Relation Extraction and Machine Learning for IE

Feiyu Xu

feiyu@dfki.de

Language Technology-Lab
DFKI, Saarbrücken

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 …

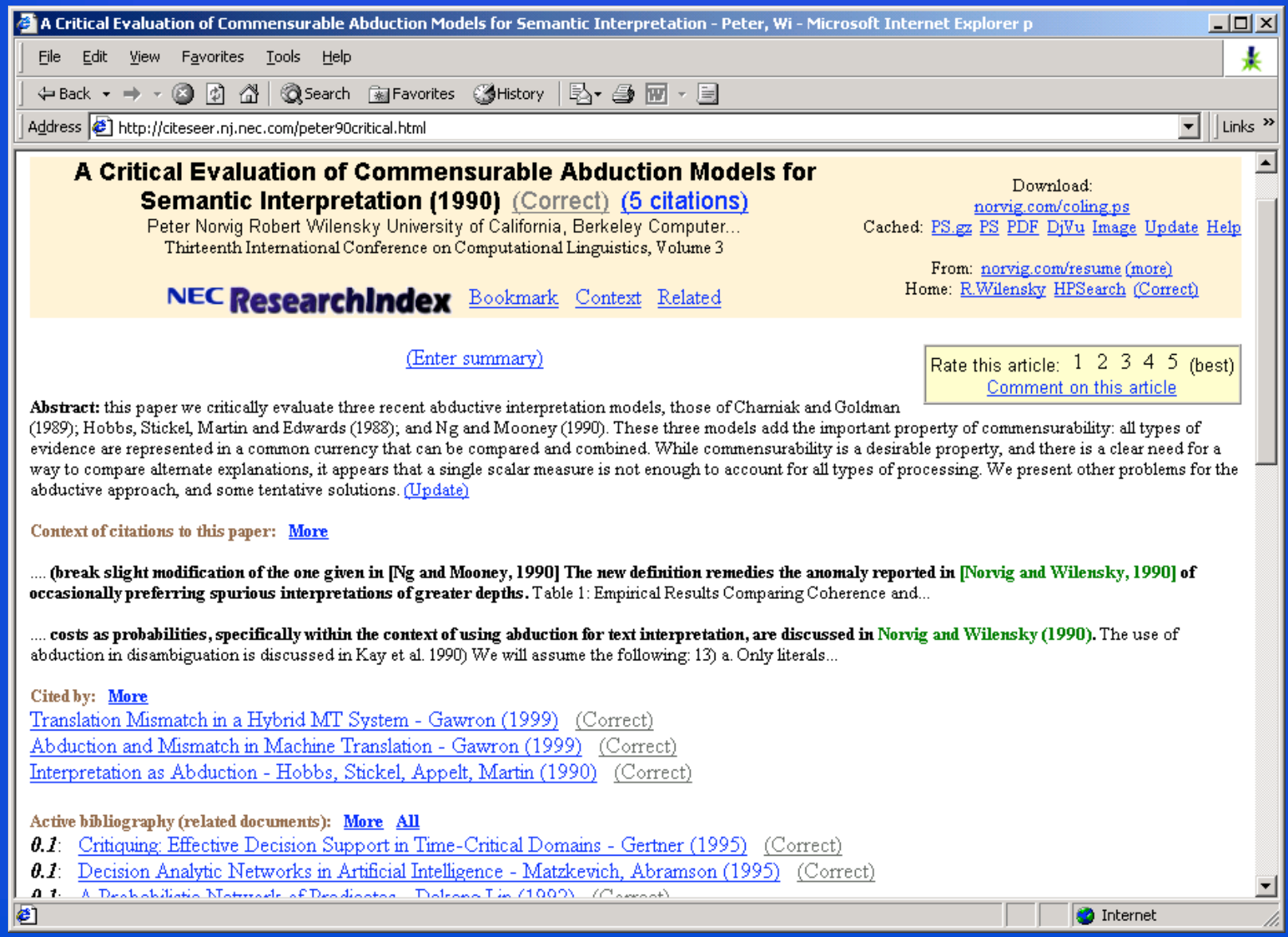
provided by NIST:
[http://www-nlpir.nist.gov/related_projects/muc/]

Information Extraction: A Pragmatic Approach

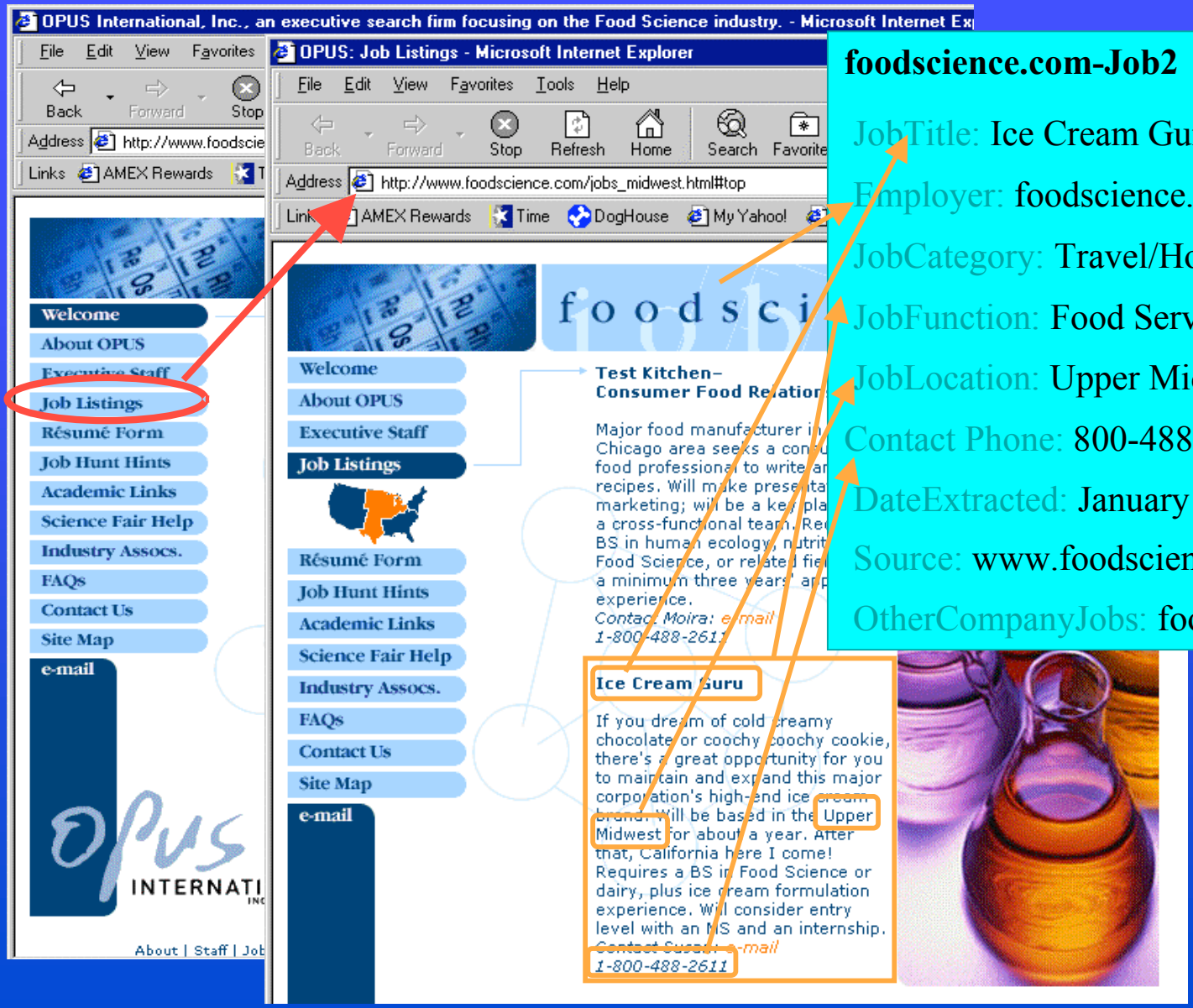
- Identify the types of entities that are relevant to a particular task
- Identify the range of facts that one is interested in for those entities
- Ignore everything else

