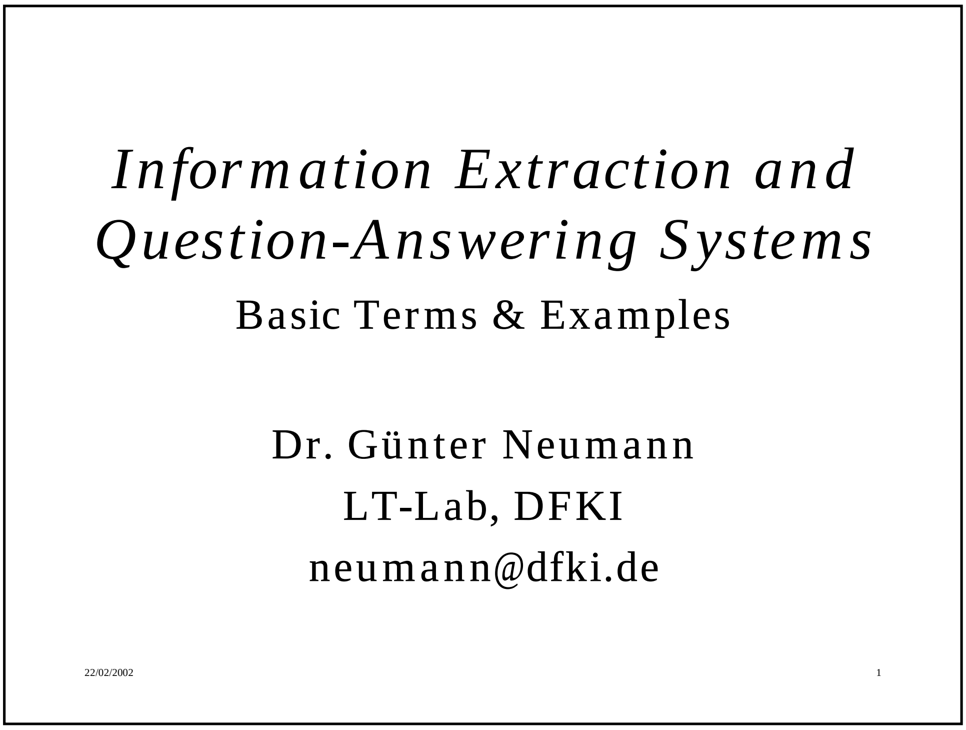
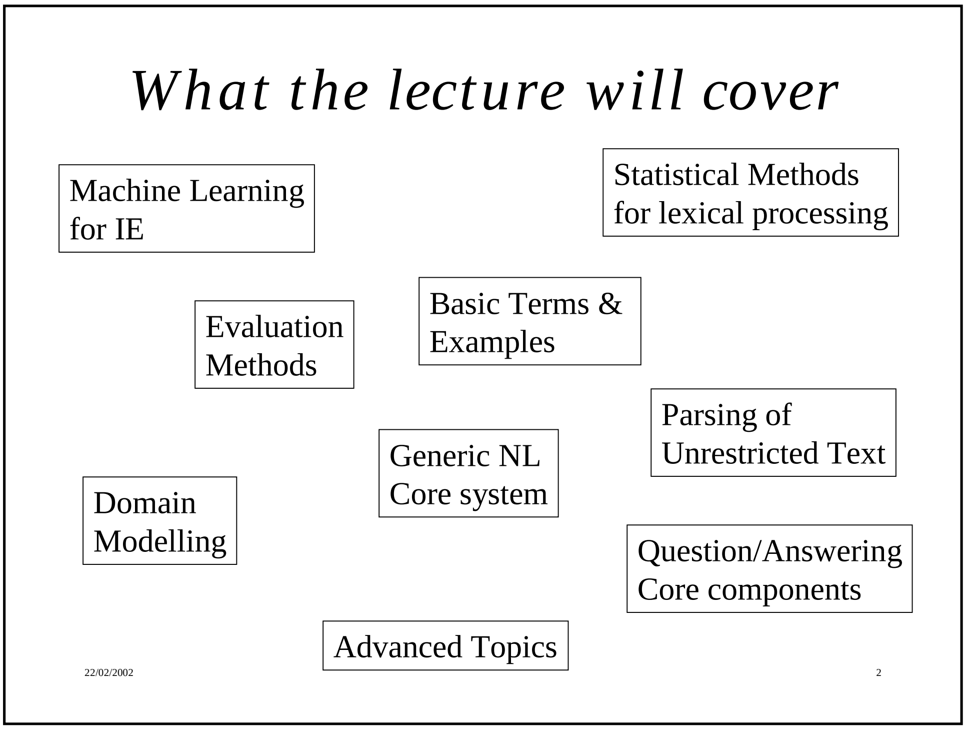


Information Extraction and Question-Answering Systems

Basic Terms & Examples

Dr. Günter Neumann
LT-Lab, DFKI
neumann@dfki.de



What the lecture will cover

Machine Learning for IE

Statistical Methods for lexical processing

Evaluation Methods

Basic Terms & Examples

Parsing of Unrestricted Text

Generic NL Core system

Domain Modelling

Question/Answering Core components

Advanced Topics

Basic Terms & Examples

We will focus on extraction of information from NL texts.

- Information Retrieval vs. Information Extraction vs. Answer Extraction
- Data vs. Information
- NLP as normalization

Information retrieval (IR)

- Deals with representation, storage, organization of and access to information items (the user‘s interest).
- Information items are translated to a *query* consisting of keywords (word forms) which summarizes the description of the user information needed.
- Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user.

Examples are Search Engines, like Google™ which retrieve documents on the Web containing the keywords, and return a *ranked* list of relevant indices to documents.
Search Engines are *word form* based and often analyse the link structure of the WWW.

