RWTH AACHEN UNIVERSITY

Interspeech 2021, Brno, Czech Republic, 30 Aug. – 03 Sep., 2021
Keynote

Forty Years of Speech and Language Processing: From Bayes Decision Rule to Deep Learning

Hermann Ney

RWTH Aachen University, Aachen, Germany

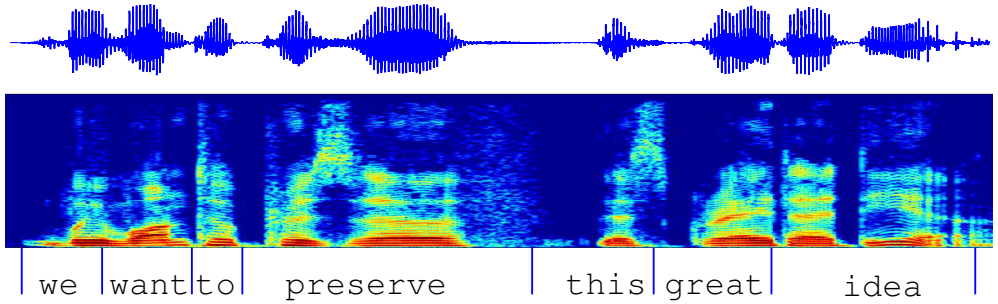
AppTek, Aachen, Germany & McLean, VA

My Presentation

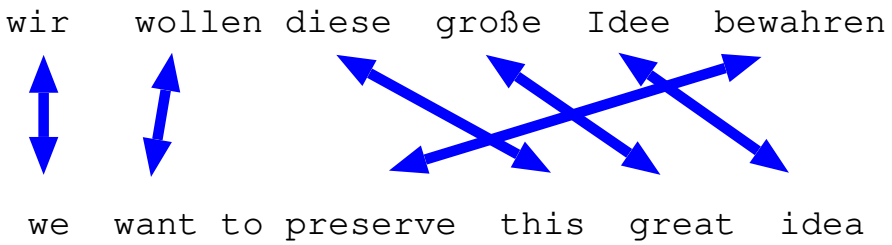
- my experience over the last 40 years:
  - share my interpretations and views (maybe controversial!)
  - mainly ASR and MT
- research history: many concepts
  - models: generative HMM, ANN and hybrid HMM, CTC, end-to-end, attention models, ...
  - training criteria: max.lik., MMI (and sMBR), cross-entropy/ANN, seq.disc.training, ...
- requirement/criterion of science:
  - what is the unifying concept and framework?
  - my answer: statistical/data-driven decision theory and Bayes decision rule
- deep learning:
  - yes, ... has been very successful
  - has a context of 30-year history (sometimes: wheel was re-invented!)
  - but there has been, is and will be life outside deep learning!
  - provides one out of many possible modelling structures

Tasks in Human Language Technology
Sequence-to-Sequence Processing

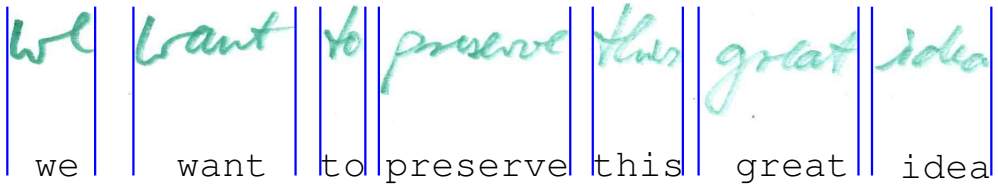
Automatic Speech Recognition (ASR)
(speech signal processing)



Machine Translation (MT)
(symbol or text processing)



Handwriting Recognition (HWR)
(image signal processing)



more tasks:
– sign language (gesture) recognition
– syntactic or semantic tagging (NLU)
– ...

Large-Scale Projects 1984-2020

- 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
  - partners: Daimler, Philips, Siemens, DFKI, KIT Karlsruhe, RWTH, U Stuttgart, ...
- TC-STAR 2004-2007: funded by EU
  - domain: recognition and translation of speeches given in EU parliament
  - task: speech translation: ASR + MT (+TTS)
  - challenge: MT robust wrt ASR errors $\rightarrow$ data-driven methods
  - first research prototype for unlimited domain and real-life data
    - fully automatic, not real time
    - without deep learning!
  - partners: KIT Karlsruhe, RWTH, CNRS Paris, UPC Barcelona, 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 Karlsruhe, RWTH**
- **BABEL 2012-2017: funded by US IARPA**
  - **recognition (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, ...)
- **EU ERC advanced grant 2017-2021:**
  emphasis on basic research for speech and language

ASR: first research 1975-1980

ASR is sequence-to-sequence processing:

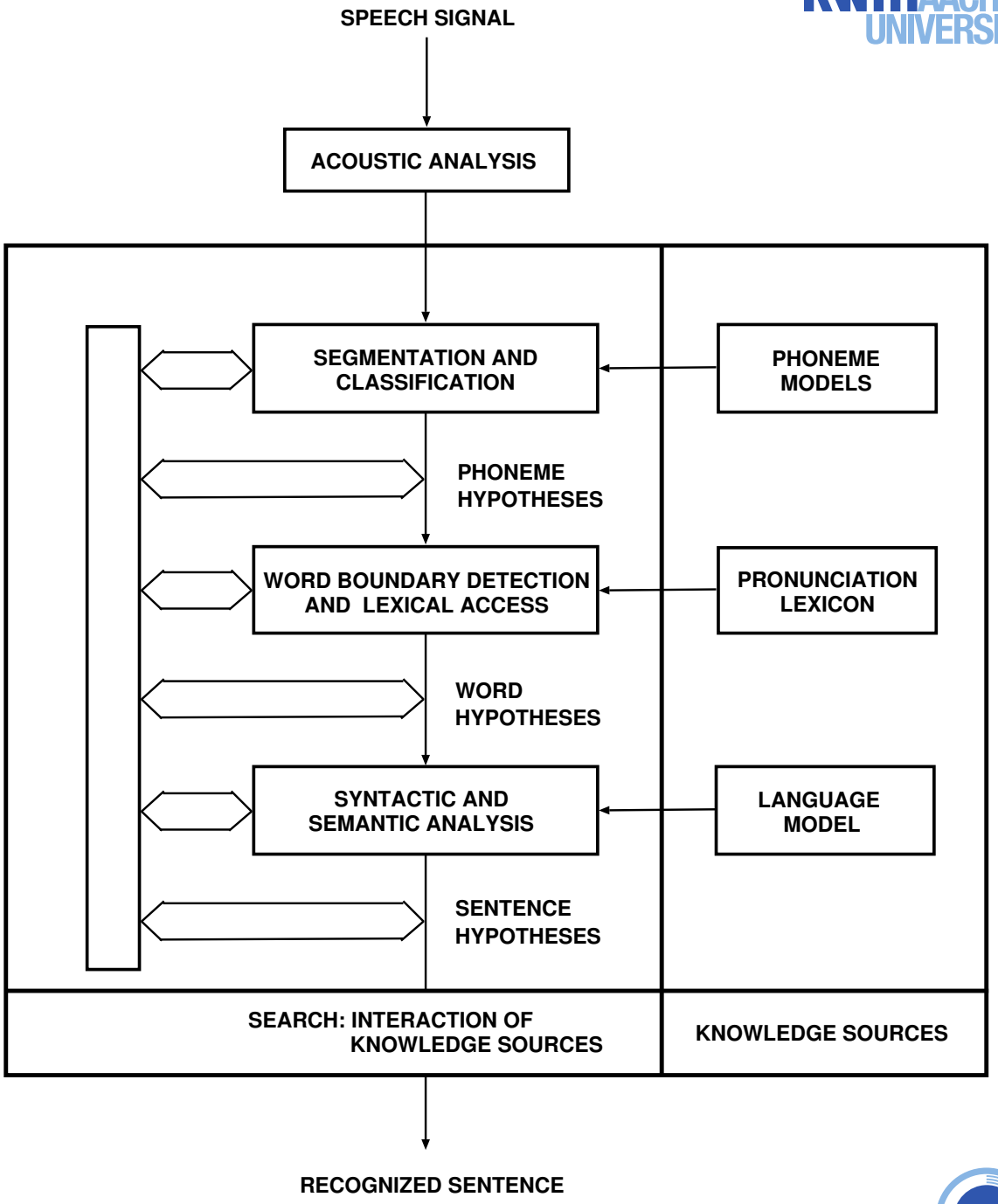
- **sequence of 10-ms acoustic vectors**
- **sequence of sounds/phonemes**
- **sequence of letters**
- **sequence of words**

problems:

- **ambiguities at all levels**
- **interdependencies of decisions**

approach 1975-1980 (Baker/CMU and Jelinek/IBM):

- **probabilitistic modelling**
- **holistic approach: single criterion and Bayes decision rule**



1985 Baseline Statistical Approach to ASR

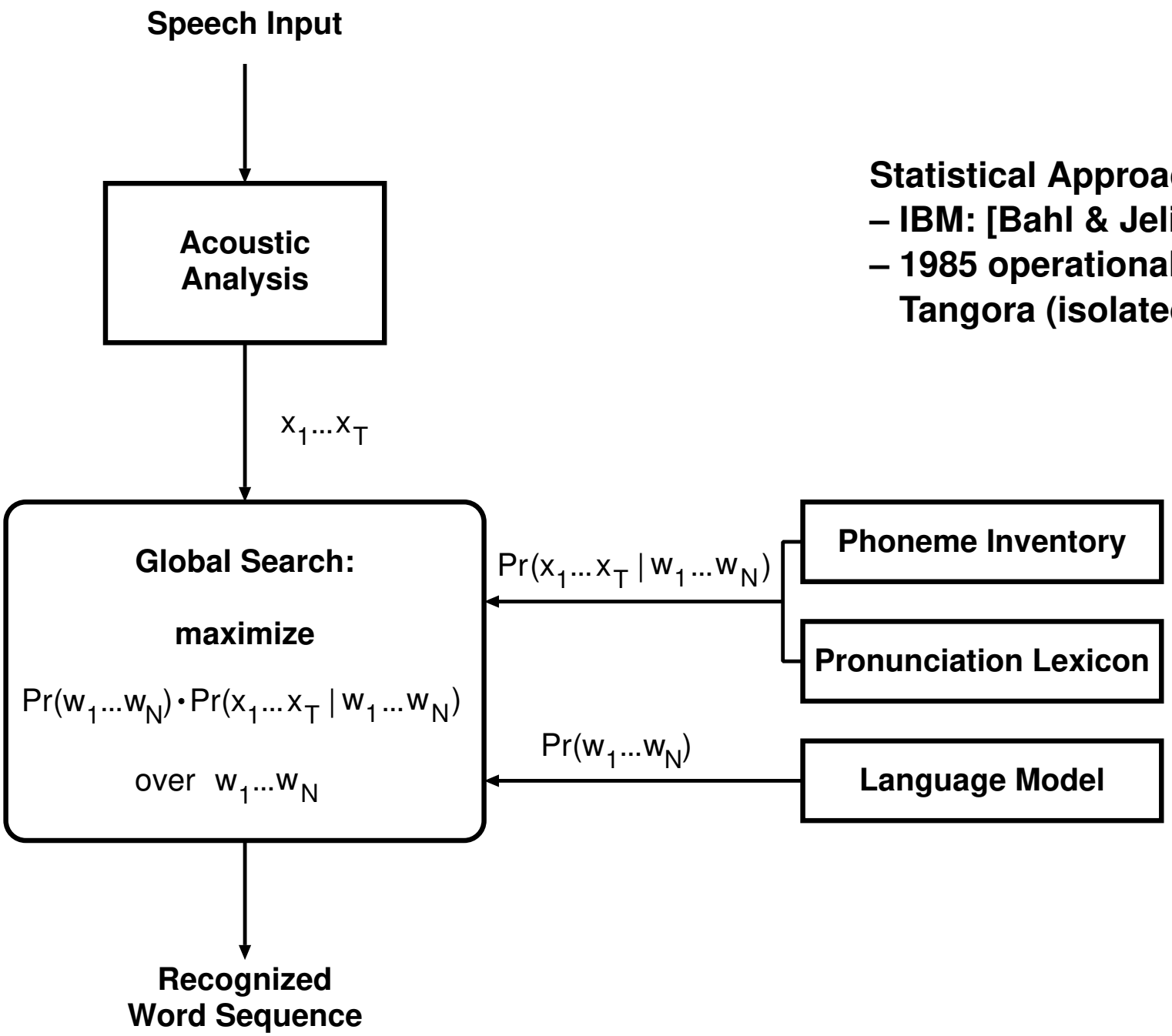
- modelling: probability distributions/statistical models with

$$\begin{aligned} \text{acoustic vectors:} \quad & x_1^T = x_1 ... x_t ... x_T \qquad x_t \in \mathbb{R}^D \\ \text{word strings:} \quad & w_1^N = w_1 ... w_n ... w_N \end{aligned}$$

- consider joint generative model:

$$p(w_1^N, x_1^T) = p(w_1^N) \cdot p(x_1^T | w_1^N)$$

- language model $p(w_1^N)$: word trigram model, learned from text only $[w_1^N]$
- acoustic (-phonetic) model $p(x_1^T | w_1^N)$: learned from annotated audio $[x_1^T, w_1^N]$
  (generative) hidden Markov model with: VQ, Gaussians, Gaussian mixtures, ...
  - structure: first-order dependence and mathematically nice
  - training: ('efficient') EM algorithm with sort of closed-form solutions
    but difficult from a maching learning point-of-view
- decoding/generation: Bayes decision rule (simplified form)
  = use single criterion and avoid local decisions



Statistical Approach to ASR
– IBM: [Bahl & Jelinek+ 83]
– 1985 operational research system: Tangora (isolated words, speaker dep.)

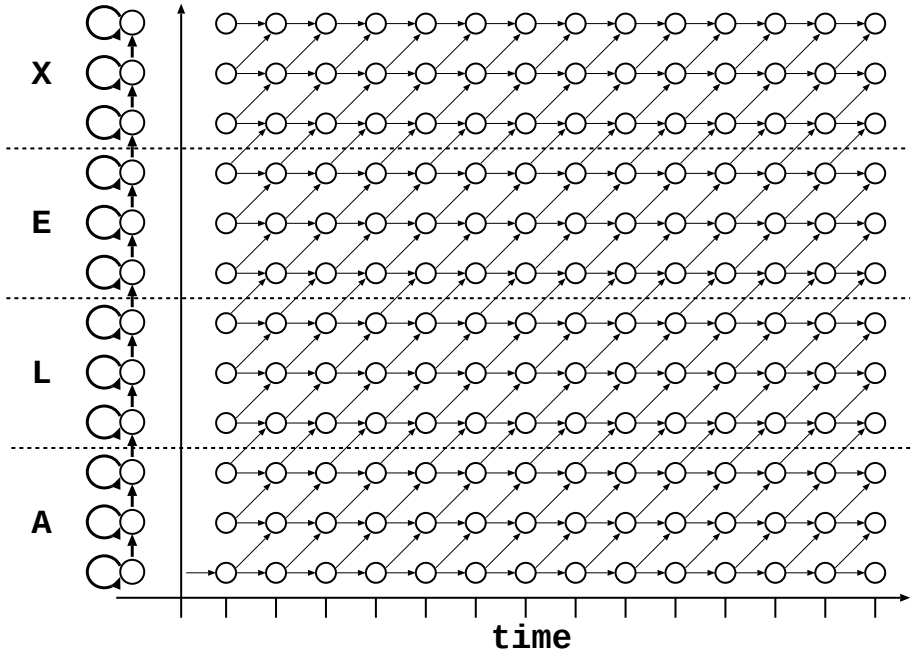
ASR History: Operational Research Systems

- steady increase of challenges:
  - vocabulary size: 10 digits ... 1000 ... 5.000 ... 20.000 ... 500.000 words
  - speaking style: isolated words ... read/colloquial/spontaneous speech
- steady improvement of statistical methods:
  HMM, Gaussians and mixtures, statistical trigram language model,
  adaptation methods, hybrid HMM (ANN), ...
- 1985-93: criticism about statistical approach/machine learning
  - too many parameters and saturation effect
  - concept of rule-based approach: acoustic-phonetic expert systems
    problem: coverage and consistency of rules
- important methodology in ASR since 1990:
  standard public data (training and testing)
  (..., TIMIT, RM/1k, WSJ/5k, WSJ/20k, NAB/64k, ..., Switchboard/tel., ...)
- 1993-2000 NIST/DARPA: comparative evaluation of operational systems:
  - virtually all systems: generative HMMs and refinements
  - 1994 Robinson: hybrid HMM with RNN (singularity!)
- 2009 ICDAR handwriting competition: CTC by Graves
- 2012 and later: mainstream = ANNs and deep learning

Finite-State Models: Hidden Markov Model (HMM)

– sequence of acoustic vectors:
$X = x_1^T = x_1 ... x_t ... x_T$ over time t

– sequence of states $s = 1, ..., S$
$s_1^T = s_1 ... s_t ... s_T$ over time t
with CART (phonetic) labels:
$a_1^S = a_1 ... a_s ... a_S$
$= W$: word sequence



HMM: time alignment between input and output

- classical HMM: generative model for x_1^T:

$$q_\vartheta(x_1^T | W = a_1^S) = \sum_{s_1^T} \prod_t q(s_t | s_{t-1}, W, \vartheta) \cdot q_t(x_t | a_{s=s_t}, \vartheta)$$

