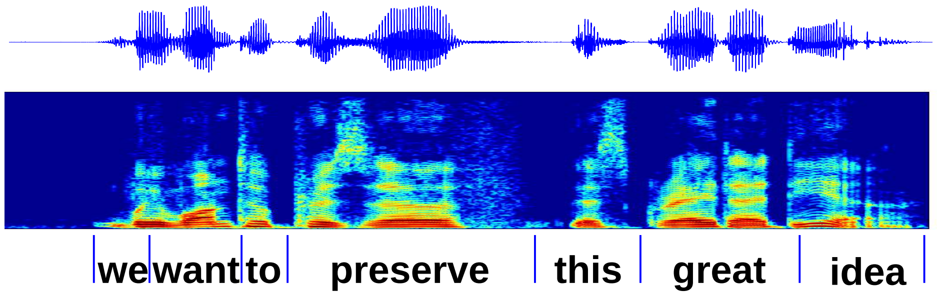
Loria & Université de Lorraine, Colloquium, Nancy, France
27-January-2017

Human Language Technology and Machine Learning:
From Bayes Decison Theory to Deep Learning

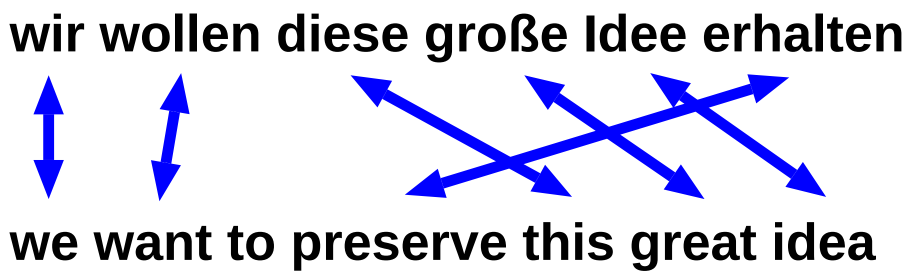
Hermann Ney
Human Language Technology and Pattern Recognition
RWTH Aachen University, Aachen, Germany

IEEE Distinguished Lecturer 2016/17

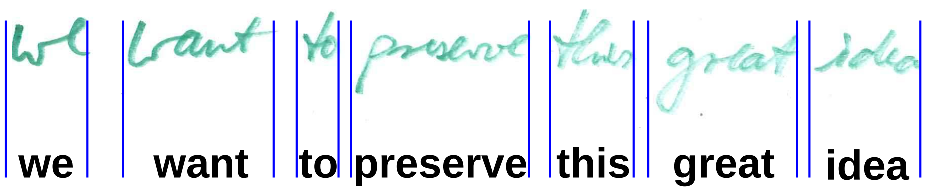
Speech Recognition



Machine Translation



Handwriting Recognition (Text Image Reognition)



tasks:
- speech recognition
- machine translation

- handwriting recognition
  (+ sign language,...)

Human Language Technology:
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 string $\rightarrow$ output string**
- **output string: string of words/letters in a natural language**
- **models of context and dependencies: strings in input and output**
  – within input and output string
  – across input and output string

Projects

activities of my team (RWTH, Philips until 1993) in large-scale joint projects:

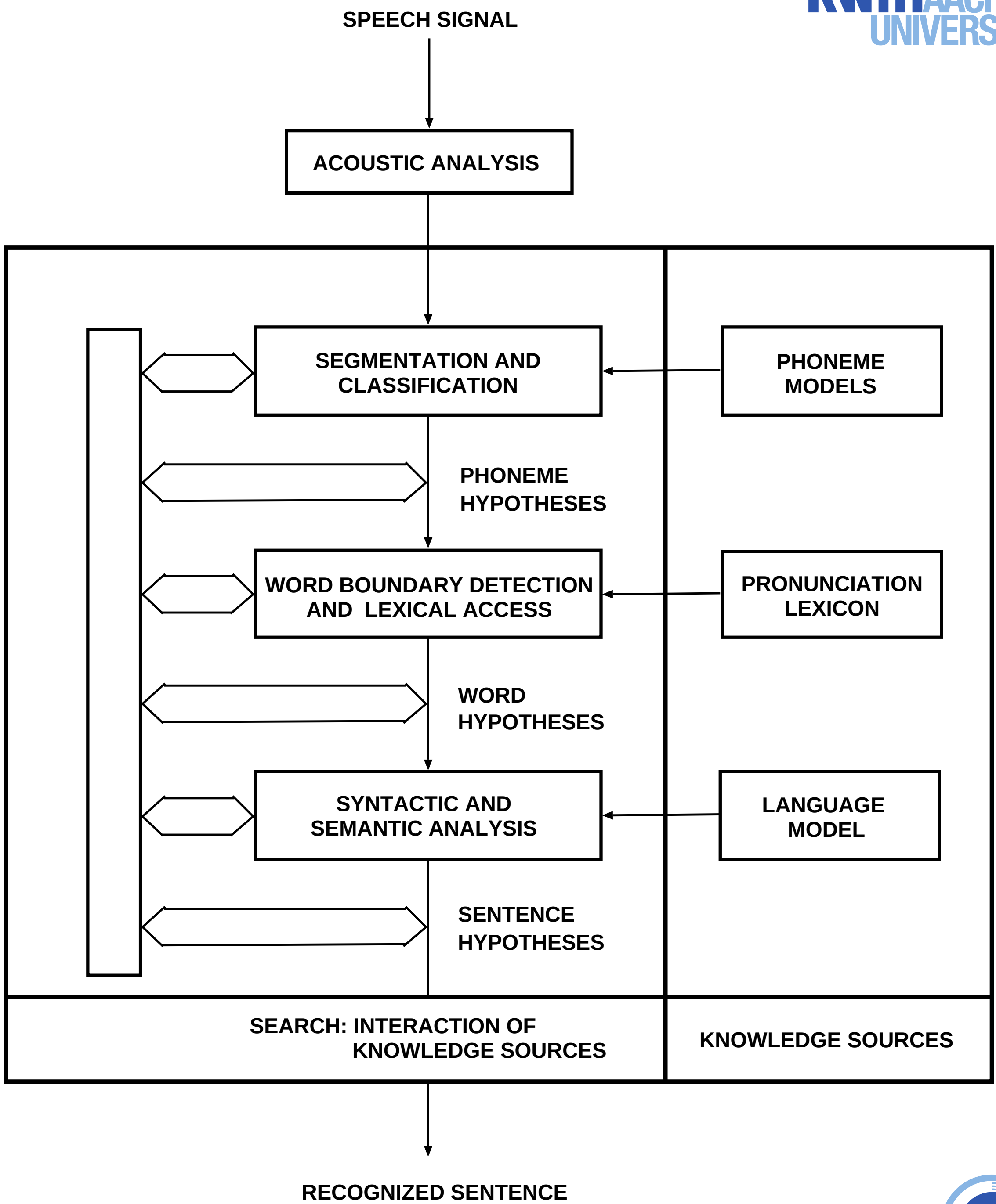
- SPICOS 1984-1989: speech recognition und understanding
  - conditions: 1000 words, continuous speech, speaker dependent
  - funded by German BMBF: Siemens, Philips, German universities
- Verbmobil 1993-2000: funded by German BMBF
  - domain: appointment scheduling, recognition and translation, German-English, limited vocabulary (8.000 words)
  - large project: 10 million DM per year, about 25 partners
  - German partners: Daimler, Philips, Siemens, DFKI, KIT, RWTH, U Stuttgart, ...
- TC-STAR 2004-2007: funded by EU
  - recognition and translation of speeches given in EU parliament
  - first research system for SPEECH TRANSLATION on real-life data
  - partners: UPC Barcelona, RWTH, CNRS Paris, KIT Karlsruhe, IBM-US Research, ...
- GALE 2005-2011: funded by US DARPA
  - recognition, translation and understanding for Chinese and Arabic
  - largest project ever on HLT: 40 million USD per year, about 30 partners
  - US partners: BBN, IBM, SRI, CMU, Stanford U, Columbia U, UW, USCLA, ...
  - EU partners: CNRS Paris, U Cambridge, RWTH

