Intelligent Information Extraction

Günter Neumann & Feiyu Xu

LT-lab, DFKI

Saarbrücken, Germany

Course Overview

- Lecture 1.
  - General overview (Günter)
- Lecture 2.
  - Core technology (Feiyu)
- Lecture 3.
  - Machine Learning for Named Entities (Günter)
- Lecture 4.
  - Learning Template Filling Rules (Feiyu)
- Lecture 5.
  - Advanced topics (Feiyu & Günter)

Introduction: Part 1

- What is IE?
- Applications of IE
- IE tasks
- Evaluation
  - MUC
  - ACE

Basic Terms & Examples

- Information Extraction from NL texts
- Data vs. Information
- NLP as normalization

Information Extraction (IE)

The goal of IE research is to build systems that find and link *relevant* information from NL text ignoring irrelevant information.

Core Functionality

Input

- Templates coding relevant information, e.g. company, product, medical information
- set of real world texts

Output

- set of instantiated templates filled with relevant text fragments (normalized to a canonical form)

Example: Terrorists actions

Salvadoran President-elect Afredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of crime.

Incident:	KILLING
Perpetrator:	„terrorist“
Confidence:	__
Human Target:	„Roberto Garcia Alvarado“

Incident:	KILLING
Perpetrator:	FMLN
Confidence:	Accused by Authorities
Human Target:	__

Incident:	KILLING
Perpetrator:	FMLN
Confidence:	Accused by Authorities
Human Target:	„Roberto Garcia Alvarado

Example: Company's turnover

Lübeck (dpa) - Die Lübecker Possehl-Gruppe, ein im Produktions-, Handel- und Dienstleistungsbereich tätiger Mischkonzern, hat 1994 den Umsatz kräftig um 17 Prozent auf rund 2,8 Milliarden DM gesteigert. In das neue Geschäftsjahr sei man ebenfalls „mit Schwung" gestartet. Im 1. Halbjahr 1995 hätten sich die Umsätze *des Konzerns* im Vergleich zur Vorjahresperiode um fast 23 Prozent auf rund 1,3 Milliarden erhöht.

Type:	turnover
C-name:	Possehl1
Year:	1994
Amount:	2.8e+9DM
Tendency:	+
Diff:	+17%

Type:	turnover
C-name:	Possehl1
Year:	1995/1
Amount:	1.3e+9DM
Tendency:	+
Diff:	+23%

Relationship of IE to other NL-related application areas

(1) Information Retrieval (IR)

Identify and extract documents as answers of an information request.

(2) Passage Retrieval

Identify and extract document snippets as answers of an information request.

(3) Information Extraction (IE)

Identify and extract relevant textual passages used for filling up a pre-defined data record/template.

(4) Textual Question-Answering

Answer an arbitrary question by using textual documents as knowledge base: Fact retrieval, combination of IR & IE.

(5) Text understanding

Interpret texts like humans do: Artificial Intelligence

Interpretation of NL-documents

(1) Information Retrieval (IR)

User

(2) Passage Retrieval

User

(3) Information Extraction (IE)

System (static, pre-defined)

(4) Textual Question/Answering

System (dynamic, facts/relations)

(5) Text understanding

System (complete)

IE is interesting for NLP, because

- IE tasks are well defined
- IE uses real-world text
- IE poses difficult and interesting NLP problems
- IE needs interface specifications between NL and domain knowledge
- IE performance can be compared to human performance on the same task

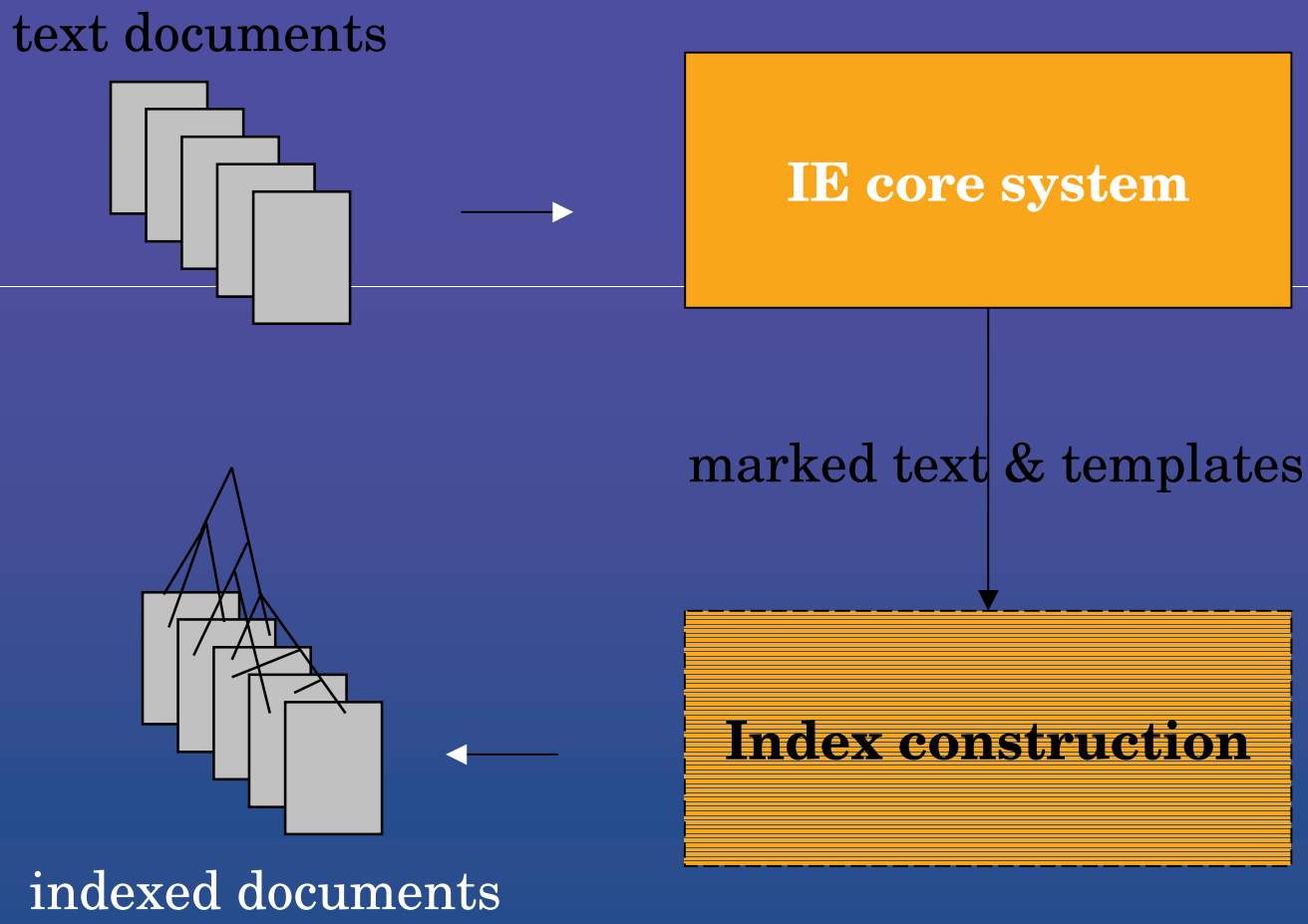
„IE systems are a key factor in encouraging NLP researchers to move from small-scale systems and artificial data to large-scale systems operating on human language.“ (Cowie&Lehnert, 1996)

