Relation Extraction

Feiyu Xu

Outline

- Introduction to relation extraction
- Brief history of information extraction
- Machine learning approches to relation extraction
  - State-of-the-art
  - DARE System (http://dare.dfki.de)
- Big Text Data Analytics
- References

Big Textual Data

- In a single day on the Internet, PayPal processes over $315 million in transactions, Facebook takes in about 350 million uploaded images, and Twitter blasts out 400 million tweets from its users
- From the perspectives of data content processing and data mining, language data belongs to unstructured data like images or videos because of the complexities of their interne structures.
- Technologies such as information retrieval and text analytics are needed to ease human users to access the relevant information in a realistic time from a big amount of data

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]

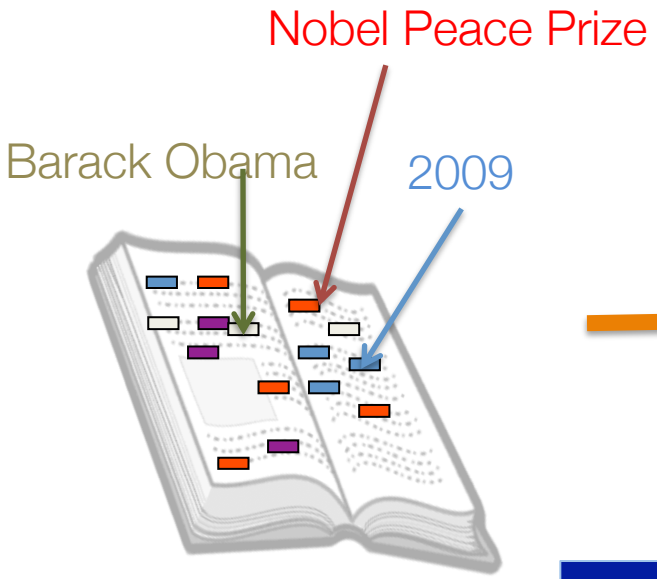
Types of Information Extraction in LT

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

What is Relation Extraction

Feiyu Xu

- Given an unstructured text, a relation extraction (RE) tool should be able to automatically recognize and extract relations among the relevant entities or concepts that are salient to the user's needs



RE

- Walter Kohn, Nobel, Chemistry,1998
- J.M. Coetzee, Nobel, Literature, 2003
- Barack Obama, Nobel, Peace, 2009

Linguistic Patterns:

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

Types of Information Extraction in LT

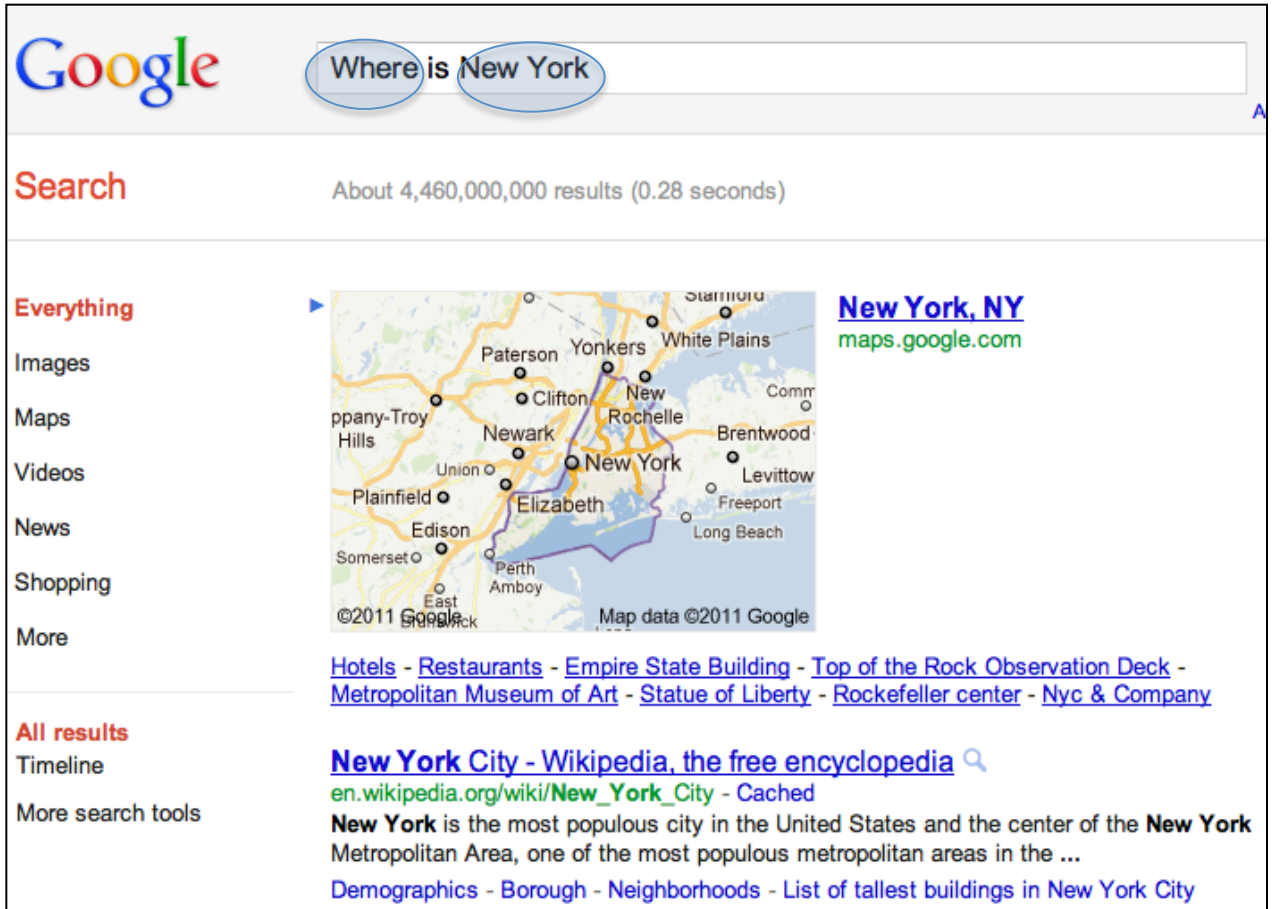
- Topic Extraction
- Term Extraction
- Named Entity Extraction
- Binary Relation Extraction
- N-ary Relation Extraction
- Event Extraction
- Answer Extraction
- Opinion Extraction
- Sentiment Extraction

