Afeka Conference for Speech Processing, Tel Aviv, July 7/8, 2014

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

1 History and Projects

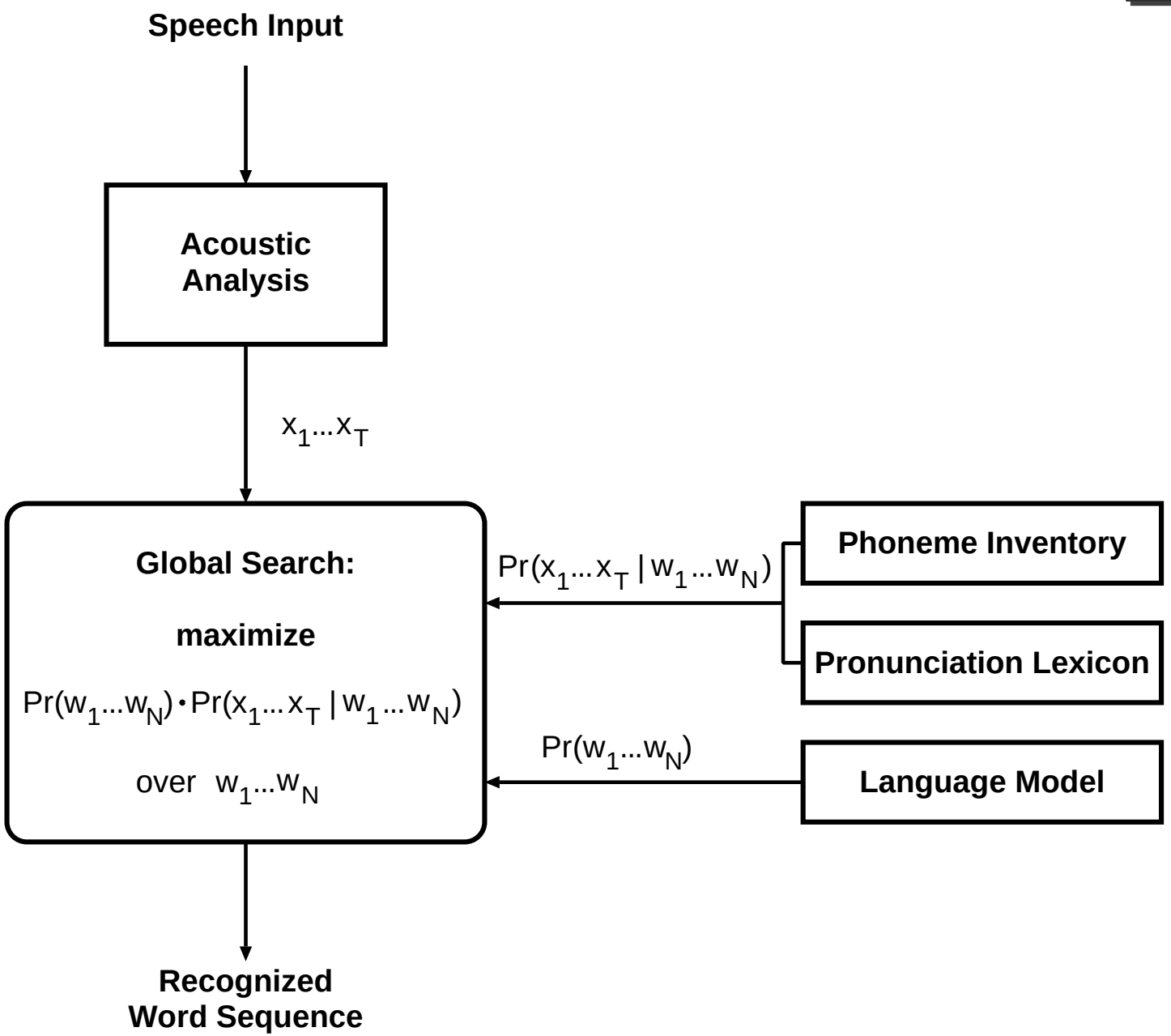
terminology: tasks in speech and natural language processing (NLP)

- **automatic speech recognition (ASR)**
- **optical character recognition (OCR: printed and handwritten text)**
- **machine translation (MT)**
- **additional tasks:**
  text classification, information retrieval, understanding, ...

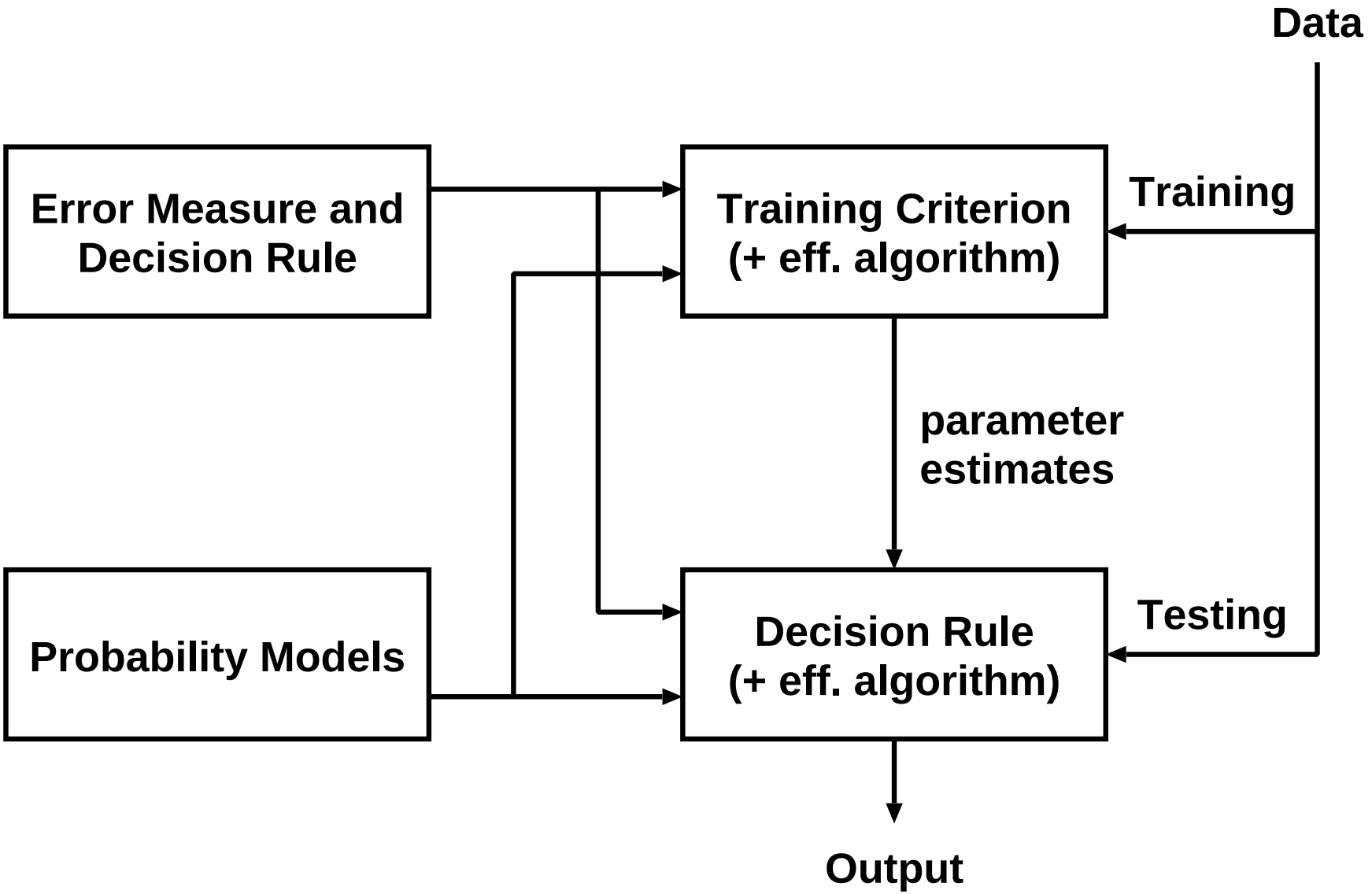
characteristic properties of these tasks (ASR, OCR, MT):