IE has a high Application Impact

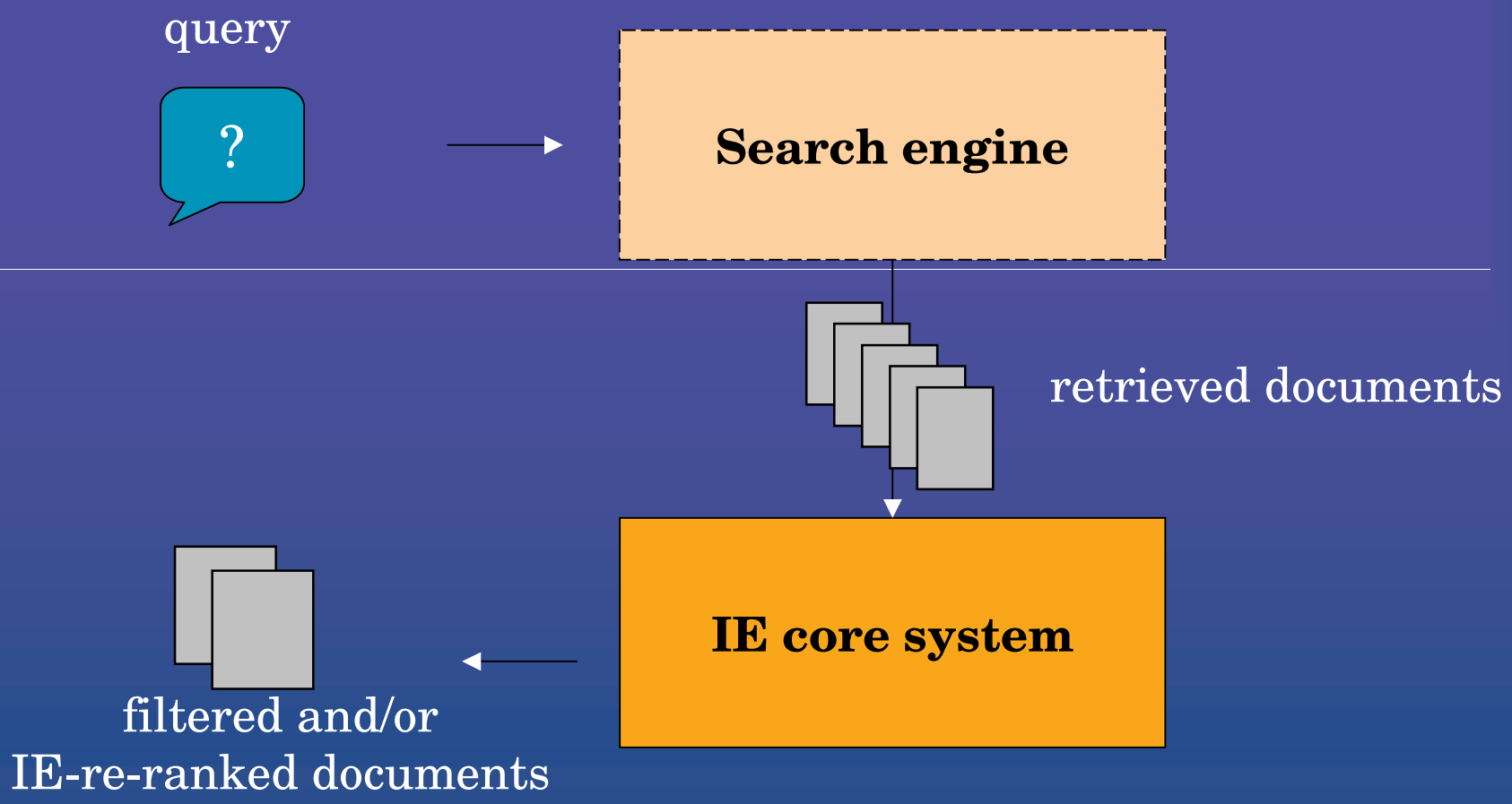
IE interacts with a number of areas

- Text classification: getting fine-grained decision rules
- Information retrieval: construction of sensitive indices which are more closely linked to the actual meaning of a text
- Text mining: improve quality of extracted structured information
- Data-base systems: improve semi-structured DB approaches
- Knowledge-base systems: combine extracted information with KB
- Question Answering: combine IE and full parsing

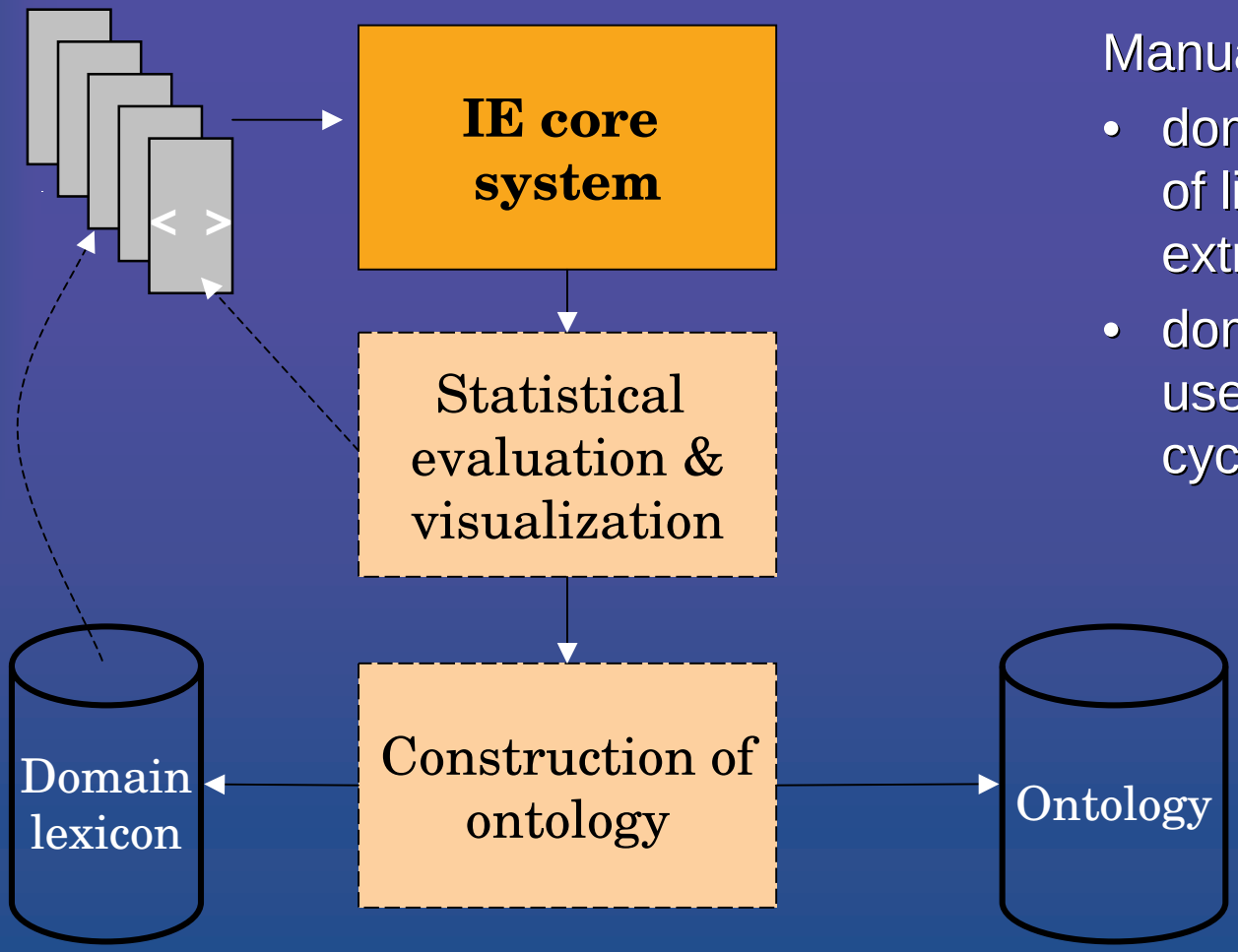
IE improves indexing



IE improves retrieval



IE supports incremental engineering of ontologies



Manual construction of:

- domain ontology on basis of linguistic information extracted from texts
- domain lexicon which is used in next acquisition cycle