- **BOLT 2011-2015: funded by US DARPA**
  - **follow-up to GALE**
  - **emphasis on colloquial language for Arabic and Chinese**
- **QUAERO 2008-2013: funded by OSEO France**
  - **recognition and translation of European languages, more colloquial speech, handwriting recognition**
  - **French partners (23): Thomson, France Telecom, Bertin, Systran, CNRS, INRIA, universities,...**
  - **German Partners (2): KIT, RWTH**
- **BABEL 2012-2016: funded by US IARPA**
  - **key word spotting with noisy and low-resource training data**
  - **rapid development for new languages (e.g. within 48 hours)**
- **EU projects 2012-2014: EU-Bridge, TransLectures**
  emphasis on recognition and translation of lectures (academic, TED, ...)

speech recognition:
what is the problem?
– ambiguities at all levels
– interdependencies of decisions

approach [CMU and IBM 1975]:
– hypothesis scores
– probabilistic framework
– statistical decision theory

modern terminology:
machine learning



Statistical Approach: 1975-2015

- two strings: input $x_1^T := x_1...x_m...x_T$ and output $c_1^N := c_1...c_n...c_N$
  with a probabilistic dependence: $p(c_1^N|x_1^T)$
- performance measure or loss (error) function: $L[\tilde{c}_1^{\tilde{N}}, c_1^N]$
  between true output $\tilde{c}_1^{\tilde{N}}$ and hypothesized output c_1^N
- Bayes decision rule minimizes expected loss:

$$x_1^T \to \hat{c}_1^{\hat{N}}(x_1^T) \;:=\; \arg\min_{N,c_1^N} \left\{ \sum_{\tilde{N},\tilde{c}_1^{\tilde{N}}} p(\tilde{c}_1^{\tilde{N}}|x_1^T) \cdot L[\tilde{c}_1^{\tilde{N}}, c_1^N] \right\}$$

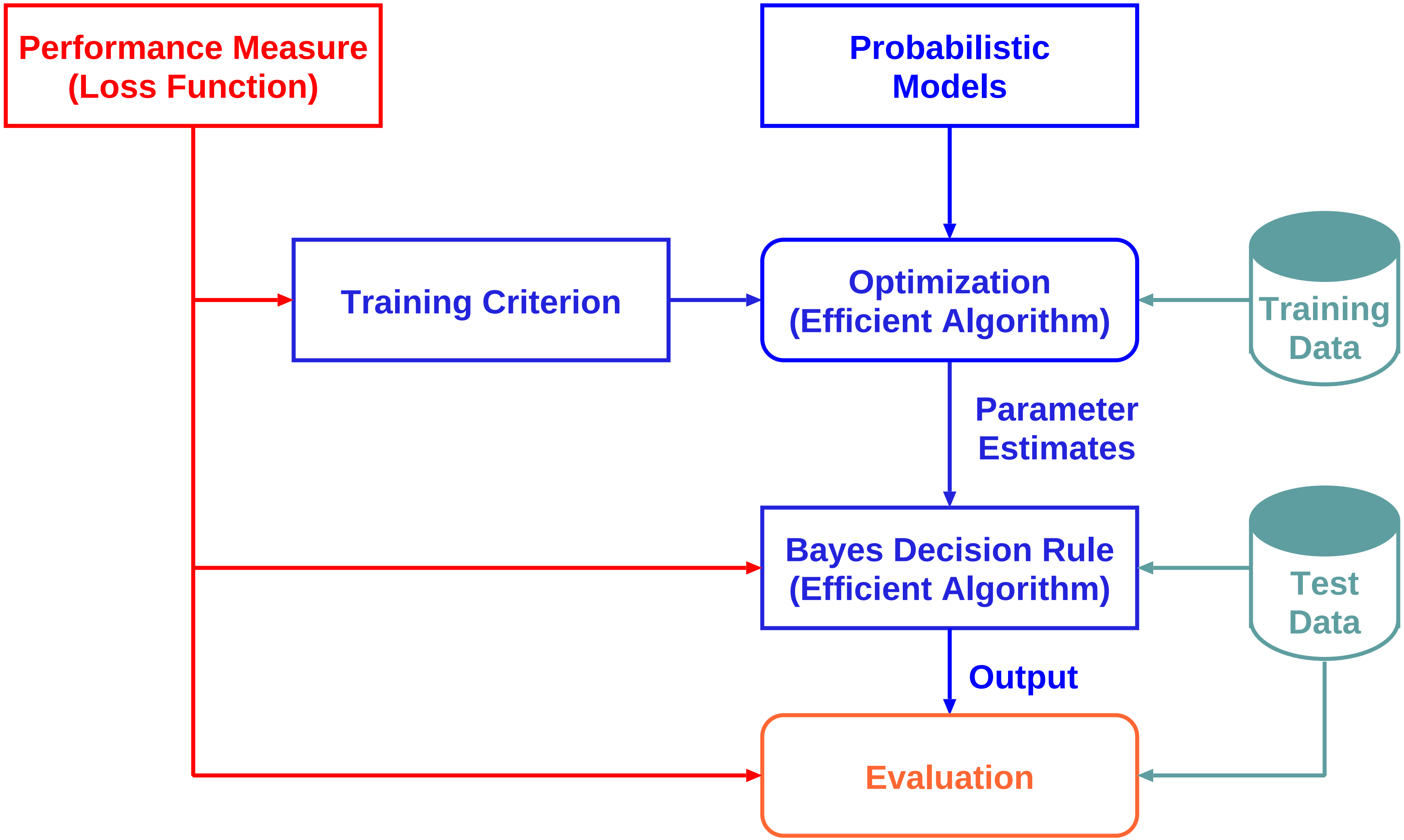
simplified rule (minimum string error): $x_1^T \to \hat{c}_1^{\hat{N}}(x_1^T) := \arg\max_{N,c_1^N} \left\{ p(c_1^N|x_1^T) \right\}$

- from true to model distribution: separation of language model $p(c_1^N)$

$$p(c_1^N|x_1^T) = p(c_1^N) \cdot p(x_1^T|c_1^N) \Big/ p(x_1^T)$$

  – advantage: huge amounts of training data without annotation
  – extension: log-linear modelling

Statistical Approach to HLT Tasks



four ingredients:

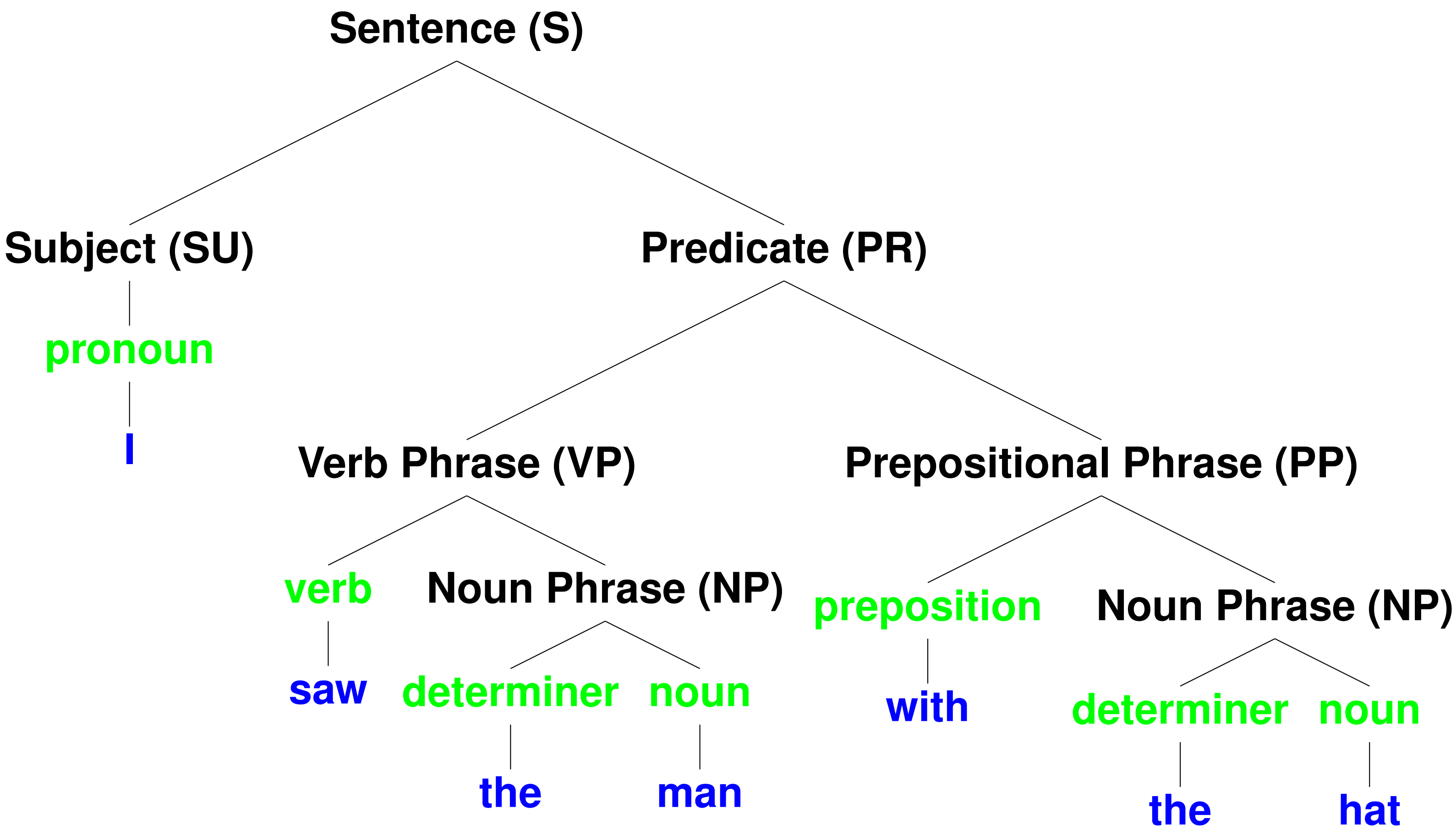
- **performance measure: error measure (e.g. edit distance)**
  we have to decide how to judge the quality of the system output
- **probabilistic models with suitable structures (*machine learning*):**
  to capture the dependencies within and between input and output strings
  - **elementary observations: Gaussian mixtures, log-linear models, support vector machines (SVM), artificial neural nets (ANN), ...**
  - **strings: n-gram Markov chains, CRF, Hidden Markov models (HMM), recurrent neural nets (RNN), LSTM RNN, ANN-based models of attention, ...**
- **training criterion (*machine learning*):**
  to learn the free model parameters from examples
  - **ideally should be linked to performance criterion *(end-to-end training)***
  - **might 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, ...

History Speech Recognition 1975-2015

- steady increase of challenges:
  - vocabulary size: 10 digits ... 1000 ... 10.000 ... 500.000 words
  - speaking style: read speech ... colloquial/spontaneous speech
- steady improvement of statistical methods:
  HMM, Gaussians and mixtures, statistical trigram language model,
  adaptation methods, artificial neural nets, ...
- 1985-93: criticism about statistical approach
  - too many parameters and saturation effect
  - ... 'will never work for large vocabularies' ...
- remedy(?) by rule-based approach:
  - language models (text): linguistic grammars and structures
  - phoneme models (speech): acoustic-phonetic expert systems
  - limited success for various reasons:
    huge manual effort is required!
    problem of coverage and consistency of rules
- evaluations: experimental tests:
  - the same evaluation criterion on the same test data
  - direct comparison of algorithms and systems

- principle:



- extensions along many dimensions

From Speech Recognition to Machine Translation (MT)

dichotomy until 1990:
– speech: signals $\rightarrow$ statistics (engineers, industrial labs)
– text: symbols $\rightarrow$ rules (linguists, universities)

use of statistics has been controversial in text processing
(symbolic processing and computational linguistics):

- 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 (rule-based)
  human language technology and artificial intelligence

history of statistical approach to MT:

- 1989-94: pioneering work at IBM Research
  key people (R. Mercer, P. Brown) left for *Renaissance Technologies* (hedge fund)
- since 1995: only a few teams advocated statistical MT:
  RWTH, UP Valencia, HKUST Hong Kong, CMU Pittsburgh
- around 2004: from singularity to mainstream in MT
  F. Och (and more RWTH PhD students) joined Google
- 2008 service *Google Translate*

Hidden Markov Models for MT: Word Alignments (Canadian Parliament; IBM 1993)