[Appelt, 2003]

IE from Research Papers



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

On the Notion *Relation Extraction*

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

Types of Information Extraction in LT

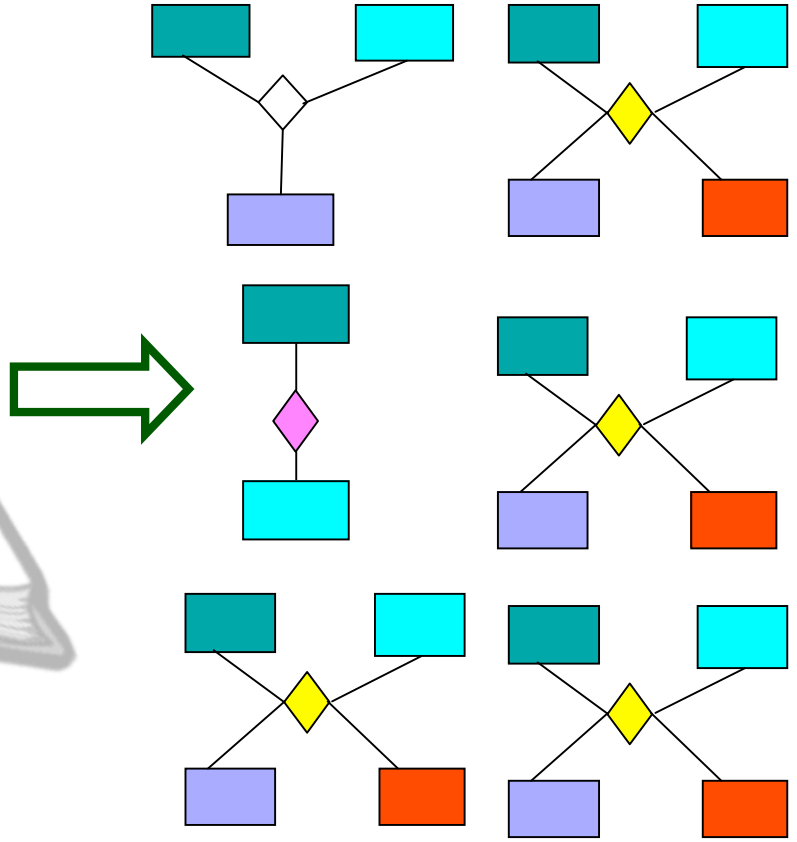
- Topic Extraction
- Term Extraction
- Named Entity Extraction
- Binary Relation Extraction
- N-ary Relation Extraction
- Event Extraction
- Answer Extraction
- Opinion Extraction
- Sentiment Extraction