Types of Relation Extraction

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



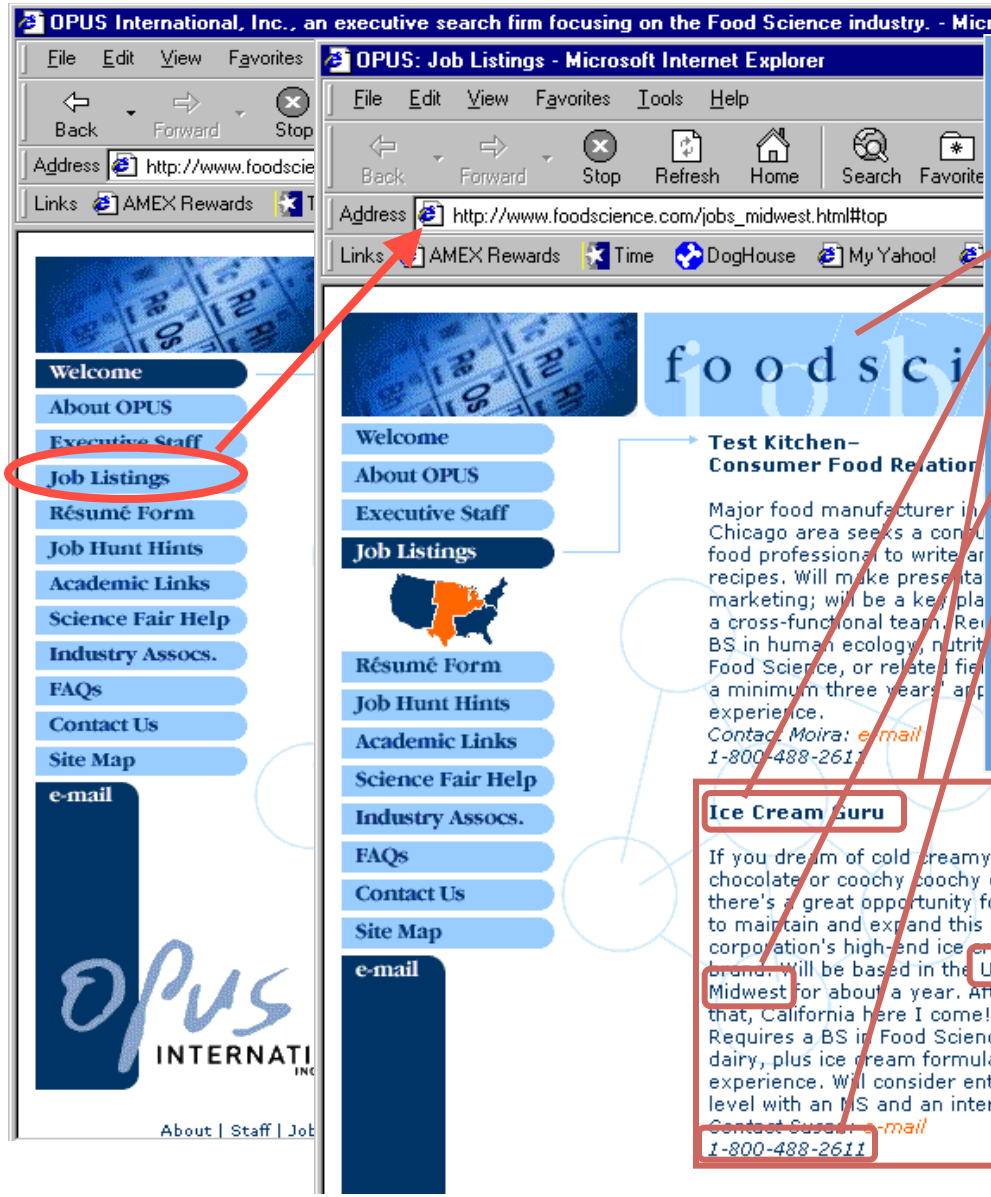
General application task 2:

☆ Information acquisition for information provider

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

RE

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

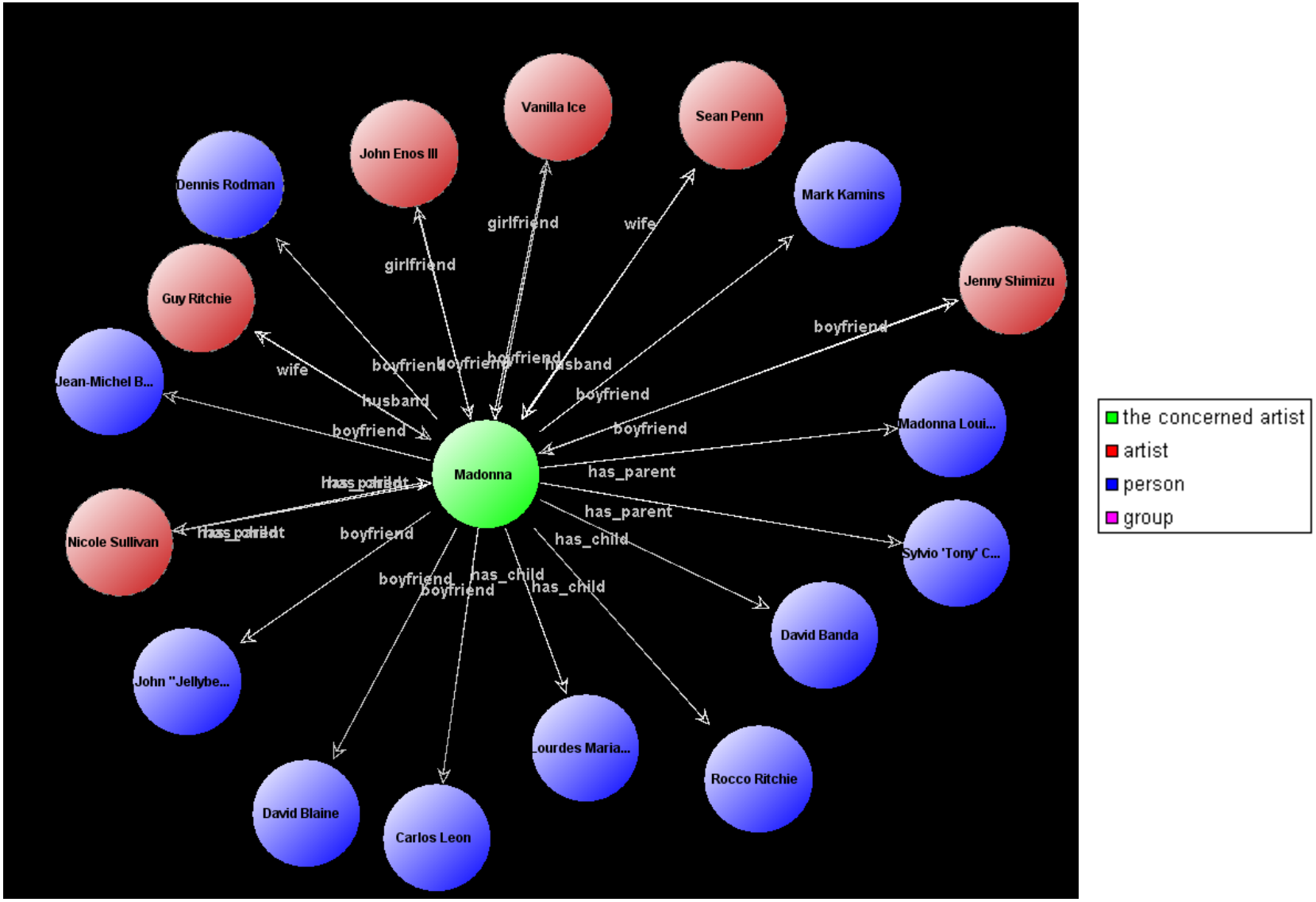
DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.html

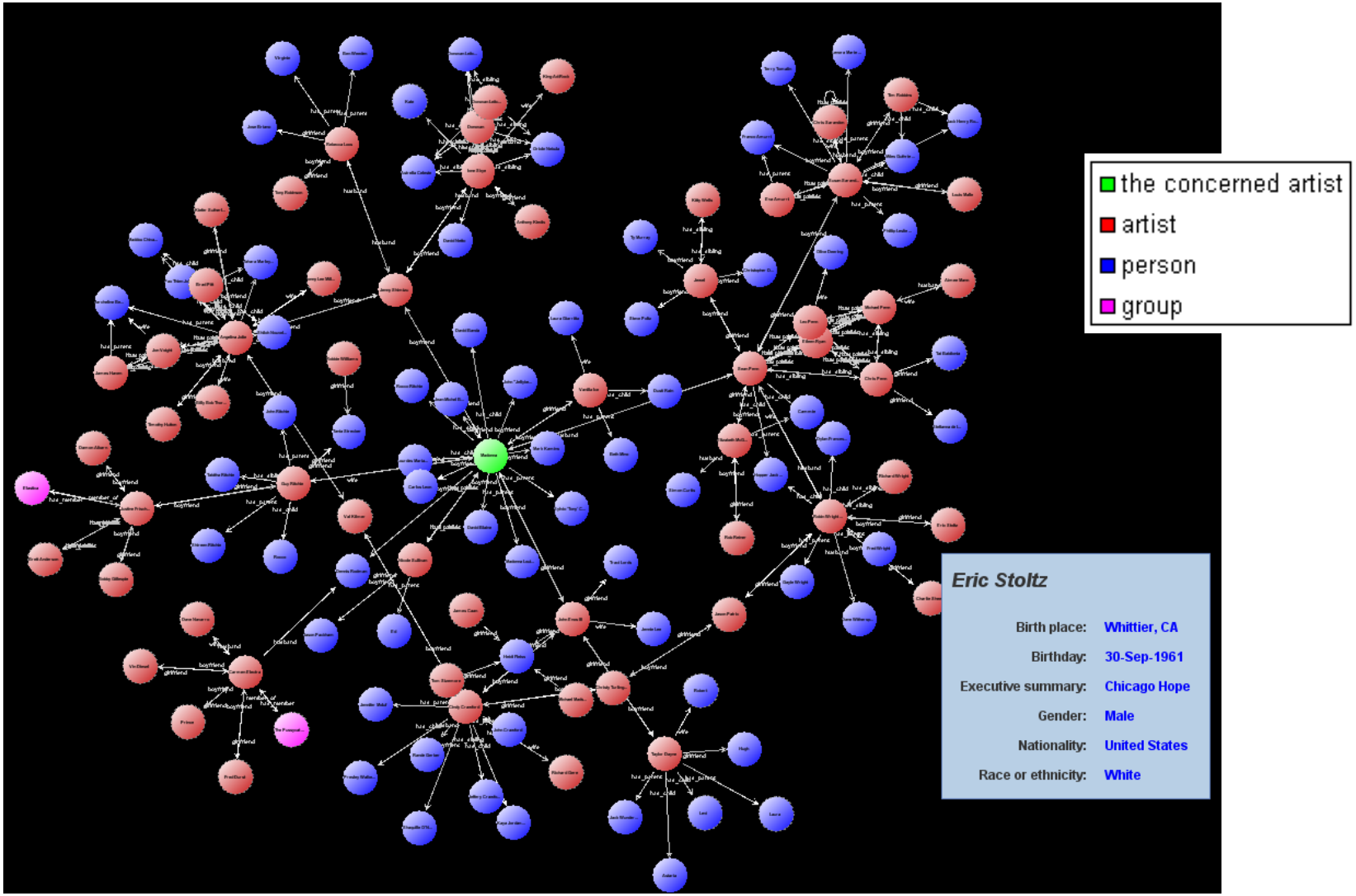
OtherCompanyJobs: foodscience.com-Job1

Acquisition of Social Network of Pop Stars from Web

Social Network of "Madonna" (Depth = 1)