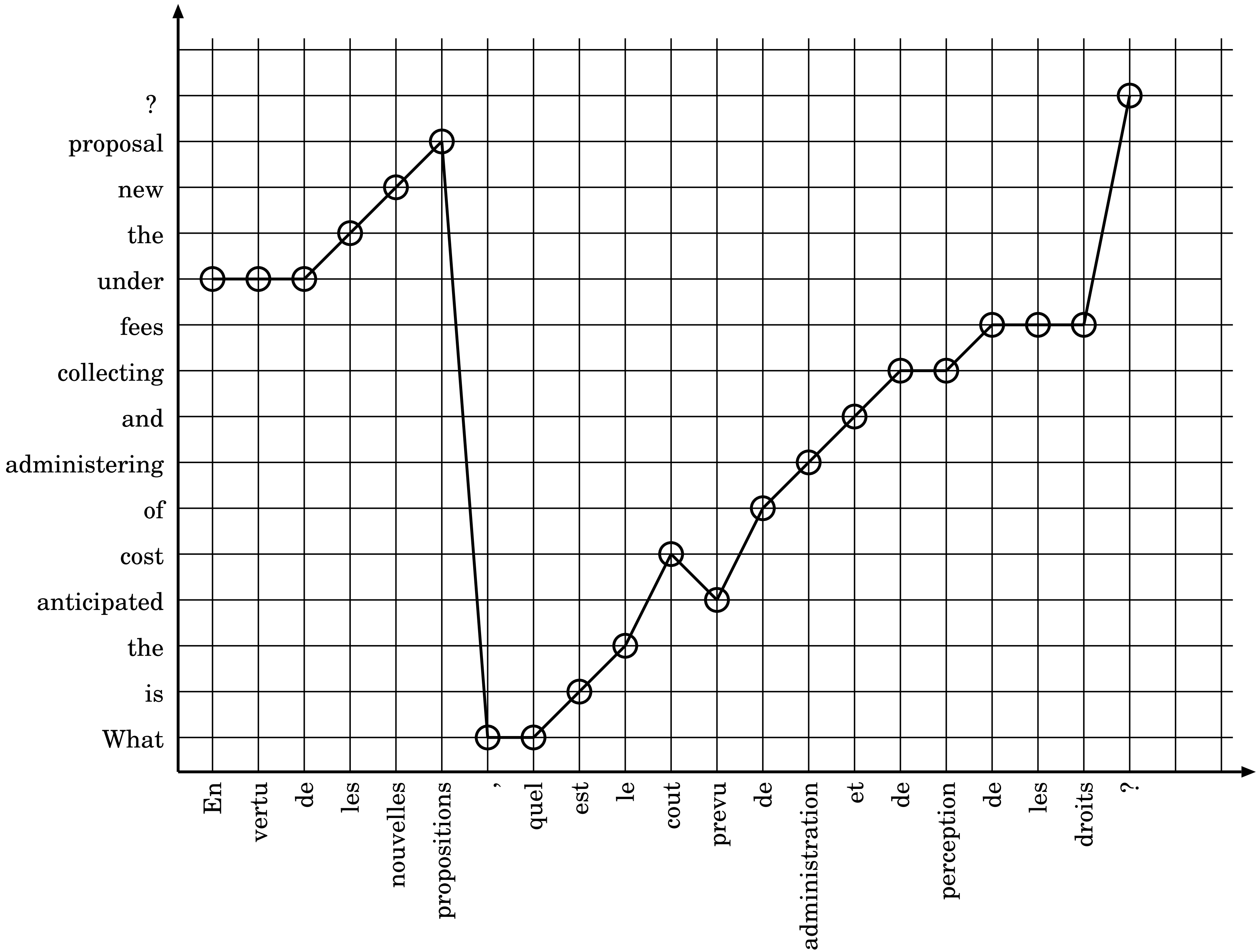
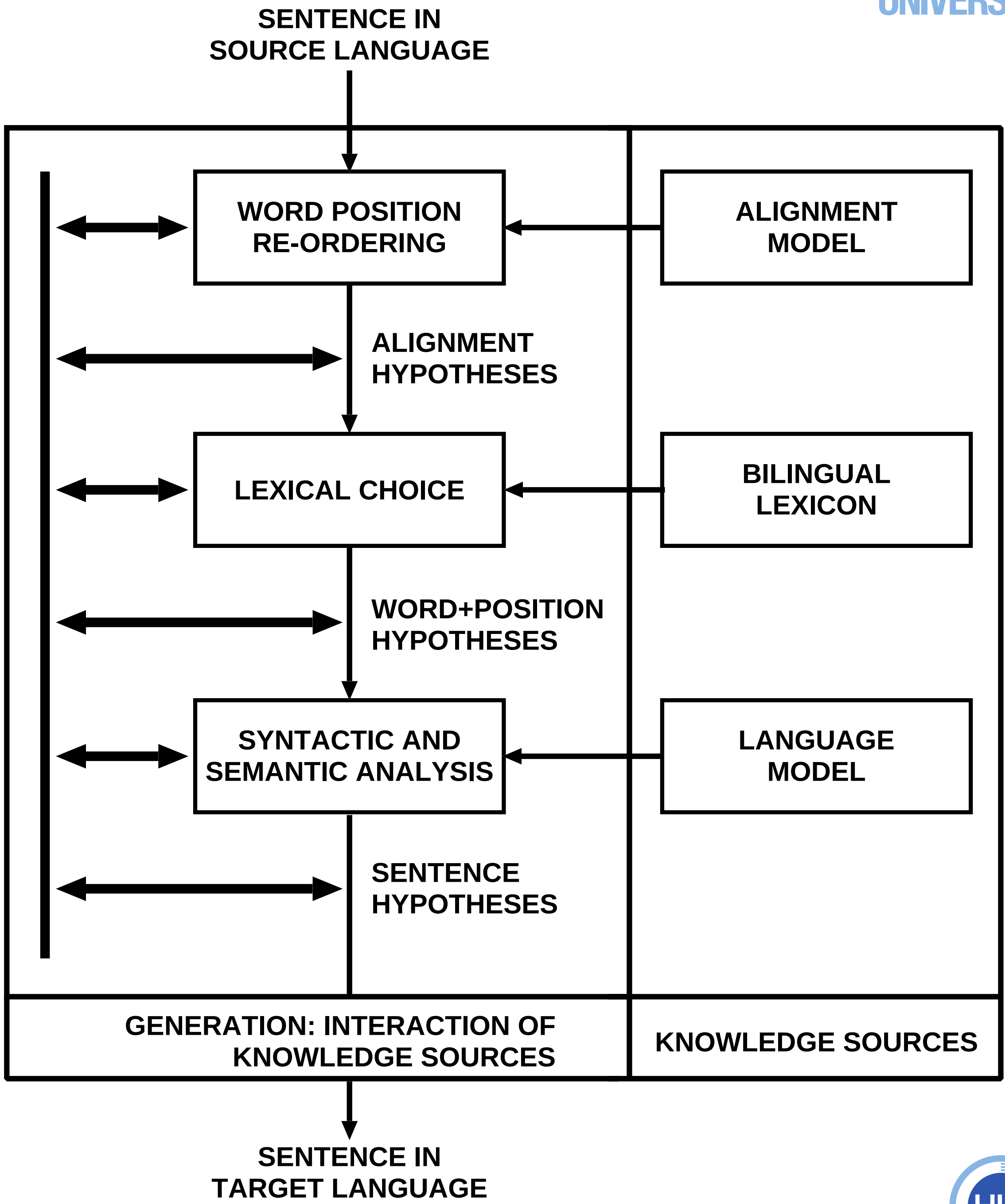


illustration: machine translation

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

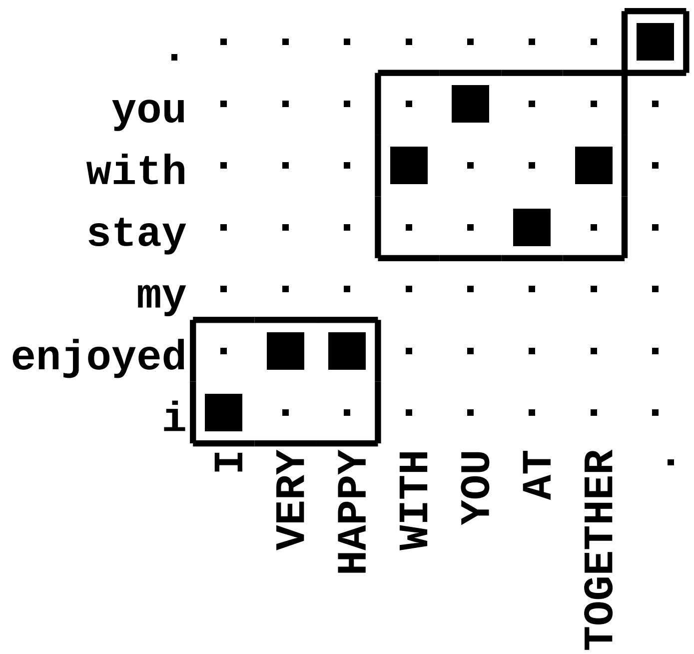
- handle interdependences, ambiguities and conflicts by Bayes decision rule as for speech recognition



From Single Words to Word Groups (Phrases) (RWTH 1998-2002)

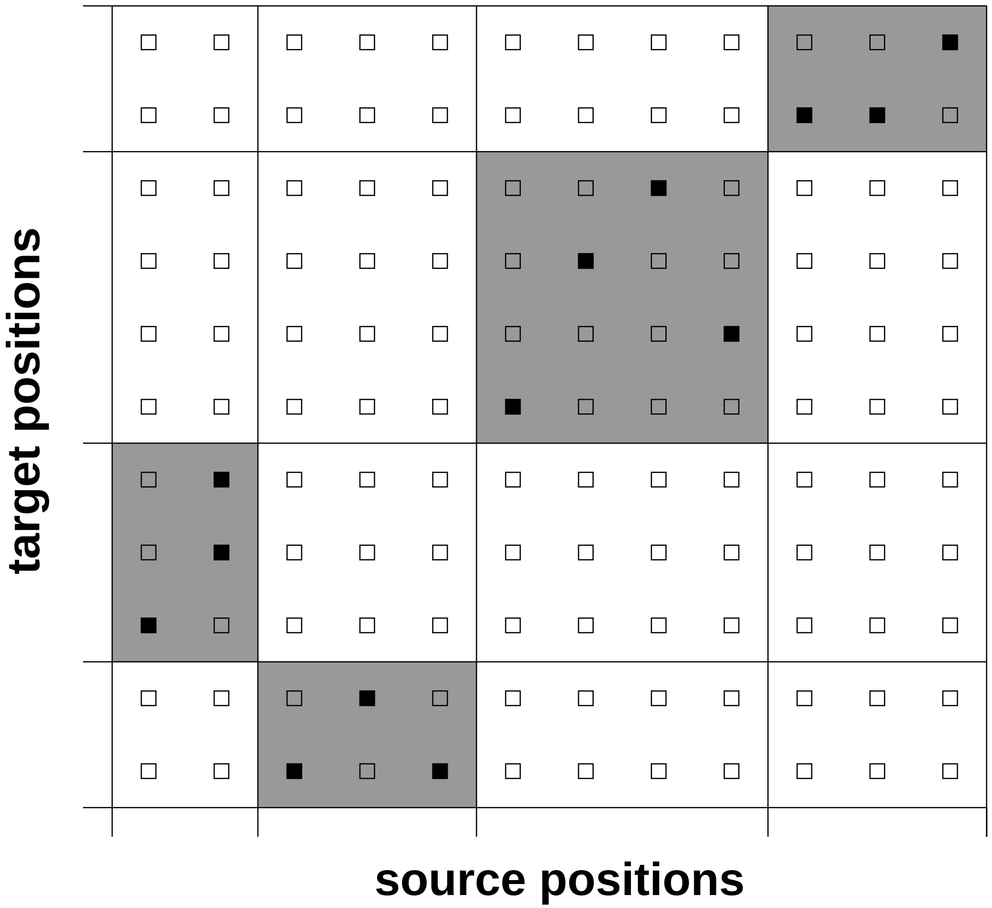
source sentence	我 很 高兴 和 你 在 一起 .
gloss notation	**I VERY HAPPY WITH YOU AT TOGETHER .**
target sentence	**I enjoyed my stay with you .**