- **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!)

Jelinek, IBM 1975: Statistical Approach to Automatic Speech Recognition (ASR)



four key ingredients:



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:

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

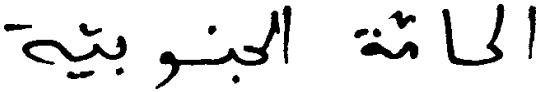
Short History of ASR

- start of statistical approach around 1975 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 (for language modelling)
- CART and phonetic decision trees
- adaptation (unsupervised and supervision light training)
- discriminative training criteria:
  MMI, (Poveys's) MPE, MCE, ...
- machine learning,
  neural networks and log-linear modelling

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	

2 From Speech to Language

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**
- **1994 (informal) evaluation:**
  – comparable to 'conventional' approaches (Systran)
  – results only for French $\rightarrow$ English
- **team went off to Renaissance Technologies (Hedge Fund)**

Statistical MT: First Years

statistical MT: a singularity until 2000

- **IBM research 1988-1994:**
  - **pioneering work: alignment/lexicon models**
  - **basis: single words**
- **period 1995-2000: only a few teams**
  - **HKUST (Dekai Wu)**
  - **CMU (Alex Waibel)**
  - **UP Valencia: EUTrans**
  - **RWTH Aachen: EUTrans, Verbmobil**
- **RWTH: first PhD students on SMT:**
  - **Cristoph Tillmann: $\rightarrow$ IBM research**
  - **Stephan Vogel: $\rightarrow$ QCRI in Qatar**
  - **Franz J. Och: $\rightarrow$ Google (Google Translate)**
  - **more students: $\rightarrow$ Google**
- **DARPA TIDES and evaluations (since 2000):**
  - **teams: IBM, CMU, ISI, ...**
  - **external: RWTH superior results (due to phrase-based approach)**

after 2002: mainstream for machine translation

From Text to Speech Translation: EU Project TC-Star (2004-2007)

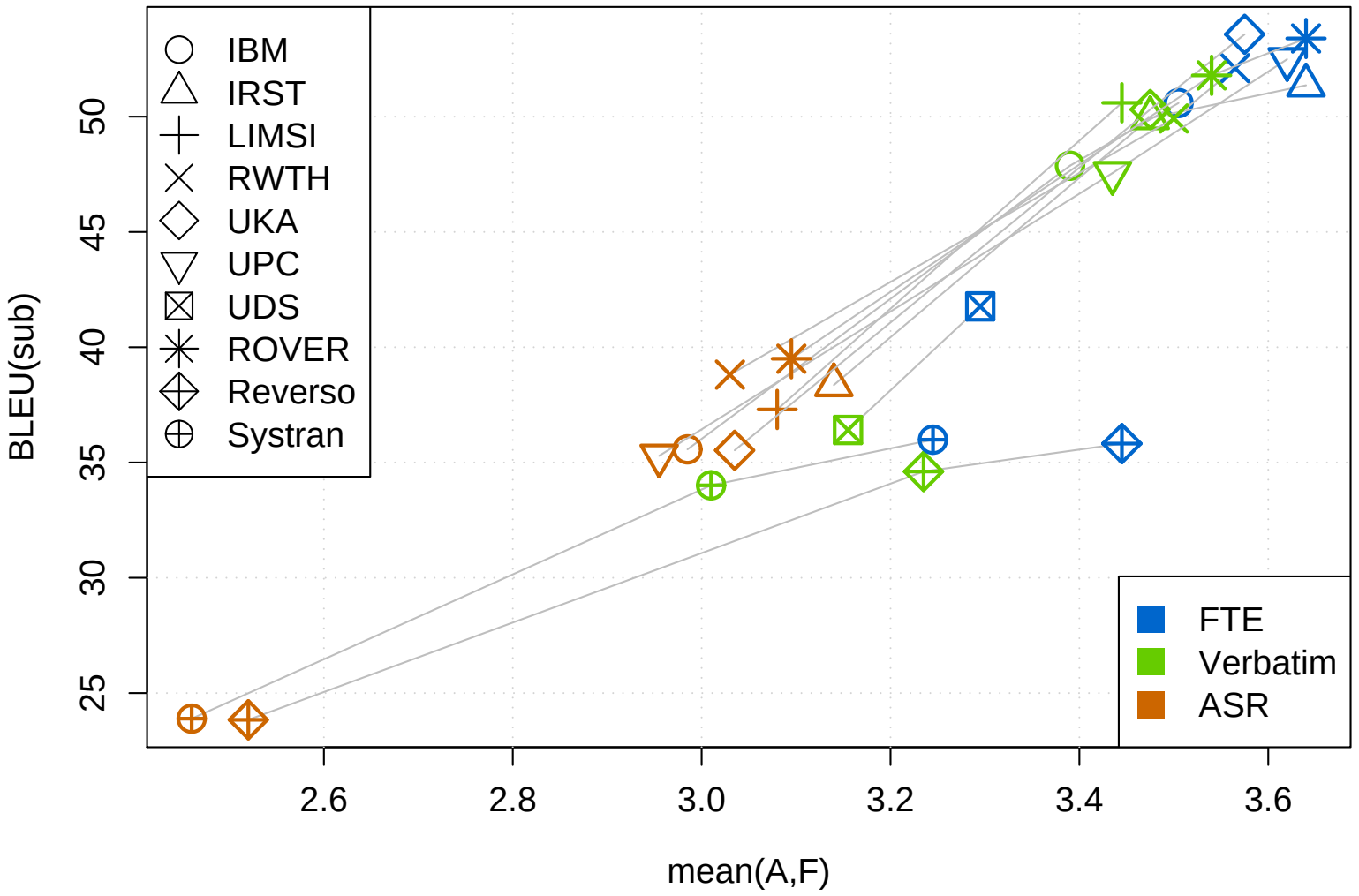
domain: SPEECHES given in the European Parliament

