Department of Computer Science and Engineering,
Department of Electronic and Computer Engineering,
HKUST, Hong Kong, Dec. 04, 2012

The Statistical Approach to Speech Recognition and Natural Language Processing: Achievements and Open Problems

Hermann Ney
Human Language Technology and Pattern Recognition

RWTH Aachen University, Aachen
DIGITEO Chair, LIMSI-CNRS, Paris

Outline

ISCA: International Speech Communication Association

- **ISCA started as ESCA (European Speech Communication Association):**
  March 27, 1988 by Rene Carree.
- **purpose:**
  to promote Speech Communication Science and Technology,
  both in the industrial and academic areas,
  covering all the aspects of Speech Communication
  (acoustics, phonetics, phonology, linguistics, natural language processing,
  artificial intelligence, cognitive science, signal processing, pattern recognition, etc.
- **ISCA offers a wide range of services;**
  in particular Interspeech, ISCA workshops, SIGs (special interest groups)

ISCA Objectives:

- **to stimulate scientific research and education,**
- **to organize conferences, courses and workshops,**
- **to publish, and to promote publication of scientific works,**
- **to promote the exchange of scientific views in the field of speech communication,**
- **to encourage the study of different languages,**
- **to collaborate with all related associations,**
- **to investigate industrial applications of research results,**
- **and, more generally, to promote relations between public and private, and between science and technology.**

1 History and Projects

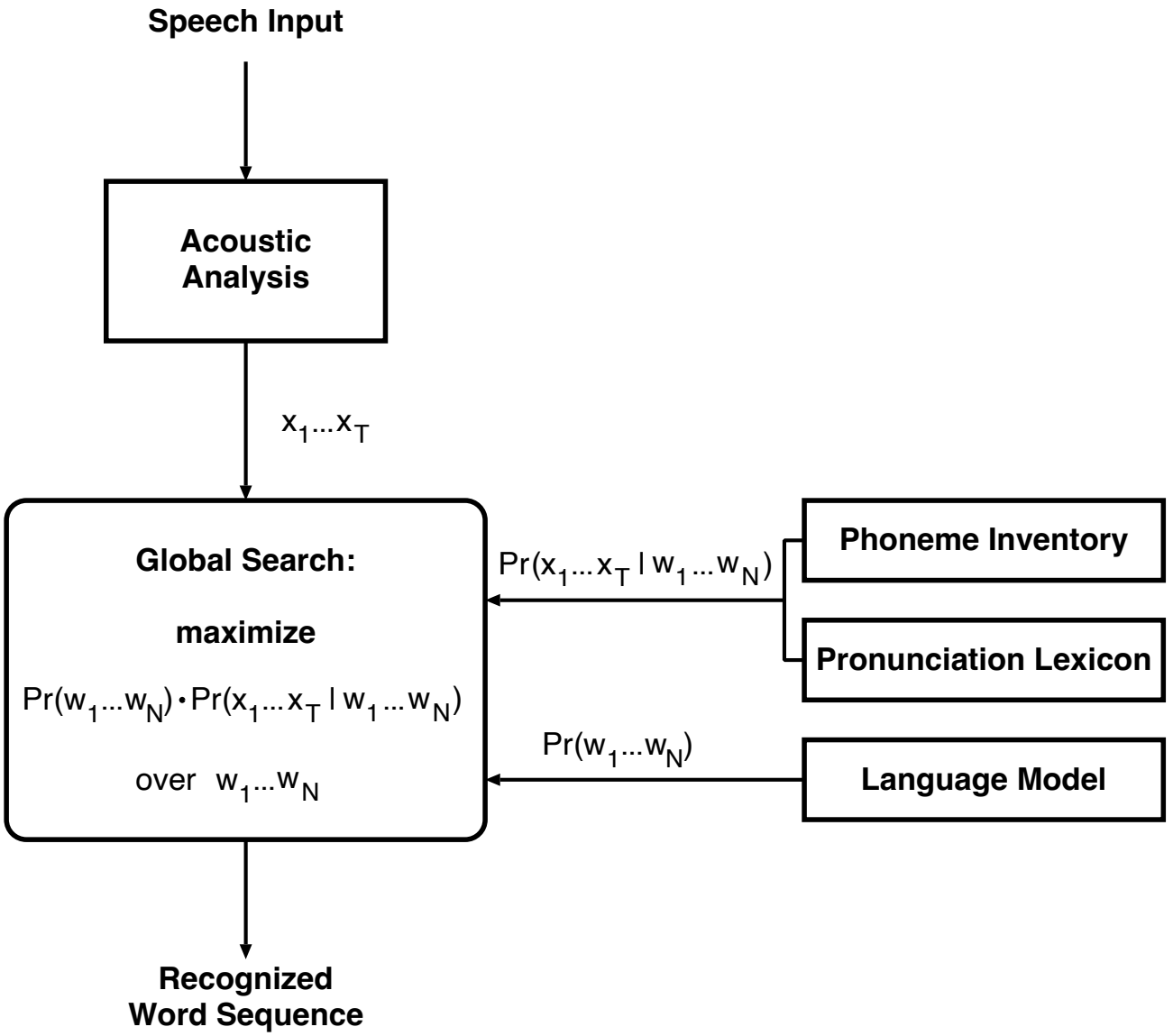
terminology: tasks in speech and natural language processing (NLP)

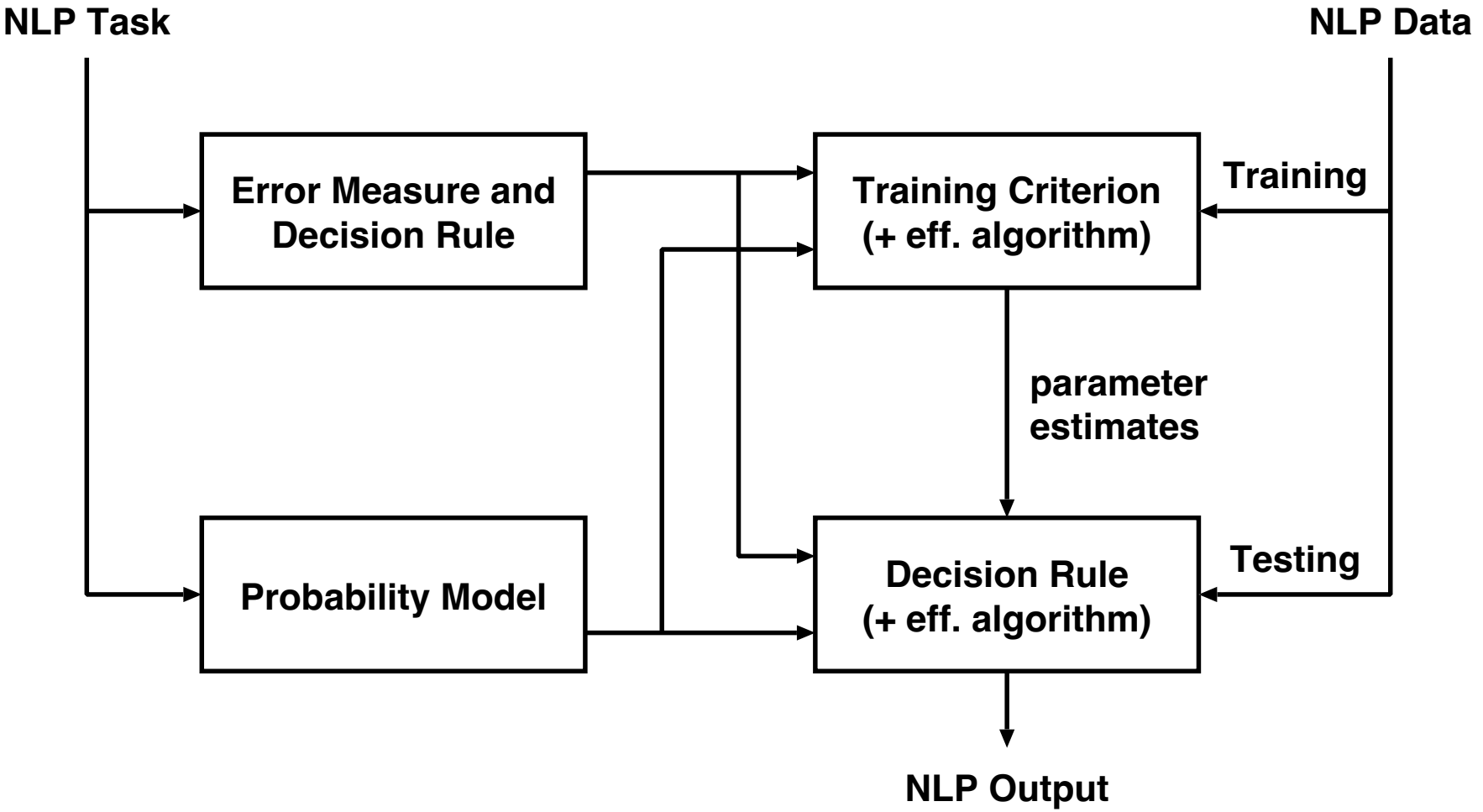
- **automatic speech recognition (ASR)**
- **optical character recognition (OCR: printed and handwritten text)**
- **machine translation (MT)**
- **document classification**
- **understanding of speech or language**

characteristic properties of these tasks (ASR, OCR, MT):

- **well-defined 'classification' tasks:**
  – due to 5000-year history of (written!) language
  – well-defined classes: letters or words of the language
- **easy task for humans**
  (ASR, OCR: at least in their native language!)
- **hard task for computers**
  (as the last 40 years have shown!)

