15th International Conference on Frontiers in Handwriting Recognition
ICFHR 2016, Shenzhen, China; October 23-26, 2016

Handwriting and Speech Recognition: From Bayes Decision Rule to Deep Neural Networks

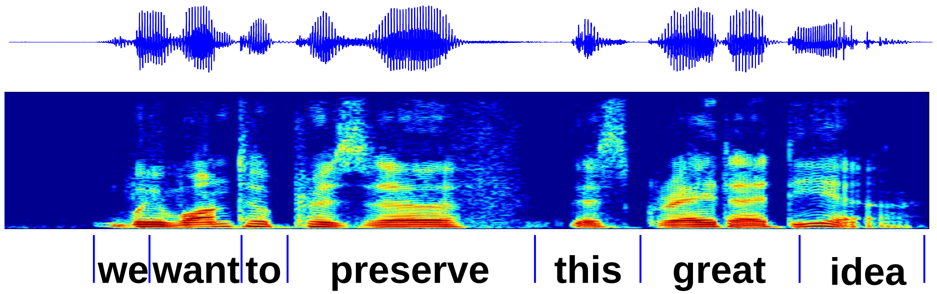
Hermann Ney
(joint work with P. Doetsch, P. Voigtlaender et al.)

Human Language Technology and Pattern Recognition
RWTH Aachen University, Aachen, Germany

IEEE Distinguished Lecturer 2016/17

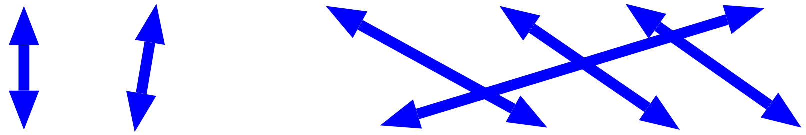
Sequence-to-Sequence Conversion and Recognition: Human Language Technology (HLT)

Automatic Speech Recognition (ASR)



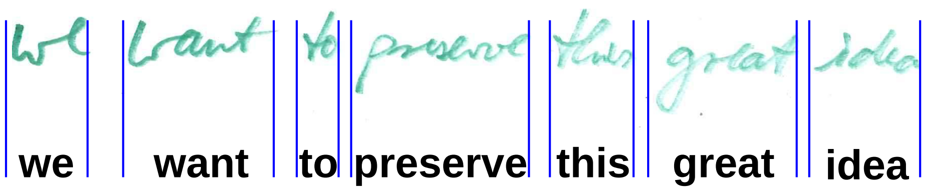
Statistical Machine Translation (SMT)

wir wollen diese große Idee erhalten



we want to preserve this great idea

Handwriting Recognition (HWR) (Text Image Recognition)



tasks:
- speech recognition
- handwriting recognition

- machine translation
  (+ sign language processing)

Sequence-to-Sequence Conversion and Recognition
Speech and Language

characteristic properties:

- **well-defined 'classification' tasks:**
  – due to 5000-year history of (written!) language
  – well-defined goal: letters or words (= full forms) of the language
- **easy task for humans (in native language!)**
- **hard task for computers**
  (as the last 50 years have shown!)

unifying view:

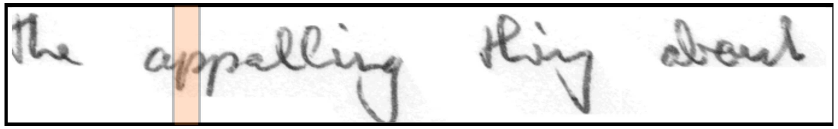
- **formal task: input sequence $\rightarrow$ output sequence**
- **output sequence: sequence of words/letters in a natural language**
- **models of context and dependencies:**
  – within input and output sequences
  – across input and output sequence

RWTH Large-Scale Joint Projects (1993-2016)

- **VERBMOBIL 1993-2000: funded by German BMBF**
  toy task (8000-word vocabulary): recognition and translation for appointment scheduling
- **TC-STAR 2004-2007: funded by EU**
  – real-life task: first research system for speech translation (EU parliament)
  – partners: KIT Karlsruhe, FBK Trento, LIMSI Paris, UPC Barcelona, IBM-US Research, ...
- **GALE 2005-2011: funded by US DARPA**
  emphasis on Chinese and Arabic speech and text
- **BOLT 2011-2015: funded by US DARPA**
  emphasis on colloquial text for Arabic and Chinese
- **QUAERO 2008-2013: funded by OSEO France (CNRS, INRIA, ...)**
  European languages, more colloquial speech, handwriting
- **EU projects 2012-2014: EU-Bridge, TransLectures**
  emphasis on recognition and translation of lectures (academic, TED, ...)
- **BABEL 2012-2016: funded by US IARPA**
  speech recognition for low-resource languages (and noisy audio!)

Similarities: Recognition of Speech and Handwriting

define sequence of vertical windows over horizontal axis:



result: one-dimensional approximation to handwriting recognition

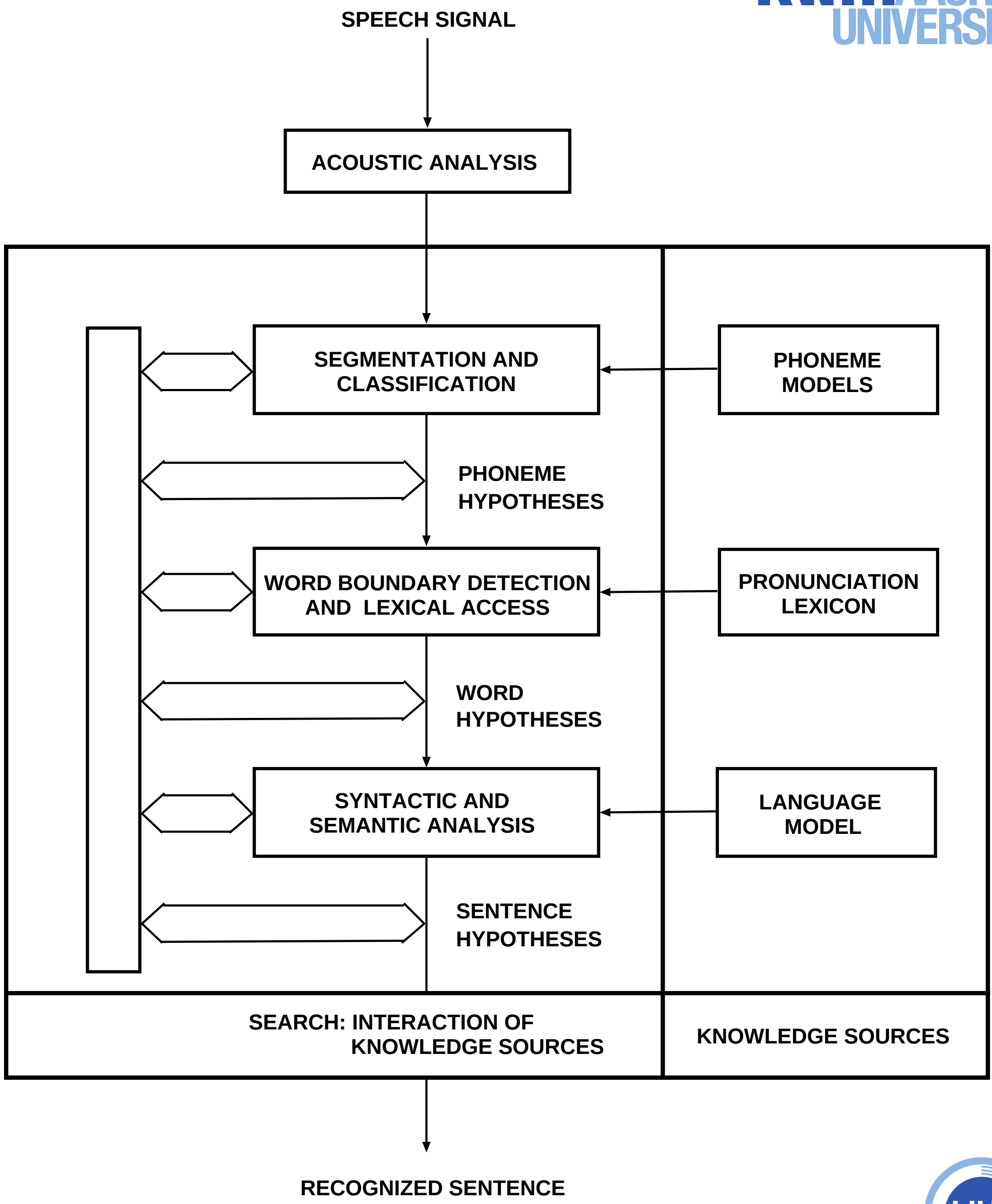
comparison: speech vs. handwriting (text image):

- **sequence of observation vectors:**
  – speech: signal segments, spectral analysis or PCA,...
  – handwriting: geometric features, PCA, pixels, ...
- **models of sounds/characters:**
  how to convert the observation vectors into hypotheses about sounds/characters?
- **lexical model: how to convert the sequence of sounds/characters into hypotheses about words?**
  – speech: pronunciation lexicon along with an orthographic dictionary
  – handwriting: only orthographic dictionary
- **language model: syntax and semantics**
  how to convert the sequence of words into hypotheses about "good" sentences?

ASR: what is the problem?
- **ambiguities at all levels**
- **interdependencies of decisions**

approach [CMU and IBM 1975]:
- **score hypotheses**
- **probabilistic framework**
- **statistical decision theory and Bayes decision rule**

important consequence:
the sequence context implies a combinatorial search problem:
- **interpret a 10-ms vector**
- **as a part of a sound/character**
- **as a part of a word**
- **as a part of a sentence**



Bayes Decision Rule Revisited

- closed world: consider a large, but finite set of (observation, label) pairs:

$$(X_r, W_r),\ r = 1, ..., R$$

- decision rule: for each observation sequence X,
  we want to guess or generate the label sequence W:

$$X \rightarrow \hat{W}(X) = ?$$

  complications: the same sequence X in the given set can have different sequences W;
  a perfect guess cannot be guaranteed!

- therefore: define performance measure or loss function (e. g. edit or Levenshtein distance) between correct output sequence W and hypothesized output sequence $\tilde{W}$:

$$L[W, \tilde{W}]$$

- for an observation X, what is the expected loss of the decision rule $X \rightarrow \hat{W}(X)$:
  answer: $\sum_W pr(W|X) \cdot L[W, \hat{W}(X)]$

  by using the posterior distribution derived from the joint empirical distribution:

$$pr(W, X) = 1/R \cdot \sum_r \delta(W, W_r) \cdot \delta(X, X_r)$$

- optimum performance: Bayes decision rule minimizes the expected loss:

$$X \rightarrow \hat{W}(X) \ := \ \arg\min_{\tilde{W}} \left\{ \sum_W pr(W|X) \cdot L[W, \tilde{W}] \right\}$$

optimum performance: Bayes decision rule minimizes the expected loss:

$$X \rightarrow \hat{W}(X) \;:=\; \arg\min_{\tilde{W}} \Big\{ \sum_{W} pr(W|X) \cdot L[W, \tilde{W}] \Big\}$$

Under these two conditions:

$$\begin{aligned} L[W, \tilde{W}] : &\quad \text{satisfies triangle inequality} \\ \max_{W} \{pr(W|X)\} &> 0.5 \end{aligned}$$

we have the MAP rule (MAP = maximum-a-posteriori) [Schlüter & Nussbaum$^+$ 12]:

$$X \rightarrow \hat{W}(X) \;:=\; \arg\max_{W} \Big\{ pr(W|X) \Big\}$$

Since [Bahl & Jelinek$^+$ 83], this simpified Bayes decision rule is widely used
for speech recognition, handwriting recognition, machine translation, ...

from closed world of finite sample, switch to arbitrary pairs of (observation, label) sequences:
introduce models of distributions $p_\vartheta(W|X)$ with free parameters ϑ

Modelling Approaches: Generative, Discriminative, Log-Linear...

For the unknown distribution in Bayes decision rule,
assume suitable model distributions $p_\vartheta(W)$ and $p_\vartheta(X|W)$ with free parameters ϑ:

$$p_\vartheta(W|X) = \frac{p_\vartheta(W) \cdot p_\vartheta(X|W)}{\sum_{\tilde{W}} p_\vartheta(\tilde{W}) \cdot p_\vartheta(X|\tilde{W})} \qquad \text{or} \qquad p_\vartheta(W|X) = \frac{q_\vartheta^{\lambda}(W) \cdot q_\vartheta^{1-\lambda}(W|X)}{\sum_{\tilde{W}} q_\vartheta^{\lambda}(\tilde{W}) \cdot q_\vartheta^{1-\lambda}(\tilde{W}|X)}$$

generalization: log-linear combination of models $q_\vartheta(W)$ and $q_\vartheta(W|X)$

important property: decomposition into two separate models:
– language model $p_\vartheta(W)$: depends on text data only!
  advantage: huge amounts available, no annotation needed!
– observation model (speech, text image) $p_\vartheta(X|W)$:
  depends on (observation, label) pairs!

learning from data:

- models $p_\vartheta(W)$ and $p_\vartheta(X|W)$ with unknown parameters ϑ
- training data: set of (observation, label) pairs $(X_r, W_r), r = 1, ..., R$

Traditional Training Criteria for ASR

- generative model (joint probability): maximum likelihood (along with EM/Viterbi algorithm for Hidden Markov models):