- **work on a real-life task:**
  - **'unlimited' domain**
  - **large vocabulary**
- **speech input:**
  - **cope with non-grammatical input and disfluencies**
  - **handle recognition errors**
  - **sentence segmentation**
- **FIRST research prototype ever on speech translation for unlimited domain and real-life data**

experimental results:

good performance

TC-Star 2007: E → S: Human vs. Automatic Evaluation



RWTH Projects 2001 – 2014

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

- **TC-Star: 2004-2007 funded by EU:**
  EU Parliament: text and speech
- **GALE 2005-2011 (and BOLT 2012-2014)**
  funded by DARPA (funding: 40 Mio US$ per year):
  – text (newswire + webtext) and speech (broadcast news + conversations)
  – Arabic/Chinese to English
  – ASR, MT and information extraction ('question answering')
  – performance measure: HTER (= human translation error rate)
- **QUAERO 2008-2013 funded by OSEO France:**
  – text and speech (news, lectures, discussions, ...)
  – more colloquial text and speech
- **recent EU projects:**
  – SIGNSPEAK: sign language recognition and translation
  – TransLectures: ASR and MT for academic lectures
  – EU-Bridge (2012-2015): ASR and MT for BN and TED Lectures
  – ...
- **BABEL (funded by IARPA): speech recognition only (spoken term detection)**

Evaluation 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

Evaluation 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

3 Inside Statistical MT

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:

- ***lexical choice:* select the right target word**
- ***re-ordering:* select the right position for the target word**
- ***syntax and semantic:* 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

key ideas of statistical approach:

- **MT (like most 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 decision rule:

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

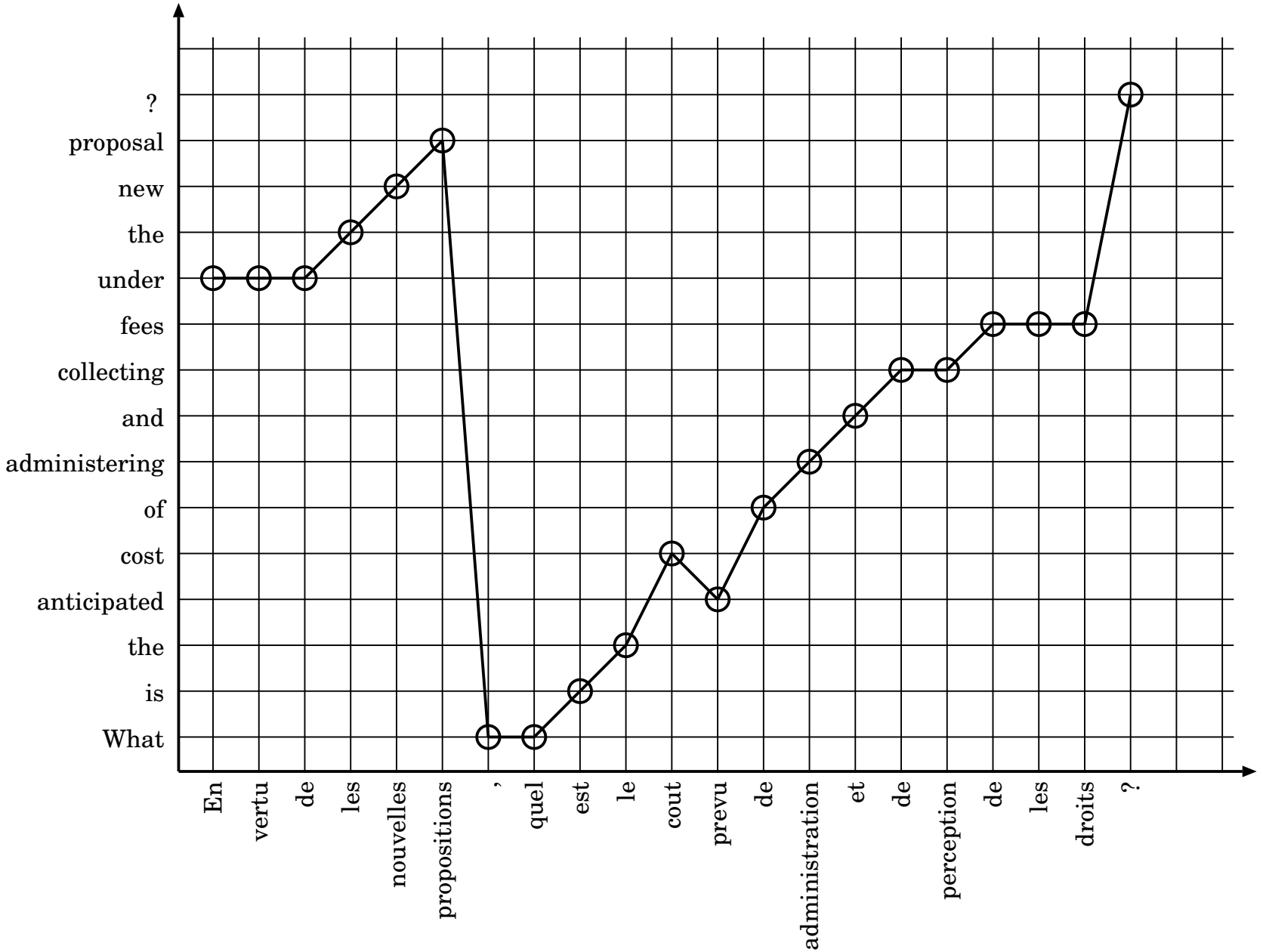
  Why this decomposition?
  each of these can be modelled separately

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

- 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

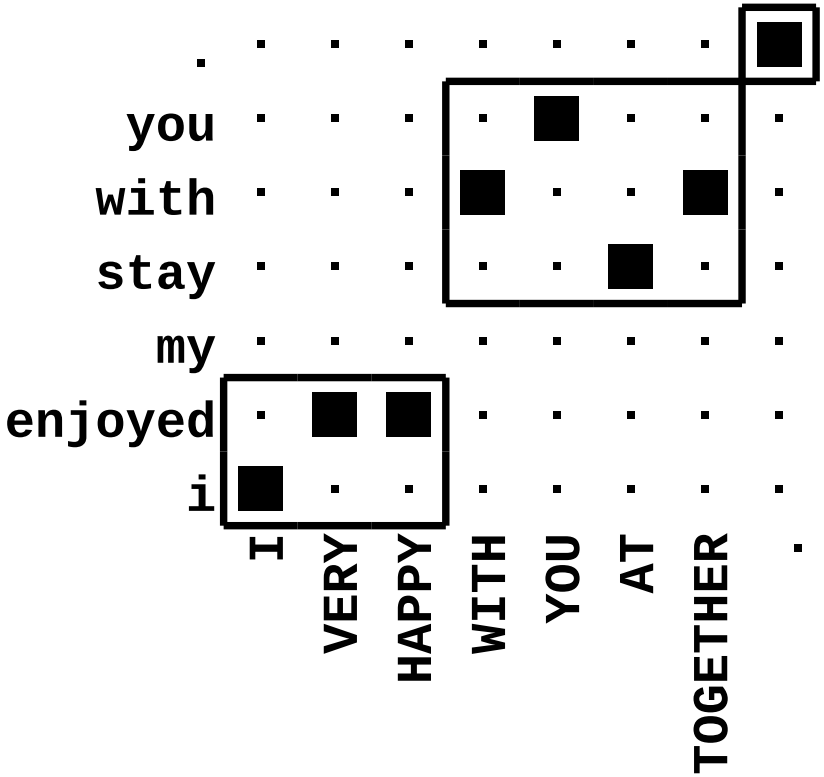
speech recognition	text translation
$Pr(x_1^T \mid T, w) =$ $\sum_{s_1^T} \prod_t [p(s_t \mid s_{t-1}, S_w, w)\ p(x_t \mid s_t, w)]$	$Pr(f_1^J \mid J, e_1^I) =$ $\sum_{a_1^J} \prod_j [p(a_j \mid a_{j-1}, I)\ p(f_j \mid e_{a_j})]$
time $t = 1, ..., T$ observations x_1^T with acoustic vectors x_t states $s = 1, ..., S_w$ of word w path: $t \rightarrow s = s_t$ always: monotone	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: montone
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. $p(f_j \mid e_{a_j})$