Types of 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**

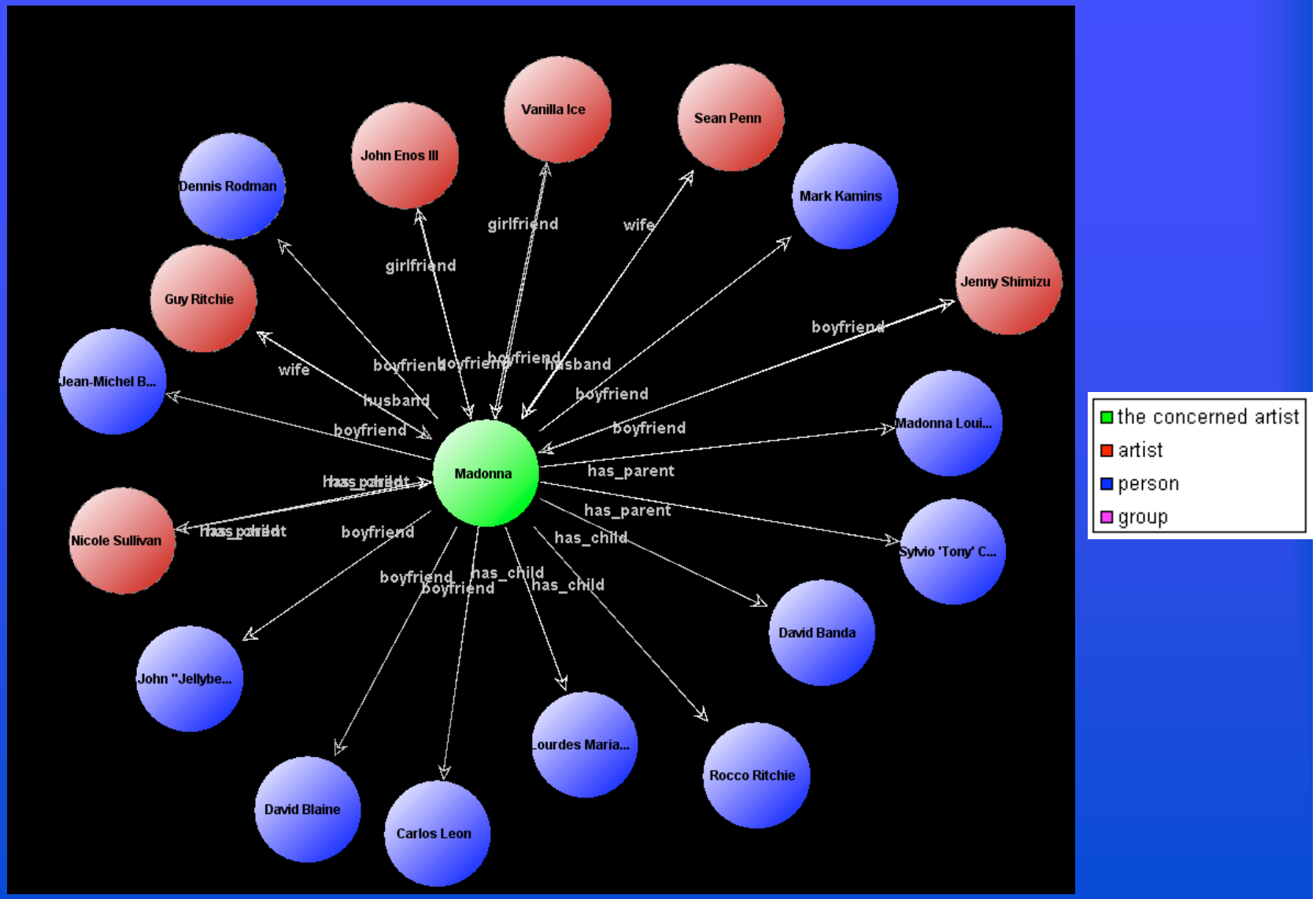
Types of Relation Extraction

Relation Extraction is a demanding sub-area of Information Extraction



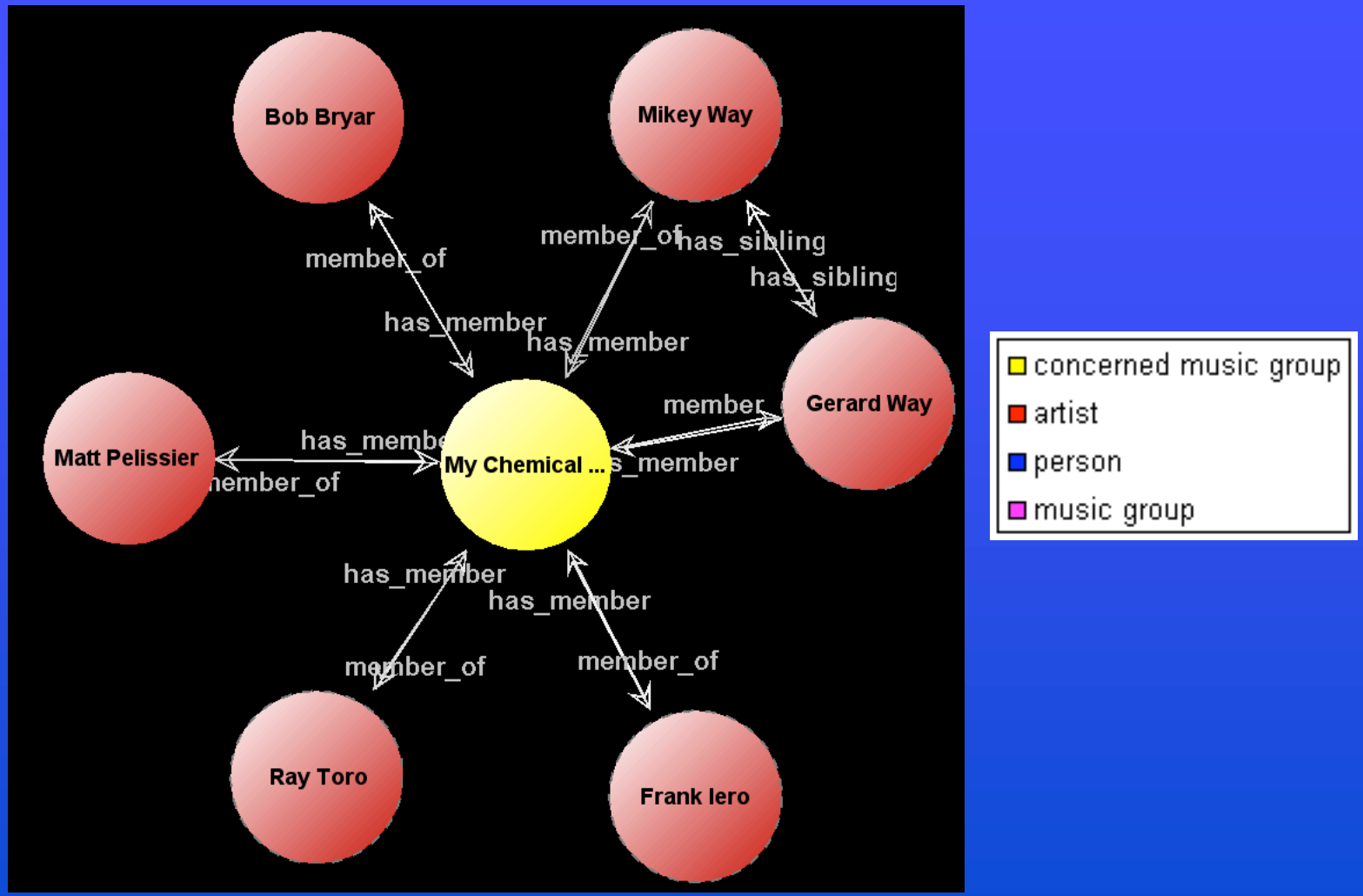
Example of Binary Social Relations

Social Network of “Madonna” (Depth = 1)



