Information Extraction and Question-Answering Systems

Foundations and methods

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

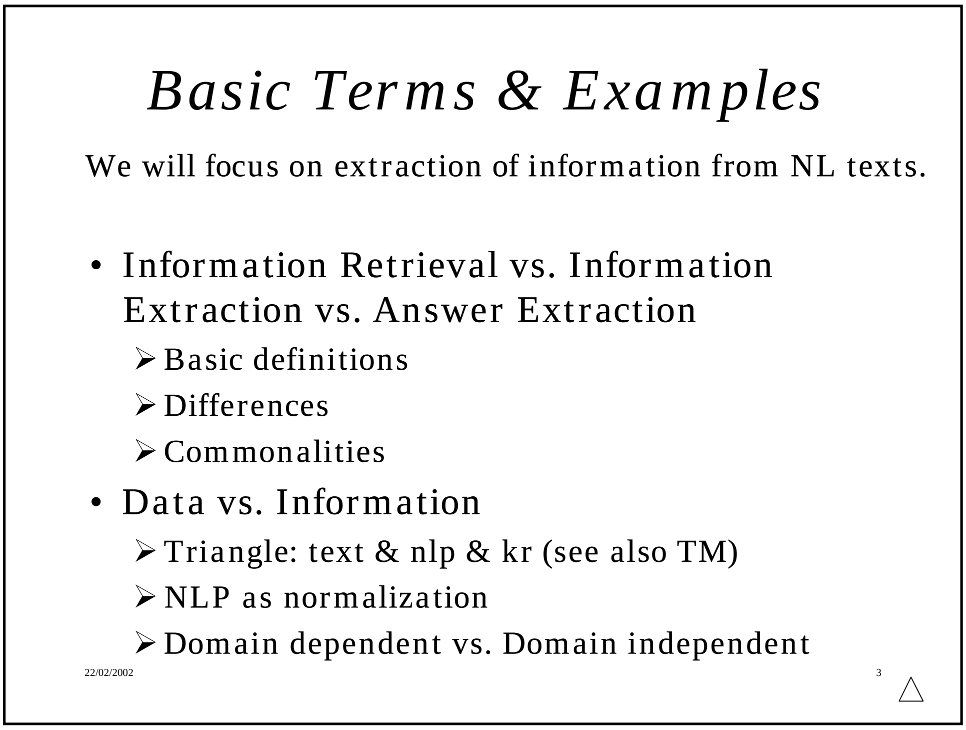
Parsing of Unrestricted Text

Generic NL Core system

Domain Modelling

Question/Answering Core components

Advanced Topics

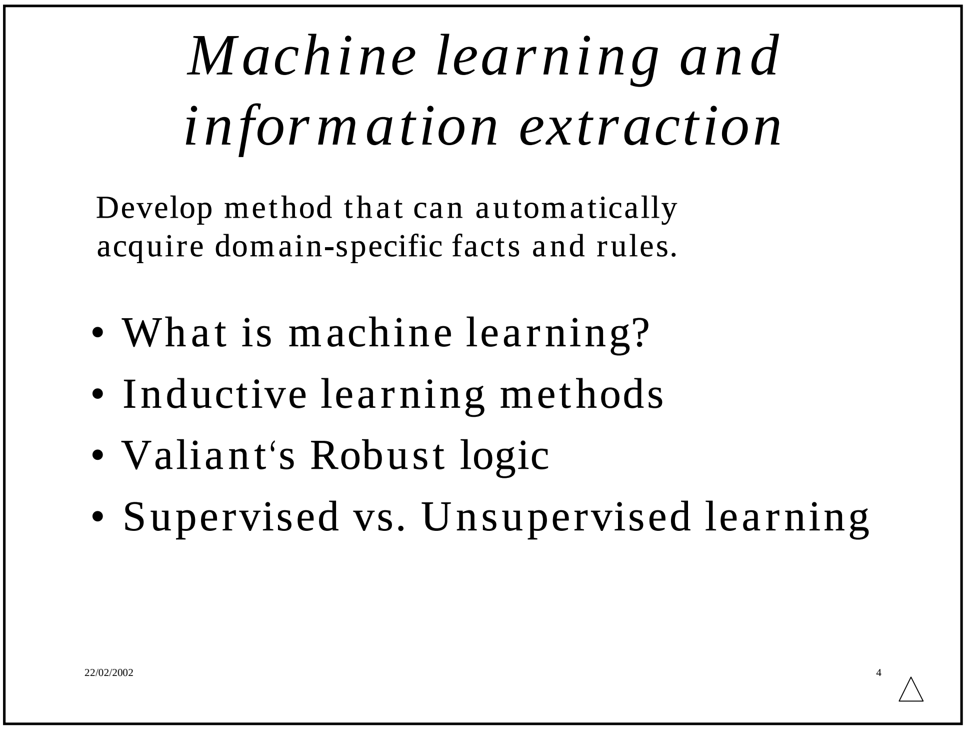


Basic Terms & Examples

We will focus on extraction of information from NL texts.