Find more information in
Baeza-Yates & Ribeiro-Neto,
Modern Information Retrieval,
Addision Wesley, 1999 url

IR and NLP

- IR usually deals with NL text which is not always well structured and could be semantically ambiguous
- IR deals with very large sets of documents
  - ➢ High amount of robusteness, efficiency
  - ➢ Domain-independent & multi-linguality
- IR considers NL text mainly from a lexical view
  - ➢ Identifying possible word forms
  - ➢ Elimination of stop words (e.g., closed class word, *the*, *der*, *ab*, *zu*, ...)
  - ➢ Stemming (e.g., *supporting*, *supported* ➜ *support*)
  - ➢ Selection of index terms
  - ➢ Term categorization structure

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: Company‘s turnover

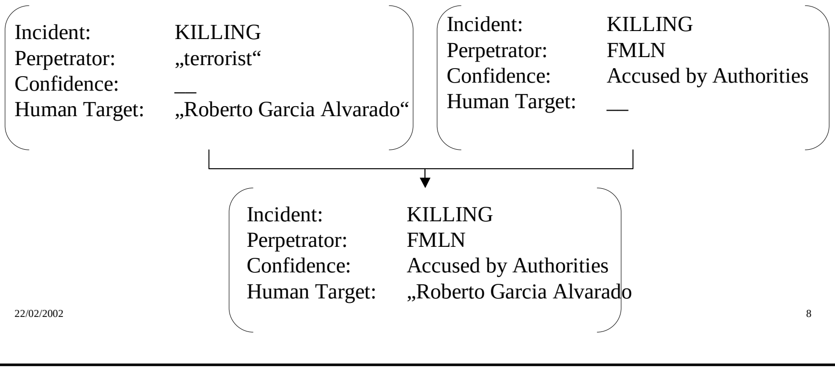
Lübeck (dpa) - Die Lübecker Possehl-Gruppe, ein im Produktions-, Handel- und Dienstleistungsbereich tätiger Mischkonzern, hat 1994 den Umsatz kräftig um 17 Prozent auf rund 2,8 Milliarden DM gesteigert. In das neue Geschäftsjahr sei man ebenfalls „mit Schwung“ gestartet. Im 1. Halbjahr 1995 hätten sich 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%

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: Scientific papers on molecular biology

Results: We have determined the crystal structure of a triacylglycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/beta-hydrolase fold and a catalytic triad comprising of residues Ser87, His286 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site. This is in contrast to the structure of PgL and Pet in which the active site is buried under a closed or partially opend ´lid´, respectively.

Filled templates for protein structure

<Residue-56>:=

ResidueType:	SERINE
ResidueNo:	87
InProtein:	<Protein-2>
Site/Function:	„active site“, „catalytic“, „interfacial activation“, „calcium-binding site“
SecondStruct:	alpha-helix
Region:	´lid´
Article:	<Article-1>

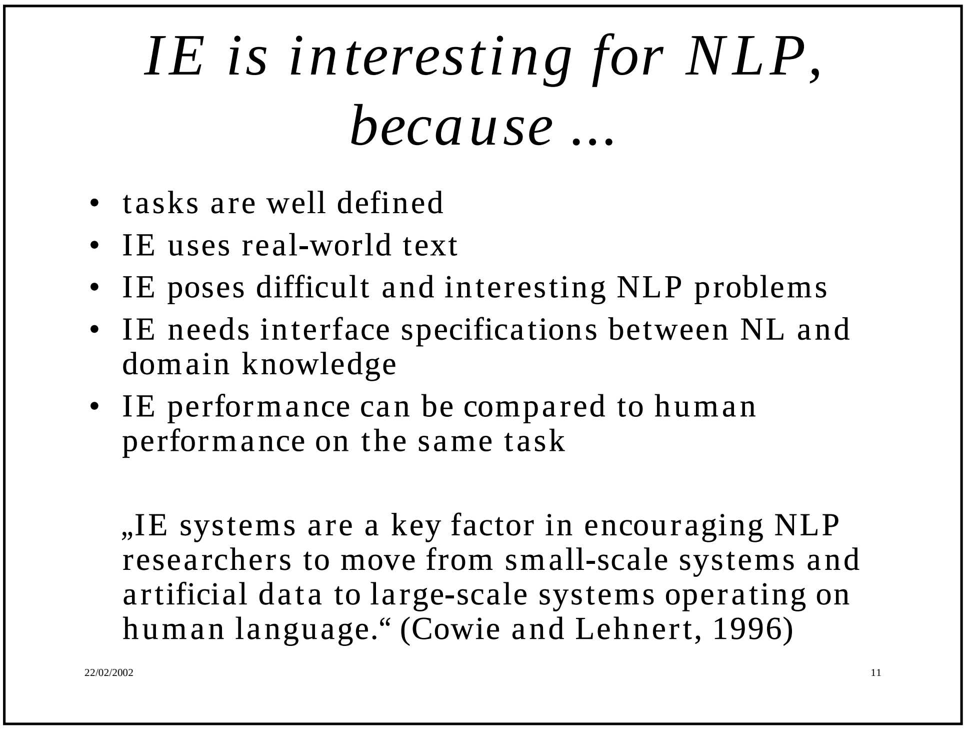
<Protein-2>:=

Name:	triacylglycerol lipase
ScopClass:	lipase
PDBCode:	1LGY
InSpecies:	<Species-4711>