best alignment for source → target language:

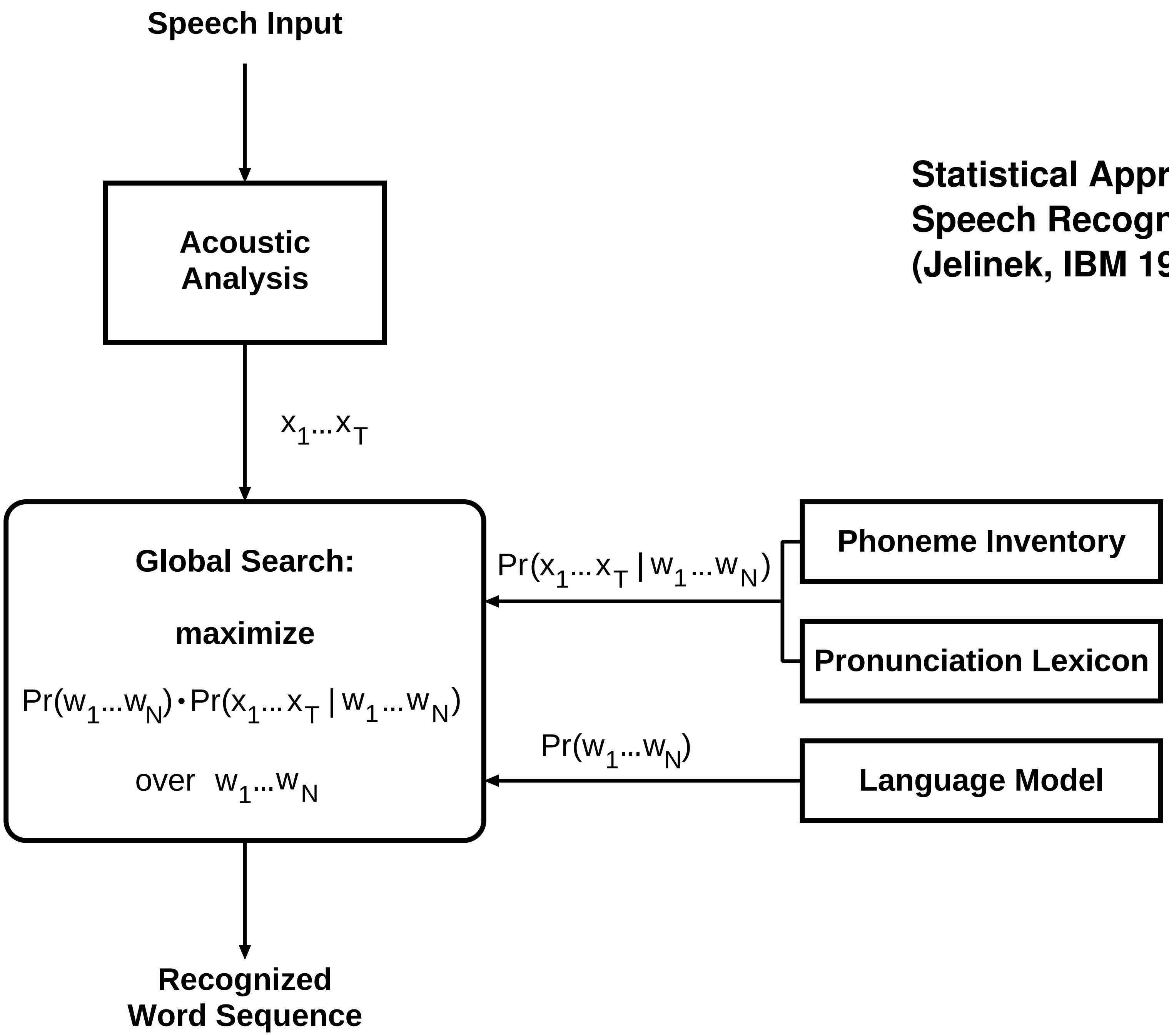


phrase-based approach:

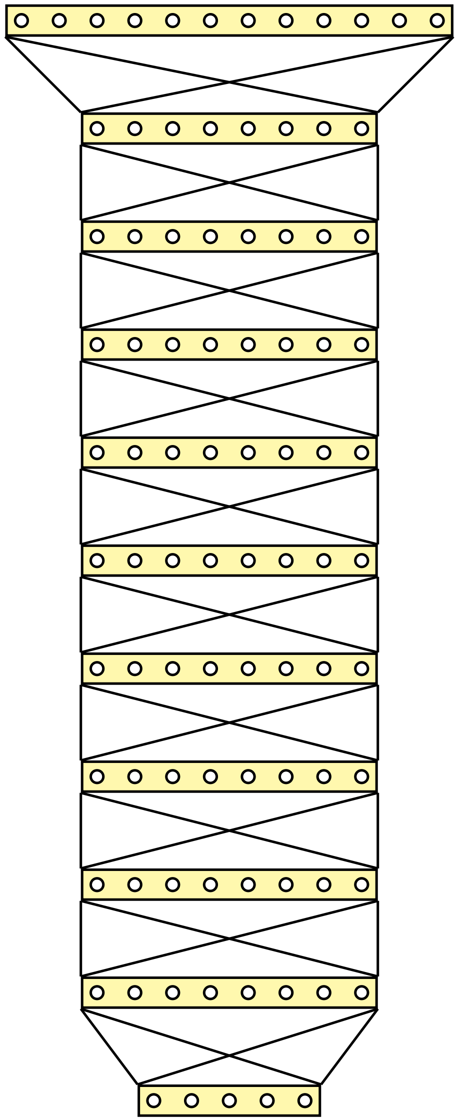
- training: extraction of phrase pairs (= two-dim. 'blocks') after alignment/lexicon training
- translation process: phrases are the smallest units



Statistical Approach to Automatic Speech Recognition (ASR) (Jelinek, IBM 1983)



Artificial Neural Networks (ANN): What is Different Now after 25 Years?



important property:

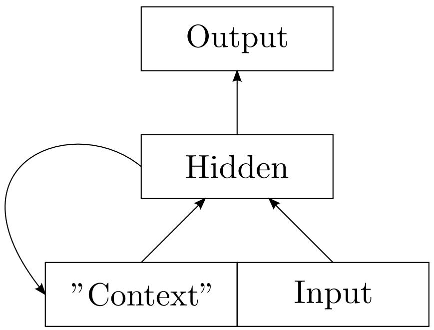
ANN outputs are probability estimates

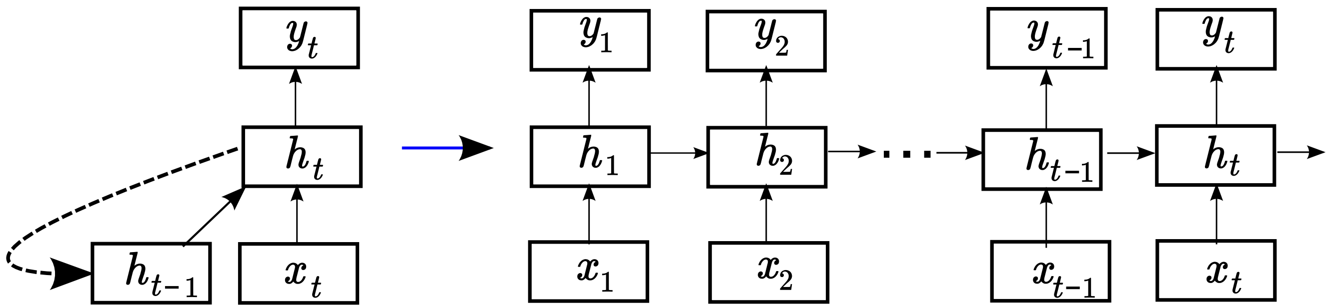
today: huge improvements by ANN:
– image object recognition
– speech recognition
– machine translation ?

comparison for ASR: today vs. 1989-1994:

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

principle for string 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]

hybrid approach:

replace emission probability of an hidden Markov model by ANN ouput

three types of hidden Markov models:

– 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

ANNs in Language Modelling

goal of language modelling: compute the prior $p(c_1^N)$ of a word sequence c_1^N
– how plausible is this word sequence c_1^N?
– measure of language model quality: perplexity PP, i.e. effective vocabulary size

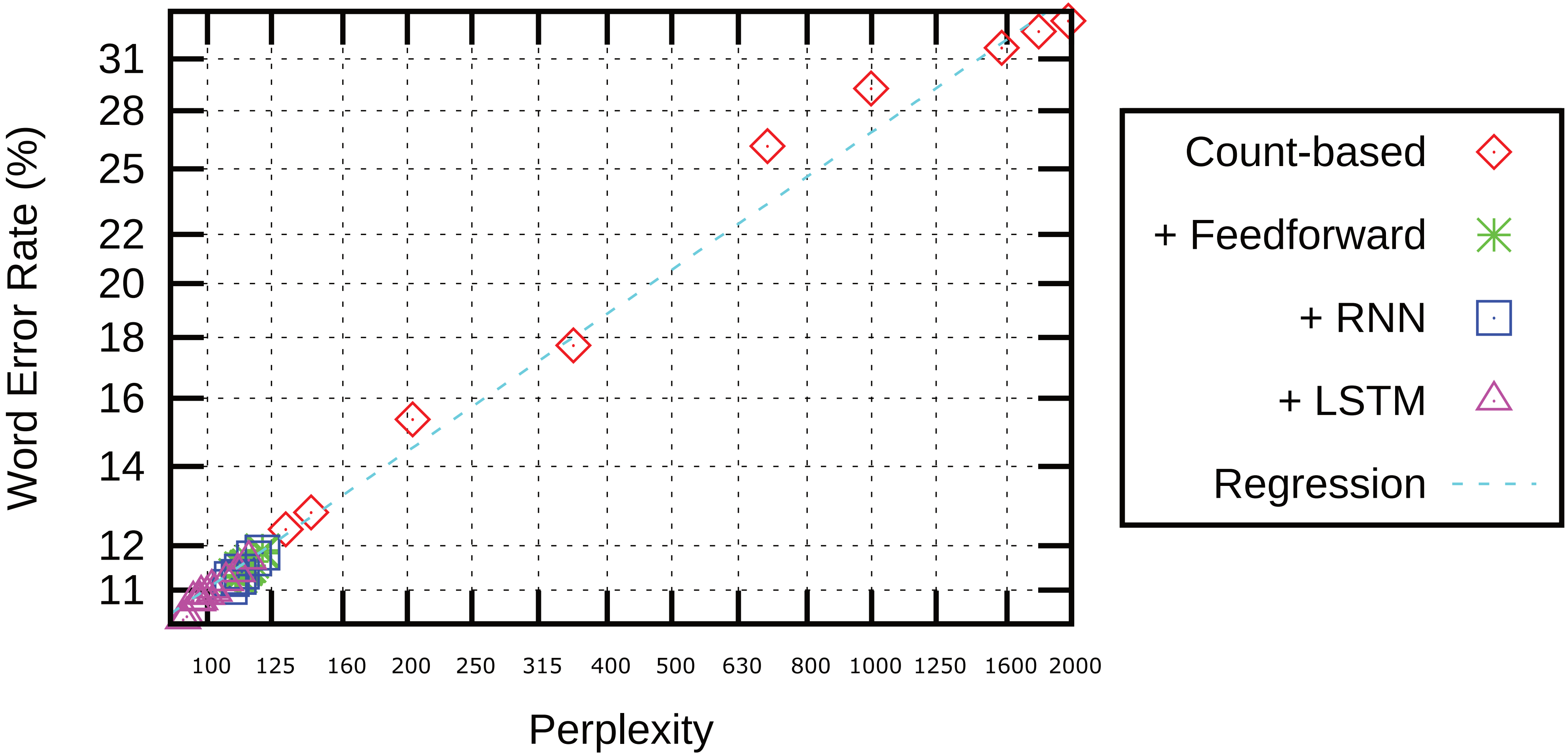
results on QUAERO English (like before):
– vocabulary size: 150k words
– training text: 50M words
– test set: 39k words

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

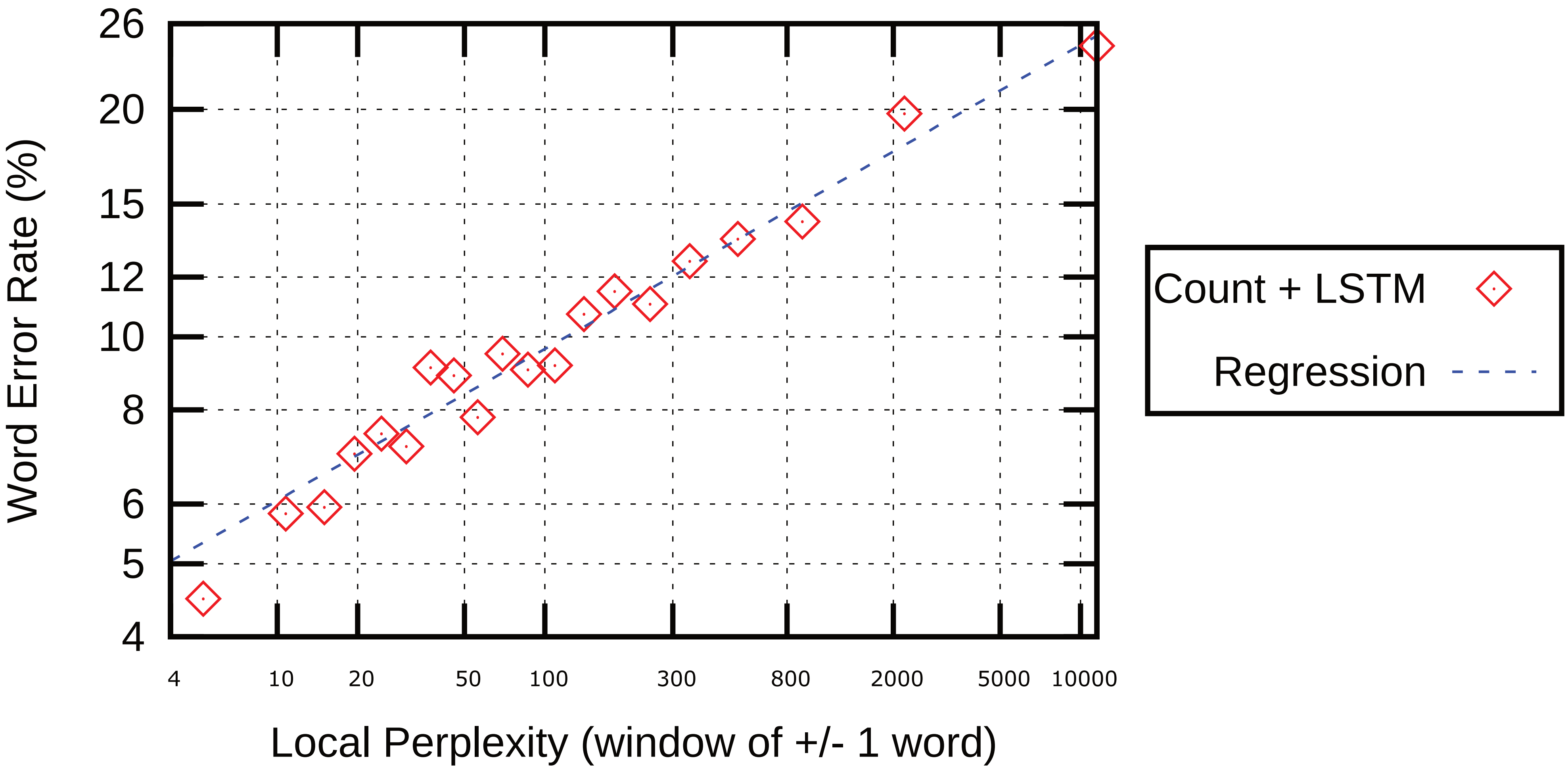
important result: improvement of PP by 40%