Statistical Approach to Automatic Speech Recognition (ASR)




NLP Task
NLP Data
Error Measure and Decision Rule
Training Criterion (+ eff. algorithm)
Training
parameter estimates
Probability Model
Decision Rule (+ eff. algorithm)
Testing
NLP Output

four ingredients of the statistical approach to ASR:

- **decision procedure (Bayes decision rule):**
  - **minimizes the decision errors**
  - **consistent and holistic criterion**
  - **no explicit segmentation**
- **models of probabilistic dependencies:**
  - **problem-specific (in lieu of 'big tables')**
  - **textbook statistics and much beyond ...**
- **model parameters are learned from examples:**
  - **statistical estimation and (any type of) learning**
  - **suitable training criteria**
- **search or decoding:**
  find the most 'plausible' hypothesis

statistical approach to ASR:

$$\textbf{ASR} = \textbf{Modelling} + \textbf{Statistics} + \textbf{Efficient Algorithms}$$

Short History of ASR

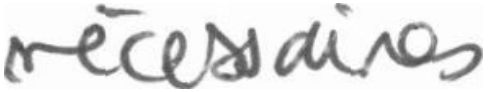
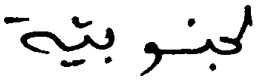
- start of statistical approach around 1972 at IBM research
- steady improvement of statistical methods over 40 years
- controversial issues: about usefulness of
  - 'existing' theories/models from phonetics and linguistics
  - rule-based approaches from classical artificial intelligence

40 years of progress by improving the statistical methods (along with training criteria):

- Hidden Markov models (HMM) along with EM algorithm
- smoothing/regularization
- CART and phonetic decision trees
- discriminative training:
  MMI, Poveys's MPE, MCE, ...
- adaptation (unsupervised and supervision light training)
- neural networks and log-linear modelling
- machine learning?

Automatic Recognition: From Speech to Characters

image text recognition:
– define vertical slots over horizontal axis
– result: image signal = (quasi) one-dim. structure like speech signal

Language	Database	Example
French	**RIMES**	
Arabic	**IfN/ENIT**	
English	**IAM**	

From Speech Recognition to Machine Translation

from subsymbolic to symbolic processing:

- **so far: recognition of signals: speech and image**
- **consider the problem of translation:**
  - **– convert the text from a source language to a target language**
  - **– problem of symbolic processing**

machine translation: why a statistical approach?
answer: we need decisions along various dimensions:

- **select the right target word**
- **select the position for the target word**
- **make sure the resulting target sentence is well formed**

interaction: Bayes decision rule handles the interdependencies of decisions

conclusion: MT (like other NLP tasks) amounts to making decisions

scientific framework for making good decisions:
probability theory, statistical classification, statistical learning

From Speech Recognition to Machine Translation

use of statistics has been controversial in NLP:

- **Chomsky 1969:**
  ... the notion 'probability of a sentence' is an entirely useless one,
  under any known interpretation of this term.
- **was considered to be true by most experts in NLP and AI**

IBM's Jelinek did not care about Chomsky's ban:

- **1988: IBM starts building a statistical system for MT**
  (in opposition to linguistics and artificial intelligence)
- **task: Canadian Hansards: English/French parliamentary debates (text!)**
- **1994 DARPA evaluation:**
  – comparable to 'conventional' approaches (Systran)
  – results only for French $\rightarrow$ English
- **team went off to Renaissance Technologies (Hedge Fund)**

After IBM: 1992 – 2000

translation of SPEECH (vs. text):

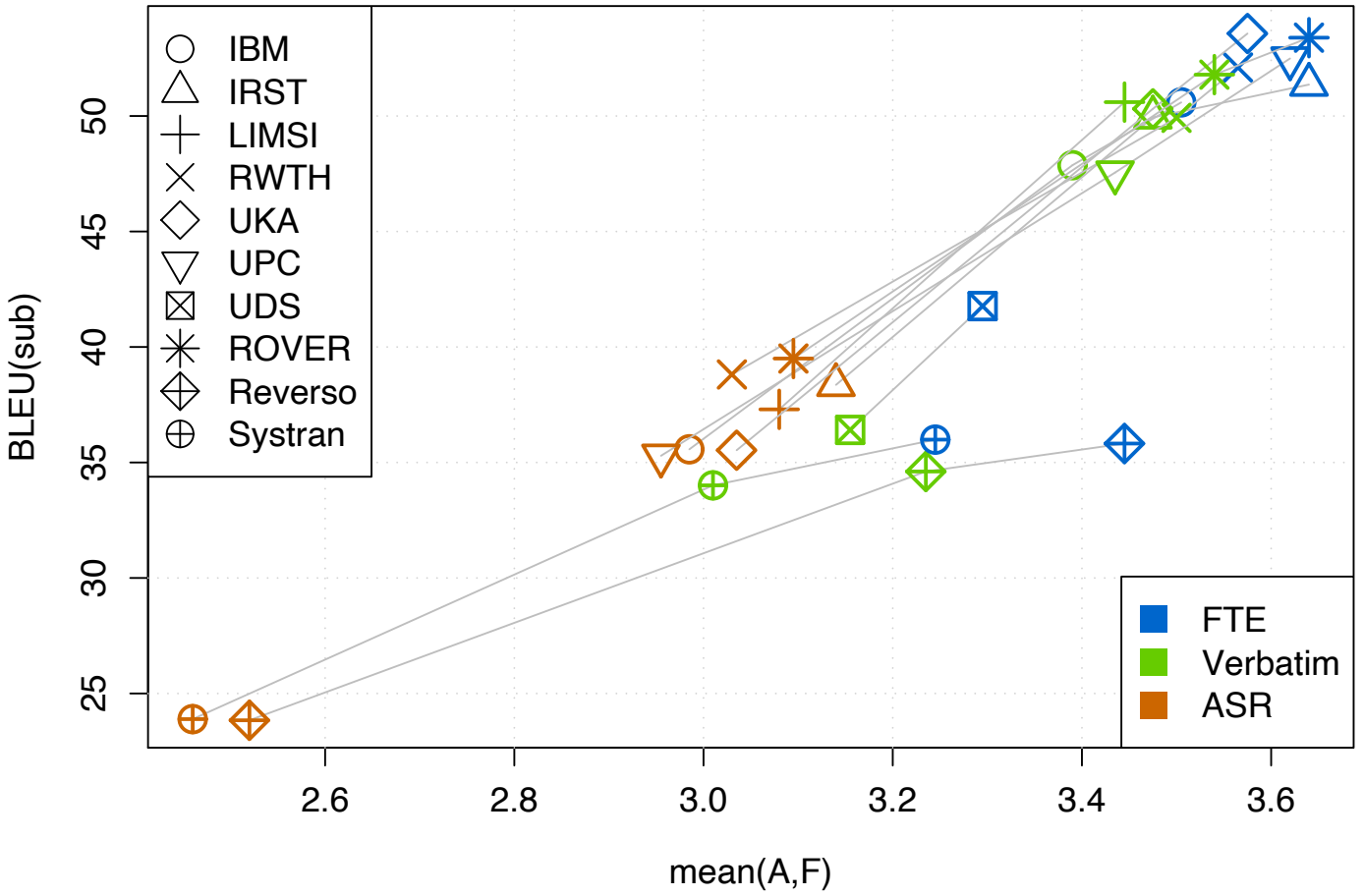
- **justification for statistical approach: robustness**
  - **cope with non-grammatical input and disfluencies**
  - **handle recognition errors**
- **projects on limited domain tasks (laboratorial data!):**
  - **CSTAR consortium**
  - **Verbmobil (German)**
  - **EU projects: Eutrans, PF-Star, LC-Star, ...**
- **EU Project TC-Star (2004-2007):**
  - **speeches given in the European Parliament**
    real-life task: unlimited domain and large vocabulary
  - **FIRST research prototype on speech translation of THIS type**

side result:

statistical approach looked promising for text, too!

E → S 2007: Human vs. Automatic Evaluation

BLEU: automatic accuracy measure
mean(A,F): human judgement of Adequacy and Fluency



More ASR and MT Projects 2001 – 2012

'unlimited' domain (real-life data) with associated evaluations:

- **TIDES 2001-04 funded by DARPA: written text (newswire):**
  MT: Arabic/Chinese to English
- **GALE 2005-2011 (and BOLT 2012-2017)**
  funded by DARPA (funding: 40 Mio US$ per year):
  – text and speech
  – Arabic/Chinese to English
  – ASR, MT and information extraction ('question answering')
- **QUAERO 2008-2013 funded by OSEO France:**
  – research track (in addition to application track)
  – multimodal data: text, handwritten text, speech, image, video, ...
  – many languages: EU languages and Arabic/Chinese
  – types of spoken language: news, lectures, discussions, ...
  – more colloquial language (for text and speech)

  – periodic evaluations: internal or external
- **more EU projects, specifically on text (after GOOGLE Translate!):**
  – EUROMATRIX and –PLUS: text MT for all EU languages
  – EU-Bridge (2012-2015): speech and language
  – ...

IWSLT 2011

- IWSLT: Int. Workshop on Spoken Language Translation
- TED lectures: from English to French
- automatic performance measures:
  - TER: error rate: the lower, the better.
  - BLEU: accuracy measure: the higher, the better.