- hybrid HMM: model of label posterior sequence a_1^S:

$$q_\vartheta(W = a_1^S | x_1^T) = \sum_{s_1^T} \prod_t q(s_t | s_{t-1}, W, \vartheta) \cdot q_t(a_{s=s_t} | x_1^T, \vartheta)$$

[Bourlard & Wellekens 89] machine learning point-of-view:
it is much(!) easier to model $q_t(a_s | x_1^T, \vartheta)$ than $q_t(x_t | a_s, \vartheta)$

ANN: probabisitic interpretation:

- **ANN outputs [Bourlard & Wellekens 89]: class posteriors**
- **softmax [Bridle 89]: softmax = posterior of (class prior + Gaussian) (assuming class-independent covariance matrix)**

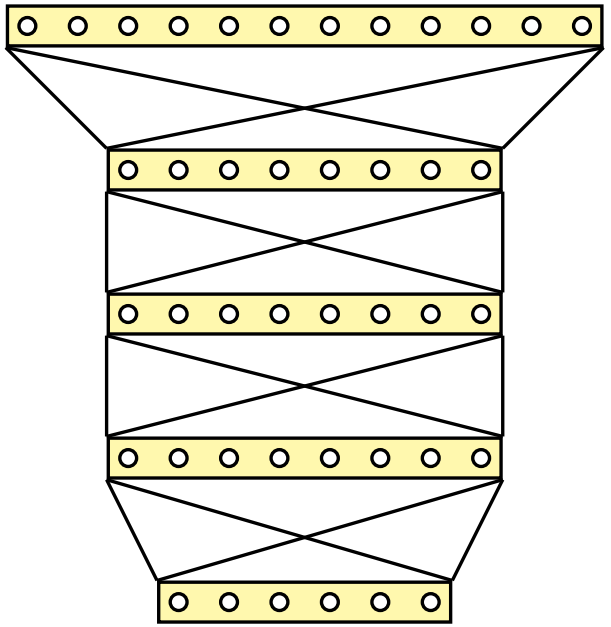
interpretation:

ANN with softmax = posterior of (class prior + Gaussian) + feature extraction

- **hidden layers perform feature extraction:**

$$z \rightarrow x = f(z)$$

with feature vector $x \in \mathbb{R}^D$ before output layer
note: no dependence on class labels $c = 1, ..., C$

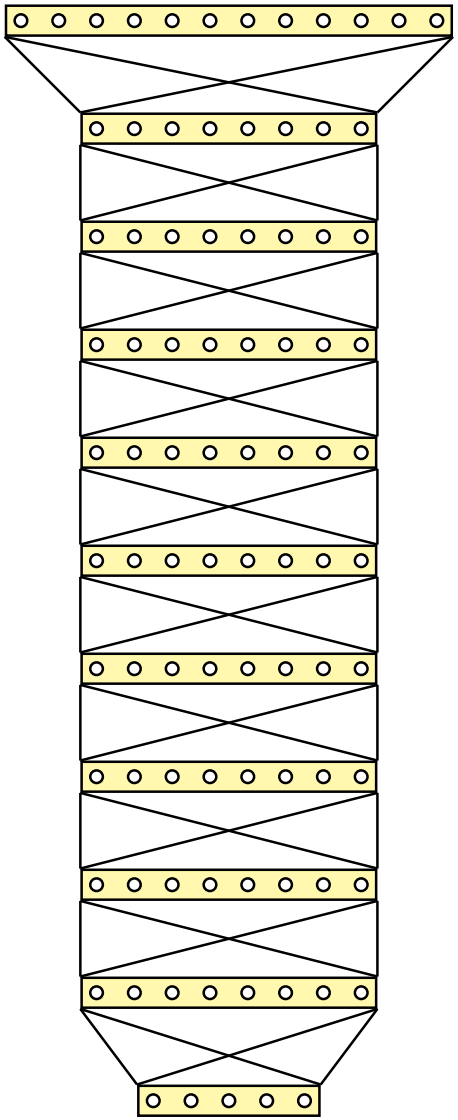


- **output layer: probability distribution over classes c**

$$p(c|x) = \frac{\exp(\alpha_c + \lambda_c^t \cdot x)}{\sum_{c'} \exp(\alpha_{c'} + \lambda_{c'}^t \cdot x)}$$

with output layer weights $\lambda_c \in \mathbb{R}^D$
and offsets (biases) $\alpha_c \in \mathbb{R}$

Artificial Neural Networks (ANN) and Deep Learning:

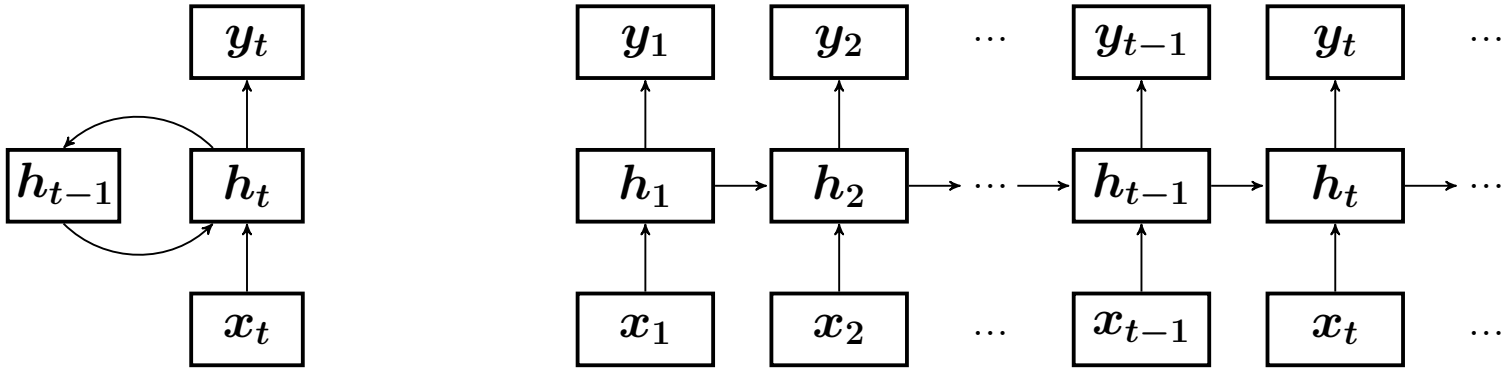


question: what is different now after 30 years?

answer: we have learned how to (better) handle a complex mathematical optimization problem:

- **more powerful hardware (e. g. GPUs)**
- **empirical recipies for optimization: practical experience and heuristics, e.g. layer-by-layer pretraining**
- **result: we are able to handle more complex architectures (deep MLP, RNN, etc.)**

ASR: sequence-to-sequence processing

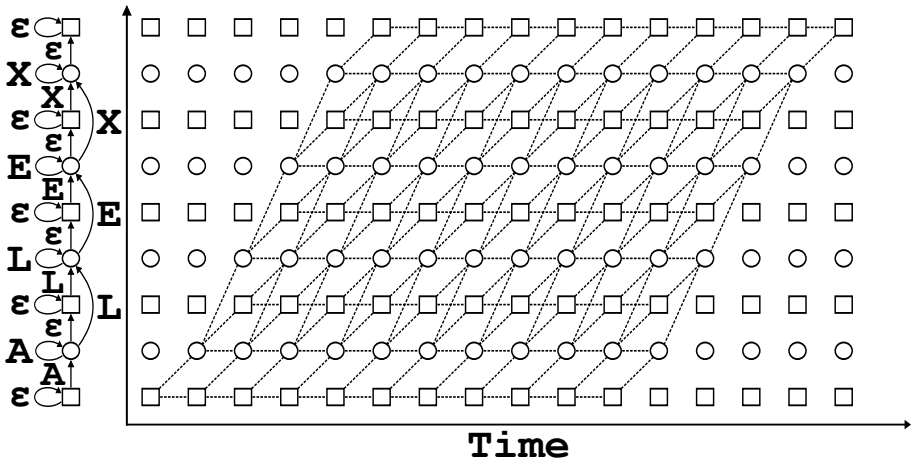


from simple ANN to RNN:
– introduce a memory (or context) component to keep track of history
– result: two types of input at time t: memory h_{t-1} and observation x_t

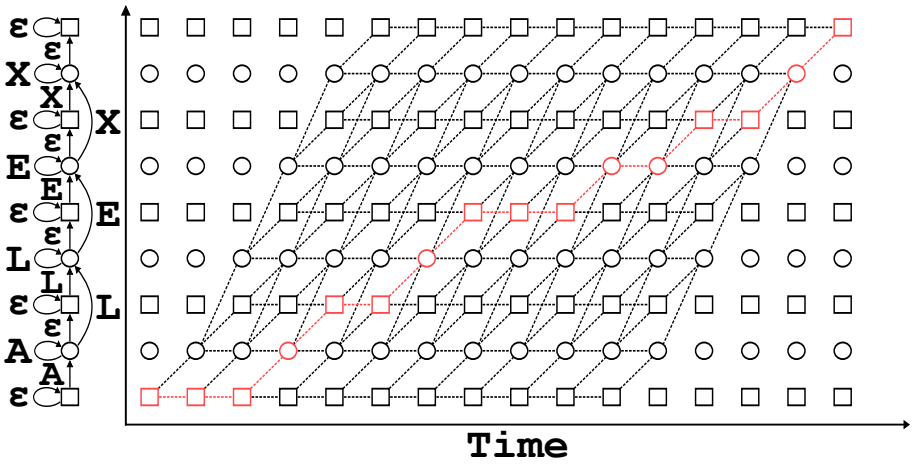
extensions:

- **(succesful!) application to ASR: [Robinson 94]**
- **bidirectional structure [Schuster & Paliwal 97]**
- **LSTM: long short-term memory [Hochreiter & Schmidhuber 97, Gers & Schraudolph$^+$ 02]**

From Hybrid HMM to CTC: Connectionist Temporal Classification



sum over all paths



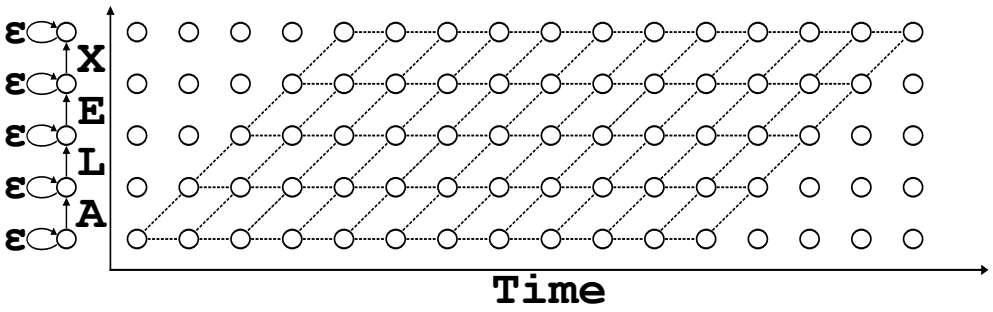
Viterbi: single best path

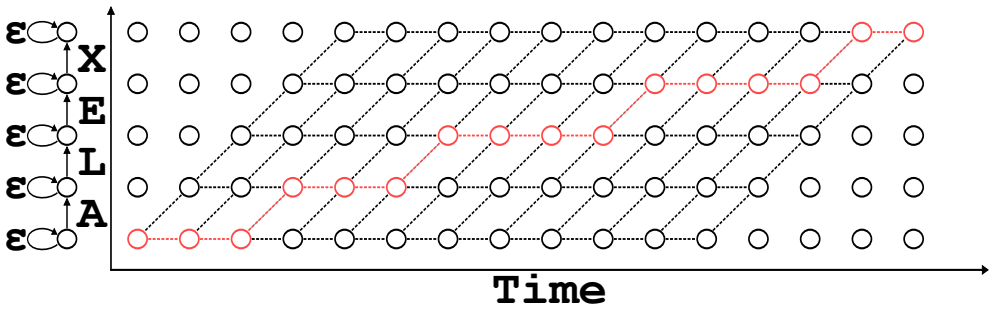
- output labels: letters or monophone labels (rather than CART labels)
  with constraints on (transition, label) pairs
- modify hybrid HMM framework:
  – no transition probabilities
  – special symbol ϵ: empty/ garbage/blank symbol, don't care symbol, repetition
- resulting model with frame label y_t for state label a_s:

$$q_\vartheta(W = a_1^S | x_1^T) = \sum_{s_1^T} \prod_t q_\vartheta(y_t = a_{s_t} | x_1^T) \qquad \text{with} \qquad \sum_{y_t \in \{a\} \cup \epsilon} q_\vartheta(y_t | x_1^T) = 1$$

- analysis (mathematical and experimental):
  the optimum solution tend to produce a maximum number of ϵ

RNN-T: RNN Transducer





from CTC to RNN-T:

– each true symbols (i. e. other than ϵ) occurs exactly once:
assigned to forward transitions only

– dependence on output context:

full context: $q_\vartheta(y_t = a_s | a_0^{s-1}, x_1^T)$

k predecessor symbols: $q_\vartheta(y_t = a_s | a_{s-k-1}^{s-1}, x_1^T)$

special case: $k = 1$ **or** 0

terminology:

– general case: RNN-T: could have vertical transitions [Graves 12]

– RNN Aligner (RNN-A): no vertical transitions [Sak & Shannon$^+$ 17]

History: ANN in Acoustic Modelling

- 1988 [Waibel & Hanazawa$^+$ 88]:
  phoneme recognition using time-delay neural networks (convolutional NNs!)
- 1989 [Bridle 89]:
  softmax operation ('Gaussian posterior') for normalization of ANN outputs
- 1989 [Bourlard & Wellekens 89]:
  – ANN outputs: can be interpreted as class posteriors
  – hybrid HMM: use frame label posteriors
- 1991 [Bridle & Dodd 91] backpropagation for HMM discriminative training at word level
- 1993 [Haffner 93]: sum over label-sequence posterior probabilities in hybrid HMMs
  (*sequence discriminative training*)
- 1994 [Robinson 94]: recurrent neural network in hybrid HMM
  (operational system, DARPA evaluations)
- 1997 [Fontaine & Ris$^+$ 97, Hermansky & Ellis$^+$ 00]:
  tandem HMM: ANN for feature extraction in a Gaussian HMM
- 2009 Graves: CTC for handwriting recognition
  (operational system, ICDAR competition 2009)
- 2012 and later: mainstream of hybrid HMM (or similar) and deep learning
- 2015 [Bahdanau & Cho$^+$ 15] attention for MT

History ASR: Tandem vs. Hybrid Approach

RWTH AACHEN UNIVERSITY

tandem approach: ANN-based explicit feature extraction + Gaussian HMM

- **2000 [Hermansky & Ellis$^+$ 00]: multiple layers of processing by combining Gaussian model and ANN for ASR**
- **2006 [Stolcke & Grezl$^+$ 06]: cross-domain and cross-language portability**
- **2007 [Valente & Vepa$^+$ 07]: 8% WER reduction on LVCSR**
- **2011 [Tüske & Plahl$^+$ 11]: 22% WER reduction on LVCSR**

hybrid approaches:

- **2009 [Graves 09]: CTC - good results on LSTM RNN for handwriting task**
- **2010 [Dahl & Ranzato$^+$ 10]: improvement in phone recognition on TIMIT**
- **2011 [Seide & Li$^+$ 11, Dahl & Yu$^+$ 12]: Microsoft Research**
  - **– fully-fledged hybrid approach**
  - **– 30% WER reduction on Switchboard 300h**
- **since 2012: other teams confirmed reductions of WER by 20% to 30%**

comparison: hybrid vs. tandem approach:

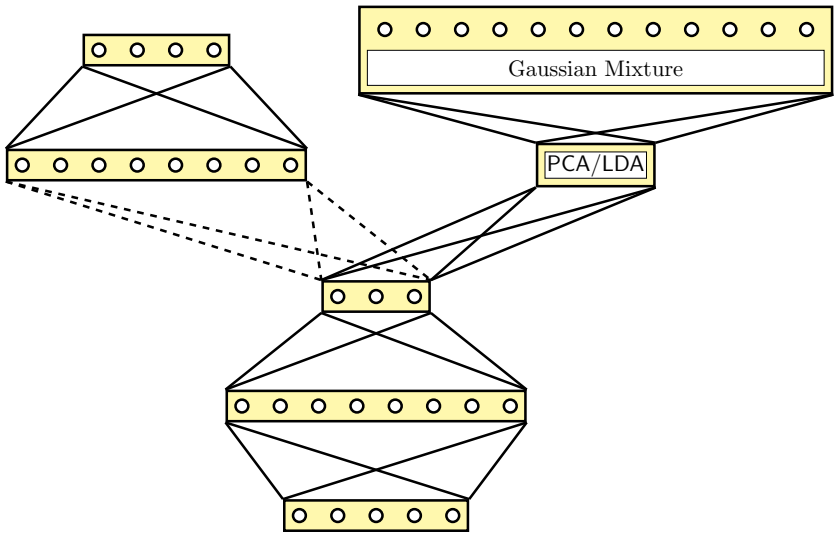
- **– hybrid approach: more monolithic and compact**
- **– the same structure in training and testing**
- **– widely used nowadays**