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



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

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

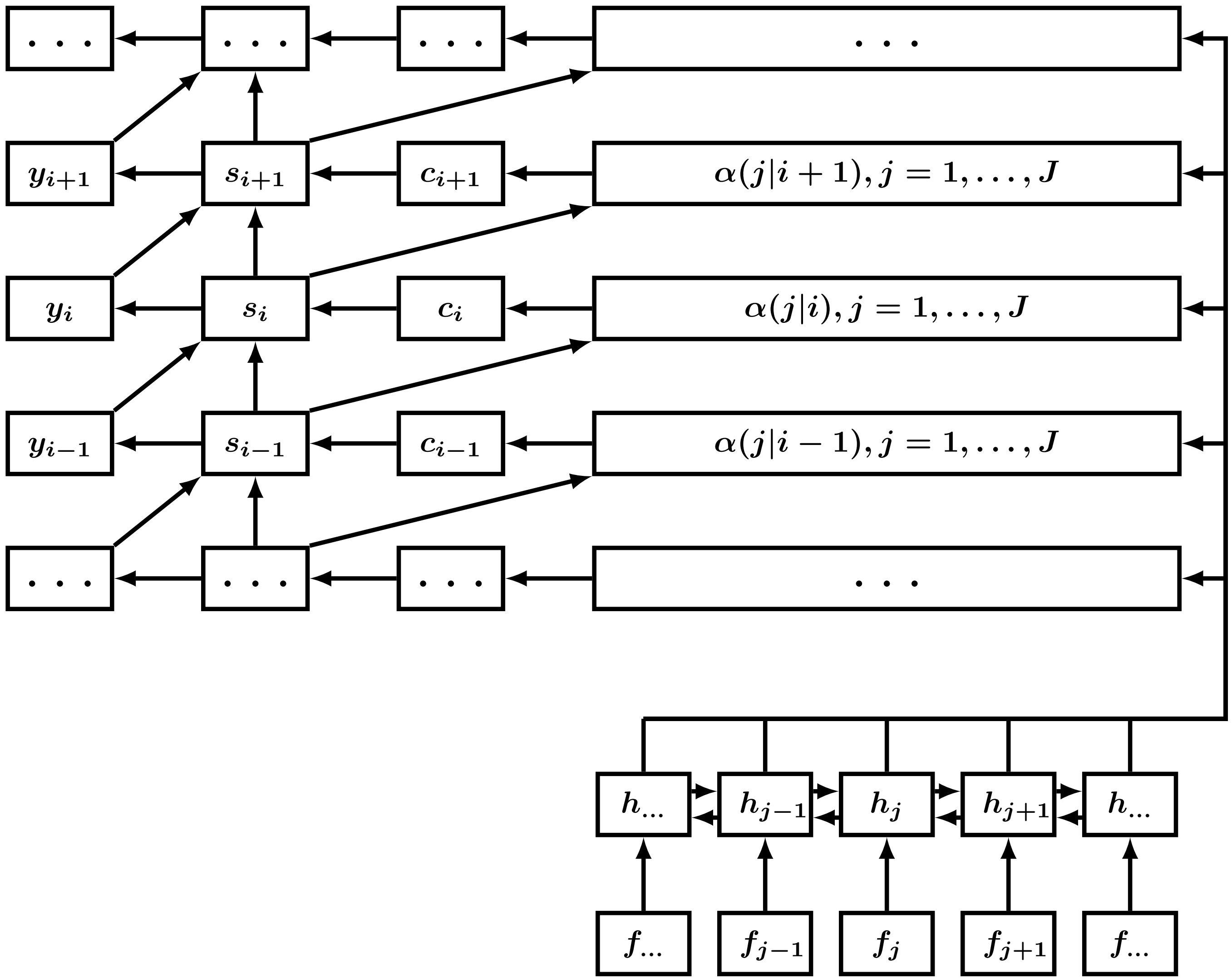


- four key ingredients:
  - choice of performance measure: errors at string, word, phoneme, frame level
  - probabilistic models at these levels and the interaction between these levels
  - training criterion along with an optimization algorithm
  - Bayes decision rule along with an efficient implementation
- about recent work on artificial neural nets (2009-15):
  - they result in significant improvements
  - they provide one more type of probabilistic models
  - they are PART of the statistical approach
- specific future challenges for statistical approach (incl. ANNs) in general:
  - complex mathematical model that is difficult to analyze
  - questions: can we find suitable mathematical approximations with more explicit descriptions of the dependencies and level interactions and of the performance criterion (error rate)?
- specific challenges for ANNs:
  - can the HMM-based alignment mechanism be replaced?
  - can we find ANNs with more explicit probabilistic structures?

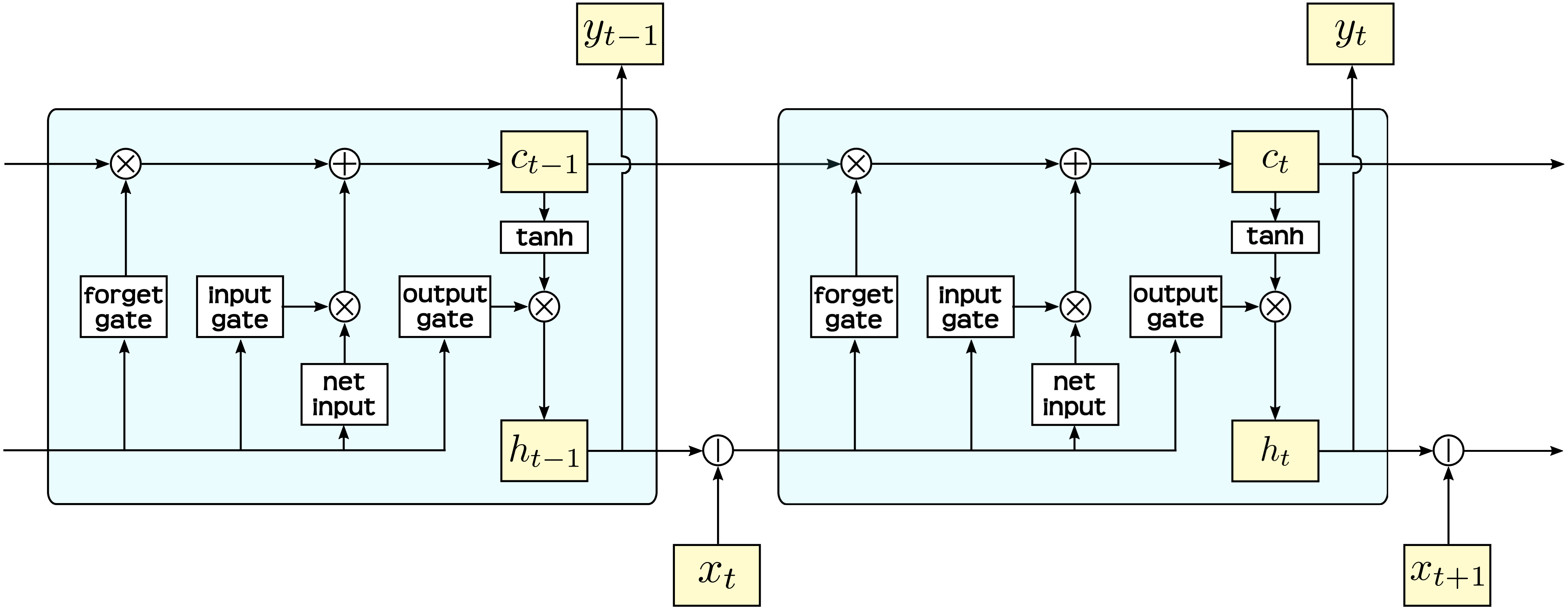
BACK-UP SLIDES

Attention-based NN MT
[?]

