAI 2008.
- Xu, Feiyu, Hans Uszkoreit, Hong Li, and Niko Felger. Adaptation of relation extraction rules to new domains. LREC 2008.
- Hans Uszkoreit, Feiyu Xu, Hong Li. Analysis and Improvement of Minimally Supervised Machine Learning for Relation Extraction. NLDB 2009. Keynote.
- Xu, Feiyu, Hans Uszkoreit Sebastian Krause and Hong Li. Boosting relation extraction with limited closed-world knowledge. COLING 2010.