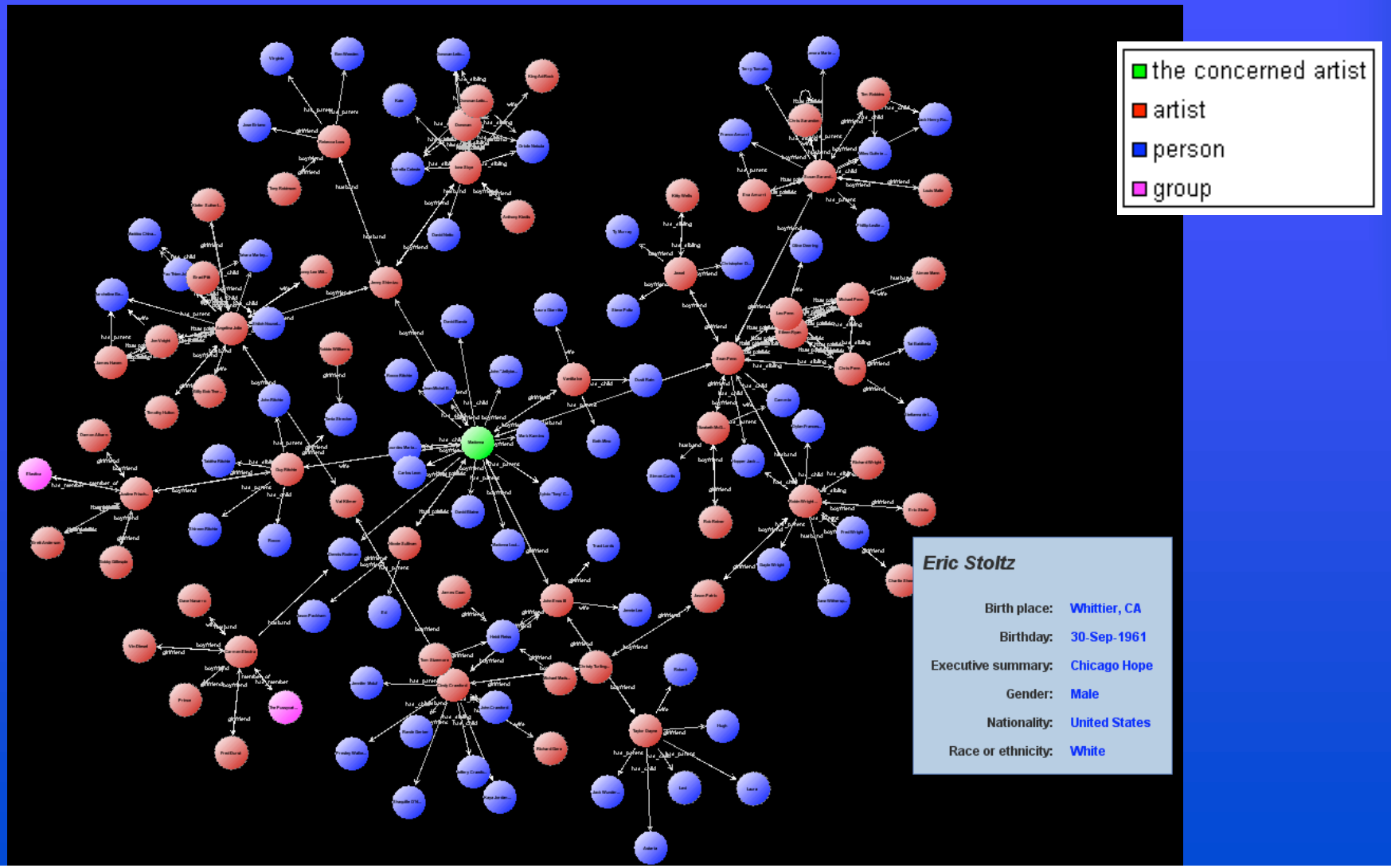
Examples of Binary Relations

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



Examples

Social Network of “Madonna” (Depth = 3)



Relation about Person, Title and Organization

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates
Microsoft

* Bill Veghte
Microsoft
VP

* Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
RichardStallman	founder	Free Soft..

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, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

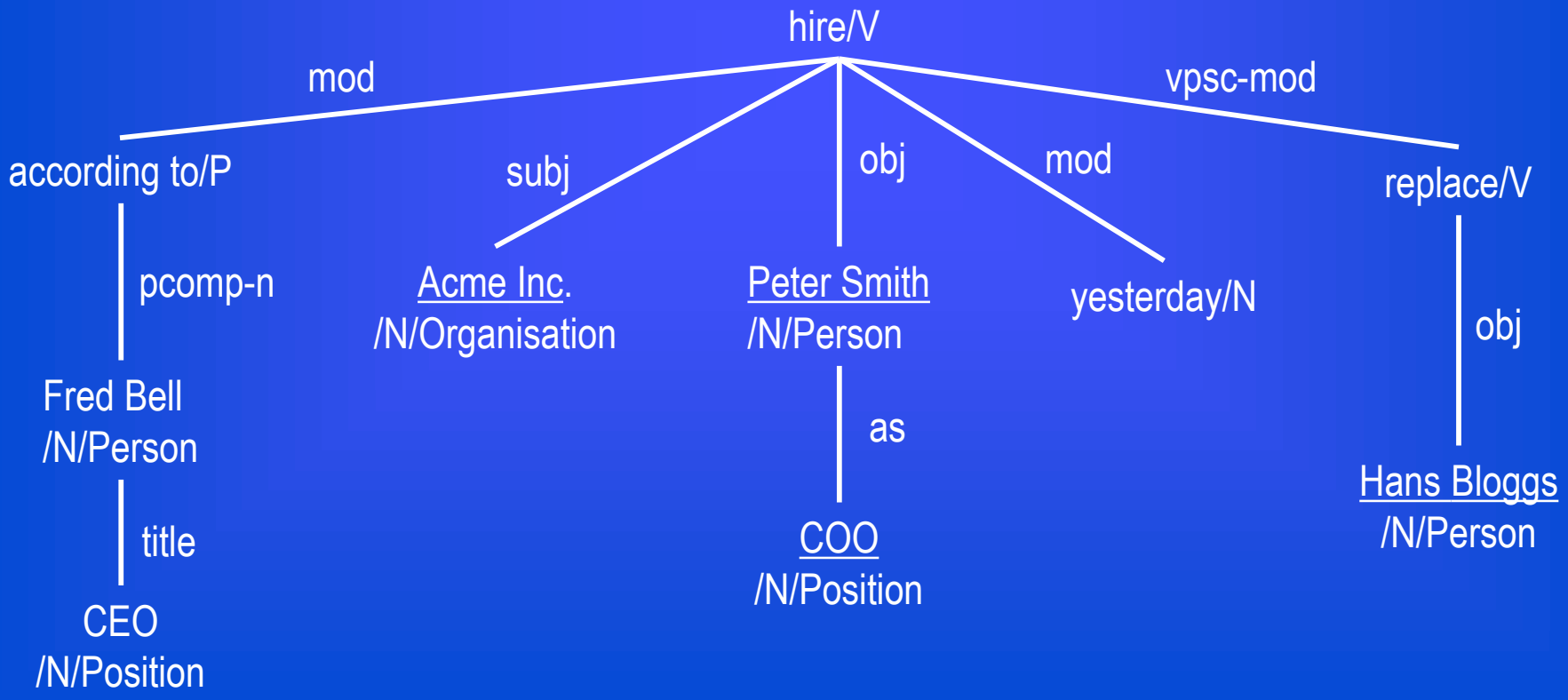
<person_in, person_out, position, organisation>

