Overview of i6:

HUMAN LANGUAGE TECHNOLOGY AND PATTERN RECOGNITION

Prof. Dr.-Ing. Hermann Ney

Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany

Contents

1 Computer Science Informatik VI (i6)

Contents

1.1 Research Topics i6

Method: Stochastic Modelling

- **Modelling Dependencies and Vague Knowledge (contrast: rule-based approach)**
- **Decision Making, in particular in Context**
- **Automatic Learning from Data/Examples**

Applications:

Human Language Technology and Pattern Recognition

Many of these applications are concerned with

Man-Machine Interface

Example: Speech Recognition

Example: Speech Recognition for Text Dictation
global task: convert the acoustic signal into a written text
decisions: which words were spoken?

different levels:

- **pure acoustic recognition:**
  convert the acoustic signal
  into a sequence of sound symbols

- **context information:**
  - **pronunciation lexicon:**
    In each language, only a small subset of all possible sound sequences is really used (or allowed).
  - **syntax + semantics:**
    The produced text should observe the syntactic and semantic constraints as they are required by the language and the application.

1.2 Tasks and Applications: Examples

- Speech recognition
  - small vocabulary
  - large vocabulary
- Machine translation
- Natural language processing
  - text/document classification
  - information retrieval
  - parsing and syntactic analysis
- Language understanding and dialog systems
- Image recognition
  - object recognition
  - handwriting recognition
- Diagnosis and expert systems

- Other applications:
  - speaker verification and identification
  - fingerprint verification and identification
  - DNA sequence identification
  - gesture recognition
  - lip reading
  - geological analysis
  - high-energy physics: bubble chamber tracks
  - ...

1.3 Projects i6

- **ARISE (EU):**
  Automatic Railway Information Systems across Europe
  – Speech Recognition and Language Modelling
- **EuTrans II (EU):**
  Translation of Spoken Language
  – Speech Recognition and Translation
- **Institut für deutsche Sprache (IdS):**
  – Language Modelling for Newspapers
- **Audio Document Retrieval (NRW):**
  – Speech Recognition and Information Retrieval
- **Verbmobil II (BMBF):**
  Speech Recognition and Translation for
  Appointment Scheduling and Traveling Information
  – Speech Recognition
  – Speech Translation
  – Prototype Modules

Projects i6

- **Image Object Recognition (RWTH):**
  - **OCR (optical character recognition)**
  - **Medical Images**
- **Advisor (EU):**
  - **Speech Recognition for German Broadcast News**
- **EGYPT follow-up (NSF):**
  - **Basic Algorithms for Statistical Machine Translation**
- **Audio Document Retrieval (NRW ?):**
  - **German Broadcast News: Recognition and Information Retrieval**
- **Bilateral Projects with Companies (including start-ups)**

Projects i6

- German DFG:
  - Improved Acoustic Modelling using Structured Models
- Coretex (EU):
  - Improving Core Technology for Speech Recognition
  - Applications: Broadcast News in Several Languages
- LC-Star (EU):
  - lexical and corpora resources for recognition, translation and synthesis
  - prototype system for machine translation of spoken sentences
- Transtype-2 (EU):
  - machine translation of written text
  - application: interactive machine-aided translation
- PF-Star (EU):
  - machine translation of spoken dialogues
  - application: tourism and travelling
- TC-Star (EU):
  - recognition and translation of EU-parliamentary speeches and debates

1.4 Courses i6

- Lectures (L4) with exercises (E2):
  - SR: Speech Recognition
  - PR1: Pattern Recognition + Neural Nets
  - SI: Signal + Image Processing
  - NLP: Natural Language Processing

- Lectures (L2) with exercises (E1):
  - PR0: 'Man-Machine Interface'
  - PR2: Advanced Topics in Pattern Recognition
  - LM: Language Modelling
  - MIP: Medical Image Processing
    ('Ringvorlesung', each WS)

- Seminars: – Diplom Degree (SS, Block)
  – Doctor Degree (WS+SS)

- Laboratory Courses (WS, Block)