IE, Data Mining, Text Mining

- IE (from text documents)

  identify, collect, and normalize prespecified information of a specific domain

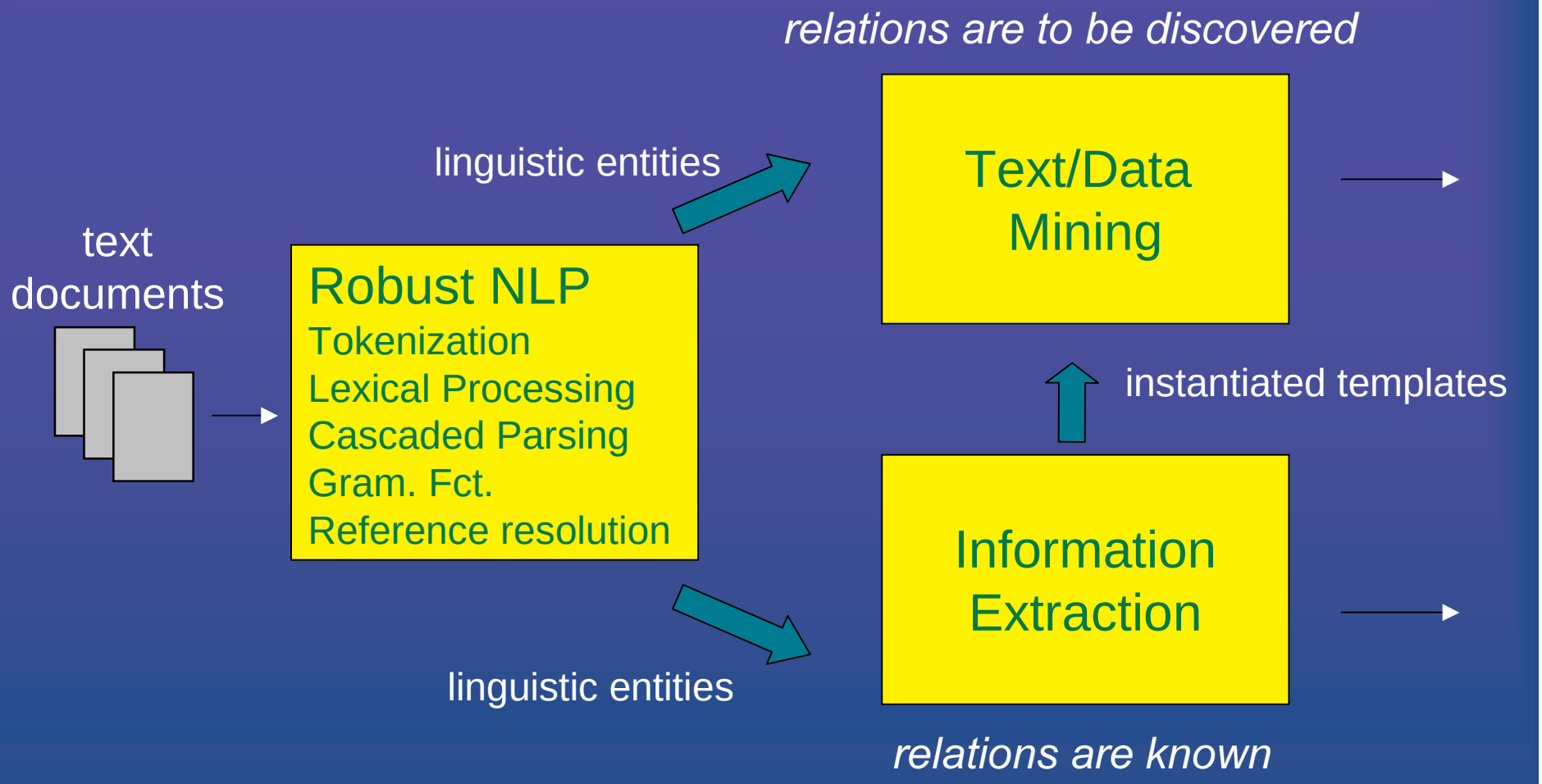
- Data Mining (from structured DB)

  information extraction and discovering of relational links

- Text Mining (from text documents)

  data mining using domain-independent shallow text processing

NL System as a Preprocessor for IE & Text/Data mining



Data - Knowledge - Information

- Main task of an information system
  - Maintain knowledge in digitilised form as data
  - Provide knowledge as useful information to a user

Data – Knowledge - Information

Information = Data + Knowledge.

- Data:
  - recorded facts or figures
- Knowledge:
  - the understanding required to convert data into information and to apply it to real-world situations
- Information:
  - the value derived from data through the application of knowledge

Data vs. Knowledge

28081749 New Dehli‘s latitude

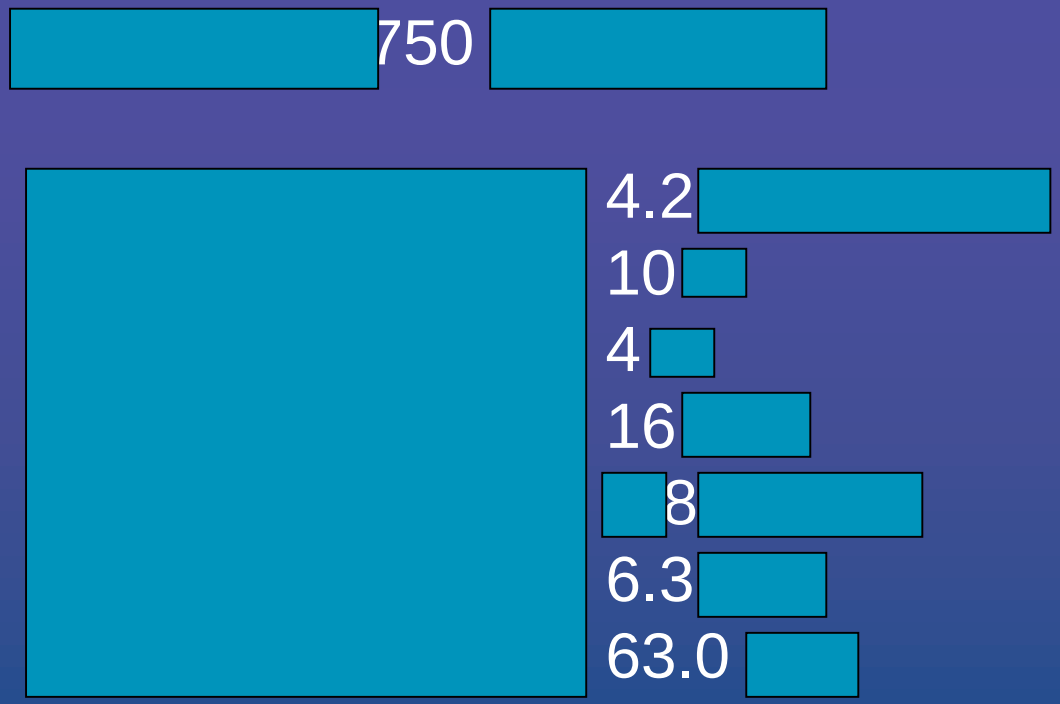
Character sequence Birthday of Goethe

Knowledge are data with meaning, e.g., a property (or feature) of an object (size of a human, name of a company). Note that the same data element might have several possible interpretations.