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

- sentence posterior probability (MMI = maximum mutual information) [Bahl & Brown$^+$ 86],[1991 Normandin]:

$$F(\vartheta) = \sum_r \log p_\vartheta(W_r|X_r)$$

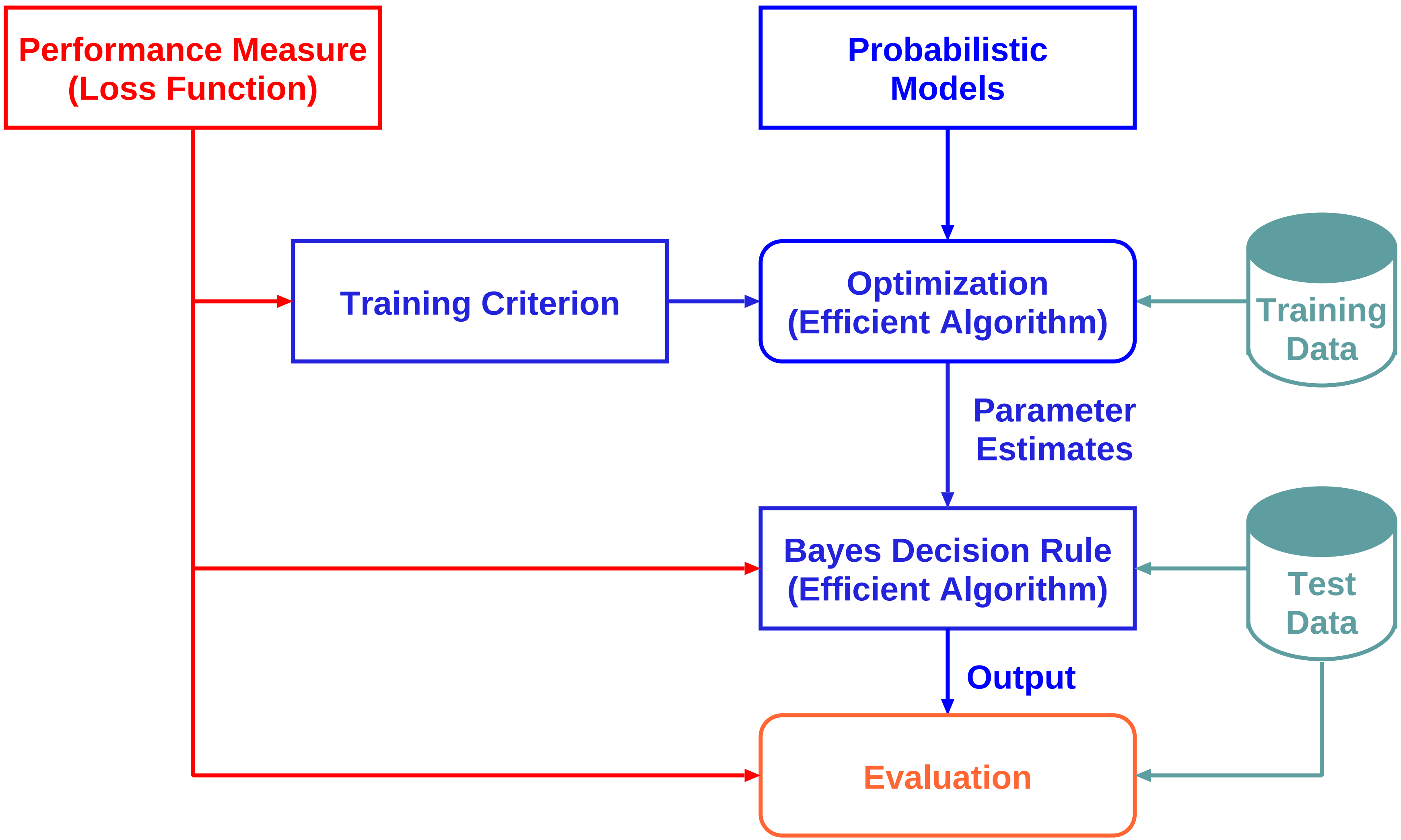
- [Povey & Woodland 02] MWE/MPE: minimum word/phoneme error (= expected 'accuracy'):

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

with the accuracy $A(W, W_r)$ of hypothesis W for correct sentence W_r:
:= sequence discriminative training

remarks:
– complex optimization problem: sum over all sentences in denominator
– approximation: word lattice, many shortcuts, ...
– experiments: relative improvement by 5-10% over maximum likelihood

Sequence-to-Sequence Recognition: Statistical Approach to HLT Tasks



Statistical Approach and Machine Learning

four ingredients:

- performance measure: error measure (e.g. edit distance)
  we have to decide how to judge the quality of the system output
  (ASR + HWR: edit distance; SMT: edit distance + block movements)

- probabilistic models with suitable structures:
  to capture the dependencies within and between input and output sequences
  – elementary observations: Gaussian mixtures, log-linear models, support vector machines (SVM), multi-layer perceptron (MLP), ...
  – sequences: n-gram Markov chains, CRF, Hidden Markov models (HMM), recurrent neural nets (RNN), LSTM RNN, CTC, ...

- training criterion:
  to learn the free model parameters from examples
  – ideally should be linked to performance criterion
  – typically result in complex mathematical optimization (efficient algorithms!)
  – extreme situation: number of free parameters vs. observations

- Bayes decision rule:
  to generate the output word sequence
  – combinatorial problem (efficient algorithms)
  – should exploit structure of models
  examples: dynamic programming and beam search, A* and heuristic search, ...
  (public toolkits for ASR/HWR: RWTH, Kaldi, ...)

Bayes Decision Rule: Principal Issues

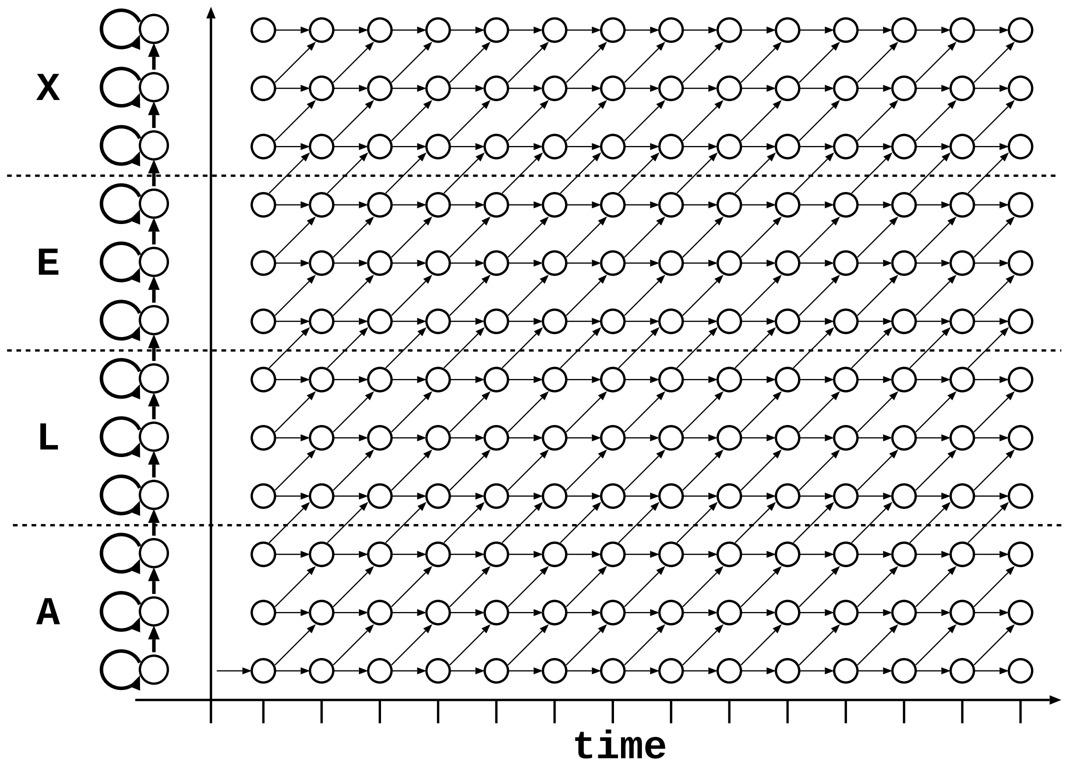
ongoing work at RWTH:

- **form of Bayes decision rule:**
  MAP rule vs. exact rule: justification?
- **mismatch conditions:**
  – optimality of Bayes rule: holds for TRUE distribution
  – what about a model distribution learned from data? optimality?
- **relation between performance (classification error) and training criteria**
- **peformance at various levels:**
  frames, phonemes, words, sentences
  – suitable training criteria at each level
  – interaction betweeen these levels
  (end-to-end training)

some results by RWTH team:
[Ney 03, Schlüter & Nussbaum^{+} 12, Schlüter & Nussbaum-Thom^{+} 13, Beck & Schlüter^{+} 15]

Acoustic Modelling: HMM and ANN

– why HMM? mechanism for time alignment (or dynamic time warping)
– critical bottleneck: emission probability model requires density estimation!
– hybrid approach: replace HMM emission probabilty by label posterior probabilities, i. e. by ANN output after suitable re-scaling

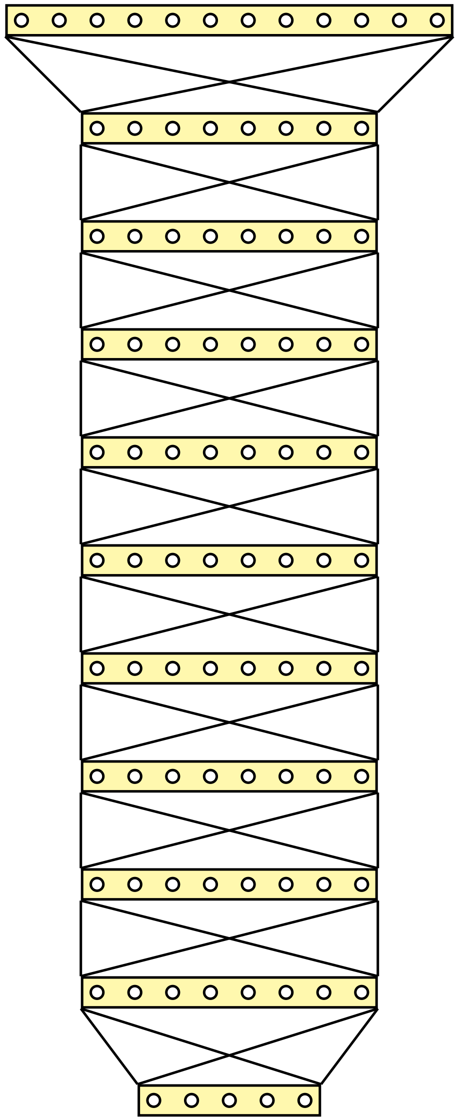


History: ANN in Acoustic Modelling

- **1988 [Waibel & Hanazawa$^+$ 88]:**
  phoneme recognition using time-delay neural networks (and CNNs!)
- **1989 [Bridle 89]:**
  softmax operation for probability normalization in output layer
- **1990 [Bourlard & Wellekens 90]:**
  - **for squared error criterion, ANN outputs can be interpreted as class posterior probabilities (rediscovered: Patterson & Womack 1966)**
  - **they advocated the use of MLP outputs to replace the emission probabilities in HMMs**
- **1993 [Haffner 93]: sum over label-sequence posterior probabilities in hybrid HMMs**
- **1994 [Robinson 94]: recurrent neural network**
  - **competitive results on WSJ task**
  - **his work remained a singularity in ASR**
- **until 2011: for speech, ANNs were never really better than Gaussian mixture models**

first clear improvements over the state of the art:
– 2008 handwriting: Graves using LSTM-RNN and CTC
– 2011 speech: Hinton & Li Deng using deep FF MLP and hybrid HMM
– more ...

What is Different Now after 25 Years?



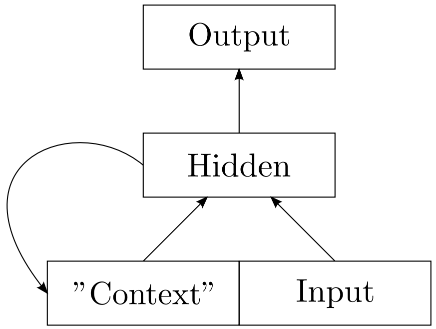
important property:
ANN outputs are probability estimates

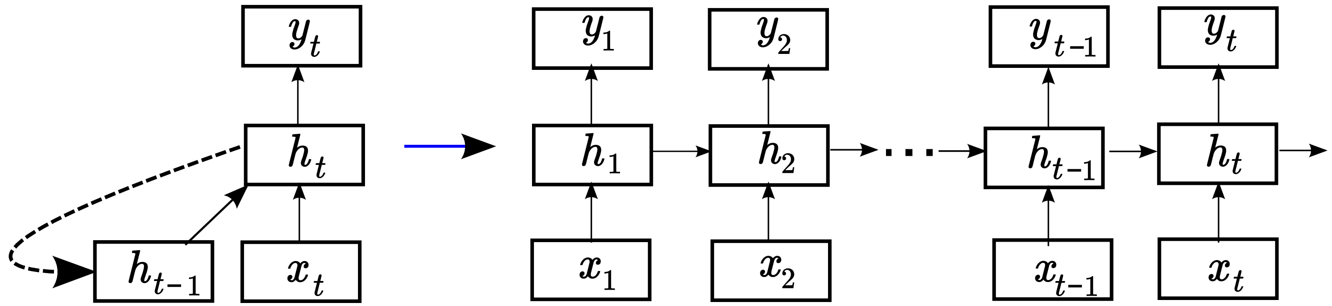
today: huge improvements by ANN:
– image object recognition
– speech and handwriting recognition
– machine translation

comparison for ASR: today vs. 1989-1994:

- number of hidden layers:
  10 (or more) rather than 2-3
- optimization strategy:
  practical experience and heuristics,
  e.g. layer-by-layer pretraining
- computation power: much higher
- specifically for ASR:
  number of output nodes (phonetic labels):
  5000 rather than 50

principle for sequence processing over time $t = 1, ..., T$:
– introduce a memory (or context) component to keep track of history
– result: there are two types of input: memory h_{t-1} and observation x_t