HLT

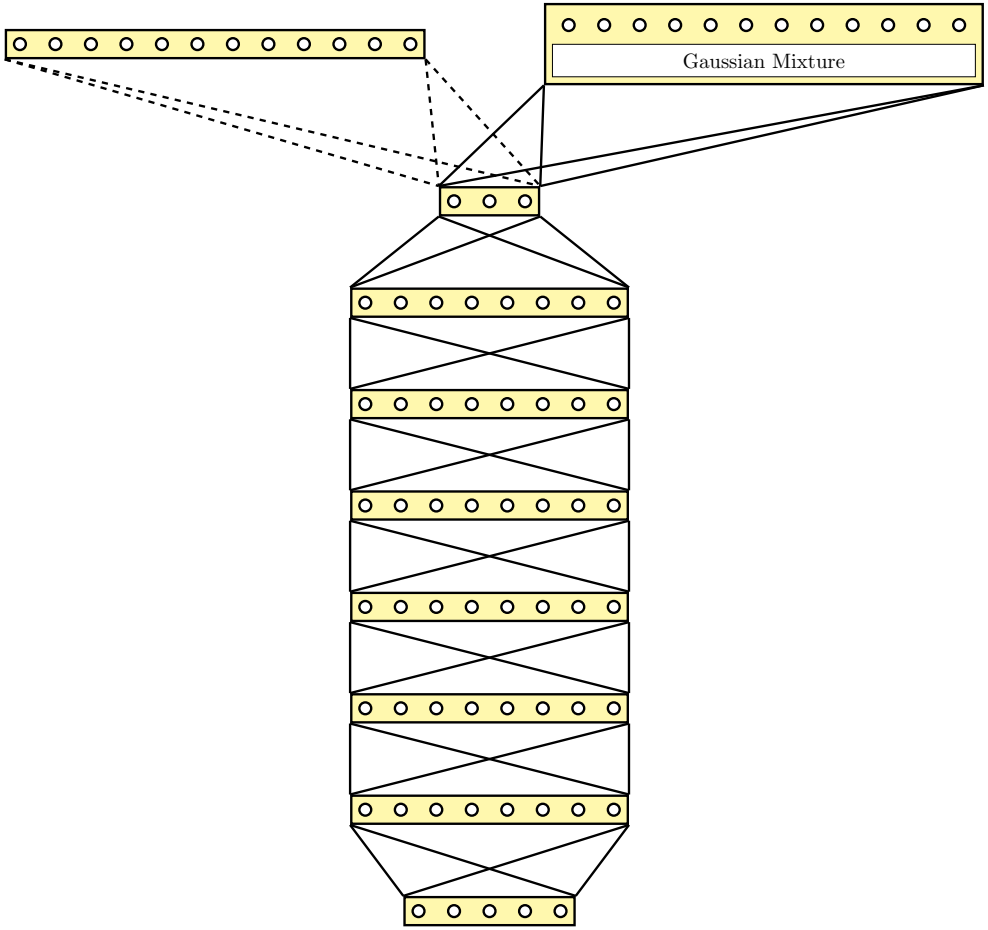
Tandem Approach: Explicit Feature Extraction

- **tandem approach: two parts:**
  MLP for feature extraction + generative HMM

  [Fontaine & Ris$^+$ 97, Hermansky & Ellis$^+$ 00]
- **extensions, e. g. bottleneck concept**
  [Stolcke & Grezl$^+$ 06, Grezl & Fousek 08],
  [Valente & Vepa$^+$ 07, Tüske & Plahl$^+$ 11]



RWTH's Tandem Structure
[Tüske & Plahl$^+$ 11]



Unifying View: Seq-to-Seq Processing for HLT

- modelling of varying structures:
  generative HMM, neural finite-state models (hybrid HMM, CTC, RNN-T), attention, ...
  – goal: to capture inherent dependences in the strings
    using task-specific knowledge about speech and language
  – task comparable to physics (or natural sciences):
    to capture dependencies in nature
- training criteria: a large set of confusing concepts:
  – maximum likelihood and EM algorithm
  – cross-entropy
  – sum criterion (loss function in CTC)
  – MMI and sequence discriminative training
  – (so-called) minimum Bayes risk
    (with variants: sMBR, Povey's phoneme/word error rate, ...)

questions:
– how can we organize this (chaotic) set of training criteria?
– what is the unifying concept?

my view: we should use the general principles formulated already in 1970s (or before):
example: textbook by Duda & Hart 1973, pp. 11-16
not explicitly meant for ASR or string processing

Bayes Decision Theory: Puristic Mathematical View (independent of practical aspects)

- concept: imagine a "huge huge" database of (input,output) string pairs (x, c) along with empirical distribution (with no parameters!:

$$[x_r, c_r], \quad r = 1, ...R \qquad\qquad pr(x, c) := 1/R \cdot \sum_{r=1}^{R} \delta(x, x_r)\delta(c, c_r)$$

- guessing game: guess c from knowing x: $\quad x \rightarrow c = \hat{c}(x)$
  terminology: classify the input data or generate the output data

- perfect solution is not possible:
  – we want to convert a relation $[x, c]$ into a function $x \rightarrow c = \hat{c}(x)$
  – need for an error measure or loss function $L[c, \hat{c}(x)]$ for each pair $[x, c]$
  popular error measures for strings:
  – in general: 0/1 loss function = error count
  – ASR/HWR: WER = word/symbol error rate/count defined by edit distance
  – MT: TER = translation error rate = edit distance + swap count

- key question: how to generate the output string?
  best compromise: for each input x (which might exist in several pairs $[x = x_r, c_r]$!), select an output string that minimizes the expected loss:

$$x \rightarrow c_*(x) \;:=\; \arg\min_{c} \Big\{ \sum_{\tilde{c}} pr(\tilde{c}|x) \cdot L[\tilde{c}, c] \Big\}$$

  using the posterior distribution $pr(c|x)$ of the data

Decision Rule: From True Bayes to Pseudo Bayes

- ***true*** **Bayes decision rule: "closed world":**
  no generalization for input x beyond training data
- ***pseudo*** **Bayes decision rule: generalization to unseen test data:**

  replace $pr(c|x)$ by a model $p_\vartheta(c|x)$

  with a (huge) parameter set ϑ to be learned from data

- **result: *pseudo* Bayes decision rule using a model $p_\vartheta(c|x)$:**

$$\textbf{general loss:} \quad x \rightarrow c_\vartheta(x) := \arg\min_c \left\{ \sum_{\tilde{c}} p_\vartheta(\tilde{c}|x) \cdot L[\tilde{c}, c] \right\}$$

$$\textbf{0/1 loss:} \quad x \rightarrow c_\vartheta(x) := \arg\max_c \left\{ p_\vartheta(c|x) \right\}$$

  textbook: widely used, optimal only for string error (= 0/1 loss)

- **two principal questions:**
  – how much does the exact loss function matter ?
  – what are suitable training criteria for learning $p_\vartheta(c|x)$?

  independent practical aspect: computational complexity (of both questions)

Bayes Decision Rule: Does the exact form of the loss function matter?

- two forms of (pseudo) Bayes decision rule:

$$\text{general loss:} \quad x \rightarrow c_\vartheta(x) := \arg\min_c \left\{ \sum_{\tilde{c}} p_\vartheta(\tilde{c}|x) \cdot L[\tilde{c}, c] \right\}$$

$$\text{0/1 loss:} \quad x \rightarrow c_\vartheta(x) := \arg\max_c \left\{ p_\vartheta(c|x) \right\}$$

- mathematical equivalence of the two rules
  [Schlüter & Scharrenbach$^+$ 05, Schlüter & Nussbaum$^+$ 11]:
  – conditions: a metric loss function $L[\tilde{c}, c]$ and

$$\max_c p_\vartheta(c|x) \geq 0.5$$

  – theoretical refinements beyond the threshold of 0.5
  – experimental results: hard to find a difference
    e. g. for high error rates: from 41% to 40%

- special case for edit distance: improvements beyond 0/1 loss
  by position-dependent symbol posterior probabilities
  [Xu & Povey$^+$ 10, Schlüter & Nussbaum$^+$ 11]

True vs. Pseudo Bayes Decision Rule: Training Criteria

considerations:

– **use annotated string pairs** $[x_r, c_r],\ r = 1, ..., R$ **for training**

– **heuristic concept:** $p_\vartheta(c|x) \rightarrow pr(c|x)$
exact distance measure? relation to associated classification error?

mathematical analysis for string error (0/1 loss) [Ney 03]:

- **empirical (=true) distributions** $pr(c, x)$ **and** $pr(c|x)$**:**
  $E_* =$ **true Bayes classification error: absolute minimum using**

  true Bayes rule: $x \rightarrow \hat{c}_\vartheta(x) = \text{argmax}_c\{pr(c|x)\}$

- **probability model** $p_\vartheta(c|x)$ **with set of parameters** ϑ**:**
  $E_\vartheta =$ **model-based classification error using:**

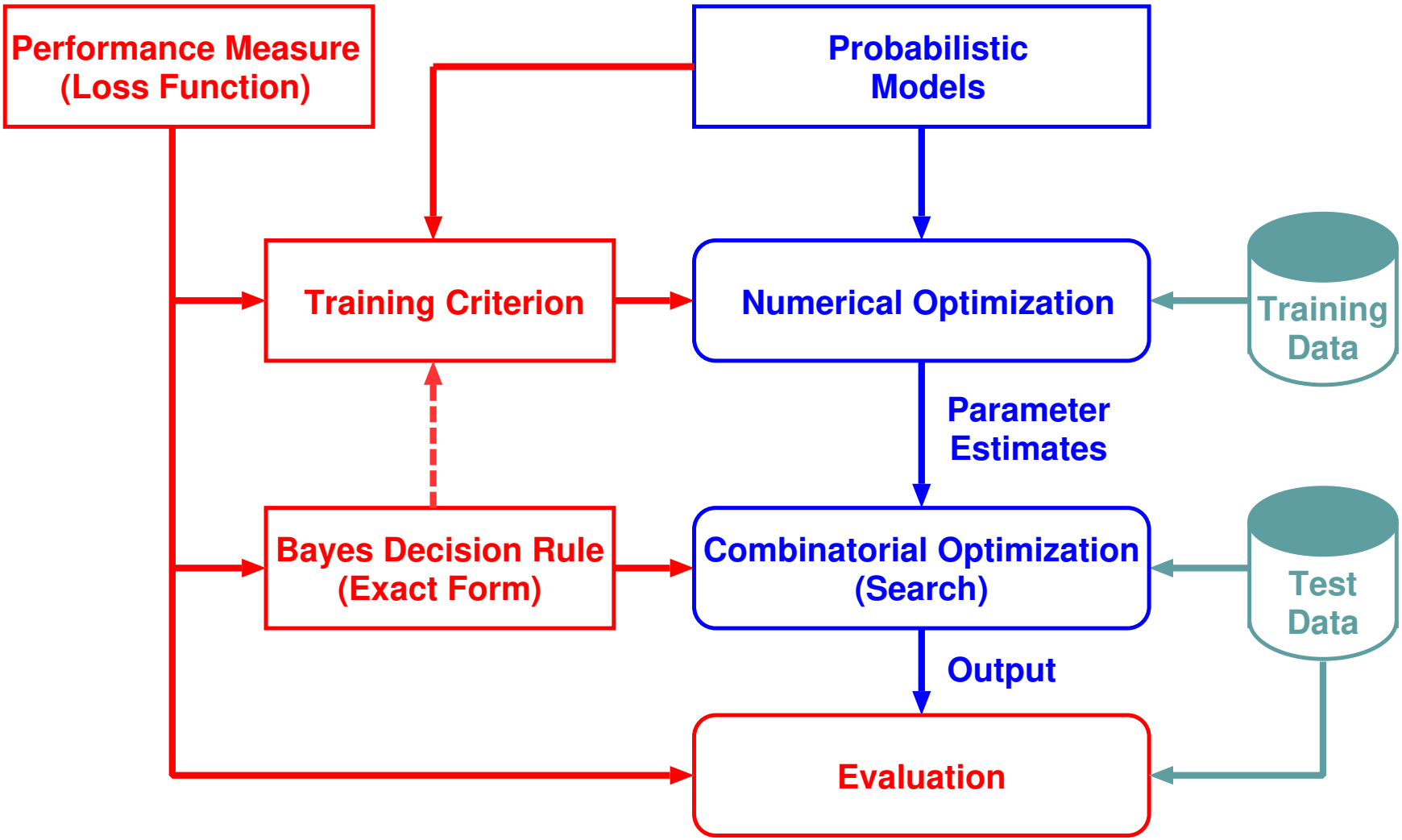
  pseudo Bayes rule: $x \rightarrow \hat{c}_\vartheta(x) = \text{argmax}_c\{p_\vartheta(c|x)\}$

upper bound on the squared difference between these errors
(= Kullback-Leibler divergence or relative entropy):

$$1/2 \cdot \left[E_* - E_\vartheta\right]^2 \leq \sum_x pr(x) \sum_c pr(c|x) \log \frac{pr(c|x)}{p_\vartheta(c|x)} = \frac{1}{R} \sum_{r=1}^{R} \log \frac{pr(c_r|x_r)}{p_\vartheta(c_r|x_r)}$$

criterion: minimize this upper bound $\rightarrow$ **cross-entropy criterion**
(other upper bounds: binary divergence and squared error)

Machine Learning: Statistical Decision Theory



principal ingredients:

- **performance measure, error measure, loss function:**
  – how to judge the quality of the system output
  – examples: ASR: edit distance; MT: TER or BLEU
- **probabilistic models (with a suitable structure)**
  for capturing the dependencies within and between input and output strings:
  – given synchronization: Markov chain, CRF, (LSTM) RNN, ...
  – input/output synchronization: generative/hybrid HMM, CTC, RNN-T, attention models, ...
- **training criterion:**
  to learn the free model parameters from examples
  – ideally should be linked to performance criterion
  – two open questions: exact form of criterion? optimization strategy?
- **Bayes decision rule: decoder/search**
  for generating the output word sequence
  – combinatorial problem (efficient algorithms)
  – should exploit structure of models
  examples: dynamic programming and beam search, A* and heuristic search, ...
  [Vintsyuk 68, Velichko & Zagoruyko 70, Vintsyuk 71] and [Sakoe & Chiba 71],
  [DRAGON/HARPY 1975] and [Bridle 82, Ney 84, Ney & Haeb-Umbach$^+$ 92]

ASR Modelling: String Posterior Probability

- complete complete model for [input,output] pair $[x_1^T, W = a_1^S]$
  consists of language model (LM) and acoustic (-phonetic) model (AM):

$$p_\vartheta(W|x_1^T) := \frac{q_\vartheta^\alpha(W) \cdot q_\vartheta^\beta(W = a_1^S|x_1^T)}{\sum_{\tilde{W}} q_\vartheta^\alpha(\tilde{W}) \cdot q_\vartheta^\beta(\tilde{W} = \tilde{a}_1^S|x_1^T)}$$

  parameter set: $\vartheta := \{\vartheta_A, \vartheta_L\}$ with ϑ_L for LM and ϑ_A for AM (and α, β)

- language model $q_\vartheta(W)$
  – prior that scores the syntactic-semantic adequacy of a word sequence
  – idea: can be learned from text data only (e. g. 100 million words)
  – no manual annotation required!

- acoustic model (AM): finite-state model (hybrid HMM, CTC, RNN-T):

$$q_\vartheta(W = a_1^S|x_1^T) = \sum_{s_1^T} \prod_t q(s_t|s_{t-1}, W, \vartheta) \cdot q_t(a_{s=s_t}|x_1^T, \vartheta)$$

  learned from manually transcribed audio data (e. g. 500 hours = 5 million words)

- remark: the log-linear combination mimicks the generative approch:

$$p_\vartheta(W|x_1^T) := \frac{p_\vartheta(x_1^T, W)}{\sum_{\tilde{W}} p_\vartheta(x_1^T, \tilde{W})} = \frac{p_\vartheta(W) \cdot p_\vartheta(x_1^T|W)}{\sum_{\tilde{W}} p_\vartheta(\tilde{W}) \cdot p_\vartheta(x_1^T|\tilde{W})}$$

Acoustic Model: Training Criterion and Procedure

suitable training criterion for [audio, word string] data $[X_r, W_r],\ r = 1, ..., R$:

$$\max_{\vartheta} \left\{ \sum_r \log\ p_\vartheta(W_r|X_r) \right\} \qquad p_\vartheta(W|X) = \frac{q^\alpha(W) \cdot q_\vartheta^\beta(W|X)}{\sum_{\tilde{W}} q^\alpha(\tilde{W}) \cdot q_\vartheta^\beta(\tilde{W}|X)}$$

terminology: *sequence discriminative training* (IBM 1986: MMI)

- **baseline approximation: ignore denominator**
  - **neural finite-state models (FSM): $q_\vartheta(W|X)$**
  - **variants: full sum or best path (frame-wise CE, Viterbi)**
    note: EM framework still works for neural FSM
  - **result: no effect of language model on the training of the acoustic model**
- **better approximation: approximate sum in denominator**
  - **word hypothesis lattice**
  - **simplifed language model (lattice-free MMI, Povey 2016)**
  result: LM affects training of AM!
- **improved training criteria (symbol error/WER in lieu of string error):**
  minimum word/phoneme error rate, state-level minimum Bayes risk (sMBR), ...

history: Bahl/IBM 1986, Normandin 1991, Valtchev 1996, Povey 2002/16, Heigold 2005/12

ASR: Sequence Discriminative Training
End-to-End Concept

reconsider training criterion for (audio,text) pairs $[X_r, W_r],\ r = 1, ..., R$:

$$\max_{\vartheta} \left\{ \sum_r \log\ p_\vartheta(W_r|X_r) \right\} \qquad p_\vartheta(W|X) := \frac{q^\alpha(W) \cdot q_\vartheta^\beta(W|X)}{\sum_{\tilde{W}} q^\alpha(\tilde{W}) \cdot q_\vartheta^\beta(\tilde{W}|X)}$$

terminology: What does *end-to-end* mean?