< Species-4711>:=

Name:	pseudomonas cepacia
NameType:	SCIENTIFIC

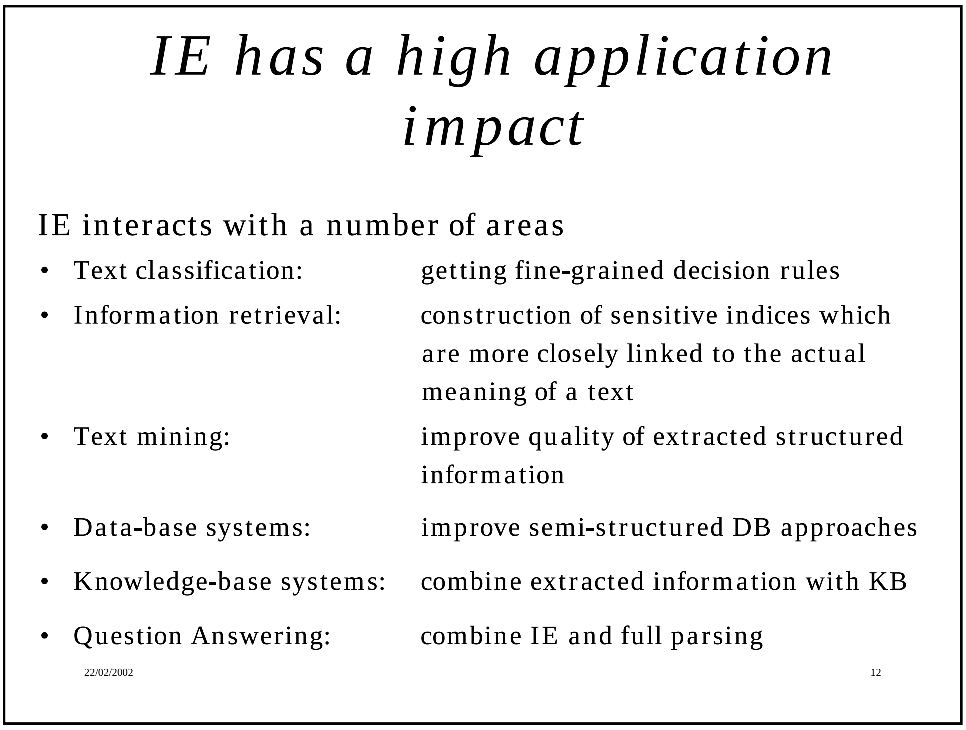
Cf. Humphreys, Demetriou, Gaizauskas (2000), url



IE is interesting for NLP, because ...

- 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 and Lehnert, 1996)

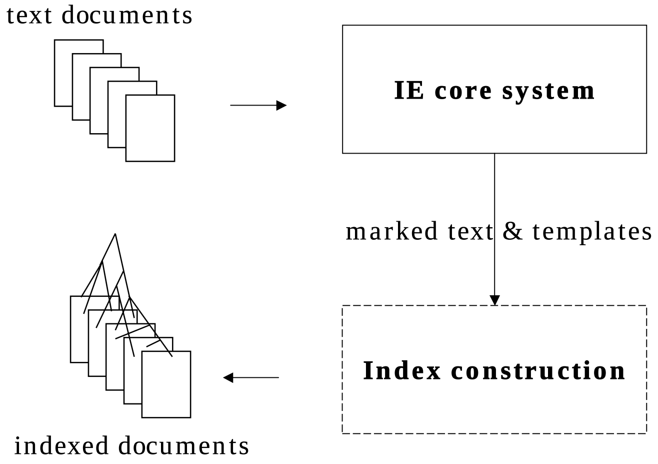


IE has a high application impact

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



text documents

IE core system

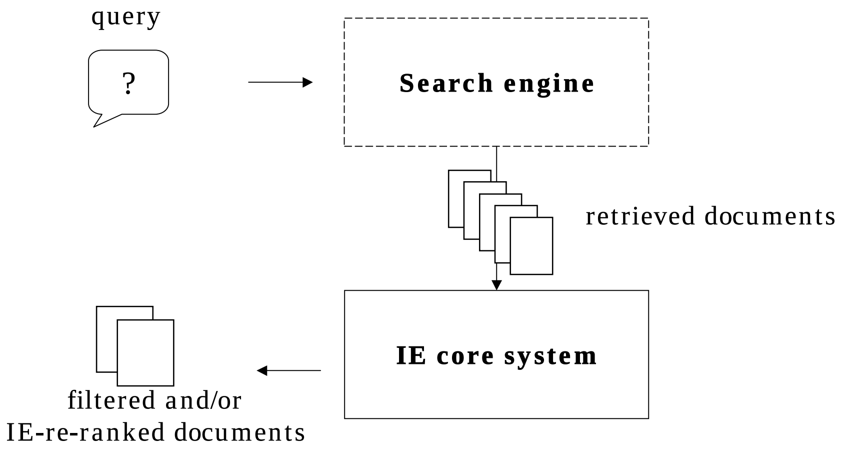
marked text & templates

Index construction

indexed documents

22/02/2002 13

IE improves retrieval



query

?