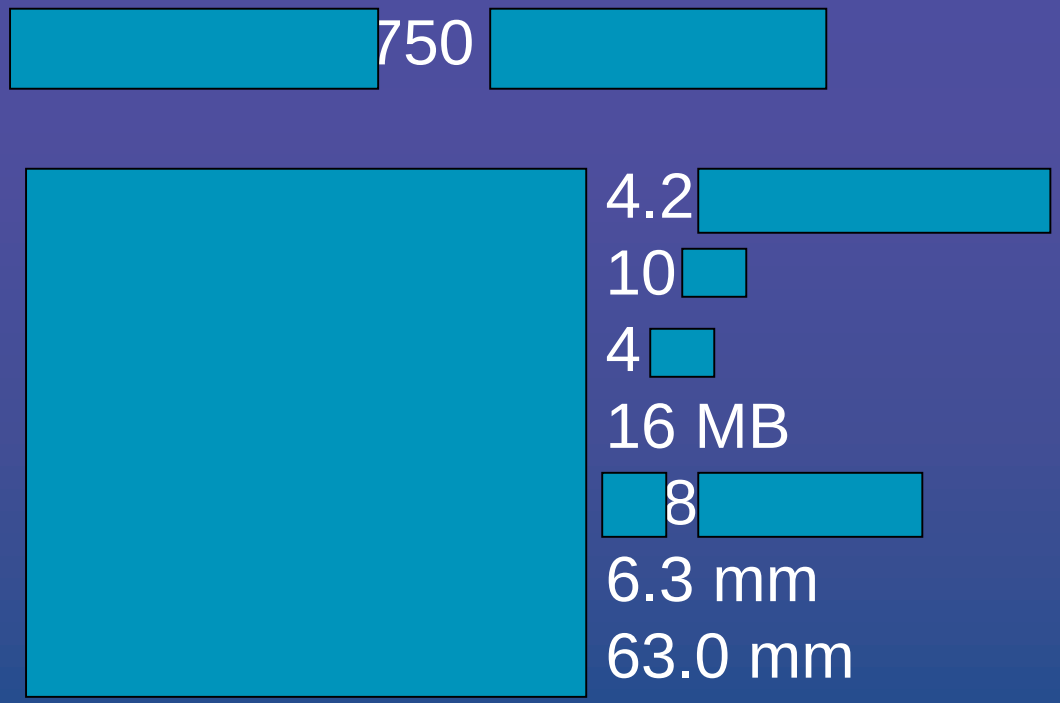
11:15

Time expression game result

?

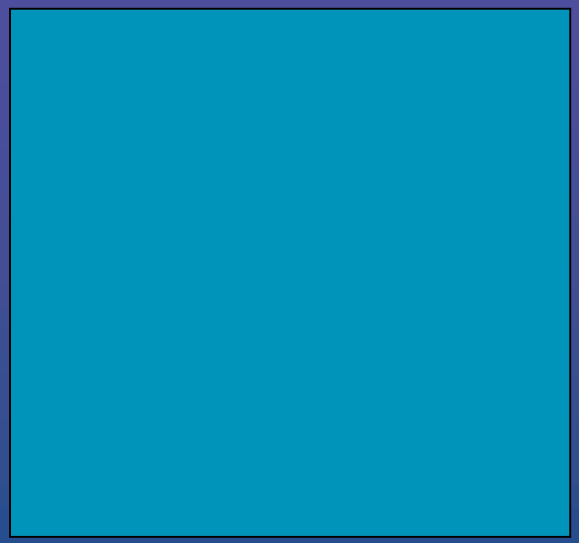

750
4.2
10
4
16
8
6.3
63.0

?


750
4.2
10
4
16 MB
8
6.3 mm
63.0 mm

?

C-750



4.2 megapixels
10 x
4 x
16 MB
F/8-2.8/3.7
6.3 mm
63.0 mm

?

Olympus C-750 Ultra Zoom

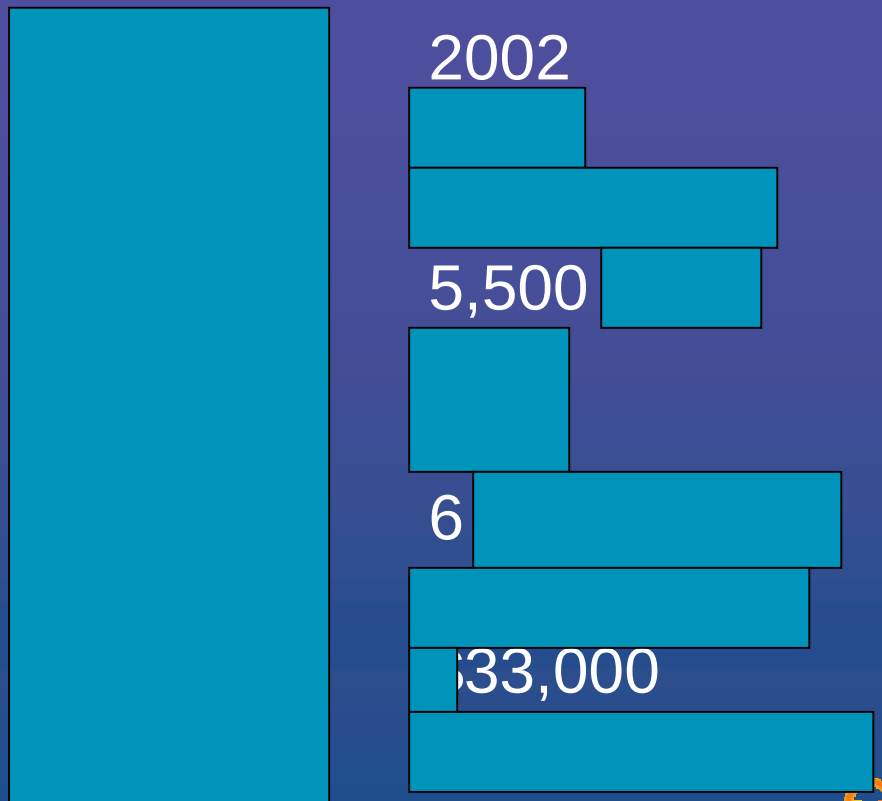
Sensor Resolution	4.2 megapixels
Optical Zoom	10 x
Digital Zoom	4 x
Installed Memory	16 MB
Lens Aperture	F/8-2.8/3.7
Focal Length min	6.3 mm
Focal Length max	63.0 mm

Digital Camera

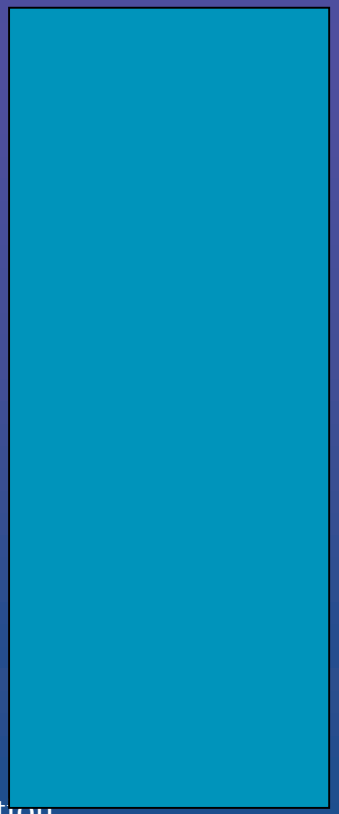
Olympus C-750 Ultra Zoom

Sensor Resolution:	4.2 megapixels
Optical Zoom:	10 x
Digital Zoom:	4 x
Installed Memory:	16 MB
Lens Aperture:	F/8-2.8/3.7
Focal Length min:	6.3 mm
Focal Length max:	63.0 mm

?


2002
5,500
6
33,000

?

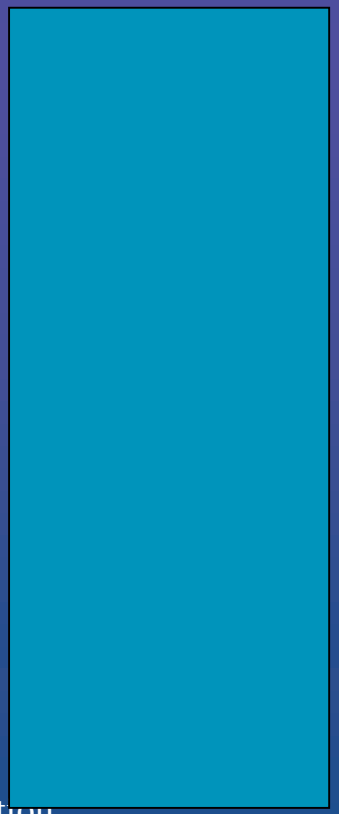


2002

5,500 miles

6 CD changer
keyless entry
$33,000
(916) 972-9117

?