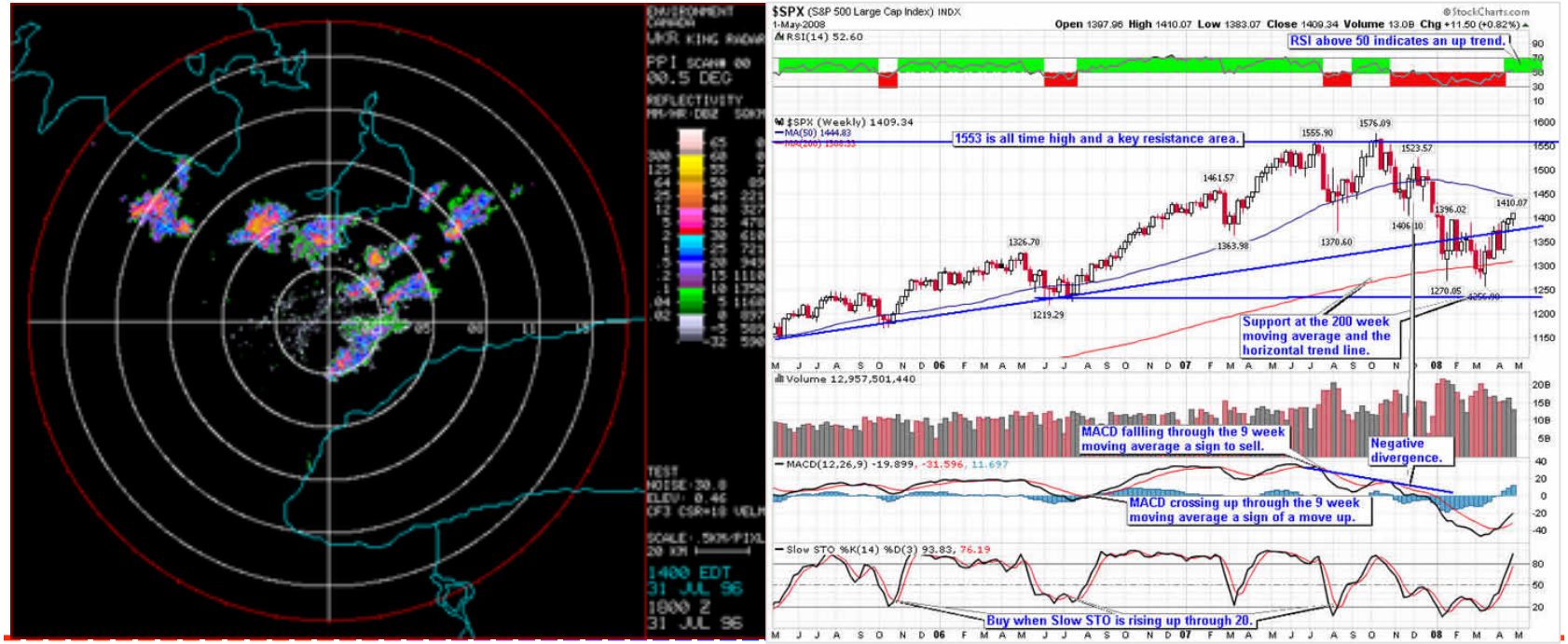


Social Network (Depth=3)