System	Results 2011	
	BLEU [%]	TER [%]
Karlsruhe IT	37.6	41.7
LIMSI Paris	36.5	43.7
RWTH Aachen	36.1	43.7
MIT Cambridge	35.3	44.0
FBK Trento	34.9	44.7
U Grenoble	34.6	44.1
DFKI Saarbrücken	34.4	45.7

WMT 2012

- WMT: ACL Workshop on Machine Translation
- text input: German to English
- domain: news
- QUAERO systems: marked by *

System	Results 2012	
	BLEU [%]	TER [%]
* QUAERO SysCom	24.4	65.4
* Karlsruhe IT	23.4	66.3
* RWTH Aachen	23.3	65.9
U Edinburgh	22.9	67.0
* LIMSI Paris	22.8	67.7
Qatar CRI	22.6	66.8
DFKI Saarbrücken	20.7	70.5
JHU Baltimore	19.7	69.4
U Prague	20.0	71.3
U Toronto	14.0	76.1

2 Inside the Statistical Approach

key ideas of statistical approach to MT:

- **MT (like ASR and other NLP tasks) is a complex task, for which perfect solutions are difficult (compare: all models in physics are approximations!)**
- **consequence: use imperfect and vague knowledge and try to minimize the number of decision errors**
- **statistical decision theory and Bayes decision rule using prob. dependencies between source sentence $F = f_1^J = f_1 ... f_j ... f_J$ and target sentence $E = e_1^I = e_1 ... e_i ... e_I$:**

$$F \rightarrow \hat{E}(F) = \arg\max_{E} \left\{ p(E|F) \right\}$$

- **resulting concept:**

MT = (Linguistic?) Modelling + Statistics + Efficient Algorithms

Statistical MT: Methodology

Bayes decision rule:

$$F \rightarrow \hat{E}(F) = \arg\max_{E} \left\{ p(E|F) \right\} = \arg\max_{E} \left\{ p(E) \cdot p(F|E) \right\}$$

important aspects in the re-written decision rule:

- **two INDEPENDENT prob. distributions (or knowledge sources):**
  $p(F|E)$**: translation model:**
  link to source sentence ('adequacy')
  $p(E)$**: language model:**
  well-formedness of target sentences ('fluency')
  i.e. its syntactic–semantic structure

- **Why this decomposition?**
  each of these models can be trained separately:
  – monolingual data: $p(E)$
  – bilingual data: $p(E|F)$

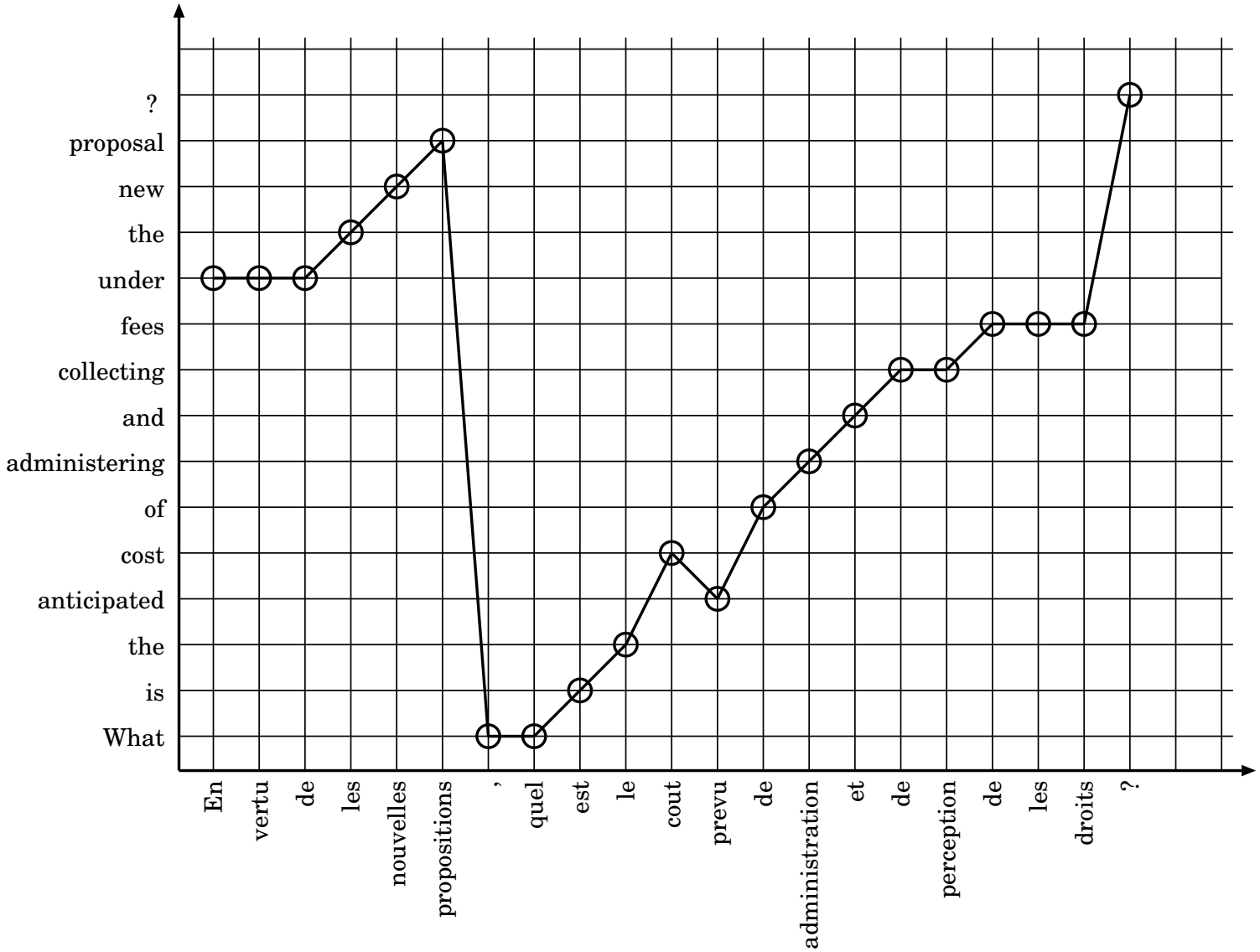
- **generation: = search = maximization over** E
  generate target sentence with the largest posterior probability

Statistical MT: Methodology

- distributions $p(E)$ and $p(F|E)$:
  - are unknown and must be learned
  - complex: distribution over strings of symbols
  - using them directly not possible (sparse data problem)!
- therefore: introduce (simple) structures by decomposition into smaller 'units'
  - that are easier to learn
  - and hopefully capture some true dependencies in the data
- example: ALIGNMENTS of words and positions: bilingual correspondences between words (rather than sentences) (counteracts sparse data and supports generalization capabilities)

$$\begin{aligned} p(F|E) &= \sum_A p(F, A|E) \\ &= \sum_A p(A|E) \cdot p(F|E, A) \end{aligned}$$

Example of Alignment (Canadian Hansards)



HMM: Recognition vs. Translation

speech recognition	text translation
$Pr(x_1^T \mid T, w) =$ $\sum_{s_1^T} \prod_t [p(s_t \mid s_{t-1}, S_w, w)\ p(x_t \mid s_t, w)]$	$Pr(f_1^J \mid J, e_1^I) =$ $\sum_{a_1^J} \prod_j [p(a_j \mid a_{j-1}, I)\ p(f_j \mid e_{a_j})]$
time $t = 1, ..., T$ observations x_1^T with acoustic vectors x_t states $s = 1, ..., S_w$ of word w path: $t \rightarrow s = s_t$ always: monotonic	source positions $j = 1, ..., J$ observations f_1^J with source words f_j target positions $i = 1, ..., I$ with target words e_1^I alignment: $j \rightarrow i = a_j$ sometimes: montonic
transition prob. $p(s_t \mid s_{t-1}, S_w, w)$ emission prob. $p(x_t \mid s_t, w)$	alignment prob. $p(a_j \mid a_{j-1}, I)$ lexicon prob. $p(f_j \mid e_{a_j})$

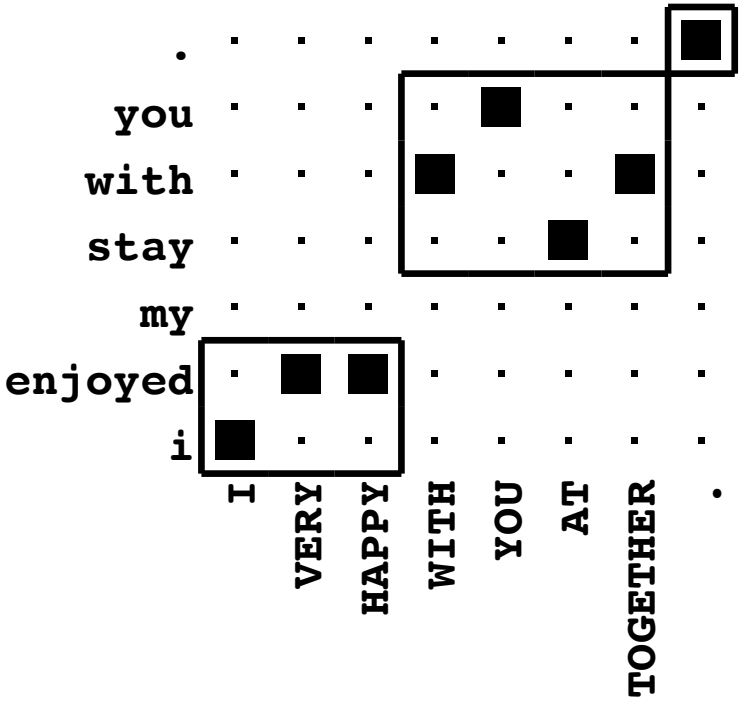
From Words to Phrases

source sentence 我 很 高兴 和 你 在 一起 .

gloss notation I VERY HAPPY WITH YOU AT TOGETHER .

target sentence I enjoyed my stay with you .