GRU: gated recurrence unit (similar to LSTM-RNN)



Recurrent Neural Network:
Details of Long Short-Term Memory



ingredients:

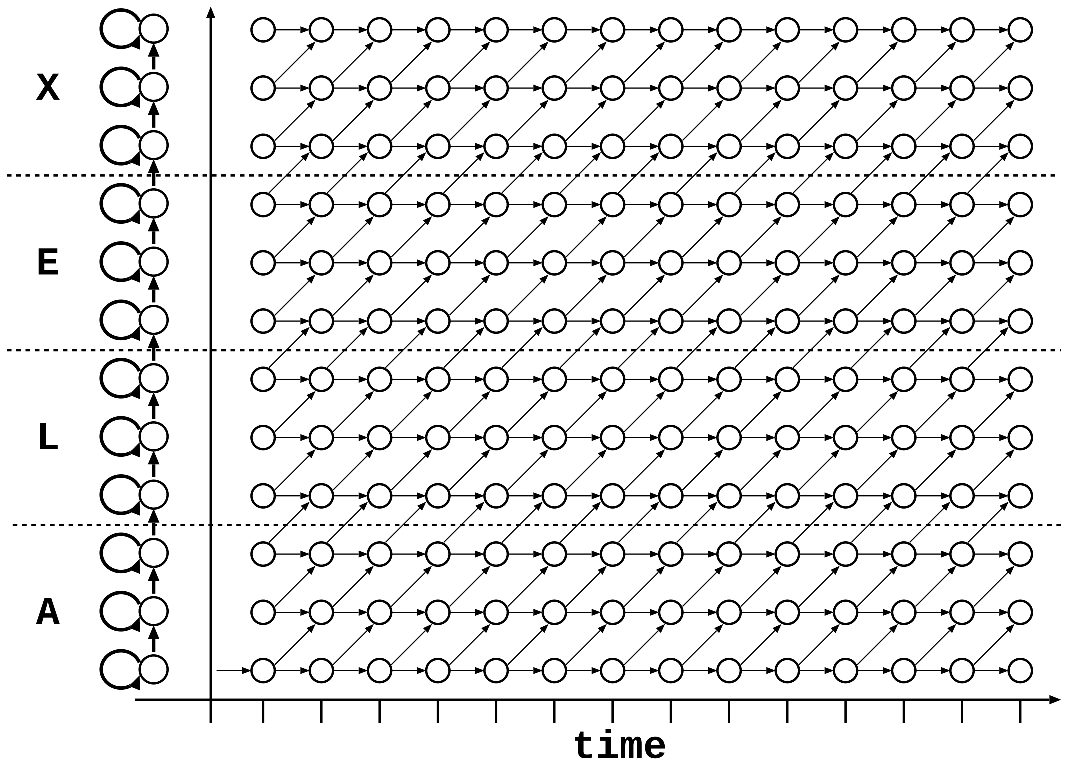
– separate memory vector c_t in addition to h_t

– use of gates to control information flow

– (additional) effect: make backpropagation more robust

Acoustic Modelling: HMM and ANN (CTC: similar [?])

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



QUAERO English Eval 2013
(competitive system)

Language Model	PP	Acoustic Model	WER[%]
Count Fourgram	131.2	Gaussian Mixture	19.2
		deep MLP	10.7
		LSTM-RNN	10.4
+ LSTM-RNN	92.0	Gaussian Mixture	16.5
		deep MLP	9.3
		LSTM-RNN	9.3

acoustic models:
– acoustic input features: optimized for model
– sequence discriminative training (MMI/MPE), not (yet) for LSTM-RNN
(end-to-end training)

remarks:
– overal improvements by ANNS: 50% relative (same amount of training data!)
– lion's share of improvement: acoustic model

- **why a separate language model?**

- **we need a model to approximate the true posterior distribution $p(w_1^N|x_1^T)$: separation of prior probability $p(w_1^N)$ of word sequence $w_1^N = w_1 ... w_n ... w_N$ in the posterior probability used in Bayes decision rule:**

$$p(w_1^N|x_1^T) = \frac{p(w_1^N) \cdot p(x_1^T|w_1^N)}{\sum_{\tilde{w}_1^{\tilde{N}},\tilde{N}} p(\tilde{w}_1^{\tilde{N}}) \cdot p(x_1^T|\tilde{w}_1^{\tilde{N}})}$$

  - **advantage: huge amounts of training data for $p(w_1^N)$ without annotation**
  - **extension: from generative to log-linear modelling**

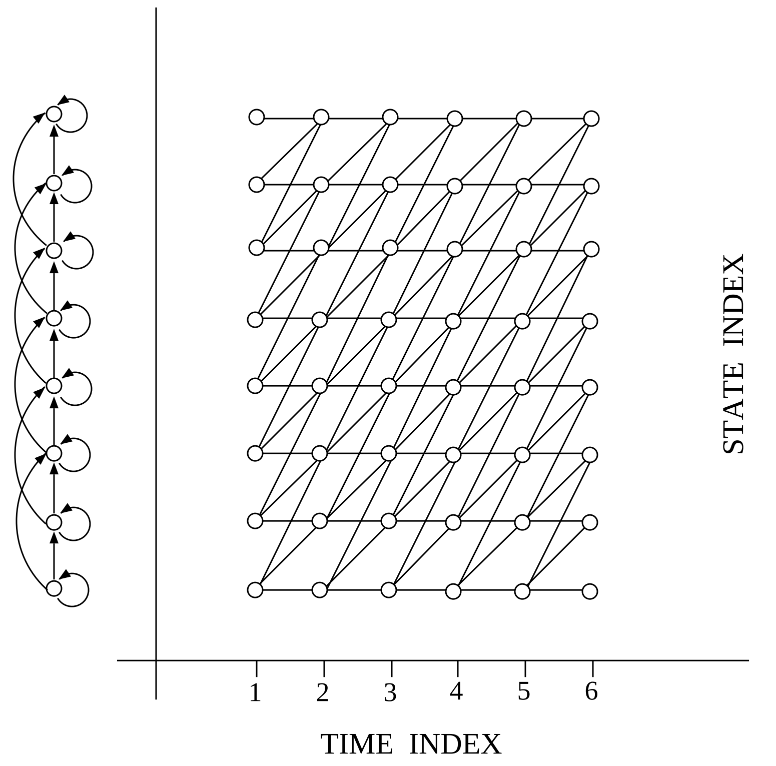
$$p(w_1^N|x_1^T) = \frac{q^{\alpha}(w_1^N) \cdot q^{\beta}(w_1^N|x_1^T)}{\sum_{\tilde{w}_1^{\tilde{N}},\tilde{N}} q^{\alpha}(\tilde{w}_1^{\tilde{N}}) \cdot q^{\beta}(\tilde{w}_1^{\tilde{N}}|x_1^T)}$$

- **note about prior $p(w_1^N)$ or $q(w_1^N)$: pure SYMBOLIC processing**

- **ANN: help here too!**

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)**

THE END