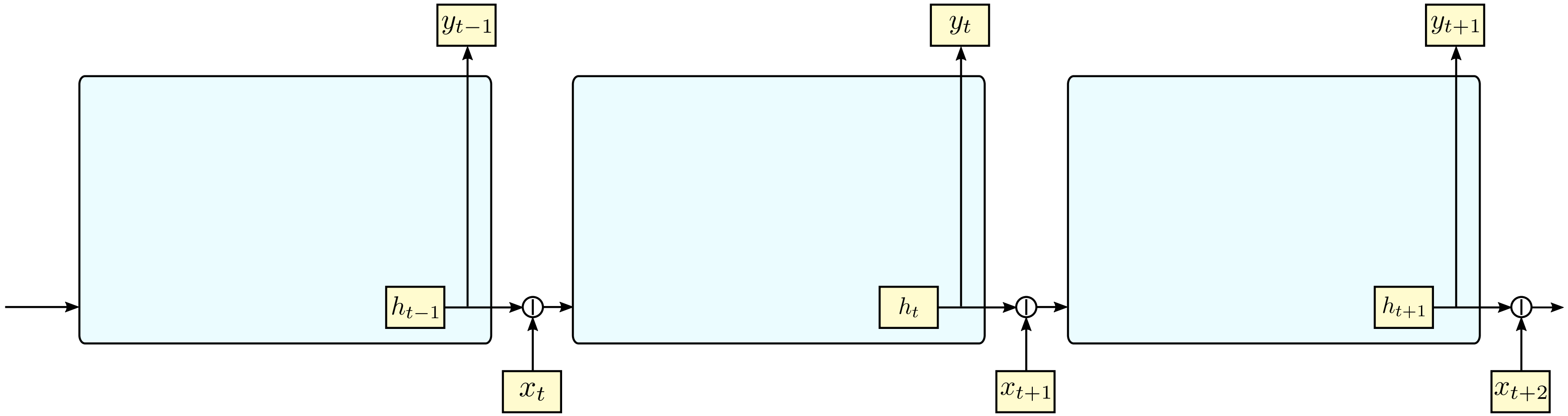


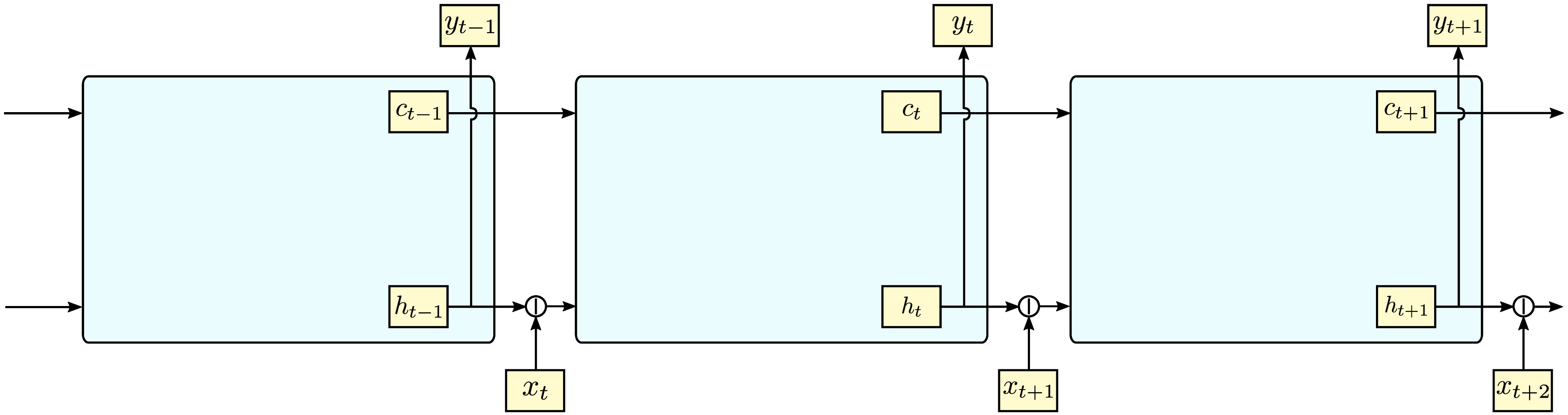


extensions:
– bidirectional variant [Schuster & Paliwal 1997]
– feedback of output labels
– long short-term memory [Hochreiter & Schmidhuber 97; Gers & Schraudolph$^+$ 02]
– stacking of recurrent-hidden layers

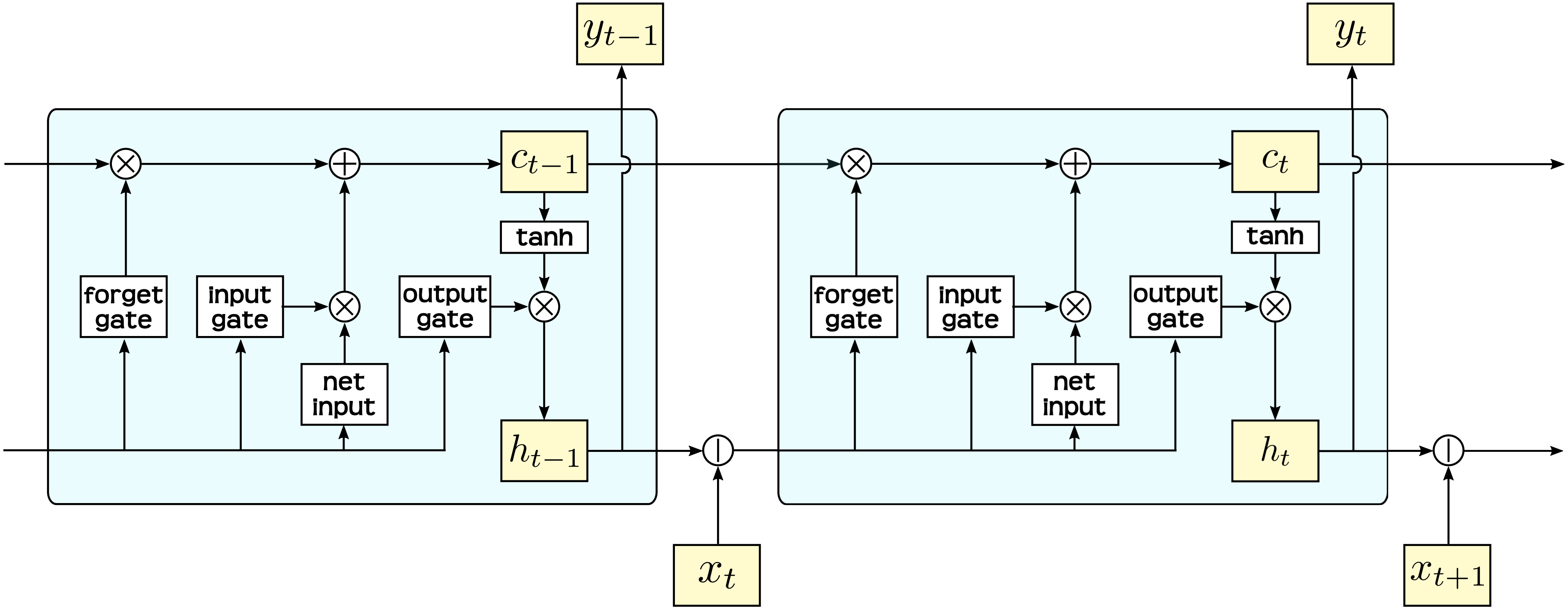
Recurrent Neural Network (RNN): Extension towards Long Short-Term Memory

add a memory cell vector c_t to hidden state vector h_t:





Recurrent Neural Network:
Details of Long Short-Term Memory



ingredients:
– **separate memory vector c_t in addition to h_t**
– **use of gates to control information flow**
– **(additional) effect: make backpropagation more robust**

hybrid approach:

replace emission probability of an hidden Markov model by ANN ouput

three types of emission models in HMMs:
– GMM: Gaussian mixture model
– MLP: deep multi-layer perceptron
– LSTM RNN: recurrent neural network with long short-term memory

experimental results for QUAERO English 2011:

approach	layers	WER[%]
conventional: best GMM	–	30.2
hybrid: best MLP	9	20.3
hybrid: best LSTM RNN	6	17.5

remarks:

- comparative evaluations in QUAERO 2011:
  competitive results with LIMSI Paris and KIT Karlsruhe
- best improvement over Gaussian mixture models
  by 40% relative using an LSTM RNN

History:

- **1989 [Nakamura & Shikano 89]:**
  English word category prediction based on neural networks.
- **1993 [Castano & Vidal$^+$ 93]:**
  Inference of stochastic regular languages through simple recurrent networks
- **2000 [Bengio & Ducharme$^+$ 00]:**
  A neural probabilistic language model
- **2007 [Schwenk 07]: Continuous space language models**
  2007 [Schwenk & Costa-jussa$^+$ 07]: Smooth bilingual n-gram translation (!)
- **2010 [Mikolov & Karafiat$^+$ 10]:**
  Recurrent neural network based language model
- **2012 RWTH Aachen [Sundermeyer & Schlüter$^+$ 12]:**
  LSTM recurrent neural networks for language modeling

today: ANNs in language (and translation!) show competitive results.

ANNs in Language Modelling

goal of language modelling: compute the prior $p_\vartheta(w_1^N)$ of a word sequence w_1^N
– how plausible is this word sequence w_1^N (independently of observation X!) ?
– measure of language model quality: perplexity PP_ϑ, i. e. effective vocabulary size

$$\log PP_\vartheta \;=\; -1/N \cdot \sum_{n=1}^{N} \log\, p_\vartheta(w_n|w_0^{n-1})$$

perplexity PP on test data:

approach	PP
baseline: count model	163.7
10-gram MLP	136.5
RNN	125.2
LSTM RNN	107.8
10-gram MLP with 2 layers	130.9
LSTM RNN with 2 layers	100.5

results on QUAERO English (like before):
– vocabulary size: 150k words
– training text: 50M words
– test set: 39k words

important result: improvement of PP by 40%

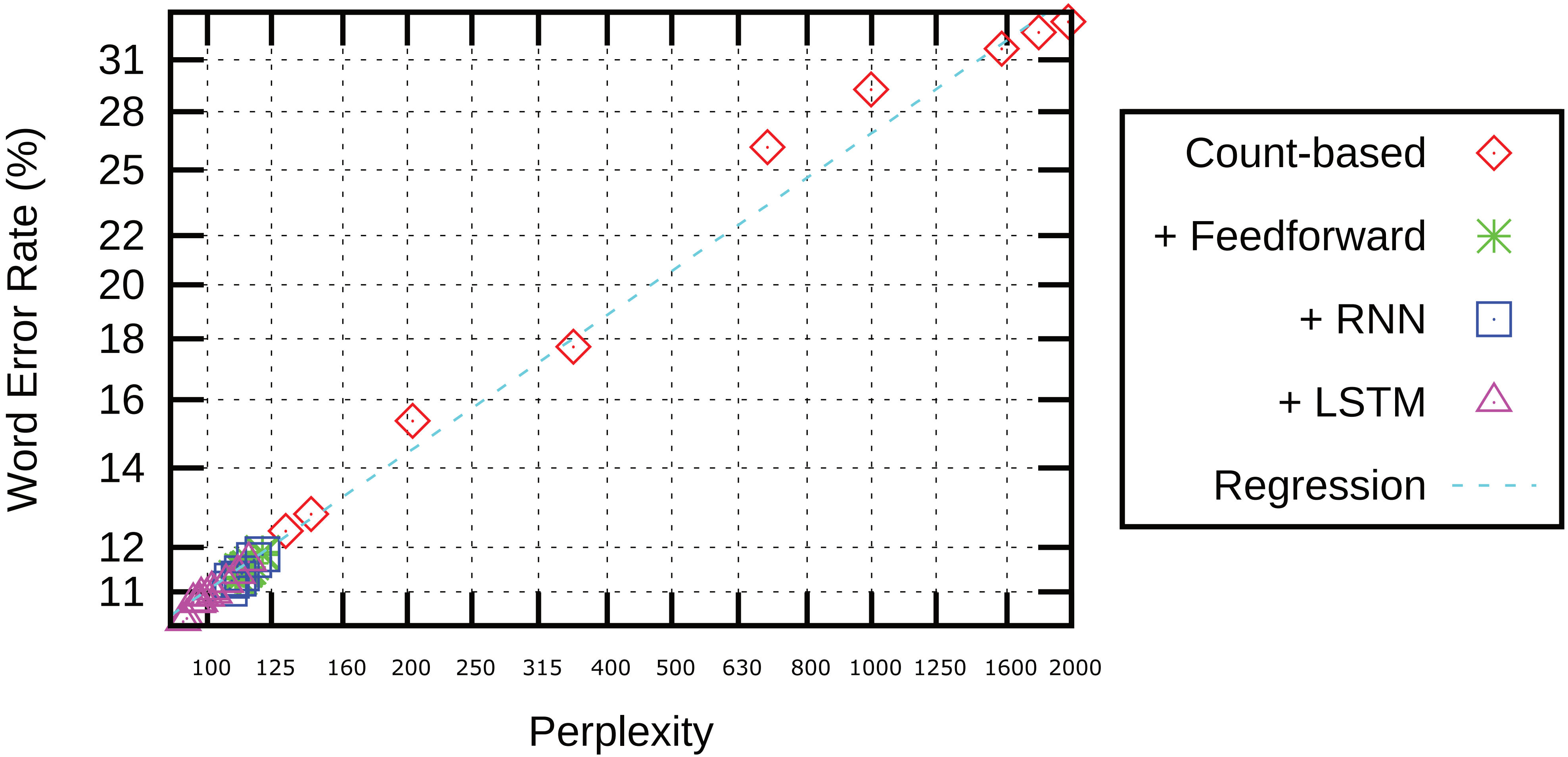
Interpolated Language Models: Perplexity and WER

- linear interpolation of TWO models: count model + ANN model
- recognition experiments:
  due to unlimited history, RNN language models require re-design of ASR search
- perplexity and word error rate on test data:

Models	PP	WER[%]
count model	131.2	12.4
+ 10-gram MLP	112.5	11.5
+ Recurrent NN	108.1	11.1
+ LSTM RNN	96.7	10.8
+ 10-gram MLP with 2 layers	110.2	11.3
+ LSTM RNN with 2 layers	92.0	10.4