Search engine

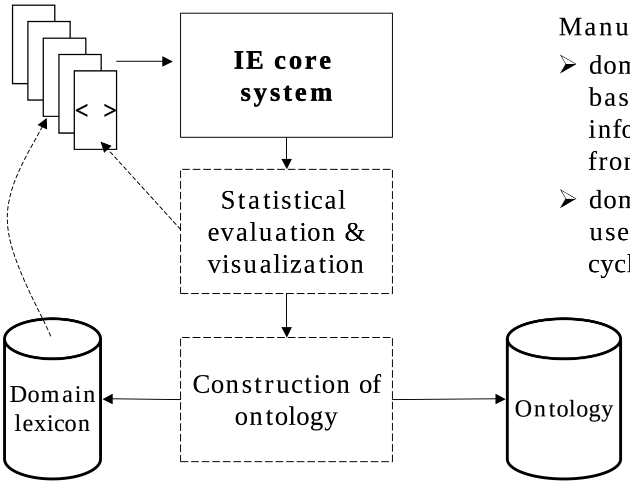
retrieved documents

IE core system

filtered and/or
IE-re-ranked documents

22/02/2002 14

IE supports incremental engineering of ontologies



IE core system

Statistical evaluation & visualization

Construction of ontology

Domain lexicon

Ontology

< >

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

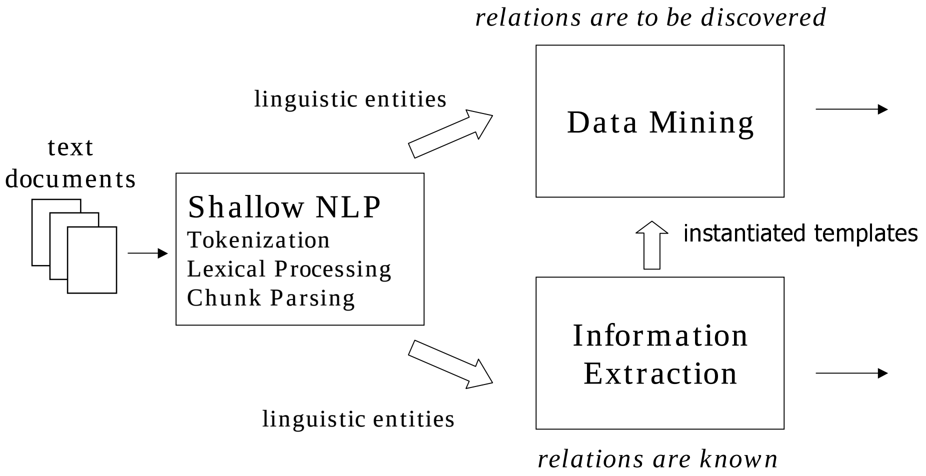
- Data Mining (from structured DB)

  information extraction and discovering of relational links

- Text Mining (from text documents)

  data mining using domain-independent shallow text processing

Shallow NL system as a preprocessor for IE & Text mining



relations are to be discovered

linguistic entities

Data Mining

text documents

Shallow NLP
Tokenization
Lexical Processing
Chunk Parsing

instantiated templates

Information Extraction

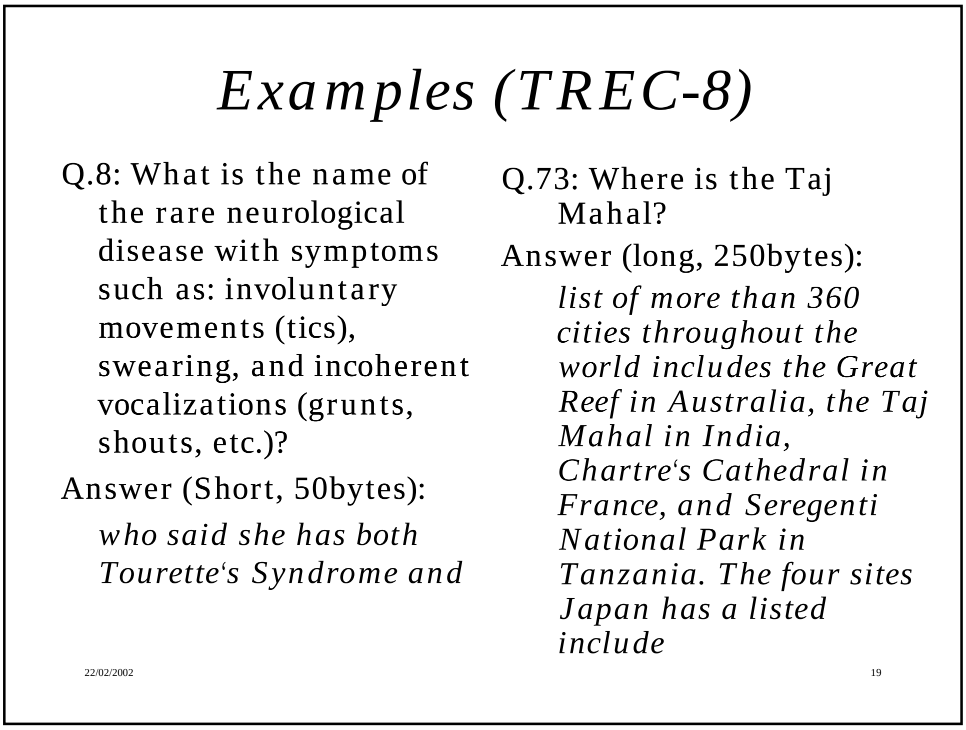
linguistic entities

relations are known

Textual Question Answering

Given a NL query, find the answer by returning a small fragment of text, where the answer actually lies.

- Identify answers of question in large collections of on-line documents
- Highlight only a short piece of text, accounting for the answer.
- Questions expressed in natural language, are not constrained to a specific domain or type of question (i.e. more then *who, what, whom, where, why* Q-types)



Examples (TREC-8)