- Information Retrieval vs. Information Extraction vs. Answer Extraction
  - ➢ Basic definitions
  - ➢ Differences
  - ➢ Commonalities
- Data vs. Information
  - ➢ Triangle: text & nlp & kr (see also TM)
  - ➢ NLP as normalization
  - ➢ Domain dependent vs. Domain independent

22/02/2002 3

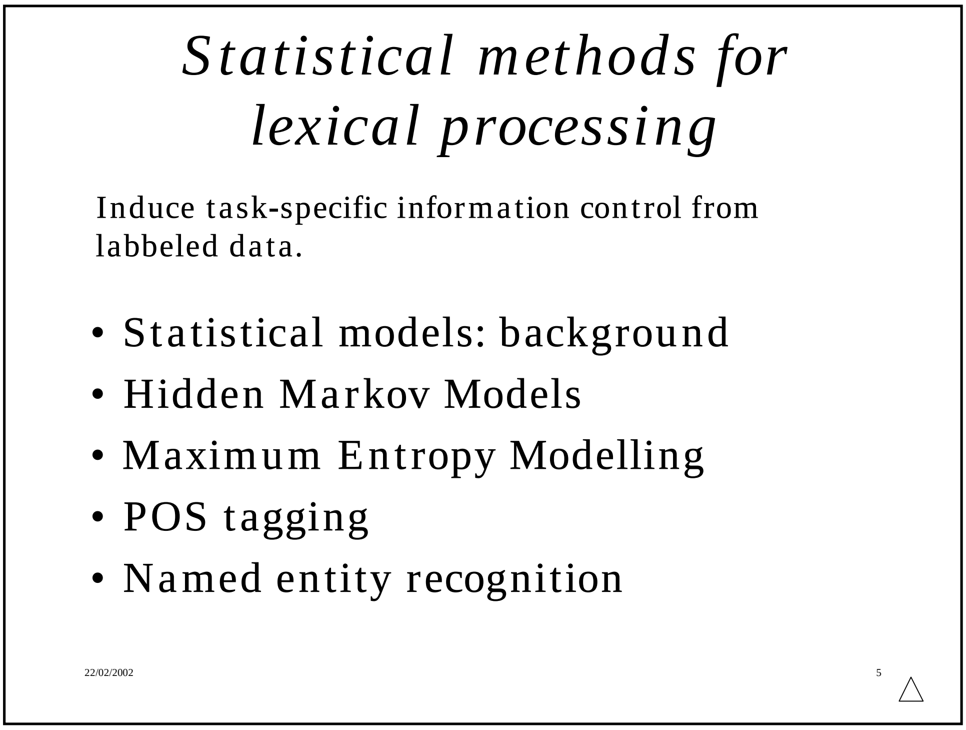


Machine learning and information extraction

Develop method that can automatically acquire domain-specific facts and rules.

- What is machine learning?
- Inductive learning methods
- Valiant‘s Robust logic
- Supervised vs. Unsupervised learning