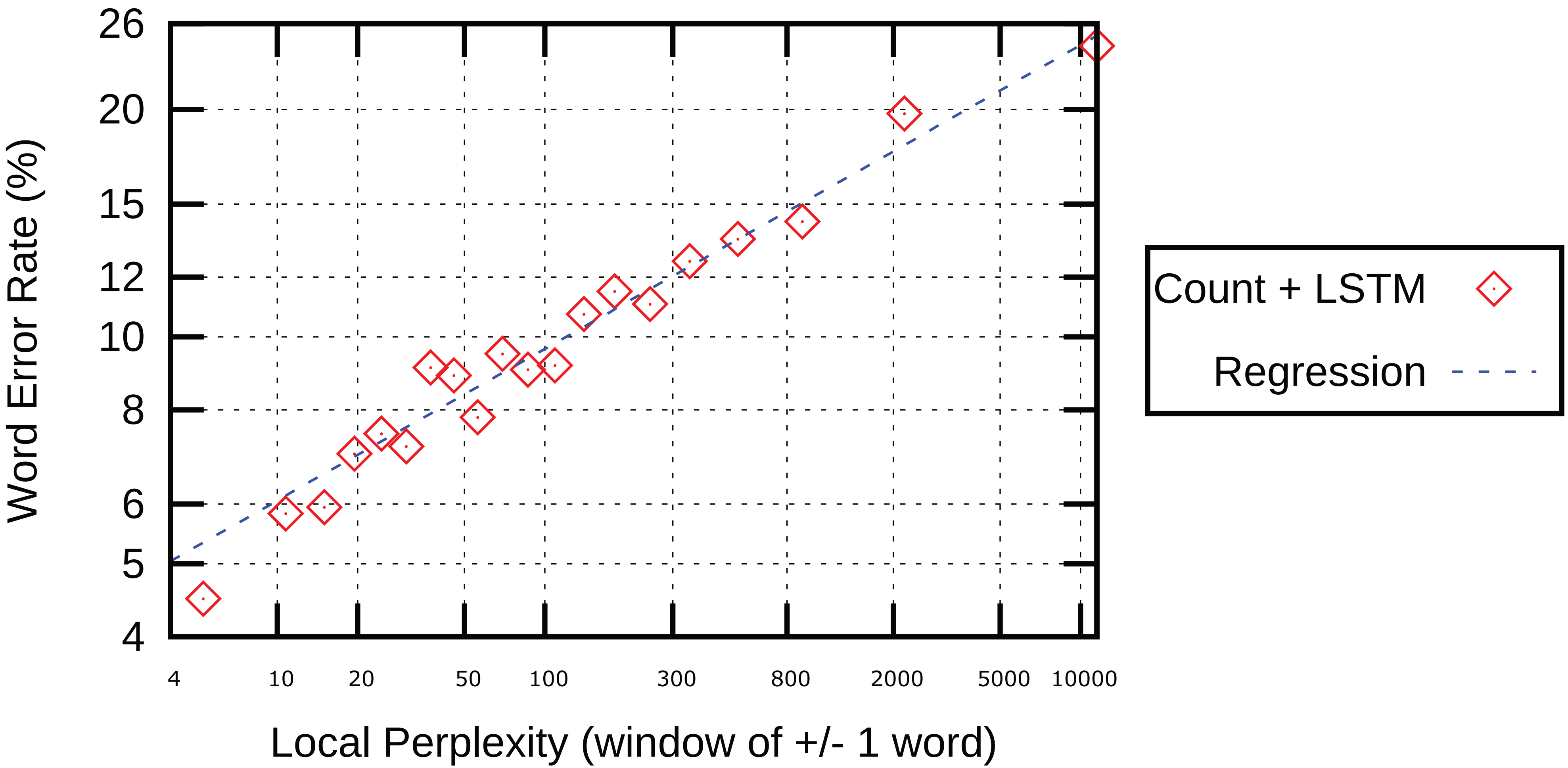
- experimental result:
  - significant improvements by ANN language models
  - best improvement in perplexity: 30% reduction (from 131 to 92)
  - empirical observation:
    power law between perplexity and WER (cube to square root)
    [Klakow & Peters 02]

empirical power law: $WER = \alpha \cdot PP^{\beta}$



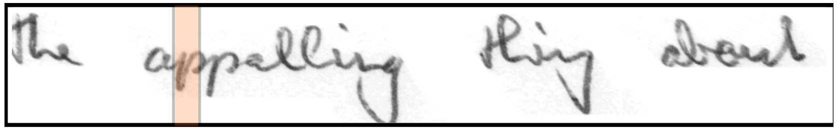
Word Error Rate vs. Local Perplexity
(3-word window, 20 bins)

empirical power law: $WER = \alpha \cdot PP^{\beta}$



RWTH AACHEN UNIVERSITY

- consider sequence of vertical windows over horizontal axis (maybe after normalization and preprocessing):



  – approximate two-dimensional problem by one-dimensional problem
  – ... looks like a problem of speech recognition
  – so far most successful

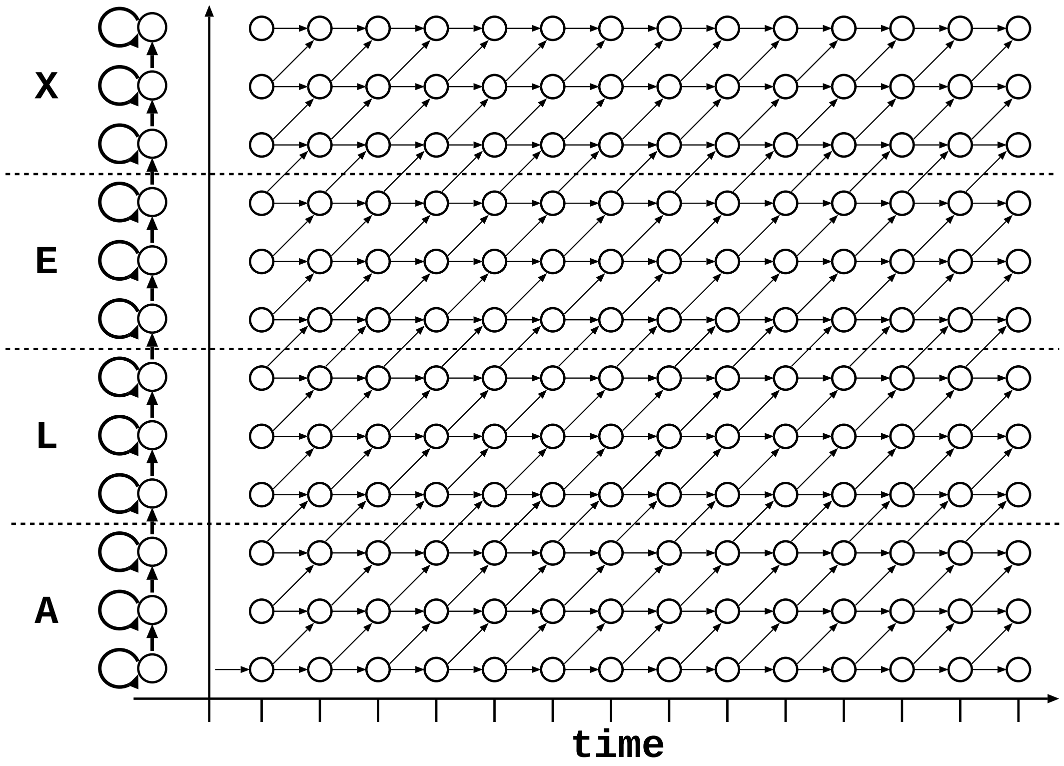
- history: dynamic time warping/HMM for character recognition
  – 1992 Pieraccini & Levin; 1993 Agazzi & Kuo
  – 1997 Kaltenmeier et al.
  – 1998 BBN Byblos: Schwartz et al. [Lu & Bazzi^{+} 98]

- history (no langauage model):
  interdependence of segmentations, alignment and decisions:
  – 1968 Kovalevsky for characer recognition (*sequential optimization*)
  – 1971 Vintsyuk for speech recognition
  Work was overlooked in Europe and USA.

Hybrid HMM Revisited

for each class symbol (sound or character),
define HMM:
– sequence of states (e g. three) with
  label state posterior prob.
– set of transitions
  with transition prob.

main purpose: time alignment



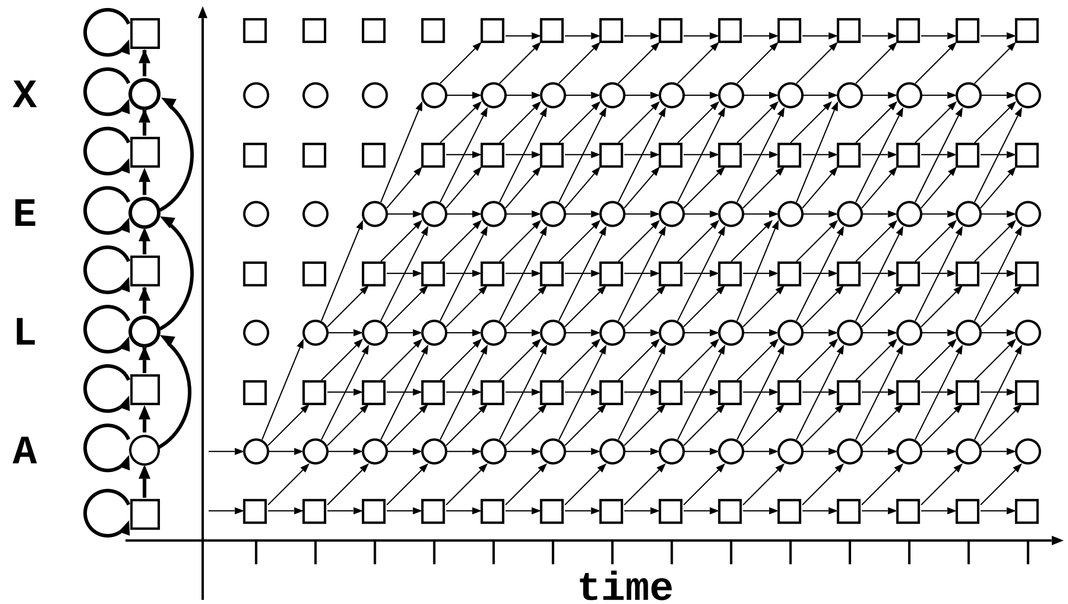
training criterion for a single (!) sequence of ovbservations $x_1^T := x_1...x_t...x_T$
with state label sequences $s_1^T := s_1...s_t...s_T$:

$$\max_{...} \left\{ \log \sum_{s_1^T} \prod_t \Big(p(s_t|s_{t-1}) \cdot p_t(s_t|x_1^T) \,/\, p(s_t) \Big) \right\}$$

simplification: best path (Viterbi) in lieu of exact sum [Haffner 93]

CTC: Connectionist Temporal Classification [Graves & Fernandez^{+} 06]

simplify HMM structure:
- **two states only**
- **tie second state across all symbols (*white space*)**
- **drop transition probabilities**
- **drop prior probabilities**



resulting training criterion for a single (!) sequence x_1^T with state label sequence s_1^T:

$$\max_{\dots} \left\{ \log \sum_{s_1^T} \prod_t p_t(s_t|x_1^T) \right\}$$

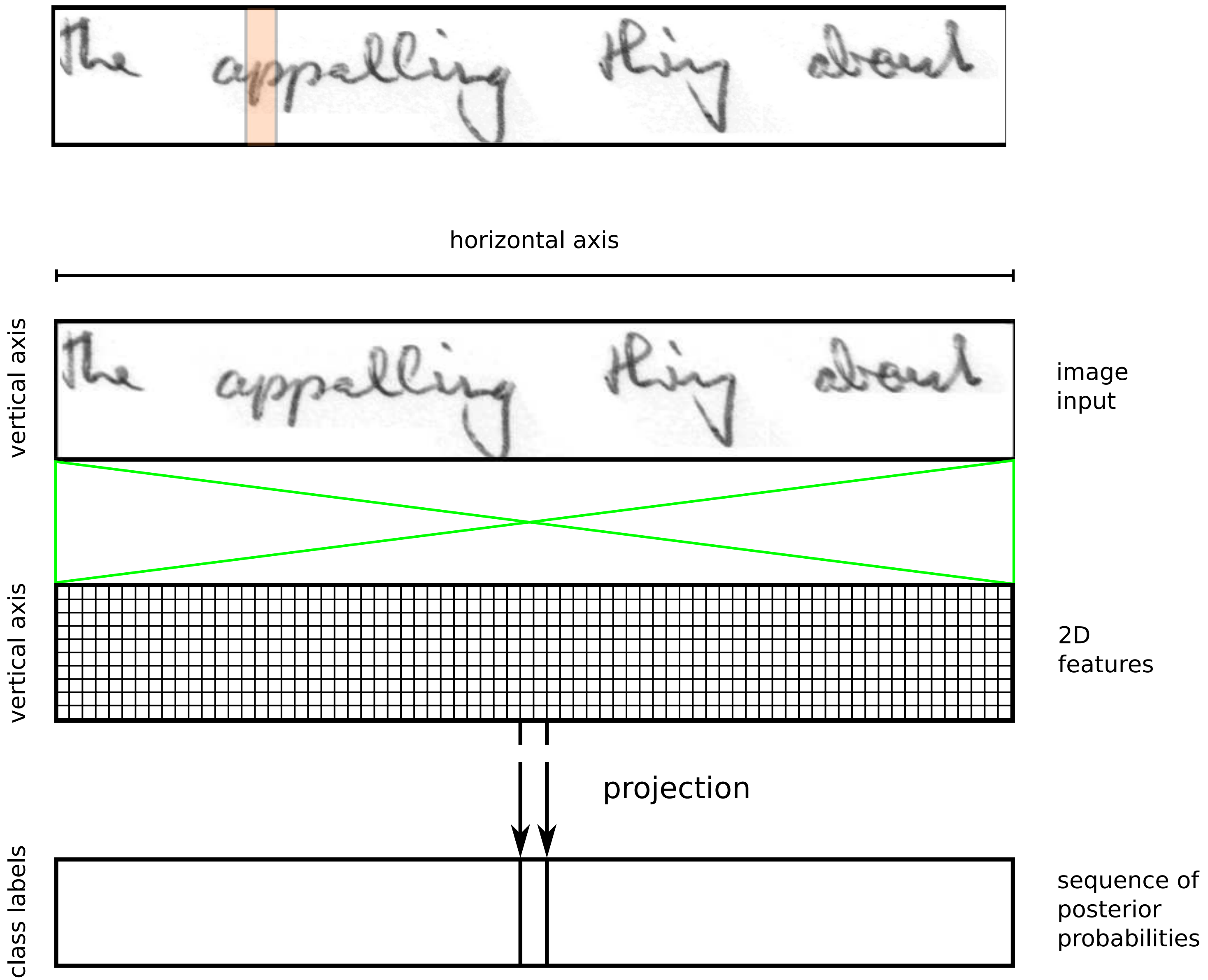
comparison of CTC with hybrid HMM and full sum:
- **effect of many simplifications: unclear ?**
- **is it the criterion or the optimization strategy ?**
- **shortcoming: no language model $\rightarrow$ weaker than seq.discr. training**

LSTM RNN: From 1D to 2D Processing

more information at this ICFHR:

- **paper with more details (Monday, oral session, 14:20):**
  P. Voigtlaender, P. Doetsch et al.:
  Handwriting Recognition with Large Multidimensional LSTM RNNs.
- **competition organized by J. A. Sánchez et al. (Wednesday, oral session, 17:30):**
  ***ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset*.**
  RWTH participated with excellent results.