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

<person_in, person_out, position, organisation>

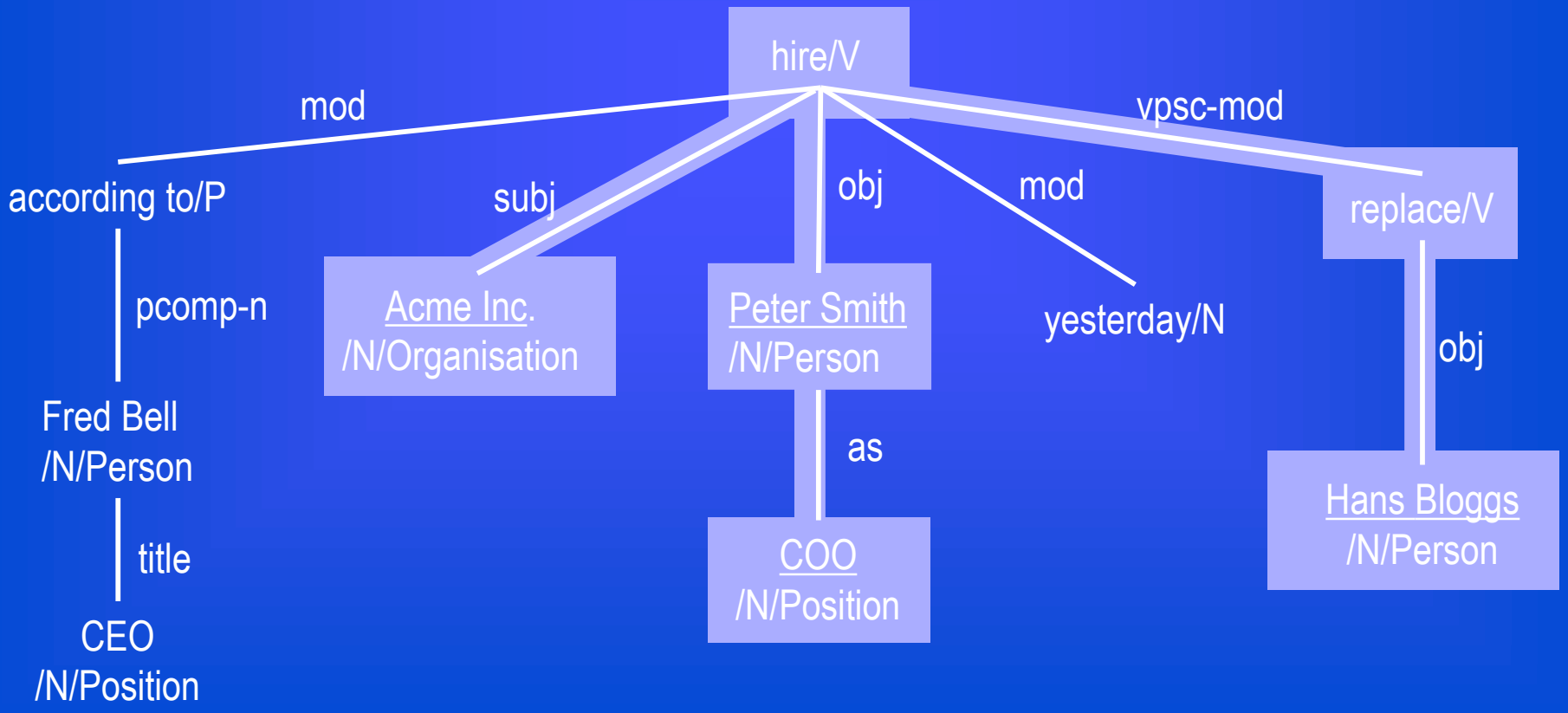
According to CEO Fred Bell, **Acme Inc**. hired **Peter Smith** as **COO** yesterday, replacing **Hans Bloggs**.

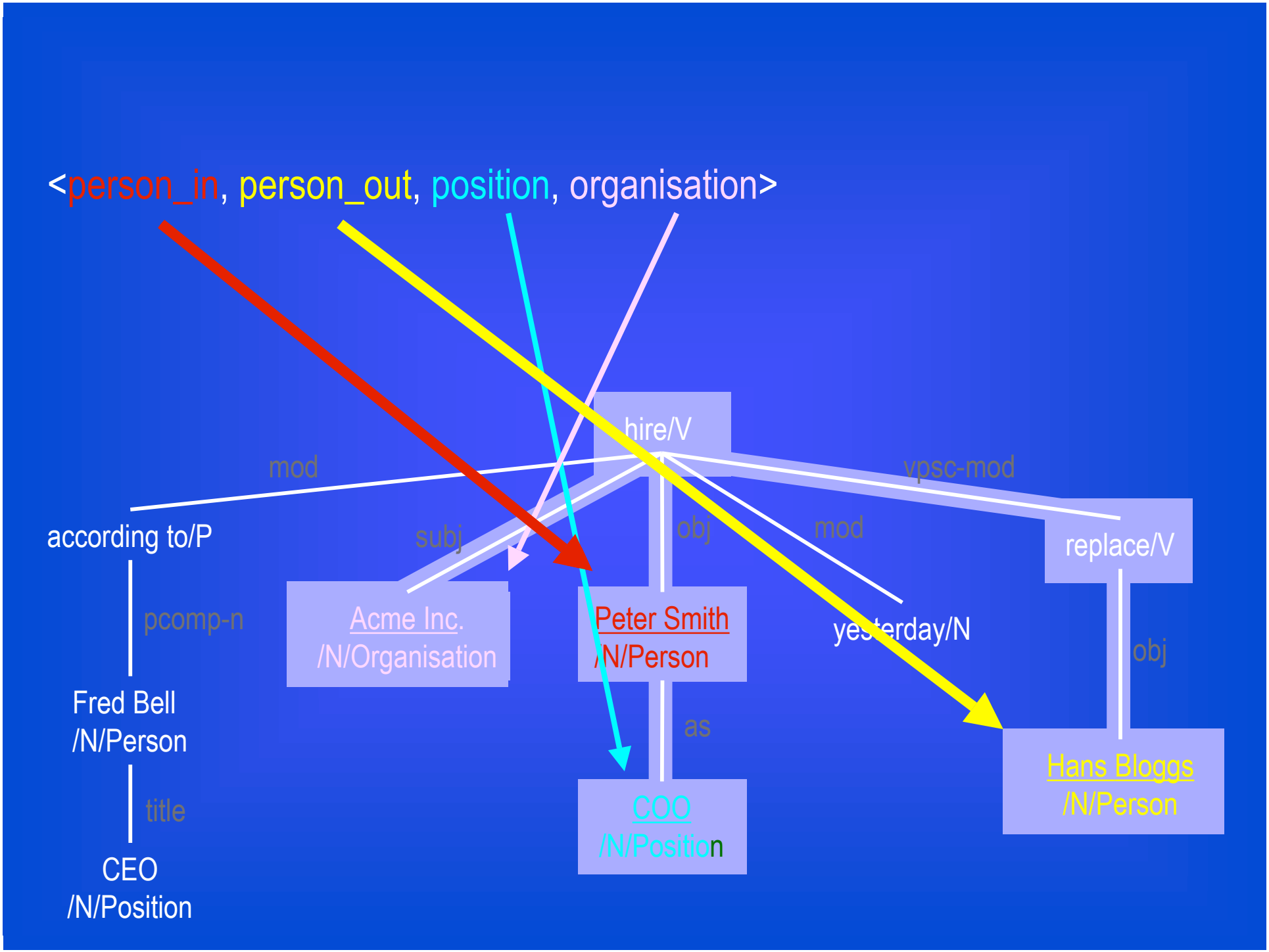
<person_in, person_out, position, organisation>