Viterbi alignment for $F \rightarrow E$:

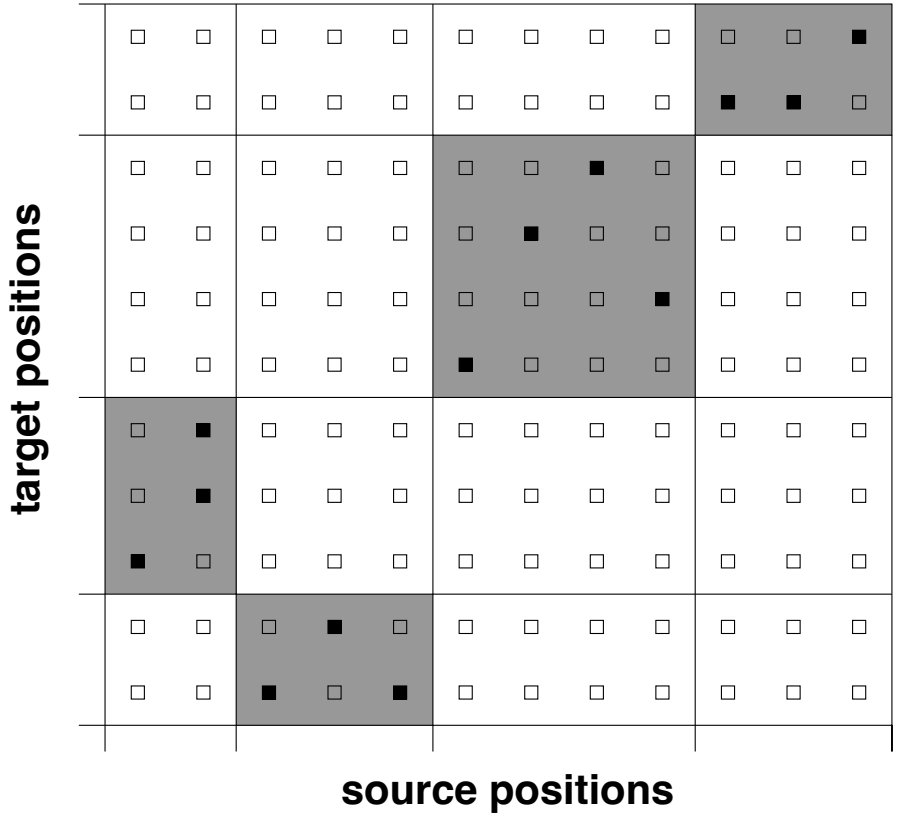


From Words to Phrases (Segments)

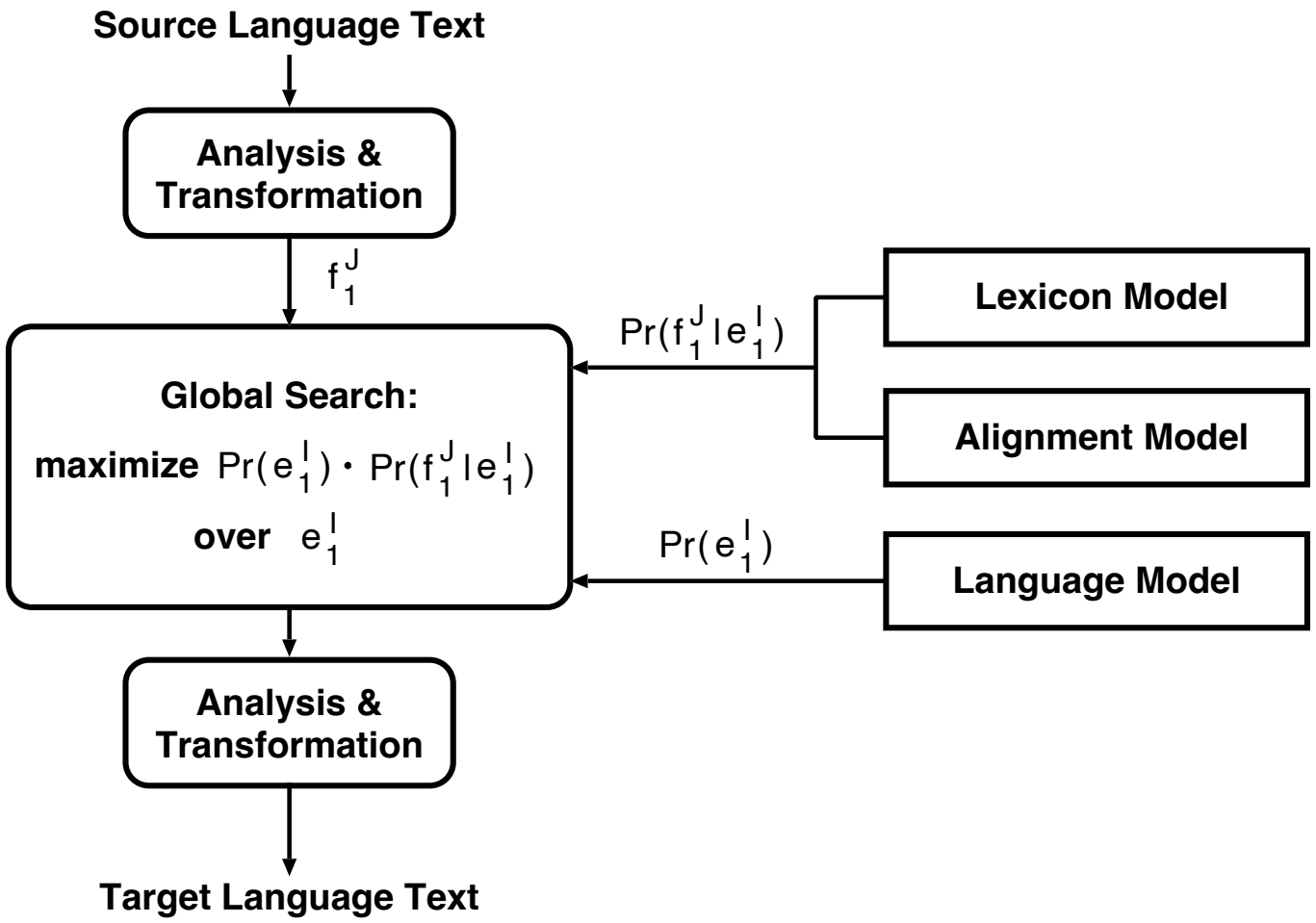
use of into two-dim. 'blocks':
beyond original IBM approach

blocks have to be "consistent"
with the word alignment:

- **words within the phrase cannot be aligned to words outside the phrase**
- **unaligned words are attached to adjacent phrases**



Architecture of a Statistical MT System



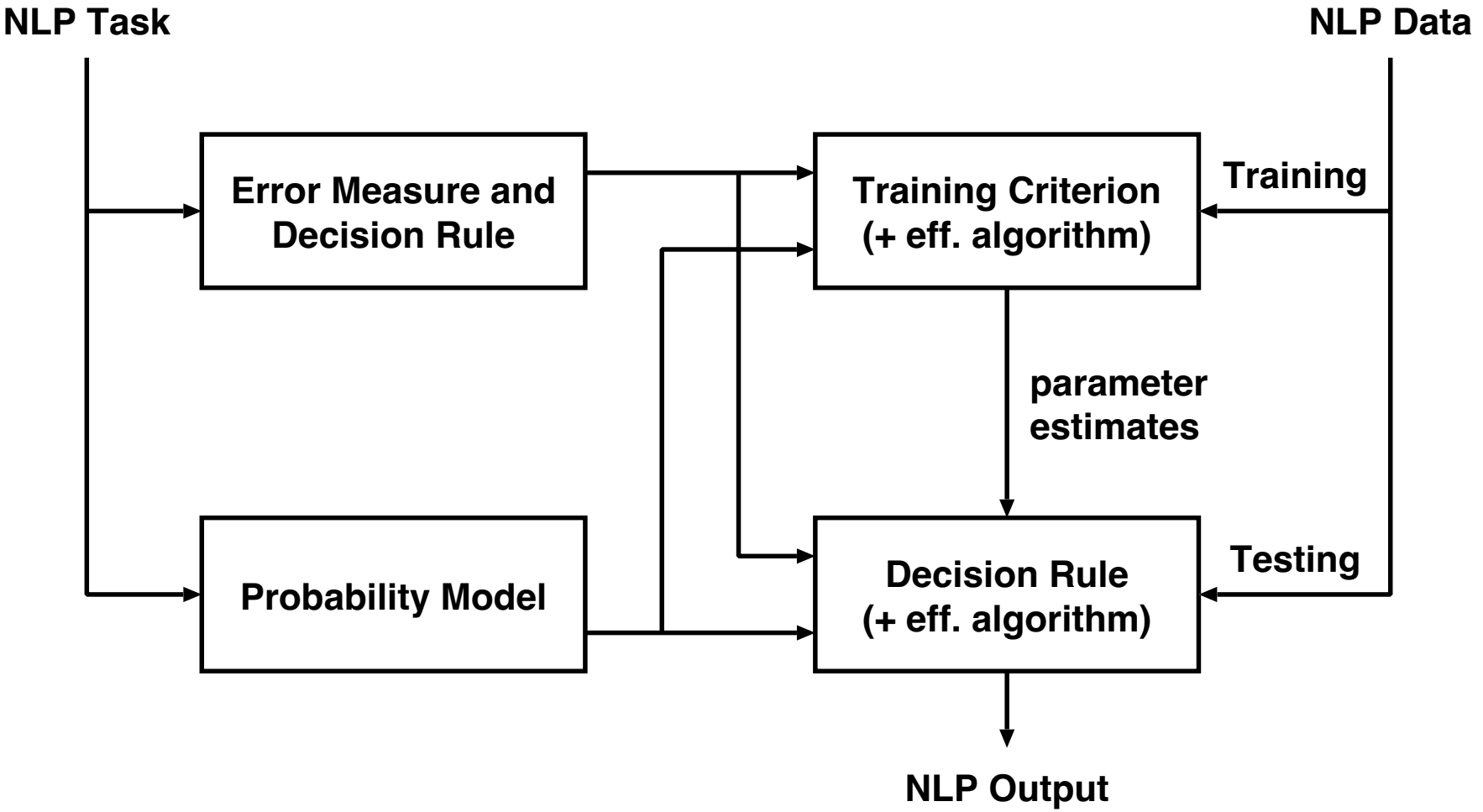
Statistical Approach Revisited

common properties of tasks in HLT:
– strings: input and output
– relevance of context information

four key ingredients:

- **form of Bayes decision rule:**
  cost function = performance measure
- **probability models:**
  (mutual) dependencies between data and within data
  $\rightarrow$ problem-specific knowledge (e.g. from phonetics and linguistics)
- **training criterion**
  along with optimization strategy
- **generation ('search', 'decoding')**
  along with efficient strategy

Why does a system make errors?
none of the four components is perfect!