- **training criterion: a single global criterion for optimum performance, independent of model structure**
- **monolithic structure of a model: simplicity/elegance of programming? what about adequacy/performance?**

remarks:

- **ASR: training of acoustic model and language model:**
  - **– transcribed audio: 500 hours = 5 million words**
  - **– text (from press, books, internet,...): 100 million words and more**
- ***end-to-end* concept:**
  - **– for the training criterion: yes**
  - **– for the structure: can it reflect the training data situation?**

Systematic Experiments on QUAERO English Eval 2013 (Tüske et al. RWTH 2017)

RWTH AACHEN UNIVERSITY

QUAERO task: broadcast news/conversations, podcasts, TED lectures

Word error rates [%] on QUAERO English Eval 2013
(note: acoustic input features were optimized for acoustic model):

Acoustic Model: hybrid HMM			Language Model	
Model		Criterion	Count	Count+ANN
			PP=131.1	PP=92.0
Gaussian Mixtures		Max.Lik.	20.7	
		seq.disc. training	19.2	16.1
Neural Net	FF MLP	frame-wise CE	11.6	
		seq.disc. training	10.7	9.0
	LSTM RNN	frame-wise CE	10.6	
		seq.disc. training	9.8	8.2

observations:
– improvements by acoustic ANNs: 50% relative
– improvement by language model ANN: 15% relative
– total improvements by deep learning: 60% relative (from 19.2% to 8.2%)

HLT

Tasks: Switchboard and Call Home

- conversational speech: telephone speech, narrow band;
  challenging task: initial WER: 60% (and higher) on Switchboard

- training data for acoustic model: Switchboard corpus
  - about 300 hours of speech
  - about 2400 two-sided recordings with an average of 200 seconds
  - 543 speakers
- test set Hub5'00
  - 20 telephone recordings form Switchboard studies (SWB)
  - 20 telephone conversations from Call-Home US English Speech (CHM)
  - total: 3.5 hours of speech
- training data for language model
  - vocabulary size fixed to 30k
  - Switchboard corpus: 2.9M running words
  - Fisher corpus: 21M running words

ASR: Results (April 2019) on Switchboard & Call Home

RWTH AACHEN UNIVERSITY

baseline models:
- **language model: 4-gram count model**
- **acoustic model: hybrid HMM with CART (allophonic) labels: LSTM bi-RNN with frame-wise cross-entropy training**
- **speaker/channel adaptation: i-vector [Dehak & Kenny+ 11]**
- **affine transformation [Gemello & Manai+ 06, Miao & Metze 15]**

word error rates [%]:

adaptation	methods	SWB	CHM	average
no	baseline approach	9.7	19.1	14.4
	+ seq. discr. training (sMBR)	9.6	18.3	13.9
	+ LSTM-RNN language model	7.7	15.8	11.7
yes (i-vector)	baseline approach	9.0	18.0	13.5
	+ seq. discr. training (sMBR)	8.4	17.2	12.8
	+ LSTM-RNN language model	6.8	15.1	10.9
+ adaptation by affine transformation		6.7	13.5	10.2

overall improvements over baseline:
- **33% relative reduction in WER**
- **by seq. discr. training, LSTM-RNN language model and adaptation**

HLT

Best Results on Call Home (CHM) and Switchboard (SWB) (best word error rates [%] reported)

team	CHM	SWB	training data, remarks
Johns Hopkins U 2017	18.1	9.0	300h, no ANN-LM, single model, data perturbation
Microsoft 2017	17.7	8.2	300h, ResNet, with ANN-LM
ITMO U 2016	16.0	7.8	300h, with ANN-LM, model comb., data perturbation
Google 2019/arXiv	14.1	6.8	300h, attention models
RWTH U 2017	15.7	8.2	300h, with ANN-LM, model comb.
RWTH U 2019/arXiv	13.5	6.7	300h, single models, adaptation
Microsoft 2017	12.0	6.2	2000h, model comb.
IBM 2017	10.0	5.5	2000h, model comb.
Capio 2017	9.1	5.0	2000h, model comb.

ASR: Librispeech Task: Hybrid HMM vs. Attention
(Vassil Panayotov & Daniel Povey)

RWTH AACHEN UNIVERSITY

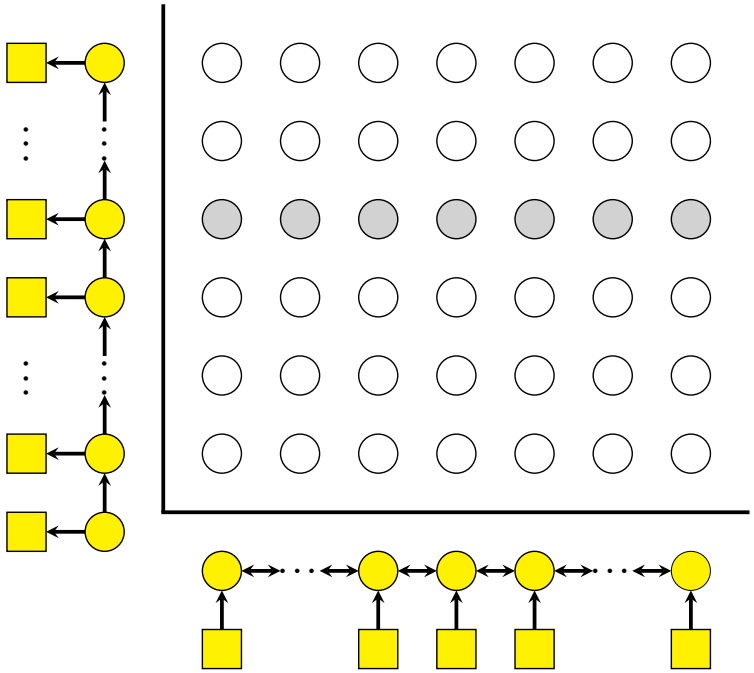
speech data: read audiobooks from the LibriVox project
with training data:
– acoustic model: 960 hrs of speech
– language model: 800 million words

word error rates[%]:

		dev		test	
team	approach	1st half	2nd half	1st half	2nd half
Irie, Zeyer et al. RWTH (Interspeech 2019)	attention with BPE units, 'no' LM	4.3	12.9	4.4	13.5
	+ LSTM-RNN LM	3.0	9.1	3.5	10.0
	+ transformer LM	2.9	8.8	3.1	9.8
Lüscher, Beck et al. RWTH (Interspeech 2019)	hybrid HMM, CART, 4g LM	4.3	10.0	4.8	10.7
	+ seq. disc. training	3.7	8.7	4.2	9.3
	+ LSTM-RNN LM	2.4	5.8	2.8	6.2
	+ transformer LM	2.3	5.2	2.7	5.7
Zeghidour et al., FB 2018	gated CNN with letters/words	3.2	10.1	3.4	11.2
Irie et al., Google 2019	attention with WPM units	3.3	10.3	3.6	10.3
Park et al., Google 2019	attention ... data augmentation	-	-	2.5	5.8

common properties in both approaches:

– **bi-directional LSTM RNN over input: acoustic vectors):** $x_t,\ t = 1, ..., T$
with internal representations $h_t(x_1^T)$
– **uni-directional LSTM RNN over output: (phoneme) labels:** $a_s,\ s = 1, ..., S$



- **hybrid HMM (finite-state model): summing over the paths, i.e. probability models**

$$p(a_1^S|x_1^T) = \sum_{s_1^T} \prod_t p\Big(s_t, y_t = a_{s_t} \Big| s_{t-1}, a_0^{s_{t-1}-1}, h_1^T(x_1^T)\Big)$$

- **attention mechanism: factorization and averaging over internal RNN representations** h_t**:**

$$p(a_1^S|x_1^T) = \prod_s p(a_s|a_0^{s-1}, x_1^T) = \prod_s p(a_s|a_{s-1}, r_{s-1}, c_s)$$

$$c_s := \sum_t p(t|a_0^{s-1}, x_1^T) \cdot h_t(x_1^T)$$

with context vector c_s **and output state vector** r_s

Phoneme RNN-T: Results

results on phoneme/grapheme RNN-Transducer (RNN-T):
IBM research [Saon & Tüske$^+$ 2021] and RWTH [Zhou & Berger$^+$ 2021]

table and results from [Saon & Tüske$^+$ 2021]
on Switchboard (SWB) and Call-Home (CHM):

authors	team	approach		WER[%]	
		acoust.model	lang.model	SWB	CHM
Saon & Tüske$^+$ 2021	IBM	RNN-T	LSTM-RNN	6.3	13.1
Tüske & Saon$^+$ 2020	IBM	attention	LSTM-RNN	6.4	12.5
Park & Chan$^+$ 2019	Google	attention	LSTM-RNN	6.8	14.1
Hadiani & Sameti$^+$ 2018	JHU	LF-MMI	RNN	7.5	14.6
Irie & Zeyer$^+$ 2019	RWTH	hybrid HMM	transformer	6.7	12.9

more results on Italian and Spanish (conversational telephone speech)

conclusions based on [Saon & Tüske$^+$ 2021, Zhou & Berger$^+$ 2021]:
similar performance like hybrid HMM

goal of language modelling: compute the prior $q_\vartheta(w_1^N)$ of a word sequence w_1^N
– how plausible is this word sequence w_1^N (independently of observation x_1^T!) ?
– measure of language model quality: perplexity PP (= geometric average)
interpretation: effective vocabulary size as seen by ASR decoder/search

$$\log PP := \log 1 \Big/ \sqrt[N]{q_\vartheta(w_1^N)} = -1/N \cdot \sum_{n=1}^{N} \log q_\vartheta(w_n|w_0^{n-1})$$

interpretation: prediction task:
based on history w_0^{n-1}, predict $q_\vartheta(w_n|...)$

approaches:
– use full history: RNN or LSTM
– truncate history: $\rightarrow$ k-gram MLP

perplexity PP on test data (QUAERO)
(Sundermeyer et al.; RWTH 2012, 2015):

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%

Interpolated Language Models: Perplexity and WER

- more details and refinements:
  - use of word classes for softmax in output layer
  - unlimited history of RNN: requires re-design of ASR search
- in practice:
  interpolation of TWO models: count model (3B words) + ANN model (60M words)
- perplexity and word error rate on test data

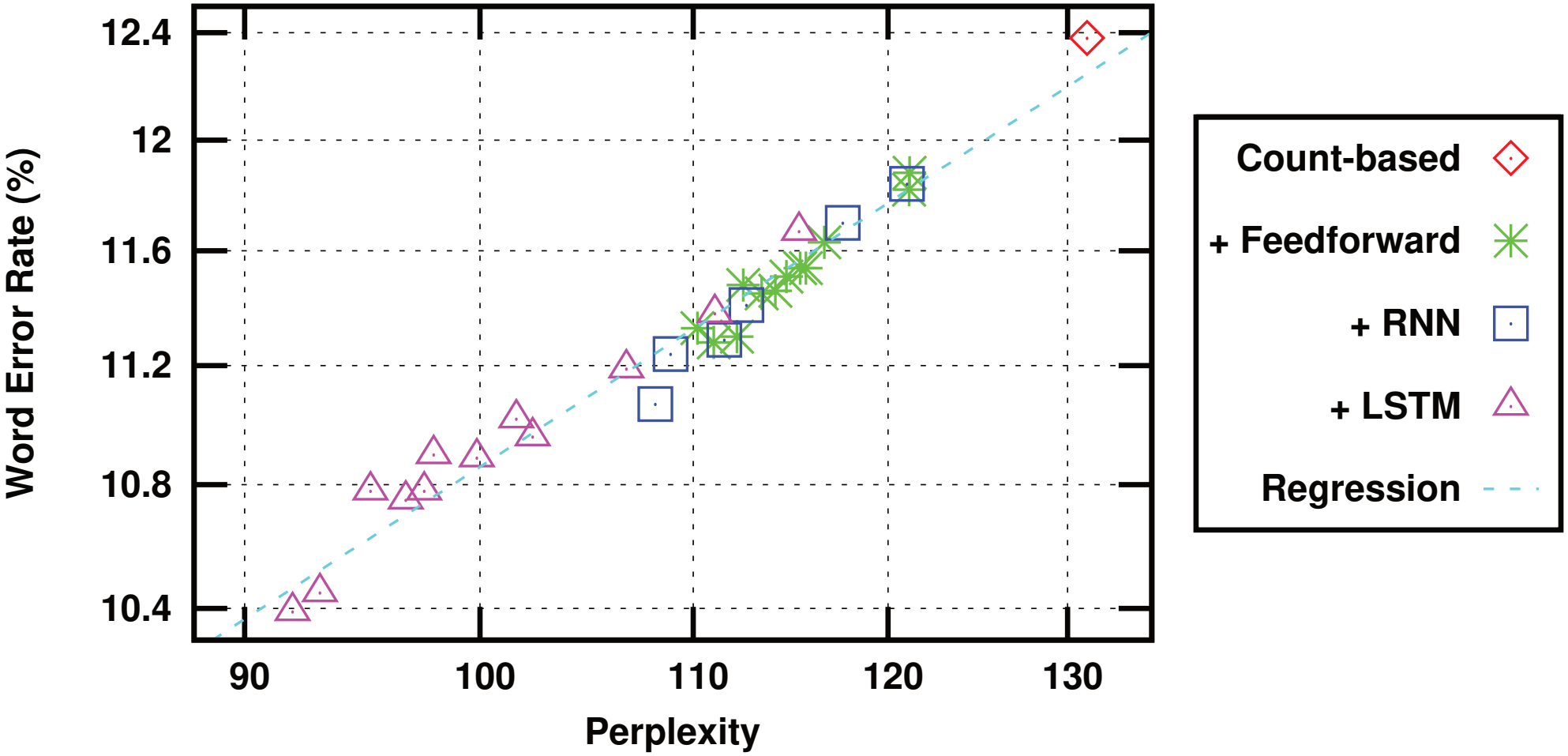
models	PP	WER[%]
count model	131.2	12.4
+ 10-gram MLP	112.5	11.5
+ Recurrent NN	108.1	11.1
+ LSTM-RNN	96.7	10.8
+ 10-gram MLP with 2 layers	110.2	11.3
+ LSTM-RNN with 2 layers	92.0	10.4

- improvements achieved:
  - perplexity: 30% reduction: from 131 to 92
  - WER: 15% reduction: from 12.4% to 10.4%

Plot: Perplexity vs. Word Error Rate

empirical law: $WER = \alpha \cdot PP^{\beta}$ **with** $\beta \in [0.3, 0.5]$

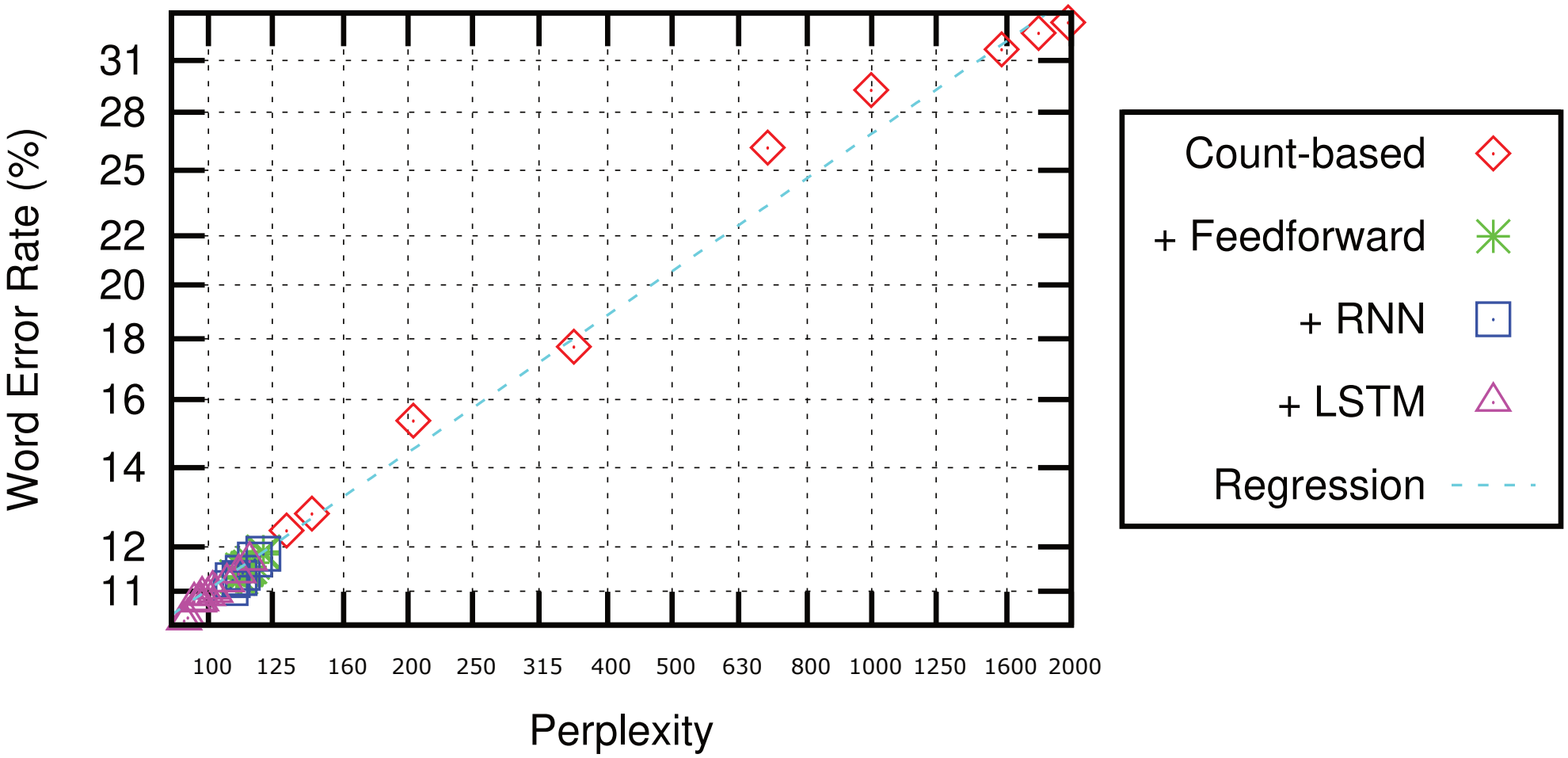
[Makhoul & Schwartz 94, Klakow & Peters 02]



Extended Range: Perplexity vs. Word Error Rate

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

open question: theoretical justification?