- Study Groups (WS+SS: speech, language, image)

course cycles of (about) two years:

year		L4	L2
01/02	WS	SR	LM
	SS	PR1	–
02/03	WS	SR	–
	SS	SI	PR2, NLP
03/04	WS	PR1	–
	SS	SR	–
04/05	WS	NLP	–
	SS	PR1	SI?
05/06	WS	SR	–
	SS	NLP	

1.5 Examinations i6

Diplom Degree:

- **area of specialization (Vertiefungsgebiet) i6 with the topics:**
  - **– Speech + Pattern Recognition**
  - **– Speech + Language Processing**
  - **– Signal + Image Processing**
  - **– ...**

  select 12 hours (SWS) out of i6 lectures

- **practical computer science (Prakt. Informatik) (3 areas):**
  recommendation: 12 hours (SWS) out of

  V4: Signal + Image Processing (SI, ...)
  V4: Language (LM, MT, NLP, ...)
  or Speech (SR, LM, ...)
  V4: i6-external lectures:
  - **– data bases**
  - **– artificial intelligence**
  - **– ...**

  additional alternatives: on demand

Examinations i6

- **Technischer Redakteur:**
  more or less similar to Diplom degree
- **Master in Computer Science (Software Systems Engineering):**
  credit system: examination after each course

2 Textbooks

Textbooks on Natural Language Processing
(using statistical/corpus-based methods)

[Manning & Schütze 99] C. D. Manning, H. Schütze:
Foundations of Statistical Natural Language Processing.
MIT Press, Cambridge, MA, 1999.
(best book on statistical methods for written language)

[Charniak 93] E. Charniak:
Statistical Language Learning.
MIT Press, Cambridge, MA, 1993.
(principles of statistical methods for written language)

[Jelinek 97] F. Jelinek:
Statistical Methods for Speech Recognition.
MIT Press, Cambridge, 1997.
(emphasis on language modelling and large vocabulary speech recognition)

Conventional Linguistics and Artificial Intelligence

[Jurafsky & Martin 00] D. Jurafsky, J. H. Martin:
Speech and Language Processing.
Prentice Hall, Englewood Cliffs, NJ, 2000.
(mixture of conventional and statistical approaches)

[Russel & Norvig 03] S. Russel, P. Norvig:
Artificial Intelligence.
2nd ed.,Prentice Hall, Upper Saddle River, NJ, 2003.
(comprehensive textbook on artificial intelligence:
– chapters 13-21: probabilistic approaches and learning,
– chapters 22-25: communication, language, perception, robotics)

[Nilsson 98] N. J. Nilsson:
Artificial Intelligence: A New Synthesis.
Morgan Kaufmann Pub., San Francisco, CA, 1998.
(standard textbook on artificial intelligence)

Textbooks on Statistical Classification and Learning (Pattern Recognition, Neural Networks, Data Mining, ...)

[Duda & Hart$^+$ 01] R. O. Duda, P. E. Hart, D. G. Stork:
Pattern Classification.
2nd ed., J. Wiley & Sons, New York, NY, 2001.
(best introduction including modern concepts)

[Ripley 96] B. D. Ripley:
Pattern Recognition and Neural Networks.
Cambridge University Press, Cambridge, England, 1996.
(emphasis on statistical concepts)

[Hastie & Tibshirani$^+$ 01] T. Hastie, R. Tibshirani, J. Friedman:
The Elements of Statistical Learning: Data Mining, Inference and Predictions.
Springer, New York, 2001.
(emphasis on modern statistical concepts)

[Devroye & Györfi$^+$ 01] L. Devroye, J. Györfi, G. Lugosi:
A Probabilistic Theory of Pattern Recognition.
Springer, New York, 1996.
(emphasis on theory and principles)

[Bishop 06] C. M. Bishop: Pattern Recognition and Machine Learning.
Springer, New York, 2006.
(pattern recognition from a machine learning point-of-view)

Textbooks on Mathematical Toolbox
(vector spaces and matrices, statistics, optimization methods, ...)

[Moon & Stirling 00] T. K. Moon, W. C. Stirling:
Mathematical Methods and Algorithms for Signal Processing.
Prentice Hall, Upper Saddle River, NJ, 2000.
(best overall summary)

[Press & Teukolsky$^+$ 92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery:
Numerical Recipes in C.
Cambridge Univ. Press, Cambridge, 2nd ed., 1992.
(good overview of numerical algorithms and implementations)

[Boyd 04 & Vandenberghe] S. Boyd, L. Vandenberghe: Convex Optimization.
Cambridge Univ. Press, Cambridge, 2004.