NLP Task
NLP Data
Error Measure and Decision Rule
Training Criterion (+ eff. algorithm)
Training
parameter estimates
Probability Model
Decision Rule (+ eff. algorithm)
Testing
NLP Output

Beyond 'Orthodox' Statistics

- huge number of free parameters:
  - statisticians prefer models with only a few parameters
  - not enough training data
  - interaction between these parameters
- performance (= error rate) of the whole system matters and not quality of parameter estimates
- task: more 'predictive' than 'descriptive'
- problem-specific knowledge required: how much?
- computational efficiency matters:
  - training procedure
  - search (or generation) process

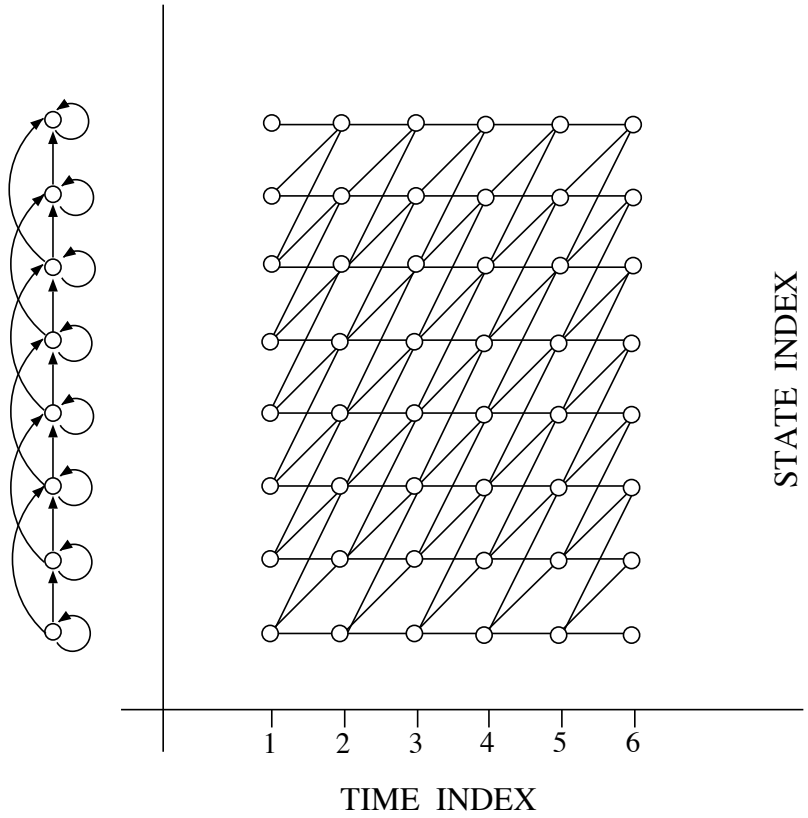
3 From Generative to Discriminative Modelling

- well-known result in classical pattern recognition
  (re-discovered in connectionism and machine learning):
  – estimation point-of-view: discriminative modelling is better
  – disadvantage: no closed-form analytic solutions
- speech recognition:
  – time-alignment problems with HMM
  – generative modelling along with Max.Lik. estimation (and EM algorithm)
  – extensions (add-on): MMI/MPE training
- posterior form of Gaussian (and other generative models):
  – strictly log-linear form
  – training criterion: convex optimization
- problem: feature functions
  – neural networks are good at that!

Log-Linear Modelling

historical development:

- **HMM: first-order model of time alignment problem**
- **emission model: Gaussian**
- **type of model: generative model**
- **training criterion:**
  - **maximum likelihood: EM algorithm (EM = Expectation Maximization)**
  - **EM: dominated the scene until 1990**
- **old discriminative concepts of pattern recognition: 'forgotten'**



observations x_1^T over time $t = 1, ..., T$ for a sentence W:

$$p(x_1^T|W) = \sum_{s_1^T} p(x_1^T, s_1^T|W) = \sum_{s_1^T} \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, W)$$

Overview: Traditional Training Criteria

notation:

- r: sentence index
- X_r: sequence of feature vectors of sentence r
- W_r: spoken word sequence of sentence r,
- W: any word sequence
- $p_\theta(\cdot)$: model with parameter set θ

- **generative model: maximum likelihood (along with EM/Viterbi):**

$$F(\theta) = \sum_r \log p_\theta(W_r, X_r) = \sum_r \log p_\theta(W_r) + \sum_r \log p_\theta(X_r|W_r)$$

  nice property: decomposition into two separate problems: language model $p_\theta(W)$ and acoustic model $p_\theta(X|W)$

- **log class posterior prob. (= MMI, maximum mutual information) [1986 Mercer, 1991 Normandin, ...]:**

$$F(\theta) = \sum_r \log p_\theta(W_r|X_r) \qquad p_\theta(W|X_r) := \frac{p_\theta(W)p_\theta(X_r|W)}{\sum_{W'} p_\theta(W')p_\theta(X_r|W')}$$

- **MCE: minimum classification error rate (old concept in pattern recognition):**

$$F(\theta) = \sum_r \frac{1}{1 + \left(\frac{p_\theta(X_r, W_r)}{\sum_{W \neq W_r} p_\theta(X_r, W)} \right)^{2\beta}}$$

  (β: smoothing constant)

- **MWE/MPE: Povey's [2004+...] minimum word/phoneme error**

$$F(\theta) = \sum_r \sum_W A(W, W_r) p_\theta(W|X_r)$$

  $A(W, W_r)$**: (approximate) accuracy of hypothesis** W **for correct sentence** W_r

HMM and EM Revisited

positive:

- **consistent framework: FULL generative model**
- **virtually closed form solutions by EM:**
  – weighted maximum likelihood estimates
  – weights: 'gammas' computed by EM

negative:

- **starting point: maximum likelihood estimation by EM:**
  more complex than really required:
  density estimation vs. classification problem!
- **extension: discriminative training**
  lots of heuristics (lattice etc)

Hybrid Approach

replace the emission probability in HMM:

- **consider the joint probability (omitting W):**

$$p(x_1^T, s_1^T) = \prod_t [p(s_t|s_{t-1}) \cdot p(x_t|s_t)]$$

- **re-write the emission probability:**

$$p(x_t|s_t) = p(x_t) \cdot \frac{p(s_t|x_t)}{p(s_t)}$$

- **for recognition purposes, the term $p(x_t)$ can be dropped**
- **result: it is sufficient to model the state posterior probability:**

$$x_t \rightarrow p(s_t|x_t)$$

  rather than the state emission distribution $p(x_t|s_t)$

- **justification:**
  – easier problem: (CART) labels $s_t = 1, ..., 5000$ vs. vectors $x_t \in \mathbb{R}^{40}$
  – disadvantage: not a generative model anymore

From Gaussian to Log-Linear Models

- key quantity in HMM: Gaussian model
- Gaussian models show up in various contexts:
  – single events: classification with no context
  – frame level in HMM approach
  – sentence level in HMM apprach
- simplified presentation for (class c, observation x):

$$p(x, c) = p(c) \cdot \mathcal{N}(x|\mu_c, \Sigma_c)$$

  with class dependent parameters:
  prior $p(c)$, mean vector μ_c and covariance matrix Σ_c
- nice decomposition:
  – prior and observation model
  – carries over to training with Max.Lik. criterion

From Gaussian to Log-Linear Models

consider class posterior probability for observation x and class c:

$$
\begin{aligned}
p(c|x) &= \frac{p(c)\,\mathcal{N}(x|\mu_c,\Sigma_c)}{\sum_{c'} p(c')\,\mathcal{N}(x|\mu_{c'},\Sigma_{c'})} = \frac{1}{Z(x)} \cdot p(c)\,\mathcal{N}(x|\mu_c,\Sigma_c) \\
&= \frac{1}{Z(x)} \cdot \frac{p(c)}{\sqrt{\det(2\pi\Sigma_c)}} \exp\Big(-\frac{1}{2}(x-\mu_c)^t\Sigma_c^{-1}(x-\mu_c)\Big) \\
&= \frac{1}{Z(x)} \cdot \exp\Big(-\frac{1}{2}x^t\Sigma_c^{-1}x + \mu_c^t\Sigma_c^{-1}x - \frac{1}{2}\mu_c^t\Sigma_c^{-1}\mu_c - \frac{1}{2}\log\det(2\pi\Sigma_c) + \log p(c)\Big) \\
&= \frac{1}{Z(x)} \cdot \exp(x^t\Lambda_c x + \lambda_c^t x + \alpha_c)
\end{aligned}
$$

with the parameters: $\alpha_c \in \mathbb{R}, \lambda_c \in \mathbb{R}^D, \Lambda_c \in \mathbb{R}^{D \cdot D}$

shift invariance: model $p(c|x)$ does not change by shifting the parameters, e.g. λ_c:

$$\lambda_c \rightarrow \lambda_c + \mu$$

Terms with μ in numerator and denominator cancel!