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 .

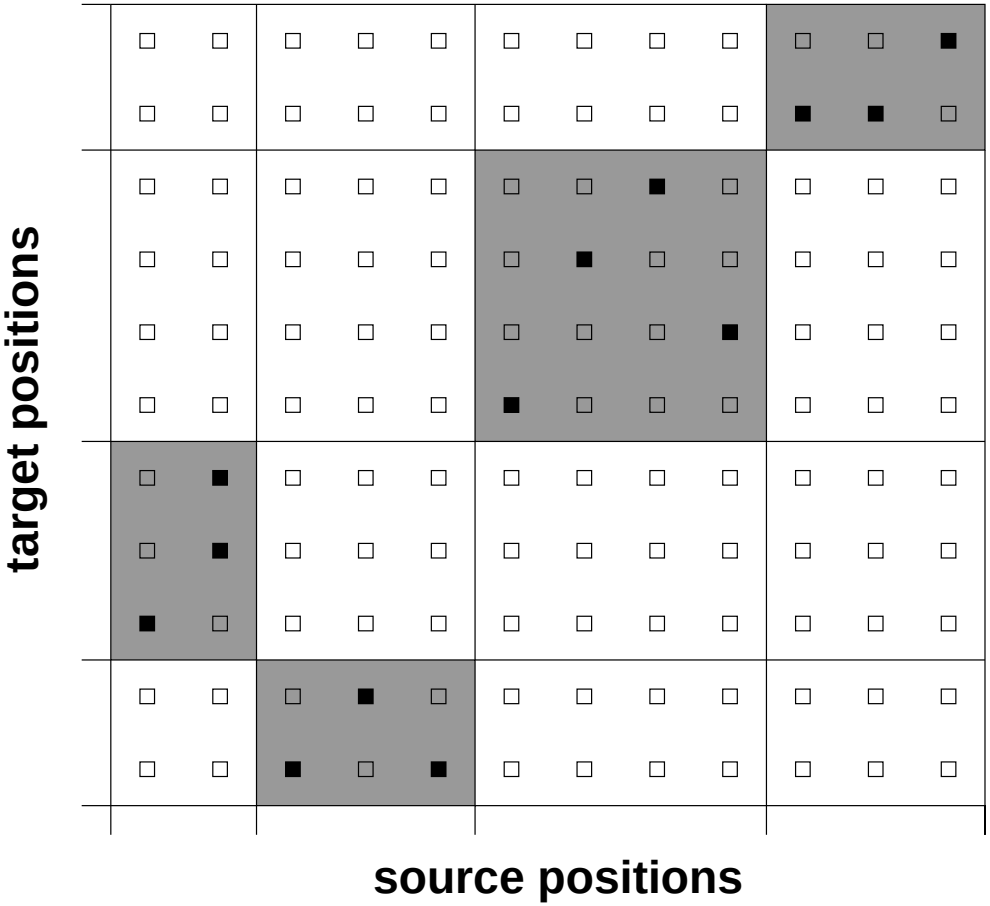
Viterbi alignment for $F \rightarrow E$:



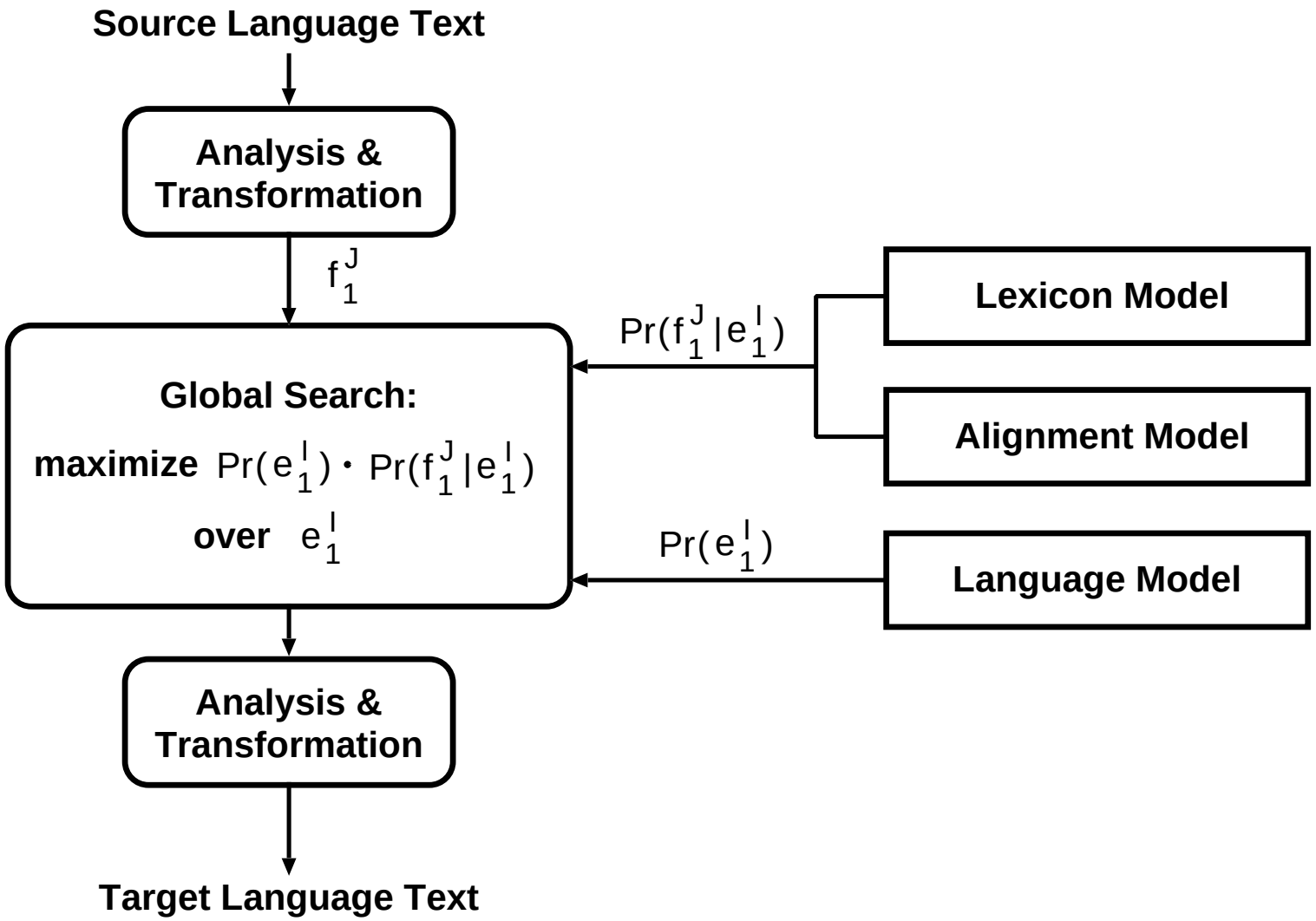
From Words to Phrases (Segments)