2002
Ford
Thunderbird
5,500 miles
Red
ABS
6 CD changer
keyless entry
$33,000
(916) 972-9117

?

Year	2002
Make	Ford
Model	Thunderbird
Mileage	5,500 miles
Features	Red
	ABS
	6 CD changer
	keyless entry
Price	$33,000
Phone	(916) 972-9117

Car Advertisement

Year	2002
Make	Ford
Model	Thunderbird
Mileage	5,500 miles
Features	Red
	ABS
	6 CD changer
	keyless entry
Price	$33,000
Phone	(916) 972-9117

Knowledge vs. Information

- Knowledge:
  - a model of the world (structural and functional properties of the real world)
- Information:
  - is that part of knowledge which is used to solve a certain problem (Information System view).
  - information only exists in concrete problem situations („What is the new email address of Dan?“).
- Information systems extract that knowledge „just in time“, a user needs in context of a given situation.
  - If the information search is done, then the information is unnecessary.
  - Seen so, information need not necessarily be stored; only if it is new knowledge. In this case information turned to knowledge.

Additional Aspects of Information

- Information theory (Shannon):
  - the information content of a message depends on its propability
- Information:
  - that part of a message which is new (low degree of redundancy), and interpretable (low degree of noise)
- Information only exists relative to an information consumer/request
- Information must be interpreted relative to already existing information
- There is no communication without information

NLP as normalization

- Template descriptions as typed objects
  - [person-in: type_of_person_name]
- Core problem for building IE systems
  - Identify general mapping between text fragments and template descriptions
- IE as normalization:
  - What are the possible ways, how a template description can be expressed in NL?
  - Determine all possible textual paraphrases for an object
- Close relationship to the problem of lexical choice in Natural Language Generation

NL analysis as step-wise normalization

- Tokenization

  9.11.2001, 11/9/2000 ⇨
  {day: 9, month: 11, year: 2000}

- Morphological analysis:
  - Determination of lexical stems
    - Inflection:
      supporting ⇨ *to support*
      Häuser ⇨ *haus*
    - German compounds:
      Informationstechnologiezentrum ⇨
      {*Information, Technologie, Zentrum*}

NL analysis as step-wise normalization

- Special phrases (word groups):
  - date and time expressions:

    18.12.98 und *Friday, December the 18th 1998*

    ⇨ <type=date, year=1998, month=12, day=18, weekday=5>

  - proper names: persons, institutions, companies, locations, products, …

  - number expressions, addresses, mathematical expressions, …

NL analysis as step-wise normalization

- General phrases:
  - nominal phrases, prepositional phrases, verb groups

    For the new economy

    ⇨ <head=for, comp=<head=economy, quant=def, mod=new>>

- complex flat sentence structure

- domain specific templates (integration of ontology)

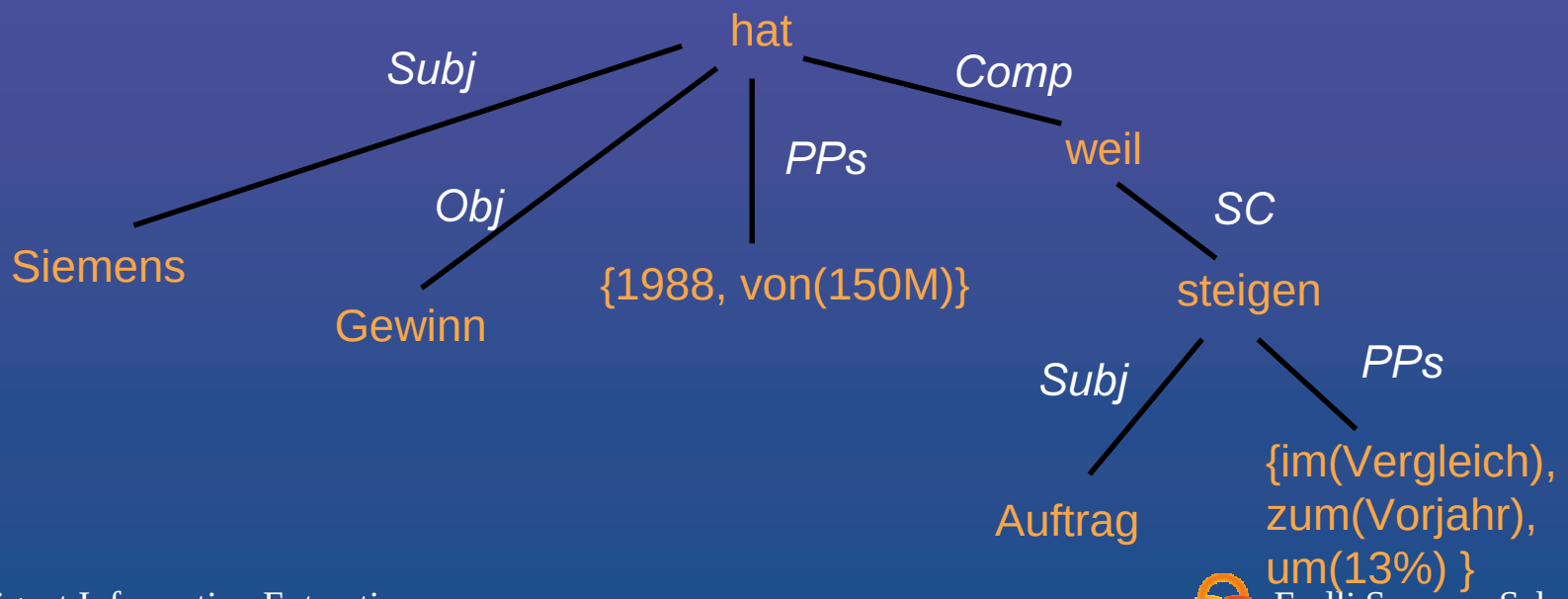
type	=	turnover	c-name	=	Possehl1
year	=	1995/1	amount	=	1.3e+9DM
tendency	=	+	diff	=	+23%

Underspecified functional description for sentences

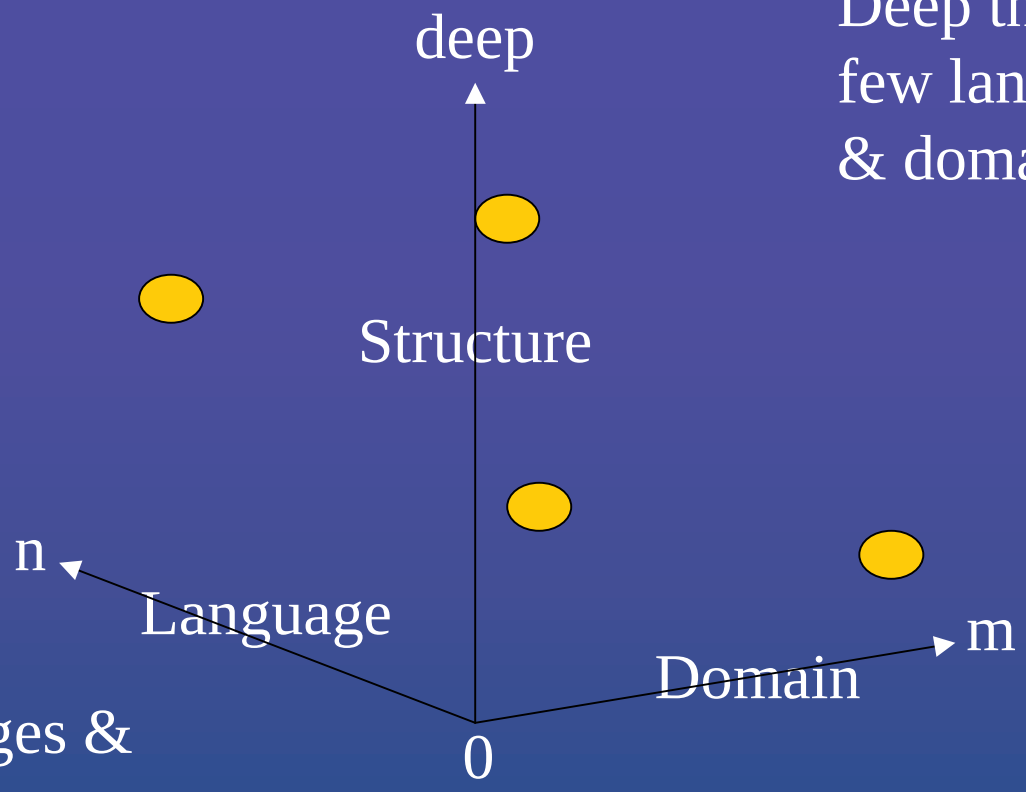
Flat dependency-based structure, only upper bounds for attachment and scoping:

[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM],
[$_{Comp}$weil] [$_{NP}$die Aufträge] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].