Machine Translation (MT): History

statistical approaches were controversial in MT (and other NLP tasks):

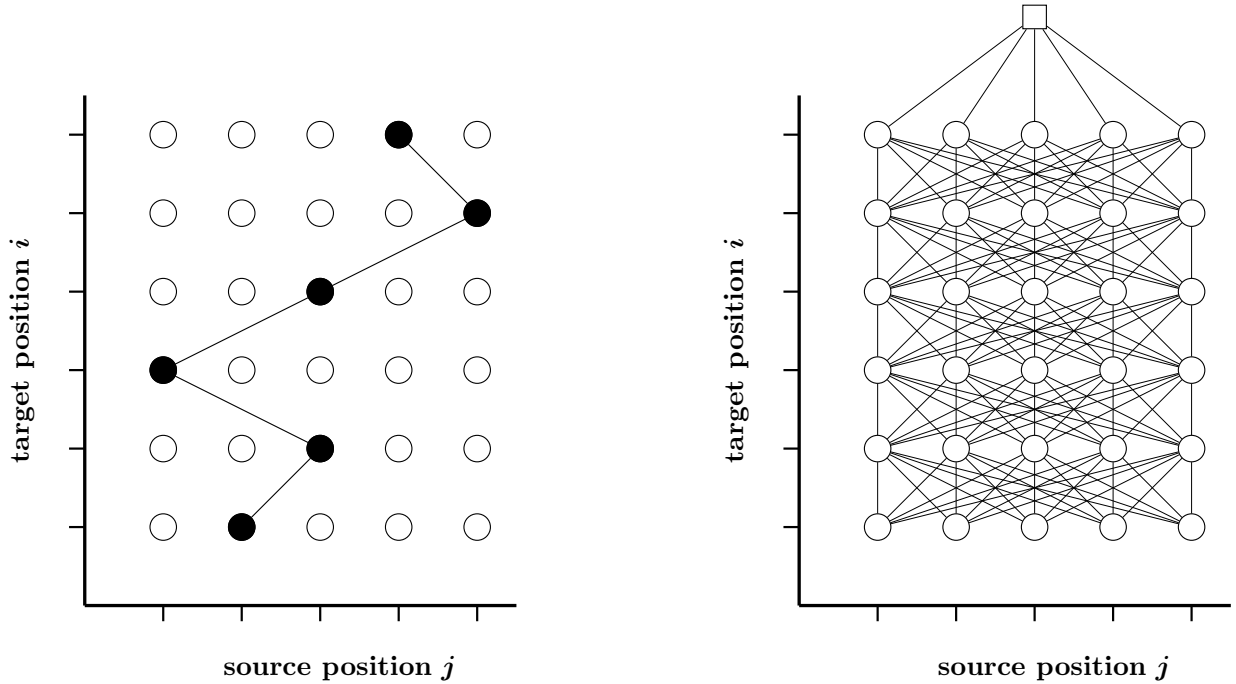
- **1969 Chomsky:**
  ... the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term.
- **result: strict dichotomy until 1995-2005:**
  – speech = spoken language: signals, subsymbolic, machine learning
  – language = written text: symbols, grammars, rule-based AI
- **until 2000: mainstream approach was rule-based**
  – result: huge human effort required in practice
  – problems: coverage and consistency of rules
- **1989-93: IBM Research: statistical approach to MT**
  1994: key people (Mercer, Brown) left for a hedge fund
- **1996-2002 RWTH: improvements beyond IBM's approach:**
  phrase-based approach and log-linear modelling
- **around 2004: from singularity to mainstream in MT**
  F. Och (and more RWTH PhD students) joined Google
  2008: service *Google Translate*
- **2015: neural MT: attention mechanism [Bahdanau & Cho$^+$ 15]**

Machine Translation: Neural HMM

- translation: from source sentence $f_1^J = f_1 ... f_j ... f_J$ to target sentence $e_1^I = e_1 ... e_i ... e_I$
- alignment direction: from target to source: $i \rightarrow j = b_i$
- first-order hidden alignments and factorization:

$$p(e_1^I|f_1^J) = \sum_{b_1^I} p(b_1^I, e_1^I|f_1^J) = \sum_{b_1^I} \prod_i p(b_i, e_i|b_{i-1}, e_0^{i-1}, f_1^J)$$

- resulting model: exploit first-order structure (or zero-order)
  training: backpropagation within EM algorithm



Experimental Results

- **WMT task: German $\rightarrow$ English:**
  - **training data: 6M sentence pairs = (137M, 144M) words**
  - **test data: (about) 3k sentence pairs = (64k, 67k) words**
- **WMT task: Chinese $\rightarrow$ English:**
  - **training data: 14M sentence pairs = (920M Chinese letters, 364M English words)**
  - **test data: (about) 2k sentence pairs = (153k Chinese letters, 71k English words)**
- **performance measures:**
  - **BLEU [%]: accuracy measure: "the higher, the better"**
  - **TER [%]: error measure: "the lower, the better"**
- **basic units for implementation:**
  - **BPE (*byte pair encoding*) units rather than full-form words**
  - **alphabet size: about 40k**

- **RWTH papers (with preliminary results):**
  [Wang & Alkhouli$^+$ 17, Wang & Zhu$^+$ 18]

Comparison: Best Results

	German→English				Chinese→English			
	test2017		test2018		dev2017		test2017	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
LSTM-RNN attention	32.1	56.3	38.8	48.1	21.4	63.6	22.9	62.0
self-attention transformer	33.4	55.3	40.4	46.8	21.8	62.9	23.5	60.1
neural HMM	31.9	56.6	38.3	48.3	20.8	63.2	22.4	61.4

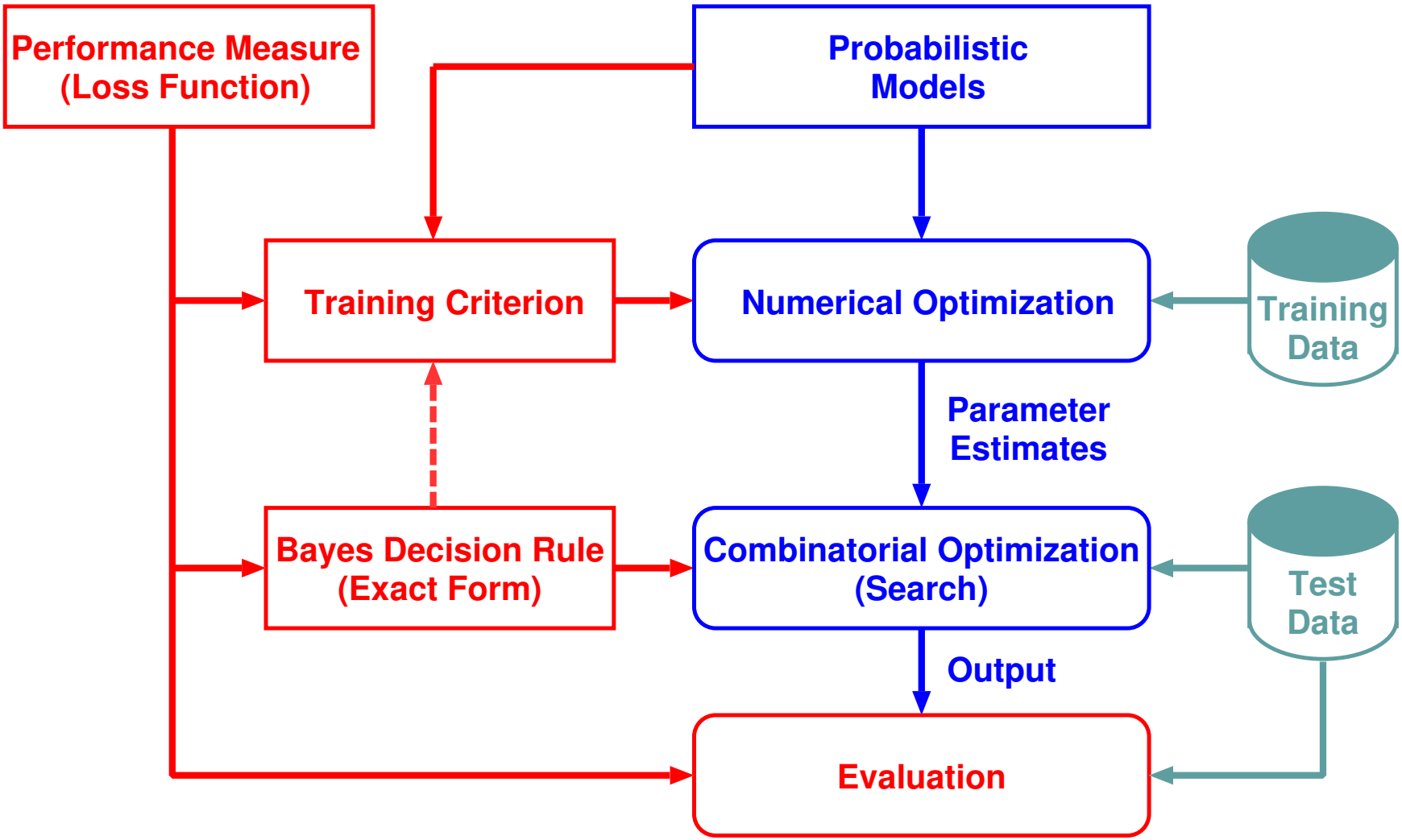
conclusions about neural HMM:
– (nearly) competitive with LSTM-RNN attention approach
– some performance gap to self-attention approach

– room for improvement of neural HMM

Summary and Conclusions

- **Bayes decision theory and statistical approach:**
  principal ingredients
  – probabilistic models at these levels and the interaction between these levels
  – training criterion along with an optimization algorithm
  – choice of performance measure: errors at sequence, word, phoneme, frame level
  – Bayes decision rule: search/decoder with an efficient implementation
- **deep learning:**
  – defines one family of probabilistic models within statistical approach
  – baseline structure: matrix-vector product + nonlinearities
  – yes, resulted in significant improvements
- **surprising success of deep learning**
  in *symbolic tasks* (e. g. MT and other NLP tasks)
- **history of machine learning and statistical classification:**
  – there has been and will be life outside deep learning
  – we must get not only the principles, but also the details right

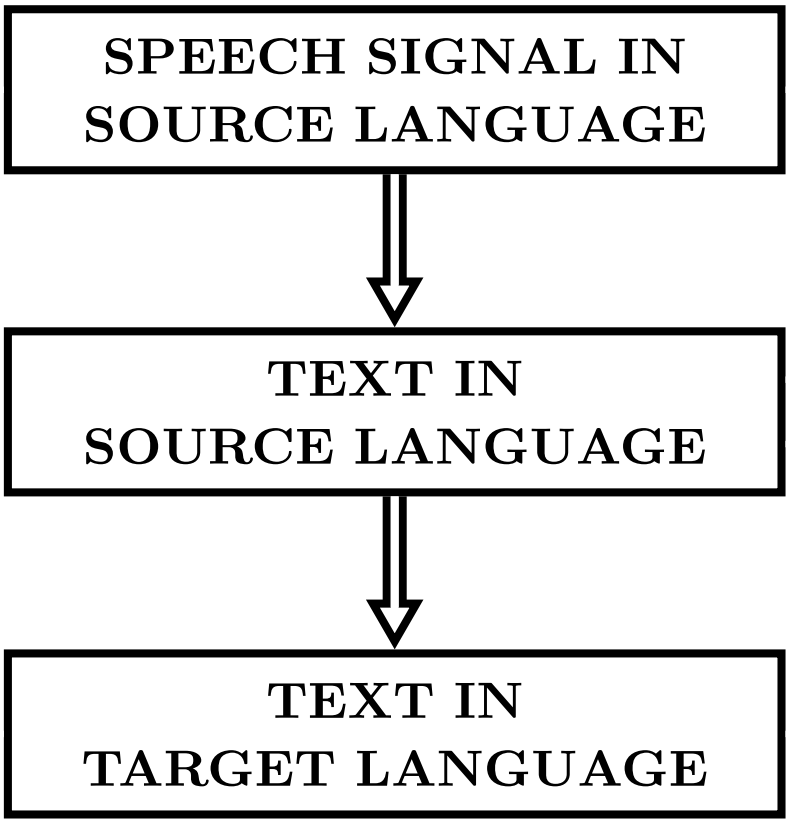
Machine Learning and Statistical Decision Theory for HLT



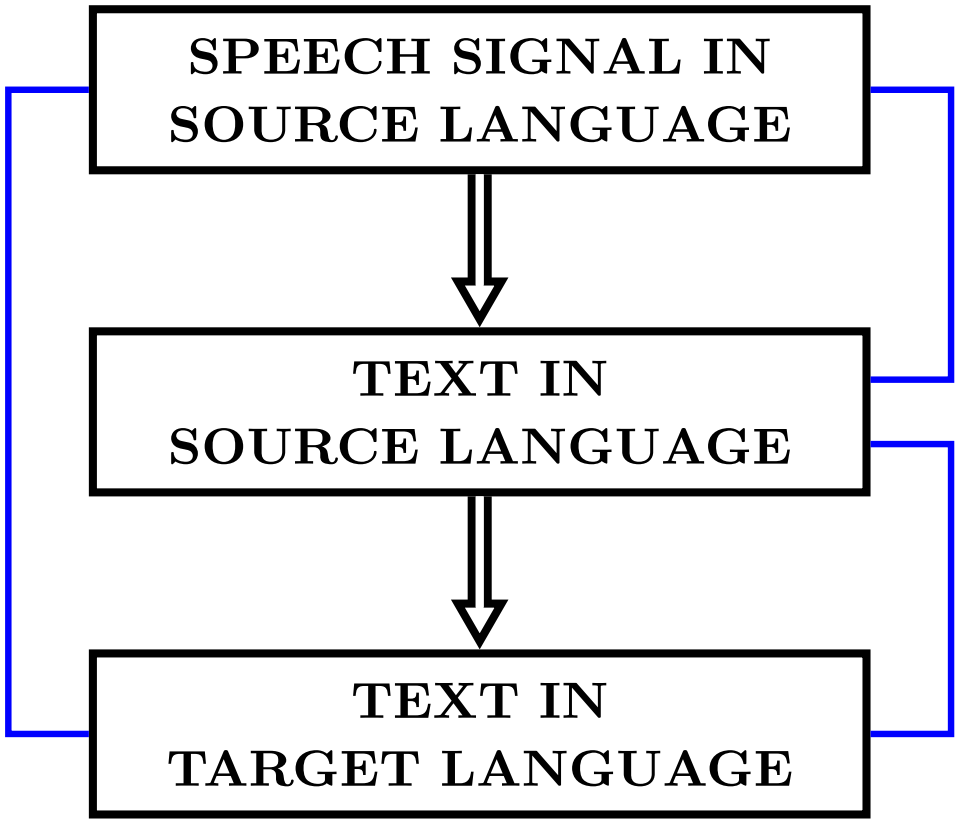
Challenges

- challenges for general machine learning:
  - mathematical optimization with huge complexity:
    we need a theoretical framework for practical aspects of gradient search
  - can we find ANNs with more explicit probabilistic structures?
  - novel structures beyond matrix-vector product + nonlinearties?
- challenges in ASR:
  - to continue the general improvements (ongoing for 40 years!)
  - task: informal colloquial speech (meetings)
  - robustness wrt acoustic conditions and language context
    (improved adaptation ?)
- unsupervised training for ASR:
  machine learning with (virtually) no labeled data?
- features for ASR beyond spectral analysis/Fourier transform:
  - recent work [Tüske & Golik 2014, Sainath et al. 2015, Baevski & Schneider$^+$ 2020]
  - real and consistent improvements over spectral analysis?
- architecture for speech translation:
  challenge: handle three types of training data