Q.8: What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

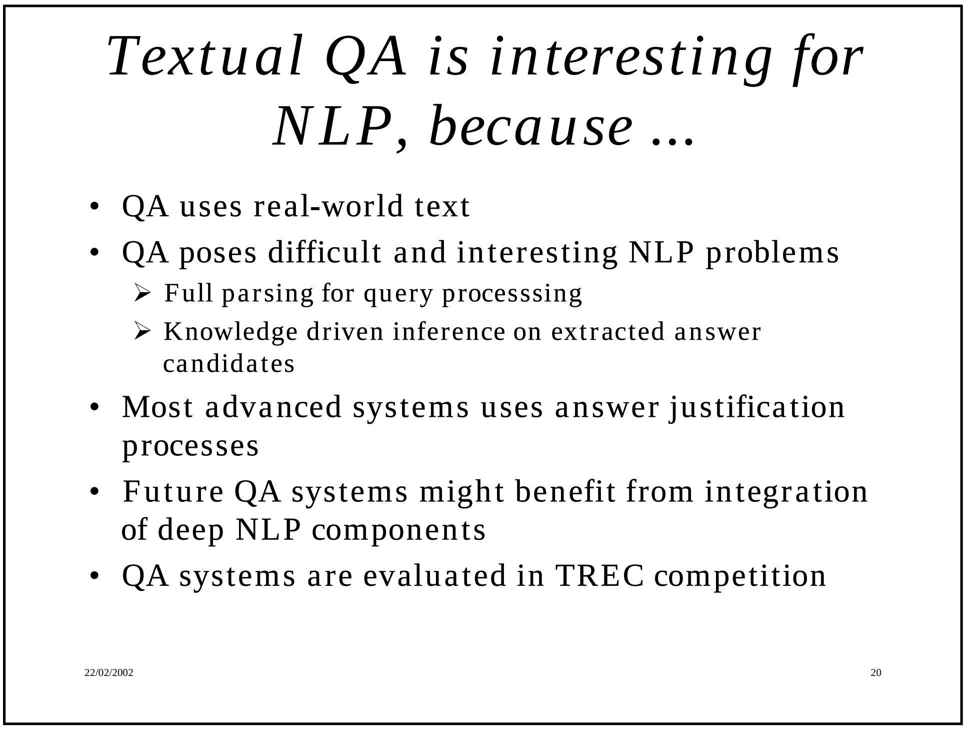
Answer (Short, 50bytes):

who said she has both Tourette‘s Syndrome and

Q.73: Where is the Taj Mahal?

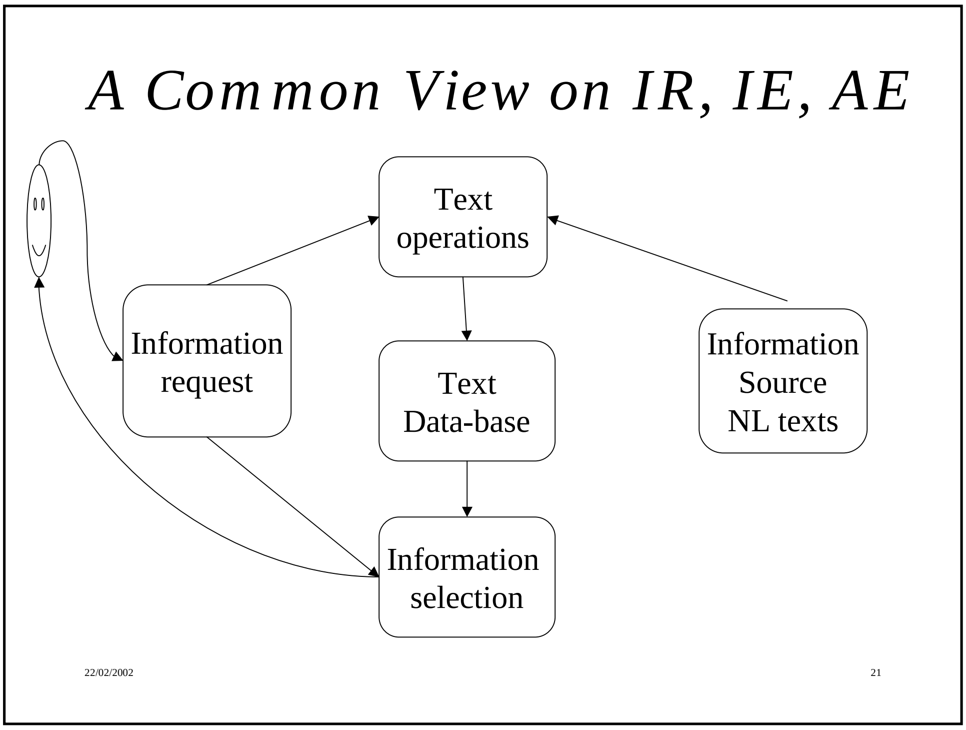
Answer (long, 250bytes):

list of more than 360 cities throughout the world includes the Great Reef in Australia, the Taj Mahal in India, Chartre‘s Cathedral in France, and Seregenti National Park in Tanzania. The four sites Japan has a listed include



Textual QA is interesting for NLP, because ...

- QA uses real-world text
- QA poses difficult and interesting NLP problems
  - ➢ Full parsing for query processsing
  - ➢ Knowledge driven inference on extracted answer candidates
- Most advanced systems uses answer justification processes
- Future QA systems might benefit from integration of deep NLP components
- QA systems are evaluated in TREC competition