phrase-based approach:

- **training: extraction of phrase pairs (= two-dim. 'blocks') after alignment/lexicon training (GIZA++)**
- **translation process: phrases are the smallest units**
- **experimental evaluation: significant improvement over IBM's single-word based models**



Architecture of a Statistical MT System



Statistical Approach Revisited

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

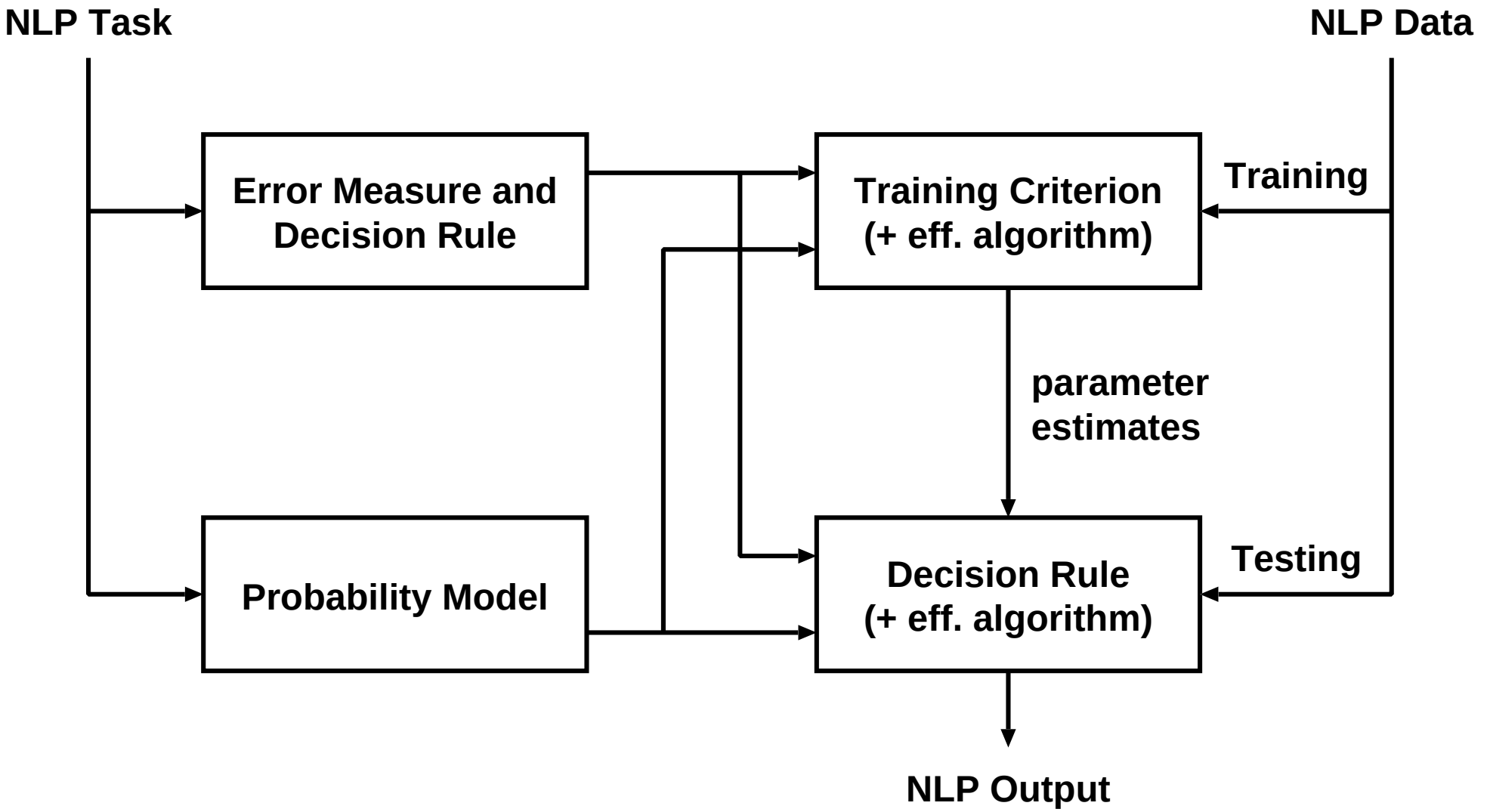
four key ingredients:

- **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!

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


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

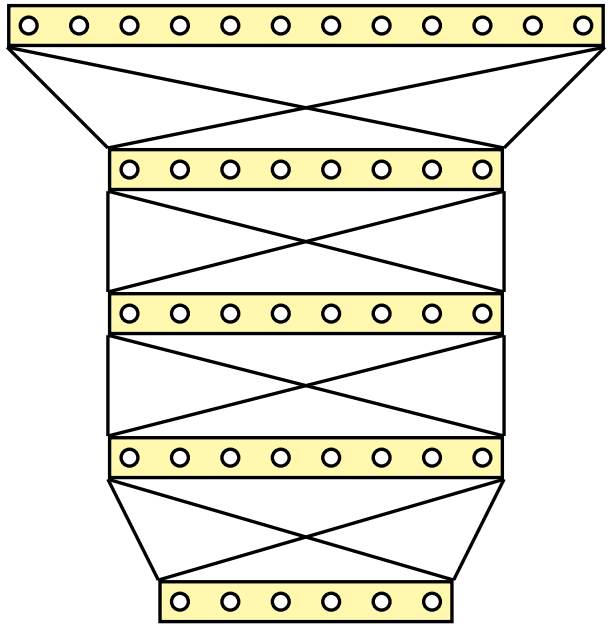
4 Revival of Artificial Neural Nets (ANN)

typical ANN structure:

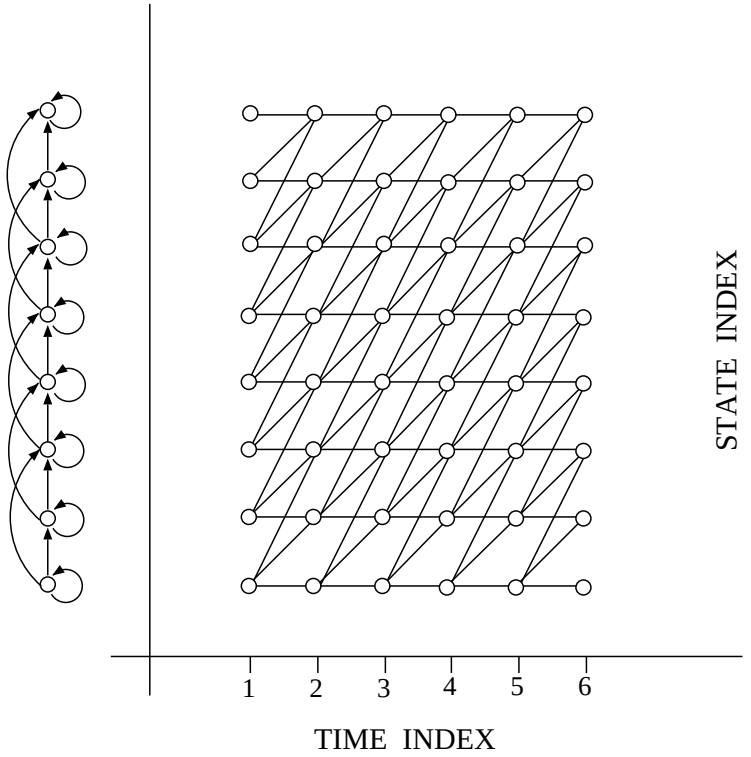
MLP: feedforward multi-layer perceptron with input, hidden, and output layers

two important properties:

- **outputs: can be shown to be (estimates of) posterior probabilities**
- **output layer: softmax operation (results from Gaussian assumptions)**



- **fundamental problem in ASR: non-linear time alignment**
- **Hidden Markov Model:**
  - **linear chain of states $s = 1, ..., S$**
  - **transitions: forward, loop and skip**
- **trellis:**
  - **unfold HMM over time $t = 1, ..., T$**
  - **path: state sequence $s_1^T = s_1 ... s_t ... s_T$**
  - **observations: $x_1^T = x_1 ... x_t ... x_T$**



link between word string hypothesis W and observations x_1^T:

- **joint probability for word string W with hidden state sequence s_1^T:**

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

- **key quantity: emission probability $p(x_t|s, W)$ of state s of word string W realized by GMM: Gaussian mixtures models (trained by EM algorithm)**
- **phonetic labels (allophones, sub-phones): $(s, W) \rightarrow \alpha = \alpha_{sW}$**

$$p(x_t|s, W) = p(x_t|\alpha_{sW})$$

Hybrid Approach

consider modelling the acoustic vector x_t:

- **re-write the emission probability for label α and acoustic vector x_t:**

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

  for recognition purposes, the term $p(x_t)$ can be dropped

- **result: model the label posterior probability by an ANN:**

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

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

- **justification:**
  – easier learning problem: labels $\alpha = 1, ..., 5000$ vs. vectors $x_t \in \mathbb{R}^{D=40}$
  – well-known result in (old) pattern recognition

- **additional property:**
  if $p(x_t|\alpha)$ is Gaussian with pooled co-variance,
  the state posterior $p(\alpha|x_t)$ results in softmax operation

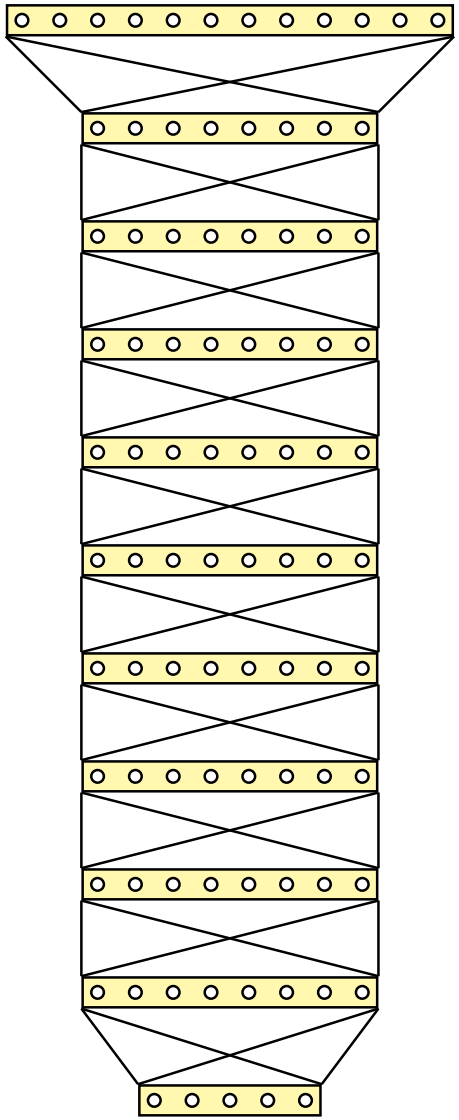
- **experimental results:**
  significant improvements ONLY recently

Hybrid Approach: What is different today?

comparison: today vs. 1989-1994:

- **number of hidden layers:**
  deeper nets, i.e. up to 10 rather than 2-3
- **number of output nodes:**
  5000 rather than 50
- **better optimization strategies,**
  e.g. layer-by-layer pretraining
  and other heuristics
- **much more computation power today**

experiments: WER reduction by 15-25%



Comparison: GMM vs. MLP [RWTH, Tüske 2013]

word error rates [%] (QUAERO English, eval 2011):

- **Gaussian mixtures model (GMM):**

method	50 hrs	250 hrs
Max.Lik.	31.2	28.6
+ f-MLLR	28.7	26.4
Min.Phone Error	30.4	

- **hybrid-MLP**

method	50 hrs	250 hrs
flat MLP: 1 hidden layer	30.6	24.5
deep MLP: 9 hidden layers + f-MLLR	25.2	20.4

conclusions:
– using a DEEP MLP is important
– MLP learns better than GMM with more data

Deep MLP: Optimizing the MLP Architecture [RWTH, Tüske 2013]

- effect of number of hidden layers

  (2000 nodes in output and hidden layers)

n	#param	50 hrs	250 hrs
1	10M	31.3	26.5
2	14M	28.3	-
3	18M	26.7	22.5
4	22M	26.1	-
5	26M	26.0	-
6	30M	25.4	21.2
7	34M	25.5	-
8	38M	25.7	-
9	42M	25.3	20.9

- effect of size of hidden layers

  (9 hidden layers, 2000 output nodes)

hidden size	#param	50 hrs	250 hrs
500	4.5M	-	22.7
1000	13M	26.1	21.1
2000	42M	25.3	20.9
3000	87M	25.2	20.9
5000	225M	25.4	20.8