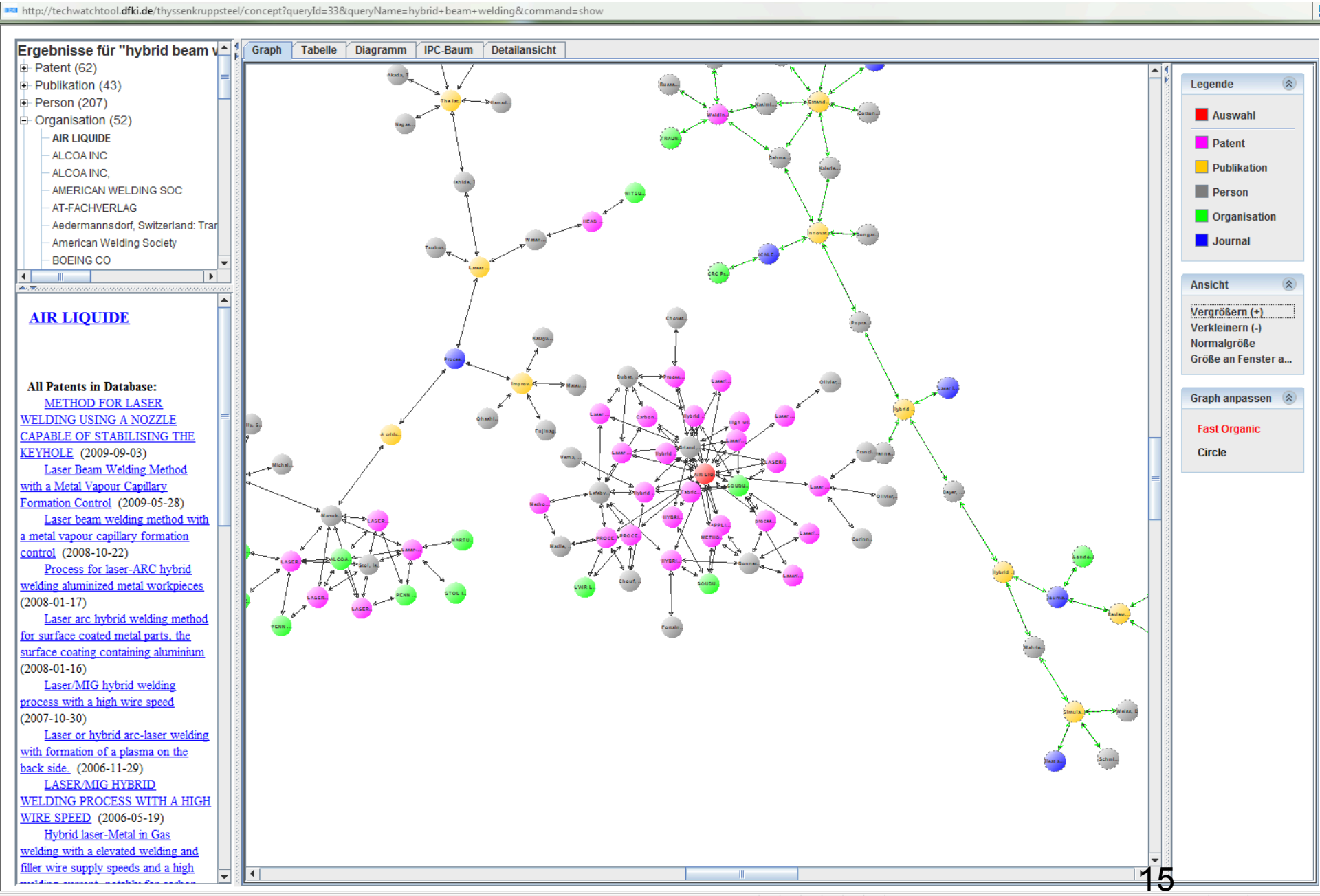


General application task 3: Big Data Analytics

- Enabling the linking between structured and unstructured data
  - Large-scale information monitoring
  - Analytics: analyses of areas, markets, trends
  - Watch: Scanning for relevant new developments



Example: Network of Innovation Keyplayers



Example in Opinion Mining

Mitten in der Euro-Krise geht ***Altkanzler Helmut Kohl*** mit *Angela Merkel* äußerst hart ins Gericht

-- Welt online, 25.08.2011

Opinion Holder

Opinion Target

Polarity

A Brief History of IE

Message Understanding Conference (MUC-6)

- 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

New Evaluation Conferences

- New terms closely related to IE research:
  - ACE (Automatic Content Analysis), Text Analytics, Text Mining
- ACE: Automatic Content Extraction (ACE) Evaluation
  - http://www.itl.nist.gov/iad/mig/tests/ace/
- TAC: Text Analytics Conference
  - http://www.nist.gov/tac/
  - Knowledge Base Population (KBP)
  - Biomedical Summarization (BiomedSumm)

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.

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

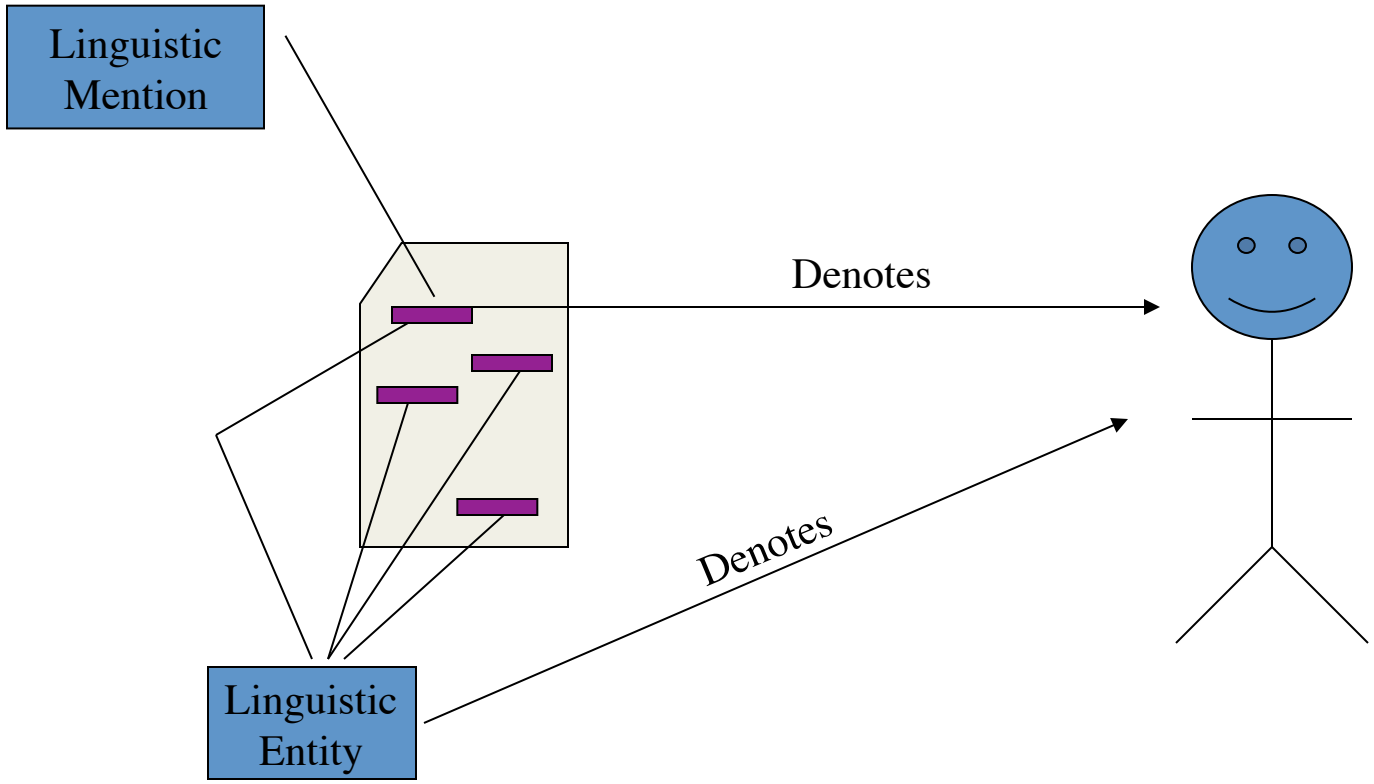
[Appelt, 2003]

The Basic Semantic Tasks of an IE System

- Recognition of linguistic mentions
- Classification of linguistic mentions 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.