A Common View on IR, IE, AE

Text operations

Information request

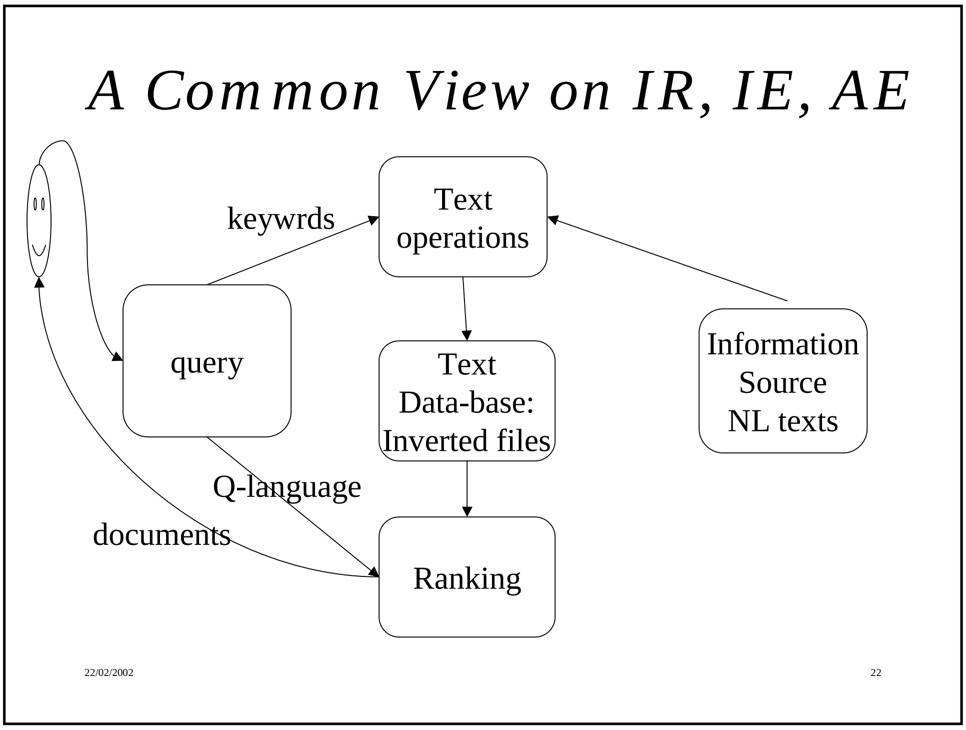
Text Data-base

Information Source NL texts

Information selection

22/02/2002

21



A Common View on IR, IE, AE

keywrds

Text operations

query

Text Data-base: Inverted files

Information Source NL texts

Q-language

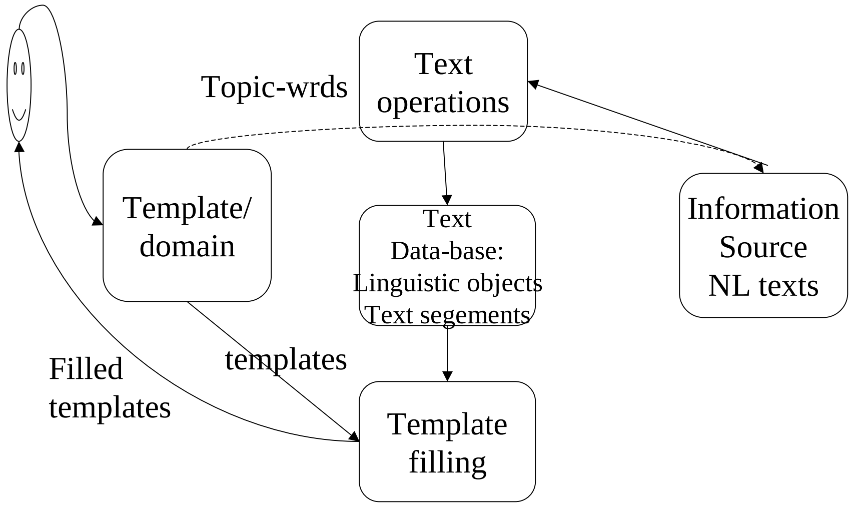
documents

Ranking

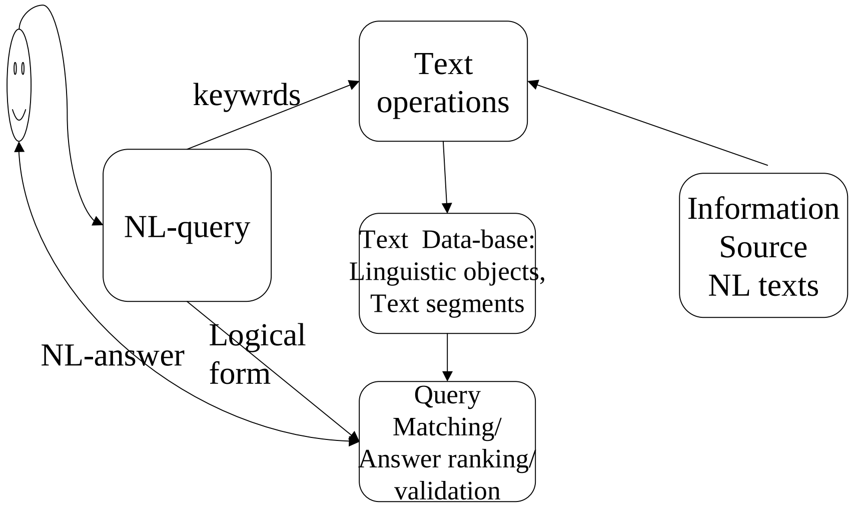
22/02/2002

22

A Common View on IR, IE, AE



A Common View on IR, IE, AE



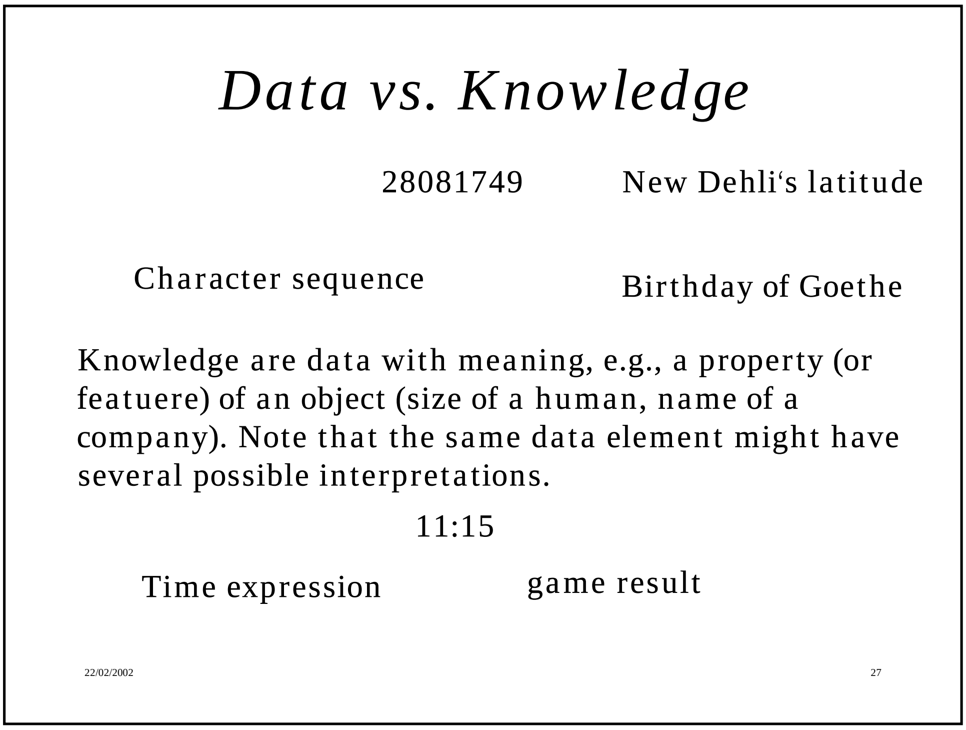
Data - Knowledge -Information

- Main task of an information system
  - ➢Maintain knowledge in digitilized form as data
  - ➢Provide knowledge as usefull information to a user