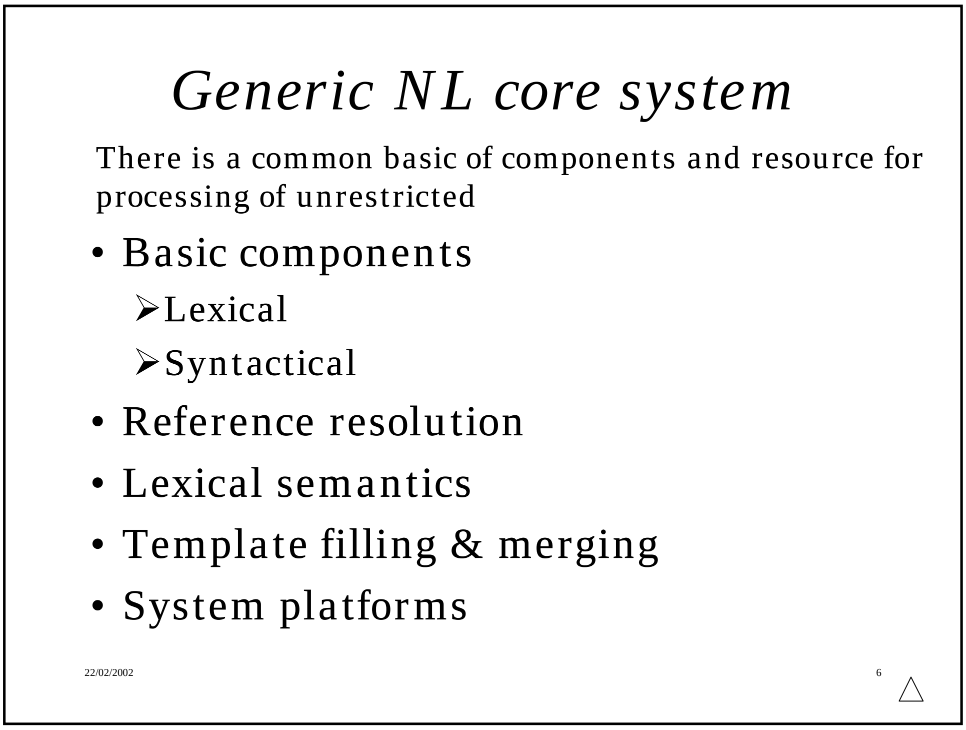
22/02/2002 4



Statistical methods for lexical processing

Induce task-specific information control from labbeled data.

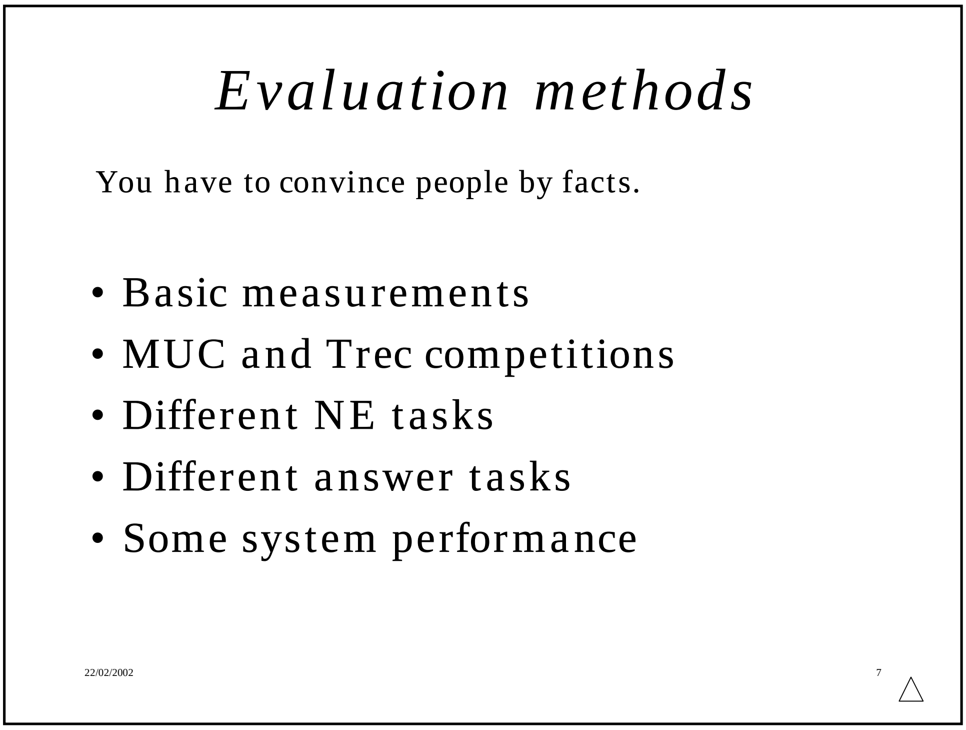
- Statistical models: background
- Hidden Markov Models
- Maximum Entropy Modelling
- POS tagging
- Named entity recognition



Generic NL core system

There is a common basic of components and resource for processing of unrestricted