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

<person_in, person_out, position, organisation>




<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
as
Fred Bell
/N/Person
Hans Bloggs
/N/Person
title
COO
/N/Position
CEO
/N/Position

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

Generic IE tasks for MUC-7

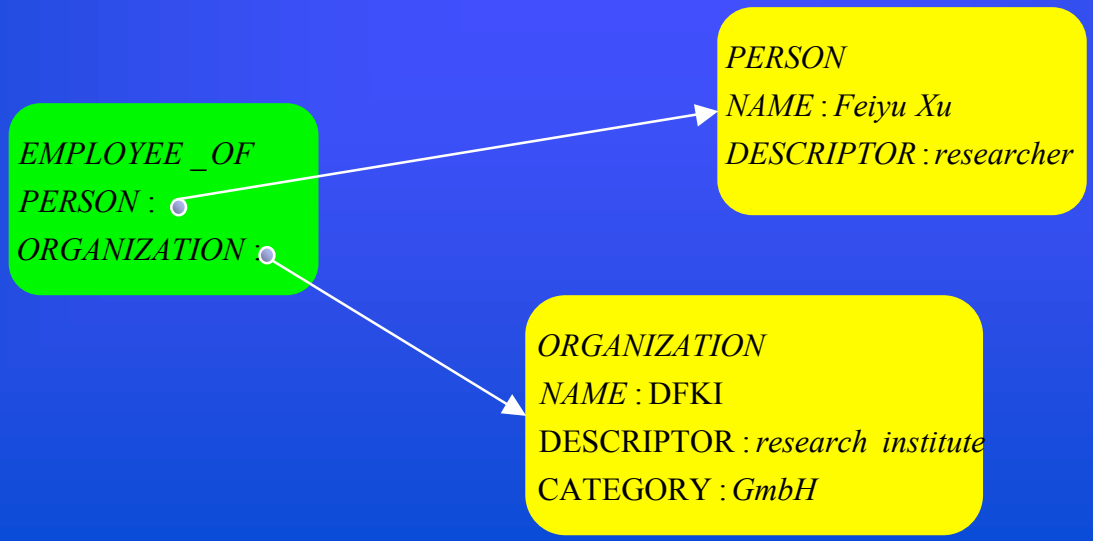
- (NE) Named Entity Recognition Task requires the identification an classification of named entities
  - organizations
  - locations
  - persons
  - dates, times, percentages and monetary expressions

- (TE) Template Element Task requires the filling of small scale templates for specified classes of entities in the texts
  - Attributes of entities are slot fills (identifying the entities beyond the name level)
  - Example: Persons with slots such as name (plus name variants), title, nationality, description as supplied in the text, and subtype.

    “Capitan Denis Gillespie, the comander of Carrier Air Wing 11”

Generic IE tasks for MUC-7

- (TR) Template Relation Task requires filling a two slot template representing a binary relation with pointers to template elements standing in the relation, which were previously identified in the TE task
  - subsidiary relationship between two companies (employee_of, product_of, location_of)



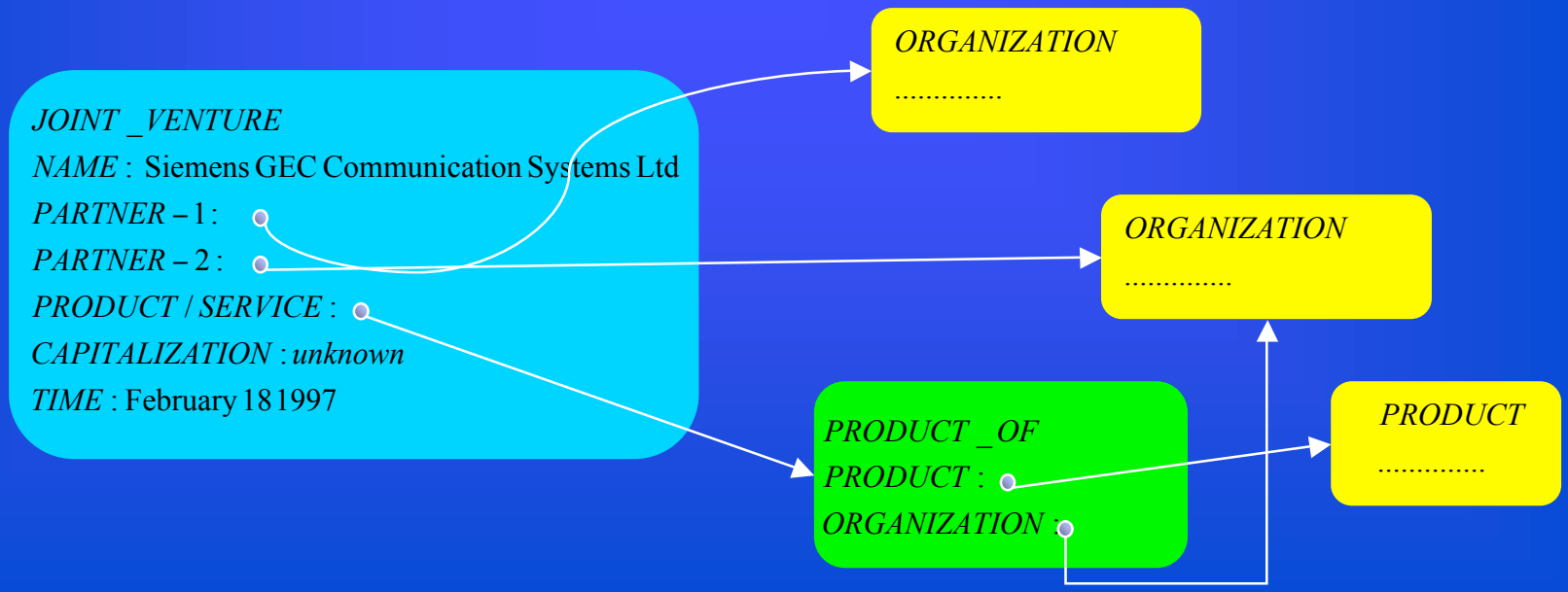
Generic IE tasks for MUC-7

- (CO) Coreference Resolution requires the identification of expressions in the text that refer to the same object, set or activity
  - variant forms of name expressions
  - definite noun phrases and their antecedents
  - pronouns and their antecedents

"The U.K. satellite television broadcaster said its subscriber base grew 17.5 percent during the past year to 5.35 million"

Generic IE tasks for MUC-7

- (ST) Scenario Template requires filling a template structure with extracted information involving several relations or events of interest
  - intended to be the MUC approximation to a real-world information extraction problem
  - identification of partners, products, profits and capitalization of joint ventures



Tasks evaluated in MUC 3-7 [Chinchor, 98]

EVAL\TASK	NE	CO	RE	TR	ST
MUC-3					YES
MUC-4					YES
MUC-5					YES
MUC-6	YES	YES	YES		YES
MUC-7	YES	YES	YES	YES	YES

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.