"The siemens company has made a revenue of 150 million marks in 1988, since the orders increased by 13% compared to last year."



Complexity of IE



Deep then
few languages
& domains

Many languages &
domains
then very shallow

Two Approaches to Building Extraction Systems

- Knowledge engineering approach
  - Grammars are constructed by hand
  - Domain patterns are discovered by a human expert through introspection and inspection of a corpus
  - Much laborious tuning and „hill climbing“
- Automatically Trainable Systems
  - Use statistical methods when possible
  - Learn rules from annotated corpora
  - Learn rules from interaction with user

Knowledge Engineering

- Advantages
  - With skill and experience, good performing systems are conceptually not hard to develop
  - The best performing systems have been hand crafted (still true for scenario patterns)
- Disadvantages
  - Very laborious development process
  - Domain adaptation might require re-configuration
  - Needs experts which have both, linguistic & domain expertise

Trainable Systems

- Advantages
  - Domain portability is relatively straightforward
  - System expertise is not required for customization
  - Data driven rule acquisition ensures full coverage of examples
- Disadvantages
  - Training data may not exist, and maybe very expensive to acquire
  - Large volume of training data may be required
  - Changes to specifications may require re-annotation of large quantities of training data

What works best?

- Use rule-based approach when
  - Resources (e.g., lexicons,lists) are available
  - Rule writers are available
  - Training data scarce or expensive to obtain
  - Extraction specs likely to change
  - Highest possible performance is critical
- Use trainable approach when
  - Resources unavailable
  - No skilled rule writers are available
  - Training data is cheap and plentiful
  - Good performance is adequate for the task

Architecture of Extraction Systems

- Domain-independent NL tools necessary
  - Major issue: robustness & efficiency
- Clean interface between domain-independent tools and domain-dependent
  - Domain modeling
  - Main Task: disambiguation
  - Easy adaptation of NL tools

Evaluation

- How can we compare human and system performance?
- How can we measure and compare different methods?
- What can we learn for future system building?

Evaluation Forums for IE

- Message Understanding Conference MUC
  - Languages considered:
    - English, Chinese, Spanish, Japanese
  - First round: 1987
- Automatic Content Extraction ACE
  - Languages considered
    - English, Chinese, Arabic
  - Pilot phase: 1999, Start phase: Oct. 2000

The Message Understanding Conference (MUC)

- Sponsored by the Defense Advanced Research Projects Agency (DARPA)
  - Developed methods for formal evaluation of IE systems
- In the form of a competition
  - participants compare their results with each other and against human annotators‘ key templates.
- Short system preparation time to stimulate portability to new extraction problems:
  - 1 month to adapt the system to the new scenario before the formal run.

MUC: Evaluation procedure

- Corpus of training texts
- Specification of the IE task
- Specification of the form of the required output
- Keys:
  - ground truth-human produced responses in output format
- Evaluation procedure
  - Blind test
  - System performance *automatically* scored against keys

MUC Tasks

- MUC-1 (87) and MUC-2 (89)
  - Messages about naval operations
- MUC-3 (91) and MUC-4 (92)
  - News articles about terrorist activity
- MUC-5 (93)
  - News articles about joint venture and microelectronics
- MUC-6 (95)
  - News articles about management changes
- MUC-7 (97)
  - News articles about space vehicle and missile launches

Events – Relations - Arguments

Examples of events or relationships to extract	Examples of their arguments/slots
Terrorist attacks (MUC-3) (example corpus/output file)	Incident_Type, Date , Location, Perpetrator, Physical_Target, Human_Target, Effects, Instrument
Changes in corporate executive management personnel (MUC-6)	Post, Company, InPerson, OutPerson,VacancyReason,OldOrganisation, NewOrganisation
Space vehicles and missile launch events (rocket launches) (MUC-7)	Vehicle_Type, Vehicle_Owner, Vehicle_Manufacturer, Payload_Type, Payload_Func, Payload_Owner,Payload_Origin,Payload_Target, Launch, Date, Launch Site, Mission Type, Mission Function, etc.

Evaluation Metrics

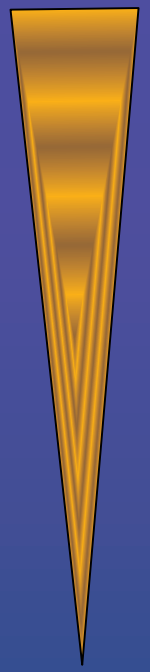
- Precision and recall:
  - Precision: correct answers/answers produced
  - Recall: correct answers/total possible answers
- F-measure
  - Where β is a parameter representing relative importance of P & R:

$$F = \frac{(\beta^2 + 1)PR}{(\beta^2 P + R)}$$

  - E.g., β=1, then P&R equal weight, β=0, then only P
- Current State-of-Art: F=.60 barrier

MUC Extraction Tasks

- Named Entity task (NE)
- Template Element task (TE)
- Template Relation task (TR)
- Scenario Template task (ST)
- Co-reference task (CO)



Named Entity Task (NE)

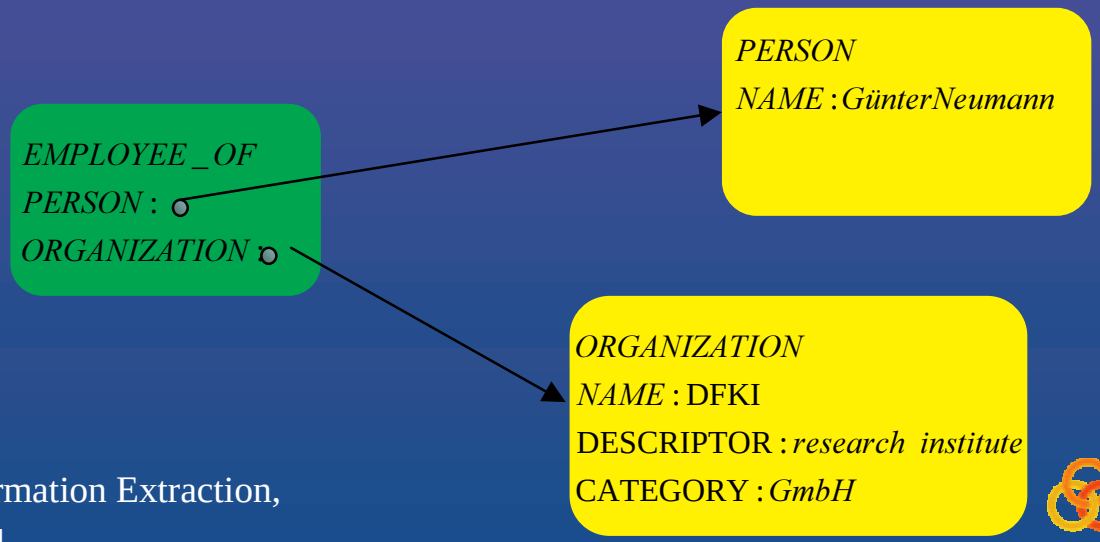
Mark into the text each string that represents a person, organization, or location name, or a date or time, or a currency or percentage figure (this classification of NEs reflects the MUC-7 specific domain and task).