Log-Linear Models: Properties

result of shift invariance [Heigold et al. 2008]:

- **exact equivalence:**
  For each log-linear posterior probability with 2nd-order features, we can define an equivalent Gaussian model (which is NOT unique!).
- **similar results for other models:**
  – count events: multinomial or Poisson model
  – string models: bigram tagging model (conditional random field)
  – hidden variables: CRF with hidden variables

log-linear modeling for Gaussians:

- **natural training criterion: log. class posterior probability**
- **possible advantages:**
  – 'easier' problem from the estimation point-of-view
  – convex optimization problem
  – full covariance matrix can be used: 'quadratic' features
  – no convergence problems (in principle!)

Log-Linear Models: General Feature Functions

for observation vector $x = [x_1, ..., x_d, ...x_D] \in \mathbb{R}^D$, define

- **polynomial features $y \in \mathbb{R}^{D_y}$:**

$$x \rightarrow y(x) := [1, x_1, ..., x_d, ..., x_D, x_1^2, ..., x_{d_1}x_{d_2}, ..., x_D^2, x_1^3, ..., x_{d_1}x_{d_2}x_{d_3}, ..., x_D^3, ...]$$

- **general feature functions $y_i(x) \in \mathbb{R}, i = 1, ..., I$, e.g. from a neural net:**

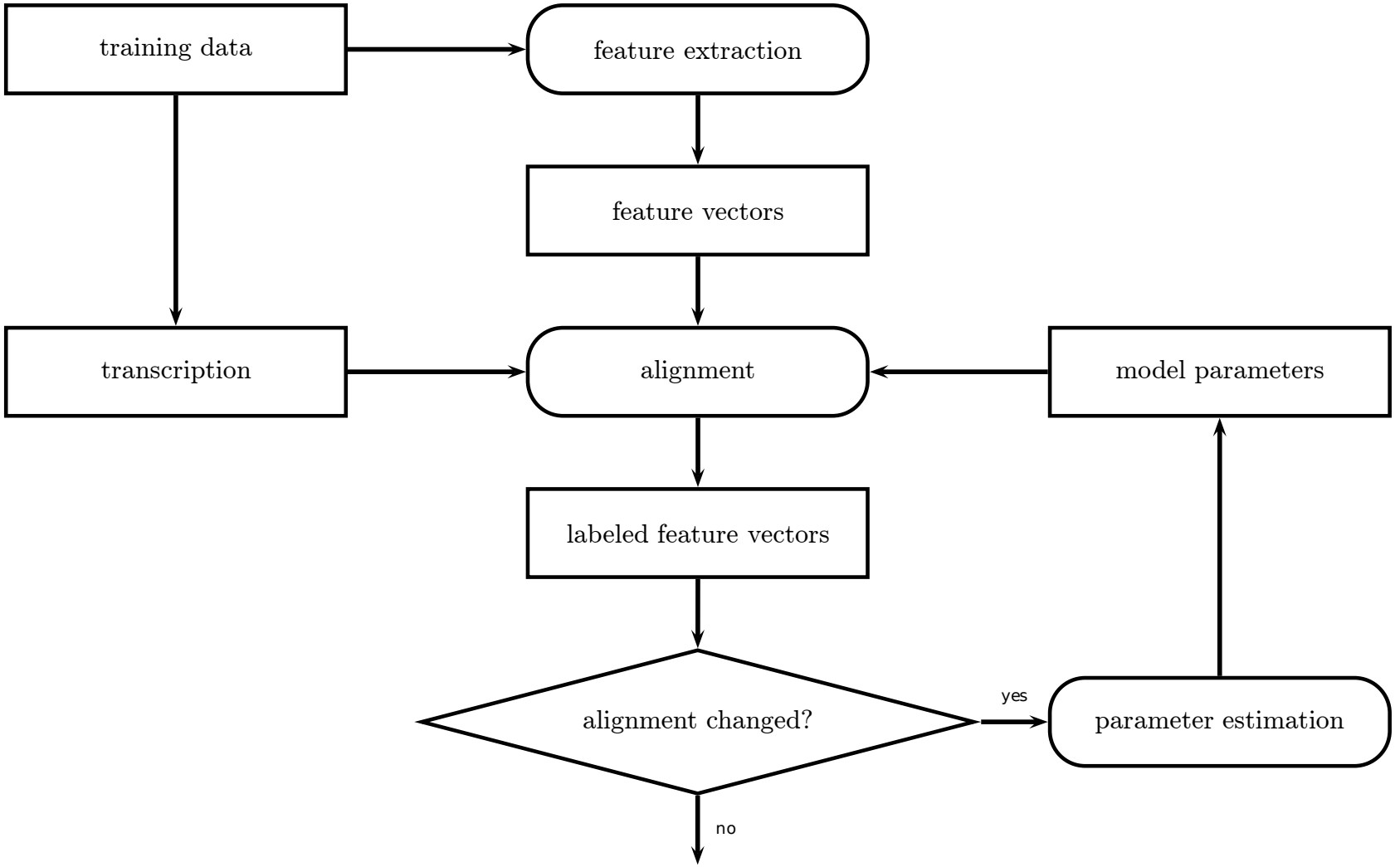
$$x \rightarrow y_i(x) := f_i(x)$$

log-linear model for class posterior probability (with dot product $\lambda_c^t y$):

$$p(c|x) = p(c|y) = \frac{\exp\left(\lambda_c^t y\right)}{\sum_{c'} \exp\left(\lambda_{c'}^t y\right)}$$

note the terminology:
– log-linear: linear in parameters λ_c
– non-linear: in feature function $x \rightarrow f_i(x)$

Hybrid Approach and Frame-based Training



Convergence Problems

high attractivity of log-linear modelling:
– class posterior probability
– convex optimization problem
– speed of convergence?

three types of normalization:

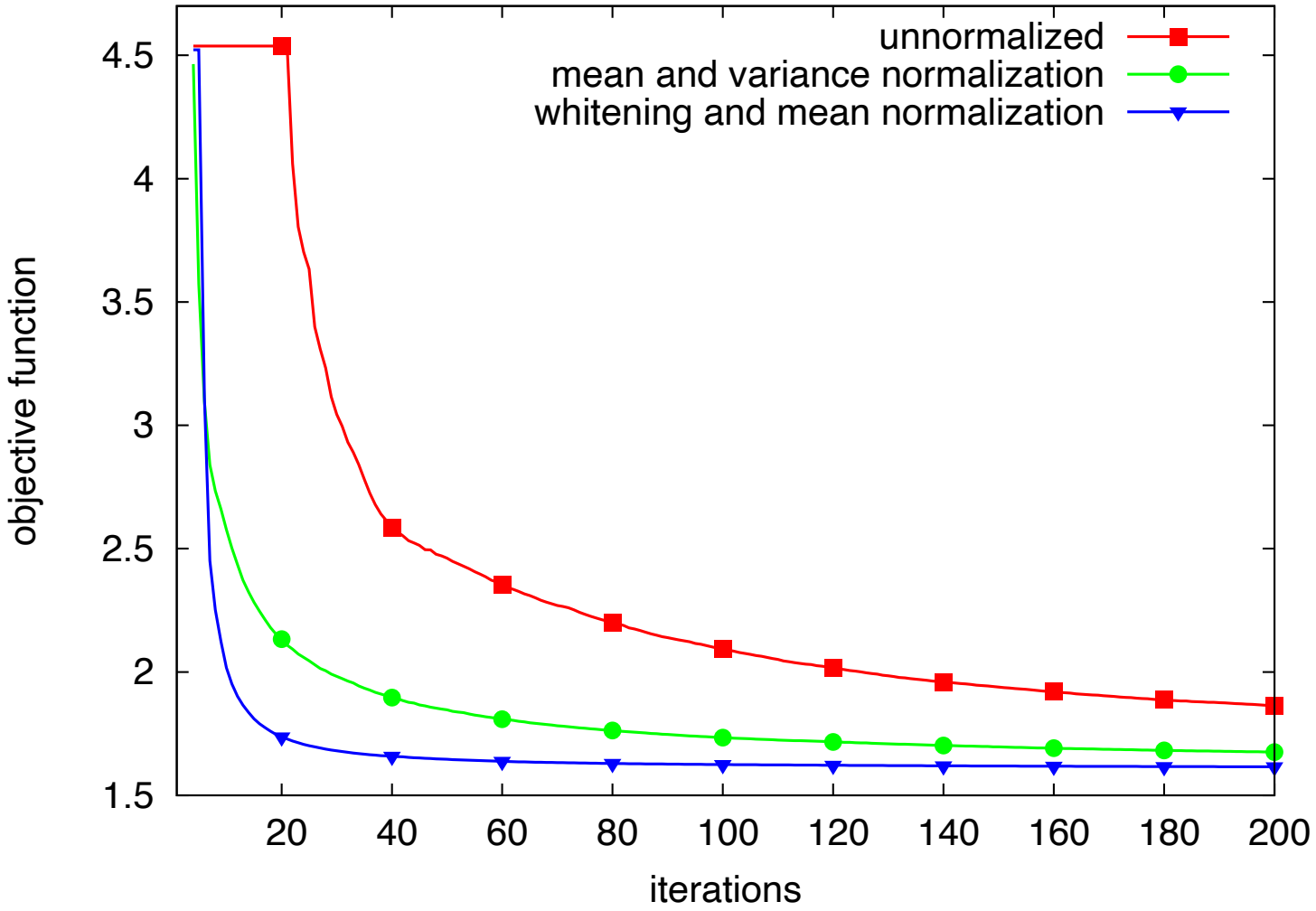
- **no normalization: justification:**
  guaranteed convergence and simplicity of the method
- **mean and variance**
- **mean and whitening (decorrelation)**

details for convergence plot on IAM:

- **observation vector:** $D = 30$ **(PCA applied to pixels)**
- **second-order features:** 495 $(= 30 + 30 \cdot 31/2)$
- **regularization: quadratic term**
- **optimization: L-BFGS**

Convergence: Plot on IAM Corpus

(second-order features)



Results on IAM Corpus (Handwriting)

signal anaysis: PCA from overlapping windows

effect of polynomial order:

Polynomial order	Number of features	WER [%]
first	30	40.2
second	465	31.6
third	5455	27.4

comparison with other approaches:

Author	Site	Method	WER[%]
Dreuw	RWTH	Gauss.Mix. + ML	39.4
Dreuw	RWTH	Gauss.Mix. + MMI	31.6
Dreuw	RWTH	Gauss.Mix. + MPE	30.0
Wiesler	RWTH	log-linear	27.4
Bertolami	U Bern	ROVER with several HMM engines	32.9
Graves	TU Munich	bi-LSTM RNN	25.9

Results: ASR QUAERO English

log-linear model in hybrid approach:

- **2nd-order features:** $1080\ (= 45 + 45 \cdot (45+1)/2)$
- **additional 'cluster' features:** 9 **frames, each with** $2^{12} = 4096$ **Gaussians**
- **training: early stopping**

word error rates [%] on QUAERO English broadcast conversations

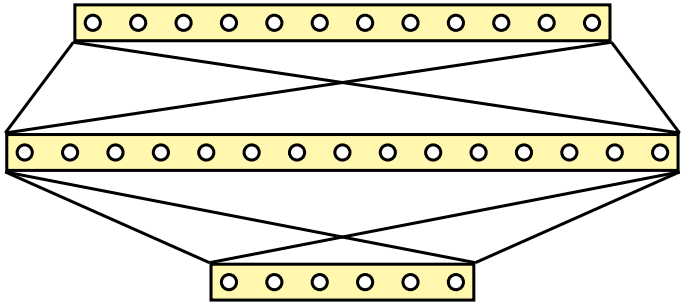
training: 103 hours and dev/eval: 3.5 hours

Method	dev-10	eval-10	eval-11
Gaussian Mixtures and Max.Lik.	25.5	25.1	32.2
+ (Povey's) MPE Training	24.0	24.0	30.6
log-linear model	24.2	24.0	30.8
system combination	22.2	22.3	28.9

promising performance of log-linear models on a large task

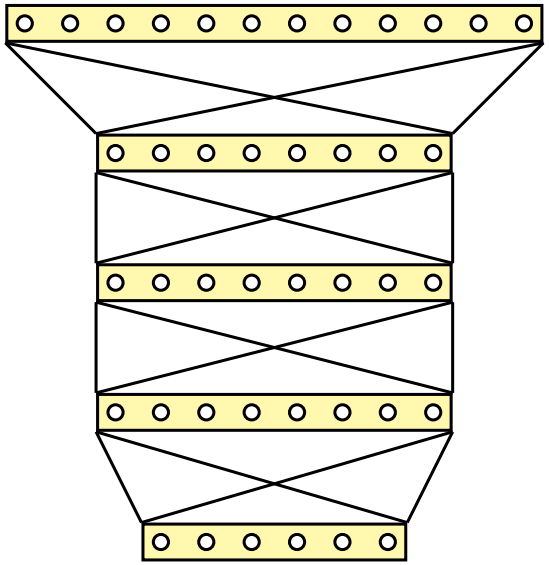
From Log-Linear Models to Neural Networks

- **open problem in log-linear models: what feature functions $f_i(x)$?**
- **neural networks in hybrid approach:**
  - **output layer: softmax = log-linear model**
  - **feature function: remaining part of neural net**
  - **training criterion: entropy (as in log-linear modelling)**
  - **price: convexity is lost!**



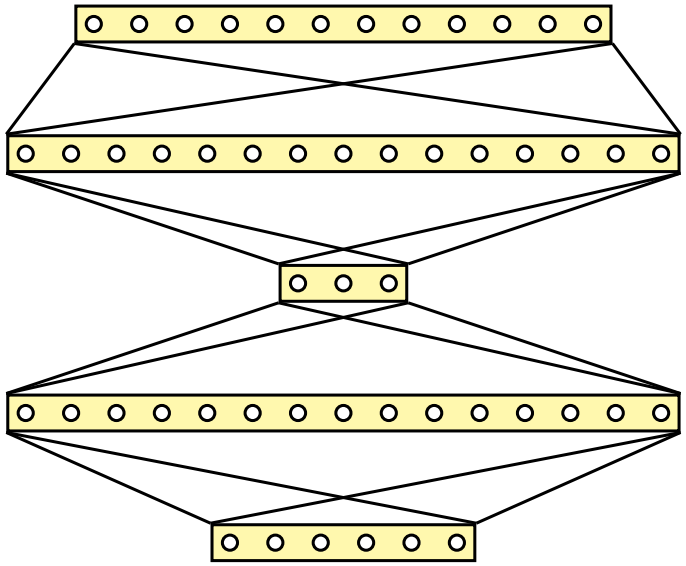
Hybrid Approach: MLP
[Bourlard 1989; Waibel 1989; Robinson 1994; Seide/Yu/Deng 2011; ...]

recent developments:
– output: 4000 CART labels
– 'deep' structure with many layers



Tandem Approach: MLP with Bottleneck
[Ellis 2001; Grezl 2007; LIMSI 2008; ...]

tandem approach:
use bottleneck level to generate feature vector for Gaussian mixtures

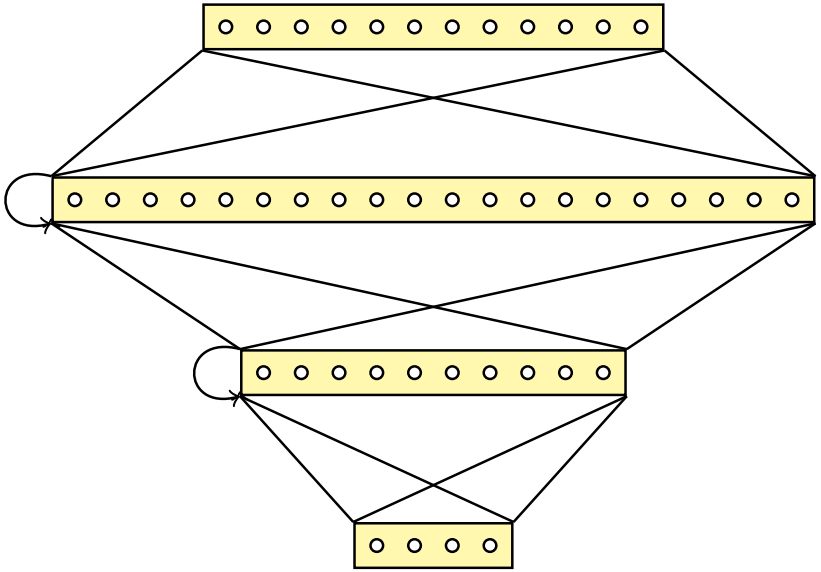


Recent and Ongoing Work
[RWTH 2011-12: Plahl, Tueske, Sundermeyer, Doetsch, ...]

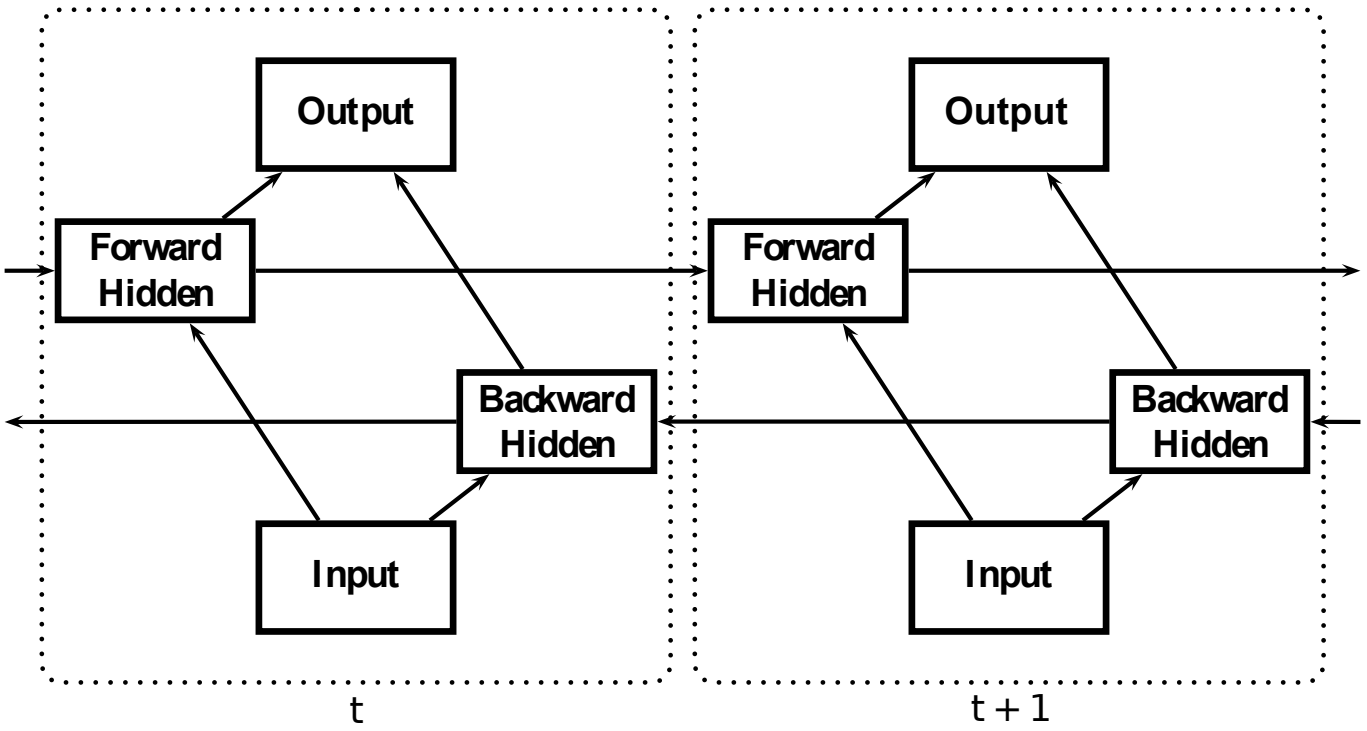
open questions in the context of high-performance systems:

- **which approach is better?**
  - **log-linear vs. hybrid vs. tandem**
  - **answer: tandem? (advantages: +SAT, +MPE)**
- **practical problems:**
  - **is non-convexity a problem?**
  - **what about training time?**
- **what tasks?**
  - **acoustic modelling in ASR**
  - **optical modelling in OCR**
  - **language modelling**
  - **machine translation (underway)**
- **which type of neural network:**
  - **MLP vs. recurrent NN vs. recurrent LSTM (long-short term memory)**
  - **answer: recurrent LSTM?**

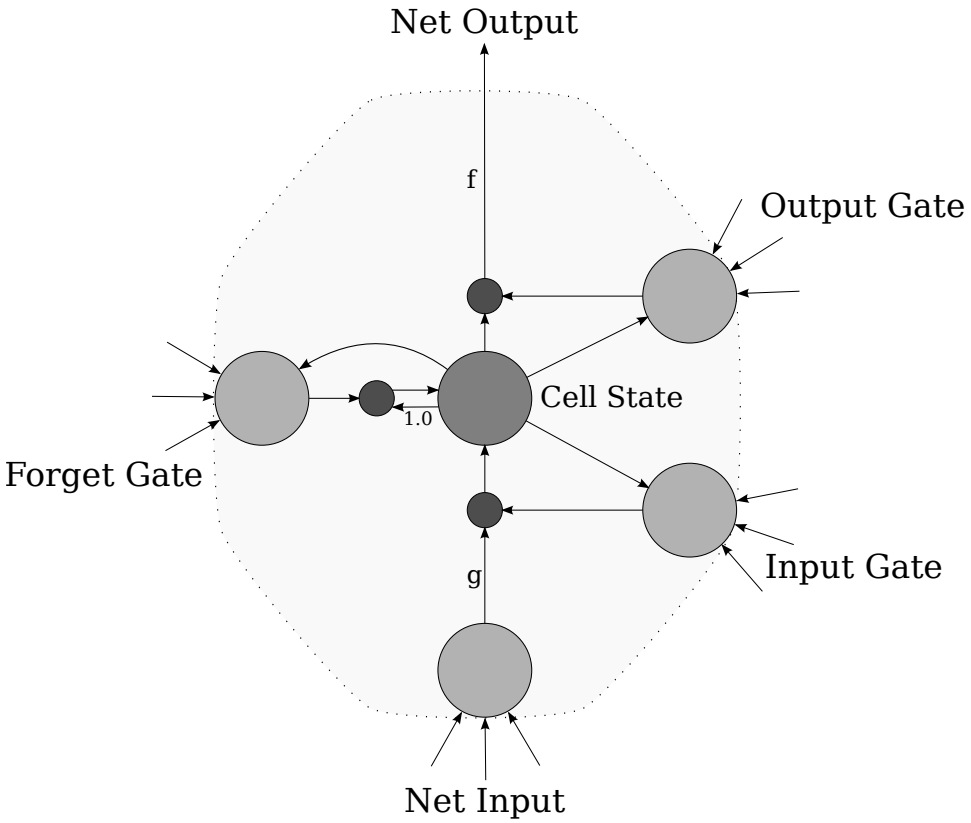
(Bidirectional) Recurrent Neural Network



Bidirectional RNN: Unfolded over Time



Long-Short-Term-Memory: LSTM Net [Graves 2009]



- **Input Gate: controls input INTO cell state**
- **Output Gate: controls output FROM cell state**
- **Forget Gate: controls 'memory' of cell state**

Models and Training Criteria Revisited

quality, correctness, adequacy etc. of models and training criteria along various dimensions:

- **do we have the right model to describe the dependencies?**
- **do we have the right criterion? ML vs. MMI vs. MCE vs ...**
  - **– good link to error rate?**
  - **– errors at which level: frames, phones, words, sentences?**
  - **– robustness of the criterion?**
- **practical problems in training: optimization task:**
  - **– do local optima pose problems?**
  - **– good convergence and efficient implementation?**

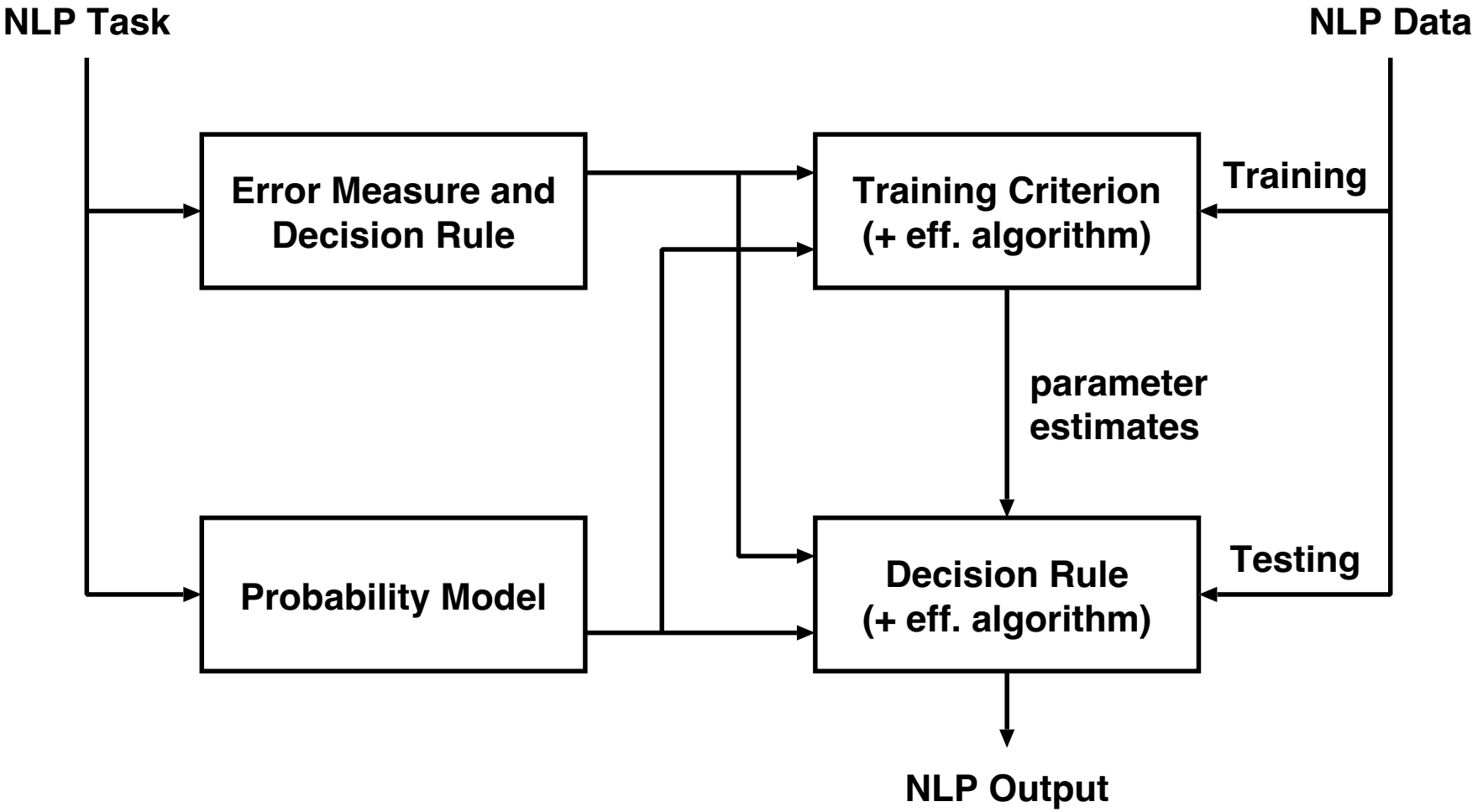
4 Conclusions

What have we learned? (focus on ASR)

- **steady improvements of models and methods (ASR: 40 years)**
- **lion's share of the improvements:**
  - **– better understanding of the modelling and the learning problems**
  - **– more efficient algorithms for learning and search ('generation')**
- **room for ongoing and future improvements:**
  - **– better understanding of interaction of levels: frames, phones, words**
  - **– from log-linear models to neural networks**
  - **– better training criteria, linked to performance**

Methodology has been sucessfully applied to a large variety of tasks:

- **speech recognition**
- **character recognition**
- **machine translation**
- **gesture recognition (sign language)**
- **...**


NLP Task
NLP Data
Error Measure and Decision Rule
Training Criterion (+ eff. algorithm)
Training
parameter estimates
Probability Model
Decision Rule (+ eff. algorithm)
Testing
NLP Output

Towards Better Models for ASR and MT

promising directions:

- **Yes, we need better problem-specific models that extract more information/dependencies from the data.**
- **These models can be related to existing acoustic, phonetic, linguistic, biological theories, but they might also be very much different.**
- **These models have to be extracted from data and verified on data!**
- **Theses models might require a DEEP integration and require research on STATISTICAL decision theory along with efficient algorithms and implementations.**

- **examples of such approaches for MT:**
  - better integration of morphosyntax
  - long-distance dependencies
  - consistent lexicon models (’phrase table’)
  - ...

THE END