The ACE Program

- "Automated Content Extraction" since 1999
- Develop core information extraction technology by focusing on extracting specific semantic entities and relations over a very wide range of texts.
- Corpora: Newswire and broadcast transcripts, but broad range of topics and genres.
  - Third person reports
  - Interviews
  - Editorials
  - Topics: foreign relations, significant events, human interest, sports, weather
- Discourage highly domain- and genre-dependent solutions

[Appelt, 2003]

Components of a Semantic Model

- 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 a Semantic Model

- 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.

Relations in Time

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

Relations vs. Features or Roles in AVMs

- Several two place relations between an entity *x* and other entities y_i can be bundled as properties of x. In this case, the relations are called roles (or attributes) and any pair <relation : y_i> is called a role assignment (or a feature).
- name <x, CR>

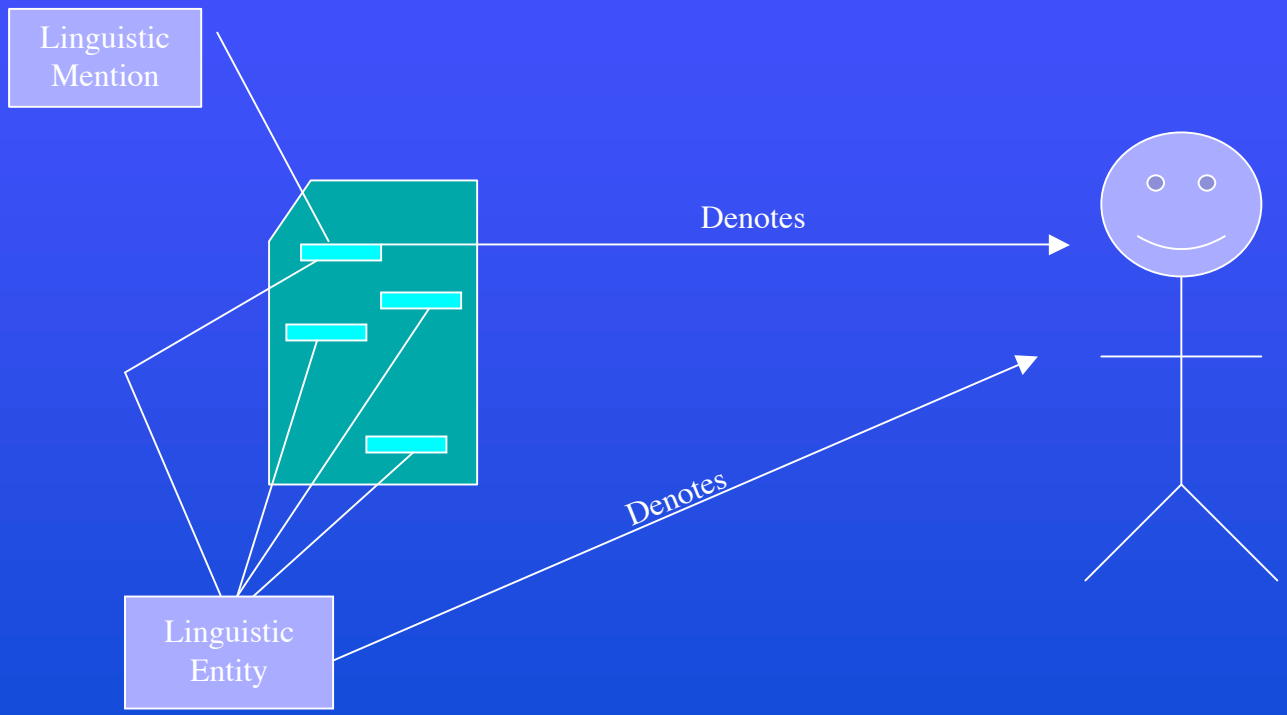
[
name: Condoleezza Rice
office: National Security Advisor
age: 49
gender: female
]

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

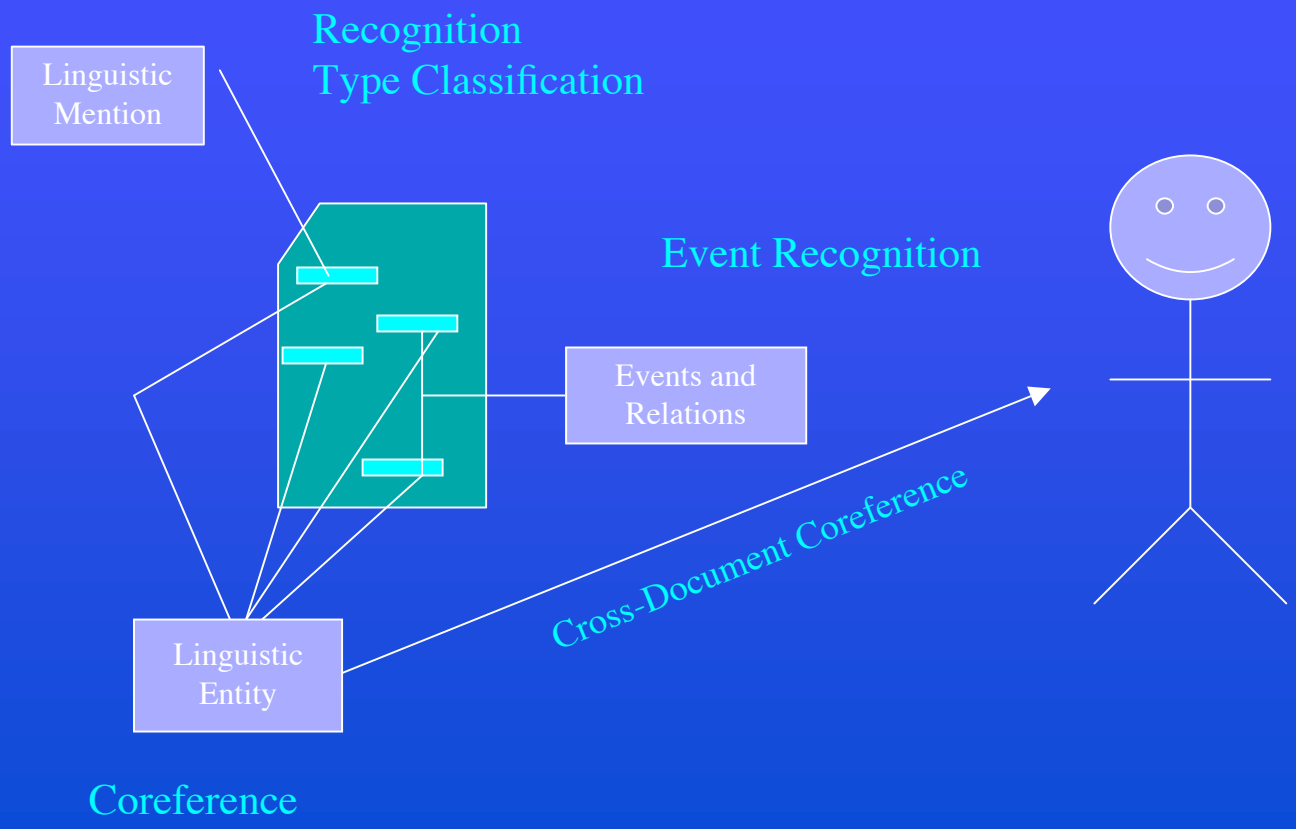
[Appelt, 2003]