LSTM RNN: From 1D to 2D Processing
[Graves 2008: Multidimensional RNN]



Database IAM

- IAM handwriting corpus [Marti & Bunke^{+} 02]
- Lexicon: 50k words
- 3-gram language model
- 80 class labels: 78 characters + whitespace + blank

corpus	#paragr.	#lines	#run. words	#run. chars	OOV[%]
train	747	6,482	53.8k	219.7k	-
dev	116	976	8.7k	31.7k	3.94
eval	336	2,915	25.4k	96.6k	3.42

IAM Results: Closed Vocabulary (OOV: 3.9 % and 3.4 %)

System	Model and Training	#params	WER[%]		CER[%]	
			dev	eval	dev	eval
Gaussian Mixtures	Max.Lik.	108.9K	10.7	-	3.8	-
LSTM RNN: 4 layers	HMM: best path	20.7M	11.2	14.5	3.3	5.3
	+ seq.disc. training		10.6	13.5	3.2	5.1
	HMM: sum	20.7M	12.7	14.6	3.8	5.5
	CTC: sum		11.3	13.1	3.7	5.3
2D LSTM RNN: 5 layers	CTC: sum	2.6M	10.1	11.7	3.1	4.0
[Pham & Bluche^{+} 14] 2D LSTM RNN	CTC: sum	142.0K	11.2	13.6	3.7	5.1

observations:
– high performance: seq.disc. training
– significant improvements for 2D LSTM RNN

IAM Results: Open Vocabulary

from closed to open vocabulary:

extend word-based language model by character-based language model so that any character sequence can be recognized
[Kozielski & Mathysiak$^+$ 14] at ICFHR 2014

... requires extension of search strategy (decoder)

System	Model and Training	#params	WER[%]		CER[%]	
			dev	eval	dev	eval
LSTM RNN: 4 layers	HMM: best path	20.7M	8.6	12.1	2.8	4.9
	+ seq.disc. training		8.3	11.7	2.8	4.7
	HMM: sum	20.7M	?	?	?	?
	CTC: sum		8.6	11.1	3.0	4.7
2D LSTM RNN: 5 layers	CTC: sum	2.6M	7.1	9.3	2.4	3.5

observations:
– in general: significant improvement by open vocabulary
– overall ranking: like closed vocabulary

Database RIMES Text Lines

- RIMES handwriting corpus [Augustin & Brodin^{+} 06]
- Lexicon: 6.7k words
- 4-gram language model
- 98 class labels: 96 characters + whitespace + blank

corpus	#paragr.	#lines	#run. words	#run. chars	OOV[%]
train	1500	11,279	82.2k	452.7k	-
eval	100	778	5.6k	31.2k	4.2

Results on RIMES Text Lines
(closed vocabulary; OOV = 4.2%)

System	Model and Training	#params	eval WER[%]	eval CER[%]
Gaussian Mixtures	Max.Lik.	47.2K	15.7	5.5
LSTM RNN: 4 layers	HMM: best path	20.7M	11.4	4.1
	+ seq.disc. training		10.9	3.8
	HMM: sum	20.7M	? 15.3	? 7.8
	CTC: sum		11.1	4.1
2D LSTM RNN: 5 layers	CTC: sum	2.6M	9.4	2.9
[Pham & Bluche+ 14] 2D LSTM RNN	CTC: sum	142.0K	12.3	3.3

observations:

– high performance (1D case): seq.disc. training and CTC

– significant improvements for 2D approach

– high fluctuations for HMM/sum: reason unclear (?)

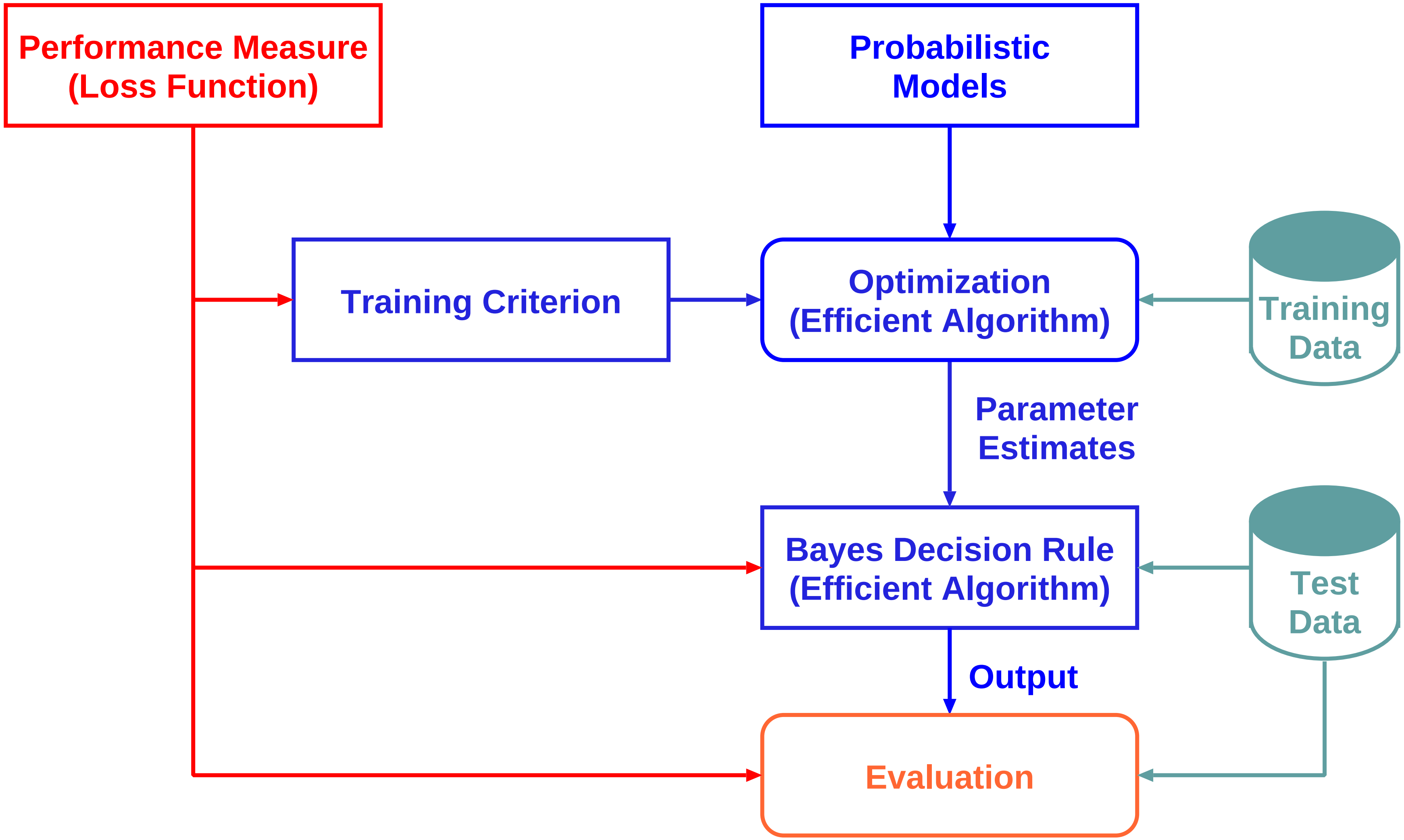
Sequence-to-Sequence Recognition: Statistical Approach and Machine Learning

- **four key ingredients:**
  - **choice of performance measure: errors at sequence, word, phoneme, frame level**
  - **probabilistic models at these levels and the interaction between these levels**
  - **training criterion along with an efficient optimization algorithm**
  - **Bayes decision rule along with an efficient search algorithm**
- **about recent work on ANNs (2011-16):**
  - **yes, ANNs result in significant improvements**
  - **ANNs provide one more type of probabilistic models**
- **shortcomings of present ANNs and challenges: too much trial and error**
  - **need of robust training and convergences**
  - **need of clear principles in designing ANN structures**

scientific challenges for the future of sequence-to-sequence recognition:

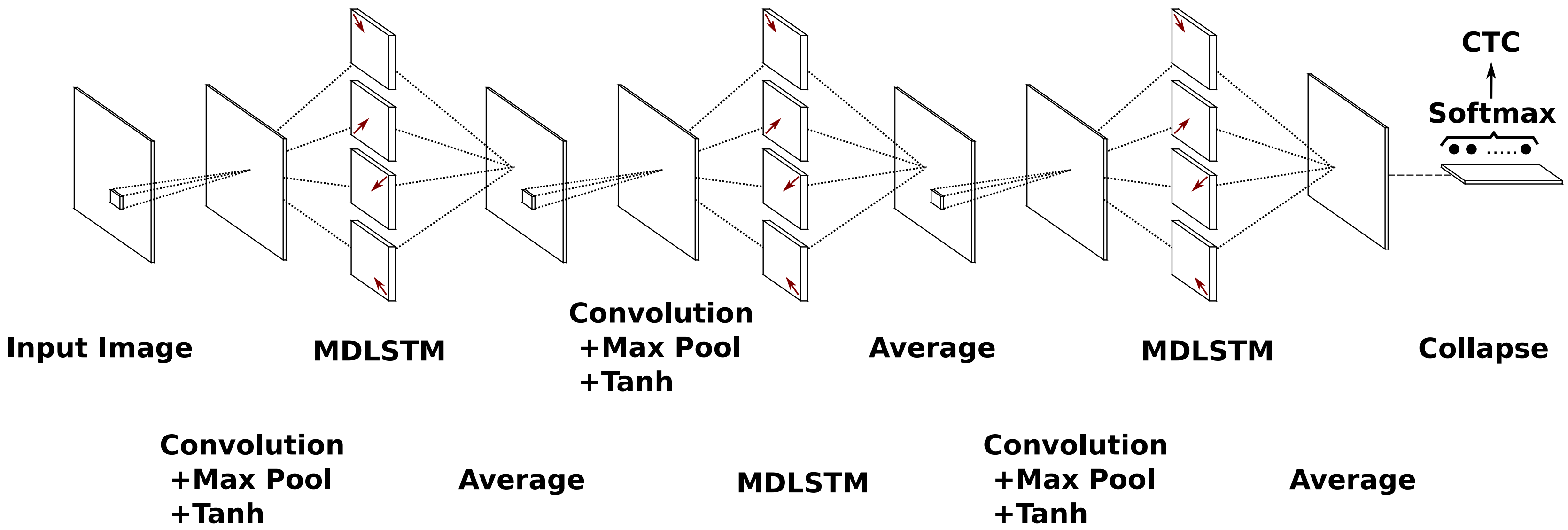
- **open lexicon: get away from closed lexicon and allow ANY sequence of characters**
- **unsupervised training:**
  e. g. ASR/HWR: observations data (without labels) + (very good) language model
- **alignment mechanism:**
  can attention-based mechanism replace first-order concepts (e.g. HMM)?

Sequence-to-Sequence Recognition: Statistical Approach to HLT Tasks

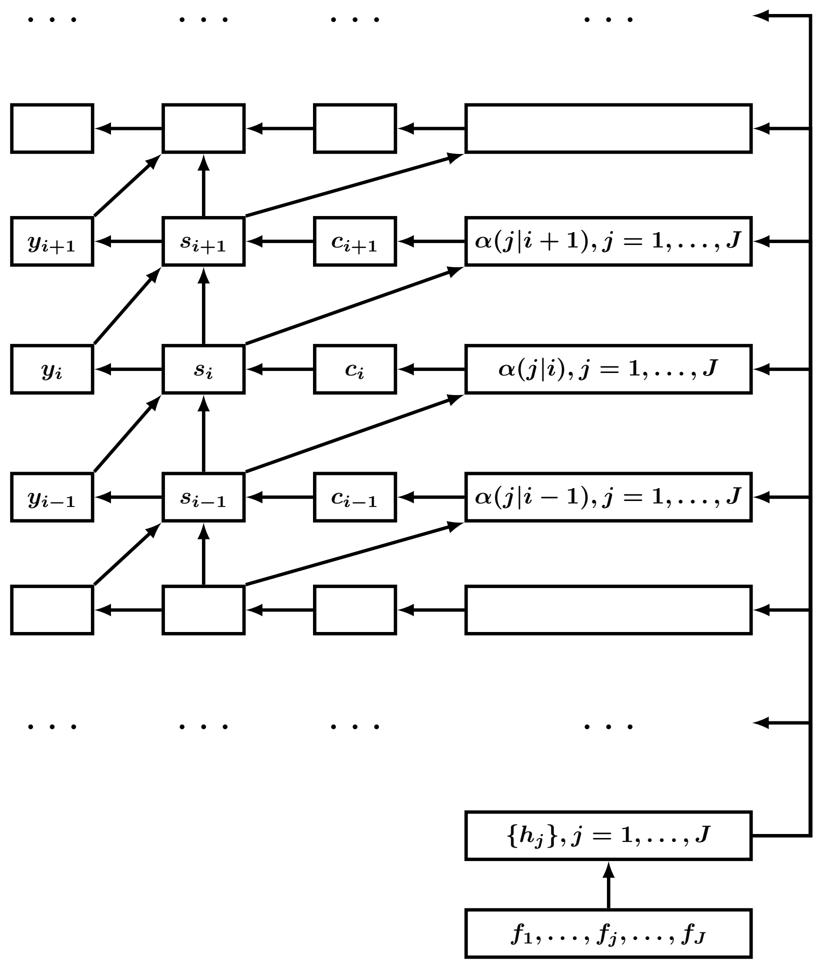


BACK-UP SLIDES
(Handwriting)