Textbooks on Statistics and Bayesian Methods

[Casella & Berger 90] G. Casella, R. L. Berger:
Statistical Inference.
Wadsworth & Brooks/Cole, Pacific Grove, CA, 1990.
(good introduction into modern general statistics)

[Berger 85] J. O. Berger:
Statistical Decision Theory and Bayesian Analysis.
Springer, New York, 2nd ed., 1985.
(topic: Bayes Methods)

[Carlin & Louis 96] B. P. Carlin, T. A. Louis:
Bayes and Empirical Bayes Methods for Data Analysis.
Chapman & Hall, London, 1996.
(topic: Bayes and Empirical Bayes Methods)

Textbooks: Speech Recognition

Textbooks on Speech Recognition:

- **emphasis on signal processing and small-vocabulary recognition:**
  L. Rabiner, B. H. Juang: Fundamentals of Speech Recognition.
  Prentice Hall, Englewood Cliffs, NJ, 1993.
- **emphasis on large vocabulary and language modelling:**
  F. Jelinek: Statistical Methods for Speech Recognition.
  MIT Press, Cambridge, 1997.
- **introduction to both speech and language:**
  D. Jurafsky, J. H. Martin: Speech and Language Processing.
  Prentice Hall, Englewood Cliffs, NJ, 2000.
- **advanced topics:**
  R. De Mori: Spoken Dialogues with Computers.
  Academic Press, London, 1998

3 Overview of Methodology

Contents

3.1 Review: HLT Tasks

Human Language Technology (HLT)

typical tasks in human language technology:

- **speech recognition:**
  convert acoustic signal to ASCII text
- **character recognition:**
  convert machine printed or handwritten characters to ASCII text
- **speech and text translation:**
  translate a written/spoken language into a target language
- **(spoken and written) language understanding:**
  generate a semantic representation (formal database query) of a written/spoken sentence
- **spoken dialog systems:**
  understanding task with a full dialog rather than a single sentence
- **speech synthesis:**
  convert written text into sequence of subword units with prosodic information
- **text summarization:**
  generate a summary of a text document or article
- **document classification:**
  classify a text document into one out of several document classes
- **...**

3.2 The Statistical Approach to HLT

principles:

- **HLT tasks are complex tasks,**
  for which perfect solutions are difficult
- **consequence: use imperfect and vague knowledge**
  and try to minimize the number of decision errors
- **formal specification:**

input:	**observation**	x
output:	**decision**	c

- **Bayes decision rule assuming probabilistic dependencies between x and c:**

$$\begin{aligned} x \rightarrow \hat{c} &= \arg\max_{c} \left\{ pr(c|x) \right\} \\ &= \arg\max_{c} \left\{ pr(c) \cdot pr(x|c) \right\} \end{aligned}$$