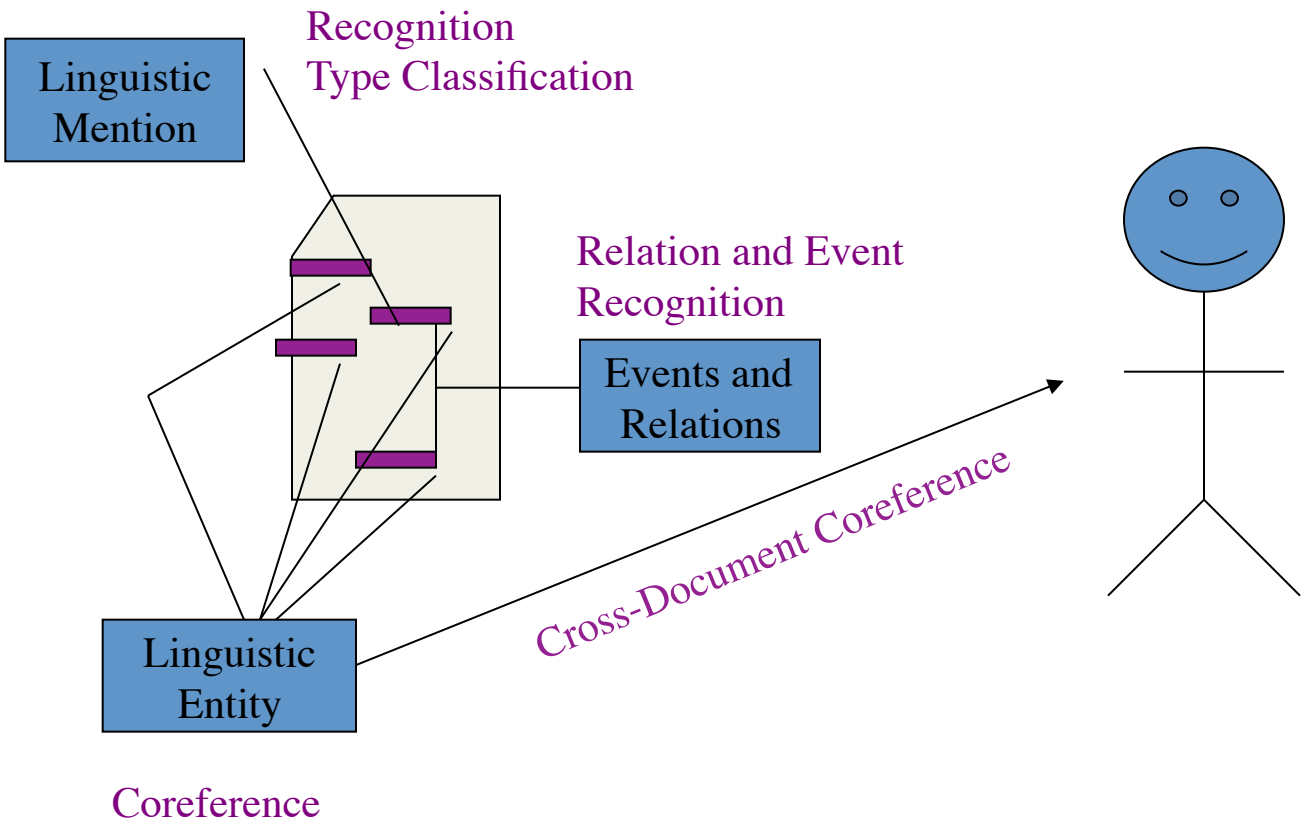
[Appelt, 2003]

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.

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.