Data – knowledge - information

Information = Data + Knowledge.

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



Data vs. Knowledge

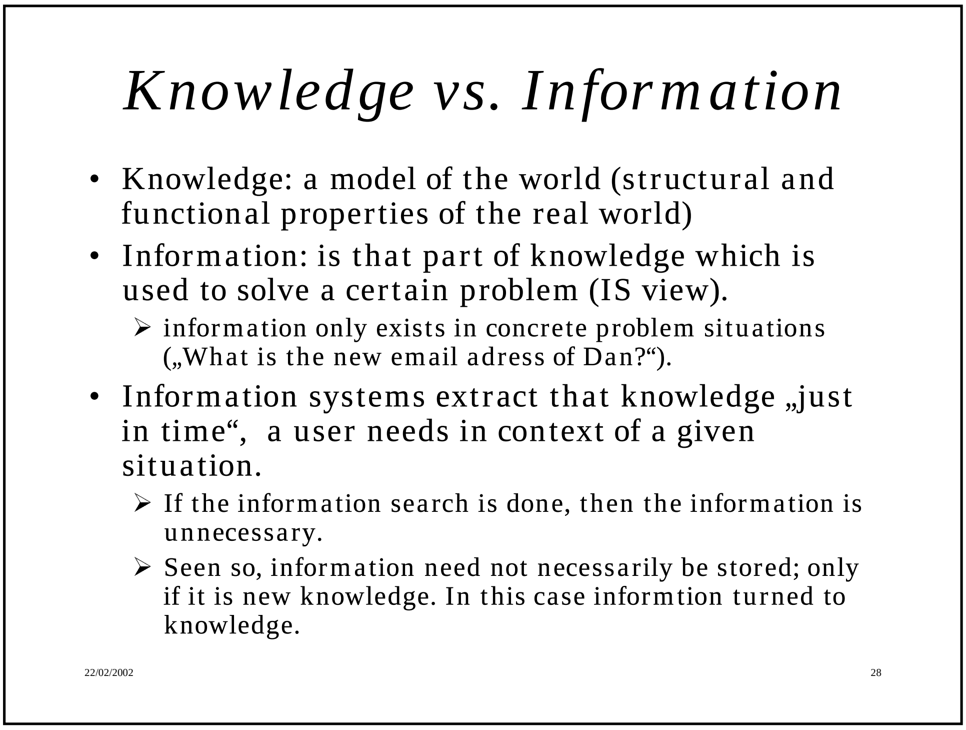
28081749 New Dehli‘s latitude

Character sequence Birthday of Goethe

Knowledge are data with meaning, e.g., a property (or featuere) 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



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 (IS view).
  - ➢ information only exists in concrete problem situations („What is the new email adress 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 informtion turned to knowledge.

Additional aspects of information

- Information theory (Shannon): the information content of a message depends on its propability
- Information is 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
- Information extraction 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:
      Häuser ⇨ *haus*
    - ▪ Compounds:
      Informationstechnologiezentrum ⇨
      {*Information, Technologie, Zentrum*}

NL analysis as step-wise normalization

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

    18.12.98 und *Freitag, der achtzehnte Dezember 1998*

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

  - ➢ proper names: persons, institutes, companies, locations

  - ➢ number expressions, adresses, formulars, aso.

NL analysis as step-wise normalization

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

    Für die deutsche Wirtschaft

    <head=für, comp=<head=wirtschaft,
    quant=def, mod=deutsch>>

- complex flat sentence structure

- domain specific templates (integration of ontology)

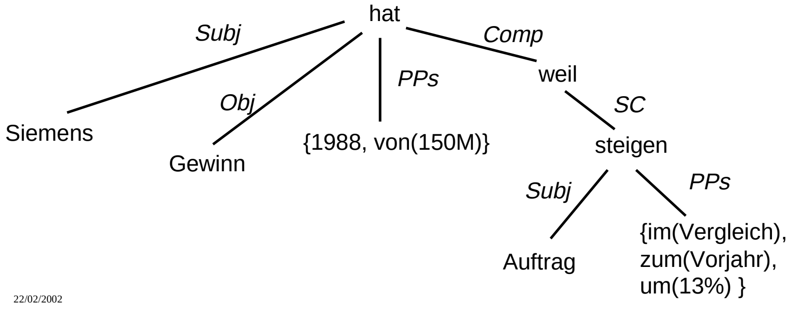
type = turnover		c-name = Possehl1
year = 1995/1		amount = 1.3e+9DM
tendency= +		diff = +23%

Underspecified functional description for sentences

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

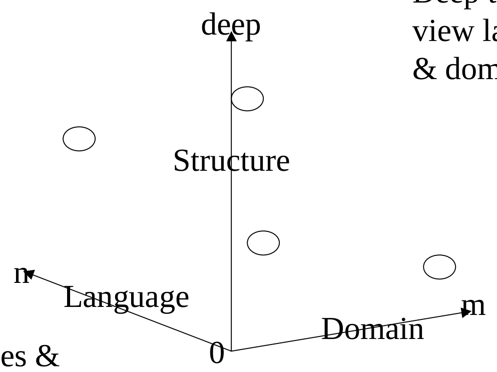
[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM], [$_{Comp}$weil] [$_{NP}$die Aufträge] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].