exercise:
give mathematically exact definitions of the various tasks in HLT

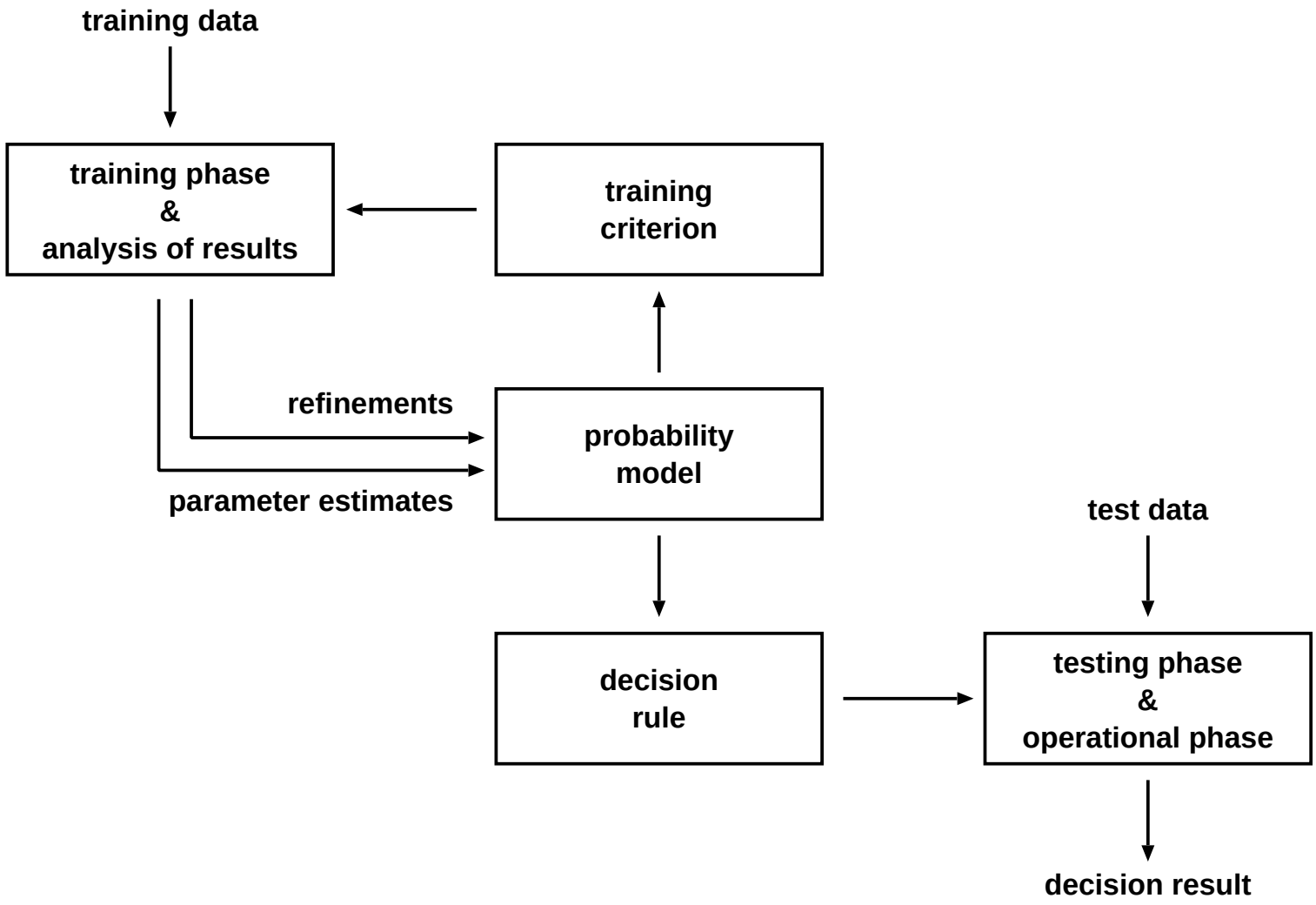
The Statistical Approach: Key Components

- **decision rule:**
  requires maximization (sometimes hard!)
  and probability distribution $pr(c|x)$, which is unknown

- **probability model $p_\theta(c|x)$ or $p_\theta(c) \cdot p_\theta(x|c)$**
  is used to replace $pr(c|x)$ or $pr(c) \cdot pr(x|c)$

- **training criterion**
  to learn the unknown parameters θ from training data

Illustration of the Statistical Approach to HLT



Examples of Probability Models

- discriminant functions (linear and nonlinear)
- neural networks: (virtually) any structure
- Gaussian classifier
- Gaussian mixtures
- models with hidden variables (path, alignment):
  – Hidden Markov models (HMM) in speech recognition
  – alignment models in machine translation
- maximum entropy models (log-linear, exponential, multiplicative)
- decision trees (CART)
- ...