- effect of size of output layer

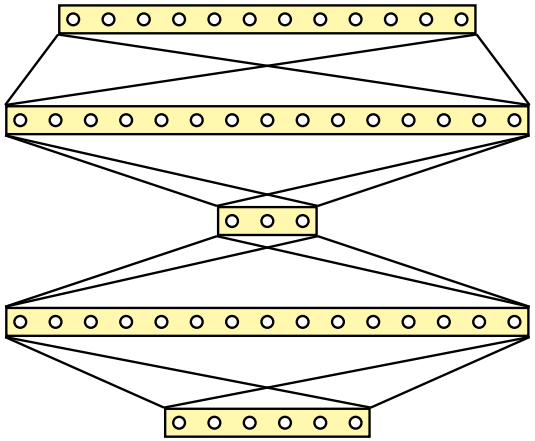
  (9 hidden layers, each with 2000 nodes)

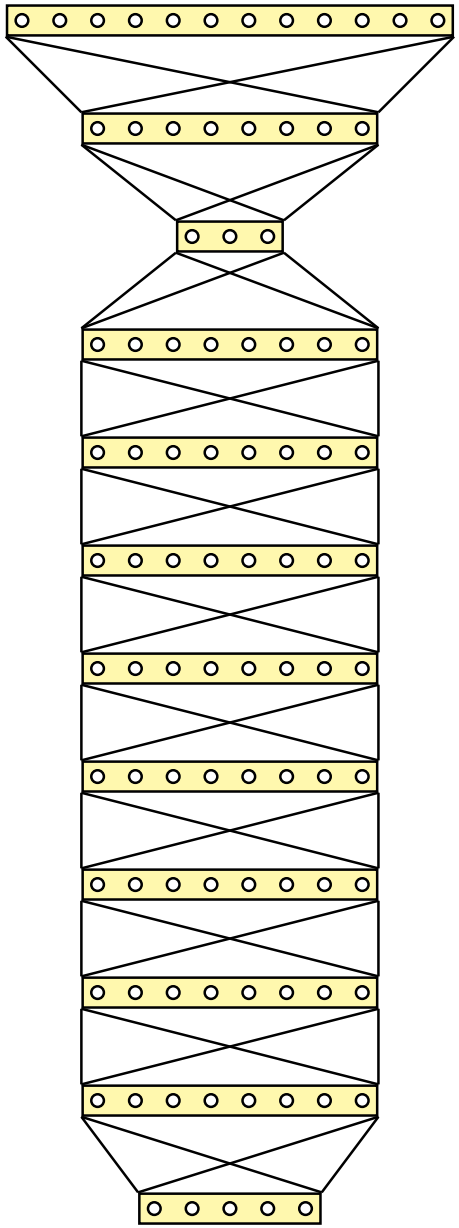
output size	#param	50 hrs	250 hrs
4500	42M	25.3	20.9
9000	51M	25.9	20.4

Tandem Approach

approach:

- **tandem: use MLP for feature extraction in an GMM (Hermansky, Ellis, Sharma 2000)**
- **bottleneck: one narrow hidden layer (Grezl et al. 2007; Grezl & Fousek 2008)**





Tandem Approach: Results [RWTH, Tüske 2013]

word error rates [%] (QUAERO English, eval 2011):

GMM with Max.Lik.	31.2	28.6
hybrid MLP tandem MLP	25.4 25.5	20.4 21.6
system combation: hybrid + tandem	23.8	19.7
more combinations (MRASTA, multi-lingual, f-MLLR, ...)		17.8
'complex' system submission in 2011		19.8

observations:
– clean architecture: best approach is deep MLP in a hybrid approach
– 'complex system': improvement by deep MLP become smaller

History: ANN in ASR

hybrid approach in ASR:
in HMM, replace emission probability by ANN output

- **1990 Bourlard & Wellekens (re-discovered):**
  – ANN outputs can be interpreted as probabilities
- **1990 Bridle: softmax operation for probability normalization**
- **1990–...: Bourlard et al.: advocated the use of MLP in HMM**
  – never competitive
- **1994 Robinson: recurrent NN**
  – competitive results on WSJ task
  – work remained a singularity in ASR
- **2000–...: work towards 'deeper' NNs**
  using TANDEM and HYBRID approach

situation:

ANNs were never really competitive with GMMs

More History: ANN in ASR

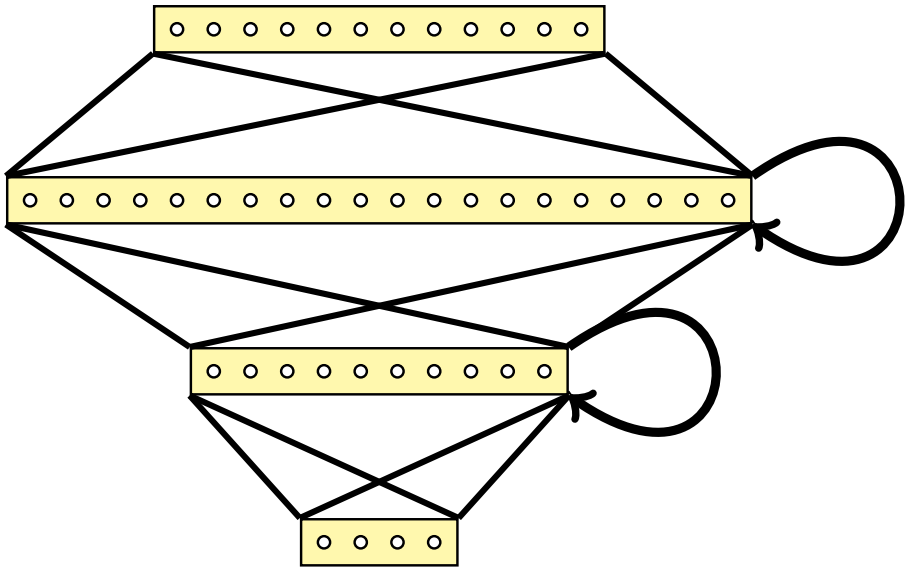
concepts of deep learning:

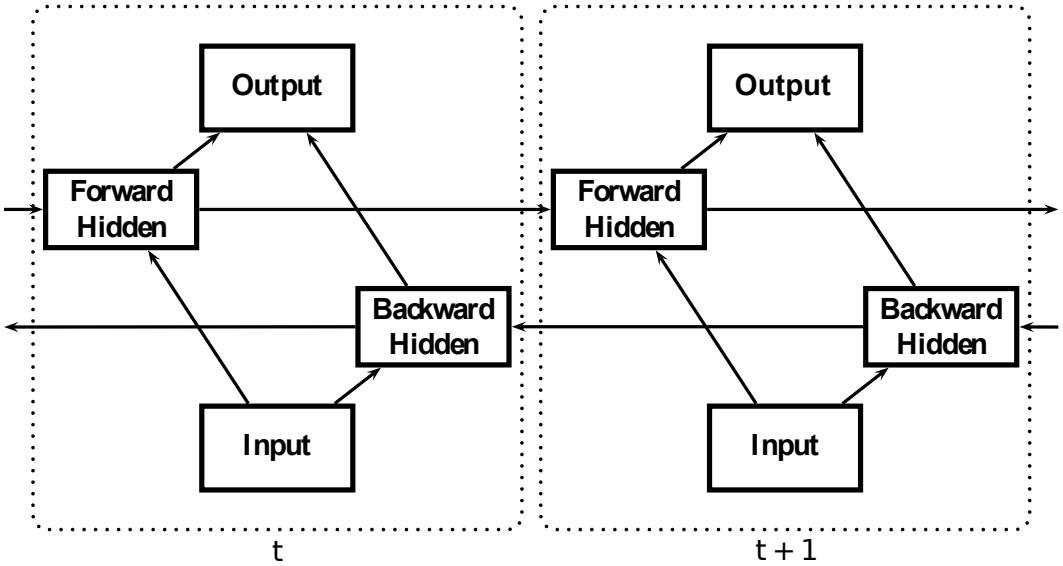
- **2002 Utgoff & Stracuzzi: many-layered learning**
- **2006 Hinton: deep learning (belief nets)**
- **2011 Seide, Deng & Yu (Microsoft):**
  – combined deep learning with hybrid approach
  – significant improvement by deep MLP on a large-scale task

related approaches:

- **1989 Waibel, Hanazawa, Hinton, Shikano, Lang:**
  Phoneme recognition using time-delay neural networks
- **1996 Fritsch, Finke, Waibel:**
  Hierarchical mixtures of experts
- **1997 Hochreiter & Schmidhuber:**
  Long short-term memory neural computation
- **...**

(Bidirectional) Recurrent Neural Network
[LSTM: Hochreuter & Schmidhuber 1997]





More ANNs

signal ('sub-symbolic') processing:
speech/audio, handwriting/OCR, image/video

'symbolic' processing for language modelling and translation:

- **1989 M. Nakamura & K. Shikano:**
  English word category prediction based on neural networks.
- **1997 J.M. Castro, F. Casacuberta, F. Prat:**
  Towards connectionist language models.
- **1997 A. Castano, F. Casacuberta:**
  A connectionist approach to machine translation.
- **2007 H. Schwenk:**
  Continuous space language models.
- **2006 H. Schwenk, M.R. Costa-jussa, J.A.R. Fonollosa:**
  Smooth bilingual n-gram translation.
- **2012 H. Son Le, A. Allauzen, F. Yvon:**
  Continuous space translation models with neural networks.

now: ANNs in language show competitive results/contributions.

Conclusions:
What did "Deep MLPs" bring? What did we learn?

deep MLPs:

- **result in significant improvements**
- **seem to learn 'better' (than GMMs),**
  e.g. multi-lingual and multi-style training
- **maybe: computationally expensive (in training),**
  but cleaner architecture of ASR system
- **overfitting: apparently no problem,**
  due to cross-validation (early stopping) and MLP structure (?)
- **general experience in ASR:**
  The correct principles are not sufficient, we must get ALL the details right.
- **long-term view:**
  it took 25 years or more ...
- **...**

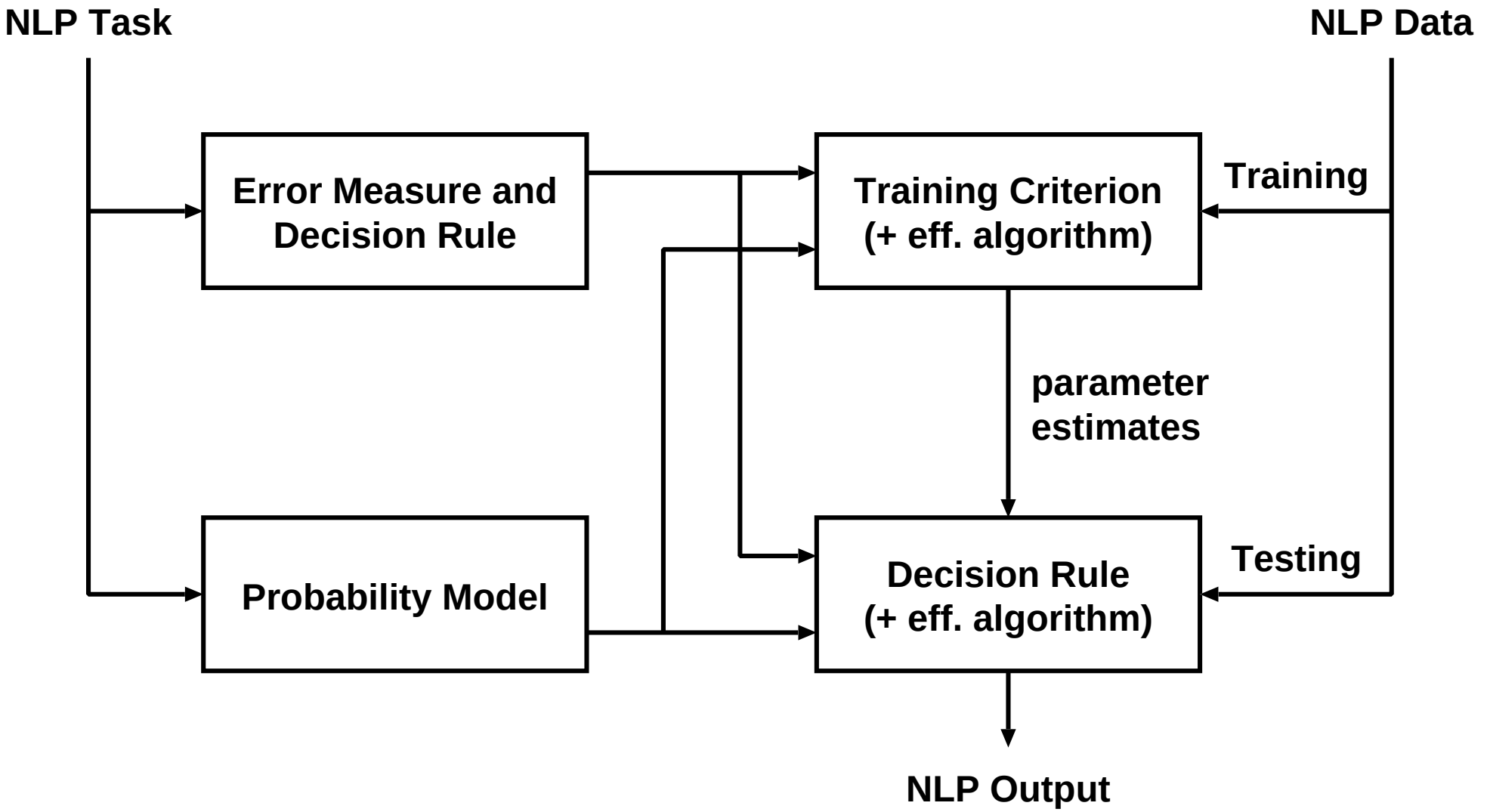
5 Conclusions

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**
  - **– better training criteria, linked to performance**
  - **– better feature extraction, e.g. by neural networks**

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

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?

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’)
  - ...

END