Template Element Task (TE)

Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text (TE consists in generic objects and slots for a given scenario, but is unconcerned with relevance for this scenario)

Template Relation task (TR)

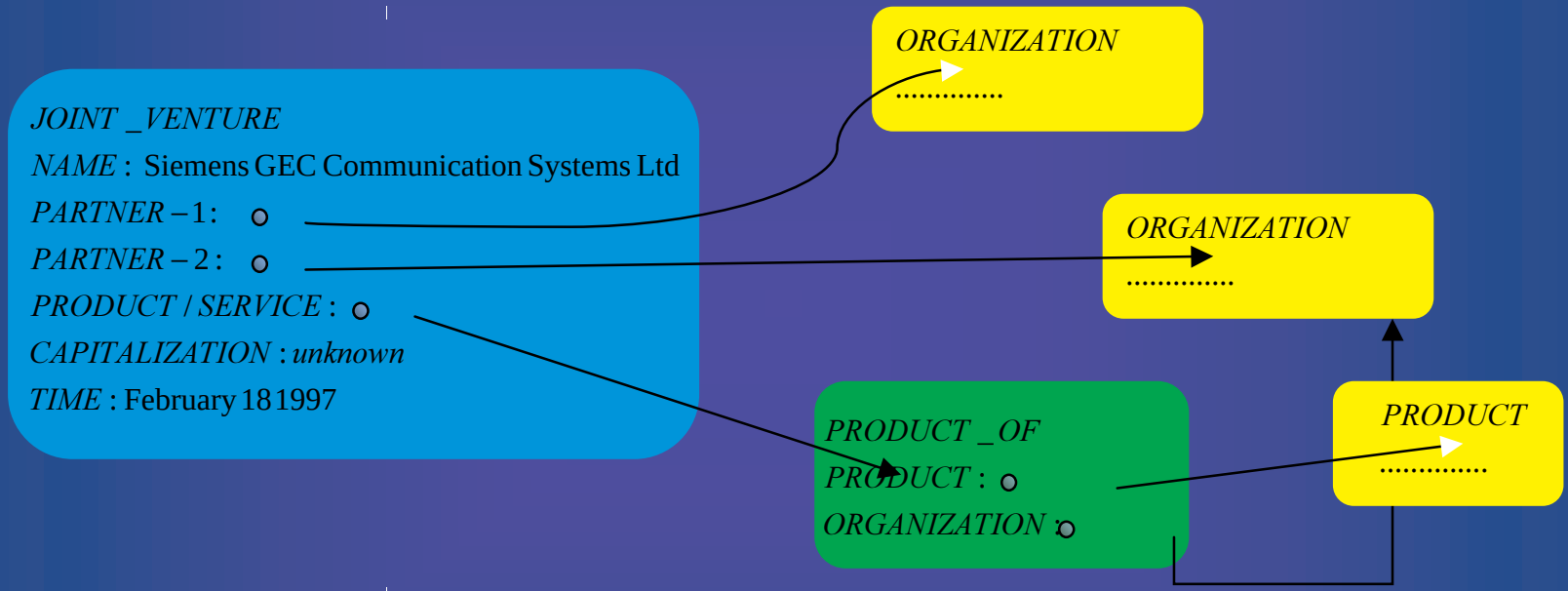
Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain-independent relationships between entities identified by TE)



Scenario Template task (ST)

Extract prespecified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations)

ST example


JOINT _VENTURE
NAME : Siemens GEC Communication Systems Ltd
PARTNER – 1 :
PARTNER – 2 :
PRODUCT / SERVICE :
CAPITALIZATION : unknown
TIME : February 18 1997
ORGANIZATION
.....
ORGANIZATION
.....
PRODUCT _OF
PRODUCT :
ORGANIZATION :
PRODUCT
.....

Coreference task (CO)

Capture information on corefering expressions, i.e. all mentions of a given entity, including those marked in NE and TE (Nouns, Noun phrases, Pronouns)

An Example

From: Tablan, Ursu, Cunningham, eurolan 2001

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

- NE: *red rocket*, *Tuesday*, *Dr. Big Head, Dr. Head,* and *We Build Rockets Inc.*
- CO: *it* refers to the rocket; *Dr. Head* and *Dr. Big Head* are the same
- TE: the rocket is *shiny red* and Head's *brainchild*
- TR: Dr. Head *works for* We Build Rockets Inc.
- ST: a *rocket launching event* occured with the various participants.

Scoring templates

- Templates are compared on a slot-by-slot basis
  - Correct: response = key
  - Partial: response ≈ key
  - Incorrect: response ≠ key
  - Spurious: key is blank
    - overgen=spurious/actual
  - Missing: response is blank