Examples of Algorithms for Decision Rule ('Search')

- **forward algorithm:**
  for HMM in speech recognition

- **dynamic programming**
  – for POS tagging and other tagging tasks:
  – for small vocabulary-speech recognition,
  – for translation using finite-state transducers

- **time-synchronous beam search and A* search:**
  for large-vocabulary speech recognition

- **position-synchronous beam search and A* search:**
  for large-vocabulary language translation

- **...**

Examples of Training Criteria

training data: labelled sequence $(x_n, c_n), n = 1, ..., N$
general criteria:

- **maximum likelihood:**

$$\arg\max_{\theta} \left\{ \sum_{n=1}^{N} \log\ p_\theta(c_n) \right\}$$

$$\arg\max_{\theta} \left\{ \sum_{n=1}^{N} \log\ p_\theta(x_n|c_n) \right\}$$

- **posterior probability:**

$$\arg\max_{\theta} \left\{ \sum_{n=1}^{N} \log\ p_\theta(c_n|x_n) \right\}$$

- **squared error criterion:**

$$\arg\min_{\theta} \left\{ \sum_{n=1}^{N} \sum_{c} \Big[p_\theta(c|x_n) - \delta(c, c_n) \Big]^2 \right\}$$

- **smoothed error count**
- **...**

Training Procedures and Algorithms for Specific Models

- **EM (expectation/maximization) algorithm:**
  maximum likelihood for hidden-variable models
  (maximum approximation: Viterbi training)

- **error back propagation:**
  squared error criterion for neural networks

- **GIS (general iterative scaling):**
  posterior probability for maximum entropy (log-linear) models

- **...**

Bayes Decision Rule: Sources of Errors

Why does a statistical decision system make errors?

To be more exact:
Why errors IN ADDITION to the minimum Bayes errors?

Reasons from the viewpoint of Bayes' decision rule:

- **incorrect input (or observation):**
  only an incomplete part or a poor transformation of the true observations is used.

- **incorrect modelling:**
  - **incorrect probability distribution**
  - **not enough training data**
  - **poor training criterion**

- **incorrect search procedure:**
  the maximum is not found

Example of Model Structures

example:
text classification: class c with word counts $x_1, ...x_d, ...x_D$

$$p(c|x_1^D) = ?$$

non-parametric method (modelfree method):
use a big table $p(c|x_1^D)$ estimated by relative frequencies

problem:
lack of generalization capabilities, depends on training data

Example: Model with Multiplicative Structure

$$
\begin{aligned}
p(c|x_1^D) &= \frac{\prod_d \alpha_d(x_d, c)}{\sum_{c'} \prod_d \alpha_d(x_d, c')} \\
&\quad \text{with} \quad \alpha_d(x_d, c) \geq 0 \quad \text{and} \quad \sum_c \alpha_d(c|x_d) = 1 \\
&= \frac{\prod_d [A_d(x_d) \cdot \alpha_d(c|x_d)]}{\sum_{c'} \prod_d [A_d(x_d) \cdot \alpha_d(c'|x_d)]} \\
&= \frac{\prod_d A_d(x_d) \cdot \prod_d \alpha_d(c|x_d)}{\prod_d A_d(x_d) \cdot \sum_{c'} \prod_d \alpha_d(c'|x_d)} \\
&= \frac{\prod_d \alpha_d(c|x_d)}{\sum_{c'} \prod_d \alpha_d(c'|x_d)}
\end{aligned}
$$

model structure: log-linear model (maximum entropy)

Example: Model with Additive Structure

$$p(c|x_1^D) = \frac{\sum_d \beta_d(x_d, c)}{\sum_{c'} \sum_d \beta_d(x_d, c')}$$

with $\beta_d(x_d, c) \geq 0$ and $\sum_c \beta_d(c|x_d) = 1$

$$= \frac{\sum_d [B_d(x_d) \cdot \beta_d(c|x_d)]}{\sum_{c'} \sum_d [B_d(x_d) \cdot \beta_d(c'|x_d)]}$$

$$= \frac{\sum_d [B_d(x_d) \cdot \beta_d(c|x_d)]}{\sum_d [B_d(x_d) \cdot \sum_{c'} \beta_d(c'|x_d)]}$$

$$= \frac{\sum_d [B_d(x_d) \cdot \beta_d(c|x_d)]}{\sum_d [B_d(x_d)]}$$

$$= \sum_d [b(d|x_1^D) \cdot \beta_d(c|x_d)] \quad \text{with} \quad b(d|x_1^D) := \frac{B_d(x_d)}{\sum_{d'} B_{d'}(x_{d'})}$$

model structure: (sort of) mixture distribution

THE END