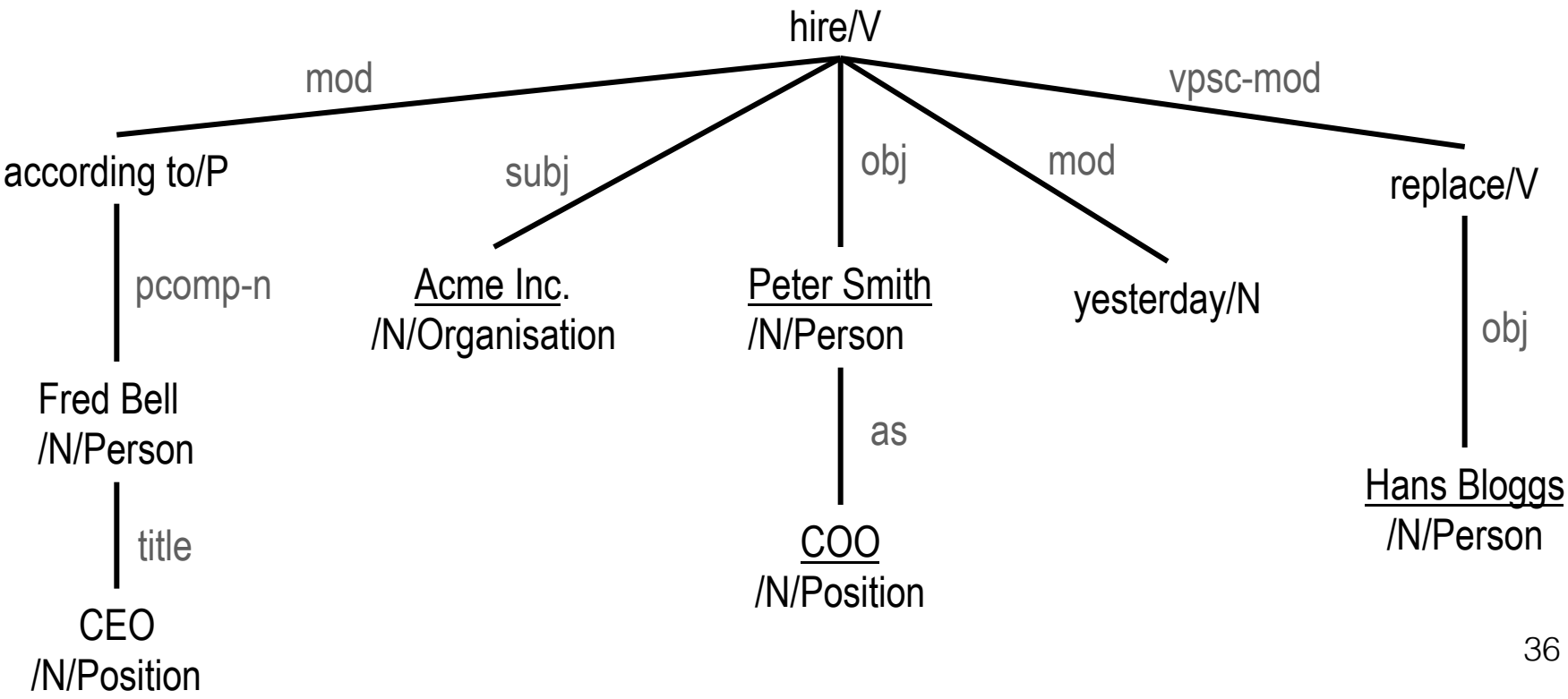
<person_in, person_out, position, organisation>

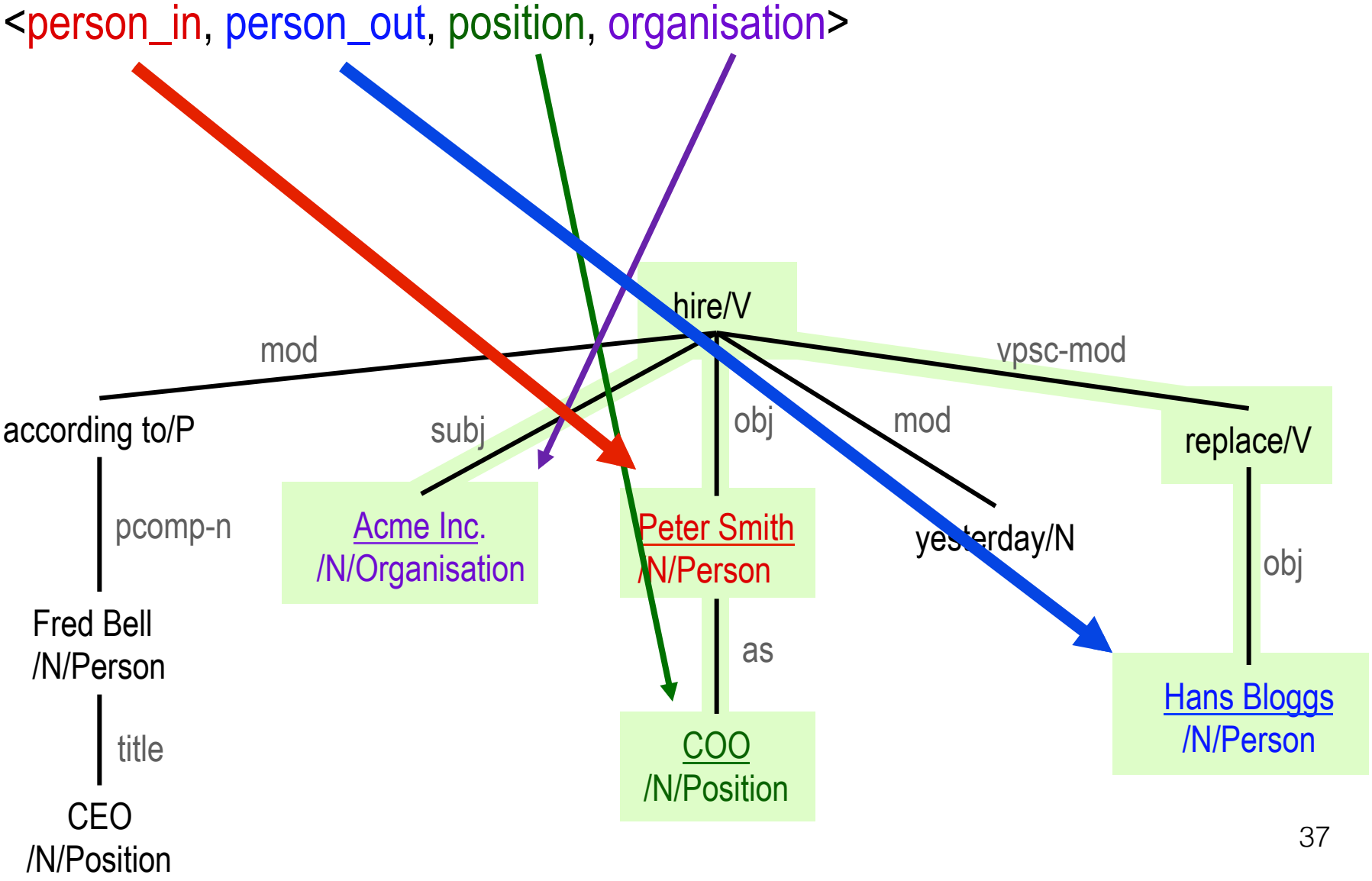
According to CEO Fred Bell, Acme Inc. hired Peter Smith as COO yesterday, replacing Hans Bloggs.