Language and World Model



[Appelt, 2003]

NLP Tasks in an Extraction System



[Appelt, 2003]

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.

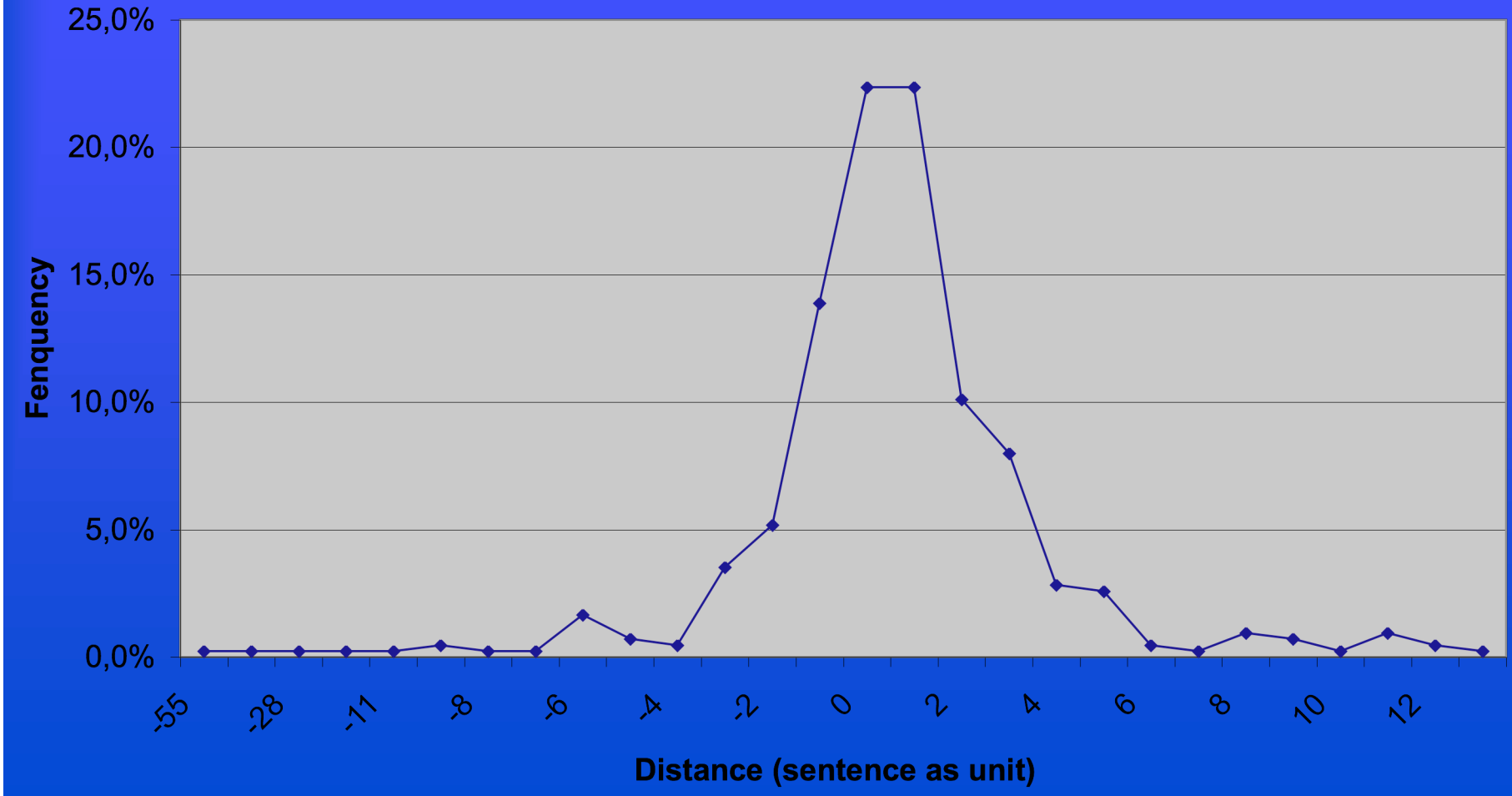
<?PERSON, Nobel, Peace, 2005>

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)



1. Two Americans have won the 2002 Nobel Prize in Economic Sciences.
2. 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.

1. Egypt honours its Nobel Prize chemist.
2. 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“

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

„*two scientists*“

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

Unification of
„*two Americans*“ and „*two scientists*“

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

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]

The ACE Ontology

- Persons
  - A natural kind, and hence self-evident
- Organizations
  - Should have some persistent existence that transcends a mere set of individuals
- Locations
  - Geographic places with no associated governments
- Facilities
  - Objects from the domain of civil engineering
- Geopolitical Entities
  - Geographic places with associated governments

[Appelt, 2003]

Why GPEs

- An ontological problem: certain entities have attributes of physical objects in some contexts, organizations in some contexts, and collections of people in others
- Sometimes it is difficult to impossible to determine which aspect is intended
- It appears that in some contexts, the same phrase plays different roles in different clauses

Aspects of GPEs

- Physical
  - San Francisco has a mild climate
- Organization
  - The United States is seeking a solution to the North Korean problem.
- Population
  - France makes a lot of good wine.

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

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.
  - “Baltimore, MD”
  - 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…

Types of ACE Relations

- ROLE - relates a person to an organization or a geopolitical entity
  - Subtypes: member, owner, affiliate, client, citizen
- PART - generalized containment
  - Subtypes: subsidiary, physical part-of, set membership
- AT - permanent and transient locations
  - Subtypes: located, based-in, residence
- SOC - social relations among persons
  - Subtypes: parent, sibling, spouse, grandparent, associate

Event Types (preliminary)

- Movement
  - Travel, visit, move, arrive, depart …
- Transfer
  - Give, take, steal, buy, sell…
- Creation/Discovery
  - Birth, make, discover, learn, invent…
- Destruction
  - die, destroy, wound, kill, damage…