Tasks evaluated in MUC 3-7

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

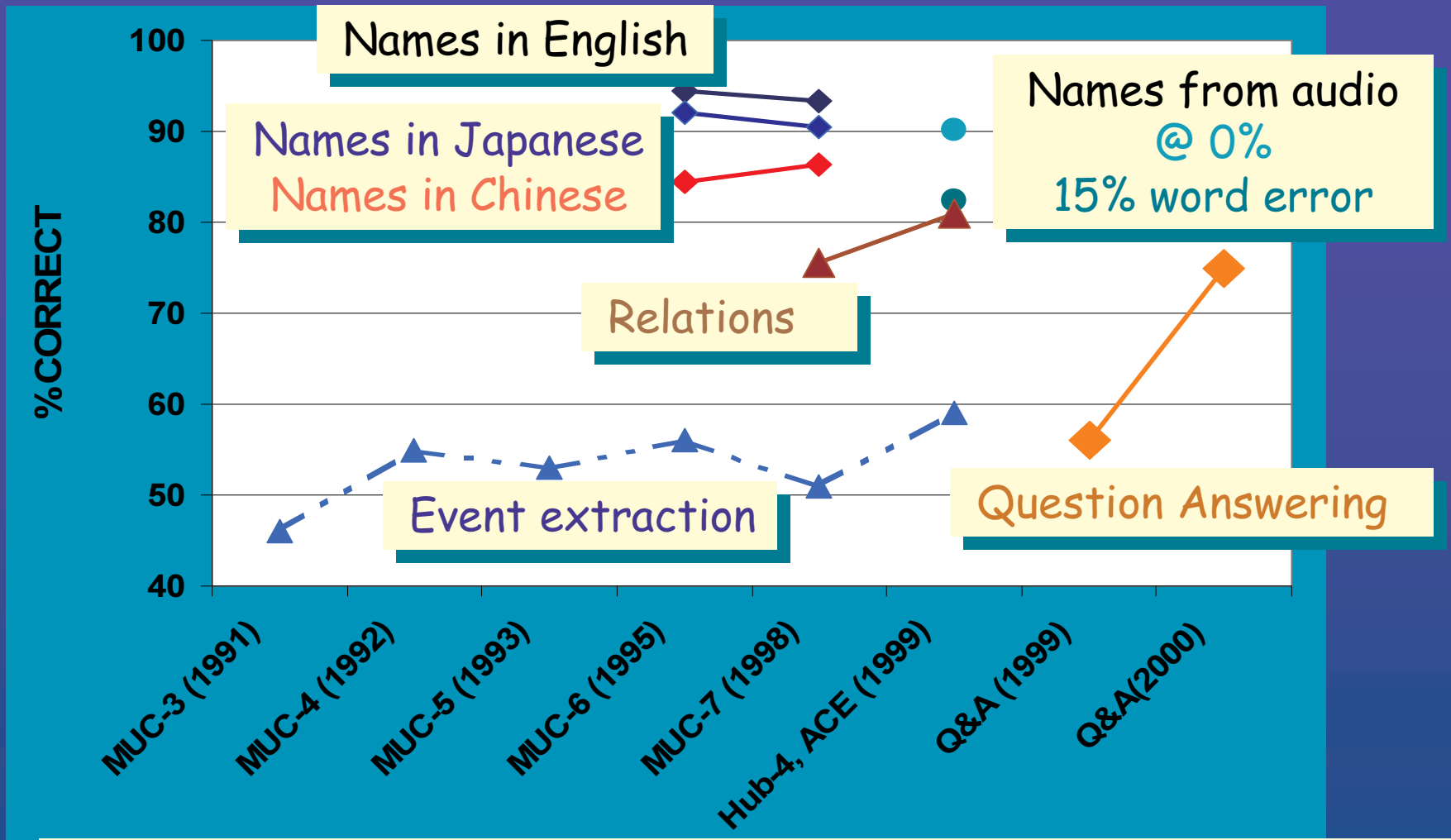
Maximum Results Reported in MUC-7

Measure\Task	NE	CO	TE	TR	ST
Recall	92	56	86	67	42
Precision	95	69	87	86	65

Human on NE task	F	R	P
Annotator 1	98.6	98	98
Annotator 2	96.9	96	98

Human on ST task: ~ 80 % F

Progress in Information Extraction



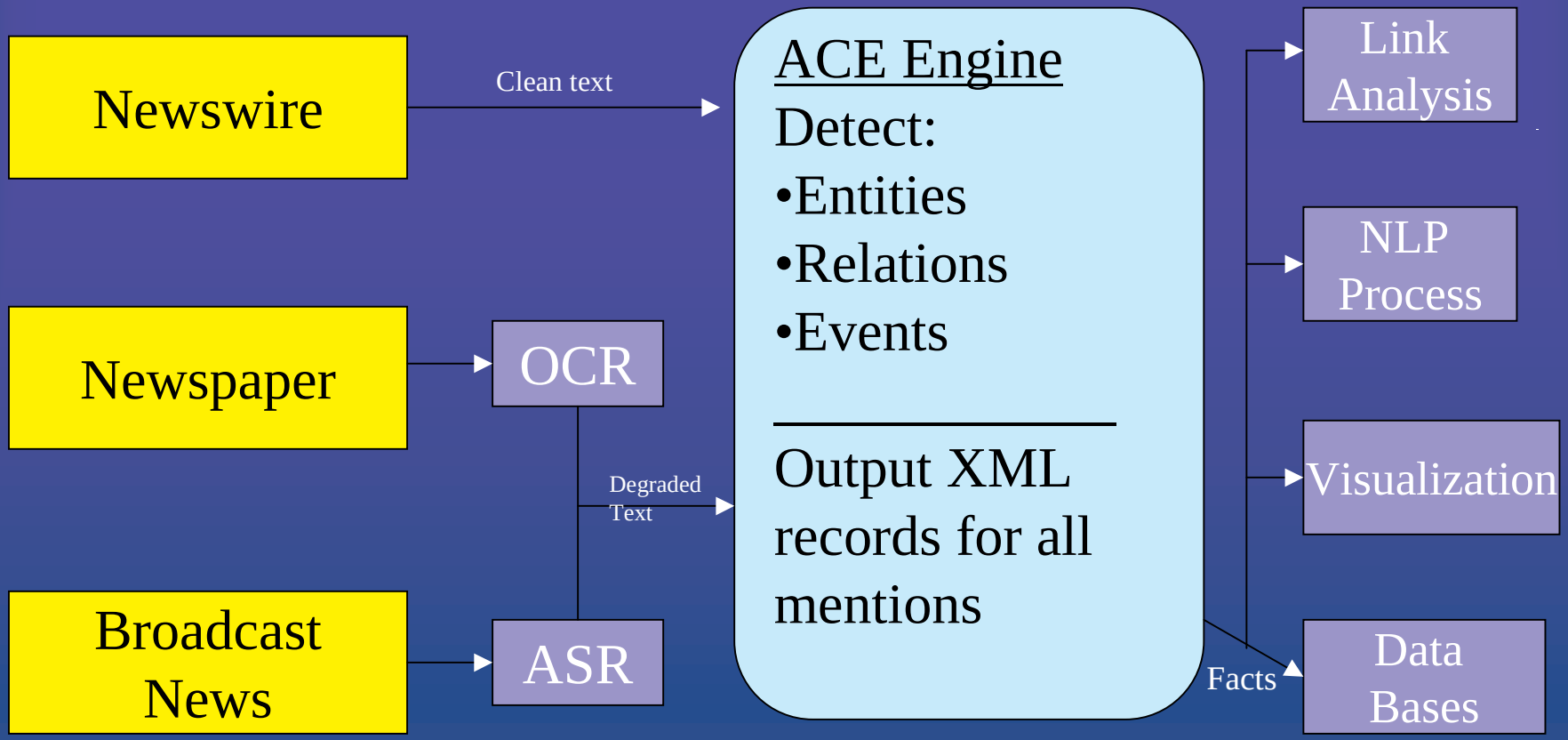
There have been many evaluations of natural language components...

The ACE Program

Partially based on Douglas Appelt, 2003

- "Automated Content Extraction"
- 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

The Technical Approach



Objectives and Performance Goals of ACE

- Objectives
  - Extract Info from Texts of Varying Quality
  - Detect Unique entities, relations, events
    - Find all mentions within documents
    - Collect all mentions by object
  - Track entities within & across documents
  - Output XML for follow-on processes
- Performance Goals
  - Extract 95% of the value in document

Core Mission: Information Gathering

- Information content is main interest of human language text
  - Semantics drives information gathering
  - Syntax is the vehicle for organizing the information
- ACE systems provide NL understanding
  - Detect each entity, relation, and event of specific type
  - Recognize all mentions of entities, relations & events
  - Resolve all mentions to the proper entity, relation, or event
- Convert information in human language into structured data
  - Extract semantics of communication
  - Output in ACE program format
- Structured data supports real world modeling & analysis

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*
- Attributes - Timeless unary properties of entities (e.g. Name)
- Temporal points and intervals
- Relations - Properties that hold of two entities over a time interval
- Events - A particular kind of relation among entities implying a change in relation state at the end of the time interval.

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
    - Co-referring noun phrases
    - Relations and events derived from different mentions, but conveying the same meaning

Choosing an Ontology for IE Semantics

- Ordinary native speakers should be able to annotate text with minimal training.
- People should have well-developed intuitions about type classification
  - Is a “museum” an organization or facility? (A FOG?)
- People should have well-developed intuitions about entity coreference
  - “Peace in the Middle East”
- Entities should be extensional, not abstract, generic, counterfactual, or fictional

The ACE Ontology and Annotation Standards

- Documents available online
  - http://www.ldc.upenn.edu/Projects/ACE/
  - Entity standards
  - Relations standards
  - Proposed event standards still under development

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

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

Entity and Mention Example

[COLOGNE, [Germany]] (AP) _ [A [Chilean] exile] has filed a complaint against [former [Chilean] dictator Gen. Augusto Pinochet] accusing [him] of responsibility for [her] arrest and torture in [Chile] in 1973, [prosecutors] said Tuesday.
[The woman, [a Chilean] who has since gained [German] citizenship], accused [Pinochet] of depriving [her] of personal liberty and causing bodily harm during [her] arrest and torture.

Person
Organization
Geopolitical Entity

Relations

- Relations hold between two entities over a time interval.
- Relations may be “timeless” or temporal interval is not specified
- Relations have inertia, I.e. they don’t change unless a relevant event happens.

Explicit and Implicit Relations

- Many relations are true in the world. Reasonable knowledge 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 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…

Summary

- Motivation for a semantic theory is a practical one driven by database filling needs
- Pick a limited ontology of core concepts, and build out, motivated by application needs
- Address a broad spectrum of semantic problems, but from a limited ontology that simplifies data annotation issues.