2D LSTM RNN: Architecture

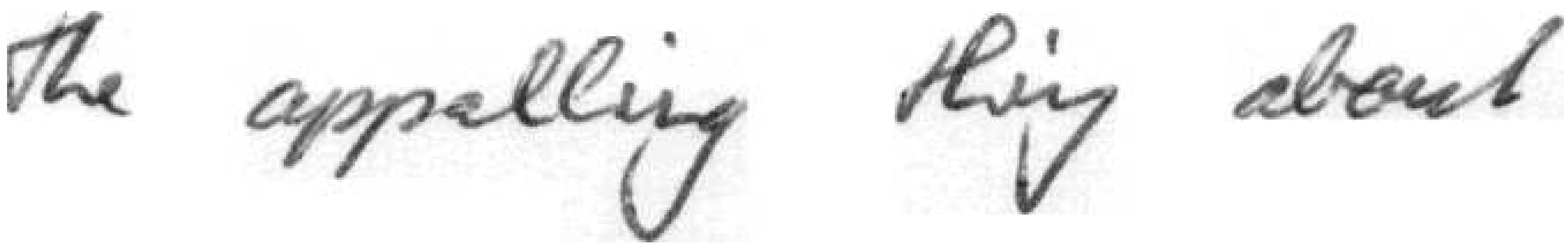


Attention-based NN MT
[Bahadanau et al. 2014]



Preprocessing: Deslanting

- **Reduce vertical distortions through shearing angle normalization**

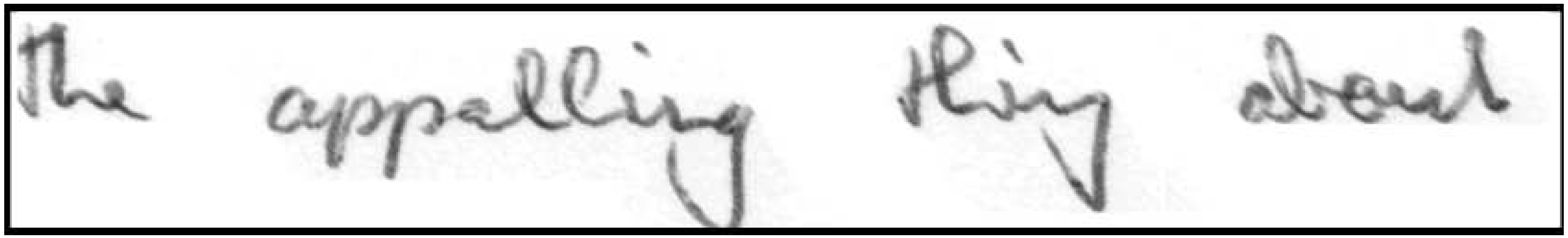


Preprocessing: Deslanting

- **Calculate vertical projection ρ for different shearing angles**
- **Choose angle with maximal score:**

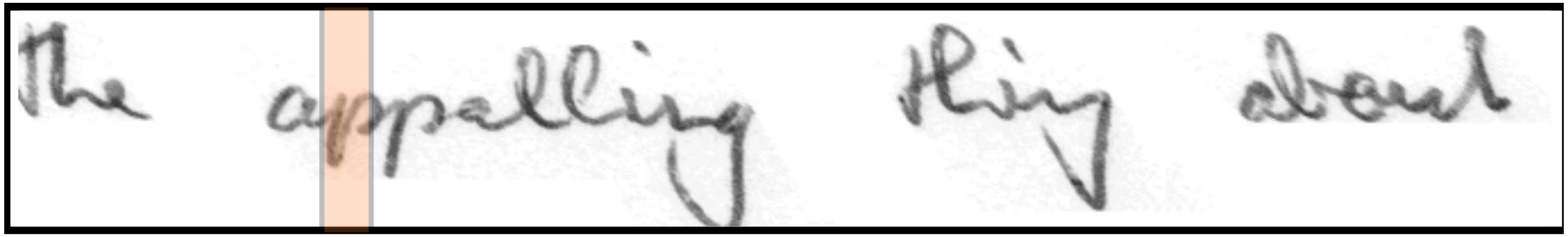
$$\chi(\rho) = \sum_{i=1}^{N-1} (\rho_i - \rho_{i+1})^2$$

Preprocessing: Deslanting



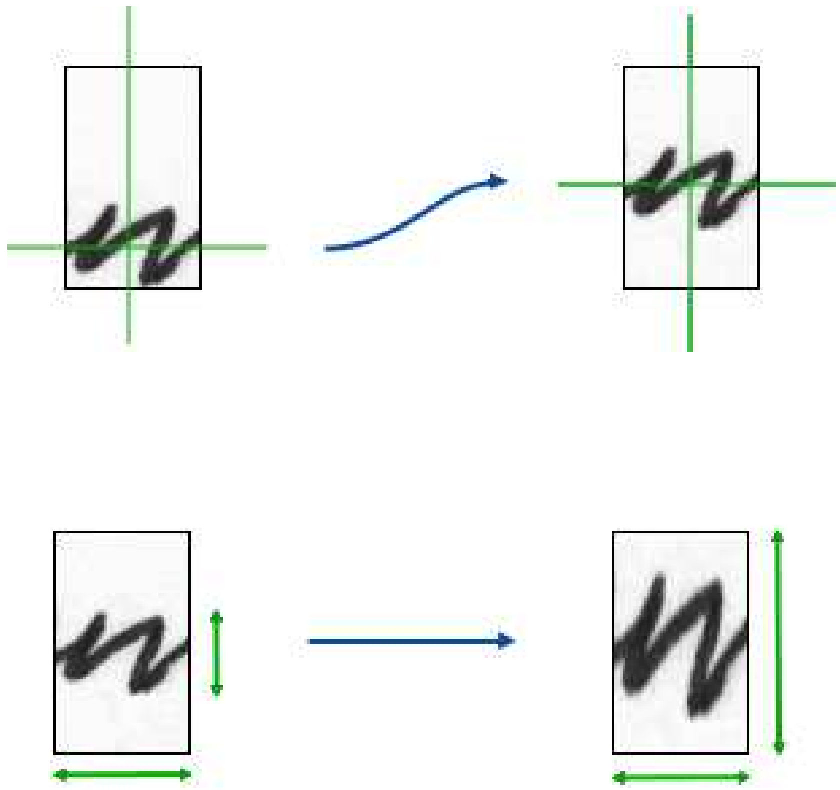
Feature extraction

- **Shift (overlapping) sliding window from left to right over the image**

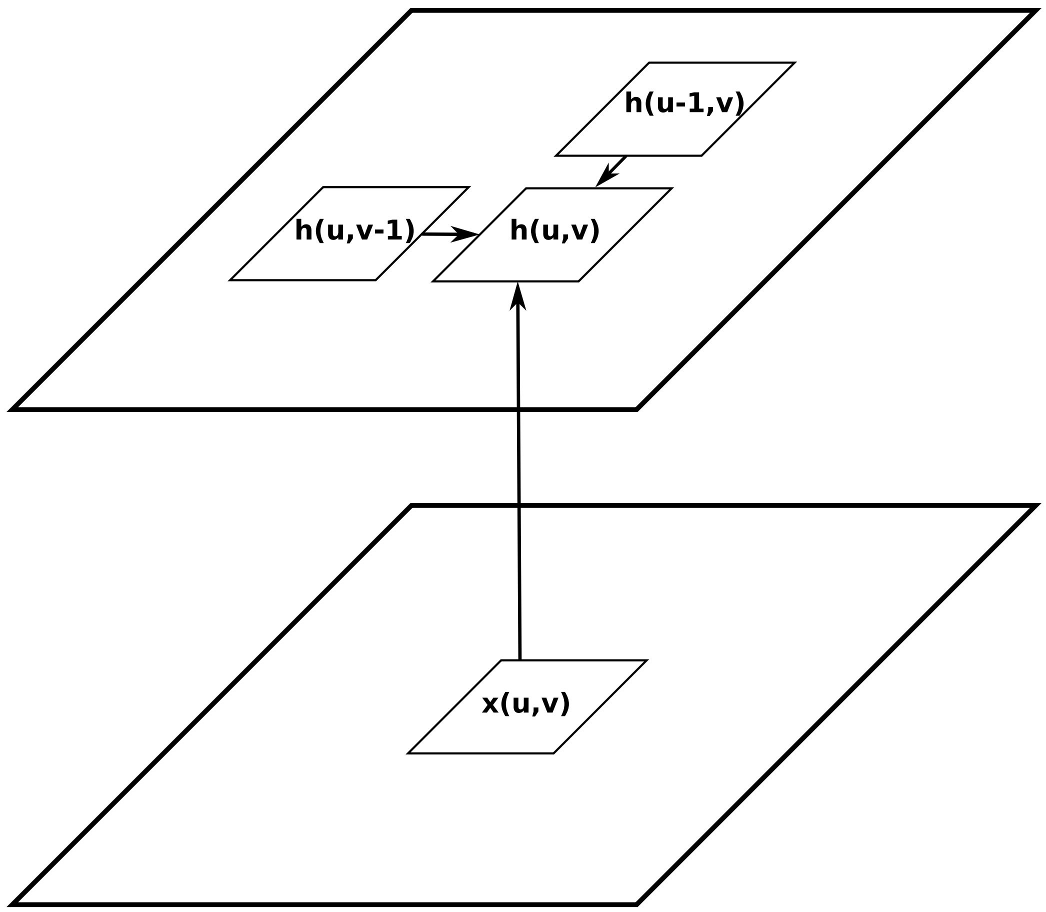


Window-based transformations

- **Normalize vertical position and scaling**

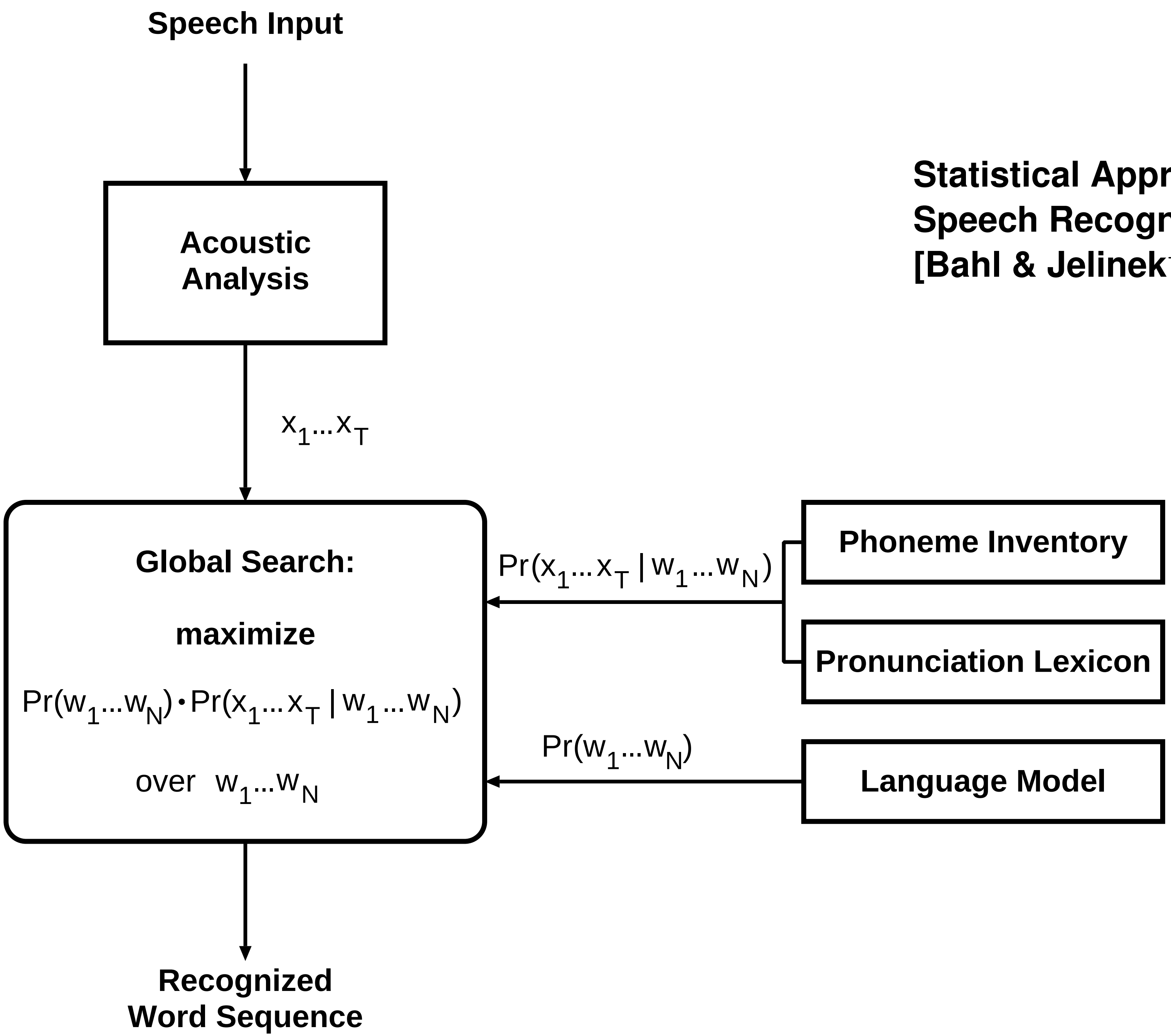


2D RNN



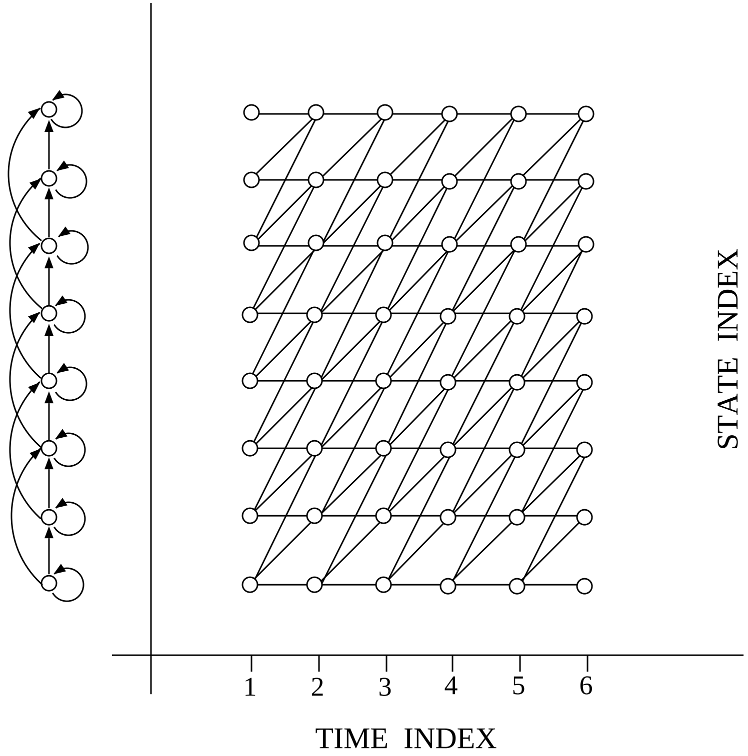
BACK-UP SLIDES
(Speech and Translation)