Tasks in Human Language Technology: Speech Translation



Tasks in Human Language Technology: Speech Translation



RWTH AACHEN UNIVERSITY

END

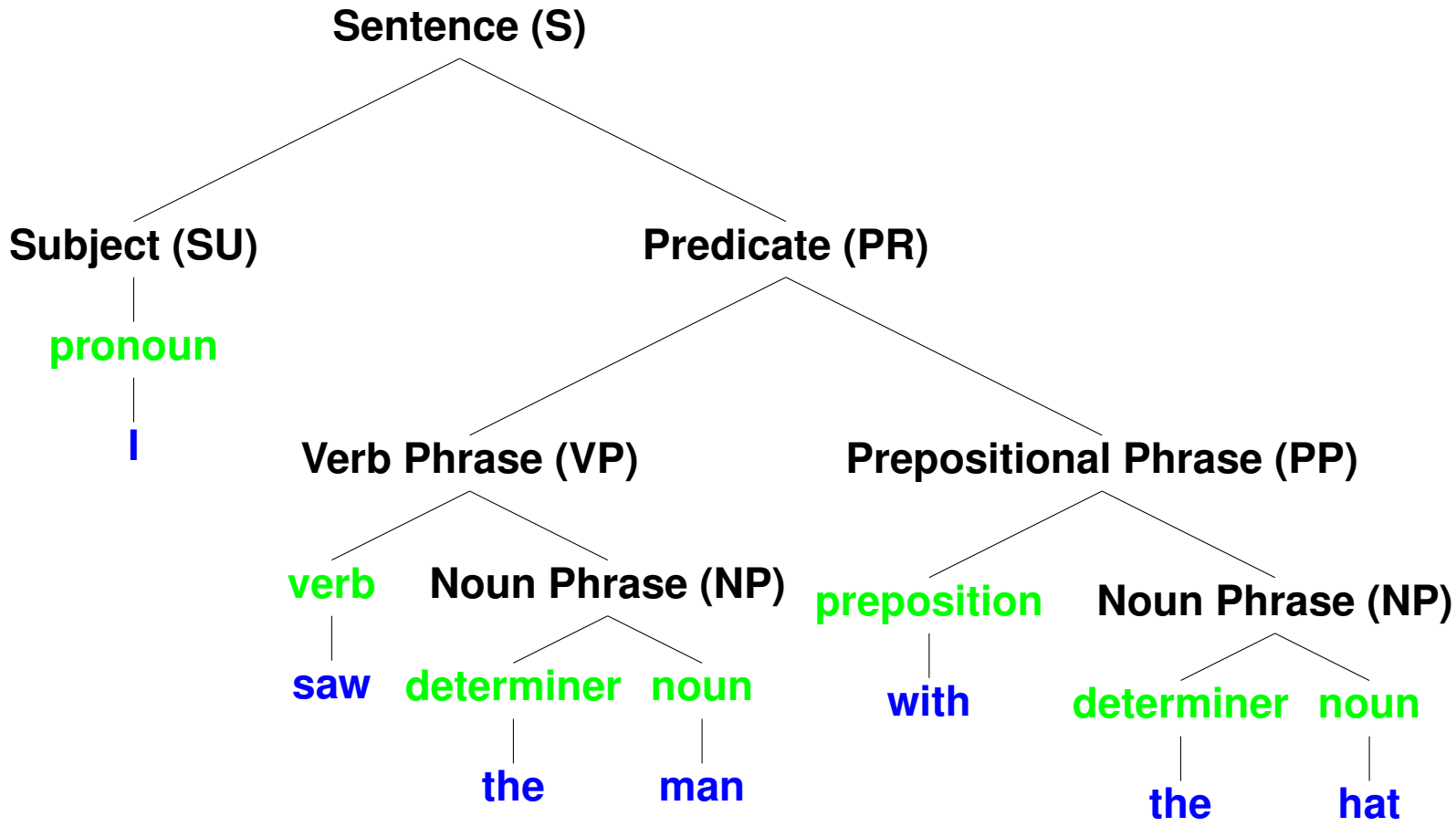
BACK-UP SLIDES

Speech & Language: From Bayes to Deep Learning

HLT

Rule-based Artficial Intelligence
Example: Grammar Rules and Syntactic Structure

- **principle:**



- **extensions along many dimensions**

Pseudo Bayes Decision Rule: Sources of Errors:
Why does a 'Bayes' decision system make errors?

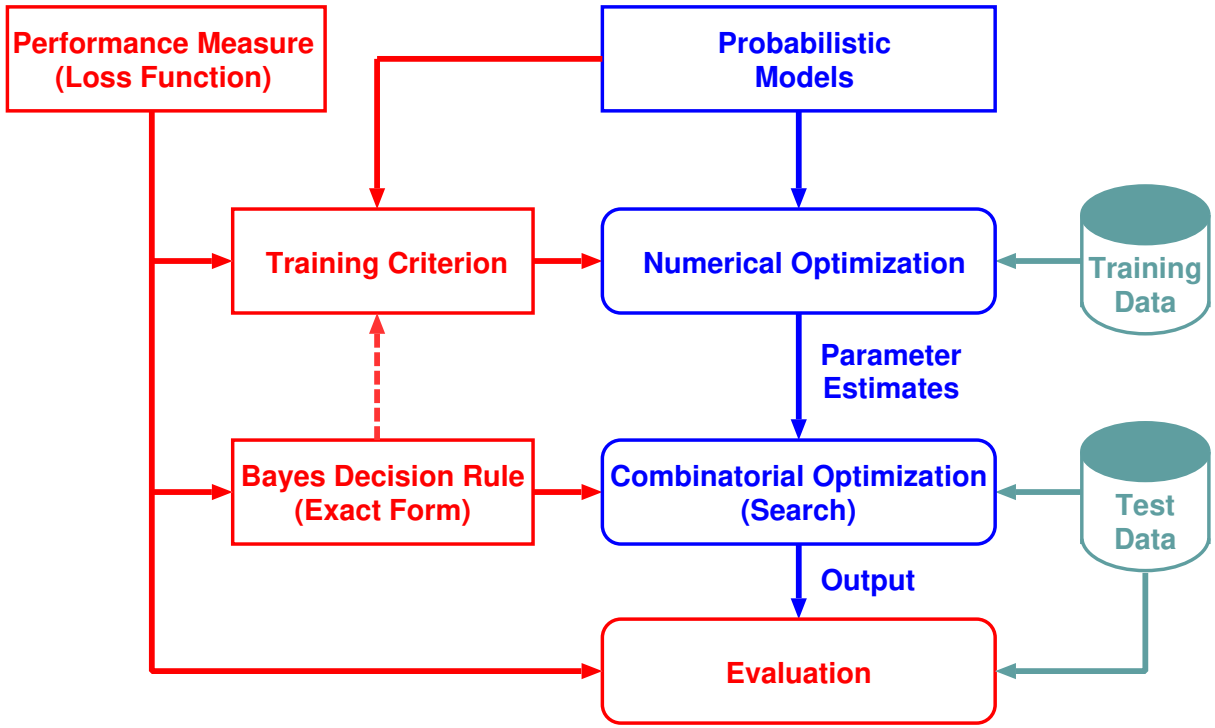
To be more exact: Why errors IN ADDITION to the so-called Bayes errors, i. e. the minimum that can be achieved?

Reasons from the viewpoint of (pseudo) Bayes decision rule:

- **probability models:**
  - **inadequate features/observations: e. g. spectral features vs. speech signal**
  - **inadequate models: e. g. acoustic model or language model**
- **training conditions:**
  - **poor training criterion: does not have strong link to performance (e.g. WER)**
  - **not enough training data**
  - **mismatch conditions between training and test data e. g. clean speech vs. noisy speech**
- **training criterion + efficient algorithm:**
  - **suboptimal algorithm for training (e. g. gradient descent)**
- **decision rule:**
  - **incorrect error measure, e. g. 0/1 loss**
- **decision rule + efficient algorithm:**
  - **suboptimal search procedure, e.g. beam search or N-best lists**

What have we learnt?

- probabilistic model:
  most important component
- training criterion:
  – is important
  – depends on prob. model
- numerical optimization:
  very important
  (compare 1993 vs. 2013!)
- loss function:
  no error weights: not critical
  low-accuracy conditions:
  more critical
- decision rule:
  – search: important
  for low-accuracy models
  – exact loss function: important
  for low-accuracy models



Neural Hidden Markov Model

- LSTM-RNN based representations for input and output:
  4 layers of encoder and 1 layer of decoder
- independent models of alignment and lexicon
  (no parameter sharing as in attention approach)

HMM	German→English						Chinese→English					
	#Par	PPL	test2017		test2018		#Par	PPL	dev2017		test2017	
			BLEU	TER	BLEU	TER			BLEU	TER	BLEU	TER
zero-order	129M	5.29	30.9	57.4	37.4	48.9	125M	8.12	20.1	65.1	20.7	64.2
first-order	136M	4.64	31.6	56.5	38.7	48.4	138M	7.63	20.1	64.0	22.0	63.2

machine translation from source source to target language:

(source: foreign) $f_1^J \rightarrow e_1^I$ **(target: English)**

key concepts for modelling posterior probability $p(e_1^I|f_1^J)$

- **direct approach: use unidirectional RNN over target positions** $i = 1, ..., I$ **with internal state vector** s_i**:**

$$p(e_1^I|f_1^J) = \prod_i p(e_i|e_0^{i-1}, f_1^J) = \prod_i p(e_i|e_{i-1}, s_{i-1}, f_1^J)$$

  interpretation: extended language model for target word sequence

- **additional component: attention mechanism for localization**

$$p(e_i|e_{i-1}, s_{i-1}, f_1^J) = p(e_i|e_{i-1}, s_{i-1}, c_i)$$

  with a context vector: $c_i := C(s_{i-1}, f_1^J)$

word embeddings and representations:

- **word embedding for target sequence:**
  - **word symbol:** e_i
  - **word vector:** $\tilde{e}_i = R_e(e_i)$ **with the embedding (matrix)** R_e
- **word embedding for source sequence:**
  - **word symbol:** f_j
  - **word vector:** $\tilde{f}_j = R_f(f_j)$ **with the embedding (matrix)** R_f
- **word representation** h_j **for source sequence using a bidirectional RNN:** $h_j = H_j(f_1^J)$

warning:
– concept: clear distinction between $f_j, \tilde{f}_j, h_j$
– notation and terminology: not necessarily consistent

approach:

- **input: bidirectional RNN over source positions j:** $f_1^J \rightarrow h_j = H_j(f_1^J)$
- **output: unidirectional RNN over target positions i:**

  $$y_i = Y(y_{i-1}, s_{i-1}, c_i)$$

  conventional notation:

  $$p(e_i|\tilde{e}_{i-1}, s_{i-1}, c_i)$$

  with RNN state vector $s_i = S(s_{i-1}, \tilde{e}_i, c_i)$ **and context vector** $c_i = C(s_{i-1}, h_1^J)$
- **context vector c_i: weighted average of source word representations:**

$$c_i = \sum_j \alpha(j|i, s_{i-1}, h_1^J) \cdot h_j \qquad \alpha(j|i, s_{i-1}, h_1^J) = \frac{\exp(A[s_{i-1}, h_j])}{\sum_{j'} \exp(A[s_{i-1}, h_{j'}])}$$

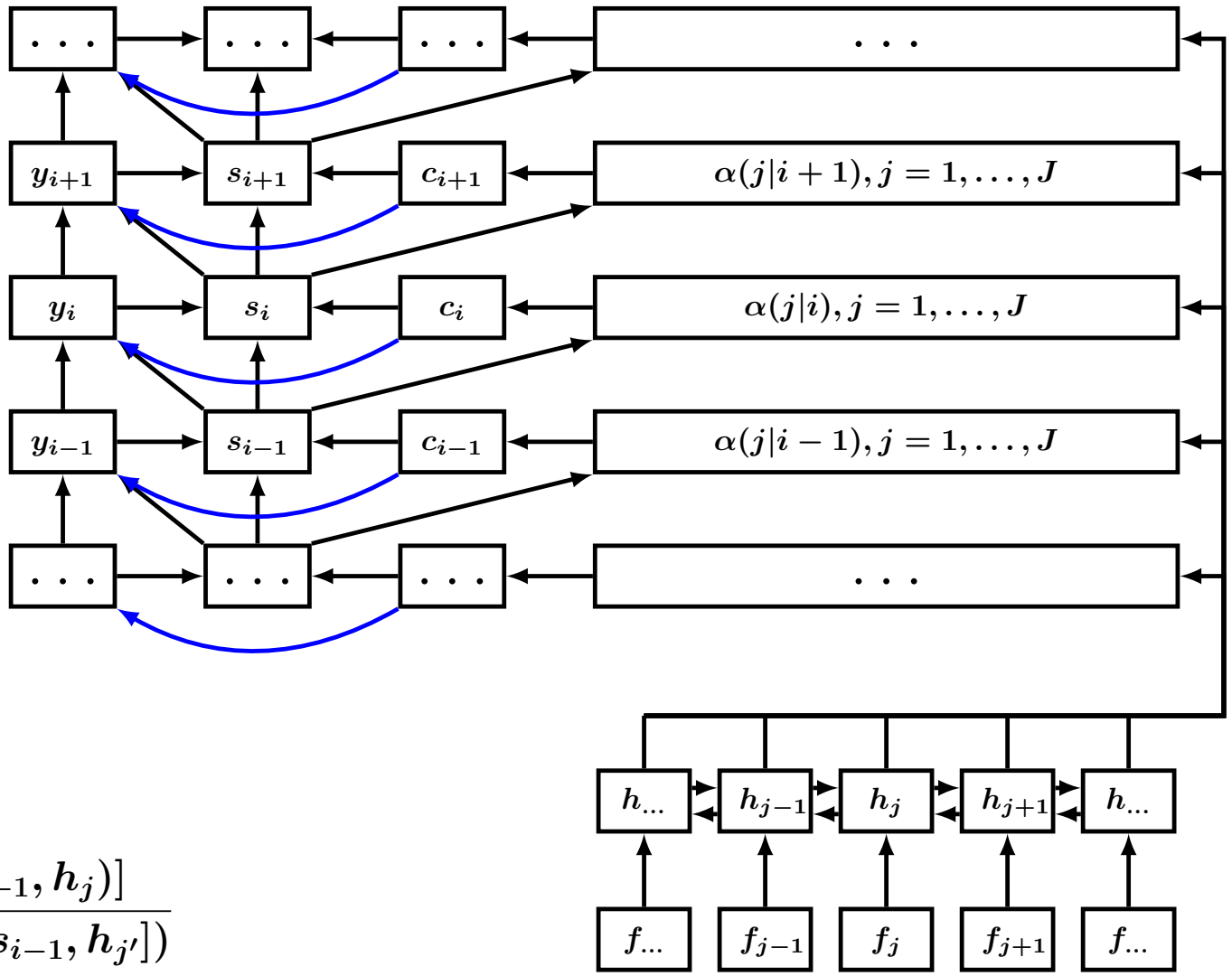
with the normalized attention weights $\alpha(j|i, s_{i-1}, h_1^J)$
and real-valued attention scores $A[s_{i-1}, h_j]$

State of the Art: Attention-based Neural MT [Bahdanau & Cho+ 15]

principle:

- **input: source sequence:**
  $f_1^J \rightarrow h_j = H_j(f_1^J)$
- **output distribution:**
  $y_i \equiv p_i(e|\tilde{e}_{i-1}, s_{i-1}, c_i)$
  notation in ANN style:
  $y_i = Y(y_{i-1}, s_{i-1}, c_i)$
- **state vector of target RNN:**
  $s_i = S(s_{i-1}, y_i, c_i)$
- **weighted context vector:**
  $c_i = \sum_j \alpha(j|i, s_{i-1}, h_1^J) \cdot h_j$
- **attention weights:**

$$\alpha(j|i, s_{i-1}, h_1^J) = \frac{\exp(A[s_{i-1}, h_j])}{\sum_{j'} \exp(A[s_{i-1}, h_{j'}])}$$

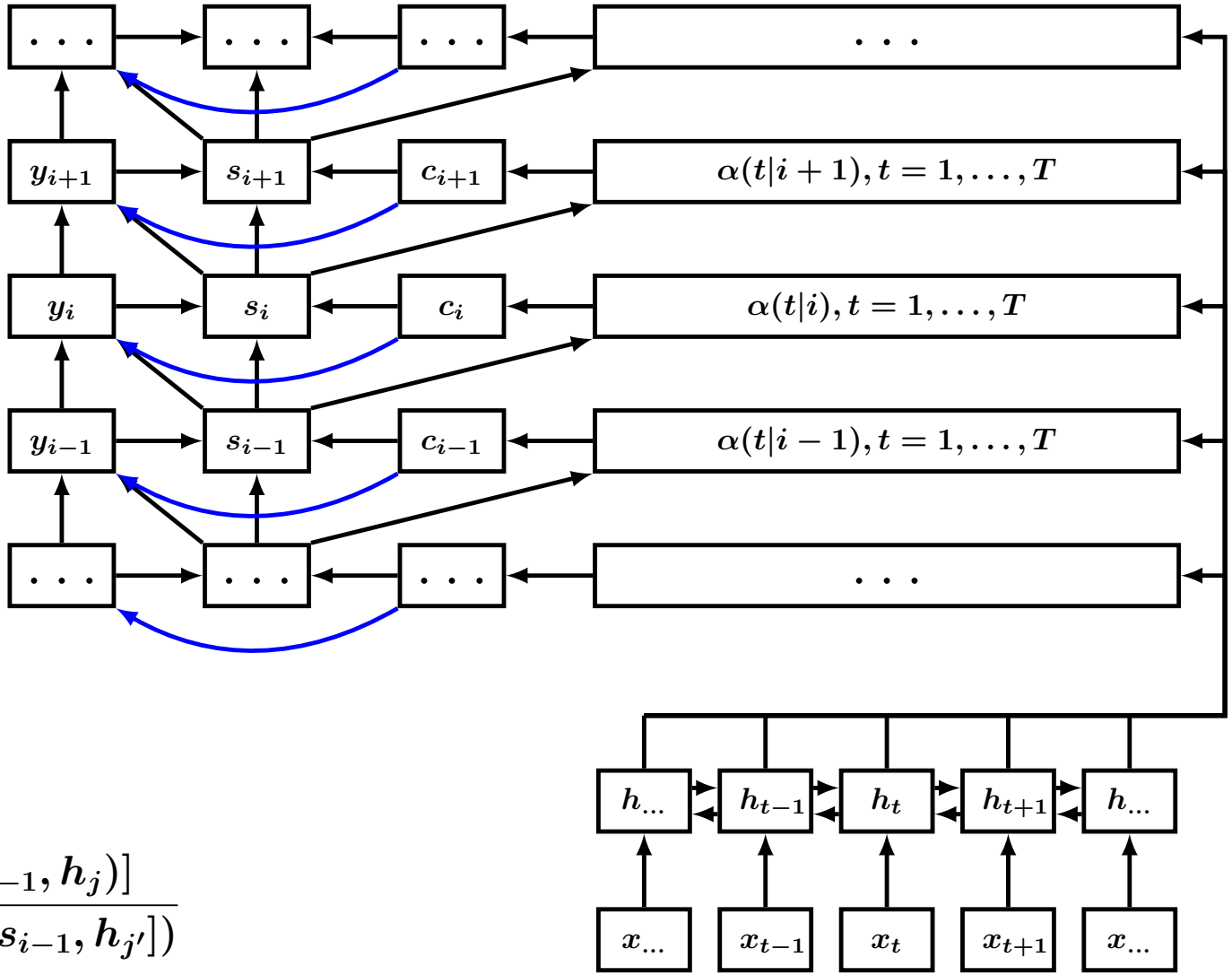


Attention-based ASR
[Bahdanau & Cho$^+$ 15]

principle:

- **input: source sequence:**
  $f_1^J \rightarrow h_j = H_j(f_1^J)$
- **output distribution:**
  $y_i \equiv p_i(e|\tilde{e}_{i-1}, s_{i-1}, c_i)$
  notation in ANN style:
  $y_i = Y(y_{i-1}, s_{i-1}, c_i)$
- **state vector of target RNN:**
  $s_i = S(s_{i-1}, y_i, c_i)$
- **weighted context vector:**
  $c_i = \sum_j \alpha(j|i, s_{i-1}, h_1^J) \cdot h_j$
- **attention weights:**

$$\alpha(j|i, s_{i-1}, h_1^J) = \frac{\exp(A[s_{i-1}, h_j)]}{\sum_{j'} \exp(A[s_{i-1}, h_{j'}])}$$



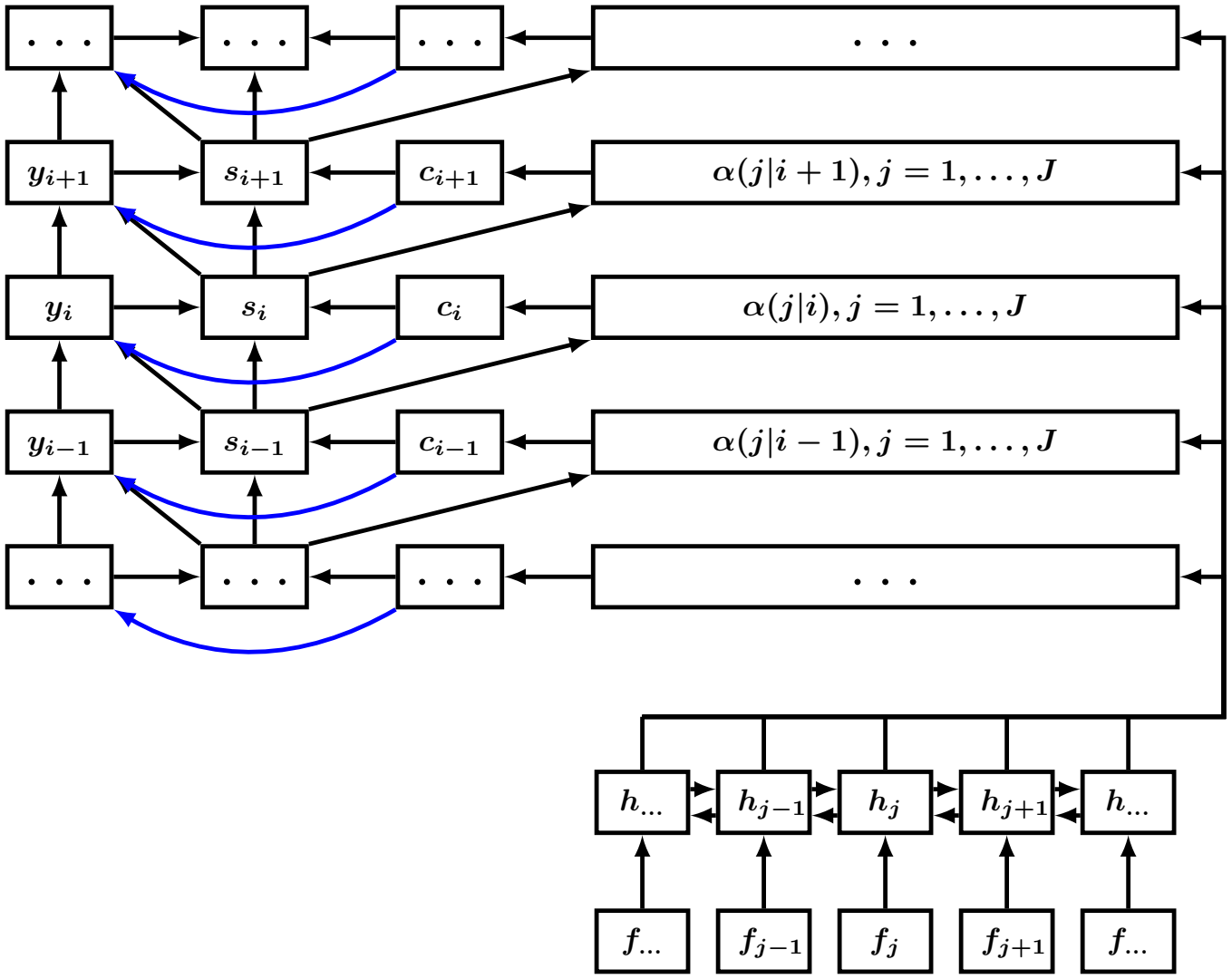
Attention-based Neural MT: Sequential Order of Operations

preparations:

- **input preprocessing:**
  $f_1^J \rightarrow h_j = H_j(f_1^J)$
- **available at position $i-1$:**
  $\tilde{e}_{i-1} \equiv y_{i-1},\ s_{i-1},\ c_{i-1}$

sequence of operations for position i:

1. **attention weights:**
   $\alpha(j|i, s_{i-1}, h_1^J) = ...$
2. **context vector:**
   $c_i = \sum_j \alpha(j|i, s_{i-1}, h_1^J) \cdot h_j$
3. **output distribution:**
   $y_i = Y(y_{i-1}, s_{i-1}, c_i)$
4. **state vector:**
   $s_i = S(s_{i-1}, y_i, c_i)$



Attention Weights
Feedforward ANN vs. Dot Product

re-consider attention weights:

$$\alpha(j|i, s_{i-1}, h_1^J) = \frac{\exp(A[s_{i-1}, h_j])}{\sum_{j'} \exp(A[s_{i-1}, h_{j'}])}$$

two approaches to modelling attention scores $A[s_{i-1}, h_j]$:

- **additive variant: feedforward (FF) ANN:**

$$A[s_{i-1}, h_j] := v^T \cdot \tanh(S s_{i-1} + H h_j)$$

  with matrices S and H and vector v
  basic implementation: one FF layer + softmax

- **multiplicative variant: (generalized) dot product between vectors:**

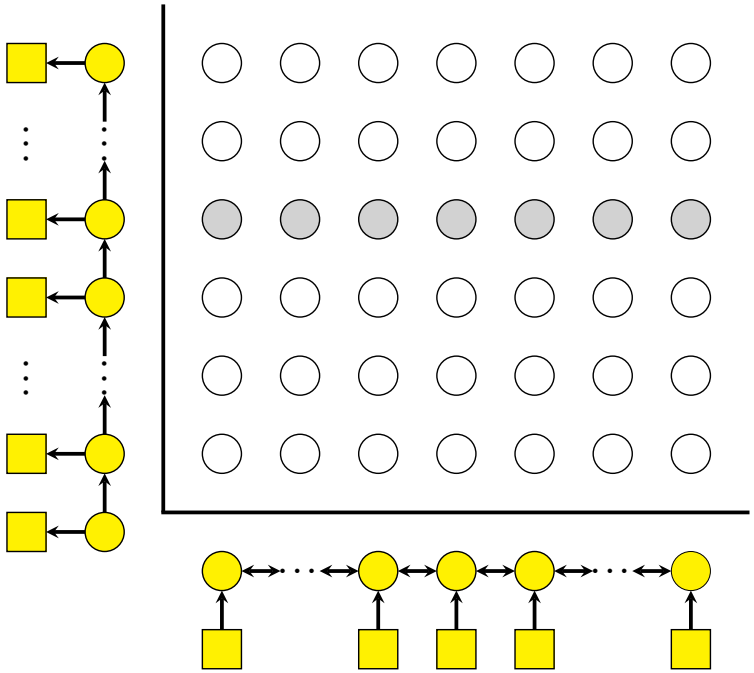
$$A[s_{i-1}, h_j] := s_{i-1}^T \cdot W \cdot h_j$$

  with a attention matrix W

experimental result: not much difference

common properties in both approaches:

– **bi-directional LSTM RNN over input words** $f_j,\ j = 1, ..., J$
– **uni-directional LSTM RNN over output words** $e_i,\ i = 1, ..., I$



- **direct HMM (finite-state model): summing over probability models**

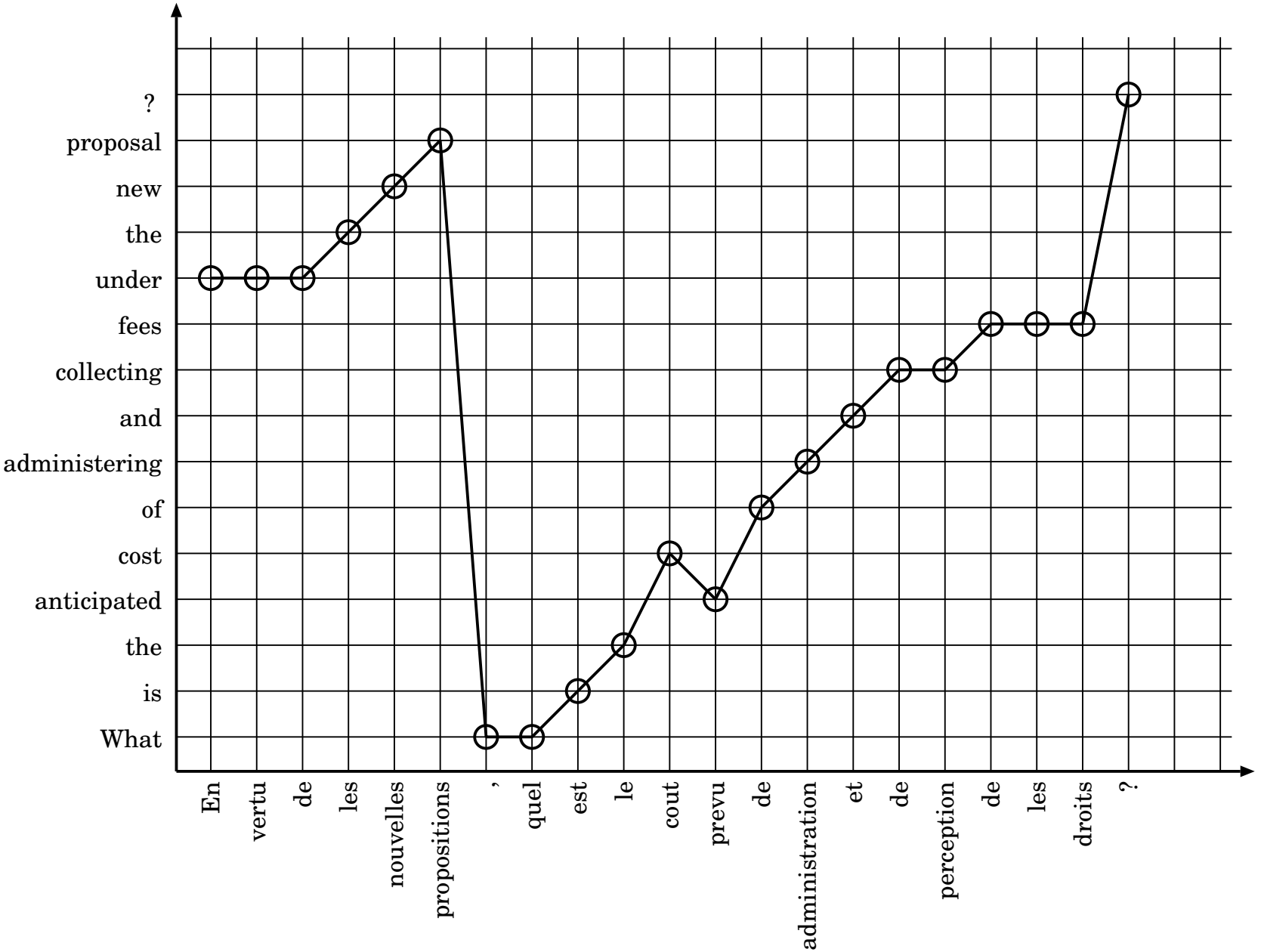
$$p(e_1^I|f_1^J) = \sum_{b_1^I} \prod_i p(b_i, e_i|b_{i-1}, e_0^{i-1}, f_1^J)$$

- **attention mechanism: averaging over internal RNN representations h_j:**

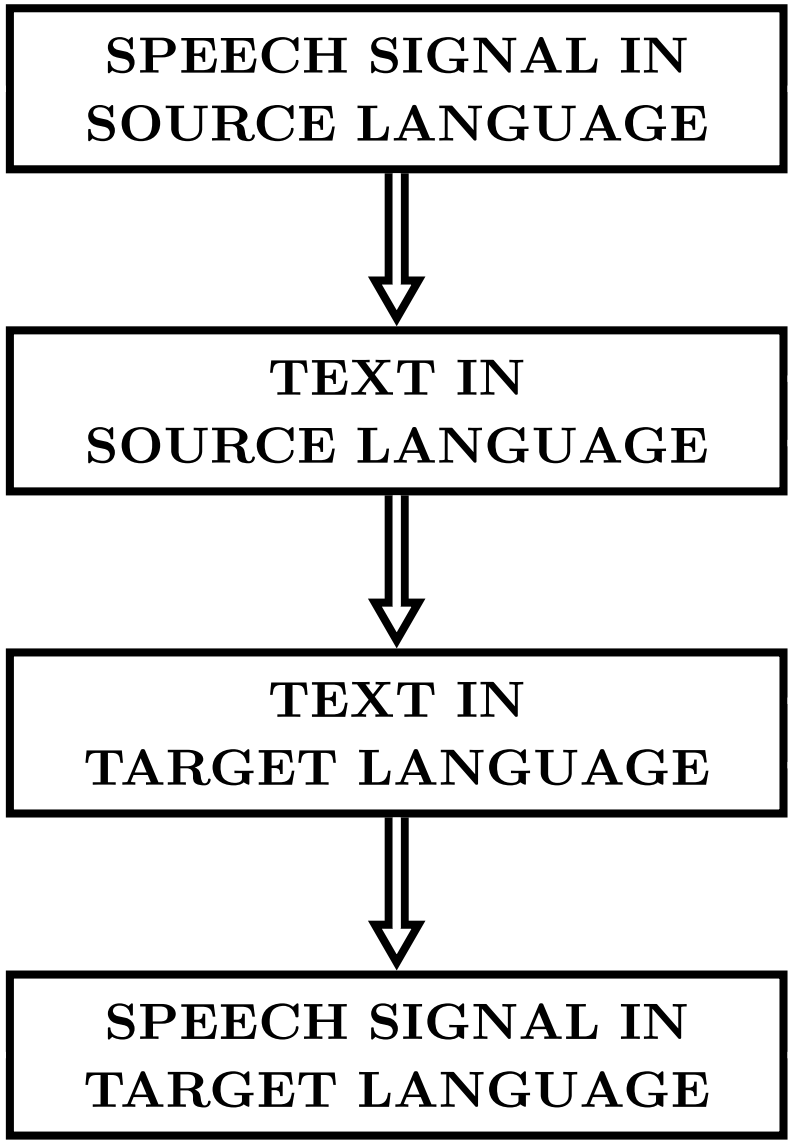
$$p(e_i|e_0^{i-1}, f_1^J) = p(e_i|e_{i-1}, s_{i-1}, c_i)$$