"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
view language
& domain



Many languages &
domains
then very shallow

Two Approaches to Building Extraction Systems

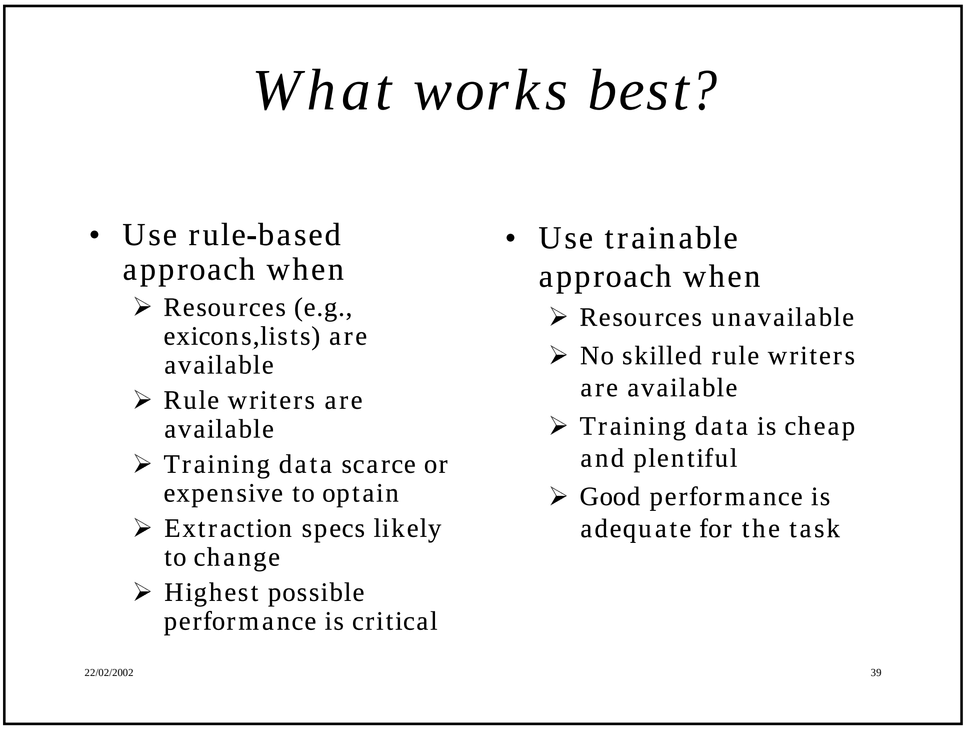
- Knowledge engineering approach
  - ➢ Grammars are constructed by hand
  - ➢ Domains patterns are discovered by a human expert through introspection and inspection of a corpus
  - ➢ Much laborious tuning and „hill climbing“
- Automatically Trainable Systems
  - ➢ Use statistical methods when possible
  - ➢ 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
- Disadvantages
  - ➢ Very laborious development process
  - ➢ Domain adaptation might require re-configuration
  - ➢ Needs experts which have both, linguistic & domain expertice

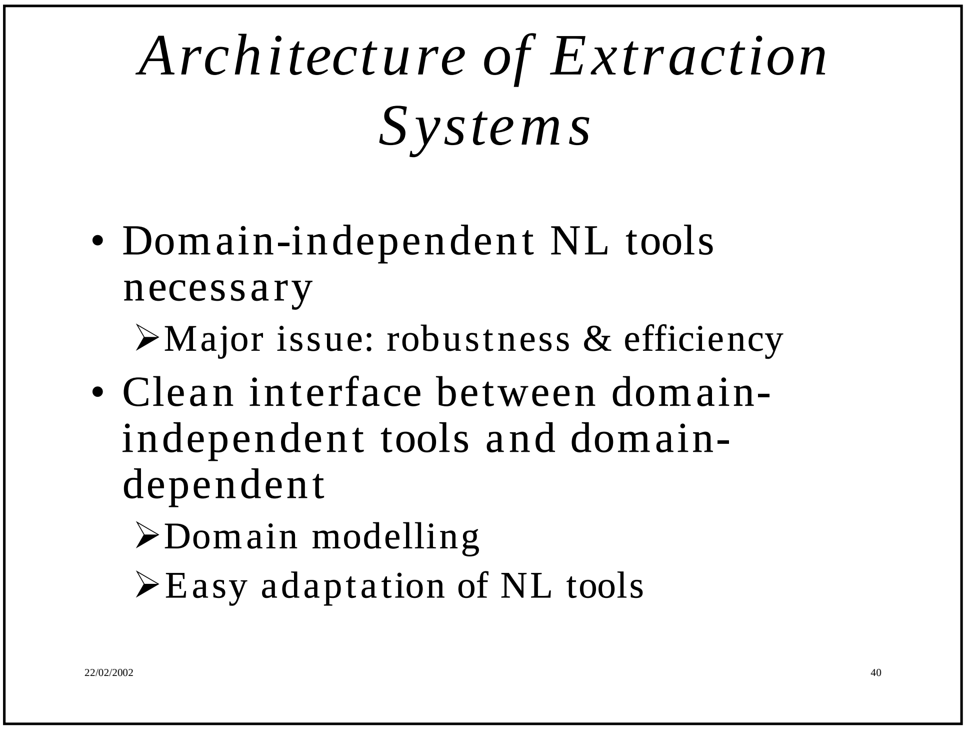
Trainable Systems

- 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 reannotation of large quantities of training data



What works best?

- Use rule-based approach when
  - Resources (e.g., exicons,lists) are available
  - Rule writers are available
  - Training data scarce or expensive to optain
  - 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 modelling
  - Easy adaptation of NL tools