Statistical Approach to Automatic Speech Recognition (ASR) [Bahl & Jelinek^{+} 83]



Hidden Markov Models (HMM)

- **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: $x_1^T = x_1...x_t...x_T$**



Hidden Markov Models (HMM)

The acoustic model $p(X|W)$ provides the link between sentence hypothesis W and observations sequence $X = x_1^T = x_1 ... x_t ... x_T$:

- **acoustic probability $p(x_1^T|W)$ using hidden state sequences s_1^T:**

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

- **two types of distributions:**
  - **transition probability $p(s|s', W)$: not important**
  - **emission probability $p(x_t|s, W)$: key quantity realized by GMM: Gaussian mixtures models (trained by EM algorithm)**

- **phonetic labels (allophones, sub-phones): $(s, W) \rightarrow \alpha = \alpha_{sW}$**

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

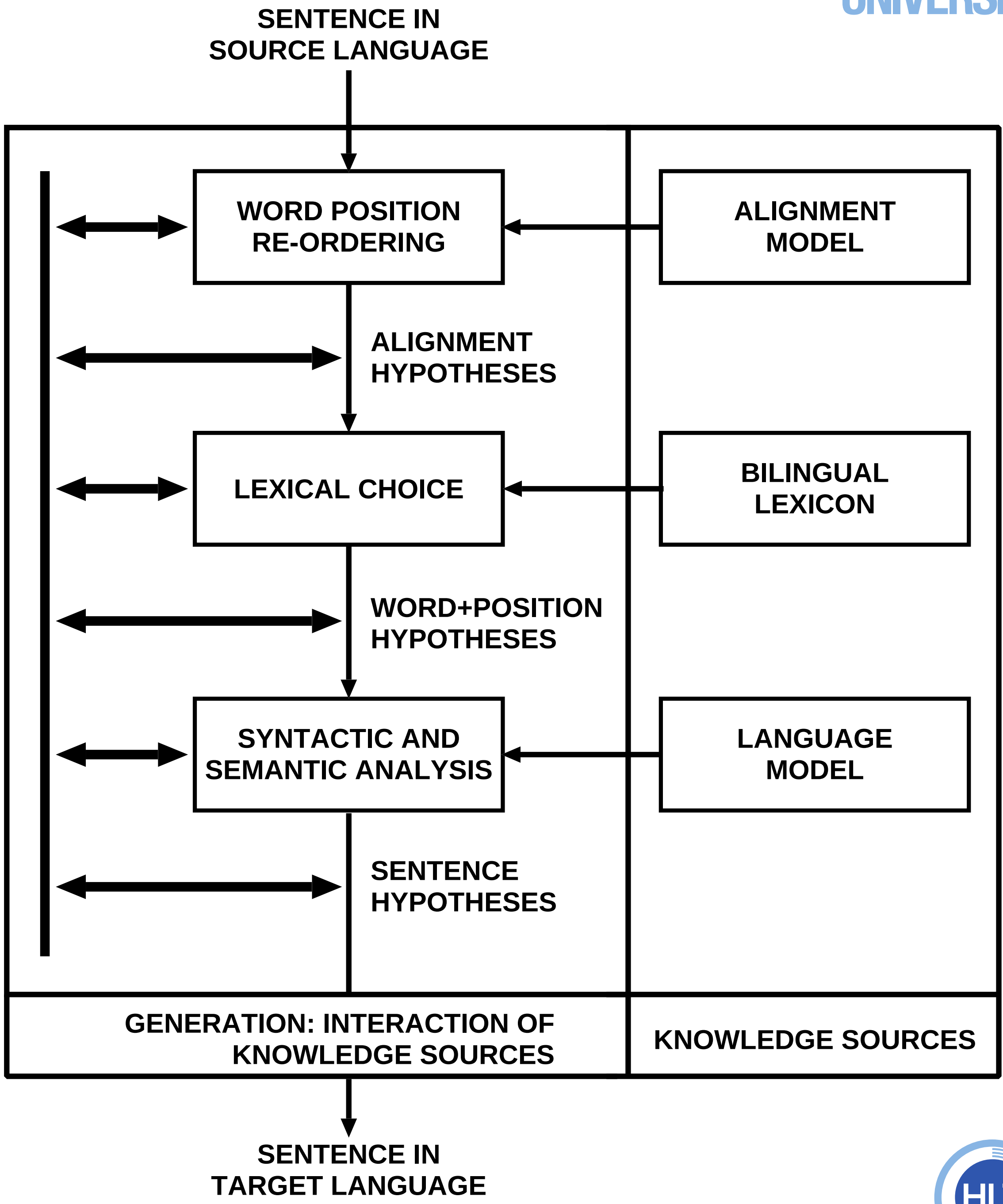
  typical approach: phoneme models in triphone context: decision trees (CART) for finding equivalence classes

- **refinements:**
  - **augmented feature vector: context window around position t**
  - **subsequent LDA (linear discriminant analysis)**

illustration: machine translation

- interaction between three models (or knowledge sources):
  – alignment model $p(A|E)$
  – lexicon model $p(E|F, A)$
  – language model $p(E)$

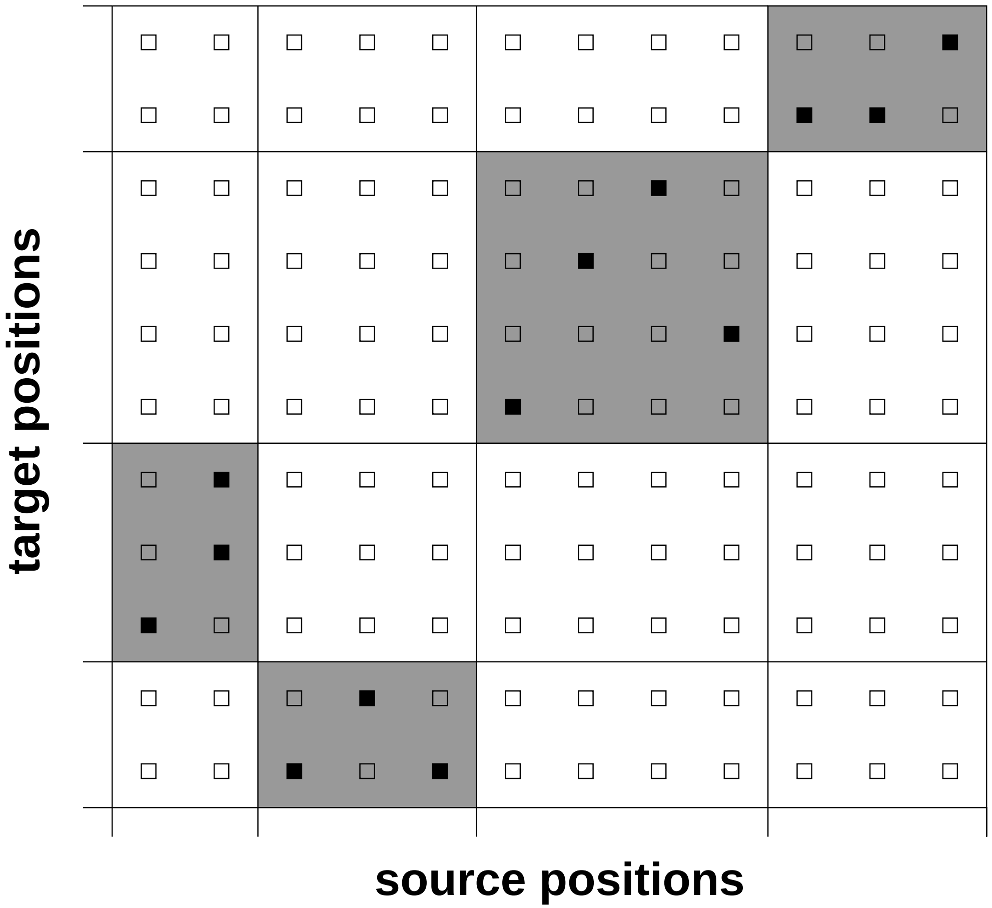
- handle interdependences, ambiguities and conflicts by Bayes decision rule as for speech recognition

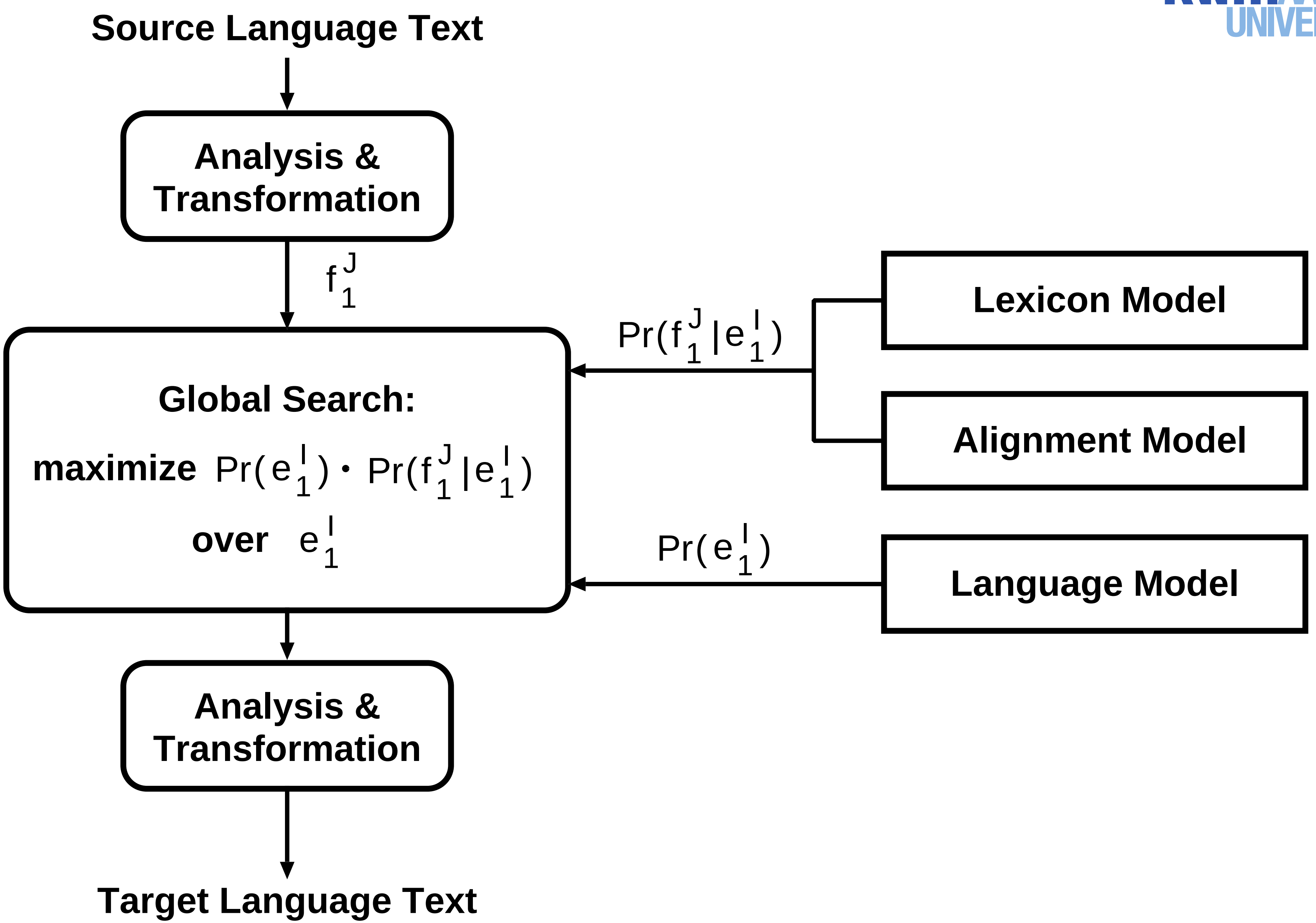


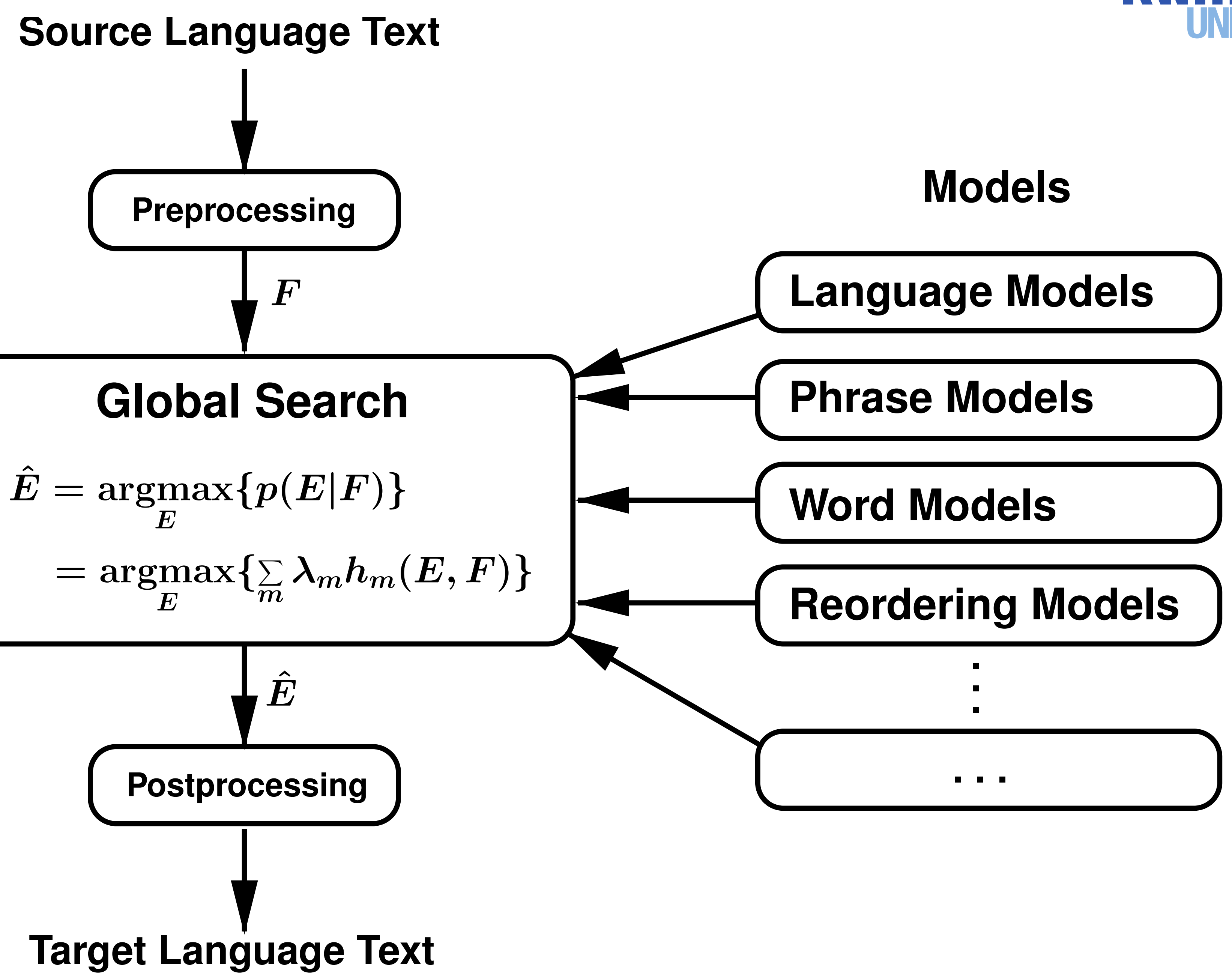
From Words to Phrases

phrase-based approach:

- **training: extraction of phrase pairs (= two-dim. 'blocks') after alignment/lexicon training**
- **translation process: phrases are the smallest units**




Source Language Text
Analysis & Transformation
f_1^J
Global Search:
maximize $Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)$
over e_1^I
$Pr(f_1^J|e_1^I)$
Lexicon Model
Alignment Model
$Pr(e_1^I)$
Language Model
Analysis & Transformation
Target Language Text


Source Language Text
Preprocessing
F
Global Search
$\hat{E} = \underset{E}{\operatorname{argmax}}\{p(E|F)\}$
$= \underset{E}{\operatorname{argmax}}\{\sum_{m} \lambda_m h_m(E,F)\}$
$\hat{E}$
Postprocessing
Target Language Text
Models
Language Models
Phrase Models
Word Models
Reordering Models
...

REFERENCES

References

[Augustin & Brodin+ 06] E. Augustin, J. Brodin, M. Carré, E. Geoffrois, E. Grosicki, F. Prêteux: RIMES evaluation campaign for handwritten mail processing. Proceedings of the Workshop on Frontiers in Handwriting Recognition, La Baule, France, Oct. 2006

[Bahl & Jelinek+ 83] L. R. Bahl, F. Jelinek, R. L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 179-190, March 1983.

[Bahl & Brown+ 86] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer: Maximum mutual information estimation of hidden Markov parameters for speech recognition. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Tokyo, pp.49-52, April 1986.

[Beck & Schlüter+ 15] E. Beck, R. Schlüter, H. Ney: Error Bounds for Context Reduction and Feature Omission, Interspeech, Dresden, Germany, Sep. 2015.

[Bengio & Ducharme+ 00] Y. Bengio, R. Ducharme, P. Vincent: A neural probabilistic language model. Advances in Neural Information Processing Systems (NIPS), pp. 933-938, Denver, CO, USA, Nov. 2000.

[Botros & Irie+ 15] R. Botros, K. Irie, M. Sundermeyer, H. Ney: On Efficient Training of Word Classes and Their Application to Recurrent Neural Network Language Models. Interspeech, pp.1443-1447, Dresden, Germany, Sep. 2015.

[Bourlard & Wellekens 90] H. Bourlard, C. J. Wellekens: 'Links between Markov Models and Multilayer Perceptrons', in D.S. Touretzky (ed.): "Advances in Neural Information Processing Systems I", Morgan Kaufmann Pub., San Mateo, CA, pp.502-507, 1989.

[Bridle 89] J. S. Bridle: Probabilistic Interpretation of Feedforward Classification Network Outputs with Relationships to Statistical Pattern Recognition, in F. Fogelman-Soulie, J. Herault (eds.): 'Neuro-computing:

Algorithms, Architectures and Applications', NATO ASI Series in Systems and Computer Science, Springer, New York, 1989.

[Brown & Della Pietra+ 93] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer: Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, Vol. 19.2, pp. 263-311, June 1993.

[Castano & Vidal+ 93] M.A. Castano, E. Vidal, F. Casacuberta: Inference of stochastic regular languages through simple recurrent networks. IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, pp. 16/1-6, Colchester, UK, April 1993.

[Castano & Casacuberta 97] M. Castano, F. Casacuberta: A connectionist approach to machine translation. European Conf. on Speech Communication and Technology (Eurospeech), pp. 91–94, Rhodes, Greece, Sep. 1997.

[Castano & Casacuberta+ 97] M. Castano, F. Casacuberta, E. Vidal: Machine translation using neural networks and finite-state models. Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI), pp. 160-167, Santa Fe, NM, USA, July 1997.

[Dahl & Yu+ 12] G. E. Dahl, D. Yu, L. Deng, A. Acero: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Tran. on Audio, Speech and Language Processing, Vol. 20, No. 1, pp. 30-42, Jan. 2012.

[Devlin & Zbib+ 14] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, J. Makhoul: Fast and Robust Neural Network Joint Models for Statistical Machine Translation. Annual Meeting of the ACL, pp. 1370–1380, Baltimore, MA,, June 2014.

[Fritsch & Finke+ 97] J. Fritsch, M. Finke, A. Waibel: Adaptively Growing Hierarchical Mixtures of Experts. NIPS, Advances in Neural Information Processing Systems 9, MIT Press, pp. 459-465, 1997.

[Gers & Schmidhuber+ 00] F. A. Gers, J. Schmidhuber, F. Cummin: Learning to forget: Continual prediction with LSTM. Neural computation, Vol 12, No. 10, pp. 2451-2471, 2000.

[Gers & Schraudolph+ 02] F. A. Gers, N. N. Schraudolph, J. Schmidhuber: Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research, Vol. 3, pp. 115-143, 2002.

[Graves & Fernandez+ 06] A. Graves, S. FernÃ¡ndez, F Gomez, J. Schmidhuber: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Int.Conf. on Machine Learning, Pittsburgh, USA, pp. 369-376, 2006.

[Haffner 93] P. Haffner: Connectionist Speech Recognition with a Global MMI Algorithm. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93) Berlin, Germany, Sep. 1993.

[Hermansky & Ellis+ 00] H. Hermansky, D. W. Ellis, S. Sharma: Tandem connectionist feature extraction for conventional HMM systems. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1635-1638, Istanbul, Turkey, June 2000.

[Hinton & Osindero+ 06] G. E. Hinton, S. Osindero, Y. Teh: A fast learning algorithm for deep belief nets. Neural Computation, Vol. 18, No. 7, pp. 1527-1554, July 2006.

[Hochreiter & Schmidhuber 97] S. Hochreiter, J. Schmidhuber: Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780, Nov. 1997.

[Klakow & Peters 02] D. Klakow, J. Peters: Testing the correlation of word error rate and perplexity. Speech Communication, pp. 19–28, 2002.

[Koehn & Och+ 03] P. Koehn, F. J. Och, D. Marcu: Statistical Phrase-Based Translation. HLT-NAACL 2003, pp. 48-54, Edmonton, Canada, May-June 2003.

[Kozielski & Mathysiak+ 14] M. Kozielski, M. Matysiak, P. Doetsch, R. Schlüter, H. Ney: Open-lexicon Language Modeling Combining Word and Character Levels. Int. Conf. on Frontiers in Handwriting Recognition (ICHFR), Crete, Greece, Sep. 2014.

[Le & Allauzen+ 12] H.S. Le, A. Allauzen, F. Yvon: Continuous space translation models with neural networks. NAACL-HLT 2012, pp. 39-48, Montreal, QC, Canada, June 2002.

[LeCun & Bengio+ 94] Y. LeCun, Y. Bengio: Word-level training of a handwritten word recognizer based on convolutional neural networks. Int. Conf. on Pattern Recognition, Jerusalem, Israel, pp. 88-92, Oct. 1994.

[Lu & Bazzi+ 98] Z.A. Lu, I. Bazzi, A. Kornai, J. Makhoul, P.S. Natarajan, R. Schwartz: A Robust language-independent OCR System. AIPR Workshop Advances in Computer Assisted Recognition, Vol. 3584 of SPIE, pp. 96–104, Jan. 1998

[Marti & Bunke+ 02] U. Marti, H. Bunke: The IAM database: an English sentence database for offline handwriting recognition. Int. Journal of Document Analysis and Recognition, pp. 39–46, 2002

[Mikolov & Karafiat+ 10] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur: Recurrent neural network based language model. Interspeech, pp. 1045-1048, Makuhari, Chiba, Japan, Sep. 2010.

[Nakamura & Shikano 89] M. Nakamura, K. Shikano: A Study of English Word Category Prediction Based on Neural Networks. ICASSP 89, p. 731-734, Glasgow, UK, May 1989.

[Ney 03] H. Ney: On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition. First Iberian Conf. on Pattern Recognition and Image Analysis, Puerto de Andratx, Spain, Springer LNCS Vol. 2652, pp. 636-645, June 2003.

[Och & Ney 03] F. J. Och, H. Ney: A Systematic Comparison of Various Alignment Models. *Computational Linguistics,* Vol. 29, No. 1, pp. 19-51, March 2003.

[Och & Ney 04] F. J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417-449, Dec. 2004.

[Och & Tillmann+ 99] F. J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. Joint ACL/SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, pp. 20-28, June 1999.

[Pham & Bluche+ 14] V. Pham, T. Bluche, C. Kermorvant, J. Louradour: Dropout improves recurrent neural networks for handwriting recognition. Int. Conf. on Frontiers in Handwriting Recognition (ICFHR) Crete, Greece, Sep. 2014

[Povey & Woodland 02] D. Povey, P.C. Woodland: Minimum phone error and I-smoothing for improved discriminative training. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 105–108, Orlando, FL, May 2002.

[Robinson 94] A. J. Robinson: An Application of Recurrent Nets to Phone Probability Estimation. IEEE Trans. on Neural Networks, Vol. 5, No. 2, pp. 298-305, March 1994.

[Schlüter & Nussbaum$^+$ 12] R. Schlüter, M. Nussbaum-Thom, H. Ney: Does the Cost Function Matter in Bayes Decision Rule? IEEE Trans. PAMI, No. 2, pp. 292–301, Feb. 2012.

[Schlüter & Nussbaum-Thom$^+$ 13] R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhouli, H. Ney: Novel Tight Classification Error Bounds under Mismatch Conditions based on f-Divergence. IEEE Information Theory Workshop, pp. 432–436, Sevilla, Spain, Sep. 2013.

[Schuster & Paliwal 97] M. Schuster, K. K. Paliwal: Bidirectional Recurrent Neural Networks. IEEE Trans. on SIgnal Processing, Vol. 45, No. 11, pp. 2673-2681, Nov. 1997.

[Schwenk 07] H. Schwenk: Continuous space language models. Computer Speech and Language, Vol. 21, No. 3, pp. 492–518, July 2007.

[Schwenk 12] H. Schwenk: Continuous Space Translation Models for Phrase-Based Statistical Machine Translation. 24th Int. Conf. on Computational Linguistics (COLING), Mumbai, India, pp. 1071–1080, Dec. 2012.

[Schwenk & Costa-jussa$^+$ 07] H. Schwenk , M. R. Costa-jussa, J. A. R. Fonollosa: Smooth bilingual n-gram translation. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 430–438, Prague, June 2007.

[Schwenk & Déchelotte$^+$ 06] H. Schwenk, D. Déchelotte, J. L. Gauvain: Continuous Space Language Models for Statistical Machine Translation. COLING/ACL 2006, pp. 723–730, Sydney, Australia July 2006.

[Solla & Levin$^+$ 88] S. A. Solla, E. Levin, M. Fleisher: Accelerated Learning in Layered Neural Networks. Complex Systems, Vol.2, pp. 625-639, 1988.

[Sundermeyer & Alkhouli$^+$ 14] M. Sundermeyer, T. Alkhouli, J. Wuebker, H. Ney: Translation Modeling with Bidirectional Recurrent Neural Networks. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 14–25, Doha, Qatar, Oct. 2014.

[Sundermeyer & Ney$^+$ 15] M. Sundermeyer, H. Ney, R. Schlüter: From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol. 23, No. 3, pp. 13–25, March 2015.

[Sundermeyer & Schlüter$^+$ 12] M. Sundermeyer, R. Schlüter, H. Ney: LSTM neural networks for language modeling. Interspeech, pp. 194–197, Portland, OR, USA, Sep. 2012.

[Utgoff & Stracuzzi 02] P. E. Utgoff, D. J. Stracuzzi: Many-layered learning. Neural Computation, Vol. 14, No. 10, pp. 2497-2539, Oct. 2002.

[Vaswani & Zhao$^+$ 13] A. Vaswani, Y. Zhao, V. Fossum, D. Chiang: Decoding with Large-Scale Neural Language Models Improves Translation. Conf. on Empirical Methods in Natural Language Processing (EMNLP, pp. 1387–1392, Seattle, Washington, USA, Oct. 2013.

[Vogel & Ney$^+$ 96] S. Vogel, H. Ney, C. Tillmann: HMM-based word alignment in statistical translation. Int. Conf. on Computational Linguistics (COLING), pp. 836-841, Copenhagen, Denmark, Aug. 1996.

[Waibel & Hanazawa$^+$ 88] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. L. Lang: Phoneme Recognition: Neural Networks vs. Hidden Markov Models. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New York, NY, pp.107-110, April 1988.

[Zens & Och$^+$ 02] R. Zens, F. J. Och, H. Ney: Phrase-Based Statistical Machine Translation. 25th Annual German Conf. on AI, pp. 18–32, LNAI, Springer 2002.

END