<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>




<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

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

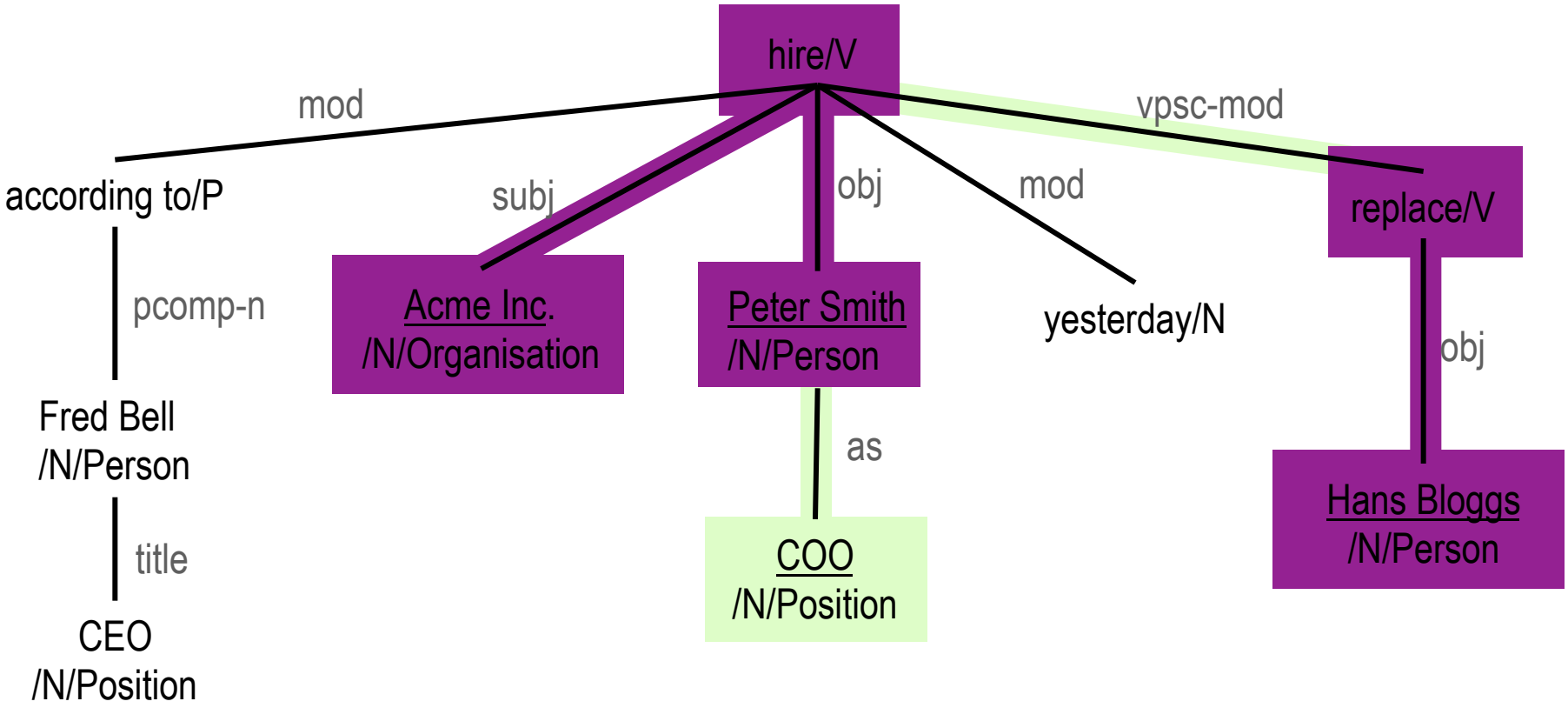
Parameters of IE Real-World Tasks

- 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
  - Distant Supervision
  - Unsupervised
- Human interaction/contribution
- Data properties
  - Domain relevance
  - Redundancy
  - Connectivity
- Evaluation/validation
  - With/without gold standard
  - Performance: recall & precision
  - Interaction among parameters

State of the Art

Binary Relation Only Approaches

- Extraction of binary relations only such as
  - *author-book*
  - *company-location*
- Do not employ the existing syntactic and semantic structures among n > 2 arguments and rely on a later component to merge binary relations into complex relations
- Approaches
  - Ravichandran and Hovy, 2002
  - Pantel et al., 2004
  - Pasca et al., 2006a; Pasca et al., 2006b


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

Surface-oriented Rule Representation

- Shallow linguistic analyses
- These formalisms are robust and efficient but only handle binary relations.
- Work best for relations whose arguments usually co-occur in close proximity within a sentence and whose mentions exhibit limited linguistic variation
- Approaches
  - Pasca et al., 2006a; Pasca et al.,2006b;
  - Kozareva et al., 2008; Hovy et al., 2009; Kozareva and Hovy, 2010

Minimally Supervised (Bootstrapping)

- Based on iterative learning with limited initial knowledge
  - Start a small number of initial examples of relation instances (or patterns)
  - Label the free texts during iterations (e.g., Agichtein and Gravano, 2000; Yangarber et al.,2000; Ravichandran and Hovy, 2002; Stevenson and Greenwood, 2005).
- Often suffer from semantic drift or the propagation of errors occurring during iterations
- Performance depends on data properties

Distant Supervision

- A massive seed-based, one step version of bootstrapping
- Rely on a large amount of trustworthy facts
- Their performance does not hinge on corpus data properties such as redundancy
- Approaches
  - (Mintz et al., 2009)
  - Others: (Wu and Weld, 2007); (Wu et al., 2008); (Weld et al., 2008); (Hoffmann et al., 2010); (Xu et al., 2011(; (Nguyen and Moschitti, 2011)

Open IE

- They do not target given relations.
- They are very useful for applications continuously faced with new relation or event types, e.g., online social media monitoring
- However, the results of these systems cannot be directly taken for filling knowledge databases, because the semantics of the new relations including the roles of the entities remains unknown.
- Example: TextRunner (Banko et al., 2007; Yates et al., 2007)

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