$$\text{with} \quad c_i = \sum_j p(j|e_0^{i-1}, f_1^J) \cdot h_j(f_1^J)$$

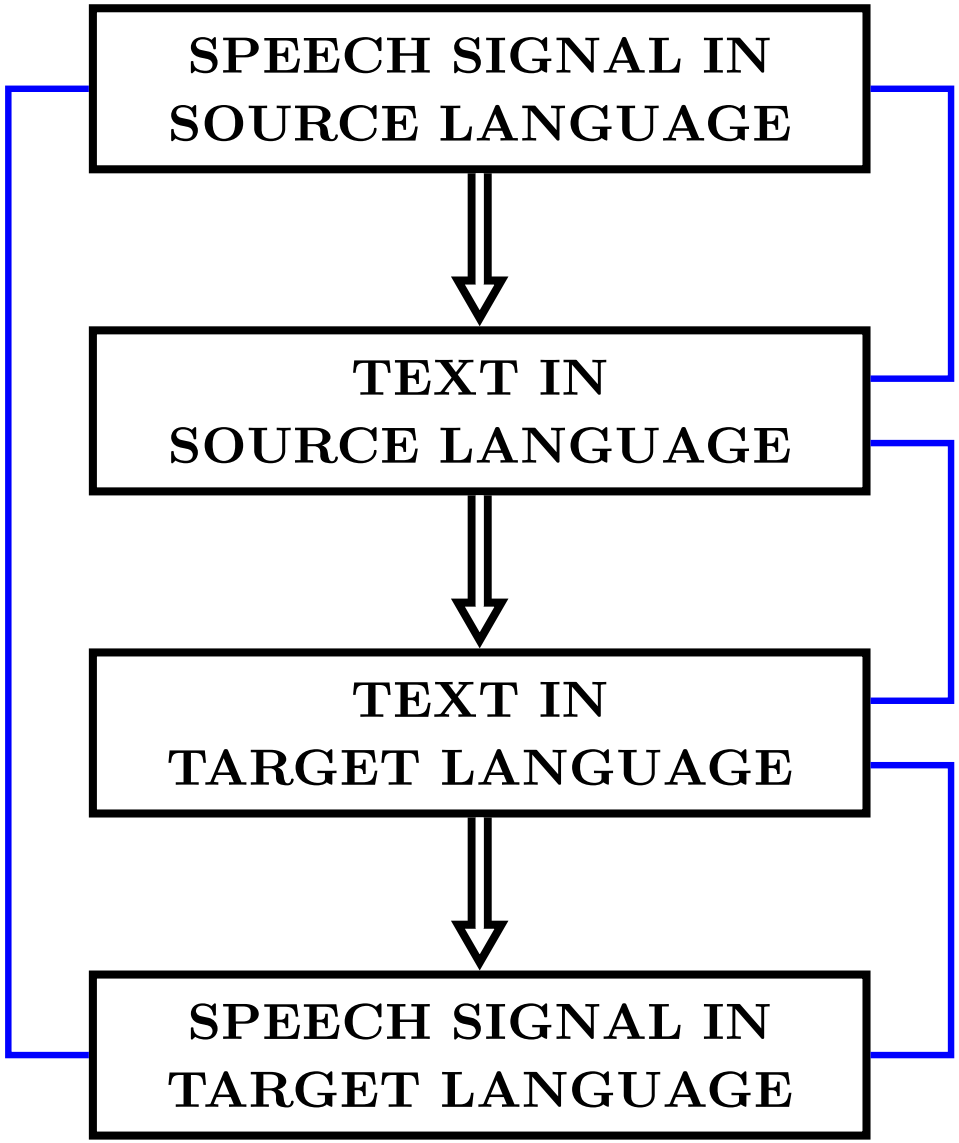
Word Alignments (based on HMM)
(learned automatically; Canadian Parliament)



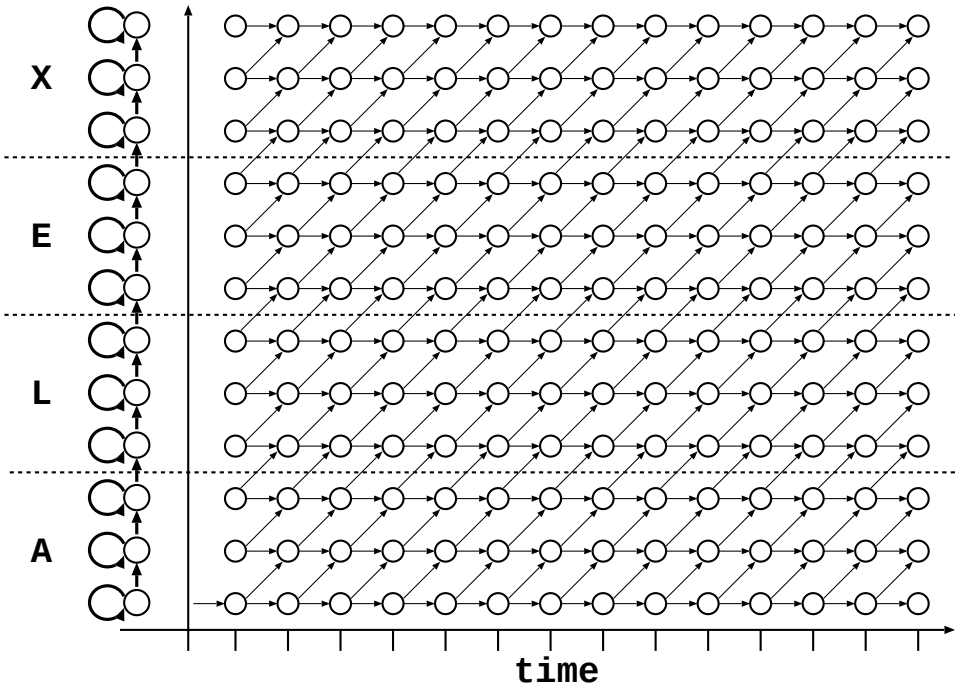
Tasks in Human Language Technology: Speech-to-Speech Translation

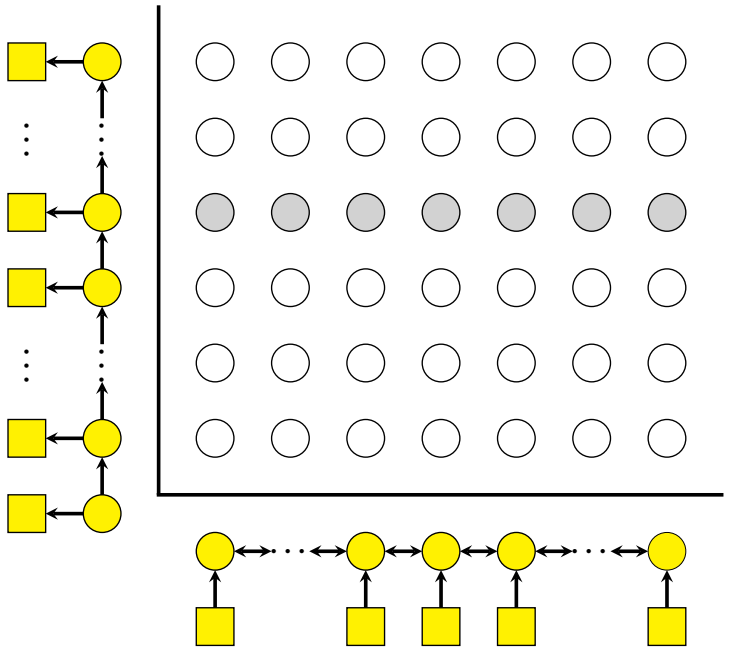


Tasks in Human Language Technology: Speech-to-Speech Translation



Sequence-to-Sequence Processing: Direct HMM and Attention Model





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 show competitive results.

History of NN based approaches to MT:

- **1997 [Neco & Forcada 97]:**
  asynchronous translations with recurrent neural nets
- **1997 [Castano & Casacuberta 97, Castano & Casacuberta$^+$ 97]:**
  machine translation using neural networks and finite-state models
- **2007 [Schwenk & Costa-jussa$^+$ 07]:**
  smooth bilingual n-gram translation
- **2012 [Le & Allauzen$^+$ 12, Schwenk 12]:**
  continuous space translation models with neural networks
- **2014 [Devlin & Zbib$^+$ 14]:**
  fast and robust neural networks for SMT
- **2014 [Sundermeyer & Alkhouli$^+$ 14]:**
  recurrent bi-directional LSTM RNN for SMT
- **2015 [Bahdanau & Cho$^+$ 15]:**
  joint learning to align and translate

References

[Baevski & Schneider+ 20] A. Baevski, S. Schneider, M. Auli: VQ-Wav2Vec: Self-Supervised Learning of Discrete Speech Representations. Facebook AI Research, Menlo Park, CA, arxiv, 16-Feb-2021.

[Bahdanau & Cho+ 15] D. Bahdanau, K. Cho, Y. Bengio: Neural machine translation by jointly learning to align and translate. Int. Conf. on Learning and Representation (ICLR), San Diego, CA, May 2015.

[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 89] 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 82] J. S. Bridle, M. D. Brown, R. M. Chamberlain: An Algorithm for Connected Word Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Paris, pp. 899-902, May 1982.

[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.

[Bridle & Dodd 91] J. S. Bridle, L. Dodd: An Alphanet Approach To Optimising Input Transformations for Continuous Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Toronot, pp. 277-280, April 1991.

[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 & Ranzato$^+$ 10] G. E. Dahl, M. Ranzato, A. Mohamed, G. E. Hinton: Phone recognition with the mean-covariance restricted Boltzmann machine. Advances in Neural Information Processing Systems (NIPS) 23, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, Eds. Cambridge, MA, MIT Press, 2010, pp. 469-477.

[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.

[Dehak & Kenny+ 11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet: Front-End Factor Analysis for Speaker Verification IEEE Trans. on audio, speech, and language processing, pp. 788-798, Vol. 19, No. 4, May 2011.

[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.

[Doetsch & Hannemann+ 17] P. Doetsch , M. Hannemann, R. Schlüer, H. Ney: Inverted Alignments for End-to-End Automatic Speech Recognition. IEEE Journal of selected topics in Signal Processing, Vol. 11, No. 8, pp. 1265-1273, Dec. 2017.

[Forcada & Carrasco 05] M. L. Forcada, R. C. Carrasco: Learning the initial state of a second-order recurrent neural network during regular language inference. Neural Computation, Vol. 7, No. 5, pp. 923-930, Sep. 2005.

[Fontaine & Ris+ 97] V. Fontaine, C. Ris, J.-M. Boite: Nonlinear discriminant analysis for improved speech recognition. Eurospeech, Rhodes, Greece, Sep. 1997.

[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.

[Gemello & Manai+ 06] R. Gemello, F. Mana, S. Scanzio, P. Lafac, R. De Mori: Adaptation of Hybrid ANN/HMM Models Using Linear Hidden Transformations and Conservative Training. IEEE Int. Conf. on Acoustics Speech and Signal Processing Proceedings, Toulouse, 2006.

[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 12] A. Graves: Sequence Transduction with Recurrent Neural Networks. U of Toronto, Canada, arxiv, 12-Nov-2012.

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

[Graves & Schmidhuber 09] A. Graves, J. Schmidhuber: Offline handwriting recognition with multidimensional recurrent neural networks. NIPS 2009.

[Grezl & Fousek 08] F. Grezl, P. Fousek: Optimizing bottle-neck features for LVCSR. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4729-4732, Las Vegas, NV, March 2008.

[Grosicki & El Abed 09] E. Grosicki, H. El Abed: ICDAR 2009 Handwriting Recognition Competition. Int. Conf. on Document Analysis and Recognition (ICDAR) 2009, Barcelona, pp. 139-1402, July 2009.

[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.

[Heigold & Macherey 05+] G. Heigold, W. Macherey, R. Schlüter, H. Ney: Minimum Exact Word Error Training. IEEE ASRU workshop, pp. 186-190, San Juan, Puerto Rico, Nov. 2005.

[Heigold & Schlüter 12+] G. Heigold, R. Schlüter, H. Ney, S. Wiesler: Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance. IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 58-69, Nov. 2012.

[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.

[Ivakhnenko 71] A. G. .Ivakhnenko: Polynomial theory of complex systems. IEEE Transactions on Systems, Man and Cybernetics, Vol. 1, No. 4, pp. 364-378, Oct. 1971.

[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.

[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.

[Makhoul & Schwartz 94] J. Makhoul, R Schwartz: State of the Art in Continuous Speech Recognition. Chapter 14, pp. 165-198, in D. B. Roe, J. G. Wilpon (Editors): Voice Communication Between Humans and Machines. National Academy of Sciences, 1994.

[Miao & Metze 15] Y. Miao. F Metze: On speaker adaptation of long short-term memory recurrent neural networks. Interspeech, Dresden, Germany, 2015.

[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.

[Mohamed & Dahl+ 09] A. Mohamed, G. Dahl, G. Hinton: Deep belief networks for phone recognition. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.

[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.

[Neco & Forcada 97] R. P. Neco, M. L. Forcada: Asynchronous translations with recurrent neural nets. IEEE Int. Conf. on Neural Networks, pp. 2535-2540, June 1997.

[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.

[Ney 84] H. Ney: The Use of a One–Stage Dynamic Programming Algorithm for Connected Word Recognition. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-32, No. 2, pp. 263-271, April 1984.

[Ney & Haeb-Umbach+ 92] H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder: Improvements in Beam Search for 10000-Word Continuous Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, San Francisco, CA, pp. 13-16, March 1992.

[Normandin & Cardin+ 94] Y. Normandin, R. Cardin, R. De Mori: High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation. IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, pp. 299-311, April 1994.

[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.

[Patterson & Womack 66] J. D. Patterson, B. F. Womack: An Adaptive Pattern Classification Scheme. IEEE Trans. on Systems, Science and Cybernetics, Vol.SSC-2, pp.62-67, Aug. 1966.

[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.

[Printz & Olsen 02] H. Printz, P. A. Olsen: Theory and practice of acoustic confusability. Computer Speech and Language, pp. 131–164, Jan. 2002.

[Raissi & Beck+ 16] T. Raissi, E. Beck, R. Schüter, H. Ney: Towards Consistent Hybrid HMM Acoustic Modeling. RWTH Aachen, Aachen, Germany, arxiv, 28-April-2021.

[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.

[Sainath & Weiss+ 16] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani: Factored Spatial and Spectral Multichannel Raw Waveform CLDNNs, Proc. ICASSP, 2016.

[Saon & Tüske+ 2021] G. Saon, Z. Tüske, D. Bolanos, B. Kingsbury: Advancing RNN Transducer Technology for Speech Recognition. IBM Research AI, Yorktown Heights, USA, arxiv, 17-Mar-2021.

[Sakoe & Chiba 71] H. Sakoe, S. Chiba: A Dynamic Programming Approach to Continuous Speech Recognition. Proc. 7th Int. Congr. on Acoustics, Budapest, Hungary, Paper 20 C 13, pp. 65-68, August 1971.

[Sak & Shannon+ 17] H. Sak, M. Shannon, K. Rao, F. Beaufays: Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping. Interspeech, Stockholm, Sweden, pp. 1298-1302, Aug. 2017.

[Schlüter & Nussbaum+ 11] R. Schlüter, M. Nussbaum-Thom, H. Ney: On the Relationship between Bayes Risk and Word Error Rate in ASR. IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 5, p. 1103-1112, July 2011.

[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.

[Schlüter & Scharrenbach+ 05] R. Schlüter, T. Scharrenbach, V. Steinbiss, H. Ney: Bayes Risk Minimization using Metric Loss Functions Interspeech, pages 1449-1452, Lisboa, Portugal, Sep. 2005.

[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.

[Seide & Li+ 11] F. Seide, G. Li, D. Yu: Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. Interspeech, pp. 437-440, Florence, Italy, Aug. 2011.

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

[Stolcke & Grezl+ 06] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, D. Vergyri: Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Toulouse, France, May 2006.

[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.

[Tüske & Plahl+ 11] Z. Tüske, C. Plahl, R. Schlüter: A study on speaker normalized MLP features in LVCSR. Interspeech, pp. 1089-1092, Florence, Italy, Aug. 2011.

[Tüske & Golik+ 14] Z. Tüske, P. Golik, R. Schlúter, H. Ney: Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR. Interspeech, ISCA best student paper award, pp. 890-894, Singapore, Sep. 2014.

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

[Valente & Vepa+ 07] F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, R. Schlüter: Hierarchical Neural Networks Feature Extraction for LVCSR system. Interspeech, pp. 42-45, Antwerp, Belgium, Aug. 2007.

[Vapnik 98] Vapnik: Statistical Learning Theory. Addison-Wesley, 1998.

[Variani & Sainath+ 16] E. Variani, T. N. Sainath, I. Shafran, M. Bacchiani: Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling. Interspeech 2016, San Francisco, CA, pp. 808-812, Sep. 2016.

[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.

[Velichko & Zagoruyko 70] V. M. Velichko, N. G. Zagoruyko: Automatic Recognition of 200 Words. Int. Journal Man-Machine Studies, Vol. 2, pp. 223-234, June 1970.

[Vintsyuk 68] T. K. Vintsyuk: Speech Discrimination by Dynamic Programming. Kibernetika (Cybernetics), Vol. 4, No. 1, pp. 81-88, Jan.-Feb. 1968.

[Vintsyuk 71] T. K. Vintsyuk: Elementwise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. Kibernetika (Cybernetics), Vol. 7, pp. 133-143, March-April 1971.

[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.

[Wang & Alkhouli+ 17] W. Wang, T. Alkhouli, D. Zhu, H. Ney: Hybrid Neural Network Alignment and Lexicon Model in Direct HMM for Statistical Machine Translation. Annual Meeting ACL, pp. 125-131, Vancouver, Canada, Aug. 2017.

[Wang & Zhu+ 18] W. Wang, D. Zhu, T. Alkhouli, Z. Gan, H. Ney: Neural Hidden Markov Model for Machine Translation. Annual Meeting ACL, Melbourne, Australia, July 2018.

[Xu & Povey+ 10] H. Xu, D. Povey, L. Mangu, J. Zhu: Minimum Bayes Risk Decoding and System Combination Based on a Recursion for Edit Distance. Computer Speech and Language, Sep. 2010.

[Zhou & Berger+ 2021] W. Zhou, S. Beger, R. Schlüter, H. Ney:
Phoneme Based Neural Transducer for Large Vocabulary Speech Recognition. RWTH Aachen, Aachen, Germany, arxiv 10-Feb-2021.

[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
REFERENCES
Speech & Language: From Bayes to Deep Learning