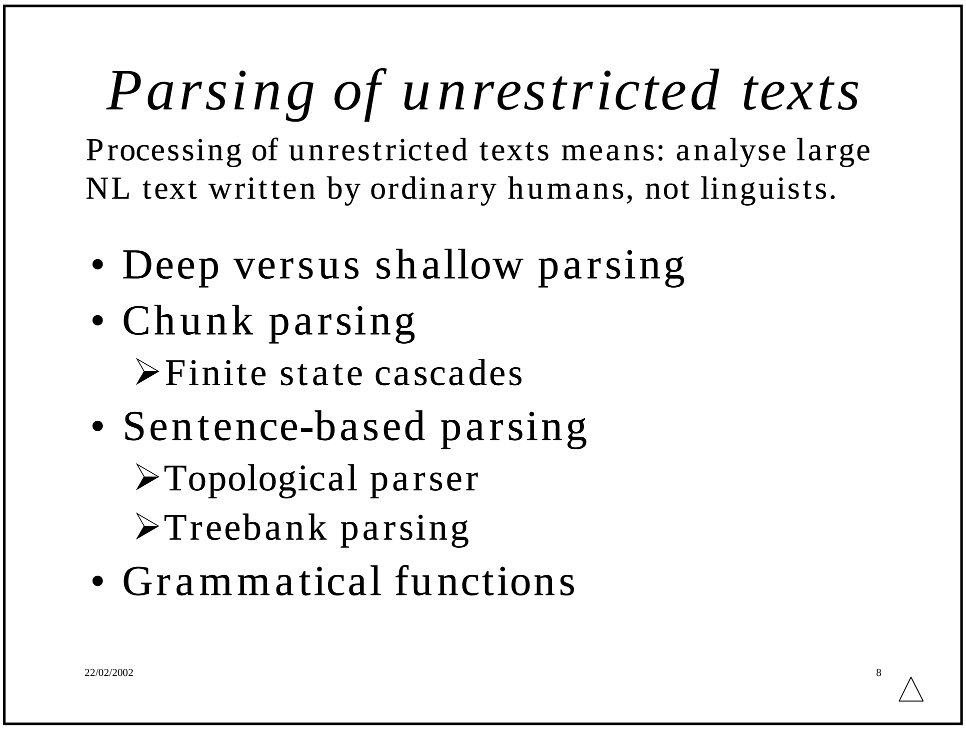
- Basic components
  - ➢Lexical
  - ➢Syntactical
- Reference resolution
- Lexical semantics
- Template filling & merging
- System platforms



Evaluation methods

You have to convince people by facts.

- Basic measurements
- MUC and Trec competitions
- Different NE tasks
- Different answer tasks
- Some system performance



Parsing of unrestricted texts

Processing of unrestricted texts means: analyse large NL text written by ordinary humans, not linguists.

- Deep versus shallow parsing
- Chunk parsing
  - ➢Finite state cascades
- Sentence-based parsing
  - ➢Topological parser
  - ➢Treebank parsing
- Grammatical functions

Domain modelling

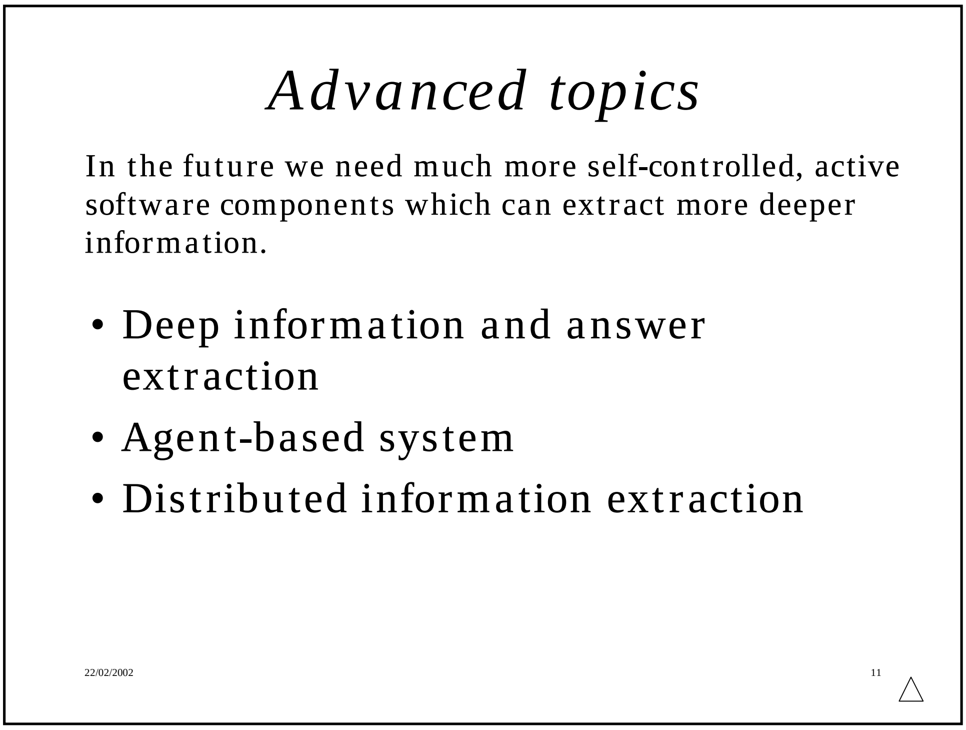
How can I systematically inform the system about the information I'm interested in?

- Template definition
- Ontologies
- Interfacing NL & ontologies
- Typed driven template processing

Question / Answering core components

Given a NL query extract its possible answer(s) from real-world NL text documents.

- Generic architecture
- Query processing
- Paragraph indexing
- Answer resolution
- Open-domain vs. domain-specific systems



Advanced topics

In the future we need much more self-controlled, active software components which can extract more deeper information.

- Deep information and answer extraction
- Agent-based system
- Distributed information extraction